Contents

NNT : 2018SACLE029

Thèse de doctorat

Statistical learning for omics association and interaction studies based on blockwise feature compression

Thèse de doctorat de l'Université Paris-Saclay
préparée à Université d'Evry-Val-d'Essonne

Ecole doctorale n°577 Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Science de la vie et de la santé

Thèse présentée et soutenue à Evry, le 04/12/2018, par

FLORENT GUINOT

Composition du Jury :

M. Avner Bar-Hen Professeur, CNAM	Rapporteur
M. Gregory Nuel Professeur, Université de la Sorbonne	Rapporteur
Mme. Florence Jaffrezic Directrice de recherche, INRA	Présidente du jury
M. Laurent Jacob Chargé de recherche, CNRS	Examinateur
M. Vincent Segura Chargé de recherche, INRA	Examinateur
M. Christophe Ambroise Professeur, Université d'Evry-Val-d'Essonne	Directeur de thèse
Mme. Marie Szafranski Maître de conférences, ENSIIE	Co-directrice de thèse
Mme. Nathalie Jourdan Bioptimize	Co-directrice de thèse

Abstract

Since the last decade, the rapid advances in genotyping technologies have changed the way genes involved in mendelian disorders and complex diseases are mapped, moving from candidate genes approaches to linkage disequilibrium mapping. In this context, Genome-Wide Associations Studies (GWAS) aim at identifying genetic markers implied in the expression of complex disease, those occurring at different frequencies between unrelated samples of affected individuals and unaffected controls. These studies exploit the fact that it is easier to establish, from the general population, large cohorts of affected individuals sharing a genetic risk factor for a complex disease than within individual families, as it is the case in traditional linkage analysis.

From a statistical point of view, the standard approach in GWAS is based on hypothesis testing, with affected individuals being tested against healthy individuals at one or more markers. However, classical testing schemes are subject to false positives, that is markers that are falsely identified as significant. One way around this problem is to apply a correction on the p-values obtained from the tests, increasing in return the risk of missing true associations that have only a small effect on the phenotype, which is usually the case in GWAS.

Although GWAS have been successful in the identification of genetic variants associated with complex multifactorial diseases (Crohn's disease, diabetes I and II, coronary artery disease,...) only a small proportion of the phenotypic variations expected from classical family studies have been explained. This missing heritability may have multiple causes amongst the following: strong correlations between genetic variants, population structure, epistasis (gene by gene interactions), disease associated with rare variants,...

The main objectives of this thesis are thus to develop new methodologies that can face part of the limitations mentioned above. More specifically we developed two new approaches: the first one is a block-wise approach for GWAS analysis which leverages the correlation structure among the genomic variants to improve statistical power in the context of univariate hypothesis testing while the second focuses on the detection of interactions between groups of metagenomic and genetic markers to better understand the complex relationship between environment and genome in the expression of a given phenotype.

General introduction

Background

The foundations of modern genetics laid down in Johann Gregor Mendel's pioneering work have resulted in the understanding that certain hereditary traits can exist in different versions (alleles), introducing the notion of homozygosity and heterozygosity. It paved the way for the comprehension of heredity mechanisms with the establishment of the first genetic maps by Thomas Hunt Morgan and the definition of genetic heritability by Ronald Fisher which suggests that the expression of a trait (phenotype) is subject to both genetic and environmental factors. These groundbreaking works led to the linkage analysis studies whose purpose is to map genes involved in the expression of diseases. These approaches, effective in locating genes involved in the expression of a simple qualitative trait, have proven less reliable in mapping complex diseases. Indeed, there may be multiple interaction between genes underlying these phenotypes and the effects of these genes may vary with exposure to environmental and other non-genetic risk factors.

These limitations have driven the development of another discipline: Genome-Wide Associations Studies (GWAS). These studies aim to identify single nucleotide polymorphisms (SNP), i.e. genetic markers that occur at different frequencies between unrelated samples of affected individuals and unaffected controls, implied in the expression of a given phenotype. These studies exploit the fact that it is easier to establish large cohorts of affected individuals sharing a genetic risk factor for a complex disease in the general population than within individual families, as it is the case with traditional linkage analysis.

In addition, recent advances in genotyping technology have made it possible to genotype the entire DNA sequence of an individual at a moderate cost and within a reasonable time. Therefore, it became necessary to develop new statistical methods able to process this type of massive data.

Problematic

From a statistical point of view, looking for these genetic markers can be supported by hypothesis testing. The standard approach in GWAS is based on univariate linear regression, with affected individuals being tested against healthy individuals at one or more loci. Classical testing schemes are subject to false positives, that is SNP that are falsely identified as significant. One way around this problem is to apply a correction for the False Discovery Rate (FDR, **?**). Unfortunately, this increases the risk of missing true associations that have only a small effect on the phenotype, which is usually the case in GWAS.

