

Université Paris-Sud

Mémoire présenté en vue de l'obtention  
du diplôme d'Habilitation à Diriger des Recherches

**Héritabilité**  
**de la régression vers la moyenne au modèle mixte**

Hervé Perdry

2017



# Table des matières

<b>Préambule</b>	<b>1</b>
1. Début en algèbre constructive . . . . .	1
2. Reconversion en génétique . . . . .	1
3. Encadrements . . . . .	4
4. Sur ce mémoire . . . . .	6
 <b>1. Au commencement</b>	 <b>7</b>
1.1. Gregor Mendel . . . . .	7
1.2. Sir Francis Galton . . . . .	13
1.3. Enfin Fisher vint . . . . .	25
 <b>2. Le modèle linéaire mixte</b>	 <b>35</b>
2.1. Introduction . . . . .	35
2.2. Les vraisemblances . . . . .	38
2.3. Tester les composantes de la variance . . . . .	42
2.4. Estimation des composantes de la variance . . . . .	43
2.5. Estimation et prédiction des effets . . . . .	45
2.6. La variance expliquée par les effets fixes . . . . .	46
2.7. L’astuce de la diagonalisation . . . . .	47
 <b>3. L’héritabilité génomique</b>	 <b>49</b>
3.1. À la recherche de l’héritabilité perdue . . . . .	49
3.2. Héritabilité génomique . . . . .	50
3.3. Prise en compte d’une structure de population dans les tests d’association	54
3.4. Performances prédictives . . . . .	55
3.5. De l’interprétation du modèle mixte . . . . .	60
 <b>4. Estimation de l’héritabilité dans une population structurée</b>	 <b>63</b>
4.1. Peut-on corriger le biais induit par la structure de population ? . . . . .	63
4.2. Analyse d’une variable simulée . . . . .	66
4.3. Analyse des variables de l’étude 3C . . . . .	67
4.4. Moralité . . . . .	72
 <b>5. Projets</b>	 <b>75</b>
5.1. L’héritabilité génomique . . . . .	76
5.2. La matrice de corrélation génétique . . . . .	81

## *Table des matières*

5.3. Modèle mixte . . . . .	83
5.4. Le dernier avatar du modèle mixte : la régression sur le LD Score . . . . .	86
5.5. Conclusion . . . . .	87
<b>Bibliographie</b>	<b>87</b>
<b>Annexes</b>	<b>101</b>
<b>A. Détails sur le modèle linéaire mixte</b>	<b>101</b>
<b>B. Sur certaines matrices aléatoires</b>	<b>109</b>

# Préambule

Ce préambule décrit rapidement mon parcours professionnel, mes activités d'encadrement, et avertit lecteurs et lectrices sur le reste du mémoire.

## 1. Débuts en algèbre constructive

Ma formation initiale est celle d'un mathématicien pur ; élève de l'École normale supérieure de Lyon, j'ai préparé ma thèse à l'Université de Franche-Comté sous la direction de Henri Lombardi, et l'ai soutenue en 2001 [1]. J'ai ensuite effectué un séjour post-doctoral à l'Université de Cantabrie en Espagne (2002 à 2004), puis à l'Université de Pise en Italie (2004 à 2006). Mes premiers travaux de recherche portent sur la théorie constructive des corps valués et sur la théorie constructive des anneaux noethériens [1–8]. C'est dans ce dernier domaine que j'ai obtenu à mon goût mon résultat le plus intéressant et le plus original : une version constructive du théorème de la base de Hilbert sans l'hypothèse usuelle de cohérence [7]. La publication de ce résultat est survenue bien après son écriture initiale, alors que j'avais déjà changé de domaine de recherche, et je n'ai pas eu le loisir de lui donner la publicité que j'aurais souhaité et qu'il aurait mérité.

## 2. Reconversion en génétique

Désirant me réorienter vers une recherche appliquée, j'ai rejoint en 2006 l'unité Inserm U535 dirigée par Françoise Clerget. J'y ai naturellement tout d'abord travaillé sur des méthodes d'analyses de données familiales qui étaient au cœur de l'activité de ce laboratoire : avec Françoise Clerget elle-même (inclusion de covariables dans un test d'association basé sur des trios [9] ; développement d'un nouveau test d'association basé sur les paires de germains concordantes [10]) et également avec Catherine Bonaïti-Pellié (estimation de la pénétrance d'une mutation à partir de données familiales, [11] ; développement d'un système de score pour le conseil génétique dans le cancer du sein, [12]). Parallèlement à ceci, j'ai collaboré avec Emmanuelle Génin et Anne-Louise Leutenegger sur les méthodes d'estimation des coefficients de consanguinité et leur utilisation dans les analyses d'association génome entier [13, 14].

## 2.1. Analyse d'association et données familiales

Le début de la thèse de Claire Dandine-Roulland a consisté à développer une méthode d'inférence bayésienne utilisant des paires de germains concordantes pour la maladie.

Ce travail est un prolongement du test d'association développé dans [10]. Il s'agit à mes yeux principalement d'une « preuve de concept », l'objet étant de démontrer l'intérêt des études familiales, alors que l'épidémiologie génétique est dominée aujourd'hui par les études cas/témoins. En effet, quand on considère des individus apparentés, au-delà de l'information génotypique nous disposons d'une information de liaison génétique, au travers notamment du nombre d'allèles partagés « identiques par descendance » (IBD, pour *Identical by Descent*). Nous démontrons que l'utilisation conjointe des données génotypiques et du nombre d'allèles IBD dans les paires de germains concordantes pour la maladie étudiée permet une inférence statistique sur un SNP causal qui n'est observé qu'à travers un SNP marqueur en déséquilibre de liaison avec lui. Ceci serait impossible avec des données cas/témoins, le modèle n'étant pas identifiable. Nous avons fait le choix d'utiliser des méthodes bayésiennes (Metropolis-Hastings) pour l'inférence.

Cette méthode est implémentée dans un package R (publié par Claire Dandine-Roulland sur le CRAN) appelé ASPBay. Un article d'exposition de la méthode est publié dans l'European Journal of Human Genetics [15] : nous l'illustrons sur des données de sclérose en plaques issus d'une étude précédente [16].

## 2.2. Méthodes de régression et arbres de décision

En collaboration avec Cyprien Mbogning et Philippe Broët, dans le cadre du projet Abi-risk, nous avons développé des méthodes de régression linéaire qui incorporent un arbre de décision inféré à partir des données.

Les méthodes basées sur les arbres de décisions sont attractives quand il s'agit de rechercher des interactions complexes entre divers facteurs, et par exemple des facteurs génétiques. Il peut cependant être nécessaire d'incorporer au modèle un terme linéaire qui permette de prendre en compte des facteurs connus, potentiellement confondants (par exemple l'ethnicité).

Une première publication expose une méthode de construction d'un tel modèle, basée sur la construction récursive d'un arbre de grande profondeur et sur l'utilisation d'un critère BIC pour son élagage [17].

Une seconde publication propose de construire un prédicteur par agrégation d'un grand nombre de tels modèles, construits par *bootstrap* à partir du jeu de données initial. Cette procédure a l'avantage d'être plus stable qu'une procédure basée sur un modèle unique, et, de ce fait, a de meilleures performances prédictives [18].

### 2.3. PestiBG : génotoxicité des pesticides

Ce projet a débuté en septembre 2014. Il est financé par l'INCa et porté par Élisabeth Boutet (INRA, Toulouse). Il fédère cinq équipes de recherche multidisciplinaires (épidémiologie, biologie, toxicologie, génotoxicité, cytogénétique et biostatistique) et vise à étudier différents biomarqueurs d'exposition et d'effets dans une cohorte de 200 agriculteurs exposés aux pesticides, ainsi que l'influence du polymorphisme génétique de gènes du métabolisme des xénobiotiques et de la réparation de l'ADN sur la génotoxicité des pesticides.

Plusieurs mesures de génotoxicité ont été effectuées, la principale étant le test des comètes (*comet assay*) réalisé à Toulouse. Nous avons travaillé avec Élisabeth Boutet à la validation d'un protocole de « haut-débit » pour ce test, en utilisant le modèle mixte pour quantifier les sources de variabilités des mesures (article en cours). Des biomarqueurs du stress oxydatif ont également été mesurés par spectrométrie de masse chez ces individus. On dispose de données de génome entier (puce Affymetrix de 600 000 SNP) — d'autres mesures de biomarqueurs sont en cours, mais les premières analyses devraient être réalisées très prochainement.

### 2.4. Modèle mixte et héritabilité

Les modèles mixtes, issus du monde de la génétique animale où ils ont été développés à des fins de sélection des reproducteurs, sont depuis quelques années très en vogue en génétique humaine. Ils ont notamment servi à des tests d'association (tests d'association avec les SNP d'un pathway ou avec un ensemble de variants rares), à des méthodes d'estimation de l'héritabilité (dite héritabilité génomique), et à la prise en compte d'une structure de population dans les études d'association de génome entier.

Claire Dandine-Roulland et moi-même avons écrit un article où nous passons en revue la méthodologie du modèle mixte, ainsi que les applications citées ci-dessus [19]. D'autre part, Claire a utilisé les données de l'Étude des Trois Cités pour calculer l'héritabilité de la stature, de l'indice de masse corporelle, mais aussi de la latitude et de la longitude du lieu de naissance [20]. Des extraits de ces deux articles seront présentés et développés dans les chapitres suivants.

### 2.5. Développement logiciel

Le développement logiciel est une part substantielle de la thèse de Claire Dandine-Roulland et de mon travail personnel. J'ai en particulier écrit *ElstonStewart* [21], une bibliothèque logicielle R qui permet d'utiliser l'algorithme d'Elston-Stewart pour le calcul de fonctions de probabilités dans les pedigrees, et Claire et moi avons écrit ensemble *gaston* [22]. Cette bibliothèque est en très grande partie écrite en C++, elle est dédiée à la manipulation de données génomiques de grande dimension et aux modèles mixtes; elle permet d'estimer les paramètres de modèles mixtes par plusieurs méthodes différentes,

d'estimer des héritabilités (y compris la composante de dominance), et de faire des études d'association de génome entier (traits binaires ou quantitatifs) avec une composante aléatoire permettant la prise en compte d'une structure de population.

J'ai créé sur `github` la librairie `gaston.utils` qui propose plusieurs fonctions qui seront dans le futur intégrées à `gaston`, notamment pour réaliser des tests d'association avec des données imputées sous formes de « dosages ». D'autres développements basés sur `gaston` sont en cours ; leur description sera esquissée dans la section suivante. Les mises à jour de `gaston` et de `gaston.utils` ont occupé depuis un an la plus grande part du temps que je n'ai pas consacré à l'enseignement et à l'encadrement.

## **3. Encadrements**

### **3.1. Stages de Master**

J'ai encadré une dizaine de stages de M1 et M2 ; il est rare dans notre discipline de pouvoir écrire et soumettre un article dans un laps de temps si court, et ça ne m'est pas encore arrivé, mais je ne perds pas espoir. Le stage de M1 d'Ozvan Bocher, qui a travaillé (au deuxième semestre 2016-17) sur les tests d'association pour variants rares, devrait en effet aboutir prochainement à la soumission d'un article. D'autre part, Ozvan et moi avons commencé l'écriture d'une bibliothèque pour l'analyse des variants rares, basée sur `gaston`.

### **3.2. Thèses doctorales**

#### **Thèse de Claire Dandine-Roulland**

Le stage de M2 de Claire Dandine-Roulland a cependant été valorisé par un article, écrit et publié au début de sa thèse [15]. Pendant sa thèse, Claire a écrit et publié deux autres articles sur le modèle mixte et les estimations de l'héritabilité, que j'ai évoqués plus haut [19, 20]. Claire a soutenu sa thèse en 2016 [23]. Son travail sera largement cité dans la suite de ce mémoire, aussi je n'en dirai pas davantage ici.

#### **Thèse de Jacqueline Milet**

Audrey Sabbagh et moi co-encadrons depuis octobre 2015 la thèse de Jacqueline Milet sur la susceptibilité génétique au paludisme simple. Jacqueline est ingénieure IRD et a une solide expérience de l'épidémiologie du paludisme et de l'analyse de données génétiques, qu'elle a naturellement souhaité valoriser par une thèse de doctorat. Ce travail repose sur l'analyse de données issues de deux cohortes de nouveaux-nés qui ont été suivies de la naissance à 18-24 mois, dans deux projets distincts menés au Bénin par l'IRD. Ces deux cohortes bien suivies sur le plan parasitologique et clinique présentent un intérêt tout particulier pour



la recherche de facteurs génétiques impliqués dans la susceptibilité au paludisme simple (par opposition au paludisme sévère qui a été beaucoup plus souvent l'objet des recherches en épidémiologie génétique). Un total d'un peu moins de 900 individus ont été génotypés par une puce Illumina (2,6 millions de SNP après contrôle qualité). Jacqueline a maintenant terminé un important travail de contrôle qualité et d'analyse par une stratégie en deux temps :

1. choix par une démarche stepwise forward des variables épidémiologiques à intégrer à un modèle (niveau d'exposition, niveau socio-économique, centre de recrutement, etc).
2. analyse par un modèle mixte des résidus du modèle construit à l'étape 1.

Pour la première étape, deux modèles ont été envisagés : une régression binomiale négative sur le nombre total d'accès palustres ; un modèle de Cox pour la modélisation d'événements récurrents avec effet individuel aléatoire (c'est le prédicteur de cet effet individuel qui est pris comme résidu à analyser à l'étape suivante). L'intérêt de cette stratégie en deux étapes est le gain de temps : la réalisation de l'étude de génome entier par les modèles non linéaires utilisés à la première étape prendrait plusieurs semaines. Des simulations ont été réalisées pour évaluer l'erreur de type I et la puissance de cette méthode.

Les deux cohortes sont étudiées séparément, dans une stratégie de type découverte / réplication. Initialement, nous avons décidé de n'analyser que les accès palustres. Il a été tardivement décidé d'ajouter un second jeu d'analyses, intégrant l'ensemble des infections (y compris donc les infections asymptomatiques, décelées par un test de « goutte épaisse »). Cela fait donc quatre analyses au total, selon le modèle considéré à l'étape 1 et selon le type d'accès palustre considéré.

Malgré la faiblesse relative des effectifs, des pics d'association qui dépassent (après imputation) le seuil de significativité de  $5 \cdot 10^{-8}$  ont été découverts, dans des gènes dont la fonction biologique laisse à penser qu'ils peuvent être impliqués dans la physiopathologie de la maladie. Un article est en court d'écriture.

Parallèlement à ce travail, Jacqueline a commencé une revue de la littérature qui porte sur les associations génétiques avec les différentes formes de paludisme (accès graves, accès simples, infections asymptomatiques). L'objectif en est de tester l'hypothèse selon laquelle les gènes déjà connus comme associés au paludisme simple sont ciblés par la sélection naturelle, comme c'est le cas pour le paludisme sévère.

### Étudiant en alternance

Anne-Louise Leutenegger et moi-même encadrons depuis septembre 2016 Isuru Haupe, un étudiant en alternance de DUT d'informatique à l'IUT de Villeteuse. Sa mission est l'écriture d'une bibliothèque logicielle basée sur *gaston* qui offre dans l'environnement de R les fonctionnalités de *FSuite* [24], un *pipeline* basé entre autres sur le logiciel *FESTim* [25]. La bibliothèque terminée permettra donc d'ajouter aux fonctionnalités offertes par *gaston* celles nécessaires à l'estimation de coefficients de consanguinité et à l'identifica-

tion de régions génomiques homozygotes « par descendance » (en anglais, *homozygous by descent*, ou HBD) chez les individus consanguins, et donc à la recherche de gènes impliqués dans les maladies mendéliennes rares par *homozygosity mapping* ou à la recherche de sous-entités mendéliennes dans les maladies multifactorielles par la stratégie HBD-GWAS [13].

## 4. Sur ce mémoire

J'ai choisi d'axer ce mémoire sur la question de l'héritabilité dite « génomique ». La plus grande part de ce qui est présenté provient de deux articles écrits avec Claire Dandine-Roulland [19, 20].

J'ai cru opportun d'y ajouter un préliminaire historique sur les théories mathématiques de l'hérédité – et je crains de ne m'être un peu laissé emporter par mon sujet. Le lecteur et la lectrice trouveront peut-être que j'exagère en leur faisant subir cinq pages sur Gregor Mendel ; mais bien que les résultats scientifiques de celui-ci soient connus de tous, il m'a semblé utile d'insister sur la pertinence de sa démarche et de ses interprétations. La figure de son contemporain Galton est également familière, mais la théorie qu'il a développée l'est moins – j'avoue avoir trouvé dans ses œuvres tout autre chose que ce à quoi je m'attendais. Enfin, présenter le modèle polygénique de Fisher était le préliminaire rêvé à la question de l'héritabilité génomique.

Le second chapitre, largement extrait d'un article de Claire [19], passe en revue la théorie du modèle mixte, qui est celle utilisée pour les calculs d'héritabilité génomique. Le troisième chapitre, extrait du même article, présente son application à l'estimation de l'héritabilité. Le quatrième chapitre, issu d'un second article [20], est une application à des données issues de l'Étude des Trois Cités. Dans ces deux chapitres j'ai essayé de questionner la pertinence du modèle.

Le chapitre final présente quelques projets futurs, toujours en se concentrant sur la question de l'héritabilité et des modèles mixtes.

# Chapitre 1.

## Au commencement

J’ai tenté de présenter ici les aspects historiques des théories de l’hérédité, et plus précisément le développement des théories mathématiques – je ne suis pas historien, et je n’ai pas fait œuvre d’historien. Certaines approximations ou lacunes sont dues à des choix délibérés de simplification et d’autres à mon ignorance ou au manque de méthode adaptée à ce travail inhabituel pour moi, qui s’est révélé bien plus ardu et plus long que ce que j’envisageais en le commençant.

On n’évoquera pas ici l’histoire connexe qui est celle des théories transformistes, de Darwin et de la théorie de la descendance avec modification, et encore moins les conceptions avancées par divers « précurseurs » fussent-ils aussi illustres qu’Aristote ou Maupertuis [26], qui ne se sont guère prêtées à la mathématisation. Nous nous contenterons de présenter de notre mieux les contributions de Mendel, de Galton et Pearson, puis de Fisher.

### 1.1. Gregor Mendel

#### 1.1.1. Jeunesse et formation

Gregor Mendel est né en 1822 à Hynčice (Heizenberg en allemand) [27], dans l’actuelle République Tchèque, qui faisait alors partie de l’Empire d’Autriche. Ses parents étaient des paysans germanophones sans grande fortune. Il reçut cependant une éducation complète, d’abord grâce aux sacrifices financiers de ses parents, puis de ses sœurs, et enfin grâce au soutien du monastère de Brno (en allemand, Brünn) où il entra en 1843 afin d’y trouver une stabilité matérielle. Il y bénéficia d’une assez grande liberté : on lui donna d’abord l’opportunité d’enseigner le latin, le grec, et les mathématiques au lycée de Znojmo, et à partir de 1851 de compléter ces études en suivant pendant deux ans des cours à l’université de Vienne. Il y suivit notamment les cours de physique dispensés par Christian Doppler et les cours de botanique de Franz Unger.

Ces détails biographiques montrent que si Mendel était bien, conformément à l’image qu’on en a le plus souvent aujourd’hui, un scientifique amateur dans la mesure où la recherche scientifique n’était pas son métier, sa formation est en revanche celle d’un professionnel. Il était préparé à utiliser la méthode expérimentale, et Franz Unger l’avait familia-

risé avec le problème de l'hybridation des végétaux. Il connaissait les travaux précurseurs d'hybrideurs (Kölreuter et Gärtner, entre autres) qui avaient observé la réapparition de caractères ancestraux dans la descendance des hybrides. Quand il revient au monastère de Brno en 1853 après avoir terminé ses études, c'est cette question qu'il est décidé à étudier de façon scientifique.

### 1.1.2. Recherches sur l'hybridation des plantes

Mendel choisit de procéder à des expériences sur le pois (*Pisum sativa*, il s'agit bien des petits pois que nous mangeons), dont la reproduction est facile à contrôler et qui produit des hybrides fertiles. Ses travaux seront communiqués à la Société des Sciences naturelles de Brno en 1865, et publiés dans les actes de cette société, sous le titre *Recherches sur l'hybridations des plantes*, en allemand : *Versuche über Pflanzen-Hybriden* [28] (traduction anglaise dans [29]).

Nous allons donner ici un rapide aperçu du contenu de ce mémoire. Mendel s'y intéresse presque exclusivement à des caractères discontinus (couleur ou forme des gousses, etc, en tout sept caractères), dont il souligne l'absence de formes intermédiaires.

#### Cas d'un caractère

La première génération d'hybrides (génération F1 en terminologie moderne) entre deux plantes issues de lignées pures qui diffèrent pour un certain caractère ne présente qu'un seul des deux caractères ancestraux, qu'il nomme le caractère *dominant* : par exemple, un hybride entre des plantes à gousses vertes et des plantes à gousses jaunes aura des gousses vertes – ce caractère est donc le caractère dominant. L'autre caractère, ici la couleur jaune, est le caractère *récessif*.

Le pois est une plante autogame, c'est à dire qu'un individu peut se reproduire avec lui-même. Mendel laisse les hybrides se reproduire par autogamie, et observe à la génération suivante (la génération F2) la réapparition du caractère récessif ; il observe que les individus présentant les caractères récessifs et dominants sont en proportion 1 : 3 (une plante récessive pour trois plantes dominantes, cf figure 1.1).

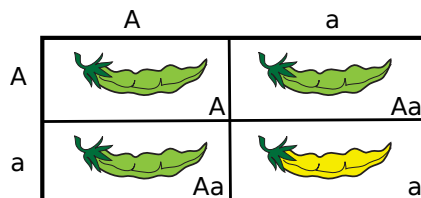


FIGURE 1.1. – Descendance d'un hybride. Les lettres en marge représentent les gamètes parentaux.

Mendel poursuit l'expérience, toujours en laissant les plantes se reproduire par autogamie. Il observe ainsi que dans la descendance d'un individu de la génération F2 présentant le caractère récessif, on n'observe plus que ce caractère ; et il observe également qu'un tiers des plantes F2 présentant le caractère dominant ont une descendance exclusivement dominante, tandis que les deux-tiers restant ont, comme l'hybride F1, une descendance où les deux caractères s'observent en proportion 1 : 3.

Mendel comprend que la proportion observée est en fait une proportion 1 : 2 : 1 de formes qu'il note A, Aa et a : A et a sont les caractères « constants » dominants et récessifs, et Aa est la forme hybride où les deux caractères sont présents, le caractère dominant étant seul visible ; il choisit de résumer ces proportions par l'expression formelle

$$A + 2 Aa + a.$$

Il montre ensuite que ce principe, aujourd'hui connu sous le nom de loi de ségrégation des caractères, permet de calculer les proportions attendues des trois types de plantes aux générations suivantes. La proportion de plantes Aa tend rapidement vers 0, ce qui est conforme aux constatations faites par ses prédécesseurs sur la réapparition des formes ancestrales dans la descendance des hybrides.

### Cas de plusieurs caractères

Mendel poursuit ses expériences en hybridant des plantes qui diffèrent par plusieurs caractères, un des parents portant des caractères A, B, etc, et l'autre des caractères a, b, etc. Il obtient à la génération F1 des plantes « dihybrides » (de type AaBb) qui ne présentent que des caractères dominants ; à la génération F2 il obtient les 4 combinaisons possibles de caractères dans les proportions 1 : 3 : 3 : 9 (soit une plante sur 16 présentant les deux caractères récessifs, 3 plantes récessives pour le premier caractère et dominantes pour le second, etc).












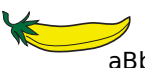




	AB	Ab	aB	ab
AB	 AB	 ABb	 AaB	 AaBb
Ab	 ABb	 Ab	 AaBb	 Aab
aB	 AaB	 AaBb	 aB	 aBb
ab	 AaBb	 Aab	 aBb	 ab

FIGURE 1.2. – Descendance d'un dihybride.

Mendel explique cette répartition en remarquant comme auparavant que certaines des plantes qui présentent un caractère dominant sont en fait des hybrides : elles portent le caractère dominant et le caractère récessif, seul le caractère dominant est observé. Les proportions de chacun des types sont obtenues en combinant formellement les deux expressions  $A + 2 Aa + a$  et  $B + 2 Bb + b$  pour obtenir

$$AB + 2 ABb + Ab + 2 AaB + 4 AaBb + 2 Aab + aB + 2 aBb + ab.$$

La figure 1.2, où les capitales A et B correspondent aux caractères dominants « gousse verte » et « gousse gonflée », associés aux caractères récessifs « gousse jaune » et « gousse étranglée », illustre le raisonnement. Cette figure peut s'obtenir en éclatant les quatre cases de la figure 1.1, où seule la couleur de la gousse est importante (le caractère A/a) en quatre cases où on fait varier la forme de la gousse (le caractère B/b) selon un motif analogue.

Ce résultat est connu sous le nom de loi de ségrégation indépendante des caractères.

### Mécanisme proposé

Mendel propose un mécanisme simple pour expliquer ses résultats : les cellules reproductrices émises par les plantes (en termes modernes, les gamètes ; Mendel utilise les mots *Keimzellen* et *Pollenzellen*, cellules germinales et cellules du pollen) ont un caractère A ou a ; et les cellules reproductrices des hybrides présentent toutes les combinaisons possibles de caractères ancestraux dans des proportions égales.

Le cas le plus simple est celui des hybrides Aa qui émettent des gamètes A et a dans les proportions 1 : 1 et l'appariement aléatoire des gamètes fait le reste. Mendel illustre ce mécanisme par l'expression

$$\frac{A}{A} + \frac{A}{a} + \frac{a}{A} + \frac{a}{a} = A + 2 Aa + a,$$

où le caractère au-dessus du trait de fraction est celui transmis par la plante mâle, celui qui est en-dessous est transmis par la plante femelle. Le même mécanisme explique ce qui est constaté pour les dihybrides, avec quatre types gamétiques AB, Ab, aB et ab (voir les marges des figures 1.1 et 1.2).

Mendel teste cette hypothèse avec succès par une expérience appelée aujourd'hui « rétrocroisement » (ou plus souvent, en anglais, *backcross*) : il s'agit simplement de croiser une plante hybride avec un de ses parents (ou une autre plante de type pur, qui n'émet des gamètes que d'un seul type). Ainsi, le croisement d'une plante Aa croisée avec une plante a produit, selon l'hypothèse de Mendel, des plantes Aa et a (qui présentent respectivement les caractères dominant et récessif) dans les proportions 1 : 1.

De même, le croisement d'une plante AaBb avec une plante ab produit des plantes AaBb, aBb, Aab et ab dans les proportions 1 : 1 : 1 : 1. Ici encore, ces quatre types de plantes sont reconnaissables par l'observation des caractères présentés.

## Caractères continus

Ces résultats concernent des caractères discontinus, sans forme intermédiaire. Mendel rapporte également des expériences sur les haricots (*Phaseolus*), pour lesquels il a notamment considéré la couleur des fleurs. Il est quelque peu dérouté par les résultats obtenus : il observe en effet un continuum de variations (du blanc au violet) et de trop rares retours à la forme récessive ou supposée telle (le blanc).

Il note pourtant :

« Même ces résultats énigmatiques, cependant, peuvent probablement s'expliquer par les lois qui régissent *Pisum* si nous supposons que la couleur des fleurs et des graines de *Ph. multiflorus* est la combinaison de deux couleurs entièrement indépendantes ou plus, qui se comportent individuellement comme n'importe quel autre caractère constant de la plante. Si la couleur de fleur A est la résultante des caractères  $A_1 + A_2 + \dots$ , qui produit la couleur violette, alors par hybridation avec le caractère distinct de couleur blanche  $a$ , on obtient un individu hybride  $A_1a + A_2a + \dots$  (...). Selon les hypothèses précédentes, ces caractères hybrides sont indépendants et vont par conséquent se développer de façon indépendante. On voit alors facilement que la combinaison de suites de caractères de cette sorte produirait une suite complète de couleurs. Si par exemple,  $A = A_1 + A_2$ , alors aux hybrides  $A_1a$  et  $A_2a$  correspondent les séries

$$A_1 + 2 A_1a + a$$

$$A_2 + 2 A_2a + a$$

dont les membres se combinent de neuf façons différentes, chacune désignant une couleur différente :

1 $A_1A_2$	2 $A_1aA_2$	1 $A_2a$
2 $A_1A_2a$	4 $A_1aA_2a$	2 $A_2aa$
1 $A_1a$	2 $A_1aa$	1 $aa$

Les nombres qui précèdent chaque combinaison indiquent combien de plantes de la couleur correspondante font partie de la série. Le total étant de 16, toutes les couleurs doivent en moyenne apparaître parmi 16 plantes, mais, on le voit, en proportions inégales. »

Il est difficile de ne pas voir, dans ce court passage, une esquisse du modèle polygénique : il aurait suffi que Mendel assigne aux différentes combinaisons énumérées ci-dessus une nuance dépendant du nombre d'allèles dominants pour l'avoir complètement formulé. C'est ce modèle que Fisher développera en 1918 — plus d'un demi-siècle plus tard.

### 1.1.3. Postérité

Le mémoire de Mendel est passionnant à lire, et je ne peux qu'inviter le lecteur à s'y référer pour compléter ce résumé. Il est étonnant de clarté et modernité ; cela s'explique en partie par le fait que le modèle proposé par Mendel, qui correspond à une certaine réalité biologique, nous est familier puisqu'il est utilisé et enseigné de nos jours. Dans le texte qui précède, il n'y a presque qu'un changement qui serait fait par un biologiste moderne : c'est l'utilisation de la notation AA au lieu de A, pour les plantes ne portant que le caractère A. Il n'en reste pas moins que Mendel a les idées admirablement claires.

Les travaux de Mendel sont passés inaperçus de son vivant. Peut-être qu'une partie de l'explication réside dans le fait qu'ils étaient présentés comme des travaux sur l'hybridation des plantes, et non sur les lois de l'hérédité, sujet qui intéressait plus particulièrement ceux qui l'auraient lu avec le plus d'intérêt ; mais surtout, Mendel étant devenu en 1868 le père supérieur de son couvent, il fut absorbé par cette tâche et n'eut guère le loisir de donner davantage de publicité à sa théorie. Il mourut en janvier 1884 d'une insuffisance rénale, et ses travaux ne furent redécouverts qu'en 1900 par Hugo de Vries, Karl Correns et Erich von Tschermak qui réalisaient des expériences similaires. Mendel n'émet aucune hypothèse sur la nature matérielle des caractères transmis ; l'hypothèse que les chromosomes en constituaient le support physique fut vite émise, et c'est à Thomas Morgan, qui établit la première carte génétique jamais réalisée – celle du génome de la drosophile –, qu'on doit la confirmation expérimentale de ce fait.

La loi de ségrégation indépendante des caractères est fausse en générale : elle n'est vraie que pour des caractères *non liés* – c'est le cas si les gènes correspondant à ces caractères sont sur des chromosomes distincts. Dans le cas contraire, la loi reste approximativement vraie si les gènes sont suffisamment éloignés les uns des autres. C'est globalement le cas des sept caractères étudiés par Mendel, à l'exception de deux (la longueur de la tige et la forme des gousses) [30], pour lesquels il est possible qu'il n'ait pas réalisé d'expérience avec des doubles hybrides.

La note finale est plus négative : en 1936, Ronald Fisher réanalyse les résultats de Mendel [31], et montre que les proportions observées par Mendel dévient trop peu de celles prédites par la théorie par rapport aux déviations aléatoires attendues. L'accumulation de déviations trop petites, d'expérience en expérience, est accablante. Voici donc Gregor Mendel sous le coup de l'infamante accusation d'avoir manipulé ses données. Il faut cependant nuancer un peu. Mendel n'avait aucune idée des ordres de grandeurs des déviations attendues ; dépourvu de l'outil mathématique (la statistique du  $\chi^2$ ) nécessaire à leur analyse, il a pu écarter de bonne foi de ses hybrides F2 des échantillons à ses yeux suspects de contamination par une fertilisation extérieure ; il a également pu réaliser plusieurs expériences et choisir de ne rapporter que celle qui correspondait le mieux à la théorie, hypothèse qui fournit des déviations en accord avec celles qui sont observées [32]. Les exigences de rigueur expérimentale en biologie ne pouvaient pas en 1850 être ce qu'elles sont de nos jours, et cette erreur ne peut être jugée avec la sévérité qui serait de mise à présent. Il faut noter en outre que ces critiques ne portent que sur la collecte ou le traitement des données issues des



expériences ; la conception des expériences est impeccable, et les rétrocroisements restent un des outils des expérimentateurs en génétique animale ou végétale.

## 1.2. Sir Francis Galton

### 1.2.1. Détails biographiques

Francis Galton est né en 1822 à Birmingham dans une famille aisée ; c'est un enfant précoce et brillant. Son père le destine à la médecine, mais les études médicales lui déplaisent et il s'oriente vers les mathématiques ; il fit en troisième année d'étude à Cambridge une dépression sévère, liée à des résultats moins bons qu'espérés dans cette matière [33]. Il se contentera de passer un examen appelé *poll degree*, sans briguer les « honneurs » qui lui auraient permis de poursuivre cette carrière. Il se tourne alors à nouveau brièvement vers la médecine, jusqu'à la mort de son père en 1844, qui le rend financièrement indépendant. Il voyage en Afrique et au Moyen-Orient, tout d'abord sans but scientifique, puis sous le patronage de la Société royale de géographie, qui lui décernera en 1852 une médaille d'or pour ses relevés cartographiques d'Afrique du Sud. À partir de ce moment, il devient un auteur prolifique, publiant chaque année plusieurs articles et ouvrages sur une foule de sujets : météorologie, avalanches, voyages...

Comme beaucoup de ses contemporains, il est passionné par l'ouvrage de Charles Darwin (qui se trouve être son cousin\*) publié en 1859, *L'Origine des espèces*. Il consacre dès lors une part croissante de son activité à la réflexion sur les lois de l'hérédité.

### 1.2.2. Les lois de l'hérédité

Dès ses voyages de jeunesse, Francis Galton manifeste un attrait pour l'anthropométrie et la biométrie naissante, attrait lié à ce qui semble être une obsession pour les chiffres. Il rapporte ainsi avoir pris à distance les mensurations de femmes hottentotes à l'aide d'un sextant (lettre à Darwin citée dans [33]), ou avoir tenté d'établir une carte de beauté des îles britanniques, en notant (sur une échelle à trois degrés) toutes les femmes qu'il croisait [34].

Il réalisera une série d'expériences pour tester la théorie de la pangénèse de Darwin, qui postulait que tous les organes émettent des *gemmules* qui s'agrègent entre elles avant d'être transmises à la descendance. Galton transfuse des lapins gris avec du sang de lapins blancs, dans l'espoir que la descendance des lapins gris présente des traits hybrides ; l'expérience n'est pas concluante [35] : le sang des lapins blancs ne transporte pas de gemmules.

Après cet échec, Galton élaborera sa propre théorie de l'hérédité [36–39], formulant tout d'abord une théorie biologique, avant de se concentrer sur la recherche d'une loi mathé-

---

\*ou plus précisément son demi-cousin, la mère de Francis Galton, Frances Darwin, étant la demi-sœur de Robert Darwin, le père de Charles Darwin. Leur grand-père commun, Erasmus Darwin, est un médecin et naturaliste célèbre.

matique de l'hérédité. Galton veut découvrir une loi qui explique aussi bien l'hérédité des traits continus (comme la stature) que celle des traits discontinus (la couleur des yeux), et en particulier pour ces derniers l'*atavisme*, c'est-à-dire la réapparition de caractères ancestraux.

### Une théorie de l'hérédité : le *stirp*

Dans *Une théorie de l'hérédité* [38], article publié en 1875, Galton propose d'appeler *stirp* (du latin *stirpes*, racine), l'ensemble des germes présents dans l'œuf fertilisé et qui sont à l'origine du développement de l'organisme. Il isole quatre postulats qui lui semblent nécessaires à une théorie organique de l'hérédité :

1. l'organisme est la juxtaposition d'un grand nombre d'unités quasi-indépendantes, qui dérivent de germes distincts ;
2. le *stirp* contient une multitude de germes, bien plus nombreux et divers que les unités organiques qui en seront dérivées, de sorte que très peu de ces germes sont finalement développés ;
3. les germes qui ne sont pas développés conservent leur vitalité et contribuent à la formation du *stirp* de la descendance de l'individu ;
4. la structure de l'organisme découle des affinités mutuelles des germes, au sein du *stirp* et au cours du développement.

Pour résumer en termes modernes la théorie développée par Galton, on peut voir le *stirp* comme une population de cellules souches, dont une partie (aléatoire) donnera naissance aux différents organes et tissus de l'organisme ; les cellules restantes se multiplient et sont transmises à la génération suivante. Galton hésite à exclure tout à fait la possibilité d'une transmission des caractères acquis, mais il ne lui concède qu'un rôle au mieux marginal, quelques cellules provenant du reste du corps pouvant réintégrer le *stirp* de façon exceptionnelle.

Ce modèle a beau être biologiquement erroné, il a de bonnes propriétés et permet à Galton de formuler des idées pertinentes. Le troisième postulat a été considéré, avec sans doute assez de justesse, comme précurseur de la théorie de la lignée germinale de Weismann [40,41]. Le modèle développé permet notamment à Galton d'insister sur le rôle du hasard dans le processus de la reproduction, et d'en faire la cause des différences entre les membres d'une fratrie. La présence dans le *stirp* de matériel qui ne se développe pas mais est transmis à la descendance permet d'expliquer l'*atavisme*. Galton attribue l'avantage de la reproduction sexuée à ce qu'elle permet de renouveler, dans le *stirp*, les cellules qui y seraient mortes, explication qui préfigure le « cliquet de Muller » (où le raisonnement porte sur le remplacement des gènes portant des mutations délétères) [42].

Cette conception de l'hérédité influencera fortement (sans qu'il en soit fait explicitement mention) la formulation, en 1897, de la loi mathématique de l'hérédité.

## Natural Inheritance

Dans *Natural Inheritance* [43], ouvrage publié en 1889, Galton reprend et développe des travaux publiés antérieurement sous forme d'articles [44–46]. Il y analyse deux jeux de données, l'un portant sur la stature des membres d'une famille, l'autre sur la couleur des yeux.

**La stature** Francis Galton a collecté des données en faisant circuler des formulaires parmi ses correspondants ou en offrant des prix aux contributeurs. Il dispose ainsi des statures de 928 enfants répartis en 205 fratries, et de la stature de leurs parents. Il corrige la différence de stature entre hommes et femmes en multipliant la stature des femmes par 1,08 ; en prenant la moyenne de la stature des deux parents, il obtient la stature du « parent-moyen » (*mid-parent*) qu'il compare à la stature des enfants du couple. Le résultat est reproduit dans la table suivante [43, 45].

TABLE I.  
NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.  
(All Female heights have been multiplied by 1.08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72.5	..	..	..	..	..	..	1	2	1	2	7	2	4	..	19	6	72.2
71.5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68.2
67.5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67.6
66.5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67.2
65.5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66.7
64.5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65.8
Below ..	1	..	2	4	1	2	1	1	..	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	..	..	..	..	..

TABLE 1.1. – Table de contingence des statures des enfants et de leurs « mid-parents »

Galton estime l'écart-type de la stature dans la population, il est de 1,7 pouce ; l'écart-type de M est estimé à 1,19 pouce, ce qui est cohérent avec un écart-type théorique (en absence d'homogamie ou d'hétérogamie) de  $\frac{1}{\sqrt{2}}1,7 = 1,21$  pouce ; et il estime l'écart-type dans les fratries à 1,5 pouce.

À partir de la table 1.1, Galton met en évidence ce qu'il appelle « régression vers la médiocrité » (*regression towards mediocrity*), qu'on traduira plus volontiers par « régression vers la moyenne » : l'écart entre la stature Y d'un individu et la stature moyenne  $\mu$  tend à être moins important que celui qu'on observe entre la stature M de son parent-moyen et  $\mu$ .

Plus précisément, on a approximativement (les notations sont de nous) \*

$$E(Y - \mu|M) = \frac{2}{3}(M - \mu).$$

Il s'agit de l'espérance de Y conditionnellement à M (dans le texte : la déviation filiale vaut *en moyenne* seulement les deux-tiers de la déviation du parent-moyen). Toujours à partir de cette table, Galton estime également la « régression de la stature du parent-moyen » :

$$E(M - \mu|Y) = \frac{1}{3}(Y - \mu).$$

Il faut dire un mot de la façon dont Galton estime ces coefficients : pour la régression de la stature des enfants, il répartit les parents-moyens en catégories (la stature est arrondie au pouce le plus proche), puis il calcule la moyenne des statures de tous les enfants d'une catégorie ; on reporte les quantités obtenues sur un graphe (figure 1.3), et, les points étant à peu près alignés, on y fait passer « au jugé » une droite dont on détermine ensuite la pente.

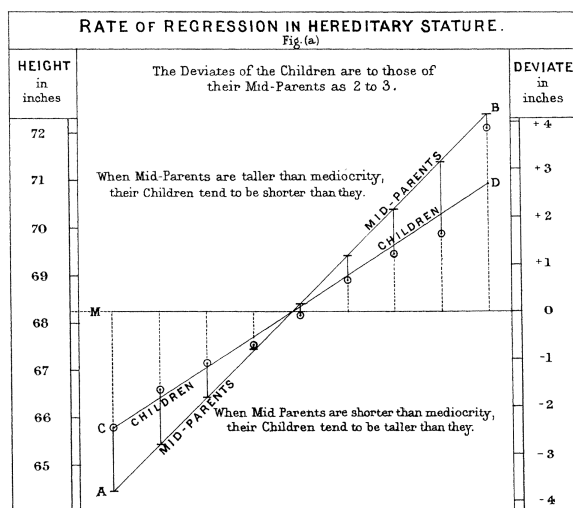


FIGURE 1.3. – La droite de régression (figure IX de [45])

On peut dès à présent livrer une interprétation moderne des résultats obtenus : la loi jointe du vecteur  $(Y - \mu, M - \mu)$  est à peu près une loi normale bivariée centrée, de matrice de variance

$$1,7^2 \times \begin{pmatrix} 1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}.$$

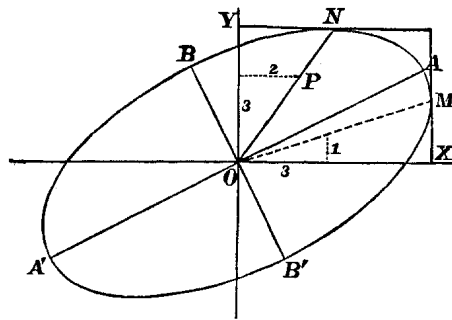
\*Pour ce coefficient de  $\frac{2}{3}$ , Galton dit avoir tout d'abord fait une estimation de  $\frac{3}{5}$ , mais avoir préféré  $\frac{2}{3}$  qui est plus simple. Est-ce parce qu'il recherche une loi naturelle qu'il pense devoir être parcimonieuse ?

Il est facile de vérifier que pour une telle loi, la variance de  $Y$  conditionnellement à une valeur donnée de  $M$  est

$$\text{var}(Y) - \left(\frac{2}{3}\right)^2 \text{var}(M) = \frac{14}{18} \times 1,7^2 \simeq 1,5^2,$$

ce qui est parfaitement cohérent (l'erreur est inférieure à un millième de pouce !) avec la valeur de l'écart-type à l'intérieur des fratries calculé par Galton.

Voici comment celui-ci aborde cette question en 1886. Tout d'abord, par un procédé de lissage des données de la table 1.1, il estime que les valeurs égales dans celle-ci sont disposées sur des ellipses concentriques, dont le centre est à la stature moyenne (soit 68,25 pouces), et telles que les tangentes horizontales à ces ellipses passent par des points  $N$  (cf figure 1.4) situés le long d'une droite de pente  $\frac{2}{3}$ , et les tangentes verticales par des points  $M$  situés le long d'une droite de pente  $\frac{1}{3}$  (ces pentes correspondent aux coefficients de la régression). Ces ellipses sont donc les courbes de niveau de la densité de la loi jointe de  $Y$  et  $M$ .



**FIGURE 1.4.** – Les ellipses de niveau déduites de la table 1.1 (figure 11 dans [43])

Pour vérifier la cohérence de ces résultats, il soumet alors à James Dickson, professeur à Cambridge, un problème que nous pouvons reformuler (et simplifier quelque peu) ainsi : si la loi de  $M$  est normale, d'espérance  $\mu$  et d'écart-type 1,22 (cette valeur correspond à une estimation préliminaire de l'écart-type plus tard estimé à 1,19 ou 1,20, nous dit-il en note), et que la loi de  $(Y - \mu)$  conditionnellement à  $(M - \mu)$  est normale d'espérance  $\frac{2}{3}(M - \mu)$  et d'écart-type 1,5, quelle est la loi de  $(Y - \mu)$  ? Quelle est l'espérance de  $(M - \mu)$  conditionnellement à  $(Y - \mu)$  ?

Dickson répond : on a

$$E(M - \mu | Y - \mu) = 0,34(Y - \mu),$$

et l'écart-type résiduel est environ 1,07. Les calculs de Dickson sont reproduits en appendice dans [43,44]. Il donne en outre des formules générales pour calculer ces valeurs, ainsi qu'une formule pour calculer la variance totale de  $Y$  – mais il ne fait pas l'application numérique qui aurait produit la valeur de  $1,71^2$ .

Galton est aux anges : « je n'avais jamais ressenti un tel sentiment de loyauté et de respect envers la souveraineté et la vaste emprise de l'analyse mathématique que quand sa

réponse est arrivée ». Les calculs de Dickson, quoique courts, ne sont pas faciles à suivre et Galton ne semble pas très à l'aise : il ne prend ni la peine de les refaire avec ses nouvelles valeurs pour l'écart-type de M (qui donnent pourtant des résultats encore plus proches de ses estimations), ni même de faire l'application numérique pour le calcul de la variance de Y, dont la concordance avec ses estimations aurait pourtant été pour lui une source de ravissement supplémentaire. Des résultats généraux sur les variables corrélées avaient déjà été énoncés par Bravais [47] ; il faudra attendre quelques années pour que Yule, Edgeworth, et bien sûr Pearson (cf par exemple [48]) reprennent et étendent ces résultats, posant des bases mathématiques solides au calcul des coefficients de régression.

**Contribution de chaque ancêtre** À partir des constatations faites ci-dessus, Galton veut estimer la contribution de chaque ancêtre à la stature de l'individu. En effet, pour Galton, il faut démêler dans ce qui précède ce qui est réellement imputable aux parents, de ce qui est imputable à une partie du stirp transmis par les parents mais provenant d'ancêtres plus lointains, et qu'on n'observe qu'indirectement chez les parents\*. Voici comment il procède (je suis ci-après le texte d'assez près sans toutefois toujours le traduire littéralement, cf [43] pp 134–136) :

« Si un parent-moyen a un écart de D à la stature moyenne, ses enfants ont un écart, en moyenne, de  $\frac{2}{3}D$ , ceci indépendamment de la contribution des ancêtres plus éloignés. D'autre part, un écart de D chez un individu implique un écart  $D' = \frac{1}{3}D$  chez son parent-moyen, qui lui même implique un écart de  $\frac{1}{3}D' = \frac{1}{9}D$  chez le parent-moyen de ce parent-moyen, c'est-à-dire chez le grand-parent-moyen de l'individu considéré ; soit, dans la totalité de la lignée de D, des écarts dont la somme est  $D + \frac{1}{3}D + \frac{1}{9}D + \dots = \frac{3}{2}D$ .

Si on suppose que la contribution de chaque ancêtre est « taxée » également, pour qu'une accumulation de contributions ancestrales dont la somme est  $\frac{3}{2}D$  produise un héritage effectif de  $\frac{2}{3}D$ , il faut que chaque contribution ait été réduite par un facteur de  $\frac{4}{9}$ , puisque  $\frac{4}{9} \times \frac{3}{2} = \frac{2}{3}$ .

Une autre possibilité est que la contribution de chaque ancêtre soit taxée de façon répétée à chaque transmission, et qu'une proportion  $\frac{1}{r}$  seulement soit transmise à chaque fois. Dans ce cas l'héritage effectif serait  $\left(\frac{1}{r} + \frac{1}{3r} + \frac{1}{9r^2} + \dots\right) D = \frac{3}{3r-1}D$ , et pour qu'il soit égal à  $\frac{2}{3}D$  il faut prendre  $\frac{1}{r} = \frac{6}{11}$ .

Selon le modèle choisi, les particularités du parent-moyen contribuent pour  $\frac{4}{9}$  aux particularités de l'enfant, ou pour  $\frac{6}{11}$ . Ces valeurs diffèrent peu de  $\frac{1}{2}$ , et leur moyenne est proche de  $\frac{1}{2}$ , donc on peut accepter ce résultat. Ainsi l'influence pure et simple du parent-moyen peut être considérée comme égale à  $\frac{1}{2}$ , celle du grand-parent à  $\frac{1}{4}$ , etc. Par conséquent l'influence d'un parent individuel est de  $\frac{1}{4}$ , d'un grand-parent de  $\frac{1}{16}$ , etc. »

On voit que dans le dernier paragraphe, Galton n'envisage plus que le second modèle, celui où la contribution des ancêtres décroît selon une série géométrique. Le moins qu'on

---

\*Galton ne fait pas explicitement référence au stirp dans le texte.

puisse dire de ces calculs est qu'ils sont quelque peu controuvés. Galton en est conscient et conclut par ces quelques mots :

« Il serait cependant hasardeux, sur cette fragile base, d'étendre cette séquence avec confiance à des générations plus éloignées. »

Le but de tout ceci est peut-être plus clair si on considère que Galton cherche à estimer dans quelles proportions le stirp d'un individu est composé des germes qui ont participé au développement de chacun de ses ancêtres. Le chapitre sur la couleur des yeux va aider (un peu) à comprendre comme Galton entend utiliser ces proportions.

**Application à la couleur des yeux** Pour Galton, la loi ébauchée ci-dessus est une loi universelle de l'hérédité, sa portée ne se limite pas à la stature. Voici comment il entend l'appliquer à la couleur des yeux. On considère un individu F qui a une « particularité » D, c'est-à-dire aussi bien un écart à la stature moyenne qu'une couleur d'yeux. La loi de la régression à la moyenne fait que chacun de ses deux parents a (en moyenne) une particularité  $\frac{1}{3}D$ , chacun des quatre grands-parents a une particularité  $\frac{1}{9}D$ , etc.

Cet individu F transmet à un de ses enfants S un quart de sa propre particularité, donc  $\frac{1}{4}D$ , mais également un seizième de la particularité de chacun de ses deux parents, soit au total (en moyenne)  $2 \times \frac{1}{16} \times \frac{1}{3}D$ , et ainsi de suite. Finalement, si on ne connaît rien des parents de F, on peut considérer qu'il contribue pour

$$\frac{1}{4}D + 2 \times \frac{1}{16} \times \frac{1}{3}D + 4 \times \frac{1}{64} \times \frac{1}{9}D + \dots = 0,3D$$

à la particularité de S. Pour Galton, ce résultat est à interpréter comme une proportion : F contribue pour 30% à l'héritage de S. L'autre parent contribue également pour 30%, et le « résidu » (sic) de 40% est dû aux ancêtres dont on ne connaît rien, mais qu'on suppose représentatifs de la population.

Voici ce que cela implique concrètement pour Galton : si on considère deux nuances pour la couleur des yeux, clairs ou foncés, avec dans la population 70% d'individus aux yeux clairs et 30% aux yeux sombres, deux parents aux yeux clairs auront un enfant aux yeux clairs avec probabilité  $0,30 + 0,30 + 0,40 \times 0,70$  (le premier terme correspond à la probabilité d'hériter la couleur des yeux du premier parent, le second terme au deuxième parent, et le troisième terme à l'héritage d'un ancêtre inconnu). De même, un parent aux yeux clairs et un parent aux yeux sombres auront un enfant aux yeux clairs avec probabilité  $0,30 + 0,40 \times 0,70$ , etc.

Le lecteur moderne aura sans doute eu quelques difficultés à suivre les calculs ci-dessus, tant la rigueur en semble ou en est absente ; j'ai pourtant fait de mon mieux pour les clarifier. La bonne impression que peut faire Galton quand il décrit les relations entre stature des parents et stature des enfants ne dure malheureusement pas.

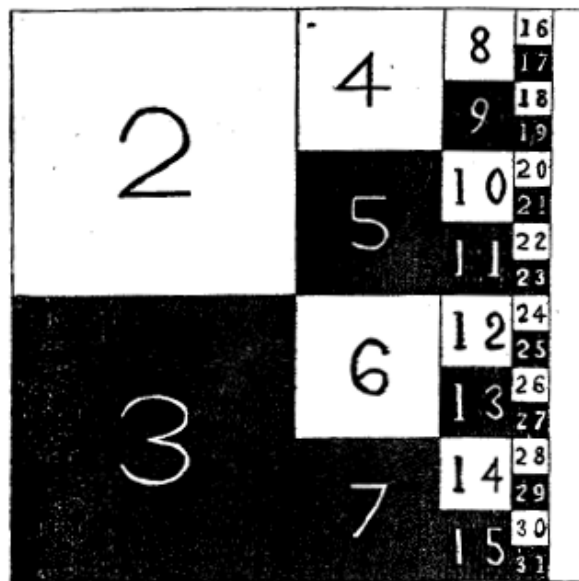
## La loi de l'hérédité de 1897

En 1897, dans *La contribution moyenne de chacun des ancêtres à l'héritage total de leur descendance* [49], Galton n'hésite plus : la loi qu'il n'a ébauchée qu'avec précaution dans *Natural Inheritance* lui semble à présent suffisamment confirmée, en particulier par l'analyse de nouvelles données sur l'héritage de la couleur du pelage des bassets. Il postule donc que les parents d'un individu contribuent à deux (en moyenne) pour moitié à l'héritage total de leur enfant, soit (en moyenne) un quart chacun ; que les quatre grands-parents contribuent (en moyenne) pour un quart, soit un seizième chacun ; etc. Avec des notations modernes, on pourra écrire

$$Y_1 = \frac{1}{4} \underbrace{(Y_2 + Y_3)}_{\text{deux parents}} + \frac{1}{16} \underbrace{(Y_4 + Y_5 + Y_6 + Y_7)}_{\text{quatre grands-parents}} + \frac{1}{64} \underbrace{(Y_8 + \dots + Y_{15})}_{\text{huit bisaïeux}} + \dots \quad (1.1)$$

La séduction de cette loi aux yeux de Galton tient beaucoup au fait que la somme des contributions ancestrales vaut 1, ce qui permet notamment d'appliquer la loi au phénotype ou à sa déviation d'avec la moyenne. En effet, on a

$$1 = \frac{1}{4} \underbrace{(1 + 1)}_{\text{deux termes}} + \frac{1}{16} \underbrace{(1 + 1 + 1 + 1)}_{\text{quatre termes}} + \frac{1}{64} \underbrace{(1 + \dots + 1)}_{\text{huit termes}} + \dots$$



**FIGURE 1.5.** – Un diagramme de l'hérédité. Il faut mentalement compléter le carré par une multitude de carrés de plus en plus petits. Figure de Meston [50], reproduite par Galton [51].

Dans une lettre à *Nature* de 1898 [51], Galton reproduit la figure 1.5 créée par Meston [50], qui lui paraît propre à illustrer et à populariser cette loi. Comme nous l'avons déjà dit plus haut à propos des lois énoncées dans *Natural Inheritance*, tout devient beaucoup plus clair



si on considère que Galton veut en fait élucider la composition du stirp, même s'il n'en fait plus mention : la figure en question représente le stirp d'un individu, mosaïque des stirps de ses ancêtres.

Bien que l'égalité énoncée ci-dessus paraisse déterministe, une variabilité subsiste : un individu n'est pas totalement déterminé par l'ensemble de ses ancêtres – il faut expliquer les différences entre les membres d'une même fratrie. Pour Galton ces différences proviennent de variations dans la valeur des proportions  $\frac{1}{4}$ ,  $\frac{1}{16}$ , etc, qui donnent la contribution de chacun des ancêtres au stirp de l'individu, et ne sont que des valeurs moyennes ; en pratique elles peuvent s'écarter de ces valeurs (mais leur somme restera égale à 1).

Karl Pearson, qui travaille sur l'hérédité et la corrélation entre apparentés depuis quelques années, se saisit immédiatement de cette loi, qu'il appelle *Galton's Law of Ancestral Heredity*. Dans un article [52] publié en janvier 1898 et dédié à Galton (*A New Year's Greeting to Francis Galton*), il l'écrit sous une forme similaire à l'équation (1.1), et l'interprète comme la régression de  $Y_1$  sur l'ensemble des valeurs du trait chez ses ancêtres, donc comme une espérance conditionnelle. On peut s'attarder un instant sur la différence d'interprétation : pour Pearson, les coefficients sont fixés, et si la valeur de  $Y_1$  n'est pas totalement déterminée par les autres  $Y_i$ , cela provient de l'existence d'une variance résiduelle.

Pearson considère (sans l'écrire explicitement) une forme plus générale de la loi :

$$Y_1 = \frac{1}{2}\gamma\beta'(Y_2 + Y_3) + \frac{1}{4}\gamma\beta'^2(Y_4 + Y_5 + Y_6 + Y_7) + \frac{1}{8}\gamma\beta'^3(Y_8 + \dots + Y_{15}) + \dots$$

tout en conservant la contrainte  $\sum_{i \geq 1} \gamma\beta'^i = 1^*$ , c'est-à-dire  $(1 + \gamma)\beta' = 1^\dagger$ . Cette variante correspond à des proportions différentes de celles supposées par Galton dans la constitution du matériel héréditaire, ce que Galton et Pearson à sa suite appellent « la taxe sur l'héritage ».

Si les lois marginales sont gaussiennes, de même espérance  $\mu$  et variance  $\sigma^2$ , s'il n'y a ni homogamie, ni hétérogamie (c'est-à-dire qu'il n'y a pas de corrélation entre les phénotypes des individus qui forment un couple), cette équation suffit à décrire la loi jointe des  $Y_i$ . Pearson montre que la corrélation entre deux individus ancêtre l'un de l'autre et distants de  $\ell$  générations est

$$r_\ell = c \left( \frac{\beta'(1 + \gamma)}{2} \right)^\ell,$$

\*La notation  $\beta'$  est de Pearson, nous conservons le « prime » à l'intention du lecteur qui voudrait se référer à [52]. La variante est introduite p. 393 équation (x) – via  $\beta = \beta'/\sqrt{2}$  – et la contrainte est explicitée p. 403.

†De façon étonnante Pearson se trompe dans le calcul de la somme de cette série et affirme que cette contrainte est équivalente à  $\gamma\beta' = \frac{1}{2}$  ; l'erreur sera corrigée deux ans plus tard dans [53]. Dans le texte nous tenons bien sûr compte de la correction.

où  $c$  dépend également de  $\gamma$  et  $\beta'$ . En tenant compte de la contrainte mentionnée plus haut, on a  $r_\ell = c \left(\frac{1}{2}\right)^\ell$  et on peut calculer

$$c = \frac{\beta'^2 - 3\beta' + 2}{\beta'^2 - 2\beta' + 2}. \quad (1.2)$$

On vérifie que  $c$  prend toutes les valeurs entre 0 et 1 quand  $\beta_1$  va de 1 à 0 ; en pratique on peut donc choisir une valeur arbitraire pour  $c$ .

La loi telle que Galton l'envisage correspond à  $\beta' = \frac{1}{2}$  et donne  $c = 0,6$ . Pearson note que c'est précisément la première valeur trouvée par Galton pour la stature\*, avant qu'il ne la corrige en  $\frac{2}{3}$ .

Pearson calcule également les corrélations entre collatéraux (frères, cousins, etc). Ce sont ces coefficients de corrélation qui l'intéressent en pratique ; dans des travaux ultérieurs (cf par exemple [54]) il s'attachera à estimer leurs valeurs pour diverses mesures anthropométriques.

### 1.2.3. Postérité

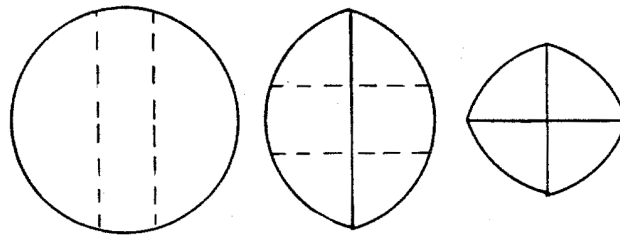
Il est frappant de comparer Galton et Mendel, parfaitement contemporains puisque nés tous les deux en 1822, et tous deux préoccupés par la découverte des lois de l'hérédité. La pauvreté de Mendel fut en grande partie la cause de ce qu'il ne put diffuser et faire reconnaître son travail scientifique, malgré son immense qualité ; l'aisance financière de Galton lui laissa au contraire tout loisir de donner une large publicité à ses théories.

Il est en effet difficile de se faire une idée de la célébrité atteinte par Galton de son vivant ; il était considéré comme un des plus grands scientifiques de son temps. À la fin de sa longue carrière (il publia plusieurs articles par an jusqu'à sa mort en janvier 1911), il écrivait à l'éditeur de *Nature* comme d'autres aujourd'hui tiennent un blog. Je ne résiste pas au plaisir de faire part ici du contenu d'une lettre qu'il lui adressa en 1906 [55] – il avait alors 84 ans. Dans cette lettre intitulée *Couper un gâteau sur la base de principes scientifiques*, il décrit une manière de partager un gros gâteau rond qu'on prévoit de manger en plusieurs fois, de façon à éviter que l'entame du gâteau ne rassisse : on commence par manger d'abord une bande centrale du gâteau (cf figure 1.6) ; les deux morceaux restants sont accolés l'un à l'autre, ainsi le gâteau ne sèche pas ; le lendemain on recommence avec une bande coupée perpendiculairement à la jointure, et il reste quatre petites parts à manger le troisième jour.

Ajoutons qu'Arthur Jensen, qui fut au siècle dernier un des champions de l'héritabilité du QI, fait allusion à cette lettre dans ces termes : *il a utilisé les mathématiques de la géométrie dans l'espace pour découvrir la façon optimale de découper un gâteau de n'importe quelle forme et dimensions en un nombre quelconque de morceaux, de façon à préserver la fraîcheur*

---

\*On vérifie simplement que le coefficient de la régression sur le parent-moyen est égal à  $c$ .



Broken straight lines show intended cuts. Ordinary straight lines show the cuts that have been made. The segments are kept in apposition by a common elastic band that encloses the whole. In the above figures about one-third of the area of the original disc is removed by each of the two successive operations.

FIGURE 1.6. – Comment couper un gâteau rond. *Nature*, 1906.

de chaque pièce [56]. Cette double anecdote est révélatrice : la réputation de Galton est énorme et on tend à lui prêter plus qu'il n'a réalisé.

Il est ainsi systématiquement présenté comme pionnier des statistiques. C'est indubitable, mais il faut préciser : des statistiques descriptives. Comme déjà signalé plus haut, Galton estime les coefficients de régression graphiquement. Il y a peu de mathématiques dans l'œuvre de Galton, et quand il fait appel à l'aide d'un mathématicien comme Dickson, il n'a pas l'air de comprendre le détail des calculs qu'il reçoit en réponse. Outre la définition du coefficient de régression, on lui doit celle de la corrélation (dite aujourd'hui corrélation de Pearson), qu'il définit comme le coefficient de régression d'une grandeur sur une autre, les deux grandeurs ayant été exprimées en nombre d'écart-types\* [57]. Il est également un pionnier de la biométrie ; son seul prédécesseur illustre est Adolphe Quételet, qui avait notamment remarqué le premier que les mesures anthropométriques sont réparties comme une loi normale. Galton est le premier à s'intéresser de façon systématique aux corrélations entre apparentés, ouvrant une voie de recherche féconde. Est-ce bien suffisant pour ranger Galton au premier rang des savants de son temps, au même titre par exemple que Darwin, Cantor, Maxwell ?

Revenons à la loi de l'hérédité. La façon dont Galton prétend la déduire des données expérimentales laisse rêveur — l'argumentation de Swinburne [58], qui remarque que Galton avait déjà énoncé cette loi en 1865 [59] sans produire d'arguments pour la démontrer, et n'a fait par la suite que chercher à confirmer son intuition première est assez convaincante. Nous avons essayé de montrer que la loi de l'hérédité de Galton se comprend beaucoup mieux si on considère que bien qu'il parle de la valeur des phénotypes, ce qu'il a à l'esprit c'est la composition du stirp.

De façon spectaculaire cependant, la version finalement produite par Pearson produit des résultats compatibles avec les corrélations observées entre apparentés, et avec les prédictions obtenues sous le modèle polygénique que Fisher proposera en 1918 — le tout évoque la conception classique selon laquelle une nouvelle théorie scientifique inclut la précédente comme cas particulier. Mais comme on l'a vu Pearson réinterprète la loi de Galton à sa fa-

\*En fait Galton utilise l'écart inter-quartiles pour standardiser les mesures, mais c'est secondaire

çon : tout d'abord, il lit l'égalité énoncée par Galton comme une espérance conditionnelle, ce qui en change profondément le sens, puis il introduit discrètement un paramètre qui permet d'ajuster la loi pour expliquer des niveaux de corrélations différents – c'est pour nous un point de détail, mais cela contredit l'idée de Galton d'un mécanisme unique à l'œuvre pour tous les traits étudiés. Elle ne peut pas non plus être appliquée à certains traits discrets qui sont beaucoup mieux expliqués par les lois de Mendel,\* ce qui ruine les espoirs de Galton d'expliquer du même coup le phénomène de l'atavisme.

Mais Galton reste également dans l'histoire comme le fondateur de l'eugénisme. Dès la première page d'*Hereditary Genius* [62], son premier livre sur l'hérédité, le décor est planté :

« De même qu'il est facile d'obtenir des races de chiens ou de chevaux particulièrement douées sur le plan de la course, ou sur tout autre plan, il serait faisable, par des mariages judicieux pendant plusieurs générations consécutives, de produire une race d'hommes extrêmement douée. »

L'eugénisme se préoccupe donc du moyen d'améliorer les qualités d'une population, en éliminant le mauvais matériel héréditaire (c'est-à-dire déficients mentaux, pauvres, vagabonds, alcooliques, voleurs...) au profit du bon. Les moyens d'arriver à cette fin sont variés, comme on le sait. Galton se contente quant à lui de préconiser de bannir socialement les mariages peu souhaitables d'un point de vue eugénique, et de faire de l'eugénisme une forme de nouvelle religion laïque ; à son crédit, il recommande prudemment de ne pas se hâter de prendre des mesures vouées à l'échec et qui risqueraient de discréditer la science [63].

Même si la question n'est apparemment pas le propos premier de l'eugénisme, Galton a son avis sur la hiérarchie des races. Dans *Hereditary Genius*, il esquisse une échelle qui va des mélanésien aux athéniens de l'époque classique – race admirable qui surclasse la race anglaise d'autant que celle-ci surclasse la race nègre, et dont il attribue la décadence à des mœurs dissolues et à une immigration qui l'ont empêchée de maintenir « la pureté de la race ». Cette décadence est pour Galton regrettable, car si la population athénienne avait pu se multiplier et s'étendre sur de larges territoires en « déplaçant les populations inférieures » qui s'y trouvaient, elle aurait sûrement accompli des choses qui « transcendent les pouvoirs de notre imagination » [62].

Cette préoccupation de préserver la pureté de la race, ou de n'y introduire que des individus de valeur génétique supérieure, restera une des grandes préoccupations des eugénistes – on se référera par exemple aux travaux de Pearson sur l'immigration des juifs en Angleterre [64], qui envisage la question sous l'angle de la valeur génétique de ceux-ci. Il paraît impossible d'absoudre Galton et les eugénistes de leur responsabilité dans le succès que connut l'idée d'« hygiène raciale ». On répondra qu'il faut replacer les choses dans leur contexte. Il est sans doute vrai que Galton reproduit les préjugés de son époque, mais, ce faisant, il les pare d'un vernis de respectabilité scientifique, et les renforce.

---

\*Les biométriciens suivirent pourtant les traces de Galton en l'appliquant à des traits discrets parfaitement mendéliens, comme l'albinisme : par une série d'expériences réalisées entre 1902 et 1904 Darbishire entend démontrer que la proportion d'albinos dans la descendance du croisement de deux souris n'est pas en concordance avec les lois de Mendel, mais bien avec la loi de l'hérédité ancestrale [60]. Il changera rapidement d'avis, voir Provine [61] pour les détails de cette histoire.

## 1.3. Enfin Fisher vint

### 1.3.1. Biométriciens et mendéliens

La redécouverte des lois de Mendel en 1900 survient alors qu'une controverse scientifique oppose d'un côté les biométriciens Karl Pearson et Raphael Weldon, et de l'autre, William Bateson. Cette controverse porte sur la nature des mécanismes de l'évolution et de la spéciation.

Pour Bateson, la spéciation est un phénomène discontinu ; il met l'accent sur la segmentation des individus ou des parties du corps. Une nouvelle espèce apparaît quand « une mutation » crée des individus qui ne peuvent plus s'hybrider avec l'ancienne espèce. Pour Weldon, qui entraîne à sa suite son collègue Pearson, l'évolution est un phénomène graduel et continu ; la spéciation nécessite que deux populations soient isolées l'une de l'autre et évoluent dans des directions différentes. La querelle n'est pas que scientifique, c'est également un affrontement de personnalités, une détestation réciproque de Weldon et Bateson.

Bateson se fait le champion du mendélisme, qui est la théorie rêvée pour la modélisation de phénomènes discontinus dans l'hérédité ; de leur côté, les biométriciens rejettent violemment ces nouvelles idées qu'ils jugent incapables d'expliquer l'hérédité des traits continus. Le mendélisme devient sinon l'enjeu principal de la controverse, du moins le plus visible ; le compromis paraît longtemps impossible et on est sommé de choisir son camp – certains en changeront, comme Arthur Darbishire (le livre de William Provine, *The Origins of theoretical population genetics* [61]), raconte cette histoire de façon passionnante).

Il est intéressant que de constater que, du même coup, les premiers généticiens seront également souvent critiques à l'égard des théories eugénistes soutenues par les biométriciens : Bateson ne rejette pas en bloc toutes les thèses des eugénistes, mais il leur demande [65] avec malice si, plutôt que de prôner la stérilisation des criminels, il ne conviendrait pas de s'attaquer aux fournisseurs de l'armée et à leurs complices, les patriotes de salle de rédaction ;\* en 1925, Thomas Hunt Morgan, un autre pionnier de la génétique, insiste sur le rôle de l'environnement dans les traits comportementaux ou mentaux, et affirme que les preuves apportées par les eugénistes dans ce domaine sont très insuffisantes ( [66], pp 198–207).

Cette hostilité réciproque entre biométriciens et généticiens, à peine troublée par les efforts inachevés de G. Udny Yule [67] eut pour effet qu'il fallut attendre 1918 pour que Ronald Aylmer Fisher fasse la synthèse entre les deux théories.

### 1.3.2. Le modèle de Fisher

Nous n'allons pas suivre fidèlement l'article de Fisher [68], dont la lecture est particulièrement ardue. Il paraît préférable d'essayer de restituer sa démarche tout en utilisant un formalisme plus moderne. Nous ne traiterons pas non plus la totalité de son contenu, car dès

---

\*dans le texte : *army contractors and their accomplices the newspaper patriots*

cet article Fisher traite le cas de locus multi-alléliques, génétiquement liés, de l'homogamie parentale, etc.

### La démarche de Fisher

Fisher suppose qu'un grand nombre de facteurs mendéliens indépendants contribuent à la valeur mesurée, qui en est la somme ; il montre qu'alors la loi de cette valeur est normale.\* Chacun de ces facteurs mendéliens est de la forme

$$X_u = \begin{cases} u_0 & \text{si le génotype est } AA \\ u_1 & \text{si le génotype est } Aa \\ u_2 & \text{si le génotype est } aa \end{cases}$$

Notons  $p$  et  $q = 1 - p$  les fréquences des allèles  $A$  et  $a$  ; sous l'hypothèse d'indépendance des allèles parentaux (que nous appelons aujourd'hui le modèle d'Hardy-Weinberg), les trois génotypes ont pour fréquences  $p^2$ ,  $2pq$  et  $q^2$ .†

On peut tenter de reconstituer la démarche de Fisher ainsi : le but étant de calculer la corrélation entre apparentés, on commence par s'intéresser à la corrélation entre parent et enfant, pour un facteur mendélien  $X_u$ . Toujours dans le modèle d'Hardy-Weinberg, les probabilités conjointes pour les génotypes parent-enfant sont les suivantes :

	AA	Aa	aa
AA	$p^3$	$p^2q$	0
Aa	$p^2q$	$p^2q + pq^2$	$pq^2$
aa	0	$pq^2$	$q^3$

TABLE 1.2. – Probabilités conjointes des génotypes parent-enfant

Notons  $X_u^p$  et  $X_u^e$  les valeurs de  $X_u$  chez le parent et l'enfant. Leur covariance est alors

$$\text{cov}(X_u^p, X_u^e) = p^3 u_0^2 + 2p^2 q u_0 u_1 + (p^2 q + pq^2) u_1^2 + 2pq^2 u_1 u_2 + q^3 u_2^2 \quad (1.3)$$

$$- (p^2 u_0 + 2pqu_1 + q^2 u_2)^2 \\ = pq (p(u_1 - u_0) + q(u_2 - u_1))^2 \quad (1.4)$$

\*Au lieu de considérer comme on aurait pu s'y attendre que le résultat est connu et ne nécessite pas de produire des arguments – la traduction française de l'article de Tchebychev [69] date de 1890 – Fisher calcule pour cela le troisième et le quatrième moment et montre qu'ils s'approchent des moments d'une loi normale.

†Cette fois, Fisher utilise ce fait sans démonstration, et sans faire référence à Hardy.

Cette factorisation quasi-miraculeuse conduit à considérer le cas particulier des facteurs mendéliens additifs, c'est-à-dire le cas  $u_1 - u_0 = u_2 - u_1 = a$  ; on calcule alors

$$\begin{aligned}\text{cov}(X_u^p, X_u^e) &= pqa^2 \\ \text{var}(X_u) &= 2pqa^2\end{aligned}$$

et la corrélation entre les effets génétiques vaut  $\frac{1}{2}$ .

### Décomposition en composante additive et composante de dominance

Pour traiter le cas général, Fisher décompose les facteurs mendéliens en une partie additive et un facteur résiduel :

« La contribution des facteurs mendéliens imparfaitement additifs se décompose, à des fins statistiques, en deux parties : une partie additive qui reflète la nature génétique sans distorsion et est à l'origine des corrélations observées ; et un résidu qui se comporte à peu près de la même façon que l'erreur arbitraire introduite dans les mesures. »

Cette décomposition est réalisée en considérant un facteur mendélien additif  $X'$  tel que  $X - X'$  ait la plus petite variance possible ; la décomposition de  $X$  est alors  $X = X' + (X - X')$ , et la variance de  $X$  se décompose en une composante additive et une composante de dominance,  $\text{var}(X) = \text{var}(X') + \text{var}(X - X')$ , dont il calcule la valeur.