Although GWAS have been successful in the identification of genetic variants associated with complex multifactorial diseases (Crohn's disease, diabetes I and II, coronary artery disease…(**?**)), only a small proportion of the phenotypic variations expected from classical family studies have been explained (**?**). This missing heritability may have multiple causes amongst the following: strong correlations between genetic variants, population structure, epistasis (gene by gene interactions), disease associated with rare variants…

Objectives

The main objectives of this thesis are to develop new methodologies, in the context of GWAS, that can face part of the limitations mentioned above. More specifically we developed two new approaches: the first one, entitled LEOS, is a blockwise approach for GWAS analysis which leverages the correlation structure among the genomic variants to reduce the number of statistical hypotheses to be tested, while the second, named SICOMORE, focuses on the detection of interactions between groups of metagenomic and genetic markers to better understand the complex relationship between environment and genome in the expression of a given phenotype.

Contributions

This thesis work gave rise to the writing of two scientific articles, one for each methodology. The method LEOS described in Chapter 4 is under minor review in the journal BMC bioinformatics while the method SICOMORE described in Chapter 5 has been published as an article of a national conference (50^{th} Journées de la statistique) but the extended version was still in a preprint status at the time this manuscript was written.

The proposed methods have been implemented in computer programs: LEOS is proposed as a webserver tool while SICOMORE is available through an R package (a vignette, added at the end of the manuscript, is available for this package).

This work has also led to several oral communications and poster presentations in the following conferences:

- *Statistical Methods for Post Genomic Data* in 2017 (poster presentation LEOS)
- *International Society for Computational Biology conference* in 2017 (poster presentation LEOS)
- *Statistical Methods in Biopharmacy* in 2017 (oral presentation LEOS).
- *Journées de statistique* in 2018 (oral presentation SICOMORE)

Contents of the manuscript

This manuscript is composed of five different chapters. The first three chapters will focus on the genetic, statistical and GWAS context while our two proposed methodologies will be presented in Chapters **??** and **??**. Chapter **??** will remind the genetic precepts fundamental to the understanding of our work while Chapter **??** will introduce the concept of statistical learning and Chapter **??** will provide an extensive introduction to GWAS by presenting some state-of-the-art statistical methods. We will also discuss the results obtained on our proposed approaches at the end of Chapters **??** and **??** before providing a general conclusion in a last section.

Notations

Notations	Explanation
n or N	number of observations
D	number of variables
d	index of the variables in $[1, ..., D]$
K	number of classes
k	index of the classes in $[1, ..., K]$
$\mathcal{T}$	training set space
$\mathcal{X}$	space of variables
X	random variable from $\mathcal{X}$
x	realization of X
$\mathbf{X}$	matrix of variables in $\mathbb{R}^{n \times D}$
x_i	column vector of the i^{th} observation

x_{id}

scalar of the i^{th} observation and the d^{th} variable of $\mathbf{X}$

$\mathcal{Y}$

space of response

Y

random variable from $\mathcal{Y}$

y

realization of Y

$\mathbf{y}$

observed response vector in $\mathbb{R}^n$

y_i

scalar of the i^{th} observation of $\mathbf{y}$

η_i

linear predictor of i^{th} observation

$f : X \to Y$

some fixed but unknown function of $\mathbf{X}$

$s(x)$

smooth monotonic function of x

$||.||_1$

ℓ_1 norm

$||.||_2^2$

ℓ_2 squared norm

ξ

knot

K_d

number of knots of variable d

B(x)

B-spline basis functions of x

m

polynomial order

$\mathbf{B}$

B-spline basis function matrix in $\mathbb{R}^{n\times(k+m+1)}$

$N(x)$

natural spline basis functions of x

$\mathbf{N}$

natural splines basis function matrix in $\mathbb{R}^{n \times n}$

$\mathbf{W}$

penalty matrix of natural splines in $\mathbb{R}^{n \times n}$

Ω

training set to classify

S

number of levels in the hierarchical tree of Ω

s

index of the levels of the hierarchy in $[1, ..., S]$

h_s

height of the s^{th} level of the hierarchy

$\mathcal{G}$

group partition of all S levels of the hierarchy

G_s

number of groups at s^{th} level of the hierarchy

$\hat{G}_s^*$

optimal number of groups

g

index of number of groups in $[1, ..., G_s]$

$\mathcal{G}^s$

a partition of Ω at the s^{th} level

$\mathcal{G}_g^s$

g^{th} group of variables at the s^{th} level

ρ_s

weights attributed to group partition $\mathcal{G}^s$

$\tilde{\mathbf{X}}^{(s)}$

matrix of supervariables at s^{th} level of the hierarchy

$\tilde{\mathbf{X}}_{\mathcal{G}^s}^{(s)}$

matrix of supervariables for partition $\mathcal{G}^s$

$\mathbf{X}_{\mathcal{G}}$

variables matrix of concatenated group partition

$\mathbf{X}^{s}_{\mathcal{G}^s}$

variables matrix of group partition $\mathcal{G}^s$

G

genomic view

G

matrix of genotype data

Z

matrix of additively coded SNPs

M

metagenomic view

M

matrix of metagenomic data

$\mathcal{M}$

group structure of metagenomic data

Δ_{GM}

matrix of interaction terms between genome and metagenome data

Abbreviations

Abbreviations

Explanation

BH

Benjamini-Hochberg

CA

Cochran-Armitage

CNP

Copy Number Polymorphism(s)