Nous en donnerons ici une interprétation géométrique : en identifiant l'espace des variables aléatoires de la forme  $X_u$  à un espace euclidien de dimension 3, avec le produit scalaire  $\langle X, Y \rangle = E(XY)$ ,  $X'$  est la projection de  $X$  sur le plan engendré par les variables aléatoires  $X_1$ , définie par  $u_0 = u_1 = u_2 = 1$ , et  $X_{012}$ , définie par  $u_0 = 0, u_1 = 1$  et  $u_2 = 2$ . Ce plan est bien l'ensemble des facteurs additifs. On en construit une base orthonormale  $(X_1, X_a)$  en orthogonalisant la base  $(X_1, X_{012})$  : on obtient

$$X_a = \frac{1}{\sqrt{2pq}} (X_{012} - 2qX_1) = \begin{cases} \frac{1}{\sqrt{2pq}}(0 - 2q) & \text{si AA} \\ \frac{1}{\sqrt{2pq}}(1 - 2q) & \text{si Aa} \\ \frac{1}{\sqrt{2pq}}(2 - 2q) & \text{si aa} \end{cases} \quad (1.5)$$

On peut compléter cette base en une base orthonormale de l'espace des facteurs mendéliens en y ajoutant

$$X_d = \begin{cases} \frac{q}{p} & \text{si AA} \\ -1 & \text{si Aa} \\ \frac{p}{q} & \text{si aa} \end{cases} \quad (1.6)$$

On peut réinterpréter le fait que  $\langle X_1, X_a \rangle = 0$  et  $\|X_a\|^2 = 1$  en  $E(X_a) = 0$  et  $\text{var}(X_a) = 1$ , c'est-à-dire que  $X_a$  est centrée et réduite ; il en va de même pour  $X_d$ . Sur cette base,  $X_u$  s'écrit

$$X_u = \mu X_1 + \alpha X_a + \delta X_d$$

avec

$$\mu = p^2 u_0 + 2pq u_1 + q^2 u_2 \quad (1.7)$$

$$\alpha = \sqrt{2pq} (p(u_1 - u_0) + q(u_2 - u_1)) \quad (1.8)$$

$$\delta = pq(u_0 + u_2 - 2u_1) \quad (1.9)$$

On a bien sûr  $E(X_u) = \langle X_1, X_u \rangle = \mu$  et  $E(X_u^2) = \|X_u\|^2 = \mu^2 + \alpha^2 + \delta^2$ , donc  $\text{var}(X_u) = \alpha^2 + \delta^2$  ; la variance de  $X_u$  a été décomposée en variance additive et en variance de dominance.

Finalement, si on considère  $r$  facteurs mendéliens indépendants  $X_{(1)}, \dots, X_{(r)}$ , la variance de chacun d'eux se décompose de la même façon :  $\text{var}(X_{(i)}) = \alpha_i^2 + \delta_i^2$ , et la variance de leur somme est  $\tau = \tau_a + \tau_d$ , avec  $\tau_a = \sum_i \alpha_i^2$  et  $\tau_d = \sum_i \delta_i^2$ .

## Corrélation entre apparentés

Revenons à la corrélation entre apparentés, et d'abord à la corrélation parent-enfant.

On suppose que le phénotype  $Y$  est la somme  $Y = G + E$  d'effets génétiques  $G$  et environnementaux  $E$  indépendants,  $G$  se décomposant lui-même en une somme de facteurs mendéliens indépendants,  $G = X_{(1)} + \dots + X_{(r)}$ , et  $E$  étant supposé normal.\* La variance de  $Y$  est donc  $\text{var}(Y) = \text{var}(G) + \text{var}(E) = \tau + \sigma^2 = \tau_a + \tau_d + \sigma^2$ .

On note avec des exposants  $p$  et  $e$  les valeurs correspondantes chez le parent et l'enfant considérés. D'après les équations (1.4) et (1.8) on a, pour chacun des facteurs mendéliens impliqués,

$$\text{cov}(X^p, X^e) = \frac{1}{2} \alpha^2.$$

C'est ce résultat qui était la motivation de la décomposition en composantes additives et dominantes. Si on décompose  $X^p$  en  $X^p = \mu + \alpha X_a^p + \delta X_d^p$  et  $X^e$  en  $X^e = \mu + \alpha X_a^e + \delta X_d^e$ , ce résultat est équivalent au fait qu'on a

$$\begin{aligned} \text{cov}(X_a^p, X_a^e) &= \frac{1}{2} & \text{cov}(X_a^p, X_d^e) &= 0 \\ \text{cov}(X_d^p, X_a^e) &= 0 & \text{cov}(X_d^p, X_d^e) &= 0 \end{aligned}$$

Ces valeurs peuvent bien entendu être calculées directement, ce qui fournit une preuve un peu plus simple de l'équation (1.4).

---

\*Fisher envisage bien l'existence d'un effet environnemental, mais ne lui accorde que peu de place ; dans les données qu'il analyse il trouve qu'on peut le négliger.



Les facteurs mendéliens étant supposés deux à deux indépendants, on a alors  $\text{cov}(G^p, G^e) = \frac{1}{2}(\alpha_1^2 + \dots + \alpha_r^2) = \frac{1}{2}\tau_a$ . En supposant l'indépendance des effets environnementaux  $E^p$  et  $E^e$ , on calcule

$$\text{cor}(Y^p, Y^e) = \frac{1}{2}h^2$$

où

$$h^2 = \frac{\tau_a}{\tau_a + \tau_d + \sigma^2}$$

est l'héritabilité restreinte ; c'est donc le rapport de la variance génétique additive sur la variance totale du phénotype. L'héritabilité large est

$$H^2 = \frac{\tau_a + \tau_d}{\tau_a + \tau_d + \sigma^2}.$$

Ni la notation  $h^2$  ni l'intérêt pour cette grandeur en particulier ne sont dus à Fisher ; comme Pearson il s'intéresse avant tout à la corrélation entre apparentés. C'est Sewall Wright qui le premier note  $h$  le ratio de l'écart-type de l'effet du génotype sur l'écart-type total du phénotype [70, 71], sans envisager précisément le problème de l'additivité des effets.

On généralise facilement ce résultat à d'autres formes d'apparentement. Fisher envisage les ascendances directes — un individu et ses grands-parents, arrière-grands-parents, etc ; il envisage également le cas des germains (frères ou sœurs), des cousins germains, des doubles cousins germains, dressant à chaque fois une table analogue à la table 1.2.

Dans un formalisme moderne, on peut caractériser une relation d'apparentement entre deux individus par les probabilités  $\zeta_0, \zeta_1, \zeta_2$  de chacun des trois « états IBD » possibles, qu'on peut définir comme ceci : deux allèles d'un même gène sont IBD (pour *Identical By Descent* : identique par origine) si ils sont hérités d'un ancêtre commun ; en une région du génome donnée, l'état IBD des deux individus est  $\text{IBD} = 0$  s'ils n'ont pas d'allèles IBD,  $\text{IBD} = 1$  s'ils ont exactement un allèle IBD, et  $\text{IBD} = 2$  s'ils ont deux allèles IBD.

On utilise souvent la paramétrisation équivalente  $\phi$  et  $\psi$  (table 1.3) :

- $\phi$  est le coefficient d'apparentement, défini comme la probabilité pour que deux allèles d'un même gène autosomal, tirés au hasard chez chacun d'eux, soit hérités d'un ancêtre commun. On a  $\phi = \frac{1}{4}\zeta_1 + \frac{1}{2}\zeta_2$ .
- $\psi = \zeta_2$  est la probabilité pour que les deux individus partagent deux allèles IBD.

Si  $\text{IBD} = 0$ , les valeurs  $X^{(1)}$  et  $X^{(2)}$  d'un facteur mendélien chez les deux individus considérés sont indépendantes ; si  $\text{IBD} = 2$ , on a  $X^{(1)} = X^{(2)}$  ; et le cas  $\text{IBD} = 1$  correspond au cas parent-enfant. On a donc pour les facteurs  $X_a$  et  $X_d$

$$\begin{aligned} \text{cov}(X_a^{(1)}, X_a^{(2)}) &= \frac{1}{2}\zeta_1 + \zeta_2 = 2\phi & \text{cov}(X_a^{(1)}, X_d^{(2)}) &= 0 \\ \text{cov}(X_d^{(1)}, X_d^{(2)}) &= 0 & \text{cov}(X_d^{(1)}, X_a^{(2)}) &= \zeta_2 = \psi \end{aligned}$$

Relation	$\zeta$	$\psi$	$\phi$
Parent/enfant	$(0, 1, 0)$	0	$\frac{1}{4}$
Grand-parent/petit-enfant	$(0, \frac{1}{2}, 0)$	0	$\frac{1}{8}$
Germain	$(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	$\frac{1}{4}$	$\frac{1}{4}$
Oncle/neveu	$(\frac{1}{2}, \frac{1}{2}, 0)$	0	$\frac{1}{8}$
Cousins germains	$(\frac{3}{4}, \frac{1}{4}, 0)$	0	$\frac{1}{16}$
Doubles cousins germains	$(\frac{9}{16}, \frac{6}{16}, \frac{1}{16})$	$\frac{1}{16}$	$\frac{1}{8}$

**TABLE 1.3.** – Valeurs de  $\phi$  et  $\psi$  pour quelques relations d'apparentement

et la covariance entre  $X^{(1)}$  et  $X^{(2)}$  est

$$\text{cov}(X^{(1)}, X^{(2)}) = 2\phi\alpha^2 + \psi\delta^2.$$

La covariance entre les composantes génétiques chez les deux individus est

$$\text{cov}(G^{(1)}, G^{(2)}) = 2\phi\tau_a + \psi\tau_d,$$

et on a

$$\text{cor}(Y^{(1)}, Y^{(2)}) = 2\phi h^2 + \psi \frac{\tau_d}{\tau_a + \tau_d + \sigma^2} = 2\phi h^2 + \psi(H^2 - h^2).$$

### Études familiales et études de jumeaux

Les héritabilités  $h^2$  et  $H^2$  peuvent donc être estimées à partir de la corrélation observée entre apparentés – par exemple, entre germains d'une part, entre parents et enfants d'autre part. Une des limitations les plus évidentes du modèle de Fisher est l'hypothèse d'absence de corrélation entre les effets environnementaux : si on considère deux apparentés proches, de phénotypes  $Y^a$  et  $Y^b$  avec

$$Y^a = G^a + E^a$$

$$Y^b = G^b + E^b,$$

en l'absence de corrélation gène-environnement, la covariance de  $Y^a$  et  $Y^b$  est

$$\text{cov}(Y^a, Y^b) = \text{cov}(G^a, G^b) + \text{cov}(E^a, E^b).$$

Les calculs qui précèdent et les procédures d'estimation de l'héritabilité qui en découlent supposent le terme  $\text{cov}(E^a, E^b)$  nul ; s'il est positif, comme il est plausible qu'il le soit, l'estimation de  $\text{cov}(G^a, G^b)$  est biaisée vers le haut, et celle de l'héritabilité l'est également.

Les études de jumeaux sont une solution pour minimiser ce problème. Dans le cas de jumeaux monozygotes, on a  $\phi = \frac{1}{2}$  et  $\psi = 1$ , d'où on tire la valeur de la corrélation phénotypique :

$$r_{\text{MZ}} = h^2 + (H^2 - h^2) + \rho_{\text{MZ}}$$

où  $\rho_{\text{MZ}}$  est la corrélation entre les environnements des jumeaux monozygotes. De même, la corrélation phénotypique entre jumeaux dizygotes est

$$r_{\text{DZ}} = \frac{1}{2}h^2 + \frac{1}{4}(H^2 - h^2) + \rho_{\text{DZ}},$$

où  $\rho_{\text{DZ}}$  est la corrélation entre les environnements des jumeaux dizygotes. Si on suppose que ces termes de corrélations environnementales sont égaux,  $\rho_{\text{MZ}} = \rho_{\text{DZ}}$ , on a

$$2(r_{\text{MZ}} - r_{\text{DZ}}) = h^2 + \frac{3}{2}(H^2 - h^2).$$

Si on suppose en outre que le terme de dominance est nul, c'est-à-dire,  $H^2 = h^2$ , on a  $2(r_{\text{MZ}} - r_{\text{DZ}}) = h^2$ . Cette formule est connue sous le nom de formule de Falconer.

Si il y a un terme de dominance non nul,  $2(r_{\text{MZ}} - r_{\text{DZ}}) = H^2 + \frac{1}{2}(H^2 - h^2)$  surestime l'héritabilité large. Même en l'absence de dominance, l'hypothèse  $\rho_{\text{MZ}} = \rho_{\text{DZ}}$  est très critiquable : il y a beaucoup de raisons, biologiques, comportementales ou sociétales, qui peuvent être cause qu'elle n'est pas vérifiée. Les biais dus à la présence d'environnement partagé entre apparentés restent donc possibles.

### 1.3.3. Retrouver la loi de Galton dans le modèle de Fisher

Dans le modèle de Fisher, la corrélation entre deux individus dont l'un est ancêtre de l'autre, les deux étant séparés par  $\ell$  générations, vaut  $r_\ell = \left(\frac{1}{2}\right)^\ell h^2$ , ce qui coïncide avec ce qu'a calculé Pearson. Pour clore ce chapitre, nous allons montrer rapidement, en supposant l'absence d'homogamie ou d'hétérogamie, qu'on peut en effet utiliser ces corrélations pour retrouver la loi de Galton, ou plus exactement la généralisation qui en a été faite par Pearson.

On peut poser le problème de la régression ainsi : les coefficients de régression  $a_1, a_2, \dots$  de  $Y_1$  sur les ancêtres moyens sont les valeurs qui minimisent la variance de

$$Y_1 - \left( \frac{1}{2}a_1(Y_2 + Y_3) + \frac{1}{4}a_2(Y_4 + \dots + Y_7) + \frac{1}{8}a_3(Y_8 + \dots + Y_{15}) + \dots \right)$$

## Chapitre 1. Au commencement

On note  $\sigma^2$  la variance commune aux  $Y_i$ . Les corrélations mentionnées impliquent que cette variance est égale à

$$\phi(a) = \sigma^2 \left( 1 + \sum_i 2^{-i} a_i^2 - 2h^2 \sum_i 2^{-i} a_i + 2h^2 \times \left( a_1 \sum_{j>1} 2^{-j} a_j + a_2 \sum_{j>2} 2^{-j} a_j + \dots \right) \right)$$

Pour que cette quantité soit finie, il faut que  $\sum_i 2^{-i} a_i$  converge.

On prend  $\frac{1}{2\sigma^2} \frac{\partial}{\partial a_k}$  de cette expression :

$$\frac{1}{2\sigma^2} \frac{\partial \phi}{\partial a_k} = 2^{-k} a_k - 2^{-k} h^2 + h^2 \times \left( 2^{-k} a_1 + 2^{-k} a_2 + \dots + 2^{-k} a_{k-1} + \sum_{j>k} 2^{-j} a_j \right)$$

et donc, en notant  $S_k = \sum_{i=1}^k a_i$  et  $T_k = \sum_{j>k} 2^{-j} a_j$ , on cherche une solution aux équations

$$a_k = h^2 - h^2 S_{k-1} - 2^k h^2 T_k \quad (k \geq 1), \quad (1.10)$$

(on pose par convention  $S_0 = 0$  pour le cas  $k = 1$ ). On en déduit, pour  $k \geq 2$ ,

$$\begin{aligned} a_{k+1} - 3a_k + 2a_{k-1} &= -h^2(S_k - 3S_{k-1} + 2S_{k-2}) - 2^k h^2(2T_{k+1} - 3T_k + T_{k-1}) \\ &= -h^2(a_k - 2a_{k-1}) - 2^k h^2(2^{-k} a_k - 2 \times 2^{-(k+1)} a_{k+1}) \\ &= -h^2(a_k - 2a_{k-1}) - h^2(a_k - a_{k+1}) \end{aligned}$$

d'où

$$(1 - h^2)a_{k+1} + (2h^2 - 3)a_k + 2(1 - h^2)a_{k-1} = 0 \quad (1.11)$$

Le cas  $h^2 = 1$  est particulier : pour tout  $k > 1$  cette équation devient  $a_k = 0$  et pour  $k = 1$ , l'équation (1.10) donne  $a_1 = \frac{1}{2}$ .

Supposons  $h^2 < 1$ . On note  $\xi_1 < \xi_2$  les deux racines de

$$(1 - h^2)\xi^2 + (2h^2 - 3)\xi + 2(1 - h^2) = 0.$$

Notons qu'on a

$$\xi^2 - 3\xi + 2 = h^2(\xi^2 - 2\xi + 2)$$

ce qui est l'équation de Pearson pour la valeur de  $c$  ; ses notations correspondent à  $c = h^2$  et  $\beta' = \xi$ .

Les suites qui satisfont à la condition 1.11 sont des combinaisons linéaires des suites  $\xi_1^k$  et  $\xi_2^k$ . On vérifie facilement que quand  $h^2$  varie de 0 à 1,  $\xi_1$  varie de 1 à 0 et  $\xi_2$  de 2 à  $+\infty$  ; pour que  $\sum_i 2^{-i} a_i$  converge, il faut prendre  $a_k = \alpha \xi^k$ . En passant à la limite  $k \rightarrow \infty$  dans (1.10), on obtient

$$0 = h^2 - h^2 \sum_{k=1}^{\infty} a_k ;$$

c'est la condition de Pearson,  $\sum_k a_k = 1$ . On en tire  $\alpha = \frac{1-\xi}{\xi}$ . On peut également retrouver ce résultat en prenant, par exemple,  $k = 1$  dans l'équation (1.10).

Cette méthode permet aussi de calculer les coefficients de régression dans le cas où on ne considère qu'un nombre fini d'ancêtres (dans ce cas il y a une composante non nulle en  $\xi_2^k$  dans la valeur de  $a_k$ ).



# Chapitre 2.

## Le modèle linéaire mixte

### 2.1. Introduction

Un problème de génétique animale, l'évaluation de la valeur reproductive des taureaux des races laitières à partir de la production laitière de leurs filles [72, 73], fut une des motivations à l'origine du développement du modèle linéaire mixte. Très utilisés en génétique animale, ces modèles ont également été très tôt utilisés en épidémiologie génétique humaine (par exemple pour l'analyse de liaison, [74]) ; ils le sont également couramment en épidémiologie « classique », par exemple pour modéliser l'hétérogénéité des centres de recrutement des patients par des effets aléatoires (on parle alors de modèle à intercepts aléatoires). Mais ces dernières années, de nouvelles applications (tests d'association et estimation d'héritabilité) leur ont donné une popularité sans précédent.

Quand nous avons décidé, Claire Dandine et moi, de nous livrer à des expériences avec ces méthodes, nous nous sommes tout d'abord heurtés à un problème dans le choix de l'outil : la bibliothèque logicielle de référence en R pour le modèle mixte est `lme4` [75], qui ne fonctionne pas de façon satisfaisante pour les utilisations dont nous avons besoin ; en effet, il est optimisé pour les matrices de covariables creuses, c'est-à-dire à dire avec beaucoup d'entrées nulles – c'est le cas dans les utilisations classiques du modèle mixte où les covariables sont des indicatrices utilisées pour modéliser un « effet centre », un « effet patient », etc, mais pas pour le calcul de l'héritabilité. Aucun autre package R ne nous a paru donner des résultats satisfaisants dans ce cas. Inversement, le logiciel GCTA [76] a été écrit pour ce cas de figure, mais d'une part il a un côté « boîte noire », d'autre part c'est un logiciel en ligne de commande, qui fonctionne en lisant des fichiers d'entrée et en écrivant des fichiers de sortie, ce qui alourdit les expériences – on s'habitue vite à la souplesse offerte par l'environnement de travail de R.

Une solution possible nous a semblé être de créer notre propre bibliothèque logicielle. Un second obstacle est alors apparu : il fallait commencer par comprendre les méthodes propres aux modèles mixtes, et en particulier à l'estimation des paramètres. Dans l'ensemble, les sections « méthodes » des articles de génétique humaine [76–79] ne sont guère de nature à éclairer ni leurs lecteurs, ni leurs lectrices – c'est ce qui motivait en tout premier lieu notre désir d'expérimentation. Les premiers points d'entrée que nous avons trouvés dans la

littérature mathématique ne nous ont pas aidés davantage. Nous avons fini par y voir clair petit à petit, grâce en particulier au livre classique de Searle, Casella et McCulloch [80].

Nous avons donc fini par passer en revue la théorie du modèle mixte, tout en cherchant à simplifier son exposé au maximum (et j'aime à croire que nous y sommes parvenus) ; une grande partie de ce chapitre est issue des annexes de [19]. Outre la simplification de l'exposé, quelques points sont originaux, par exemple l'interprétation de l'algorithme EM comme une ascension de gradient, ou le calcul de la variance attribuable aux effets fixes.

Cette clarification nous a permis d'écrire la bibliothèque logicielle dont nous faisons le projet ; elle s'appelle *gaston* et nous en avons parlé brièvement dans le chapitre introductif de ce mémoire.

### 2.1.1. Notations et introduction du modèle

On considère  $n$  individus ; pour chacun d'eux, un trait (ou phénotype)  $Y_i$  a été mesuré, ainsi que deux jeux de covariables  $X_{i1}, \dots, X_{ip}$  et  $Z_{i1}, \dots, Z_{iq}$ , qui vont intervenir dans la valeur de l'espérance de  $Y_i$  selon un modèle linéaire. Les effets des covariables  $X_{ij}$ , notés  $\beta_1, \dots, \beta_p$ , sont des paramètres du modèle, appelés *effets fixes*. En revanche, les effets des covariables  $Z_{ij}$ , notés  $u_1, \dots, u_q$ , sont des variables aléatoires indépendantes, tirées dans une loi  $\mathcal{N}(0, \tau)$ . La variance  $\tau$  de cette distribution est l'unique paramètre qui permet de modéliser ces effets.

Les effets aléatoires sont indépendants de l'indice  $i$  de l'individu ; on peut considérer qu'ils sont tirés au hasard au début de l'expérience, lors du choix des covariables  $Z_{ij}$ . La relation entre  $Y$  et les covariables est linéaire :

$$Y_i = X_{i1}\beta_1 + \dots + X_{ip}\beta_p + Z_{i1}u_1 + \dots + Z_{iq}u_q + e_i$$

pour tout  $i = 1, \dots, n$ , où les termes d'erreur résiduelle  $e_1, \dots, e_n$  sont tirés indépendamment dans une loi normale  $\mathcal{N}(0, \sigma^2)$ .

Il est préférable d'écrire ce modèle sous forme matricielle :

$$Y = X\beta + Zu + e, \quad (2.1)$$

où  $Y = (Y_1, \dots, Y_n)'$  est le vecteur des variables d'intérêt,  $X$  est la matrice  $n \times p$  des covariables à effets fixes  $\beta = (\beta_1, \dots, \beta_p)'$ ,  $Z$  la matrice  $n \times q$  des covariables à effets aléatoires  $u = (u_1, \dots, u_q)' \sim \mathcal{N}(0, \tau I_q)$ , et le vecteur d'erreurs résiduelles  $e = (e_1, \dots, e_n)' \sim \mathcal{N}(0, \sigma^2 I_n)$ .

On peut étendre le modèle en considérant plusieurs matrices  $Z_1, Z_2, \dots$  de dimensions  $n \times q_i$  auxquelles on associe des effets aléatoires de variances  $\tau_1, \tau_2, \dots$  :

$$Y = X\beta + Z_1u_1 + \dots + Z_ku_k + e \quad (2.2)$$

avec

$$u_1 \sim \mathcal{N}(0, \tau_1 I_{q_1}), \dots, u_k \sim \mathcal{N}(0, \tau_k I_{q_k}), e \sim \mathcal{N}(0, \sigma^2 I_n). \quad (2.3)$$



Une formulation équivalente fait intervenir des « valeurs génétiques » aléatoires  $\omega_1$  à  $\omega_k$  de matrices de covariance connue.

$$Y = X\beta + \omega_1 + \cdots + \omega_k + e \quad (2.4)$$

avec

$$\omega_1 \sim \mathcal{N}(0, \tau_1 K_1), \dots, \omega_k \sim \mathcal{N}(0, \tau_k K_k), e \sim \mathcal{N}(0, \sigma^2 I_n), \quad (2.5)$$

où pour tout  $\ell$ ,  $K_\ell = Z_\ell Z'_\ell$ . La valeur génétique  $\omega_\ell$  correspond au terme  $Z_\ell u_\ell$ . Ces deux écritures induisent la même distribution du vecteur  $Y$ , c'est-à-dire une loi normale multivariée d'espérance  $X\beta$  et de variance

$$V = V(\tau_1, \dots, \tau_k, \sigma^2) = \tau_1 K_1 + \cdots + \tau_k K_k + \sigma^2 I_n. \quad (2.6)$$

Dans la suite, nous omettrons les paramètres et noterons cette matrice  $V$ .

Le traitement du modèle linéaire mixte a nécessité le développement d'outils qui lui sont propres, et que nous allons détailler dans la suite, en particulier :

- La vraisemblance restreinte : les estimateurs du maximum de vraisemblance des composantes  $\tau$  et  $\sigma^2$  de la variance sont biaisés. Pour obtenir des estimateurs non biaisés, on utilise la vraisemblance restreinte, qui est la vraisemblance du modèle obtenu après projection sur un espace où l'espérance de  $Y$  est nulle. On parle d'estimateurs REML, pour *Restricted Maximum Likelihood*.
- L'algorithme du second ordre classiquement utilisé pour la maximisation de la vraisemblance restreinte, n'est ni l'algorithme de Newton-Raphson, ni le Fisher scoring : pour diminuer le fardeau calculatoire inhérent aux calculs de la matrice d'information observée ou d'information de Fisher, on utilise la matrice d'information moyenne. On parle d'algorithme AIREML, pour *Average Information Restricted Maximum Likelihood*.
- Les tests basés sur la vraisemblance ne sont pas les tests habituels, pour plusieurs raisons : d'une part, on n'est pas dans le cadre théorique habituel où on considère des mesures répétées indépendantes ; d'autre part, pour tester par exemple l'hypothèse nulle  $\tau = 0$ , on se place au bord de l'espace des paramètres.
- Même si les effets aléatoires  $u$  ne sont pas des paramètres du modèle, on peut prédire leur valeur en se basant sur la valeur de  $Y$  observée, simplement par  $\hat{u} = E(u|Y)$  – c'est exactement la même chose que la prédiction de la taille d'un individu conditionnellement à celle de ses parents, quand on connaît la loi jointe de ces valeurs. On parle de BLUP.

## 2.2. Les vraisemblances

### 2.2.1. La vraisemblance

Le vecteur des observations  $Y$  suit une loi normale multivariée d'espérance  $X\beta$  et de variance  $V = \tau_1 K_1 + \dots + \tau_k K_k + \sigma^2 I_n$ . La vraisemblance est simplement la densité de cette distribution (à une constante multiplicative près), interprétée comme une fonction des paramètres  $\beta, \tau_1, \dots, \tau_k, \sigma^2$  :

$$L(\beta, \tau_1, \dots, \tau_k, \sigma^2) = \frac{1}{\sqrt{|V|}} \exp \left( -\frac{1}{2} (Y - X\beta)' V^{-1} (Y - X\beta) \right).$$

La log-vraisemblance est

$$\ell(\beta, \tau_1, \dots, \tau_k, \sigma^2) = -\frac{1}{2} \log |V| - \frac{1}{2} (Y - X\beta)' V^{-1} (Y - X\beta). \quad (2.7)$$

Estimer les composantes de la variance  $\tau_1, \dots, \tau_k$  et  $\sigma^2$  en maximisant cette vraisemblance mène à des estimations biaisées. La solution classique est d'utiliser la « vraisemblance restreinte », que nous présenterons plus loin ; avant cela, nous allons calculer les dérivées premières de la log-vraisemblance, car elles seront utiles par la suite.

### 2.2.2. Les dérivées de la log-vraisemblance

Le gradient en  $\beta$  est :

$$\frac{\partial \ell}{\partial \beta} = X' V^{-1} (Y - X\beta). \quad (2.8)$$

Pour calculer les dérivées premières suivant  $\tau_i$ , on remarque que  $\frac{\partial}{\partial \tau_i} V = K_i$  et  $\frac{\partial}{\partial \tau_i} \log |V| = \text{tr}(V^{-1} K_i)$ . Donc on a

$$\frac{\partial \ell}{\partial \tau_i} = -\frac{1}{2} \text{tr}(V^{-1} K_i) + \frac{1}{2} (Y - X\beta)' V^{-1} K_i V^{-1} (Y - X\beta). \quad (2.9)$$

Le paramètre  $\sigma^2$  se comporte comme un des  $\tau_i$ , il suffit de remplacer  $K_i$  par la matrice identité  $I_n$  :

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{1}{2} \text{tr}(V^{-1}) + \frac{1}{2} (Y - X\beta)' V^{-1} V^{-1} (Y - X\beta). \quad (2.10)$$

### 2.2.3. Une vraisemblance profilée

Bien qu'il n'y ait pas de forme close pour les estimateurs du maximum de vraisemblance, on trouve facilement la valeur de  $\beta$  qui annule le gradient en  $\beta$  (2.8) :

$$\widehat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y \quad (2.11)$$

Alors on a

$$\begin{aligned} Y - X\beta &= (I_n - X(X'V^{-1}X)^{-1}X'V^{-1})Y \\ &= V(V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})Y \\ &= VPY \end{aligned}$$

où

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \quad (2.12)$$

La matrice  $P$  dépend des paramètres  $\tau_1, \dots, \tau_k, \sigma^2$ . C'est une matrice symétrique qui vérifie  $PVP = P$  et  $PX = 0$ .

La vraisemblance profilée s'obtient en introduisant  $\widehat{\beta}$  dans la vraisemblance (2.7) :

$$\ell^{pr}(\tau, \sigma^2) = \ell(\widehat{\beta}, \tau, \sigma^2) = -\frac{1}{2} \log |V| - \frac{1}{2} Y'PY. \quad (2.13)$$

### 2.2.4. La vraisemblance restreinte

#### Un cas particulier élémentaire

Considérons le cas particulier d'un échantillon de  $n$  variables normales indépendantes  $Y_1, \dots, Y_n$ , de moyenne  $\mu$  et de variance  $\sigma^2$ . Cela correspond à un vecteur  $Y \in \mathbb{R}^n$  avec moyenne  $X\beta = \mu \mathbf{1}_n$  où  $\mathbf{1}_n = (1, \dots, 1)' \in \mathbb{R}^n$  et de variance  $V = \sigma^2 I_n$ . Dans ce cas, la log-vraisemblance se réduit à

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2,$$

et les estimateurs du maximum de vraisemblance sont

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n} \sum_i Y_i \\ \widehat{\sigma}^2 &= \frac{1}{n} \sum_i (Y_i - \widehat{\mu})^2. \end{aligned}$$

L'estimateur  $\widehat{\sigma}^2$  est bien connu pour être biaisé. Si  $\mu$  est connu, en remplaçant  $\widehat{\mu}$  par  $\mu$  dans la formule ci-dessus pour  $\widehat{\sigma}^2$ , on obtient un estimateur sans biais : le biais est attribuable à la présence de l'estimateur  $\widehat{\mu}$ .

Pour supprimer le biais, une solution est d'introduire une matrice  $C \in \mathbb{R}^{(n-1) \times n}$  telle que  $C\mathbf{1}_n = 0$  et  $CC' = I_{n-1}$ . Il y a beaucoup de façons de choisir une telle matrice, puisqu'il suffit que les lignes de  $C$  soient une base orthonormale de l'espace vectoriel orthogonal au vecteur  $\mathbf{1}_n$ .

Considérons le vecteur  $\mathcal{Y} = CY$ , qui suit une loi normale de moyenne connue 0 et de variance  $C(\sigma^2 I_n)C' = \sigma^2 I_{n-1}$ . On a maintenant un échantillon de  $n - 1$  observations indépendantes de moyenne nulle, et  $\sigma^2$  est estimé sans biais par

$$\frac{1}{n-1} \sum_i \mathcal{Y}_i^2 = \frac{1}{n-1} \mathcal{Y}' \mathcal{Y} = \frac{1}{n-1} Y' C' C Y$$

Les conditions imposées sur  $C$  impliquent que  $C'C = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n'$  (Annexe A.1), et on a  $Y'C'CY = \sum_i Y_i^2 - \frac{1}{n} (\sum_i Y_i)^2$ ; la formule ci-dessus correspond à l'estimateur sans biais usuel  $\widehat{\sigma}^2 = \frac{1}{n-1} \left( \sum_i Y_i^2 - \frac{1}{n} (\sum_i Y_i)^2 \right)$ .

### Le cas général

Le cas général est basé sur la même idée. On fixe une matrice de contrastes  $C \in \mathbb{R}^{(n-p) \times n}$  telle que  $CX = 0$  et  $CC' = I_{n-p}$  (les lignes de  $C$  sont une base orthonormale de l'espace vectoriel orthogonal à l'espace engendré par les colonnes de  $X$ ). On travaille sur  $\mathcal{Y} = CY$ , qui a une espérance 0 et variance

$$W = CVC' = \tau_1 CK_1 C' + \dots + \tau_k CK_k C' + \sigma^2 I_{n-p}.$$

On obtient la log-vraisemblance restreinte pour  $\tau_1, \dots, \tau_k, \sigma^2$  :

$$\begin{aligned} \ell^{\text{re}}(\tau_1, \dots, \tau_k, \sigma^2) &= -\frac{1}{2} \log |W| - \frac{1}{2} \mathcal{Y}' W^{-1} \mathcal{Y} \\ &= -\frac{1}{2} \log |CVC'| - \frac{1}{2} Y' C' (CVC')^{-1} CY \end{aligned}$$

Pour toute matrice de contraste  $C$ , on a  $C' (CVC')^{-1} C = P$  avec  $P$  comme dans (2.12) (annexe A.1); on a également

$$\log |CVC'| = \log |V| + \log |X'V^{-1}X| + \text{constante}$$

(annexe A.1), et finalement

$$\ell^{\text{re}}(\tau_1, \dots, \tau_k, \sigma^2) = -\frac{1}{2} \log |V| - \frac{1}{2} \log |X'V^{-1}X| - \frac{1}{2} Y' P Y + \text{constant}. \quad (2.14)$$

### Dérivées de la log-vraisemblance restreinte

Plutôt que de dériver (2.14), on remplace  $Y$  par  $CY$ ,  $V$  par  $CVC'$ ,  $K_i$  par  $CK_iC'$  et  $X\beta$  par 0 dans les formules (2.9), (2.10) ; après substitution de  $C' (CVC')^{-1} C$  par  $P$ , les dérivées premières sont

$$\frac{\partial \ell^{\text{re}}}{\partial \tau_i} = -\frac{1}{2} \text{tr}(PK_i) + \frac{1}{2} Y' PK_i PY \quad (2.15)$$

et

$$\frac{\partial \ell^{\text{re}}}{\partial \sigma^2} = -\frac{1}{2} \text{tr}(P) + \frac{1}{2} Y' PPY. \quad (2.16)$$

On calcule également les dérivées secondes :

$$\frac{\partial^2 \ell^{\text{re}}}{\partial \tau_i \partial \tau_j} = \frac{1}{2} \text{tr}(PK_i PK_j) - Y' PK_i PK_j PY \quad (2.17)$$

Comme déjà noté, le paramètre  $\sigma^2$  se comporte comme un des  $\tau_i$ , la matrice  $K_i$  étant remplacée par  $I_n$  ; on obtient donc facilement les dérivées en  $\sigma^2$  à partir de l'équation précédente.

L'espérance de (2.17), avec  $Y \sim \mathcal{N}(X\beta, V)$ , est :

$$\begin{aligned} E \left( \frac{\partial^2 \ell^{\text{re}}}{\partial \tau_i \partial \tau_j} \right) &= \frac{1}{2} \text{tr}(PK_j PK_i) - E(Y' PK_i PK_j PY) \\ &= \frac{1}{2} \text{tr}(PK_j PK_i) - \text{tr}(PK_i PK_j PV) \\ &= \frac{1}{2} \text{tr}(PK_j PK_i) - \text{tr}(PK_j PVPK_i) \\ &= -\frac{1}{2} \text{tr}(PK_j PK_i) \end{aligned} \quad (2.18)$$

puisque  $PVP = P$ . On obtient l'information de Fisher :

$$I(\tau_1, \dots, \tau_k, \sigma^2) = \begin{bmatrix} \frac{1}{2} \text{tr}(PK_1 PK_1) & \cdots & \frac{1}{2} \text{tr}(PK_1 PK_k) & \frac{1}{2} \text{tr}(PK_1 P) \\ \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2} \text{tr}(PK_k PK_1) & \cdots & \frac{1}{2} \text{tr}(PK_k PK_k) & \frac{1}{2} \text{tr}(PK_k P) \\ \frac{1}{2} \text{tr}(PPK_1) & \cdots & \frac{1}{2} \text{tr}(PPK_k) & \frac{1}{2} \text{tr}(PP) \end{bmatrix}, \quad (2.19)$$

L'information observée, c'est-à-dire l'opposé de la hessienne de la vraisemblance restreinte, est

$$\begin{aligned} J(\tau_1, \dots, \tau_k, \sigma^2) &= \\ &= -I(\tau_1, \dots, \tau_k, \sigma^2) + \begin{bmatrix} Y' PK_1 PK_1 PY & \cdots & Y' PK_1 PK_k PY & Y' PK_1 PPY \\ \vdots & \ddots & \vdots & \vdots \\ Y' PK_k PK_1 PY & \cdots & Y' PK_k PK_k PY & Y' PK_k PPY \\ Y' PPK_1 PY & \cdots & Y' PPK_k PY & Y' PPPY \end{bmatrix}. \end{aligned} \quad (2.20)$$

## 2.3. Tester les composantes de la variance

On suit ici [81, 82] pour construire un test du score pour  $\tau = 0$ , dans le cas  $k = 1$ . On va voir qu'on peut indifféremment utiliser la vraisemblance « classique » ou la vraisemblance restreinte. Commençons par la vraisemblance classique. La dérivée logarithmique suivant  $\tau$  est (équation 2.9)

$$\frac{\partial \ell}{\partial \tau}(\beta, \tau, \sigma^2) = -\frac{1}{2} \text{tr}(V^{-1}K) + \frac{1}{2}(Y - X\beta)'V^{-1}KV^{-1}(Y - X\beta)$$

Le test est construit en remplaçant  $\beta$  et  $V$  par leurs estimateurs du maximum de vraisemblance sous la contrainte  $\tau = 0$ . On a  $\widehat{V} = \widehat{\sigma}^2 I_n$ , et donc

$$\begin{aligned} \frac{\partial \ell}{\partial \tau}(\widehat{\beta}, 0, \widehat{\sigma}^2) &= -\frac{1}{2} \text{tr}(\widehat{V}^{-1}K) + \frac{1}{2}(Y - X\widehat{\beta})'\widehat{V}^{-1}K\widehat{V}^{-1}(Y - X\widehat{\beta}) \\ &= -\frac{1}{2\widehat{\sigma}^2} \text{tr}(K) + \frac{1}{2(\widehat{\sigma}^2)^2}(Y - X\widehat{\beta})'K(Y - X\widehat{\beta}). \end{aligned} \quad (2.21)$$

Estimer  $Y - X\widehat{\beta}$  sous la contrainte  $\tau = 0$  revient à l'estimer dans un modèle linéaire  $Y = X\beta + e$ , et on sait que  $Y - X\widehat{\beta} = P_0 Y$  où

$$P_0 = (I_n - X(X'X)^{-1}X').$$

Le terme  $\text{tr}(K)$  dans (2.21) ne dépend pas des observations  $Y$ , on peut donc l'abandonner. On néglige la variabilité de  $\widehat{\sigma}^2$  et on utilise la statistique de test

$$Q = (Y - X\widehat{\beta})'K(Y - X\widehat{\beta}) \quad (2.22)$$

$$= Y'P_0KP_0Y \quad (2.23)$$

Si on part de la vraisemblance restreinte, on obtient de la même façon

$$\frac{\partial \ell^{\text{re}}}{\partial \tau}(0, \widehat{\sigma}^2) = -\frac{1}{2} \text{tr}(\widehat{P}K) + \frac{1}{2}Y'\widehat{P}K\widehat{P}Y,$$

où  $\widehat{P}$  est donné par l'équation (2.12) avec  $\tau = 0$  et  $\sigma^2 = \widehat{\sigma}^2$ , c'est-à-dire  $\widehat{P} = \frac{1}{\widehat{\sigma}^2}P_0$ , et à nouveau on obtient une statistique de test construite sur  $YP_0KP_0Y$  comme dans (2.23).

Pour obtenir la distribution de  $Q$  sous l'hypothèse nulle, on écrit  $Y = X\beta + \sigma Z$  avec  $Z \sim \mathcal{N}(0, I_n)$ ; on a alors  $P_0 Y = \sigma P_0 Z$ . La statistique  $Q$  se ré-écrit

$$Q = Z'(\sigma^2 P_0 K P_0)Z.$$

Soit  $(\sigma^2 P_0 K P_0) = U\Lambda U'$  la décomposition en éléments propres de  $(\sigma^2 P_0 K P_0)$ , où  $\Lambda$  est la matrice diagonale des valeurs propres  $\lambda_1, \dots, \lambda_n$ , et  $U$  est une matrice orthogonale. La

distribution de  $U'Z$  reste  $\mathcal{N}(0, I_n)$ , et on en déduit que  $Q$  est une combinaison linéaire de variables  $\chi^2(1)$  indépendantes :

$$Q \sim \lambda_1 \chi^2(1) + \dots + \lambda_n \chi^2(1).$$

En pratique, on remplace  $\sigma^2$  par son estimateur  $\widehat{\sigma}^2$  – la distribution ci-dessus n'est plus alors qu'une approximation. Certains auteurs préfèrent utiliser la statistique de test  $Q' = \frac{1}{\widehat{\sigma}^2} Y' P_0 K P_0 Y$ , qui est distribuée de la même façon – en prenant cette fois ci comme coefficients de la combinaison linéaire de  $\chi^2(1)$  les valeurs propres de  $P_0 K P_0$ .

Il peut être utile de noter que si  $K$  est de la forme  $K = ZZ'$ , alors  $P_0 K P_0 = (P_0 Z)(Z' P_0)$  et  $Z' P_0 Z = (Z' P_0)(P_0 Z)$  ont les mêmes valeurs propres ; on peut utiliser la plus petite de ces deux matrices pour le calcul des valeurs propres.

Le calcul de la fonction de répartition de cette combinaison linéaire de  $\chi^2(1)$ , nécessaire à l'obtention d'un degré de significativité, n'est pas chose facile. L'usage s'est établi d'utiliser la méthode de Davies [83].

## 2.4. Estimation des composantes de la variance

### 2.4.1. L'algorithme EM-REML

L'algorithme EM vient naturellement à l'esprit quand on considère le modèle mixte : il est adapté à la maximisation de la vraisemblance de modèles « à variables latentes », ou à variables non observées ; ici, les variables latentes sont les effets aléatoires. Cet algorithme consiste à alterner deux étapes jusqu'à convergence : dans l'étape E, on calcule la distribution des variables latentes, qui dépend des valeurs courantes des paramètres ; dans l'étape M, on utilise cette distribution pour estimer de nouvelles valeurs des paramètres.

Nous montrons dans l'annexe A.2 que, pour la maximisation de la vraisemblance restreinte, cet algorithme est une forme d'« ascension de gradient ». Il consiste à itérer

$$\begin{aligned} \tau_{1,(r+1)} &= \tau_{1,r} + \left( \frac{2\tau_{1,r}^2}{r(K)} \right) \frac{\partial \ell^{\text{re}}}{\partial \tau} (\tau_{1,r}, \dots, \tau_{k,r}, \sigma_r^2) \\ &\vdots \\ \tau_{k,(r+1)} &= \tau_{k,r} + \left( \frac{2\tau_r^2}{r(K)} \right) \frac{\partial \ell^{\text{re}}}{\partial \tau} (\tau_{1,r}, \dots, \tau_{k,r}, \sigma_r^2) \\ \sigma_{r+1}^2 &= \sigma_r^2 + \frac{2(\sigma_r^2)^2}{n} \frac{\partial \ell^{\text{re}}}{\partial \sigma^2} (\tau_{1,r}, \dots, \tau_{k,r}, \sigma_r^2). \end{aligned}$$

où  $r(K)$  est le rang de  $K$ . Le pas de cette ascension de gradient dépend de la valeur courante des paramètres ; l'avantage est qu'on est assuré que les contraintes  $\tau_1, \dots, \tau_k, \sigma^2 > 0$  sont

vérifiées à chaque étape. Malheureusement la convergence de l'algorithme est (atrocement) lente, mais il reste intéressant de faire quelques étapes d'EM pour s'approcher du maximum avant d'utiliser une méthode du second ordre, comme l'algorithme AI-REML présenté ci-après.

### 2.4.2. L'algorithme AI-REML

L'algorithme « *Average Information* » REML peut être vu comme l'hybridation de deux algorithmes classiques de maximisation de la vraisemblance : l'algorithme de Newton-Raphson, et le Fisher Scoring.

Notons  $\theta$  le vecteur des paramètres de la vraisemblance (dans notre cas,  $\theta = (\tau_1, \dots, \tau_r, \sigma^2)$ ). L'algorithme de Newton-Raphson consiste à itérer la suite

$$\theta^{r+1} = \theta^r + J(\theta^r)^{-1}U(\theta^r)$$

où le score  $U(\theta)$  est le gradient de la log-vraisemblance, et  $J(\theta)$  est la matrice d'information observée (2.20) (donc l'opposée de la hessienne de la log-vraisemblance), alors que le Fisher Scoring itère

$$\theta^{r+1} = \theta^r + I(\theta^r)^{-1}U(\theta^r)$$

où  $I(\theta)$  est la matrice d'information de Fisher (2.19) (donc l'espérance de  $J(\theta)$ ).

Le calcul de  $I(\theta)$  et  $J(\theta)$  est très lourd. L'algorithme AI-REML utilise, en place de  $J(\theta)$  ou  $I(\theta)$ , l'information moyenne  $AI(\theta) = \frac{1}{2} (I(\theta) + J(\theta))$ , c'est-à-dire

$$AI(\tau_1, \dots, \tau_k, \sigma^2) = \frac{1}{2} \begin{bmatrix} Y'PK_1PK_1PY & \dots & Y'PK_1PK_kPY & Y'PK_1PPY \\ \vdots & \ddots & \vdots & \vdots \\ Y'PK_kPK_1PY & \dots & Y'PK_kPK_kPY & Y'PK_kPPY \\ Y'PPK_1PY & \dots & Y'PPK_kPY & Y'PPPY \end{bmatrix}.$$

Une fois que  $P$  a été calculé, chaque entrée de cette matrice peut être calculée en  $O(n^2)$  opérations, alors que les termes de la forme  $\text{tr}(PPK_1)$  qui apparaissent dans  $I(\theta)$  et dans  $J(\theta)$  se calculent en  $O(n^3)$  opérations.

Le calcul de l'inverse de  $AI(\theta)$  est en  $O(k^3)$ ;  $k$  étant de taille modérée, ça n'est pas ce terme qui domine la complexité algorithmique, mais le calcul de  $P$  qui reste très coûteux : il est en  $O(n^3)$  également. Comme la connaissance de cette matrice est nécessaire pour calculer le gradient de la log-vraisemblance, on peut raisonnablement penser qu'on ne peut pas significativement alléger les calculs simplement en remplaçant la matrice  $AI(\theta)$  par une autre matrice définie positive (comme le font les méthodes dites de quasi-Newton).



## 2.5. Estimation et prédiction des effets

Nous nous intéressons maintenant au problème de l'estimation des effets fixes, et de la prédiction des effets aléatoires. Supposons tout d'abord que les valeurs des paramètres  $\tau_1, \dots, \tau_k, \sigma^2$  soient connues. Les effets fixes sont estimés par

$$\widehat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y$$

ainsi que nous l'avons remarqué plus haut (équation 2.11). La variance de  $\widehat{\beta}$  est

$$\text{var}(\widehat{\beta}) = (X'V^{-1}X)^{-1}. \quad (2.24)$$

Les estimateurs  $\widehat{u}_1, \dots, \widehat{u}_k, \widehat{e}$  des termes aléatoires doivent vérifier

$$Z_1\widehat{u}_1 + \dots + Z_k\widehat{u}_k + \widehat{e} = Y - X\widehat{\beta} = VPY.$$

Les valeurs les plus vraisemblables de ces variables sont celles qui maximisent la densité jointe de  $(u_1, \dots, u_k, e)$ , sous cette contrainte. Cette densité est proportionnelle à

$$\exp\left(-\frac{1}{2}\left(\frac{1}{\tau_1}u_1'u_1 + \dots + \frac{1}{\tau_k}u_k'u_k + \frac{1}{\sigma^2}e'e\right)\right).$$

Il suffit donc de minimiser  $\left(\frac{1}{\tau_1}u_1'u_1 + \dots + \frac{1}{\tau_k}u_k'u_k + \frac{1}{\sigma^2}e'e\right)$ . La méthode des multiplicateurs de Lagrange donne

$$\begin{cases} \frac{1}{\tau_1}\widehat{u}_1 = Z_1'\lambda \\ \vdots \\ \frac{1}{\tau_k}\widehat{u}_k = Z_k'\lambda \\ \frac{1}{\sigma^2}\widehat{e} = \lambda \end{cases}$$

avec  $\lambda \in \mathbb{R}^n$ . On en déduit

$$Z_1\widehat{u}_1 + \dots + Z_k\widehat{u}_k + \widehat{e} = \tau_1 Z_1 Z_1' \lambda + \dots + \tau_k Z_k Z_k' \lambda + \sigma^2 \lambda = V\lambda,$$

et il suffit de poser  $\lambda = PY$  pour satisfaire la contrainte. Finalement les valeurs cherchées sont

$$\begin{aligned} \widehat{u}_1 &= \tau_1 Z_1' PY \\ &\vdots \\ \widehat{u}_k &= \tau_k Z_k' PY \\ \widehat{e} &= \sigma^2 PY. \end{aligned} \quad (2.25)$$

Ce sont les *Best Linear Unbiased Predictors (BLUP)*, des prédicteurs des variables aléatoires non observées. Les lois étant normales, c'est aussi le cas des lois conditionnelles considérées ci-dessus, et leur mode coïncide avec leur espérance : ils peuvent également être définis comme l'espérance des  $u_i$  conditionnellement à la valeur observée de  $Y$ .

Chacun des  $\widehat{u}_i$  est dans l'espace vectoriel engendré par les lignes de  $Z_i$ . Si  $n \ll q_i$ , ce qui est souvent le cas dans les applications considérées en génomique, c'est un tout petit sous-espace de  $\mathbb{R}^n$ , et  $\widehat{u}_i$  est également comme la prédiction de la projection de  $u_i$  sur ce sous-espace.

Quand les valeurs des paramètres ne sont pas connues, on les remplace dans les équations (2.25) par leurs estimateurs du maximum de la vraisemblance restreinte : on obtient  $\widehat{u}_1 = \widehat{\tau}_1 Z_1' \widehat{P} Y$ , etc. On appelle ces prédicteurs les *eBLUP (empirical BLUP)*.

Remarquons pour finir que les prédicteurs des valeurs génétiques  $\omega_1, \dots, \omega_k$  sont

$$\begin{aligned} \widehat{\omega}_1 &= Z_1 \widehat{u}_1 = \tau_1 K_1 P Y \\ &\vdots \\ \widehat{\omega}_k &= Z_k \widehat{u}_k = \tau_k K_k P Y \end{aligned} \tag{2.26}$$

## 2.6. La variance expliquée par les effets fixes

En pratique, il est fréquent que les covariables  $X$  et  $Z$  incluses dans le modèle ne soient pas fixées par l'expérimentateur, mais mesurées en même temps que la variable d'intérêt  $Y$ . Dans une telle « étude observationnelle », il est naturel de s'intéresser à la variance de  $Y$  dans la population. Pour ce faire, la démarche naturelle est de traiter les différentes mesures  $Y_i$  comme indépendantes. À première vue, cette démarche ne s'inscrit pas naturellement dans le cadre du modèle mixte, qui cherche au contraire à modéliser la covariance entre les  $Y_i$ .

Nous montrons ici comment analyser le calcul de la variance des  $Y_i$ , et comment décomposer le résultat en variance expliquée par les effets fixes et variance expliquée par les différents effets aléatoires.

La variance empirique des composantes du vecteur  $Y \in \mathbb{R}^n$  est

$$\text{ev}(Y) = \frac{1}{n-1} \left( Y'Y - \frac{1}{n} (1_n' Y)^2 \right) \tag{2.27}$$

Définissons une forme linéaire  $\Psi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$  par

$$\Psi(A) = \frac{1}{n-1} \left( \text{tr}(A) - \frac{1}{n} 1_n' A 1_n \right). \tag{2.28}$$

C'est la moyenne des termes diagonaux de  $A$ , moins la moyenne des termes non-diagonaux. On a  $\text{ev}(Y) = \Psi(Y Y')$  pour tout  $Y \in \mathbb{R}^n$ .

La linéarité de  $\Psi$  implique

$$\begin{aligned}
 E(\text{ev}(Y)) &= E(\Psi(YY')) \\
 &= \Psi(E(YY')) \\
 &= \Psi(E(Y)E(Y)' + \text{Var}(Y)) \\
 &= \text{ev}(E(Y)) + \Psi(\text{Var}(Y))
 \end{aligned} \tag{2.29}$$

Dans notre cadre avec  $E(Y) = X\beta$  et  $\text{Var}(Y) = V = \tau_1 K_1 + \dots + \tau_k K_k + \sigma^2 I_n$ , on a

$$E(\text{ev}(Y)) = \text{ev}(X\beta) + \Psi(V). \tag{2.30}$$

Le terme  $\text{ev}(X\beta)$  est la variance expliquée par les covariables incluses dans  $X$ , et le terme  $\Psi(V)$  est la variance due aux différents effets aléatoires. Par linéarité de  $\Psi$ , on a

$$\Psi(V) = \tau_1 \Psi(K_1) + \dots + \tau_k \Psi(K_k) + \sigma^2 \Psi(I_n) = \tau \Psi(K) + \sigma^2.$$

Notons que si les matrices  $K_\ell$  ont sur leur diagonale des termes proches de 1, et que les termes hors diagonale sont proches de 0, on a  $\Psi(K_\ell) \simeq 1$  et

$$E(\text{ev}(Y)) \simeq \text{ev}(X\beta) + \tau_1 + \dots + \tau_k + \sigma^2.$$

Cette situation sera typique de certaines matrices calculées dans les applications en génétique humaine (en particulier, dans les estimations d'héritabilité).

Les estimateurs du maximum de vraisemblance restreinte permettent d'estimer sans biais la part de variance attribuée aux effets aléatoires. En revanche, en substituant  $\hat{\beta}$  à  $\beta$  dans  $\text{ev}(X\beta)$ , on obtient un estimateur biaisé

$$\begin{aligned}
 E(\text{ev}(X\hat{\beta})) &= \text{ev}(X\beta) + \Psi(\text{Var}(X\hat{\beta})) \\
 &= \text{ev}(X\beta) + \Psi(X\text{Var}(\hat{\beta})X') \\
 &= \text{ev}(X\beta) + \Psi\left(X(X'V^{-1}X)^{-1}X'\right).
 \end{aligned}$$

Pour réduire le biais, on estimera  $\text{ev}(X\beta)$  par :

$$\text{ev}(X\hat{\beta}) - \Psi\left(X(X'\hat{V}^{-1}X)^{-1}X'\right).$$

## 2.7. L'astuce de la diagonalisation

Nous nous plaçons ici dans le cas où  $k = 1$  (on peut omettre l'indice sur  $Z = Z_1$ ,  $K = K_1$ ). Pour faciliter l'ensemble des calculs présentés dans les sections qui précèdent, il est possible de réécrire le modèle sous forme diagonale en utilisant la décomposition en éléments

simples de  $K$ ,  $K = U\Sigma^2U'$ , où  $U$  est orthogonale (i.e.  $UU' = U'U = I_n$ ) et  $\Sigma$  est une matrice diagonale. La variance de  $Y_D = (U'Y)$  est donc

$$U'(\tau K + \sigma^2 I_n)U = \tau \Sigma^2 + \sigma^2 I_n$$

qui est une matrice diagonale ; son inverse se calcule en  $O(n)$  opérations, ce qui accélère grandement le calcul de la vraisemblance restreinte (2.14) et son optimisation. Le coût du calcul préliminaire, la décomposition de  $K$  en éléments simples, est cependant non négligeable – il est en  $O(n^3)$  – et cette méthode est avant tout utile quand la décomposition de  $K$  est par ailleurs nécessaire pour d'autres buts (par exemple, l'analyse en composantes principales des données génomiques contenues dans  $Z$ ) ou quand on doit estimer les paramètres de plusieurs modèles faisant intervenir la même matrice  $K$ , avec par exemple des effets fixes différents.

Cette stratégie a été utilisée dans EMMA [84, 85] ou dans FaST [86]. En annexe A.3, nous développons l'utilisation de cette astuce en détail, avec une attention particulière au cas où  $X$  comprend certaines des composantes principales de  $Z$ , qui sont les colonnes de  $U\Sigma$ .

## Chapitre 3.

# L'héritabilité génomique

Dans ce chapitre, dont le contenu est largement tiré de l'article plus vaste mais parfois moins détaillé écrit avec Claire Dandine [19], nous décrivons deux applications maintenant communes du modèle linéaire mixte dans la pratique contemporaine de la génétique humaine : l'estimation de l'héritabilité génomique, et la prise en compte d'une structure de population dans les études d'association de génome entier. Ces deux applications utilisent sensiblement le même modèle. Nous nous intéressons également aux performances prédictives du modèle linéaire mixte pour la prédiction de la contribution de l'ensemble du génome à un phénotype quantitatif.

### 3.1. À la recherche de l'héritabilité perdue

Après Galton, Pearson et Fisher, on a longtemps continué à estimer l'héritabilité en se basant sur la corrélation entre apparentés. Les études de jumeaux en particulier ont été très populaires, car elles permettaient de prendre en compte (dans une certaine mesure) l'environnement partagé dans les familles, dont la présence biaise positivement les estimations. Cependant, si les effets environnementaux sont mieux corrélés chez les jumeaux monozygotes que chez les jumeaux dizygotes, le biais subsiste [87–89].

Par ailleurs, les études d'association avec le génome entier ont permis la recherche des polymorphismes génétiques à la source de la variabilité phénotypique. Des centaines de SNP ont été associés avec des phénotypes polygéniques [90, 91]. Cependant la part de variance expliquée par ces SNP est très inférieure à ce que prédisent les études familiales ; la différence a été appelée « héritabilité manquante » [79, 92–96].

Cette idée d'héritabilité manquante a été une des motivations au développement de méthodes, basées sur l'analyse de données de génome entier par un modèle mixte, pour estimer l'héritabilité à partir d'individus non apparentés [78, 79].

## 3.2. Héritabilité génomique

Nous allons exposer ici l'utilisation du modèle mixte pour estimer « l'héritabilité génomique » à partir non plus de données familiales, mais d'échantillons de population. Cette méthode est souvent présentée comme ayant l'avantage de régler le problème de l'environnement partagé : les individus ne sont pas des apparentés proches, ce qui exclut le biais introduit par les environnements familiaux. Pour autant, cela n'exclut pas l'existence de corrélations gène-environnement (nous y reviendrons au chapitre suivant). Plusieurs méthodes similaires ont été proposées pour l'estimation de l'héritabilité génomique, [76–79, 97]. Ici nous nous contenterons de présenter rapidement ce qui est utilisé le plus souvent en pratique, et implémenté dans des logiciels comme GCTA.

### 3.2.1. Rappel : le modèle de Fisher

Rappelons que le modèle de Fisher pour la valeur d'un phénotype quantitatif chez un individu d'indice  $i$  s'écrit, en omettant les termes de dominance (nous y reviendrons) :

$$Y_i = \mu + \underbrace{\alpha_1 X_{ai1} + \dots + \alpha_r X_{air}}_{G_i} + \varepsilon_i \quad (3.1)$$

où les  $X_{aij}$  ( $j = 1, \dots, r$ ) sont les composantes additives des facteurs mendéliens impliqués (définis par l'équation 1.5), les effets  $\alpha_j$  sont des constantes et les résidus  $\varepsilon_i$ , qui représentent les effets environnementaux (ou purement aléatoires), sont gaussiens et indépendants :  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . Dans ce modèle, ce sont les génotypes  $X_{aij}$  qui sont considérés aléatoires ; ils sont centrés et réduits par définition. On suppose également que les génotypes aux SNP  $j$  et  $j'$  (avec  $j \neq j'$ ) sont indépendants (pas de déséquilibre gamétique) :  $\text{cov}(X_{aij}, X_{aij'}) = 0$  et donc  $\text{var}(G_i) = \alpha_1^2 + \dots + \alpha_r^2 = \tau_a$ . On suppose en outre que les  $X_{aij}$  et  $\varepsilon_i$  sont indépendants (pas de corrélation gène-environnement), de sorte que  $\text{var}(Y_i) = \tau_a + \sigma^2$ . Rappelons que l'héritabilité (au sens étroit) est

$$h^2 = \frac{\tau_a}{\tau_a + \sigma^2}.$$

Quand deux individus  $i$  et  $i'$  ont pour coefficient de parenté  $\phi_{ii'}$ , on a, pour tout  $j$ ,  $\text{cov}(X_{ij}, X_{i'j}) = 2\phi_{ii'}$ , donc  $\text{cov}(G_i, G_{i'}) = 2\phi_{ii'}\tau_a$ . L'indépendance des facteurs environnementaux implique  $\text{cov}(Y_i, Y_{i'}) = 2\phi_{ii'}\tau_a$ , et  $\text{cor}(Y_i, Y_{i'}) = 2\phi_{ii'}h^2$ . Plus généralement, si  $Y = (Y_1, \dots, Y_n)'$  est un vecteur de phénotypes mesuré chez des individus de coefficients de parenté  $\phi_{ii'}$ , si on note  $\Phi$  la matrice de terme général  $\phi_{ii'}$ , on a  $\text{var}(Y) = \tau \cdot (2\Phi) + \sigma^2 I_n$ . De façon équivalente,  $Y$  suit un modèle mixte écrit comme en (2.4),  $Y = \mathbf{1}_n \mu + \omega + \varepsilon$ , avec  $\text{var}(\omega) = \tau \cdot (2\Phi)$  et  $\text{var}(\varepsilon) = \sigma^2 I_n$ .

### 3.2.2. L'héritabilité génomique

#### L'héritabilité restreinte

L'estimation de l'héritabilité génomique par le modèle mixte suppose qu'on dispose, pour  $n$  individus (non apparentés), des génotypes en  $r$  SNP, couvrant la totalité des autosomes. Le modèle utilisé repose sur la même écriture (3.1) que le modèle de Fisher, interprétée comme un modèle linéaire mixte : les génotypes  $X_{aij}$ , qui étaient aléatoires, sont maintenant considérés comme des constantes, et ce sont les effets  $\alpha_j$  qui sont des variables aléatoires indépendantes de variance  $\frac{1}{r}\tau$ ,  $\alpha_j \sim \mathcal{N}\left(0, \frac{1}{r}\tau\right)$ . En notant  $\mathbf{X}_a$  la matrice des  $X_{aij}$ , on peut l'écrire sous forme matricielle

$$\mathbf{Y} = \mathbf{1}_n\mu + \mathbf{X}_a\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \quad (3.2)$$

avec  $\boldsymbol{\alpha} \sim \mathcal{N}\left(0, \frac{\tau}{r}\mathbf{I}_r\right)$  et  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ . La variance de  $\mathbf{Y}$  est  $\text{var}(\mathbf{Y}) = \tau\mathbf{K} + \sigma^2\mathbf{I}_n$ , avec  $\mathbf{K} = \frac{1}{r}\mathbf{X}_a\mathbf{X}_a'$ .

Malgré la similarité des écritures, il y a de grandes différences entre le modèle de Fisher et celui-ci :

- on raisonne conditionnellement aux génotypes, ce qui élimine la nécessité de supposer l'équilibre gamétique ;
- on suppose que tous les SNP génotypés ont un effet, alors que dans le modèle de Fisher il suffit que le nombre  $r$  de locus avec un effet soit assez grand pour que  $\sum_{j=1}^r \alpha_j X_{aij}$  soit approximativement normal ;
- les effets  $\alpha_j$  sont tous pris dans la même loi, avec le même écart-type ; la définition des  $X_{aij}$  implique que l'effet allélique au SNP  $j$  est  $\frac{\alpha_j}{\sqrt{p_j q_j}}$ , et son écart-type est donc inversement proportionnel à  $\sqrt{p_j q_j}$ .

En pratique, la matrice  $\mathbf{X}_a$  est construite à partir des génotypes « bruts »  $X_{ij}$ , codés par 0, 1 ou 2, au moyen de la transformation

$$X_{aij} = \frac{X_{ij} - 2q_j}{\sqrt{2p_j q_j}}, \quad (3.3)$$

où  $p_j$  et  $q_j$  sont les fréquences (empiriques) des allèles de référence et alternatif, respectivement ; le dénominateur est l'écart-type de  $X_{ij}$  sous le modèle d'Hardy-Weinberg (cf 1.5). Ainsi, les colonnes de  $\mathbf{X}_a$  sont centrées et (approximativement) réduites. La matrice  $\mathbf{K} = \frac{1}{r}\mathbf{X}_a\mathbf{X}_a'$  est appelée en anglais *Genetic Relationship Matrix*, ou GRM – on pourra en français parler de « matrice de corrélation génétique ».

On estime les paramètres  $\tau$  et  $\sigma^2$  du modèle par la méthode du maximum de vraisemblance restreinte, et l'héritabilité  $h^2$  par

$$\widehat{h^2} = \frac{\widehat{\tau}}{\widehat{\tau} + \widehat{\sigma^2}}.$$

En présence de covariables connues ayant un effet sur le génotype, on peut les inclure dans le modèle avec un effet fixe, en ajoutant un terme  $X\beta$  au modèle. Dans ce cas, on estime la plupart du temps l'héritabilité par la même formule, c'est-à-dire comme la proportion de la variance *résiduelle* (non expliquée par les covariables) qui est due à des facteurs génétiques additifs [93]. Il est également possible de partitionner la variance de  $Y$  en trois composantes (la variance expliquée par les covariables, celle qui est due aux effets génétiques aléatoires, et la variance résiduelle), comme décrit en section 2.6 (nous sommes bien dans le cas où  $\Psi(K) \simeq 1$  à laquelle il est fait allusion dans le texte).

#### **Le modèle mixte : astuce mathématique ou description du réel ?**

Dans le cas d'individus apparentés, la GRM  $K$  est un estimateur de la matrice  $2\Phi$  – cela découle immédiatement des calculs effectués à la fin de la section 1.3.2. Si dans 3.3 on utilise les fréquences  $p_j$  de la population dont est extrait l'échantillon, et non les fréquences empiriques, c'est un estimateur non biaisé :  $E(K) = 2\Phi$ . Dans le cas général, cela reste un estimateur convergent. Cette remarque peut servir à justifier (dans une certaine mesure) le changement de modèle : dans cette optique, le modèle mixte n'est pas à prendre au pied de la lettre, c'est un artifice mathématique qui n'a pour intérêt que de produire la « bonne » structure de variance. Dans le cas d'un échantillon sans apparentements proches, la GRM peut s'interpréter comme une matrice qui capture des apparentements cryptiques, ou du moins, si on préfère éviter le terme « apparentement », qui estime une covariance génomique supposée être, en espérance, la même en tout point du génome.

Cependant la tentation est forte de se laisser aller à une interprétation plus littérale. L'efficacité pratique de cette méthode semble avoir poussé certains chercheurs à se laisser aller à une forme de réalisme scientifique, en considérant le modèle mixte comme une image fidèle de la réalité, sinon comme un reflet exact de celle-ci.

Si on adopte ce point de vue, l'efficacité du modèle tient à sa capacité à capturer une composante polygénique, à travers le vecteur des effets aléatoires (ou modélisés comme tels)  $\alpha$ . Le postulat que l'effet allélique du SNP  $j$  a une variance proportionnelle à  $\frac{1}{p_j q_j}$  n'a alors plus aucune raison d'être. D'autres relations entre l'effet allélique et les fréquences alléliques, par exemple de la forme  $(p_j q_j)^\gamma$  avec  $\gamma$  arbitraire, peuvent être postulées [97]. Alternativement, on peut créer des groupes de SNP sur la base de la fréquence de l'allèle mineur, et étendre le modèle de façon à avoir un paramètre de variance pour chaque groupe de SNP [77]. Identifier les héritabilités calculées par ces méthodes à l'héritabilité définie dans le modèle de Fisher est délicat – dans le cas d'individus apparentés, les matrices de covariances impliquées ne sont plus des estimateurs de la matrice  $2\Phi$ , et la relation entre l'héritabilité et la corrélation phénotypique entre apparentés disparaît.

La méthode est souvent vantée et perçue comme non biaisée, en particulier parce que, les individus n'étant pas apparentés, il n'y aurait pas d'environnement partagé [98]. Il a cependant été démontré très tôt que la présence d'une structure de population conduit à une sur-estimation de l'héritabilité [99]. En effet, un facteur environnemental associé au



phénotype étudié peut différer d'une strate de population à une autre. La structure de population peut également correspondre à un continuum : certains facteurs, comme le niveau d'exposition aux ultra-violets, peuvent varier avec la latitude, ce qui est aussi le cas de la fréquence allélique de nombreux SNP. Dans ce cas de figure, ce facteur associé à la fois au phénotype et à la variation génétique entre les strates de population agit comme une variable de confusion. La solution communément proposée pour corriger ce biais est l'inclusion dans le modèle de quelques composantes principales génomiques, avec des effets fixes [100–104]. Nous verrons au chapitre 4 que cela n'est pas toujours suffisant.

### Estimation de la variance de dominance

Il n'est pas difficile d'étendre la méthode au calcul de la variance de dominance [105] par le biais d'un modèle mixte. Dans le modèle de Fisher la dominance correspond à des termes  $X_d$  que nous avons omis dans (3.1) :

$$Y_i = \mu + \underbrace{\alpha_1 X_{ai1} + \dots + \alpha_r X_{air}}_{\text{composante additive}} + \underbrace{\delta_1 X_{di1} + \dots + \delta_r X_{dir}}_{\text{composante de dominance}} + \varepsilon_i. \quad (3.4)$$

Les  $X_d$  sont des termes aléatoires, obtenus à partir des génotypes par l'équation (1.6) et les effets  $\delta_1, \dots, \delta_r$  sont des paramètres du modèle. À nouveau on procède à un échange entre les statuts de constantes et de variables aléatoires. Ceci revient à définir une matrice  $X_d$  contenant les termes  $X_{dij}$ , et à considérer le modèle mixte

$$Y = \mathbf{1}_n \mu + X_a \alpha + X_d \delta + \varepsilon \quad (3.5)$$

avec  $\alpha \sim \mathcal{N}\left(0, \frac{\tau_a}{r} I_r\right)$ ,  $\delta \sim \mathcal{N}\left(0, \frac{\tau_d}{r} I_r\right)$ , et  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ . On note  $D = X_d X_d'$ , et la variance de  $Y$  est  $\text{var}(Y) = \tau_a K + \tau_d D + \sigma^2 I_n$ .

Dans le cas d'individus apparentés, de même que les composantes de  $K$  estiment le double des coefficients d'apparentement, les composantes de  $D$  estiment la probabilité  $\psi$  que les individus soient  $\text{IBD} = 2$  (cf section 1.3.2).

Il ne reste qu'à estimer les paramètres du modèle par l'algorithme AIREML. On peut ensuite estimer l'héritabilité restreinte par

$$\widehat{h^2} = \frac{\widehat{\tau}_a}{\widehat{\tau}_a + \widehat{\tau}_d + \widehat{\sigma}^2}$$

et une composante de dominance, correspondant au terme  $H^2 - h^2$ , par

$$\widehat{h_d^2} = \frac{\widehat{\tau}_d}{\widehat{\tau}_a + \widehat{\tau}_d + \widehat{\sigma}^2}.$$

Notons que l'orthogonalité de  $X_a$  et  $X_d$  définis par (1.5) et (1.6) n'implique pas nécessairement que les estimations  $\tau_a$  et  $\tau_d$  sont indépendantes l'une de l'autre.

### 3.3. Prise en compte d'une structure de population dans les tests d'association

Les petits effets des variants fréquents ciblés dans les analyses d'association du génome entier (qui n'ont la plupart du temps que des effets indirects dûs à l'existence d'un déséquilibre de liaison avec un variant causal non génotypé) et le grand nombre de variants à analyser rendent nécessaire d'inclure dans ces études un grand nombre d'individus de façon à atteindre une puissance statistique satisfaisante. Pour analyser des échantillons d'une taille aussi importante, il est nécessaire de prendre en compte la structure de population pour éviter les « fausses associations » créées par cette dernière. Une solution populaire et simple à mettre en œuvre est la régression sur les composantes principales (PCR), qui inclut quelques composantes principales des données génomiques dans un modèle linéaire [100] :

$$Y = X\beta + PC_1\gamma_1 + \dots + PC_k\gamma_k + \varepsilon \quad (3.6)$$

où la matrice  $X$  intègre des covariables cliniques et le génotype d'un SNP dont on teste l'association avec le phénotype (par exemple, par un test de Wald). Ici, les effets  $\gamma_1, \dots, \gamma_k$  des composantes principales sont des effets fixes. Les premières composantes principales reflétant bien la structure de population [106, 107], leur présence dans le modèle réduit l'impact de celle-ci.

D'autres auteurs [84–86, 108, 109] ont proposé d'utiliser le modèle mixte

$$Y = X\beta + X_a\alpha + \varepsilon$$

dans le même but. La matrice  $X$  est ici la même que dans le modèle (3.6), et  $X_a$  est, comme en (3.3), la matrice  $n \times r$  des génotypes centrés réduits, en excluant éventuellement la région génomique (ou tout le chromosome) où se trouve le SNP à tester. Le terme  $X_a\alpha$  est interprété comme une composante polygénique, modélisant l'effet de l'ensemble du génome.

Cependant une autre interprétation est possible : ce modèle peut être relié à la PCR par la décomposition en valeurs singulières (SVD) de  $\frac{1}{\sqrt{r}}X_a$  [110], qui s'écrit  $\frac{1}{\sqrt{r}}X_a = U\Sigma L'$ , où  $U$  est une matrice orthogonale de dimensions  $n \times n$ ,  $\Sigma$  est une matrice diagonale de dimensions  $n \times n$  également, et  $L$  est une matrice de dimensions  $q \times n$  qui vérifie  $L'L = I_n$ .

Les composantes principales génomiques utilisées dans la PCR sont justement les colonnes de  $U\Sigma$  ; les colonnes de  $L$  contiennent les *loadings* correspondant, c'est-à-dire qu'on a  $U\Sigma = ZL$ . Soit  $w = (w_1, \dots, w_n)' = \sqrt{r}L'\alpha$  ; c'est un vecteur aléatoire qui suit une loi normale multivariée, d'espérance nulle et de variance  $rL' \left( \frac{\tau}{r} I_q \right) L = \tau I_n$ . Alors le modèle mixte considéré se ré-écrit

$$\begin{aligned} Y &= X\beta + X_a\alpha + \varepsilon \\ &= X\beta + (U\Sigma)(\sqrt{r}L'\alpha) + \varepsilon \\ &= X\beta + PC_1w_1 + \dots + PC_nw_n + \varepsilon, \end{aligned} \quad (3.7)$$

avec  $w = (w_1, \dots, w_n) \sim \mathcal{N}(0, \tau I_n)$ .

On peut donc voir le modèle mixte comme une simple extension de la PCR qui intègre au modèle *toutes* les composantes principales, avec des effets aléatoires en lieu et place des effets fixes. Son efficacité peut être attribuée à sa capacité à modéliser la structure de population à l'aide d'un grand nombre de composantes principales, tout en évitant le surajustement qui rendrait le modèle (3.6) inopérant si  $k$  était trop grand.