ddNTP

dideoxyNucleotide TriPhosphate

DNA

DeoxyriboNucleic Acid

FDP

False Discovery Proportion

FDR

False Discovery Rate

FN

False Negative

FP

False Positive

FWER

Family-Wise Error Rate

GAM

Generalized Additive Model

GCV

General Cross-Validation

GL

Group-Lasso

GLM

Generalized Linear Model

GLMM

Generalized Linear Mixed Model

GWAS

Genome-Wide Association Study(ies)

HAC

Hierarchical Agglomerative Clustering

HCAR

Hierarchical Clustering and Averaging Regression

HGAM

High-dimensional Generalized Additive Models

HLA

Human Leukocyte Antigen

HWE

Hardy-Weinberg Equilibrium

LARS

Least Angle Regression

LASSO

Least Absolute Shrinkage and Selection Operator

LD

Linkage Disequilibrium

MCMC

Markov Chain Monte Carlo

MLGL

Multi-Layer Group Lasso

MSE

Mean Squared Error

MWAS

Metagenome-Wide Association Study(ies)

OLS

Ordinary Least Squares

OR

Odds Ratio

OTU

Operational Taxonomic Unit(s)

PCA

Principal Component Analysis

PCR

Polymerase Chain Reaction

P-IRLS

Penalized-Iteratively Reweighted Least Squares

RFLP

Restriction Fragment Length Polymorphism(s)

RNA

RiboNucleic Acid

ROC

Receiver Operating Characteristic

RR

Relative Risk

RSS

Residual Sum of Squares

SASA

Single Aggregated-SNP analysis

SIComORe

Selection of Interaction effects in COmpressed Multiple Omics Representation

SKAT

Sequence Kernel Association Test

SMA

Single Marker Analysis

SNP

Single Nucleotide Polymorphism(s)

SSLP

Simple Sequence Length Polymorphism(s)

TN

True Negative

TP

True Positive

WGA

Whole-Genome Association

WTCCC

Wellcome Trust Case-Control Consortium

Chapter 1

Basic concepts of molecular genetics

The purpose of this chapter is to provide the basic concepts in genetics necessary to the understanding of the Genome-Wide Associations Studies. The first section focuses on the description of the genome, sequencing analysis and introduces the notion of genetic mapping. The second section brings some concepts in population genetics necessary to a good understanding of linkage disequilibrium and association studies. The last section gives the definition of linkage disequilibrium and its origins. This section also explains the notion of haplotype structure, which is a key feature of the human genome that we leveraged with the methodology described in Chapter **??**.

1.1 Genome description

The common point to all organisms is to own a genome containing the biological information necessary to their construction, maintenance and survival. Most genomes are made of DNA (deoxyribonucleic acid), with the exception of viruses that have an RNA (ribonucleic acid) genome. DNA and RNA are both polymeric molecules composed of chains of monomeric subunits called *nucleotides*.

The human genome, which is representative of the genomes of all multicellular animals, consists of two parts:

- The nuclear genome including about 3.2×10^9 nucleotides of DNA divided into 24 linear molecules, the chromosomes. These 24 chromosomes consist of 22 autosomes and two sex chromosomes, X and Y.
- The mitochondrial genome is a circular DNA molecule present in multiple copies in the organelles called *mitochondria*. The human mitochondrial

genome contains 37 genes.

In the animalia taxon, the vast majority of cells are diploid which means that each autosome are present in two copies plus two sex chromosomes, XX for females and XY for males. These cells are known as somatic cells in contrast to sex cells, or gametes, which are haploid and possess only one copy of each chromosome. The use of the biological information contained in the DNA requires the coordinated action of several proteins participating in a series of complex biochemical reactions referred to as genome expression. The direct product of genome expression is the transcriptome, a collection of RNA molecules derived from the protein-coding genes. The transcriptome is maintained by the process of transcription, in which individual genes are copied into RNA molecules. The indirect product of genome expression is the proteome, the cell's collection of proteins. The proteins constituting the proteome are synthesized by translation of the individual RNA molecules present in the transcriptome.