Cette méthode semble plus efficace que la PCR en présence de structure de population [85, 101], alors que selon certains auteurs la PCR peut être plus efficace en présence de variables de confusion environnementales [101]. Il est possible d'ajouter quelques composantes principales avec des effets fixes au modèle linéaire (3.7) [101] :

$$Y = X\beta + PC_1\gamma_1 + \dots + PC_k\gamma_k + PC_{k+1}w_{k+1} + \dots + PC_nw_n + \varepsilon,$$

où  $\gamma_1, \dots, \gamma_k$  sont des effets fixes et  $w_{k+1}, \dots, w_n \sim \mathcal{N}(0, \tau)$  des effets aléatoires.

Les deux interprétations possibles du modèle (composante polygénique ou généralisation de la PCR) peuvent conduire à des choix différents de  $X_a$  : la première interprétation mène à inclure dans  $X_a$  le plus grand nombre de SNP possibles, y compris des variants rares [100], alors que dans la pratique de la PCR il est recommandé d'enlever les régions génomiques où existe un déséquilibre de liaison étendu, et même de ne conserver qu'un ensemble de SNP en faible LD mutuel, c'est-à-dire  $r^2 < 0,2$  ou  $r^2 < 0,1$  [111, 112]. L'impact de ce choix ne semble pas avoir été particulièrement commenté dans la littérature.

## 3.4. Performances prédictives

### 3.4.1. Performance des prédictions avec les BLUP

Nous considérons ici l'utilisation du modèle (3.2) pour la prédiction d'un phénotype centré  $Y$  (on prend donc  $\mu = 0$ ). On divise un échantillon de  $n$  individus en un échantillon d'apprentissage de taille  $n_1$ , dont les génotypes centrés-réduits sont rassemblés dans une matrice  $X_{a1}$  de dimension  $n_1 \times r$ , et un échantillon de test de taille  $n_2$ , dont les génotypes associés sont dans la matrice  $X_{a2}$  de taille  $n_2 \times r$ .

La GRM de l'échantillon entier est

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

avec  $K_{ij} = \frac{1}{q} X_{ai} X'_{aj}$  (on a  $K'_{21} = K_{12}$ ). Pour rendre les calculs littéraux praticables, nous supposons ici que les valeurs de  $\tau$  et de  $\sigma^2$  sont connues (en pratique, on utiliserait leurs estimations  $\hat{\tau}$  et  $\hat{\sigma}^2$ ).

En utilisant le modèle mixte  $Y_1 = X_{a1}\alpha + \varepsilon_1$  pour estimer les effets génétiques  $\alpha$  par leur BLUP  $\hat{\alpha}$ , on peut prédire les phénotypes du deuxième échantillon par  $\hat{Y}_2 = X_{a2}\hat{\alpha}$ . En remplaçant  $\hat{\alpha}$  par sa valeur (équation 2.25), on obtient

$$\begin{aligned}\hat{Y}_2 &= \tau K_{21} \left( \tau K_{11} + \sigma^2 I_{n_1} \right)^{-1} Y_1 \\ &= K_{21} \left( K_{11} + \frac{1-h^2}{h^2} I_{n_1} \right)^{-1} Y_1.\end{aligned}\tag{3.8}$$

C'est simplement l'espérance de  $Y_2 = X_{a2}\alpha + \varepsilon_2$  conditionnellement à la valeur observée de  $Y_1$  ; on peut l'obtenir directement en utilisant le fait que  $Y = (Y_1', Y_2')'$  suit une loi  $\mathcal{N}(0, \tau K + \sigma^2 I_n)$ .

On définit deux coefficients d'ajustement basés sur l'écart quadratique moyen entre les valeurs  $Y_{2i}$  du phénotype dans l'échantillon de test, et sa valeur prédite :

$$1 - R_{\text{gen}}^2 = \frac{E \left( \left( \hat{Y}_{2i} - X_{a2i}\alpha \right)^2 \right)}{\text{var}(X_{a2i}\alpha)} \quad \text{et} \quad 1 - R_{\text{tot}}^2 = \frac{E \left( \left( \hat{Y}_{2i} - Y_{2i} \right)^2 \right)}{\text{var}(Y_{2i})}.$$

On a  $\text{var}(X_{a2i}\alpha) = \tau$  et  $\text{var}(Y_{2i}) = \tau + \sigma^2$ . De plus, comme  $E(\hat{Y}_{2i}) = E(Y_{2i}) = E(\alpha) = 0$ , on a

$$\begin{aligned}R_{\text{gen}}^2 &= \frac{1}{\text{var}(X_{a2i}\alpha)} \left( 2 \text{cov}(\hat{Y}_{2i}, X_{a2i}\alpha) - \text{var}(\hat{Y}_{2i}) - E(\hat{Y}_{2i} - X_{a2i}\alpha)^2 \right) \\ &= \frac{1}{\tau} \left( 2 \text{cov}(\hat{Y}_{2i}, X_{a2i}\alpha) - \text{var}(\hat{Y}_{2i}) \right). \\ R_{\text{tot}}^2 &= \frac{1}{\text{var}(Y_{2i})} \left( 2 \text{cov}(\hat{Y}_{2i}, Y_{2i}) - \text{var}(\hat{Y}_{2i}) - E(\hat{Y}_{2i} - Y_{2i})^2 \right) \\ &= \frac{1}{\tau + \sigma^2} \left( 2 \text{cov}(\hat{Y}_{2i}, Y_{2i}) - \text{var}(\hat{Y}_{2i}) \right)\end{aligned}$$

Les BLUP  $\hat{\alpha}$  sont indépendants de  $\varepsilon_2$ , donc  $\hat{Y}_{2i}$  l'est également. On a donc  $\text{cov}(\hat{Y}_{2i}, Y_{2i}) = \text{cov}(\hat{Y}_{2i}, X_{a2i}\alpha)$ . On a donc  $R_{\text{tot}}^2 = h^2 R_{\text{gen}}^2$  ; on ne s'intéressera donc qu'à  $R_{\text{gen}}^2$ , qui est plus facilement interprétable.

Si la valeur  $Y_2$  du phénotype des individus de l'échantillon de test est connue,  $\text{cov}(\hat{Y}_{2i}, X_{a2i}u) = \text{cov}(\hat{Y}_{2i}, Y_{2i})$  est estimée par

$$\frac{1}{n_2} Y_2' \hat{Y}_2.$$

On peut calculer l'espérance de cette expression. En utilisant (3.8), on écrit

$$\begin{aligned} E\left(\frac{1}{n_2} Y_2' \widehat{Y}_2\right) &= \frac{1}{n_2} E\left(\text{tr}\left(Y_2' K_{21} \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)^{-1} Y_1\right)\right) \\ &= \frac{1}{n_2} \text{tr}\left(K_{21} \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)^{-1} E(Y_1 Y_1')\right) \\ &= \frac{1}{n_2} \tau \text{tr}\left(K_{21} \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)^{-1} K_{12}\right). \end{aligned} \quad (3.9)$$

D'autre part,

$$\text{var}(Y_1) = \tau \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)$$

et on obtient

$$\text{var}(\widehat{Y}_2) = \tau K_{21} \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)^{-1} K_{12}.$$

L'expression (3.9) est donc également égale à l'espérance de la variance empirique des composantes de  $\widehat{Y}_2$ , si on la calcule naturellement par

$$\text{var}(Y_{2i}) = \frac{1}{n_2} \widehat{Y}_2' \widehat{Y}_2.$$

On a pour finir

$$E(R_{\text{gen}}^2) = \frac{1}{n_2} \text{tr}\left(K_{21} \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)^{-1} K_{12}\right). \quad (3.10)$$

Cette formule permet le calcul de l'espérance de  $R_{\text{gen}}^2$  et  $R_{\text{tot}}^2 = h^2 R_{\text{gen}}^2$  au seul moyen d'une GRM  $K$  et de l'héritabilité  $h^2$ . Dès que la taille  $n_2$  de l'échantillon de test dépasse quelques centaines d'individus, cette valeur est très stable, le choix de la façon dont on divise l'échantillon en un échantillon d'apprentissage et un échantillon de test n'introduisant que peu de variabilité.

Une approximation très grossière de cette quantité peut être obtenue en remarquant que la matrice  $h^2 K_{11} + (1-h^2) I_{n_1}$  est proche de  $I_{n_1}$ , les termes non diagonaux de  $K_{11}$  étant généralement petits, et donc

$$E(R_{\text{gen}}^2) \simeq n_1 h^2 \eta \quad (3.11)$$

où  $\eta$  est la variance des termes dans  $K_{12}$ , c'est-à-dire la variance des termes hors diagonale de  $K$ . Cette approximation simpliste n'est évidemment pas valable pour de grandes valeurs de  $n_1$ , puisqu'elle n'est pas bornée. Il paraît inévitable de devoir utiliser la totalité d'une GRM pour calculer la valeur de (3.10) pour différentes tailles d'apprentissage  $n_1$ .

Il n'est pas facile d'obtenir des GRM calculées sur de grands échantillons issus d'une population peu ou modérément structurée. La distribution naturelle pour des matrices de co-

variances aléatoires est la distribution de Wishart ; la comparaison avec des données réelles montre cependant que le spectre des GRM n'est pas similaire au spectre des matrices de Wishart, ce qui disqualifie ce modèle. La section suivante propose un modèle qui produit des matrices très similaires à celles qu'on observe.

### 3.4.2. Un modèle stochastique de matrices de corrélation génétique

Nous proposons le modèle suivant pour des GRM aléatoires :

$$K = U\Lambda U'$$

où  $U$  est une matrice orthogonale aléatoire de dimensions  $n \times n$  (tirée selon la loi uniforme sur le groupe des matrices orthogonales, c'est-à-dire l'unique loi invariante par multiplication par toute matrice orthogonale  $\Gamma$ ), et  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  est une matrice diagonale.

Pour le choix du vecteur  $\Lambda$  de valeurs propres, nous nous sommes basés sur les données observées sur 6000 individus issus de l'étude des Trois Cités (en abrégé 3C), présentées au chapitre suivant et dans [20]. Notons tout d'abord que si  $\Lambda$  est fixé, on peut montrer (annexe B) que l'espérance des termes diagonaux de  $K$  est  $E(\Lambda) \stackrel{\text{def}}{=} \frac{1}{n} \sum_i \lambda_i$  et que les termes hors diagonale sont centrés, de variance  $\simeq \frac{1}{n} \text{var}(\Lambda) \stackrel{\text{def}}{=} \frac{1}{n} \sum_i (\lambda_i - E(\Lambda))^2$ .

Les valeurs des  $\lambda_i$  doivent donc être choisies de façon à ce que  $E(\Lambda) = 1$  et  $\text{var}(\Lambda) = n\eta$  où  $\eta$  est la variance observée sur termes hors diagonale des données des 3C, c'est-à-dire  $\eta = 1,63 \cdot 10^{-5}$ .

On constate en analysant les GRM obtenues sur les données 3C que, à l'exception d'un petit nombre de grandes valeurs propres associées aux toutes premières composantes principales, les  $\lambda_i$  sont très proches des quantiles d'une distribution de Guerrero-Johnson SB [113], c'est-à-dire des valeurs

$$\xi + \lambda \expit\left(\frac{1}{\delta}(z_\alpha - \gamma)\right), \quad \text{pour } \alpha = \frac{1}{n+1}, \dots, \frac{n}{n+1} \quad (3.12)$$

où les  $z_\alpha$  sont les quantiles d'une loi normale standard,  $\xi$  et  $\lambda$  sont des paramètres de localisation et d'échelle choisis pour que les conditions sur l'espérance et la variance des  $\lambda_i$  soient vérifiées, et où on a choisi empiriquement les paramètres de forme comme suit :

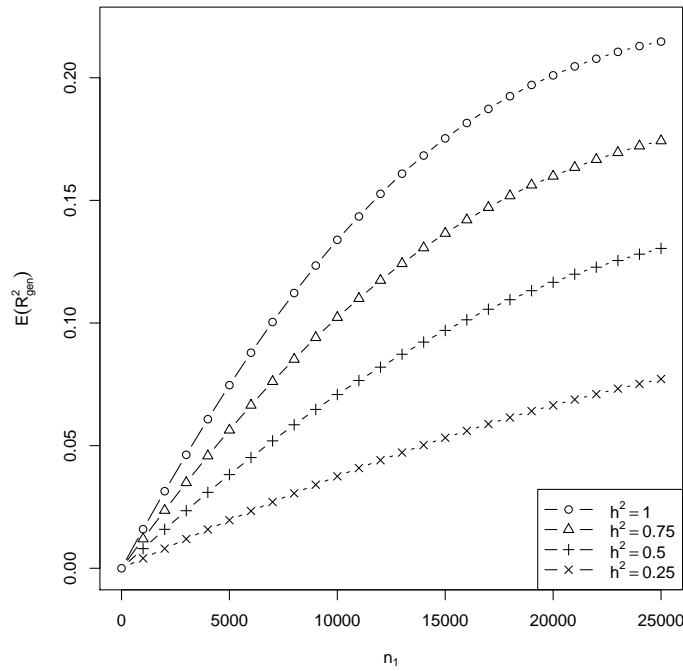
$$\begin{aligned} \gamma &= 0,39045 + n/13580 \\ \frac{1}{\delta} &= 1,1074 + n/175000. \end{aligned} \quad (3.13)$$

### 3.4.3. Application au calcul des performances de prédiction

Nous avons utilisé cette procédure pour générer des GRM de taille  $n = n_1 + n_2$  avec  $n_1$  allant de 500 à  $n_1 = 25\,000$ ,  $n_2 = 1000$ . Les matrices orthogonales  $U$  ont été générées par

l'algorithme décrit dans [114] ; on constate que les valeurs de  $R_{\text{gen}}^2$  (équation 3.10) sont très peu sensibles au choix de  $U$ , c'est-à-dire qu'elles sont très concentrées autour de la valeur moyenne.

Pour les valeurs de  $n_1 \leq 5000$ , on vérifie que le résultat obtenu est le même que celui qu'on obtient en utilisant directement les données 3C ; pour les valeurs supérieures, les résultats présentés figure 3.1 sont le fruit d'une extrapolation éhontée, mais qui nous paraît crédible. Pour  $n_1 \leq 10\,000$ ,  $E(R_{\text{gen}}^2)$  croît de façon linéaire, de façon étonnamment proche de ce qui est donné par l'approximation (3.11).



**FIGURE 3.1.** – Prédiction avec les BLUP sur des données de génome entier : espérance du coefficient d'ajustement en fonction de la taille de l'échantillon d'apprentissage  $n_1$ , pour  $h^2 = 0.25, 0.5, 0.75$  and 1.

On voit que le moins qu'on puisse dire est que les performances prédictives sont très pauvres. Il n'est pas surprenant de constater que plus l'héritabilité est faible, moins le processus d'apprentissage est efficace. Le scénario utilisé est optimiste, puisqu'on a supposé que  $Y$  suivait exactement le modèle linéaire mixte utilisé pour sa prédiction.

Notons qu'en l'absence d'homogamie et d'hétérogamie, en prédisant un trait  $Y$  (avec une héritabilité  $h^2$  positive) par la demi-somme des valeurs observées chez les parents, on obtient  $R_{\text{gen}} = \frac{1}{2}$  indépendamment de  $h^2$  (si  $h^2$  est nul, il n'est pas défini).

Le fait que les BLUP  $\hat{\alpha} = X_a'PY$  ne permettent pas la prédiction des phénotypes est en partie lié au fait qu'ils ne prédisent en fait que la projection de  $\alpha$  sur le sous-espace de dimension  $n$  engendré par les lignes de  $X_a$  (section 2.5), qui est « tout petit » dans  $\mathbb{R}^p$ . Considérons un individu pour lequel on veut prédire la valeur de  $Y$  : s'il y a plusieurs individus

de l'échantillon d'apprentissage avec lesquels sa corrélation génomique est élevée, il n'est pas nécessaire de bien estimer les effets  $\alpha$  pour obtenir une bonne prédiction du phénotype – on peut penser au « cas limite » où il a un ou même plusieurs jumeaux (corrélation  $2\phi = 1$ ) dans l'échantillon d'apprentissage. L'autre cas limite serait celui d'une corrélation nulle avec tous les individus de l'échantillon d'apprentissage : alors, pour bien prédire  $Y$  il faut avoir bien prédit  $\alpha$ . Les faibles niveaux de corrélations observés dans nos échantillons nous rapprochent de ce second cas limite.

Un autre problème potentiel pour la prédiction de  $\alpha$  est la possibilité d'une hétérogénéité des tailles d'effet d'une région du génome à l'autre. Pour y remédier, certains auteurs ont proposé de définir de façon dynamique des régions génomiques avec un paramètre de variance pour chacune d'elle [115]. Une solution plus simple (et probablement tout aussi efficace) serait d'inclure dans le modèle un score construit sur les SNP qui ont des effets très significatifs, et d'utiliser les BLUP pour améliorer les capacités de prédiction de ce score.

Dans le contexte de l'évaluation de la valeur reproductive du bovin laitier, où l'environnement est beaucoup plus homogène que dans les populations d'animaux sauvages ou dans les populations humaines, et où la diversité génétique est moindre, il a été montré que le succès de la sélection assistée par marqueurs est dû à l'utilisation d'individus étroitement apparentés, et que la précision de la prédiction décroît rapidement avec le niveau d'apparentement entre les individus [116, 117]. Ceci implique à nos yeux que la prédiction de phénotypes polygéniques à partir d'échantillons d'individus faiblement apparentés est vouée à rester inefficace.

### 3.5. De l'interprétation du modèle mixte

Nous nous sommes interrogés sur les différences entre le modèle de Fisher et le modèle mixte (ou « modèle de Visscher » ?), passant largement sous silence les problèmes posés par les hypothèses d'additivité, d'absence de termes d'interaction, etc, qui sont bien sûr criticables au premier chef [118, 119]. Cependant même si on décide d'accepter le modèle de Fisher et ses hypothèses, le glissement au modèle mixte n'est pas sans conséquences, et il n'est pas acquis que les deux modèles soient équivalents, les différences étant purement techniques.

Dans les applications classiques des modèles linéaires mixtes, les effets aléatoires sont utilisés par exemple quand des mesures répétées sont faites sur des individus aléatoires (inclusion d'un « effet individu » aléatoire) ; c'est le cas du suivi longitudinal, ou de la production laitière des différentes filles de taureaux dont la valeur reproductive est modélisée par un effet aléatoire. Un autre exemple classique est celui où des mesures sont faites sur plusieurs individus de groupes aléatoires (inclusion d'un « effet groupe »), ou de façon similaire la modélisation des effets expérimentaux dans les dosages biologiques. Dans tous ces cas, la modélisation de ces effets comme des variables aléatoires gaussiennes est naturelle.



En revanche, dans les applications que nous avons présentées ici, la modélisation de l'effet d'un SNP comme aléatoire est peu naturelle, même si on pourrait argumenter en rappelant que chaque SNP est le produit d'une mutation aléatoire. La supposition faite que les écart-types des effets aléatoires sont inversement proportionnels à l'écart-type des génotypes codés par le nombre d'allèle alternatif n'est au mieux qu'un postulat commode – certains auteurs tentés par une interprétation réaliste du modèle mixte ont d'ailleurs envisagé de s'en affranchir. L'interprétation que nous appelons réaliste consiste à considérer que le modèle mixte est une approximation raisonnable de la réalité, et que le terme aléatoire  $X_a\alpha$  est propre à modéliser une composante polygénique. Ne pas adopter l'interprétation réaliste ne veut pas dire qu'on rejette l'usage du modèle, mais plutôt qu'il faut s'interroger sur l'existence d'autres interprétations qui peuvent expliquer son efficacité pratique.

Nous avons mentionné une des pistes qu'il est possible de suivre pour répondre à ce questionnement : le modèle mixte sert à produire une structure de covariance entre individus, au moyen de la GRM, qui est une matrice de corrélation empirique. Chaque coefficient de la GRM est la moyenne de coefficients de corrélation calculés SNP par SNP. Si on considère que, même quand les individus ne sont que très peu apparentés, il existe une corrélation allélique entre eux, identique en espérance en tout point du génome, la GRM est alors un moyen d'estimer cette corrélation. Il y a un moyen de tester cette hypothèse : si elle est vraie, les estimations faites sur deux moitiés du génome devraient être bien corrélées. L'expérience montre qu'en-deçà du seuil  $2\phi \approx 0,025$ , qui sert en pratique à sélectionner des paires d'individus non apparentés, une telle corrélation semble exister, même si elle est très faible ( $r \approx 0,01$ ). Cette piste mérite sans doute d'être approfondie.

Il peut également être intéressant de revenir sur l'héritabilité génomique (section 3.2.2) à la lumière de l'interprétation que nous avons donnée du modèle quand il est utilisé pour corriger une structure de population (section 3.3) : le terme  $X_a\alpha$  se réécrit comme un terme  $PC_1w_1 + \dots + PC_nw_n$ , où toutes les composantes principales de  $X_a$  sont incluses dans le modèle, faisant dans la méthode une variante ou une extension de la régression sur les composantes principales (PCR). L'efficacité de la PCR tient au fait que les premières composantes principales capturent bien l'origine géographique des individus ; celle du modèle mixte pourrait relever du même principe. Cela soulève une question : est-ce que l'ensemble des composantes principales, ou du moins un grand nombre d'entre elles, et non seulement les premières, sont susceptibles de refléter la structure de population dans un modèle mixte ? Nous essaierons d'y répondre au chapitre suivant.



## Chapitre 4.

# Estimation de l'héritabilité dans une population structurée

Ce chapitre reprend une part du travail exposé dans l'article [20] co-écrit avec Claire Dandine. Dans cet article nous avons appliqué la méthode de l'héritabilité génomique aux données que le comité scientifique de l'étude des Trois-Cités a accepté de partager avec nous. Le problème auquel nous avons cherché à répondre est celui du biais induit par l'existence d'une structure de population.

### 4.1. Peut-on corriger le biais induit par la structure de population ?

Nous avons décrit au chapitre précédent la méthode utilisée pour estimer l'héritabilité génomique (restreinte) à l'aide d'un modèle mixte. Ainsi que nous l'avons mentionné, la présence d'une structure de population peut créer un biais dans les estimations. La solution recommandée est de procéder comme pour la correction du biais dans les études d'association sur le génome entier, en incluant dans le modèle les 10 ou 20 premières composantes principales génomiques avec un effet fixe.

Il n'y a cependant aucune base théorique sur la base de laquelle ce nombre de composantes est choisi ; on peut en inclure beaucoup plus sans compromettre la qualité des estimations. Nous avons donc exploré la façon dont les estimations de l'héritabilité varient avec le nombre de PCs incluses dans le modèle, et la capacité de la méthode à corriger le biais inclus par la structure de population.

Pour ce faire, nous utilisons des données simulées ainsi qu'un jeu de 6 000 individus de l'étude des Trois Cités [120] (3C). Nous analysons la latitude et la longitude du lieu de naissance des individus de l'étude 3C comme des traits quantitatifs. Les valeurs de ces coordonnées géographiques ne sont évidemment pas déterminées par des facteurs génétiques, et leur héritabilité réelle est nulle. Cependant, la proximité génétique entre les individus n'est pas indépendante de la proximité géographique entre leurs lieux de naissance ; nous sommes en présence d'une structure de population qui doit avoir pour conséquence une

surestimation de l'héritabilité par le modèle mixte. Comme d'autre part, les premières composantes principales sont bien corrélées avec les coordonnées du lieu de naissance [106], on s'attend à ce que leur inclusion dans le modèle avec un effet fixe corrige efficacement le biais. Cet exemple peut donc aider à se faire une idée du nombre de composantes principales qu'il est nécessaire d'inclure pour parvenir à une correction satisfaisante du biais.

Nous analysons également plusieurs traits anthropométriques mesurés sur les participants de l'étude 3C : la stature, qui est l'exemple historique étudié par Galton, et le poids, l'IMC, le tour de tête et le rapport taille/hanches pour lesquels on attend une héritabilité positive. Les études d'association avec le génome entier ont découvert de nombreux SNP associés avec la stature [121–127], ou avec des traits associés à l'obésité comme le poids, l'IMC, le ratio taille/hanches [128–137]. La table 4.1 rassemble quelques estimations de la littérature récente. Pour tous ces traits, nous calculons l'héritabilité génomique et examinons la façon dont elle évolue avec l'inclusion de composantes principales avec effets fixes dans le modèle.

	Données familiales		Études de jumeaux		Données génomiques	
Stature	0.92	[138]	0.68 à 0.94	[139–142]	0.44 à 0.62	[77, 78, 97, 142–145]
Poids	-		0.37	[142]	0.19	[143]
					0.26	[142]
IMC	0.24 à 0.81	[146]	0.47 à 0.90	[146]	0.16 à 0.27	[77, 142, 143, 145]
			0.28 [142]			
			0.45 à 0.84 [147]			
Rapport taille/hanches	-		-		0.16 (H), 0.18 (F)	[145]†
Tour de tête	0.66	[148]	0.75	[147]	-	

**TABLE 4.1.** – Estimations de l'héritabilité dans la littérature

† Valeur ajustée sur l'IMC et stratifiée sur le sexe. La référence [146] est une méta-analyse de 88 études de jumeaux et de 27 études familiales.

#### 4.1.1. L'étude des Trois Cités

L'étude des Trois Cités est une étude de cohorte avec suivi longitudinal, incluant 9294 personnes de 65 ans ou plus, recrutées entre 1999 et 2012 à Bordeaux, Dijon et Montpellier. Les participants ont été génotypés au Centre National de Génotypage avec des puces Illumina Human610-Quad, comme décrit dans [149].

Le contrôle qualité de ces données a consisté à retirer des données :

- les individus dupliqués et les individus dont le sexe génomique ne concorde pas avec le sexe rapporté ;

#### 4.1. Peut-on corriger le biais induit par la structure de population ?

- les individus avec plus de 5% de génotypes manquants ;
- les individus avec un taux d'hétérozygotie à plus de 3 écart-types de la moyenne observée ;
- les individus qui ne sont pas nés en France ;
- les individus d'ascendance non-européenne (identifiés par une analyse en composantes principales incluant les européens de HapMap) ;
- les SNP avec un taux de génotypage inférieur à 99% ou un écart aux proportions de Hardy-Weinberg avec  $p < 10^{-8}$ .

Pour éliminer les apparentements cryptiques nous avons éliminé un individu de chaque paire dont l'apparement génomique (mesuré par les entrées  $k_{ij}$  de la GRM K) est supérieur à 0,025. Le jeu de données final contient 5 793 individus et 509 931 SNP autosomaux.

##### 4.1.2. Modèle et notations

Le modèle utilisé est

$$Y = X\beta + \sum_{i=1}^p PC_i \gamma_i + X_a \alpha + \varepsilon \quad (4.1)$$

où

- $Y \in \mathbb{R}^n$  est la variable quantitative analysée
- $X \in \mathbb{R}^{n \times q}$  est une matrice de covariables à inclure dans le modèle avec effets fixes (l'âge, le sexe) ;
- $PC_1, \dots, PC_p \in \mathbb{R}^n$  sont les  $p$  premières composantes principales (on fera varier  $p$  de 0 à 2000, pour  $n \simeq 6000$ ) ;
- $X_a$  est comme à la section 3.2.2 la matrice des génotypes centrés réduits ;
- $\beta \in \mathbb{R}^q$  et  $\gamma_1, \dots, \gamma_p$  sont des effets fixes ;
- $\alpha \sim \mathcal{N}\left(0, \frac{\tau}{r} I_r\right)$  et  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$  sont des effets aléatoires.

Les composantes principales  $PC_1, \dots, PC_p$ , destinées à corriger le biais induit par la structure de population, peuvent être calculées à partir de la matrice  $X_a$  entière, ou à partir d'une sous-matrice de  $X_a$ , choisie de façon à n'avoir qu'un faible déséquilibre de liaison entre les SNP conservés [112]. Dans la suite nous présenterons les résultats obtenus avec le premier choix ; ils sont similaires à ceux qu'on obtient en utilisant les « données élaguées », qui sont présentés dans l'article cité [20].

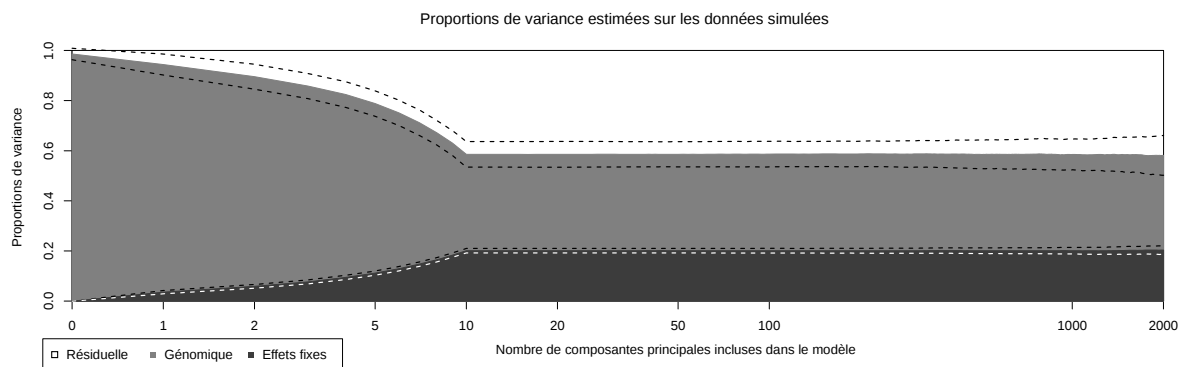
Les paramètres  $\tau$  et  $\sigma^2$  sont estimés par l'algorithme AIREML, et l'héritabilité par  $h^2 = \frac{\tau}{\tau + \sigma^2}$ , ce qui revient à négliger la contribution du terme  $X\beta$  à la variance de  $Y$ . En outre, nous décomposons la variance de  $Y$  en trois parties, comme décrit à la section 2.6 : la variance expliquée par les covariables présentes dans  $X$ , la variance génétique  $\tau$ , et la variance résiduelle  $\sigma^2$ .

## 4.2. Analyse d'une variable simulée

Nous avons simulé une variable quantitative suivant le modèle (4.1), en utilisant les données des 3C pour la matrice  $X_a$ , qui est la matrice de dimension  $5793 \times 509\,931$  des génotypes centrés-réduits. On a inclus les  $p = 10$  premières composantes principales, avec  $\gamma_1, \dots, \gamma_{10}$  choisis de façon à ce que chacune d'elle explique 2% de la variance. Le reste de la variance est réparti à égalité entre la composante génomique et la composante résiduelle : on a  $\sigma^2 = \tau$ ,  $h^2 = 0.5$ , et la proportion de variance totale expliquée par chacune de ces deux composantes est de 40%.

Cette variable simulée a été analysée avec  $p$  variant de 0 à 2000. L'expérience a été répétée 100 fois afin d'estimer l'espérance et l'écart-type des estimateurs. Les résultats sont représentés figure 4.1. L'axe des abscisses correspond au nombre  $p$  de composantes principales incluses dans le modèle. Sur l'axe des ordonnées, on peut lire la proportion de variance expliquée par chacune des composantes : en gris foncé, la proportion de variance expliquée par les  $p$  composantes principales, en gris clair, la proportion de variance expliquée par la composante génomique, et en blanc la proportion de variance résiduelle. Les lignes pointillées correspondent à l'espérance  $\pm 1$  écart-type.

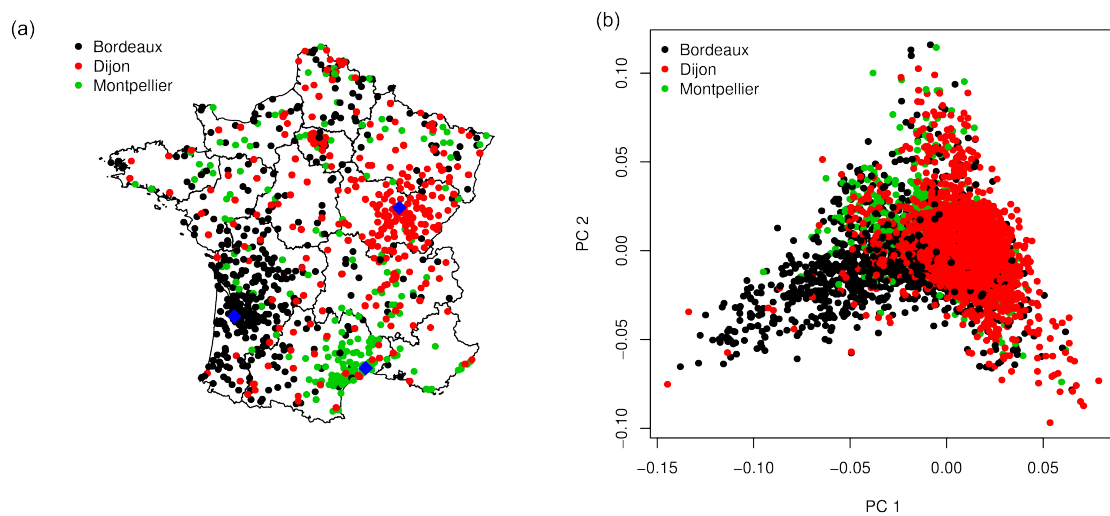
Quand  $p < 10$ , comme on s'y attend la proportion de variance génomique est sur-estimée. En revanche, dès qu'on a inclus  $p \geq 10$  composantes principales pour l'estimation des paramètres, les vraies proportions (20% pour les effets fixes, 40% pour chacun des deux termes aléatoires) sont bien estimées, avec un écart type proche de 0,05, qui n'augmente que lentement avec  $p$ , la moyenne des estimations restant stable.



**FIGURE 4.1.** – Proportions de variances estimées sur les données simulées, en fonction du nombre  $p$  de composantes principales incluses pour l'analyse avec le modèle mixte.

### 4.3. Analyse des variables de l'étude 3C

Nous avons analysé plusieurs variables quantitatives provenant de l'étude des Trois Cités. Tout d'abord, nous avons considéré la latitude et la longitude du lieu de naissance, obtenues à partir du code postal de ce dernier. Bien qu'il n'y ait que trois centres de recrutement, les lieux de naissance sont bien répartis (figure 4.2). Nous avons également analysé plusieurs phénotypes anthropométriques : la stature, le poids, l'IMC, rapport taille/hanches. La moyenne et l'écart-type des variables sont récapitulés dans la table 4.2, pour les hommes, les femmes, et tous les individus ensemble. Le sexe et l'âge ont été inclus comme covariables pour l'estimation de l'héritabilité des phénotypes anthropométriques.



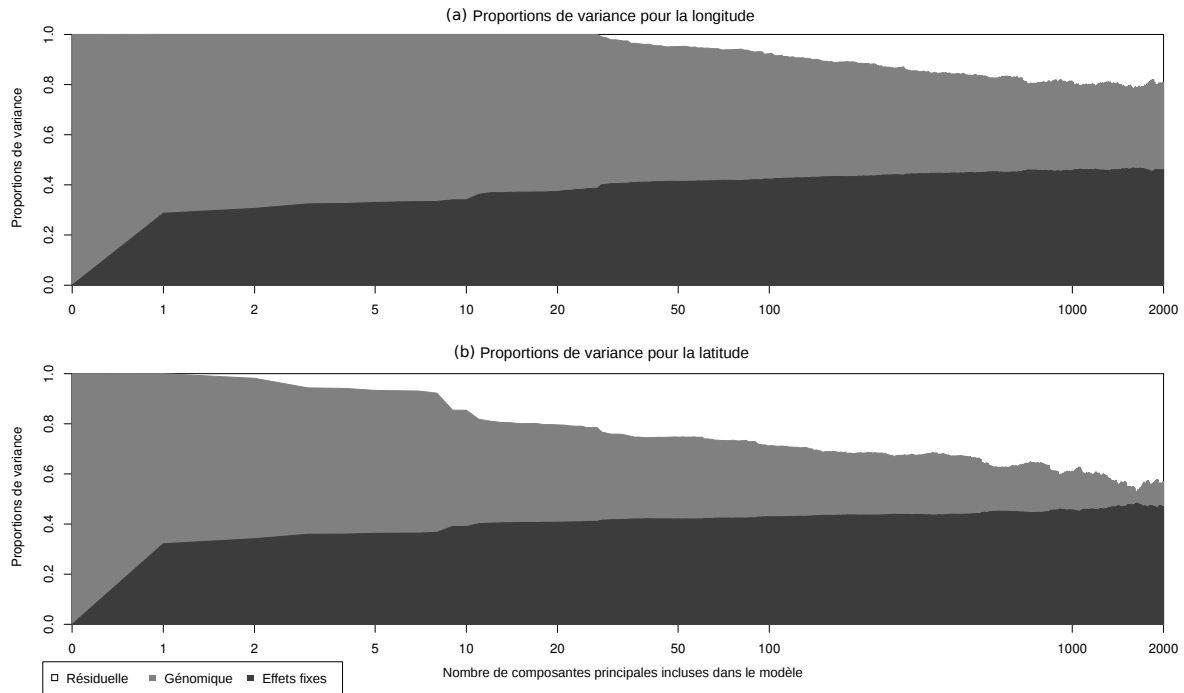
**FIGURE 4.2.** – (a) Distribution des lieux de naissances des individus de l'étude 3C, et (b) les deux premières composantes principales. La couleur des points correspond au centre de recrutement : Bordeaux (1499 individus), Dijon (3676 individus) et Montpellier (618 individus). Les centres sont représentés par des carrés bleus.