DNA is a polymer, a polynucleotide, in which the monomeric subunits are four chemically distinct nucleotides linked together in chains that can reach length of thousands, even millions of units in length (Figure **??**). Each nucleotide in a DNA polymer is made up of three components: a deoxyribose, which is a pentose, a nitrogenous base (cytosine, thymine, adenine or guanine) and a phosphate group. A molecule made up of just the pentose and base is called a *nucleoside* and adding a phosphate group converts it into a nucleotide.

What makes DNA such a unique molecule is its famous double-helix structure discovered by (**?**). The key feature of the double-helix structure that convinced biologists that genes are made of DNA is the constrained base pairing between the nucleotides. Indeed, the limitation that adenosine can only be paired with thymine, and guanine with cytosine, means that DNA replication can result in perfect copies of a parent molecule simply by using the sequences of the pre-existing strands to build the sequences of the new strands.

1.2 Genome sequencing

1.2.1 DNA sequencing

Several methods for DNA sequencing exist, among them the chain termination method first developed by (**?**) is the most popular but alternative techniques such as chemical degradation sequencing (**?**) and pyrosequencing (**?**) are also used.

Chain termination method is based on the principle that single-stranded DNA molecules that differ in length by just a single nucleotide can be separated by polyacrylamide gel electrophoresis[1]. This procedure is illustrated and explained

[1]Polyacrylamide gel electrophoresis (PAGE) is a technique widely used in genetics to separate biological macromolecules, sush as nucleic acids, according to their electrophoretic mo-

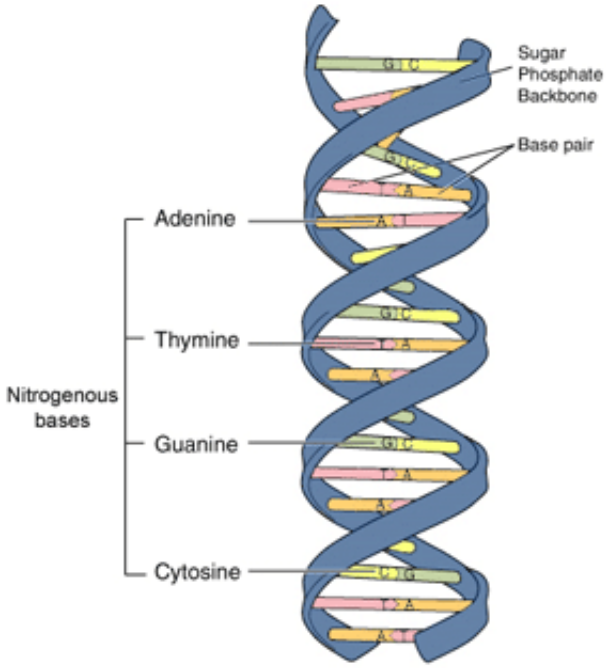


Figure 1.1: Double-Helix structure of DNA molecule. ©University of Leicester / Licence Creative commons

in Figure **??**.

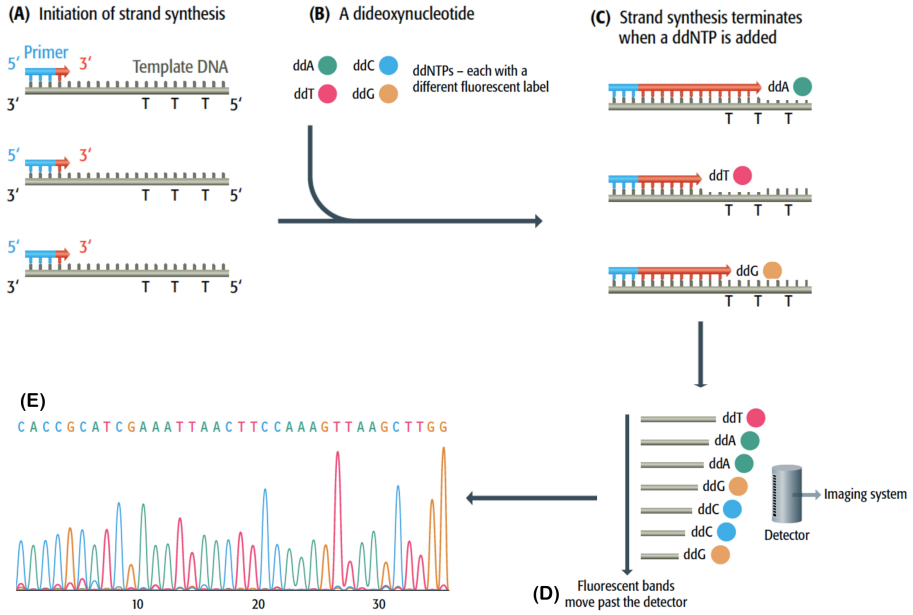


Figure 1.2: Chain termination DNA sequencing (**?**). (A) Use of universal primers for the synthesis of DNA complementary of a single-stranded template. (B) Incorporation of small amount of fluorescent dideoxynucleotides (ddATP, ddTTp, ddCTP and ddGTP), each with a different fluorescent label. (C) The ddNTP block the synthesis of DNA because they have a hydrogen atom rather than a hydroxyl group attached to the 3' carbon. (D) Each labelled DNA strand passes through a polyacrylamide gel electrophoresis, migrating more or less according to their length, and after separation a fluorescent detector is capable of discriminating the labels attached to the ddNTP. (E) The information is passed to the imaging system and a sequence of DNA is printed out. The sequence is represented by a series of peaks, one for each nucleotide position.