Variable	Hommes (N = 2298)			Femmes (N = 3495)			Tous (N = 5793)		
	Moyenne	É.-t.	n	Moyenne	É.-t.	n	Moyenne	É.-t.	n
Age	74,15	5,56	2298	74,39	5,49	3495	74,30	5,52	5793
Latitude	46,78	1,73	2090	46,76	1,63	3171	46,77	1,67	5261
Longitude	3,32	2,40	2090	3,35	2,45	3171	3,34	2,43	5261
Stature	169,58	6,35	2290	156,60	6,17	3461	161,77	8,91	5751
Poids	75,58	11,27	2292	62,58	11,32	3485	67,74	12,97	5777
IMC	26,27	3,53	2288	25,52	4,36	3457	25,82	4,06	5745
Tour de tête	57,75	2,05	2243	55,37	2,07	3414	56,32	2,37	5657
Rapport taille/hanches	0,95	0,07	2117	0,84	0,07	3180	0,88	0,09	5297

**TABLE 4.2.** – Description des variables quantitative de l'étude 3C

### 4.3.1. Longitude et latitude

La figure 4.3 montre la décomposition de la variance de la longitude et de la latitude, sur le même principe que nous avons utilisé pour la variable simulée. Quand on n'inclut pas de composante principale dans le modèle, on estime dans les deux cas que 100% de la variance est génétique.



**FIGURE 4.3.** – Proportions de variances estimées pour les coordonnées géographiques, en fonction du nombre  $p$  de composantes principales incluses pour l'analyse avec le modèle mixte.

L'ajout progressif de composantes principales ne fait diminuer que lentement l'héritabilité estimée de la longitude ; il faut en ajouter plus de 20 pour obtenir une estimation plus petite que 100%. L'héritabilité estimée de la latitude décroît un peu plus rapidement. Quand on inclut  $p = 500$  composantes principales, l'héritabilité estimée de la longitude est de 70%, celle de la latitude est de 39% ; pour  $p = 2000$ , ces valeurs sont descendues à 63% et 19% respectivement. On peut trouver des chiffres précis dans les tables 4.3 et 4.4, ainsi que les tests du rapport de vraisemblance\* pour l'hypothèse  $h^2 = 0$ , qui sont très significatifs quand  $p$  est petit, et restent significatifs dans les deux cas pour  $p = 1000$ . Ces tables donnent aussi les écart-types des différents estimateurs ; ils augmentent avec  $p$ , mais de façon très lente – en fait il faut dépasser  $p = 1000$  pour que l'augmentation soit sensible.

\*Le test du rapport de vraisemblance suit asymptotiquement une loi de mélange  $\frac{1}{2}\chi^2(0) : \frac{1}{2}\chi^2(1)$  [150] ; le seuil de significativité à 5% est de 2,70.



		degré de			
	LRT	signification	$\hat{\tau}$ (é.t.)	$\hat{\sigma}^2$ (é.t.)	$\hat{h}^2$ (é.t.)
0 PC	1892.63	<1e-40	4.34 (0.085)	2.9e-4 (0.089)	1.000
1 PC	822.44	<1e-40	3.90 (0.076)	2.6e-4 (0.072)	1.000
2 PCs	708.77	<1e-40	3.86 (0.075)	2.6e-4 (0.070)	1.000
3 PCs	600.64	<1e-40	3.82 (0.075)	2.5e-4 (0.069)	1.000
4 PCs	598.75	<1e-40	3.82 (0.074)	2.5e-4 (0.069)	1.000
5 PCs	570.04	<1e-40	3.80 (0.074)	2.5e-4 (0.068)	1.000
10 PCs	524.18	<1e-40	3.77 (0.074)	2.5e-4 (0.067)	1.000
20 PCs	377.77	<1e-40	3.65 (0.071)	2.4e-4 (0.063)	1.000
50 PCs	207.02	<1e-40	3.19 (0.068)	0.28 (0.060)	0.919 (0.061)
100 PCs	168.31	8.7e-39	2.96 (0.068)	0.46 (0.060)	0.867 (0.064)
500 PCs	69.56	3.7e-17	2.29 (0.069)	0.98 (0.061)	0.701 (0.079)
1000 PCs	38.23	3.2e-10	2.07 (0.073)	1.14 (0.063)	0.645 (0.095)
2000 PCs	15.93	3.3e-5	2.02 (0.087)	1.18 (0.068)	0.632 (0.136)

**TABLE 4.3.** – Estimation des paramètres du modèle pour la longitude et leur écart-type, test du rapport de vraisemblance (LRT) et  $p$ -valeurs associées

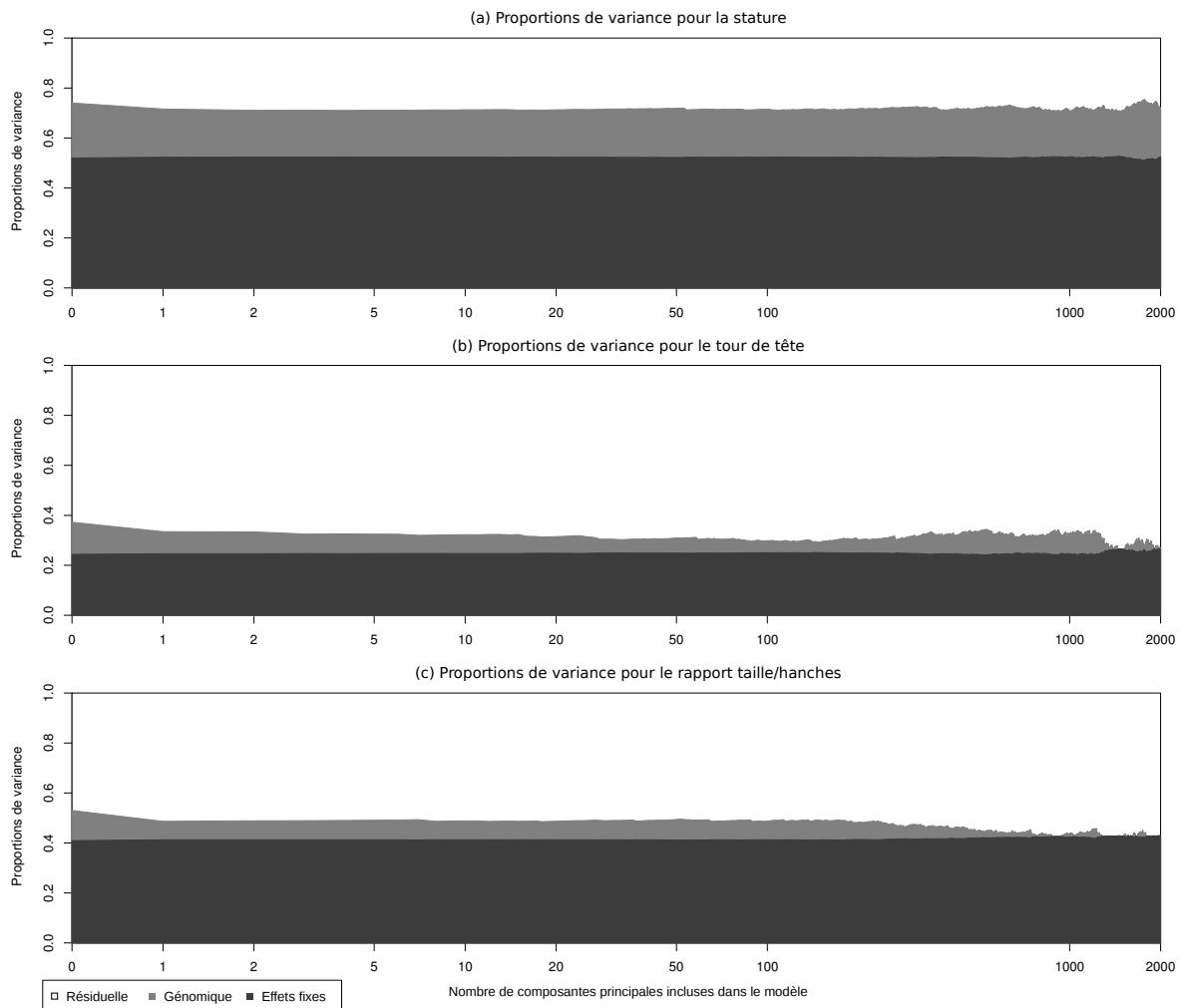
		degré de			
	LRT	signification	$\hat{\tau}$ (é.t.)	$\hat{\sigma}^2$ (é.t.)	$\hat{h}^2$ (é.t.)
0 PC	1854.15	<1e-40	2.05 (0.040)	1.4e-4 (0.039)	1.000
1 PC	558.69	<1e-40	1.81 (0.035)	1.2e-4 (0.031)	1.000
2 PCs	424.64	<1e-40	1.74 (0.035)	0.05 (0.030)	0.972 (0.050)
3 PCs	312.82	<1e-40	1.61 (0.035)	0.15 (0.030)	0.912 (0.053)
4 PCs	307.45	<1e-40	1.60 (0.035)	0.16 (0.030)	0.908 (0.053)
5 PCs	292.00	<1e-40	1.57 (0.034)	0.18 (0.030)	0.896 (0.054)
10 PCs	177.38	<1e-40	1.28 (0.033)	0.40 (0.030)	0.761 (0.058)
20 PCs	114.47	5.1e-27	1.08 (0.033)	0.57 (0.030)	0.656 (0.062)
50 PCs	74.84	2.5e-18	0.91 (0.032)	0.70 (0.030)	0.563 (0.065)
100 PCs	52.25	2.4e-13	0.79 (0.032)	0.80 (0.030)	0.496 (0.068)
500 PCs	21.08	2.2e-6	0.60 (0.033)	0.95 (0.031)	0.387 (0.082)
1000 PCs	7.10	3.9e-3	0.43 (0.036)	1.08 (0.032)	0.286 (0.104)
2000 PCs	1.18	0.139	0.28 (0.044)	1.20 (0.035)	0.189 (0.167)

**TABLE 4.4.** – Estimation des paramètres du modèle pour la latitude et leur écart-type, test du rapport de vraisemblance (LRT) et  $p$ -valeurs associées.

### 4.3.2. Variables anthropométriques

Nous avons analysé les variables anthropométriques décrites plus haut en incluant le sexe et l'âge comme covariables. La figure 4.4 montre la décomposition de la variance pour trois de ces variables : la stature, le tour de tête et le rapport taille/hanches. Le sexe et l'âge expliquent 52%, 24% et 41% de ces trois variables ; quand  $p = 0$  l'héritabilité est estimée à 46%, 17% et 20% de la variance restante. Après inclusion d'une composante principale dans le modèle ( $p = 1$ ), ces valeurs tombent à 40%, 12% et 13%, ce qui montre que la première estimation était biaisée vers le haut. Ces valeurs évoluent peu quand on inclut davantage de composantes principales (table 4.5).

Nous ne présentons pas ici de figure pour le poids et l'IMC, pour lesquels les estimations sont moins sensibles à la valeur de  $p$  (table 4.5). Les covariables expliquent 26% et 1,4% de la variance, respectivement ; les héritabilités sont estimées à 22% et 19%.



**FIGURE 4.4.** – Proportions de variances estimées pour (a) la stature, (b) le tour de tête et (c) le rapport taille/hanches

### 4.3. Analyse des variables de l'étude 3C

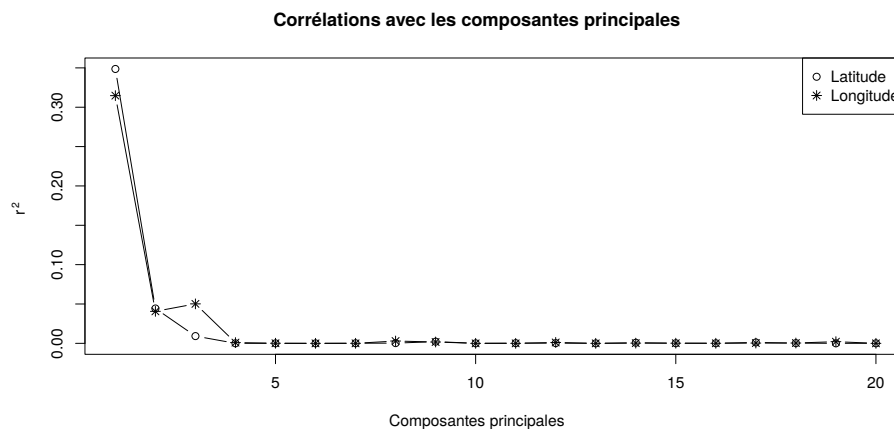
Variable		LRT	degré de signification	$\hat{\tau}$ (é.t.)	$\hat{\sigma}^2$ (é.t.)	$\hat{h}^2$ (é.t.)
Stature	0 PC	68.80	5.6e-17	17.48 (0.719)	20.64 (0.689)	0.459 (0.061)
	1 PC	41.96	4.7e-11	15.32 (0.714)	22.57 (0.691)	0.404 (0.064)
	5 PCs	37.06	5.7e-10	14.90 (0.714)	22.95 (0.691)	0.394 (0.066)
	10 PCs	37.58	4.4e-10	15.05 (0.714)	22.82 (0.691)	0.397 (0.066)
	20 PCs	37.02	5.8e-10	15.05 (0.716)	22.82 (0.691)	0.398 (0.066)
Tour de Tête	0 PC	9.85	8.5e-4	0.722 (0.078)	3.52 (0.079)	0.170 (0.057)
	1 PC	3.70	0.027	0.495 (0.078)	3.73 (0.079)	0.117 (0.062)
	5 PCs	2.67	0.051	0.437 (0.078)	3.79 (0.079)	0.103 (0.064)
	10 PCs	2.46	0.058	0.424 (0.078)	3.80 (0.079)	0.100 (0.065)
	20 PCs	1.88	0.085	0.376 (0.078)	3.84 (0.079)	0.089 (0.065)
Rapport talle/hanches	0 PC	13.09	1.5e-4	9.2e-04 (8.6e-5)	3.6e-3 (8.7e-5)	0.205 (0.061)
	1 PC	3.54	0.030	5.6e-04 (8.6e-5)	3.9e-3 (8.7e-5)	0.126 (0.068)
	5 PCs	3.98	0.023	6.0e-04 (8.6e-5)	3.9e-3 (8.7e-5)	0.133 (0.068)
	10 PCs	3.65	0.028	5.6e-04 (8.6e-5)	3.9e-3 (8.7e-5)	0.125 (0.066)
	20 PCs	3.41	0.032	5.6e-04 (8.6e-5)	3.9e-3 (8.7e-5)	0.126 (0.070)
Poids	0 PC	13.92	9.6e-5	28.24 (2.32)	97.23 (2.32)	0.225 (0.062)
	1 PC	13.80	1.0e-4	28.26 (2.32)	97.21 (2.32)	0.225 (0.062)
	5 PCs	13.07	1.5e-4	27.90 (2.32)	97.53 (2.32)	0.222 (0.062)
	10 PCs	13.45	1.2e-4	28.38 (2.33)	97.11 (2.32)	0.226 (0.063)
	20 PCs	12.41	2.1e-4	27.55 (2.33)	97.82 (2.32)	0.220 (0.063)
IMC	0 PC	10.44	6.2e-4	3.23 (0.302)	13.09 (0.302)	0.198 (0.063)
	1 PC	9.34	1.1e-3	3.12 (0.302)	13.18 (0.302)	0.191 (0.064)
	5 PCs	9.65	9.4e-4	3.18 (0.302)	13.12 (0.303)	0.195 (0.064)
	10 PCs	10.21	7.0e-4	3.29 (0.303)	13.04 (0.303)	0.201 (0.064)
	20 PCs	8.26	2.0e-3	3.00 (0.303)	13.28 (0.303)	0.184 (0.065)

TABLE 4.5. – Estimation des paramètres du modèles pour les mesures anthropométriques et leur écart-type, test du rapport de vraisemblances (LRT) et  $p$ -valeurs associées.

## 4.4. Moralité

L'analyse des données simulées sous le modèle mixte montre qu'il est possible d'inclure un grand nombre de composantes principales dans le modèle sans compromettre la précision de l'estimation de l'héritabilité – bien sûr cela dépend de la taille de l'échantillon, mais dès qu'elle dépasse quelques milliers d'individus, on peut prendre  $p = 100$  ou 500 composantes principales sans inconvénient.

Il est surprenant que l'estimation « naïve », sans correction, produise une héritabilité de 100% pour la latitude et la longitude : si on s'attend bien à estimer une héritabilité positive, les coordonnées du lieu de naissance étant corrélées aux premières composantes principales (figure 4.5), une valeur aussi élevée était inattendue. On pouvait également s'attendre à ce que l'héritabilité estimée se rapproche de 0 dès que quelques composantes principales sont incluses dans le modèle ; on a vu qu'il n'en est rien. Non seulement les héritabilités estimées restent positives, mais les tests du rapport de vraisemblance sont très significatifs.



**FIGURE 4.5.** – Corrélation ( $r^2$ ) des coordonnées géographiques avec les 20 premières composantes principales.

Browning et Browning [99] ont obtenu des résultats similaires avec des données cas/témoins simulées à partir des génotypes du Welcome Trust Case Control Consortium. Leurs simulations font l'hypothèse d'un recrutement biaisé : 90% des individus originaires d'Écosse et du Pays de Galles sont assignés au groupe des cas, et seulement 10% des individus originaires d'Angleterre ; le groupe des témoins est formé avec les individus restant. Dans leur réponse, Goddard *et coll.* ont défendu la méthode de l'héritabilité génomique en arguant du caractère extrême de ce scénario. Le même argument pourrait resservir à propos de notre étude ; il est cependant réaliste d'imaginer qu'un phénotype quantitatif soit sous l'influence d'un facteur environnemental qui varie avec la latitude ou la longitude. Il ressort de nos analyses que l'estimation de l'héritabilité d'un tel trait serait sur-estimée, sans que l'inclusion des premières composantes principales dans l'analyse permette de corriger le biais.

Les estimations de l'héritabilité obtenues pour les traits anthropométriques sont globalement compatibles avec les résultats de la littérature. Trois de ces traits (la taille, le rapport taille/hanches, et le tour de tête) semblent affectés par la présence d'une structure de population. Ils sont par ailleurs significativement associés aux coordonnées géographiques ; dans [20], nous avons également estimé leur héritabilité en intégrant au modèle la latitude et la longitude du lieu de naissance, comme covariable d'ajustement ; cela fait disparaître la sensibilité des estimations à la valeur de  $p$ , ce qui conforte l'idée de la présence d'une structure de population.

De plus, dans cette analyse complémentaire, l'héritabilité du rapport taille/hanches tombe à 8% (au lieu de 13% avec l'analyse présentée plus haut), et le test du rapport de vraisemblance n'est plus significatif. Cela laisse penser que pour ce trait, la correction par l'inclusion des composantes principales n'est pas suffisante pour corriger le biais dû à la structure de population ; mais les tenants du modèle polygénique pourront rétorquer qu'après tout, rien ne prouve qu'il n'y a pas de polymorphismes génétiques impliqués dans la valeur de cette variable, et dont la fréquence varie avec la géographie. De façon générale, en présence de corrélation gène-environnement, démêler ce qui relève de l'un ou de l'autre est une gageure.

Pour finir, revenons à la question que nous avons posée à la fin du chapitre précédent : est-ce que seules les premières composantes principales génomiques sont affectées par la structure de population, ou est-ce que celle-ci se répercute sur un grand nombre de celles-là ? De façon étonnante, c'est la seconde alternative qui est vraie. Les composantes principales étant calculées sur  $\mathbf{X}_a$  tout entière, le modèle (4.1) peut être écrit

$$Y = X\beta + PC_1\gamma_1 + \dots + PC_p\gamma_p + PC_{p+1}w_{p+1} + \dots + PC_nw_n + \varepsilon$$

où les effets aléatoires  $w_j$  sont tirés dans une loi  $\mathcal{N}(0, \tau)$  ; la valeur de  $\tau$  est donc estimée sur les composantes  $p + 1$  à  $n$ . Ce constat ne suffit évidemment pas à reléguer les estimations de l'héritabilité génomique au rang d'artefacts méthodologiques, mais cela devrait inciter à la prudence quand il s'agit d'interpréter les valeurs produites par cette méthode.



# Chapitre 5.

## Projets

Dans ce dernier chapitre je vais décrire rapidement plusieurs projets de recherche et de collaborations en me concentrant sur divers aspects de l'héritabilité génomique, qui est ma principale préoccupation.

Je passerai donc à peu près sous silence plusieurs projets en cours sur les tests d'association entre un phénotype et un ensemble de plusieurs SNP, malgré l'intérêt que je leur porte. Très brièvement,

- avec Jacqueline Milet, nous projetons d'évaluer les performances d'une variante du test SKAT [151,152] pour tester l'association de régions génomiques avec le paludisme simple. SKAT repose sur une matrice de similarité génomique entre individus, dans le calcul de laquelle les SNP sont pondérés en fonction de leur fréquence, les SNP les plus rares étant supposés avoir un effet potentiellement plus grand. Nous proposons d'utiliser une pondération qui dépend d'une statistique de test de sélection naturelle, les SNP les plus soumis à sélection ayant un effet potentiellement plus grand.
- avec Ozvan Bocher (étudiante en M2), Gaëlle Marenne et Emmanuelle Génin (Brest), nous avons deux projets de tests d'association avec les variants rares. Dans le premier projet, déjà très avancé après le stage de M1 d'Ozvan, nous proposons un test d'association avec une variable catégorielle à plus de deux niveaux (par exemple, deux groupes de patients et un groupe de témoins). Dans le second projet, qui devrait débiter en 2018, nous nous proposons de tester le comportement de variantes de SKAT, où la matrice de similarité génomique sera calculée d'une façon entièrement différente de celle utilisée dans le test original.

Revenons maintenant à l'héritabilité génomique et à ses inévitables compagnons, la matrice de corrélation génétique et le modèle mixte.

## 5.1. L'héritabilité génomique

### 5.1.1. Modèle oligogénique ou polygénique ?

Cette section décrit une collaboration en cours avec Anthony Herzig, Anne-Louise Leutenegger, Teresa Natile et Marina Ciullo (Paris et Naples).

Ainsi que nous l'avons mentionné, pour justifier le modèle de Fisher pour les variables quantitatives, il suffit que le nombre  $r$  de SNP intervenant dans le modèle soit assez grand pour que la somme des effets soit approximativement gaussienne ; ça sera le cas par exemple pour  $r \geq 20$ , pourvu que les fréquences alléliques ne soient pas trop faibles. Convenons de parler alors de « modèle oligogénique ». Le calcul de l'héritabilité génomique suppose au contraire implicitement qu'on est sous un « modèle polygénique », supposant que tous les SNP génotypés ont un effet ; en pratique, il semble que dès que  $r$  vaut au moins 10 000, et si les SNP causaux sont bien répartis sur le génome, ce modèle donne des résultats satisfaisants.

Considérons l'estimation de l'héritabilité par un modèle mixte

$$\text{var}(Y) = \tau_a K + \tau_d D + \sigma^2 I_n \quad (5.1)$$

où le terme général de  $K$  est  $2\phi_{ij}$ , le double du coefficient de parenté entre les individus  $i$  et  $j$ , et celui de  $D$  est  $\psi_{ij}$ , la probabilité que les individus  $i$  et  $j$  soient IBD = 2 (sections 1.3.2 et 3.2.2).

On peut s'appuyer sur deux types d'échantillons pour estimer l'héritabilité, et selon le cas sur des valeurs différentes pour  $K$  et  $D$  :

1. Pour un échantillon d'individus non apparentés, les matrices  $K$  et  $D$  seront calculées à partir de données de génome entier (section 3.2.2) ; c'est l'héritabilité génomique ;
2. Pour des individus apparentés (en particulier, issus d'une population isolée) il reste possible de calculer l'héritabilité génomique ; on peut également utiliser pour  $K$  et  $D$  les valeurs issues d'une analyse de leurs relations de parenté.

Nous formulons les hypothèses suivantes :

1. Dans le premier cas (individus non apparentés), si la variable analysée suit un modèle polygénique, on obtiendra des estimations convergentes ; mais si elle suit un modèle oligogénique, les matrices  $K$  et  $D$  calculées sur un très grand nombre de SNP seront des estimations très bruitées des « vraies » matrices de covariance du trait (celles qui seraient calculées uniquement sur les SNP causaux), et on estimera probablement une héritabilité nulle ou très inférieure à la vraie valeur.
2. Dans le deuxième cas (individus apparentés), les deux méthodes doivent produire des matrices  $K$  et  $D$  similaires, et donc des estimations proches de l'héritabilité. Ces estimations doivent être convergentes, sous le modèle polygénique comme sous le modèle oligogénique.



Dans les faits, les estimations de l'héritabilité génomique au sens étroit ont souvent donné des valeurs compatibles (étant données les grandes variations observées d'une population à l'autre) avec les valeurs obtenues sur des données familiales. Nous avons pourtant été frappés par les différences entre les résultats obtenus par Zhu et coll [105] pour la composante de dominance pour certains traits, comme la stature ou le cholestérol LDL (table 5.1), et les résultats obtenus en population isolée. Une des explications possibles à ces discordances est que la composante de dominance soit oligogénique.

	Non apparentés		Population isolée					
	Zhu et coll, 2015 [105]		Traglia et coll, 2009 [153]		Pilia et coll, 2006 [154]		Abney et coll, 2001 [155]	
	$h^2$	$H^2 - h^2$	$h^2$	$H^2 - h^2$	$h^2$	$H^2 - h^2$	$h^2$	$H^2 - h^2$
Stature	0,48	0,02	0,78	0,22	0,77	0,23	-	-
LDL	0,21	0,01	0,23	0,77	0,38	0,29	0,36	0,60

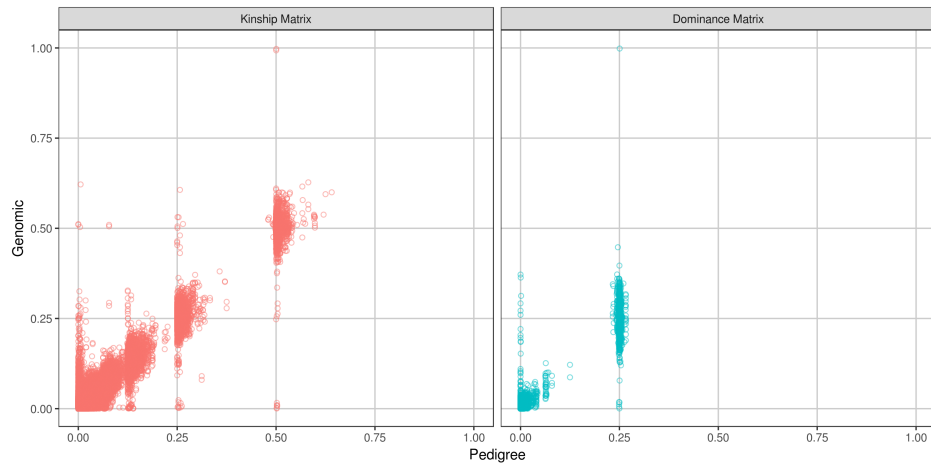
**TABLE 5.1.** – Estimation de l'héritabilité pour la stature et le cholestérol LDL.

La composante de dominance est significative dans les populations isolées étudiées.

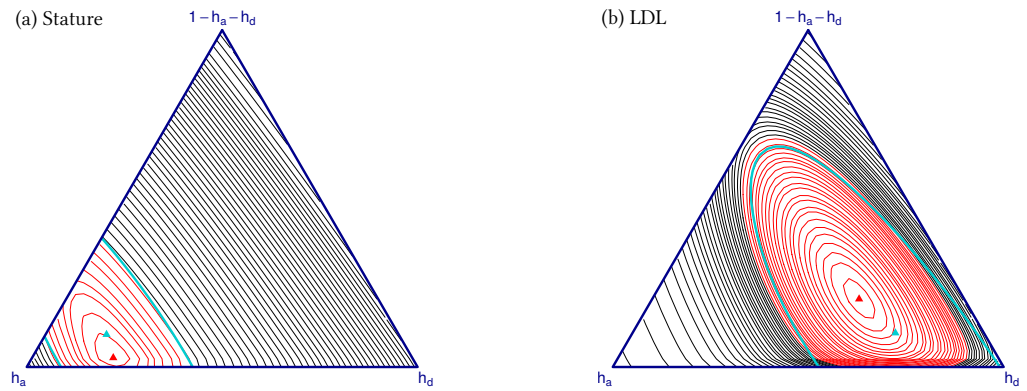
Afin de tester les hypothèses formulées plus haut et d'explorer le cas particulier de ces phénotypes, Anthony a commencé à travailler sur le cas des populations isolées. Il a tout d'abord estimé l'héritabilité de ces deux traits (et de quelques autres) en utilisant les phénotypes et les génotypes (174 000 SNP) de 1350 individus issus de la population isolée des villages de Campora, Cardile et Gioi, dans le Cilento (Italie) [156]. Pour cet échantillon, les matrices K et D estimées par les deux méthodes sont cohérentes malgré des différences relativement importantes, comme l'illustre la figure 5.1.

Les deux panneaux de la figure 5.2 montrent les courbes de niveau d'une vraisemblance profilée (analogue à celle qui est présentée en annexe A.3.2) qui s'exprime en fonction des deux paramètres  $h_a^2 = h^2$  et  $h_d^2 = H^2 - h^2$ , calculée pour les matrices K et D génomiques. La région définie par  $h_a^2 \geq 0$ ,  $h_d^2 \geq 0$ ,  $h_a^2 + h_d^2 \leq 1$ , est représentée en coordonnées barycentriques dans un triangle équilatéral ; le sommet inférieur gauche correspond à  $h_a^2 = 1$ , le sommet inférieur droit à  $h_d^2 = 1$  et le sommet supérieur à  $h_a^2 + h_d^2 = 0$ . Les courbes en rouge correspondent à une région de confiance à 95%. Le point et l'ellipse turquoise correspondent au maximum de vraisemblance et à la région de confiance obtenus avec les matrices K et D calculées à partir des relations de parenté. Il apparaît que les deux méthodes produisent des résultats compatibles pour les deux variables quantitatives analysées ici ; on constate également que l'estimation de la composante de dominance  $h_d^2$  est beaucoup moins précise que celle de  $h_a^2$ , ce qui est probablement dû au fait que les paires d'individus informatives sont moins nombreuses.

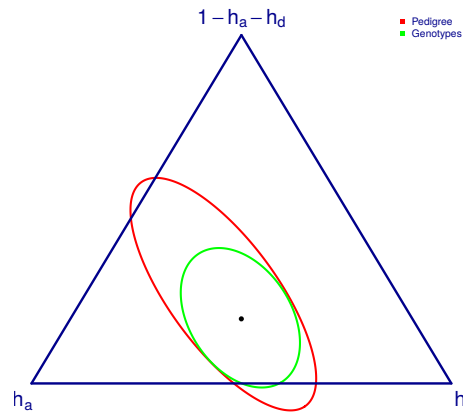
Pour comprendre les propriétés de la méthode dans le cas du modèle polygénique, Anthony a simulé une centaine de jeux de données avec des niveaux d'apparentement similaires à ceux des données réelles : il a attribué à chacun des fondateurs de la généalogie du Cilento deux haplotypes pris dans les haplotypes de 1000 Génomes pour la population de Toscane, puis il a généré les génotypes de l'ensemble de la population par « gene dropping ». Un phénotype est ensuite généré sous le modèle polygénique avec  $h_a^2 = h_d^2 = 0,4$ ,



**FIGURE 5.1.** – Comparaison des coefficients non diagonaux des matrices K et D obtenues par les deux méthodes (figure A. Herzig).



**FIGURE 5.2.** – Régions de confiance pour l'héritabilité de (a) la stature, et (b) le cholestérol LDL (figure A. Herzig)



**FIGURE 5.3.** – Ellipse représentant les régions de variation des estimations de l'héritabilité (figure A. Herzig)

les tailles d'effet des SNP étant tirées dans une loi normale. Les résultats obtenus après 100 simulations sont représentés figure 5.3 : le point noir est la valeur simulée pour  $h_a^2$  et  $h_d^2$ , l'ellipse rouge est l'ellipse de variation des estimations basées sur les relations de parenté et l'ellipse verte est celle des estimations basées sur les génotypes. C'est donc cette dernière méthode qui est la plus précise pour le modèle polygénique.

L'étape suivante de ce travail est la réalisation de simulations analogues sous le modèle oligogénique – on peut s'attendre à voir disparaître l'avantage de l'héritabilité génomique sur la méthode basée sur les relations de parenté. Il faudra ensuite simuler échantillons d'individus non apparentés, sous les deux modèles, pour comparer les estimations de l'héritabilité génomique dans ces deux situations et vérifier nos hypothèses.

### 5.1.2. Biais des estimations génomiques de l'héritabilité

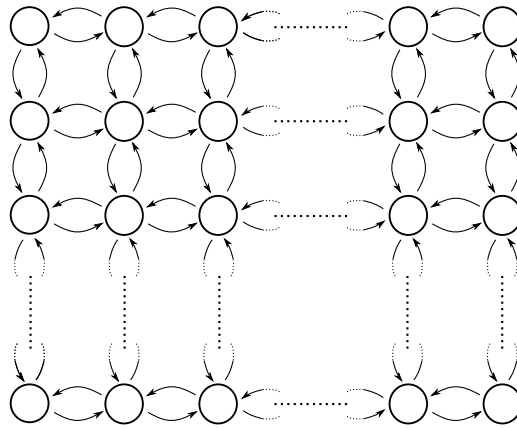
Pour diverses raisons, il me paraît important de continuer à explorer les causes de biais possibles dans l'estimation de l'héritabilité basée sur le modèle mixte.

#### La structure de population

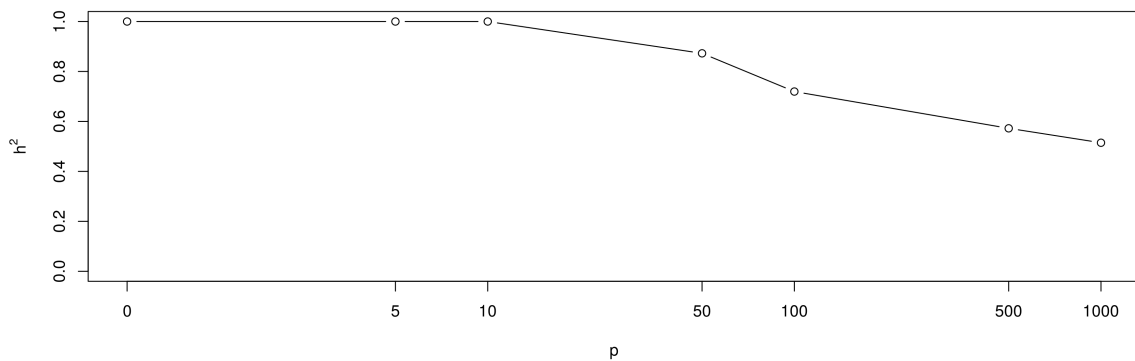
Le travail de Claire Dandine sur les données de l'étude des Trois Cités a mis en évidence de façon empirique le fait que la structure de population peut être la cause d'un biais impossible à résorber par les méthodes usuelles. Pour mieux comprendre ce phénomène, il serait intéressant de procéder à des simulations de populations, dans un modèle de type « pas japonais » ou *stepping stone*, où des sous-populations ou dèmes sont disposés sur une grille, chacun échangeant du matériel génétique avec les dèmes voisins (figure 5.4). Même si au temps  $t = 0$  la population est homogène, au fil des générations la dérive génétique va créer des différences d'une extrémité à l'autre de la grille. Les corrélations génétiques entre individus seront d'autant plus élevées qu'ils sont issus de dèmes proches, ce qui créera l'illusion que les coordonnées des individus sur la grille sont hérissables.

Il est facile de simuler l'évolution des fréquences alléliques dans chacun des dèmes, pour quelques (dizaines de) milliers de SNP indépendants. On peut ensuite générer les génotypes d'un échantillon d'individus tirés au hasard dans la grille, et calculer la matrice des corrélations génétiques entre ces individus. Cette matrice permet finalement d'estimer l'héritabilité des coordonnées du dème dont sont issus les individus, avec, comme au chapitre 4,  $p$  composantes principales génomiques en effet fixe dans le modèle. Le comportement du modèle dépend de plusieurs paramètres : la taille de la grille, la taille de chaque dème, les taux de migration, le nombre de SNP indépendants, et le nombre de générations simulées après un départ avec des fréquences alléliques homogènes. La figure 5.5 montre que ce modèle simple semble parvenir à reproduire ce que nous avons observé pour la latitude et la longitude sur les données des Trois Cités.

Il est nécessaire de réaliser des simulations plus complexes pour produire des données génomiques réalistes, notamment avec du déséquilibre de liaison. Une approche théorique



**FIGURE 5.4.** – Un modèle de pas japonais sur une grille finie.



**FIGURE 5.5.** – Héritabilité de l'indice de la ligne du dème, en fonction du nombre  $p$  de composantes principales en effet fixe. Calcul effectuée à partir 25 000 SNP pour 4 000 individus pris au hasard dans une grille de  $50 \times 50$  dèmes contenant 100 individus chacun. Chaque dème reçoit 10% d'allèles des dèmes voisins à chaque génération ; les simulations se sont étendues sur 100 générations.

du comportement de ce modèle pourrait être également tentée. Il serait pertinent d'ajouter à l'effet de la dérive génétique une pression de sélection naturelle sur de multiples locus, pression dont l'intensité dépendrait des coordonnées de la grille. Ceci devrait créer une structure de population plus importante que la dérive seule.

## L'homogamie

L'homogamie est une autre cause de structure de population. Dans une population finie, l'homogamie pour un trait quantitatif, c'est-à-dire l'existence pour ce trait d'une corrélation positive entre les parents d'un enfant, peut s'interpréter comme l'existence d'un grand nombre de strates disposées linéairement selon la valeur du trait dans la strate. Chaque strate reçoit à chaque génération des allèles en provenance des autres strates avec une pro-

tabilité d'autant plus forte qu'elles sont proches. Dans cette situation, comme dans le modèle du pas japonais, la dérive génétique doit créer des différences de fréquences alléliques qui auront pour résultat une inflation des estimateurs de l'héritabilité. Il serait naturel d'introduire dans ce modèle une transmission familiale de l'environnement.