Pyrosequencing is a method generally used for the rapid determination of very short sequence of DNA and does not required electrophoresis or any fragment separation procedure as with chemical degradation sequencing. Since it can only generate a few tens of base pairs per experiment, it is used when many short sequences must be generated as fast as possible, for instance in single-nucleotide polymorphism typing. With this technique, the template is copied in a straightforward manner without added ddNTP and, as the new strand is being made, the order in which the deoxynucleotide are incorporated can be followed (see Figure **??** for more details).

---

bility.

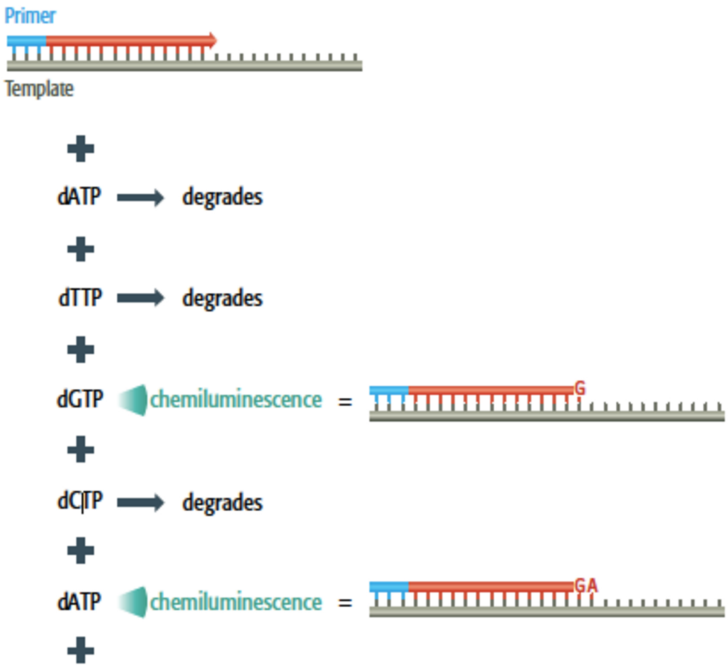


Figure 1.3: Pyrosequencing (**?**). Each deoxynucleotide is added individually, along with a nucleotidase enzyme that degrades the deoxynucleotide if it is not synthesized. The incorporation is detected by a flash of chemiluminescence induced by the pyrophosphate released from the deoxynucleotide. The order in which the deoxynucleotide are added to the growing strand can therefore be followed.

1.2.2 Sequence assembly

One of the main challenges in genome sequencing is to master the assembly of the multitude of short sequences generated by DNA sequencing techniques in order to reconstruct the complete continuous sequence of chromosome that can reach a length of several tens of megabases. The most straightforward method to sequence assembly is to build up the master sequence by directly searching for overlaps between all the short sequences. This method is known as the shotgun method (**?**). The shotgun method is the standard approach for sequencing small prokaryotic[2] genome but it is not suited to the analysis of larger genome because the required data analysis becomes too complex as the number of fragment increases (for n fragments, the number of possible overlaps is $2n^2 - 2n$). Moreover it can lead to errors when repetitive regions of a genome are analysed because when a repetitive sequence is broken into fragments, many of the resulting pieces contain the same sequence motifs.

To overcome these issues, techniques that make use of a genome map to guide the assembly are used, namely the whole-shotgun method and clone contig method (Figure **??**):

- **Whole-genome shotgun method.** This method takes the same approach as the standard shotgun procedure but uses the distinctive features on the genome map as landmark to assemble the whole sequence. Reference to the map ensures that regions containing repetitive DNA are assembled correctly.
- **Clone contig method.** In this method the genome is broken into manageable segments which are short enough to be assembled accurately by the shotgun method. Once the sequence of a segment has been completed, it is positioned at its correct location on the map

1.3 DNA polymorphism

1.3.1 Restriction Fragment Length Polymorphisms (RFLP)

RLFP are detected using a certain type of enzymes that cut DNA (restriction enzymes) at specific restriction site. Some restriction sites are polymorphic with one allele displaying the correct sequence for the restriction site while the second allele have an altered sequence so the restriction site is no longer recognized by the enzyme. The consequence is that the two adjacent restriction fragments remain linked together after treatment with the enzyme, leading to a polymorphism known as RFLP. The RFLP markers can be detected using

[2] A prokaryote is a unicellular organism that lacks a membrane-bound nucleus, mitochondria, or any other membrane-bound organelle.

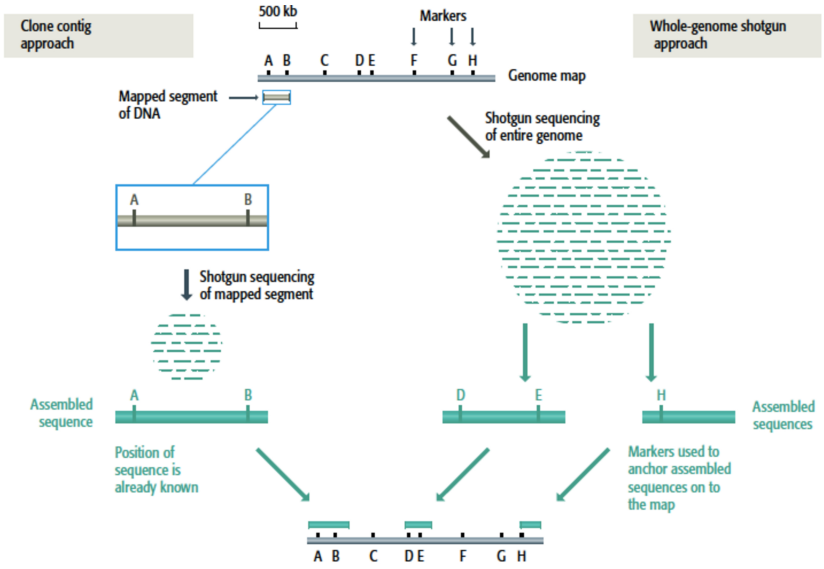


Figure 1.4: Clone contig and whole-genome shotgun for sequence assembly (**?**). To illustrate both techniques, a genome map of linear DNA molecule of 2.5 Mb has been represented together with the location of 8 known markers(A-H). On the left, the clone contig approach starts with a segment of DNA whose position on the genome is known since it contains the markers A and B. The segment is sequenced by the shotgun method and the master sequence placed at its known position on the map. On the right, the whole-genome shotgun method involves random sequence of the entire genome resulting in pieces of contiguous sequence. If a contiguous sequence contains a marker then it can be positioned on the map.

molecular biology techniques such as southern hybridization or polymerase chain reaction (PCR) (see (**?**) for more detail).

1.3.2 Simple Sequence Length Polymorphisms (SSLP)

SSLP are repeated nucleotidic sequences displaying different numbers of repeat units in each allele. There are two types of SSLP: minisatellites with repeat unit up to 25 base pair[3] (bp) in length and microsatellites with shorter repeated sequences (13 bp or less). Microsatellites are more commonly used than minisatellites because they are more frequent and evenly spread on the genome (5×10^5 with repeat units of 6 bp or less in the human genome). Furthermore, the PCR used to type a length polymorphism is more efficient and accurate with sequences less than 300 bp in length.

1.3.3 Single Nucleotide Polymorphisms (SNP)

A single nucleotide polymorphism is a variation in a single nucleotide that occurs at a specific position in the genome (see Figure **??**). In a given population, most individuals may have a specific nucleotide at one position (e.g., a C) but a minority of individuals could have a different nucleotide at the same position (e.g., a G). The two possible nucleotide variations at a particular genomic position (locus) are said to be alleles, this type of polymorphism is extremely frequent in the human genome (a few millions).

The vast majority of SNP are biallelic because they originate from a point mutation in the genome, converting a nucleotide into another. For an SNP to be more than biallelic, it would be necessary for a new mutation to appear, after the first has been fixed in the population, to exactly the same position in the genome, which is highly unlikely. SNP typing methods are based on oligonucleotide hybridization analysis where an oligonucleotide (short single-stranded DNA molecule) will hybridize with another DNA molecule only if the oligonucleotide forms a completely base-paired structure with the other molecule (under precise temperature conditions).

Oligonucleotide hybridization can discriminate between the two alleles of an SNP if there is at least one mismatch at one position between the oligonucleotide and the target DNA. Several screening methods based on oligonucleotide hybridization exists: DNA chip (microarray) which use fluorescent markers to detect hybridization, oligonucleotide ligation assay (OLA) using capillary electrophoresis and amplification refractory mutation system (ARMS test) based on PCR primers and electrophoresis.

Recent breakthroughs in microarray technology have meant that hundreds of thousands of SNP can now be densely genotyped at moderate cost. As a result,

[3] A base pair (bp) is a unit consisting of two nucleobases bound to each other by hydrogen bonds.