Pour explorer cette cause potentiellement importante de biais dans les populations humaines, il faudra concevoir un modèle de simulation en s'appuyant sur la littérature, en particulier sur des modèles d'environnement familial partagé.

## 5.2. La matrice de corrélation génétique

### 5.2.1. Modèle aléatoire pour les matrices de corrélation génétique

Le modèle proposé à la section 3.4.2 pour les matrices de corrélation génétique, qui consiste à prendre  $K = U\Lambda U'$  avec  $U$  une matrice orthogonale tirée dans la loi uniforme et  $\Lambda$  une matrice diagonale fixée, mérite d'être approfondi. Nous avons en effet utilisé une procédure empirique, calibrée sur les données des Trois Cités, pour choisir les valeurs présentes dans  $\Lambda$  (équations 3.12 et 3.13). Il serait intéressant d'avoir plus d'arguments, tant théoriques qu'empiriques, pour comprendre comment se comportent ces valeurs.

Pour les aspects théoriques, il faudrait en particulier prendre en compte le théorème d'entrelacement de Cauchy, qui énonce que les valeurs propres d'une sous-matrice de  $K$  s'insèrent entre les valeurs propres de  $K$ . Une sous-matrice  $K'$  de  $K$  étant supposée suivre une loi analogue (à la dimension près), ses valeurs propres et celles de  $K$  vérifient cette propriété d'entrelacement. En ajoutant à cette propriété les contraintes  $E(\Lambda) = 1$  et  $\text{var}(\Lambda) = n\eta$  (section 3.4.2), on doit pouvoir en déduire des propriétés intéressantes nécessairement vérifiées par  $\Lambda$ , en particulier le comportement asymptotique des grandes valeurs propres.

Pour les aspects empiriques, il faudrait utiliser d'autres matrices de corrélation génétique, issues d'autres populations et d'autres modes d'échantillonnage ; il faudrait en particulier considérer le cas où on a laissé des individus avec des apparentements plus proches dans l'échantillon. En effet, une des étapes du contrôle qualité sur les données des Trois Cités a consisté à supprimer un individu de chaque paire pour laquelle la valeur de  $k_{ij}$  excédait 0,025.

Ce travail est lié à celui envisagé à la section 5.1.2, qui donne des modèles pour des populations structurées. Il est également lié à celui envisagé à la section qui suit, la variance des valeurs présentes dans la matrice de corrélation génétique étant un des paramètres à prendre en compte dans le choix des  $\lambda_i$ .

### 5.2.2. Prise en compte du déséquilibre de liaison

Si  $\mathbf{X}_a$  est la matrice  $n \times r$  des génotypes centrés réduits, on peut analyser le calcul de  $\mathbf{K} = \frac{1}{r} \mathbf{X}_a \mathbf{X}_a'$  comme celui de la moyenne de  $r$  matrices  $\mathbf{K}_1, \dots, \mathbf{K}_r$  définies par

$$\mathbf{K}_j = \mathbf{X}_{aj} \mathbf{X}_{aj}'$$

où  $\mathbf{X}_{aj}$  est la  $j$ -ème colonne de  $\mathbf{X}_a$ , c'est-à-dire le vecteur des génotypes centrés réduits au SNP  $j$ . Si on considère que les coefficients de  $\mathbf{K}$  estiment une corrélation génétique qu'on peut appeler ou non « apparentement », les coefficients de chacun des  $\mathbf{K}_j$  sont des estimateurs de cette variable aléatoire. La moyenne empirique de ces estimateurs est un estimateur convergent.

Cependant en présence de déséquilibre de liaison entre les SNP  $j$  et  $j'$ , c'est-à-dire de corrélation entre les génotypes portés par un individu en ces SNP, les estimateurs présents dans  $\mathbf{K}_j$  et  $\mathbf{K}_{j'}$  sont corrélés. Il est alors naturel d'estimer la matrice de corrélation génétique par une moyenne pondérée

$$\tilde{\mathbf{K}} = \sum_{j=1}^r w_j \mathbf{K}_j$$

où les  $w_j$  sont choisis de façon à minimiser la variance de l'estimateur obtenu, sous la contrainte  $\sum_j w_j = 1$  nécessaire à obtenir un estimateur non biaisé.

Nous avons montré que, quand les apparentements et les taux de consanguinité sont faibles, la covariance entre les coefficients dans  $\mathbf{K}_j$  et  $\mathbf{K}_{j'}$  est  $r_{jj'}^2$ , le carré de la corrélation  $r_{jj'}$  entre les SNP  $j$  et  $j'$ . Le vecteur des poids optimaux est donc

$$\mathbf{w} = \left( \mathbf{1}_n' \mathbf{A}^{-1} \mathbf{1}_n \right)^{-1} \mathbf{A}^{-1} \mathbf{1}_n$$

où la matrice  $\mathbf{A}$  est la matrice des valeurs  $r_{jj'}^2$  pour  $j, j' = 1, \dots, r$ .

Il n'est pas question de calculer la matrice  $\mathbf{A}$  en entier, ni son inverse. En supposant que si  $|j - j'|$  est assez grand,  $r_{jj'} \simeq 0$ , plusieurs pistes sont envisageables pour le calcul de  $\mathbf{A}^{-1} \mathbf{1}_n$ .

D'autres auteurs ont considéré des sommes pondérées pour prendre en compte le déséquilibre de liaison dans le calcul de  $\mathbf{K}$  [97, 157], mais aucun n'a pris comme critère la minimisation de la variance.

## 5.3. Modèle mixte

### 5.3.1. Précision des estimateurs

Revenons sur l'astuce de la diagonalisation (annexe A.3) pour le modèle lineaire mixte dans le cas d'une variable  $Y$  centrée. En utilisant la SVD de  $Z = U\Sigma L'$ , on réécrit le modèle

$$Y = Zu + \varepsilon$$

avec  $u \sim \mathcal{N}(0, \tau I_q)$ , sous la forme

$$Y = U\Sigma w + \varepsilon$$

avec  $w \sim \mathcal{N}(0, \tau I_n)$ .

Le vecteur  $U'Y$  est le vecteur des coefficients  $(\hat{\beta}_1, \dots, \hat{\beta}_n)$  de la régression linéaire à effets fixes de  $Y$  sur les colonnes de  $U$ , qui sont les composantes principales normées de  $Z$ . Ce vecteur suit une loi  $\mathcal{N}(0, \tau\Sigma^2 + \sigma^2 I_n)$ . On a  $\text{var}(\hat{\beta}_i) = \sigma^2 + \tau\lambda_i$  où les  $\lambda_i$  sont les coefficients de la matrice diagonale  $\Lambda = \Sigma^2$ , qui contient les valeurs propres de  $K = ZZ'$ . On a donc  $E(\hat{\beta}_i^2) = \sigma^2 + \tau\lambda_i$ , et on peut estimer  $\tau$  et  $\sigma^2$  par une simple régression linéaire des  $\hat{\beta}_i^2$  sur les  $\lambda_i$  : l'intercept correspond à  $\sigma^2$  et le coefficient de régression à  $\tau$ .

Cette procédure grossière se comporte étonnamment bien, donnant des résultats presque identiques à ceux obtenus par maximum de vraisemblance : une rapide étude de simulation utilisant une matrice  $K$  simulée selon la procédure décrite en section 3.4.2 montre que les deux méthodes sont non biaisées, ont la même variance et que la corrélation entre leurs résultats est  $r = 0,98$ . Cela n'a que peu d'intérêt pratique pour l'estimation des paramètres, car la maximisation de la vraisemblance en utilisant l'astuce de la diagonalisation et la méthode de Newton (annexe A.3) est déjà extrêmement rapide.

Cependant, si nous tenons la similitude de comportement de ces deux méthodes pour acquise, l'analyse de la seconde méthode permet d'obtenir une estimation de la variance des estimateurs à partir des valeurs  $\lambda_i$ , et en particulier la façon dont cette variance varie avec la taille  $n$  de l'échantillon.

En utilisant le modèle aléatoire décrit à la section 3.4.2 pour les matrices de corrélations génétiques, à partir de  $E(\Lambda) = 1$  et  $\text{var}(\Lambda) = n\eta$  où  $\eta$  est la variance des termes non diagonaux de  $K$  (de l'ordre de  $10^{-5}$  sur les données des Trois Cités), cette heuristique montre que tant que  $n \ll \eta^{-1}$ , la variance des estimateurs est en  $n^{-2}$ . Des simulations confirment ce comportement surprenant (dans la mesure où on prend vite l'habitude que les variances soient en  $n^{-1}$ ), donnant une décroissance en  $n^{-1,9}$  pour  $n \leq 6000$ .

Tout ceci demande à être affiné, en particulier par des recherches dans la littérature scientifique qui traite du modèle mixte et de l'héritabilité\* et par un travail théorique (sans doute

---

\*De façon étonnante la question ne semble pas avoir été posée dans la littérature de génétique statistique. Les références [97] et [158] évaluent la précision des estimateurs en fonction du nombre de SNP inclus dans le modèle et de la méthode utilisée pour le calcul de la matrice  $K$ , mais se placent à taille d'échantillon fixée.

peut-on travailler directement sur la vraisemblance pour montrer le résultat annoncé). Cette question est à envisager en parallèle avec le travail projeté à la section 5.2.1 sur les matrices de corrélation génétique.

### 5.3.2. Le modèle logistique mixte

Estimer les paramètres dans un modèle logistique mixte (ou plus généralement dans un modèle linéaire généralisé mixte) est plus difficile que dans le cas linéaire, la vraisemblance faisant intervenir une intégrale qui ne peut pas être calculée de façon littérale. La solution classique est l'utilisation de la *Penalized Quasi-Likelihood (PQL)*, qu'on peut interpréter comme une approximation de Laplace pour le calcul de la vraisemblance. Malheureusement, la PQL présente des biais importants dans le cas du modèle logistique mixte [159, 160]. La bibliothèque logicielle *lme4* [75] propose l'utilisation d'une approximation de bien meilleure qualité par une quadrature gaussienne adaptative ; malheureusement ainsi que nous l'avons déjà dit cette bibliothèque est optimisée pour le cas de matrices de covariables creuses, et est très peu performante pour les données génomiques.

La procédure esquissée à la section 5.3.1 pour le modèle linéaire suggère cependant une approche heuristique pour le cas du modèle logistique mixte

$$\text{logit}(P(Y = 1)) = \alpha + Zu$$

où  $u \sim \mathcal{N}(0, \tau I_q)$ . La première étape consisterait à réaliser la régression logistique sur chacune des composantes principales normées de  $Z$  une à une, et la seconde étape utiliserait les coefficients de régression obtenus à la première étape pour estimer  $\tau$ .

Les arguments avancés en 5.3.1 pour le cas linéaire ne sont bien sûr plus valables dans le cas du modèle logistique. En particulier, les coefficients de régression sur les composantes principales normées ne sont pas indépendants les uns des autres dans le cas de la régression logistique. Il est nécessaire d'approfondir cette idée à la fois par une approche théorique, par des simulations, et si le comportement de la méthode le justifie, par une application à des données réelles.



### 5.3.3. Réduction de dimension

Le projet décrit dans cette section a été suscité par une question de Hugues Aschard, et fera l'objet d'une collaboration avec l'équipe Génétique Statistique qu'il dirige à l'Institut Pasteur.

Considérons à nouveau le modèle mixte

$$Y \sim \mathcal{N}(X\beta, \tau K + \sigma^2 I_n).$$

On s'intéresse ici au cas où  $n$  est très grand. Soit  $C \in \mathbb{R}^{n' \times n}$  une matrice semi-orthogonale, c'est-à-dire une matrice qui vérifie  $CC' = I_{n'}$  (on a nécessairement  $n' < n$ ). Alors

$$CY \sim \mathcal{N}(CX\beta, \tau CKC' + \sigma^2 I_{n'}).$$

Ce nouveau modèle mixte est une projection du premier sur un sous-espace de dimension  $n' < n$ . Il a les mêmes paramètres  $\beta$ ,  $\tau$  et  $\sigma^2$ , et peut être utilisé pour une estimation plus rapide de ceux-ci.

Si  $n' = n/k$ , la complexité des calculs étant en  $O(n^3)$ , l'estimation des coefficients sur le modèle projeté est  $k^3$  fois plus rapide que sur le modèle d'origine. Ceci permet d'envisager de réaliser plusieurs réductions de dimension de cette forme avec des matrices  $C$  différentes, et de combiner les estimateurs en en prenant la moyenne.

Plusieurs options viennent à l'esprit pour le choix des matrices  $C$  :

- On peut prendre des matrices de 0 et de 1 avec un 1 exactement sur chaque colonne, et au plus un 1 sur chaque ligne, ce qui est une façon alambiquée de dire qu'on prend un sous-échantillon de taille  $n'$  ;
- on peut choisir des matrices aléatoires uniformément dans l'espace des matrices semi-orthogonales ;
- on peut choisir  $C$  en prenant en compte la structure de  $K$ , de façon à minimiser la perte de précision de l'estimateur.

L'avantage de la première option est la rapidité du calcul de  $CY$  et  $CKC'$ . La seconde option paraît séduisante au premier abord, car de telles projections aléatoires ont de bonnes propriétés pour d'autres applications [161] – mais il n'est pas évident que ça soit le cas dans notre problème. De plus le surcoût (par rapport à l'option précédente) pour générer  $C$  et calculer  $CKC'$  n'est pas négligeable. Plusieurs pistes peuvent être imaginées pour la mise en œuvre de la troisième option ; un travail théorique et des tests seront nécessaires.

Une des questions prioritaires ici est d'analyser le comportement attendu de la méthode en terme de précision des estimateurs, pour chacun des deux choix ci-dessus. Cette analyse ne peut se faire que si on connaît la variance des estimateurs en fonction de la taille de l'échantillon, question posée à la section 5.3.1.

## 5.4. Le dernier avatar du modèle mixte : la régression sur le LD Score

Cette section présente une collaboration débutante avec Aude Saint-Pierre (Brest).

Bulik-Sullivan et coll ont proposé une méthode appelée *LD-score regression* [162]. Cette méthode se place sous le modèle mixte 3.2, que nous rappelons brièvement :

$$Y = \mathbf{1}_n\mu + \mathbf{X}_a\alpha + \varepsilon$$

avec  $\mathbf{X}_a$  la matrice  $n \times r$  des genotypes centrés réduits (ici  $r$  est le nombre total de SNP du génome, génotypés ou non),  $\alpha \sim \mathcal{N}\left(0, \frac{\tau}{r}\mathbf{I}_r\right)$  et  $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ .

On considère une statistique  $\chi_j^2$  de test d'association entre le trait considéré et le SNP d'indice  $j$  ; elle suit, sous l'hypothèse d'absence d'effet de tous les SNP du génome ( $\tau = 0$ ), une loi  $\chi^2(1)$ . Les auteurs montrent que dans le cas général

$$E(\chi_j^2) = 1 + \frac{n}{r}h^2\ell_j,$$

où  $n$  est la taille de l'échantillon et  $\ell_j$  est un score de LD, calculé par  $\ell_j = \sum_{j'} r_{jj'}^2$ , où  $r_{jj'}$  est le coefficient de corrélation entre les SNP  $j$  et  $j'$ . Il ressort de ce calcul qu'une partie de l'inflation des statistiques de test constatée dans les études d'association avec le génome entier pourrait être due non à la structure de population, mais au caractère polygénique des phénotypes étudiés. Les auteurs proposent donc de calculer la régression des valeurs des statistiques de test sur les  $\ell_j$ ,  $E(\chi_j^2) = \hat{\mu} + \hat{\beta}\ell_j$ . Un intercept  $\hat{\mu}$  plus grand que 1 est interprété comme dû à la structure de population ; la pente  $\hat{\beta}$  peut théoriquement permettre d'estimer l'héritabilité.

Comme les auteurs le notent, cette méthode repose fortement sur l'hypothèse  $\text{var}(\alpha_j) = \frac{\tau}{r}$ , qui implique que la variance de l'effet allélique est proportionnelle à  $\frac{1}{p_jq_j}$  où  $p_j$  et  $q_j$  sont les fréquences alléliques. Toujours d'après [162], si elle n'est pas vérifiée, cette méthode sous-estimera l'effet de la structure de population et sur-estimera l'héritabilité. J'avais noté à propos de l'héritabilité génomique que la seule raison d'être de cette relation entre la variance de l'effet allélique et les fréquences alléliques était sa commodité mathématique. La méthode du LD Score semble pourtant prendre ce postulat au pied de la lettre.

Nous projetons, Aude Saint-Pierre et moi-même, d'analyser en détail les propriétés de cette méthode, et de l'appliquer aux variables quantitatives des données de l'étude des Trois Cités – y compris la latitude et la longitude – et de comparer son comportement à celui du modèle mixte et de la régression sur les composantes principales (modèle 3.6).

## 5.5. Conclusion

J'espère avoir convaincu mes lecteurs et lectrices de l'intérêt de questionner l'usage du modèle mixte pour l'estimation de l'héritabilité et de s'interroger sur ses propriétés. Les questions posées dans ce chapitre sont souvent connectées les unes aux autres, et j'espère les mener à bien en compagnie de futurs doctorants et de collaborateurs d'horizons variés. Comme aimait à le dire mon directeur de thèse, Henri Lombardi : « l'avenir est radieux ! ».



# Bibliographie

- [1] Perdry, H. *Aspects constructifs de la théorie des corps valués (précédée d'un chapitre sur la noethérianité constructive)*. Thèse de doctorat, Université de Franche-Comté, 2001.
- [2] Lombardi, H, Perdry, H. *The Buchberger algorithm as a tool for ideal theory of polynomial rings in constructive mathematics*, pages 393–407. London Mathematical Society Lecture Note Series. Cambridge University Press, 1998.
- [3] Kuhlmann, FV, Perdry, H. *Dynamic computations inside the algebraic closure of a valued field*. Valuation theory and its applications, volume 2 : pages 133–156, 2003.
- [4] Perdry, H. *A generalization of Hensel's Lemma*. Valuation theory and its applications, volume 2 : pages 241–249, 2003.
- [5] Perdry, H. *Strongly Noetherian rings and constructive ideal theory*. Journal of Symbolic Computation, volume 37(4) : pages 511–535, 2004.
- [6] Perdry, H. *Henselian valued fields : a constructive point of view*. Mathematical Logic Quarterly, volume 51(4) : pages 400–416, 2005.
- [7] Perdry, H. *Lazy bases : a minimalist constructive theory of Noetherian rings*. Mathematical Logic Quarterly, volume 54(1) : pages 70–82, 2008.
- [8] Alonso, ME, Lombardi, H, Perdry, H. *Elementary constructive theory of Henselian local rings*. Mathematical Logic Quarterly, volume 54(3) : pages 253–271, 2008.
- [9] Perdry, H, Babron, MC, Clerget-Darpoux, F. *The ordered transmission disequilibrium test : detection of modifier genes*. Genetic epidemiology, volume 33(1) : pages 1–5, 2009.
- [10] Perdry, H, Müller-Myhsok, B, Clerget-Darpoux, F. *Using affected sib-pairs to uncover rare disease variants*. Human heredity, volume 74(3-4) : pages 129–141, 2012.
- [11] Bonaïti, B, Bonadona, V, Perdry, H, et coll. *Estimating penetrance from multiple case families with predisposing mutations : extension of the 'genotype-restricted likelihood'(GRL) method*. European Journal of Human Genetics, volume 19(2) : page 173, 2011.
- [12] Bonaïti, B, Alarcon, F, Andrieu, N, et coll. *A new scoring system in cancer genetics : application to criteria for BRCA1 and BRCA2 mutation screening*. Journal of medical genetics, volume 51(2) : pages 114–121, 2014.
- [13] Génin, E, Sahbatou, M, Gazal, S, et coll. *Could inbred cases identified in GWAS data succeed in detecting rare recessive variants where affected sib-pairs have failed ?* Human heredity, volume 74(3-4) : pages 142–152, 2012.

- [14] Gazal, S, Sahbatou, M, Perdry, H, et coll. *Inbreeding coefficient estimation with dense SNP data : comparison of strategies and application to HapMap III*. Human heredity, volume 77(1-4) : pages 49–62, 2014.
- [15] Dandine-Roulland, C, Perdry, H. *Where is the causal variant ? on the advantage of the family design over the case–control design in genetic association studies*. European Journal of Human Genetics, volume 23(10) : page 1357, 2015.
- [16] Babron, MC, Perdry, H, Handel, AE, et coll. *Determination of the real effect of genes identified in GWAS : the example of IL2RA in Multiple Sclerosis*. European Journal of Human Genetics, volume 20(3) : page 321, 2012.
- [17] Mbogning, C, Perdry, H, Toussile, W, Broët, P. *A novel tree-based procedure for deciphering the genomic spectrum of clinical disease entities*. Journal of Clinical Bioinformatics, volume 4(1) : page 6, 2014.
- [18] Mbogning, C, Perdry, H, Broët, P. *A bagged, partially linear, tree-based regression procedure for prediction and variable selection*. Human heredity, volume 79(3-4) : pages 182–193, 2015.
- [19] Dandine-Roulland, C, Perdry, H. *The use of the linear mixed model in human genetics*. Human heredity, volume 80(4) : pages 196–206, 2015.
- [20] Dandine-Roulland, C, Bellenguez, C, Debette, S, et coll. *Accuracy of heritability estimations in presence of hidden population stratification*. Scientific reports, volume 6, 2016.
- [21] Perdry, H. *Librairie logicielle ElstonStewart*, 2012.  
<https://cran.r-project.org/web/packages/ElstonStewart>
- [22] Perdry, H, Dandine-Roulland, C. *Librairie logicielle gaston*, 2015.  
<https://cran.r-project.org/web/packages/gaston>
- [23] Dandine-Roulland, C. *Modélisation de la composante génétique des maladies humaines : Données familiales et Modèles Mixtes*. Thèse de doctorat, Université Paris-Saclay, 2016.
- [24] Gazal, S, Sahbatou, M, Babron, MC, et coll. *FSuite : exploiting inbreeding in dense SNP chip and exome data*. Bioinformatics, volume 30(13) : pages 1940–1941, 2014.
- [25] Leutenegger, AL, Prum, B, Génin, E, et coll. *Estimation of the inbreeding coefficient through use of genomic data*. Am J Hum Genet, volume 73(3) : pages 516–523, 2003.
- [26] Rostand, J. *Esquisse d’une histoire de la biologie*. Gallimard, Paris, 1945.
- [27] Edelson, E. *Gregor Mendel and the roots of genetics*. Oxford University Press, New York and Oxford, 2001.
- [28] Mendel, G. *Versuche über Pflanzen-Hybriden*. Actes Soc Hist Nat Brünn, volume 3 : pages 3–47, 1865.
- [29] Bateson, W. *Mendel’s principle of heredity, A defence*. Cambridge University Press, Cambridge, 1902.
- [30] Reid, JB, Ross, JJ. *Mendel’s genes : Toward a full molecular characterization*. Genetics, volume 189(1) : pages 3–10, 2011.

- [31] Fisher, RA. *Has Mendel's work been rediscovered?* Annals of science, volume 1(2) : pages 115–137, 1936.
- [32] Pires, AM, Branco, JA. *A statistical model to explain the Mendel–Fisher controversy.* Statistical Science, pages 545–565, 2010.
- [33] Pearson, K. *The life, letters and labours of Francis Galton.* Cambridge University Press, Cambridge, 1914.
- [34] Galton, F. *Memories of My Life.* Methuen, London, 1908.
- [35] Galton, F. *Experiments in pangenesis, by breeding from rabbits of a pure variety, into whose circulation blood taken from other varieties had previously been largely transfused.* Proceedings of the Royal Society, volume 19 : pages 393–410, 1871.
- [36] Galton, F. *On blood-relationship.* Proceedings of the Royal Society, volume 20 : pages 394–402, 1872.
- [37] Galton, F. *On blood-relationship.* Nature, volume 6 : pages 173–176, 1872.
- [38] Galton, F. *A theory of heredity.* Contemporary Review, volume 27 : pages 80–95, 1875.
- [39] Bulmer, M. *The development of Francis Galton's ideas on the mechanism of heredity.* Journal of the History of Biology, volume 32(2) : pages 263–292, 1999.
- [40] Weismann, A. *Die Continuität des Keimplasma's als Grundlage einer Theorie der Vererbung.* Gustav Fischer, Jena, 1885.
- [41] Weismann, A. *Das Keimplasma. Eine Theorie der Vererbung.* Gustav Fischer, Jena, 1892.
- [42] Muller, HJ. *Some genetic aspects of sex.* The American Naturalist, volume 66(703) : pages 118–138, 1932.
- [43] Galton, F. *Natural Inheritance.* Macmillan and Co., London and New York, 1889.
- [44] Galton, F, Dickson, JDH. *Family likeness in stature.* Proceedings of the Royal Society of London, volume 40(242–245) : pages 42–73, 1886.
- [45] Galton, F. *Regression towards mediocrity in hereditary stature.* The Journal of the Anthropological Institute of Great Britain and Ireland, volume 15 : pages 246–263, 1886.
- [46] Galton, F. *Family likeness in eye-colour.* Proceedings of the Royal Society of London, volume 40(242–245) : pages 402–416, 1886.
- [47] Bravais, A. *Analyse mathématique sur les probabilités des erreurs de situations d'un point.* Imprimerie royale, Paris, 1844.
- [48] Pearson, K. *Mathematical contributions to the theory of evolution. III. regression, heredity, and panmixia.* Philosophical Transactions of the Royal Society of London Series A, volume 187 : pages 253–318, 1896.
- [49] Galton, F. *The average contribution of each several ancestor to the total heritage of the offspring.* Proceedings of the Royal Society of London, volume 61(369–377) : pages 401–413, 1897.

- [50] Meston, AJ. *The Galton Law of Heredity and how breeders may apply it*. Published by the author, Pittsfield, Mass., 1898.
- [51] Galton, F. *A diagram of heredity*. Nature, volume 57(1474) : page 293, 1898.
- [52] Pearson, K. *Mathematical contributions to the theory of evolution. On the law of ancestral heredity*. Proceedings of the Royal Society of London, volume 62(379-387) : pages 386–412, 1898.
- [53] Pearson, K. *Mathematical contributions to the theory of evolution. On the law of reversion*. Proceedings of the Royal Society of London, volume 66(424-433) : pages 140–164, 1900.
- [54] Pearson, K, Lee, A. *On the laws of inheritance in man : I. Inheritance of physical characters*. Biometrika, volume 2(4) : pages 357–462, 1903.
- [55] Galton, F. *Cutting a round cake on scientific principles*. Nature, volume 75 : page 173, 1906.
- [56] Jensen, AR. *The g factor : The science of mental ability*. Praeger, Westport, Connecticut and London, 1998.
- [57] Galton, F. *Co-relations and their measurement, chiefly from anthropometric data*. Proceedings of the Royal Society of London, volume 45(273-279) : pages 135–145, 1888.
- [58] Swinburne, RG. *Galton's law—formulation and development*. Annals of science, volume 21(1) : pages 15–31, 1965.
- [59] Galton, F. *Hereditary talent and character*. Macmillan's magazine, volume 12(157-166) : pages 318–327, 1865.
- [60] Darbishire, AD. *On the result of crossing Japanese waltzing with albino mice*. Biometrika, volume 3(1) : pages 1–51, 1904.
- [61] Provine, WB. *The Origins of Theoretical Population Genetics*. The University of Chicago Press, Chicago and London, 1971.
- [62] Galton, F. *Hereditary genius*. Macmillan and Company, 1869.
- [63] Galton, F. *Eugenics : Its definition, scope, and aims*. American Journal of Sociology, volume 10(1) : pages 1–25, 1904.
- [64] Pearson, K, Moul, M. *The problem of alien immigration into Great Britain, illustrated by an examination of Russian and Polish Jewish children*. Annals of Human Genetics, volume 1(1) : pages 5–54, 1925.
- [65] Bateson, W. *Commonsense in racial problems*. The Eugenics review, volume 13(1) : page 325, 1921.
- [66] Morgan, TH. *Evolution and Genetics*. Princeton University Press, Princeton, 1925.
- [67] Yule, GU. *Mendel's laws and their probable relation to intra-racial heredity*. New Phytologist, volume 1 : pages 194–207, 222–238, 1902.
- [68] Fisher, RA. *The correlation between relatives on the supposition of Mendelian inheritance*. Transactions of the Royal Society of Edinburgh, volume 52(02) : pages 399–433, 1918.



- [69] Chebyshev, PL. *Sur deux théorèmes relatifs aux probabilités*. Acta mathematica, volume 14 : pages 305–315, 1890.
- [70] Wright, S. *The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs*. Proceedings of the National Academy of Sciences, volume 6(6) : pages 320–332, 1920.
- [71] Wright, S. *Correlation and causation*. Journal of agricultural research, volume 20(7) : pages 557–585, 1921.
- [72] Henderson, CR. *Estimation of variance and covariance components*. Biometrics, volume 9(2) : pages 226–252, 1953.
- [73] Henderson, CR. *Best linear unbiased estimation and prediction under a selection model*. Biometrics, pages 423–447, 1975.
- [74] Amos, CI. *Robust variance-components approach for assessing genetic linkage in pedigrees*. Am J Hum Genet, volume 54(3) : page 535, 1994.
- [75] Bates, D, Mächler, M, Bolker, B, Walker, S. *Fitting linear mixed-effects models using lme4*. Journal of Statistical Software, volume 67(1) : pages 1–48, 2015.
- [76] Yang, J, Lee, SH, Goddard, ME, Visscher, PM. *GCTA : a tool for genome-wide complex trait analysis*. Am J Hum Genet, volume 88(1) : pages 76–82, 2011.
- [77] Yang, J, Bakshi, A, Zhu, Z, et coll. *Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index*. Nat Genet, 2015.
- [78] Yang, J, Benyamin, B, McEvoy, BP, et coll. *Common SNPs explain a large proportion of the heritability for human height*. Nat Genet, volume 42(7) : pages 565–569, 2010.
- [79] Lee, SH, Wray, NR, Goddard, ME, Visscher, PM. *Estimating missing heritability for disease from genome-wide association studies*. Am J Hum Genet, volume 88(3) : pages 294–305, 2011.
- [80] Searle, SR, Casella, G, McCulloch, CE. *Variance components*. John Wiley & Sons, New-York, 2009.
- [81] Lin, X. *Variance component testing in generalised linear models with random effects*. Biometrika, volume 84(2) : pages 309–326, 1997.
- [82] Liu, D, Lin, X, Ghosh, D. *Semiparametric regression of multidimensional genetic pathway data : Least-squares kernel machines and linear mixed models*. Biometrics, volume 63(4) : pages 1079–1088, 2007.
- [83] Davies, RB. *Algorithm AS 155 : The distribution of a linear combination of  $\chi^2$  random variables*. J R Stat Soc Ser C Appl Stat, volume 29(3) : pages 323–333, 1980.
- [84] Kang, HM, Zaitlen, NA, Wade, CM, et coll. *Efficient control of population structure in model organism association mapping*. Genetics, volume 178(3) : pages 1709–1723, 2008.
- [85] Kang, HM, Sul, JH, Service, SK, et coll. *Variance component model to account for sample structure in genome-wide association studies*. Nat Genet, volume 42(4) : pages 348–354, 2010.

## Bibliographie

- [86] Lippert, C, Listgarten, J, Liu, Y, et coll. *FaST linear mixed models for genome-wide association studies*. Nat Methods, volume 8(10) : pages 833–835, 2011.
- [87] Kempthorne, O, Osborne, RH. *The interpretation of twin data*. Am J Hum Genet, volume 13(3) : page 320, 1961.
- [88] Scarr, S. *Environmental bias in twin studies*. Eugenics Quarterly, volume 15(1) : pages 34–40, 1968.
- [89] Scarr, S, Carter-Saltzman, L. *Twin method : Defense of a critical assumption*. Behav Genet, volume 9(6) : pages 527–542, 1979.
- [90] Feero, WG, Guttmacher, AE, Manolio, TA. *Genomewide association studies and assessment of the risk of disease*. New England Journal of Medicine, volume 363(2) : pages 166–176, 2010.
- [91] Donnelly, P. *Progress and challenges in genome-wide association studies in humans*. Nature, volume 456(7223) : pages 728–731, 2008.
- [92] Maher, B. *Personal genomes : The case of the missing heritability*. Nature News, volume 456(7218) : pages 18–21, 2008.
- [93] Visscher, PM, Hill, WG, Wray, NR. *Heritability in the genomics era—concepts and misconceptions*. Nat Rev Genet, volume 9(4) : pages 255–266, 2008.
- [94] Zuk, O, Hechter, E, Sunyaev, SR, Lander, ES. *The mystery of missing heritability : Genetic interactions create phantom heritability*. Proceedings of the National Academy of Sciences, volume 109(4) : pages 1193–1198, 2012.
- [95] Manolio, TA, Collins, FS, Cox, NJ, et coll. *Finding the missing heritability of complex diseases*. Nature, volume 461(7265) : pages 747–753, 2009.
- [96] Gusev, A, Bhatia, G, Zaitlen, N, et coll. *Quantifying missing heritability at known GWAS loci*. PLoS genetics, volume 9(12) : page e1003993, 2013.
- [97] Speed, D, Hemani, G, Johnson, MR, Balding, DJ. *Improved heritability estimation from genome-wide SNPs*. Am J Hum Genet, volume 91(6) : pages 1011–1021, 2012.
- [98] Goddard, M, Lee, H, Yang, J, et coll. *Response to Browning and Browning*. American journal of human genetics, volume 89(1) : pages 193–195, 2011.
- [99] Browning, SR, Browning, BL. *Population structure can inflate SNP-based heritability estimates*. Am J Hum Genet, volume 89(1) : page 191, 2011.
- [100] Price, AL, Patterson, NJ, Plenge, RM, et coll. *Principal components analysis corrects for stratification in genome-wide association studies*. Nat Genet, volume 38(8) : pages 904–909, 2006.
- [101] Zhang, Y, Pan, W. *Principal component regression and linear mixed model in association analysis of structured samples : competitors or complements ?* Genet Epidemiol, volume 39(3) : pages 149–155, 2015.
- [102] Janss, L, de Los Campos, G, Sheehan, N, Sorensen, D. *Inferences from genomic models in stratified populations*. Genetics, volume 192(2) : pages 693–704, 2012.

- [103] Segura, V, Vilhjálmsson, BJ, Platt, A, et coll. *An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations*. Nature genetics, volume 44(7) : pages 825–830, 2012.
- [104] Price, AL, Zaitlen, NA, Reich, D, Patterson, N. *New approaches to population stratification in genome-wide association studies*. Nature Reviews Genetics, volume 11(7) : pages 459–463, 2010.
- [105] Zhu, Z, Bakshi, A, Vinkhuyzen, AA, et coll. *Dominance genetic variation contributes little to the missing heritability for human complex traits*. The American Journal of Human Genetics, volume 96(3) : pages 377–385, 2015.
- [106] Novembre, J, Johnson, T, Bryc, K, et coll. *Genes mirror geography within Europe*. Nature, volume 456(7218) : pages 98–101, 2008.
- [107] Heath, SC, Gut, IG, Brennan, P, et coll. *Investigation of the fine structure of European populations with applications to disease association studies*. Eur J Hum Genet, volume 16(12) : pages 1413–1429, 2008.
- [108] Zhou, X, Stephens, M. *Genome-wide efficient mixed-model analysis for association studies*. Nat Genet, volume 44(7) : pages 821–824, 2012.
- [109] Aulchenko, YS, De Koning, DJ, Haley, C. *Genomewide rapid association using mixed model and regression : a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis*. Genetics, volume 177(1) : pages 577–585, 2007.
- [110] Hoffman, GE. *Correcting for population structure and kinship using the linear mixed model : theory and extensions*. PloS one, volume 8(10) : page e75707, 2013.
- [111] Price, AL, Weale, ME, Patterson, N, et coll. *Long-range LD can confound genome scans in admixed populations*. Am J Hum Genet, volume 83(1) : pages 132–135, 2008.
- [112] Anderson, CA, Pettersson, FH, Clarke, GM, et coll. *Data quality control in genetic case-control association studies*. Nat Protoc, volume 5(9) : pages 1564–1573, 2010.
- [113] Guerrero, VM, Johnson, RA. *Use of the Box-Cox transformation with binary response models*. Biometrika, volume 69(2) : pages 309–314, 1982.
- [114] Diaconis, P, Shahshahani, M. *The subgroup algorithm for generating uniform random variables*. Probability in the engineering and informational sciences, volume 1(1) : pages 15–32, 1987.
- [115] Speed, D, Balding, DJ. *MultiBLUP : improved SNP-based prediction for complex traits*. Genome Res, volume 24(9) : pages 1550–1557, 2014.
- [116] Habier, D, Fernando, R, Dekkers, J. *The impact of genetic relationship information on genome-assisted breeding values*. Genetics, volume 177(4) : pages 2389–2397, 2007.
- [117] Habier, D, Tetens, J, Seefried, FR, et coll. *The impact of genetic relationship information on genomic breeding values in German Holstein cattle*. Genet Select Evol, volume 42(1) : page 5, 2010.
- [118] Nelson, RM, Pettersson, ME, Carlborg, Ö. *A century after Fisher : time for a new paradigm in quantitative genetics*. Trends in Genetics, volume 29(12) : pages 669–676, 2013.