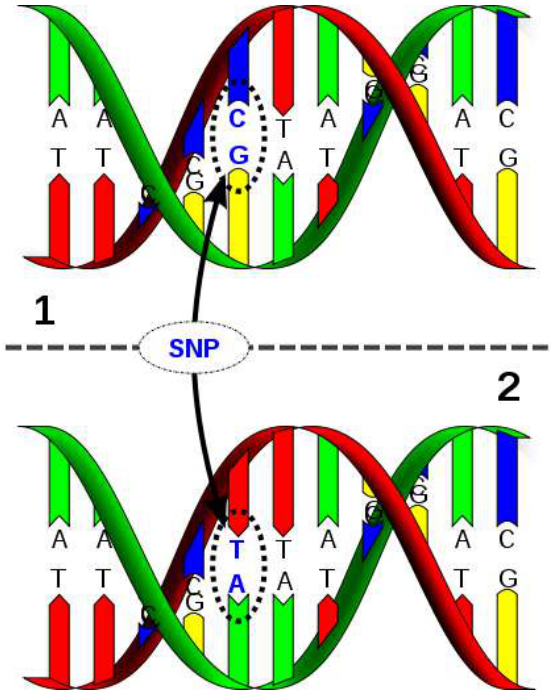


Figure 1.5: Schematic representation of a single nucleotide polymorphism. ©David Hall / Licence *Creative Commons*

it has become possible to characterize the genome of an individual with up to a million genetic markers. DNA chip technology makes use of piece of glass, or silicon, carrying many different oligonucleotides in a high-density array (Figure **??**). To prepare really high-density arrays, oligonucleotides are synthesized *in situ* on the surface of the piece of glass resulting in a DNA chip. A density of up to 300,000 oligonucleotides per cm^2 is possible and 150,000 polymorphisms can be typed in a single experiment (**?**).

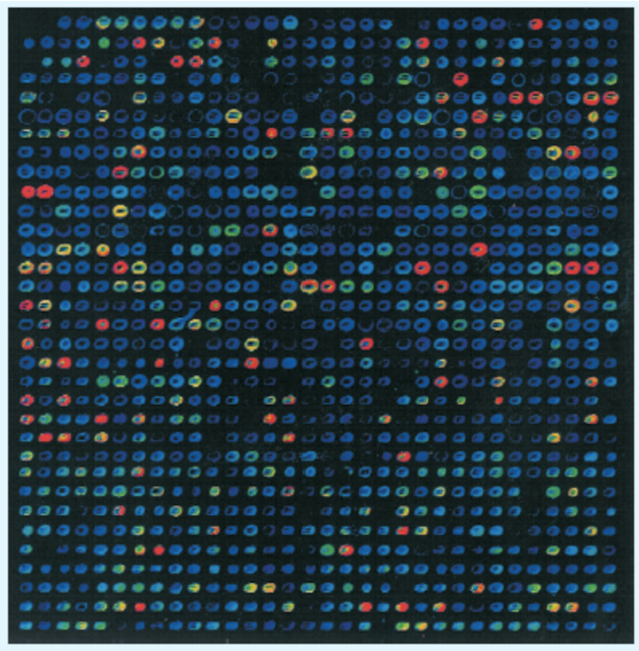


Figure 1.6: Visualization of the hybridization of a fluorescent labelled probe to a microarray. The DNA to be tested is labelled with a fluorescent marker and put onto the surface of the microarray. Hybridization is detected by examining with a fluorescence microscope the position at which the fluorescent signal is emitted indicating which oligonucleotides have hybridized with the target DNA (**?**).

Today, DNA microarrays are used in clinical diagnostic tests for some diseases. With the advent of new DNA sequencing technologies, some of the tests for which microarrays were used in the past now use DNA sequencing instead. Nevertheless, microarray tests being less expensive than sequencing, they remain used for very large studies as well as for some clinical tests.

1.4 Linkage and partial linkage for genetic mapping

Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequences features on a genome. Historically, the first markers used to construct genetic maps were genes coding for mendelian traits (qualitative traits that are highly heritable) with distinguishable phenotypes for each allele (see (**?**) for more details on early gene mapping works). Although genes are useful markers, genetic maps based only on them are not precise in large genomes due to the gaps existing between successive coding region. Furthermore, only a part of the genes exist in allelic forms that can be distinguished conventionally. That is why DNA markers having at least 2 alleles are preferable, i.e. RFLP, SSLP or SNP previously described.

Genetic mapping makes use of the principle of inheritance at first described by Gregor Mendel (**?**) and the resulting genetic linkage properties to estimate the relative position of each DNA markers on a chromosome. The principle of genetic linkage arises from the fact that, while chromosomes are inherited as intact units, the alleles of some pairs of genes located on the same chromosome should also be inherited together. However, this principle, deriving from the Second Law of Mendel which states that pairs of alleles segregate independently is not what we observe in reality. Indeed, genetically linked genes are sometimes inherited together and sometimes are not, resulting in what we call partial linkage.