- [119] Génin, E, Clerget-Darpoux, F. *The missing heritability paradigm : a dramatic resurgence of the GIGO Syndrome in genetics*. Hum Heredity, volume 79(1) : pages 10–13, 2015.
- [120] Group, CS, et coll. *Vascular factors and risk of dementia : design of the Three-City Study and baseline characteristics of the study population*. Neuroepidemiology, volume 22(6) : page 316, 2003.
- [121] Weedon, MN, Lango, H, Lindgren, CM, et coll. *Genome-wide association analysis identifies 20 loci that influence adult height*. Nature genetics, volume 40(5) : pages 575–583, 2008.
- [122] Gudbjartsson, DF, Walters, GB, Thorleifsson, G, et coll. *Many sequence variants affecting diversity of adult human height*. Nature genetics, volume 40(5) : pages 609–615, 2008.
- [123] Lettre, G, Jackson, AU, Gieger, C, et coll. *Identification of ten loci associated with height highlights new biological pathways in human growth*. Nature genetics, volume 40(5) : pages 584–591, 2008.
- [124] Allen, HL, Estrada, K, Lettre, G, et coll. *Hundreds of variants clustered in genomic loci and biological pathways affect human height*. Nature, volume 467(7317) : pages 832–838, 2010.
- [125] Visscher, PM. *Sizing up human height variation*. Nature genetics, volume 40(5) : pages 489–490, 2008.
- [126] Lanktree, MB, Guo, Y, Murtaza, M, et coll. *Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height*. The American Journal of Human Genetics, volume 88(1) : pages 6–18, 2011.
- [127] van der Valk, RJ, Kreiner-Møller, E, Kooijman, MN, et coll. *A novel common variant in DCST2 is associated with length in early life and height in adulthood*. Human molecular genetics, volume 24(4) : pages 1155–1168, 2015.
- [128] Frayling, TM, Timpson, NJ, Weedon, MN, et coll. *A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity*. Science, volume 316(5826) : pages 889–894, 2007.
- [129] Willer, CJ, Speliotes, EK, Loos, RJ, et coll. *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation*. Nature genetics, volume 41(1) : pages 25–34, 2009.
- [130] Speliotes, EK, Willer, CJ, Berndt, SI, et coll. *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index*. Nature genetics, volume 42(11) : pages 937–948, 2010.
- [131] Thorleifsson, G, Walters, GB, Gudbjartsson, DF, et coll. *Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity*. Nature genetics, volume 41(1) : pages 18–24, 2009.
- [132] Loos, RJ. *Genetic determinants of common obesity and their value in prediction*. Best practice & research Clinical endocrinology & metabolism, volume 26(2) : pages 211–226, 2012.

- [133] Scuteri, A, Sanna, S, Chen, WM, et coll. *Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits*. PLoS genetics, volume 3(7) : page e115, 2007.
- [134] Loos, RJ, Lindgren, CM, Li, S, et coll. *Common variants near MC4R are associated with fat mass, weight and risk of obesity*. Nature genetics, volume 40(6) : pages 768–775, 2008.
- [135] Heid, IM, Jackson, AU, Randall, JC, et coll. *Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution*. Nature genetics, volume 42(11) : pages 949–960, 2010.
- [136] Lindgren, CM, Heid, IM, Randall, JC, et coll. *Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution*. PLoS Genet, volume 5(6) : page e1000508, 2009.
- [137] Yoneyama, S, Guo, Y, Lanktree, MB, et coll. *Gene-centric meta-analyses for central adiposity traits in up to 57,412 individuals of european descent confirm known loci and reveal several novel associations*. Human molecular genetics, page ddt626, 2013.
- [138] Visscher, PM, Macgregor, S, Benyamin, B, et coll. *Genome partitioning of genetic variation for height from 11,214 sibling pairs*. The American Journal of Human Genetics, volume 81(5) : pages 1104–1110, 2007.
- [139] Macgregor, S, Cornes, BK, Martin, NG, Visscher, PM. *Bias, precision and heritability of self-reported and clinically measured height in Australian twins*. Human genetics, volume 120(4) : pages 571–580, 2006.
- [140] Silventoinen, K, Sammalisto, S, Perola, M, et coll. *Heritability of adult body height : a comparative study of twin cohorts in eight countries*. Twin research, volume 6(05) : pages 399–408, 2003.
- [141] Polderman, TJ, Benyamin, B, de Leeuw, CA, et coll. *Meta-analysis of the heritability of human traits based on fifty years of twin studies*. Nature genetics, 2015.
- [142] Chen, X, Kuja-Halkola, R, Rahman, I, et coll. *Dominant genetic variation and missing heritability for human complex traits : Insights from twin versus genome-wide common SNP models*. The American Journal of Human Genetics, volume 97(5) : pages 708–714, 2015.
- [143] Yang, J, Manolio, TA, Pasquale, LR, et coll. *Genome partitioning of genetic variation for complex traits using common SNPs*. Nature genetics, volume 43(6) : pages 519–525, 2011.
- [144] Visscher, PM, Yang, J, Goddard, ME. *A commentary on ‘common SNPs explain a large proportion of the heritability for human height’ by Yang et al. (2010)*. Twin Research and Human Genetics, volume 13(06) : pages 517–524, 2010.
- [145] Yang, J, Bakshi, A, Zhu, Z, et coll. *Genome-wide genetic homogeneity between sexes and populations for human height and body mass index*. Human Molecular Genetics, page ddv443, 2015.

## Bibliographie

- [146] Elks, CE, Den Hoed, M, Zhao, JH, et coll. *Variability in the heritability of body mass index : a systematic review and meta-regression*. Frontiers in endocrinology, volume 3, 2012.
- [147] Smit, DJ, Luciano, M, Bartels, M, et coll. *Heritability of head size in dutch and australian twin families at ages 0–50 years*. Twin Research and Human Genetics, volume 13(04) : pages 370–380, 2010.
- [148] Ermakov, S, Kobylansky, E, Livshits, G. *Quantitative genetic study of head size related phenotypes in ethnically homogeneous Chuvasha pedigrees*. Annals of human biology, volume 32(5) : pages 585–598, 2005.
- [149] Lambert, JC, Heath, S, Even, G, et coll. *Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer’s disease*. Nat Genet, volume 41(10) : pages 1094–1099, 2009.
- [150] Stram, DO, Lee, JW. *Variance components testing in the longitudinal mixed effects model*. Biometrics, pages 1171–1177, 1994.
- [151] Wu, MC, Kraft, P, Epstein, MP, et coll. *Powerful SNP-set analysis for case-control genome-wide association studies*. Am J Hum Genet, volume 86(6) : pages 929–942, 2010.
- [152] Wu, MC, Lee, S, Cai, T, et coll. *Rare-variant association testing for sequencing data with the sequence kernel association test*. Am J Hum Genet, volume 89(1) : pages 82–93, 2011.
- [153] Traglia, M, Sala, C, Masciullo, C, et coll. *Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes*. PloS one, volume 4(10) : page e7554, 2009.
- [154] Pilia, G, Chen, WM, Scuteri, A, et coll. *Heritability of cardiovascular and personality traits in 6,148 Sardinians*. PLoS genetics, volume 2(8) : page e132, 2006.
- [155] Abney, M, McPeck, MS, Ober, C. *Broad and narrow heritabilities of quantitative traits in a founder population*. The American Journal of Human Genetics, volume 68(5) : pages 1302–1307, 2001.
- [156] Colonna, V, Natile, T, Astore, M, et coll. *Campora : a young genetic isolate in South Italy*. Human heredity, volume 64(2) : pages 123–135, 2007.
- [157] Zou, F, Lee, S, Knowles, MR, Wright, FA. *Quantification of population structure using correlated SNPs by shrinkage principal components*. Human heredity, volume 70(1) : pages 9–22, 2010.
- [158] Lee, SH, Yang, J, Chen, GB, et coll. *Estimation of SNP heritability from dense genotype data*. The American Journal of Human Genetics, volume 93(6) : pages 1151–1155, 2013.
- [159] Breslow, NE, Lin, X. *Bias correction in generalised linear mixed models with a single component of dispersion*. Biometrika, volume 82(1) : pages 81–91, 1995.
- [160] Breslow, N. *Whither PQL ?* In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 1–22. Springer, 2004.

- [161] Dasgupta, S. *Learning mixtures of gaussians*. In *Foundations of computer science, 1999. 40th annual symposium on*, pages 634–644. IEEE, 1999.
- [162] Bulik-Sullivan, BK, Loh, PR, Finucane, HK, et coll. *LD-Score regression distinguishes confounding from polygenicity in genome-wide association studies*. *Nature genetics*, volume 47(3) : pages 291–295, 2015.





## Annexe A.

### Détails sur le modèle linéaire mixte

Nous rassemblons ici quelques calculs dont l'absence dans le texte principal n'aura fait de peine à personne.

#### A.1. Résultats techniques pour le calcul de la vraisemblance restreinte

Soit  $X \in \mathbb{R}^{n \times p}$  une matrice de rang  $p$ , et soit  $C \in \mathbb{R}^{(n-p) \times n}$  une matrice de contrastes :  $CX = 0$  et  $CC' = I_p$ . Soit  $V \in \mathbb{R}^{n \times n}$  une matrice symétrique définie positive. Nous montrons dans cette annexe deux identités matricielles faisant intervenir  $C$ ,  $X$  et  $V$ , utiles pour le calcul de la vraisemblance restreinte du modèle linéaire mixte, et nous finissons par exprimer le déterminant de  $CVC'$  en fonction de  $V$  et  $X$  uniquement.

Commençons par un résultat préliminaire :

##### A.1.1. Projections sur des espaces orthogonaux complémentaires

Soit  $A \in \mathbb{R}^{n \times (n-p)}$ ,  $B \in \mathbb{R}^{n \times p}$ , avec  $\text{rank}(A) = n - p$  and  $\text{rank}(B) = p$ . Si  $A'B = 0$ , alors

$$A(A'A)^{-1}A' + B(B'B)^{-1}B' = I_n$$

**Preuve**  $A(A'A)^{-1}A'$  est la matrice de la projection sur  $\mathfrak{I}A$  (l'espace vectoriel engendré par les colonnes de  $A$ ), et  $B(B'B)^{-1}B'$  est la matrice de la projection sur  $\mathfrak{I}B$ . Les hypothèses sont équivalentes au fait que  $\mathfrak{I}A$  et  $\mathfrak{I}B$  sont des espaces orthogonaux, et que leur somme directe est l'espace  $\mathbb{R}^n$  entier ; le résultat annoncé est alors immédiat.

### Deux identités matricielles

Soit  $X \in \mathbb{R}^{n \times p}$  de rang  $p$ , et soit  $C \in \mathbb{R}^{(n-p) \times n}$  avec  $CC' = I_p$ . On applique le résultat précédent à  $A = C'$  et  $B = X$ , ce qui donne

$$C'C = I_n - X(X'X)^{-1}X'. \quad (\text{A.1})$$

Si  $V$  est définie positive, le même résultat cette fois avec  $A = V^{\frac{1}{2}}C'$  et  $B = V^{-\frac{1}{2}}X$ , implique

$$V^{\frac{1}{2}}C'(CVC')^{-1}CV^{\frac{1}{2}} = I_n - V^{-\frac{1}{2}}X(X'V^{-1}X)^{-1}X'V^{-\frac{1}{2}}$$

et donc

$$C'(CVC')^{-1}C = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \quad (\text{A.2})$$

Cette matrice est notée  $P$  dans le contexte de la vraisemblance restreinte (cf équation 2.12). Notons que  $P$  est une matrice singulière ; on a  $\mathfrak{I}P = \mathfrak{I}C' = (\mathfrak{I}X)^\perp$ , et  $PVP = P$ .

#### A.1.2. Le déterminant de $CVC'$

Nous allons montrer que

$$|CVC'| \times |X'X| = |V| \times |X'V^{-1}X|$$

or

$$\log |CVC'| + \log |X'X| = \log |V| + \log |X'V^{-1}X|. \quad (\text{A.3})$$

**Première preuve** Pour  $t \in [0,1]$ , soit  $V(t) = (1-t)I_n + tV$ , de sorte que  $V(0) = I_n$  et  $V(1) = V$  ; pour tout  $t \in [0,1]$ ,  $V(t)$  est définie positive, et les fonctions suivantes sont définies :

$$\begin{aligned} f(t) &= \log |CV(t)C'| + \log |X'X| \\ g(t) &= \log |V(t)| + \log |X'V(t)^{-1}X|. \end{aligned}$$

On a  $f(0) = g(0)$ . Si on montre que  $f'(t) = g'(t)$ , alors  $f(1) = g(1)$  ce qui implique le résultat. Notons que  $V'(t) = V$ .

On calcule

$$\begin{aligned} f'(t) &= \text{tr} \left( (CV(t)C')^{-1} CVC' \right) \\ &= \text{tr} \left( C' (CV(t)C')^{-1} C \times V \right) \\ &= \text{tr} \left( V(t)^{-1} \times V \right) - \text{tr} \left( V(t)^{-1} X (X'V(t)^{-1}X)^{-1} X'V(t)^{-1} \times V \right), \end{aligned}$$

où on a utilisé l'équation (A.2) dans la dernière ligne. D'autre part

$$\begin{aligned} g'(t) &= \text{tr} \left( V(t)^{-1} \times V \right) + \text{tr} \left( \left( X'V(t)^{-1}X \right)^{-1} \times X' \left( -V(t)^{-1} \times V \times V(t)^{-1} \right) X \right) \\ &= \text{tr} \left( V(t)^{-1} \times V \right) - \text{tr} \left( V(t)^{-1}X \left( X'V(t)^{-1}X \right)^{-1} X'V(t)^{-1} \times V \right), \end{aligned}$$

ce qui permet de conclure.  $\square$

**Deuxième preuve** Soit  $D \in \mathbb{R}^{p \times n}$  tel que les lignes de  $D$  forment une basent orthogonales de  $\mathfrak{I}X$ , l'espace engendré par les colonnes de  $X$ . On a  $DD' = I_p$ ,  $D'D = X(X'X)^{-1}X'$ , et  $CD' = 0$ .

Soit  $U = \begin{bmatrix} C \\ D \end{bmatrix}$ ; on a  $UU' = U'U = I_n$ . Alors

$$UVU' = \begin{bmatrix} CVC' & CVD' \\ DVC' & DVD' \end{bmatrix}.$$

Le complément de Schur du bloc supérieur gauche

$$S = DVD' - DVC' (CVC')^{-1} CVD'.$$

En utilisant l'équation (2.12), on obtient

$$S = DX \left( X'V^{-1}X \right)^{-1} X'D'.$$

On a  $|V| = |CVC'| \times |S|$ . De plus

$$\begin{aligned} |S| &= |DX| \times \left| \left( X'V^{-1}X \right)^{-1} \right| \times |X'D'| \\ &= |X'D'DX| \times \left| X'V^{-1}X \right|^{-1} \\ &= |X'X| \times \left| X'V^{-1}X \right|^{-1} \end{aligned}$$

car  $D'D = X(X'X)^{-1}X'$ .  $\square$

## A.2. L'algorithme EM-REML

Nous détaillons ici les calculs nécessaires à la formulation de l'algorithme EM pour la maximisation de la vraisemblance restreinte comme une ascension de gradient avec un « pas adaptatif ». Nous commençons par le cas particulier d'un modèle sans effets fixes, avant de passer au cas général.

### A.2.1. L'algorithme EM quand il n'y a pas d'effets fixes

Considérons le modèle tel qu'écrit dans l'équation (2.4), avec  $X = 0$ . Pour simplifier l'exposé on prendra  $k = 1$ , la généralisation à  $k > 1$  étant facile. L'algorithme EM considère les effets aléatoires  $\omega$  comme des variables latentes, ou non observées. La distribution jointe de  $(Y, \omega)$  est normale, d'espérance nulle et de variance

$$\begin{bmatrix} V & \tau K \\ \tau K & \tau K \end{bmatrix}. \quad (\text{A.4})$$

Si  $\omega$  était observé, les paramètres seraient facilement estimés par

$$\tau = \frac{1}{r(K)} \omega' K^+ \omega \quad (\text{A.5})$$

où  $r(K)$  est le rang de  $K$ ,  $K^+$  est un inverse généralisé de  $K$ , et par

$$\sigma^2 = \frac{1}{n} e' e \quad (\text{A.6})$$

où on a posé  $e = Y - \omega$ .

L'algorithme EM consiste à itérer deux étapes. Dans l'étape E, en utilisant les valeurs courantes des paramètres, on écrit la distribution de  $\omega$  conditionnellement à la valeur observée  $Y$ ; dans l'étape M, paramètres sont réestimés par les espérances des expressions (A.5) et (A.6).

Soient  $\tau_r$  et  $\sigma_r^2$  les valeurs des paramètres à la  $r$ -ème itération; soit  $V_r = \tau_r K + \sigma_r^2 I_n$  la valeur correspondante de  $\text{var}(Y)$ .

#### Étape E

À cette étape, étant données les valeurs courantes des paramètres,  $\tau_r$  et  $\sigma_r^2$ , on calcule la moyenne et la variance de la distribution de  $\omega$  conditionnellement à la valeur observée  $Y$ . Il découle des propriétés classiques des lois normales multivariées, que

$$E(\omega|Y) = \tau_r K V_r^{-1} Y \quad (\text{A.7})$$

et

$$\text{var}(\omega|Y) = \tau_r K - \tau_r^2 K V_r^{-1} K. \quad (\text{A.8})$$

On calcule également l'espérance et la variance conditionnelle de  $e = Y - w$ , dont on aura besoin à l'étape M :

$$\begin{aligned} E(e|Y) &= Y - E(\omega|Y) = Y - \tau_r K V_r^{-1} Y \\ &= (V_r - \tau_r K) V_r^{-1} Y \\ &= \sigma_r^2 V_r^{-1} Y. \end{aligned} \quad (\text{A.9})$$

La variance conditionnelle de  $e$  est  $\text{var}(e|Y) = \text{var}(\omega|Y)$ . En remplaçant dans l'équation (A.8),  $\tau_r K$  par  $V_r - \sigma_r^2 I_n$  on peut la réécrire

$$\text{var}(e|Y) = \sigma_r^2 I_n - (\sigma_r^2)^2 V_r^{-1}. \quad (\text{A.10})$$

### Étape M

À cette étape, on utilise les quantités calculées à l'étape E pour estimer de nouvelles valeurs  $\tau_{r+1}$  et  $\sigma_{r+1}^2$  des paramètres (les espérances et les variances conditionnelles ci-dessous sont calculées pour  $\tau = \tau_r$  et  $\sigma^2 = \sigma_r^2$ ) :

$$\tau_{r+1} = E \left( \frac{1}{r(K)} \omega' K^+ \omega \middle| Y \right) = \frac{1}{r(K)} (E(\omega|Y)' K^+ E(\omega|Y) + \text{tr} (K^+ \text{var}(\omega|Y))) \quad (\text{A.11})$$

et

$$\sigma_{r+1}^2 = E \left( \frac{1}{n} e' e \middle| Y \right) = \frac{1}{n} (E(e|Y)' E(e|Y) + \text{tr} (\text{var}(e|Y))). \quad (\text{A.12})$$

### Rassembler les deux étapes

En utilisant les équations (A.7) et (A.8), on obtient

$$\begin{aligned} E(\omega|Y)' K^+ E(\omega|Y) &= \tau_r^2 Y' V_r^{-1} K V_r^{-1} Y \\ \text{tr} (K^+ \text{var}(\omega|Y)) &= \tau_r \text{tr} (K^+ K) - \tau_r^2 \text{tr} (K^+ K V_r^{-1} K) \\ &= \tau_r r(K) - \tau_r^2 \text{tr} (K V_r^{-1}) \end{aligned}$$

En introduisant ceci dans l'équation (A.11), on a

$$\tau_{r+1} = \tau_r + \frac{1}{r(K)} \tau_r^2 (Y' V_r^{-1} K V_r^{-1} Y - \text{tr} (K V_r^{-1})) \quad (\text{A.13})$$

$$= \tau_r + \left( \frac{2\tau_r^2}{r(K)} \right) \frac{\partial \ell}{\partial \tau} (\tau_r, \sigma_r^2). \quad (\text{A.14})$$

Similairement, en introduisant (A.9) et (A.10) dans l'équation (A.12), on obtient

$$\sigma_{r+1}^2 = \sigma_r^2 + \frac{1}{n} (\sigma_r^2)^2 (Y'V_r^{-1}V_r^{-1}Y - \text{tr}(V_r^{-1})) \quad (\text{A.15})$$

$$= \sigma_r^2 + \frac{2(\sigma_r^2)^2}{n} \frac{\partial \ell}{\partial \sigma^2} (\tau_r, \sigma_r^2). \quad (\text{A.16})$$

### A.2.2. L'algorithme EM-REML

L'itération des étapes E et M que nous venons de décrire permettent la maximisation de la vraisemblance. Pour la maximisation de la vraisemblance restreinte, on remplace Y par CY, K par CKC' et V par CVC' dans les équations (A.13) et (A.15); on obtient le résultat énoncé dans le texte principal

$$\begin{aligned} \tau_{r+1} &= \tau_r + \frac{1}{r(K)} \tau_r^2 (Y'P_r K P_r Y - \text{tr}(K P_r)) \\ &= \tau_r + \left( \frac{2\tau_r^2}{r(K)} \right) \frac{\partial \ell^{\text{re}}}{\partial \tau} (\tau_r, \sigma_r^2) \end{aligned}$$

et

$$\begin{aligned} \sigma_{r+1}^2 &= \sigma_r^2 + \frac{1}{n} (\sigma_r^2)^2 (Y'P_r P_r Y - \text{tr}(P_r)) \\ &= \sigma_r^2 + \frac{2(\sigma_r^2)^2}{n} \frac{\partial \ell^{\text{re}}}{\partial \sigma^2} (\tau_r, \sigma_r^2). \end{aligned}$$

### A.3. Détails sur l'astuce de la diagonalisation

Nous développons ici brièvement les techniques utilisées dans `gaston` pour l'estimation rapide des paramètres d'un modèle mixte quand il n'y a qu'une matrice  $K$  dans le modèle.

L'astuce repose sur l'utilisation de la décomposition en éléments simples de  $K$ . Si on ne tient pas compte du coût préliminaire de cette décomposition – il est en  $O(n^3)$  – la complexité de la maximisation de la vraisemblance est linéaire en  $n$ , au lieu d'être cubique, le calcul de  $V^{-1}$  à chaque étape de l'algorithme AIREML étant en  $O(n^3)$ .

Une place spéciale est faite à l'intégration au modèle des  $p$  premières composantes principales génomiques avec des effets fixes, qui peut se faire sans augmenter la complexité.

#### A.3.1. Réécriture du modèle sous forme diagonale

Soit  $Z \in \mathbb{R}^{n \times q}$  la matrice des covariables à effets aléatoires, et soit  $K = ZZ'$ . On note

$$Z = U\Sigma L'$$

sa décomposition en valeurs singulières. Elle est liée à la décomposition en éléments simples de  $K = U\Sigma^2U'$ , par  $L = Z'U\Sigma^{-1}$ . En pratique, il suffit de connaître  $U$  et  $\Sigma$  pour les calculs qui suivent. Les colonnes de  $U \in \mathbb{R}^{n \times n}$  sont les composantes principales normées des génotypes. On décompose  $U$  en deux blocs,  $U = [U_1 \ U_2]$  où  $U_1$  est formé des  $p$  premières colonnes et  $U_2$  des  $n - p$  dernières.

On considère ici le modèle mixte

$$Y = X\beta + Zu + e \tag{A.17}$$

avec  $u \sim \mathcal{N}(0, \tau I_q)$  et  $e \sim \mathcal{N}(0, \sigma^2 I_n)$ .

Les covariables avec effets fixes  $X$  sont faites de  $s$  « vraies covariables » (dont une colonne de 1 pour l'inclusion d'un intercept) et de  $p$  composantes principales normées :

$$X = [X_{\text{cov}} \ U_1] \in \mathbb{R}^{n \times (s+p)}$$

Soient  $Y_D = U'Y$  et  $X_D = U'X$ . On réécrit (A.17) comme

$$Y_D = X_D\beta + \Sigma w + e_D \tag{A.18}$$

où  $w = L'u \sim \mathcal{N}(0, \tau I_n)$  et  $e_D = U'e \sim \mathcal{N}(0, \sigma^2 I_n)$ .

### A.3.2. Une vraisemblance restreinte profilée

On réécrit la vraisemblance restreinte avec deux paramètres  $v$  et  $h^2$ , en posant  $\tau = h^2v$  et  $\sigma^2 = (1 - h^2)v$ . Soit

$$V_0 = h^2\Sigma^2 + (1 - h^2)I_n$$

de sorte que  $V = \text{var}(Y_D) = v \times V_0$ .

On pose

$$P = V^{-1} - V^{-1}X_D \left( X_D' V^{-1} X_D \right)^{-1} X_D' V^{-1}$$

et

$$P_0 = V_0^{-1} - V_0^{-1}X_D \left( X_D' V_0^{-1} X_D \right)^{-1} X_D' V_0^{-1} \quad (\text{A.19})$$

de sorte que  $P = \frac{1}{v} \times P_0$ .

La vraisemblance restreinte est

$$\begin{aligned} \ell^{\text{re}}(h^2, v) &= -\frac{1}{2} \left( \log |V| + \log |X_D' V^{-1} X_D| + Y_D' P Y_D \right) \\ &= -\frac{1}{2} \left( \log |V_0| + \log |X_D' V_0^{-1} X_D| + \frac{1}{v} Y_D' P_0 Y_D + (n - s - p) \log(v) \right) \end{aligned} \quad (\text{A.20})$$

On peut profiler cette vraisemblance pour faire disparaître le paramètre  $v$ . En effet la dérivée de  $\ell^{\text{re}}(h^2, v)$  suivant  $v$  est

$$\frac{\partial \ell^{\text{re}}}{\partial v}(h^2, v) = \frac{1}{2v^2} \left( Y_D' P_0 Y_D - (n - s - p)v \right)$$

qui s'annule en  $v = \frac{1}{n-s-p} Y_D' P_0 Y_D$ .

En introduisant cette valeur dans (A.20) on obtient une vraisemblance restreinte profilée

$$\ell^{\text{re.P}}(h^2) = -\frac{1}{2} \left( \log |V_0| + \log |X_D' V_0^{-1} X_D| + (n - s - p) \log(Y_D' P_0 Y_D) \right).$$

C'est une fonction d'une seule variable,  $h^2$ , qui peut être maximisée très efficacement par la méthode de Newton.

Notons que si il n'y a pas de covariables,  $s + p = 0$ , alors  $V_0$  et  $P_0$  sont des matrices diagonales et la complexité des calculs est linéaire en  $n$ . De façon générale, toutes les expressions matricielles qui apparaissent dans le calcul de la vraisemblance restreinte profilée et de ses dérivées (ainsi que des BLUPs, etc) peuvent être effectués par blocs. La méthode du complément de Schur est notamment utilisée pour calculer  $V_0^{-1}$ , et on se contente d'inverser des matrices de dimension  $s \times s$ , avec une complexité en  $O(s^3)$ . En pratique, ce terme est négligeable devant la complexité en  $O(n)$  nécessaire à la manipulation des matrices diagonales.



## Annexe B.

### Sur certaines matrices aléatoires

Nous avons proposé de générer des GRM aléatoires en posant

$$K = U\Lambda U'$$

où  $U$  est une matrice orthogonale aléatoire de dimensions  $n \times n$  (tirée selon la loi uniforme sur le groupe des matrices orthogonales, c'est-à-dire l'unique loi invariante par multiplication par toute matrice orthogonale  $\Gamma$ ), et  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  est une matrice diagonale (non aléatoire). On note (abusivement)

$$E(\Lambda) = \frac{1}{n} \sum_k \lambda_k$$
$$\text{var}(\Lambda) = \frac{1}{n} \sum_k (\lambda_k - E(\lambda))^2 = \frac{1}{n} \sum_k \lambda_k^2 - \frac{1}{n^2} \left( \sum_k \lambda_k \right)^2$$

Nous montrons ici que l'espérance des termes diagonaux de  $K$  est égale à  $E(\Lambda)$ , et que les termes hors diagonale sont centrés et de variance  $\simeq \frac{1}{n} \text{var}(\Lambda)$  (pour  $n$  grand).

Avant de montrer cela, il faut obtenir quelques résultats sur la loi des matrices  $U$ .

#### B.1. Lois uniformes sur le groupe des matrices orthogonales et sur la sphère

Nous appelons « loi uniforme sur le groupe des matrices orthogonales » l'unique loi qui est invariante par multiplication à gauche ou à droite par une matrice orthogonale  $\Gamma$  : la densité en  $U$  est égale à la densité en  $U\Gamma$  ou en  $\Gamma U$ , pour n'importe quelle matrice  $\Gamma$ .

Les matrices orthogonales incluant les matrices de rotation, on en déduit que la loi marginale d'une ligne ou d'une colonne d'une matrice orthogonale  $U$  tirée dans cette loi est la loi uniforme sur la sphère  $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ .

Les éléments de  $\mathbb{S}^{n-1}$  s'écrivent naturellement à l'aide des coordonnées sphériques  $(\phi_1, \dots, \phi_{n-1})$ , avec  $\phi_1, \dots, \phi_{n-2} \in [0, \pi]$  et  $\phi_{n-1} \in [0, 2\pi[$ .

Le changement de variable est donné par

$$\begin{aligned}x_1 &= \cos \phi_1 \\x_2 &= \sin \phi_1 \cos \phi_2 \\x_3 &= \sin \phi_1 \sin \phi_2 \cos \phi_3 \\&\vdots \\x_{n-1} &= \sin \phi_1 \dots \sin \phi_{n-2} \cos \phi_{n-1} \\x_n &= \sin \phi_1 \dots \sin \phi_{n-2} \sin \phi_{n-1}\end{aligned}$$

et l'élément de surface est  $\sin^{n-2} \phi_1 \dots \sin \phi_{n-2} d\phi_1 \dots d\phi_{n-1}$ . En coordonnées sphériques, la loi uniforme sur  $\mathbb{S}^{n-1}$  est donc particulièrement agréable, puisqu'elle admet une densité et que les coordonnées  $\phi_i$  sont indépendantes, de lois marginales proportionnelles à  $\sin^{n-2} \phi_1 d\phi_1$ ,  $\sin^{n-3} \phi_2 d\phi_2$ , etc.

## Loi marginale d'un élément

On cherche la loi marginale de  $u_{ik}$ , un élément de  $U$ . D'après ce qui précède c'est la même loi que celle de  $x_1$  quand  $x = (x_1, \dots, x_n)'$  est tiré dans la loi uniforme sur  $\mathbb{S}^{n-1}$ .

On a  $x_1 = \cos \phi_1$  et la densité de  $\phi_1$  est proportionnelle à  $\sin^{n-2} \phi_1$ . La densité de  $x_1$  est donc proportionnelle à

$$\frac{1}{|\sin(\arccos x_1)|} \sin^{n-2}(\arccos x_1) = (1 - x_1^2)^{\frac{1}{2}(n-3)}.$$

pour  $x_1 \in [-1, 1]$ . La constante d'intégration est

$$\rho_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)}.$$

On peut se servir de cette densité pour calculer  $\text{var}(x_1) = E(x_1^2) = \frac{1}{n}$ . On pouvait aussi l'avoir directement en prenant l'espérance de  $\sum_k x_k^2 = 1$ .

## Application à la projection sur un vecteur quelconque

On a trouvé la densité de  $x_1 = e_1'x$ ; la symétrie sphérique de la loi de  $x$  fait que c'est aussi la loi de  $u'x$  pour n'importe quel vecteur unitaire  $u$ . On en déduit que si  $u$  est de norme quelconque, la densité de  $t = u'x$  est

$$\frac{\rho_n}{||u||} \left(1 - \frac{t^2}{||u||^2}\right)^{\frac{1}{2}(n-3)} \quad \text{pour } |t| \leq ||u||.$$

## Loi jointe de deux éléments d'une même ligne ou d'une même colonne

À nouveau, pour obtenir la loi jointe de  $(u_{ik}, u_{jk})$ , il suffit de chercher la loi de  $(x_1, x_2)$  quand  $x = (x_1, \dots, x_n)'$  est tiré dans la loi uniforme sur  $\mathbb{S}^{n-1}$ .

On a  $(x_1, x_2) = (\cos \phi_1, \sin \phi_1 \cos \phi_2)$  et la densité de  $(\phi_1, \phi_2)$  est proportionnelle à  $\sin^{n-2} \phi_1 \sin^{n-3} \phi_2$ . La matrice jacobienne du changement de variable est

$$\begin{pmatrix} -\sin \phi_1 & 0 \\ \cos \phi_1 \cos \phi_2 & -\sin \phi_1 \sin \phi_2 \end{pmatrix}$$

d'où le jacobien :  $\sin^2 \phi_1 \sin \phi_2$ .

On a besoin du changement de variable réciproque :

$$\begin{aligned} \phi_1 &= \arccos x_1 \\ \phi_2 &= \arccos \left( \frac{x_2}{\sin \phi_1} \right) = \arccos \left( \frac{x_2}{\sqrt{1-x_1^2}} \right) \end{aligned}$$

La densité de  $(x_1, x_2)$  est proportionnelle à

$$\begin{aligned} & \frac{\sin^{n-2}(\arccos x_1) \sin^{n-3} \left( \arccos \frac{x_2}{\sqrt{1-x_1^2}} \right)}{\left| \sin^2(\arccos x_1) \sin \left( \arccos \frac{x_2}{\sqrt{1-x_1^2}} \right) \right|} \\ &= \sin^{n-4}(\arccos x_1) \sin^{n-4} \left( \arccos \frac{x_2}{\sqrt{1-x_1^2}} \right) \\ &= (1-x_1^2)^{\frac{1}{2}(n-4)} \left( 1 - \frac{x_2^2}{1-x_1^2} \right)^{\frac{1}{2}(n-4)} \\ &= \left( 1 - (x_1^2 + x_2^2) \right)^{\frac{1}{2}(n-4)} \end{aligned}$$

pour  $x_1^2 + x_2^2 \leq 1$ . La constante d'intégration est  $\frac{n-2}{2\pi}$ .

La place ne nous étant pas comptée, nous allons montrer qu'on peut aussi obtenir ce résultat par une autre méthode, plus géométrique et moins calculatoire.

La loi de  $(x_1, x_2)$  c'est aussi la loi des deux premiers éléments d'une ligne de  $U$ , disons  $(x_1, y_1)$  si  $x$  et  $y$  sont les deux premières colonnes de  $U$ . On peut tirer  $x$  et  $y$  par les étapes suivantes :

- on tire  $x$  dans  $\mathbb{S}^{n-1}$

- on choisit une matrice  $C$  de dimensions  $n \times (n-1)$ , dont les colonnes forment une base de l'hyperplan orthogonale à  $x$  :

$$C'x = 0, C'C = I_{n-1}, CC' = I_n - xx'.$$

- on tire  $y^0$  dans  $\mathbb{S}^{n-2} \subset \mathbb{R}^{n-1}$  et on pose  $y = Cy^0$ .

On veut la loi jointe de  $x_1 = e_1'x$  et  $y_1 = e_1'y = e_1'Cy^0$ . La distribution de  $y_1$  conditionnellement à  $x$  et  $C$  ne dépend en fait que de  $x$  (le choix arbitraire de  $C$  ne modifie pas la loi de  $y$ ).

On veut la densité de  $y_1 = e_1'Cy^0$ . Le théorème de Pythagore donne  $\|e_1\|^2 = \|C'e_1\|^2 + x_1^2$ , donc  $\|C'e_1\| = (1 - x_1^2)^{\frac{1}{2}}$ . D'après le résultat donné plus haut sur la projection d'un vecteur aléatoire de la sphère sur un vecteur quelconque, la densité de  $y_1$  conditionnellement à  $C$  et  $x_1$  est

$$\frac{\rho_{n-1}}{(1 - x_1^2)^{\frac{1}{2}}} \left(1 - \frac{y_1^2}{1 - x_1^2}\right)^{\frac{1}{2}(n-4)}.$$

La densité jointe de  $(x_1, y_1)$  est donc

$$\rho_n \rho_{n-1} \left(1 - (x_1^2 + x_2^2)\right)^{\frac{1}{2}(n-4)},$$

et les propriétés de la fonction  $\Gamma$  donnent  $\rho_n \rho_{n-1} = \frac{n-2}{2\pi}$ .

## B.2. Loi des éléments de la GRM

On a  $K = U\Lambda U'$ , donc  $k_{ij} = \sum_k \lambda_k u_{ik} u_{jk}$ .

On a vu que  $\text{var}(u_{ik}) = \frac{1}{n}$ , on a donc  $E(k_{ii}) = \sum_k \lambda_k E(u_{ik}^2) = \frac{1}{n} \sum_k \lambda_k$ .

Posons  $\gamma_k = u_{ik} u_{jk}$ . On a

$$\text{var}(k_{ij}) = \sum_k \lambda_k^2 \text{var}(\gamma_k) + 2 \sum_{k < \ell} \lambda_k \lambda_\ell \text{cov}(\gamma_k, \gamma_\ell).$$

A partir de la loi jointe de  $(u_{ik}, u_{jk})$  trouvée ci-dessus on calcule  $E(\gamma_k) = 0$  et  $\text{var}(\gamma_k) = \frac{1}{n(n+2)}$ .

D'autre part

$$\gamma_1 + \dots + \gamma_n = \sum_{k=1}^n u_{ik} u_{jk} = 0$$

les lignes  $u_i$  et  $u_j$  étant orthogonales. On a donc

$$\sum_k \text{var}(\gamma_k) + 2 \sum_{k < \ell} \text{cov}(\gamma_k, \gamma_\ell) = 0.$$

Par symétrie tous les termes sont égaux, et on a

$$\text{cov}(\gamma_k, \gamma_\ell) = -\frac{1}{n(n-1)(n+2)}.$$

On a alors

$$\begin{aligned} \text{var}(k_{ij}) &= \frac{1}{n(n+2)} \sum \lambda_k^2 - \frac{2}{n(n-1)(n+2)} \sum_{k < \ell} \lambda_k \lambda_\ell \\ &= \frac{1}{n(n+2)} \sum \lambda_k^2 - \frac{1}{n(n-1)(n+2)} \left( \left( \sum_k \lambda_k \right)^2 - \sum \lambda_k^2 \right) \\ &= \frac{1}{(n-1)(n+2)} \left( \sum \lambda_k^2 - \frac{1}{n} \left( \sum_k \lambda_k \right)^2 \right) \\ &= \frac{n}{(n-1)(n+2)} \text{var}(\Lambda) \\ &\simeq \frac{1}{n} \text{var}(\Lambda) \end{aligned}$$