This partial linkage property is explained by the behaviour of chromosomes during meiosis, where homologous chromosomes can undergo physical breakage and exchange fragment of DNA in a process called *crossing-over* (or recombination). These recombination events explain why linked genes and therefore linked DNA markers are sometimes not inherited together. This allows to develop a way to map the relative position of DNA markers since markers which are close together will be separated less frequently than two markers that are far away. Furthermore, the frequency with which markers on a same chromosome are unlinked by crossovers will be directly proportional to the distance between them. The recombination frequency is therefore a measure of the distance between two markers and if we estimate the frequencies for several pairs of markers, we can construct a map of their relative positions.

Comparisons between genetic maps and the actual positions of genes on DNA molecules, as revealed by DNA sequencing, have shown that some regions of chromosomes, called recombination *hotspots*, are more likely to be involved in crossovers than others. This results in shared chromosomic region among individuals of the same population, although each individual has a unique DNA sequence, and is known as haplotype structure. We will see in Section **??** that different populations have their own haplotypic structure.

Table 1.1: Expected frequency of an allele transmitted from an individual sampled from a population in Hardy-Weinberg Equilibrium

Parental Genotype	Genotype probability	Probability of transmitting A	Joint probability
AA	p^2	0	0
Aa	$2p(1-p)$	0.5	$p(1-p)$
aa	$(1-p)^2$	1	$(1-p)^2$
Total	1	-	p

1.5 Basic concepts in population genetics

1.5.1 Hardy-Weinberg equilibrium in large population

We consider a biallelic locus with alleles A and a present in a population at frequencies p and q respectively. If we assume that the two copies of the gene that an individual carries are inherited independently, then the number of copies of the allele A will follow a binomial distribution, $\mathcal{B}(2, p)$, that is, that the probabilities of the three possible genotypes (aa, aA and AA) will follow the Hardy-Weinberg law (**?**):

$$p^2 + 2pq + q^2 = 1$$

with

$$p^2 = p(AA); q^2 = p(aa) \text{ and } 2pq = p(Aa).$$

Hardy-Weinberg's law states that in an isolated population of unlimited size, not subject to selection, and in which there are no mutations, the allelic frequencies remain constant. If the couplings are panmictic (random mating), the genotypic frequencies are deduced directly from the allelic frequencies and also remain constant. The assumption of random mating says that the probability that any pair of individual mates is unrelated to their genotype (except for the X chromosome) or their ethnic origin. However, in practice it is not truly the case since couples tend to mate within their ethnic group and are likely to select partners with compatible traits, some of which may be influenced by specific genes. Such non-random mating is commonly ignored in many genetic analyses of chronic disease traits, for which its effect may be negligible. Nevertheless, to the extent that it occurs, its major effect is to slow down the rate of convergence to Hardy-Weinberg equilibrium rather than to distort the equilibrium distribution (**?**).

1.5.2 Genetic drift in small population

In small populations, the results presented above will still be true in expectation, but the allele frequencies will vary from generation to generation simply as a result of chance (sampling error). It follows that, in finite populations, the expected value of the allele frequency will remain constant but its variance will increase from one generation to the next. This means that in generation, there is a non-zero probability that one allele might not be transmitted to any offspring, in which case that allele becomes extinct and the other becomes fixed. In fact, with absence of mutation and selection, one of the alleles will eventually become extinct, and the probability that it is the allele a that disappears turns out to be simply $1 - q$.

At first glance, this might seem to contradict the claim that in expectation, the allele frequency remains constant, but in fact with probability q the allele frequency will eventually become 1 and with probability $1 - q$ it will become 0; hence in expectation, the allele frequency remains $q \times 1 + (1 - q) \times 0 = q$. This phenomenon is known as genetic drift and was first introduced by Sewall Wright, one of the founders in the field of population genetics, (**?**).

Figure **??** illustrates the effect of genetic drift on allelic frequencies for 2 alleles A and a at 1 locus for different population neither subject to selection nor mutation.

1.5.3 Concept of heritability

Sewall Wright and Ronald Fisher first introduced the concept of heritability in the context of family studies. Wright's heritability is based on the analysis of correlation and its estimate is based on the path analysis method (**?**) while the definition of Fisher is based on the analysis of variance and is defined as the proportion of total variance in a population for a particular measurement, taken at a particular time or age, that is attributable to variation in additive genetic or total genetic values (**?**).

An observed phenotype for a trait of interest can be partitioned into a statistical model representing the contribution of the unobserved genotype and unobserved environmental factors:

$$\text{Phenotype} = \text{Genotype} + \text{Environment}.$$

The variance of the observed phenotype (σ_P^2) can thus be partitioned into the sum of unobserved genotype and environmental variances (σ_G^2 and σ_E^2):

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2.$$

Following the definition of Fisher, the broad-sense heritability (H^2) can be expressed as a ratio of variances by expressing the proportion of the phenotypic

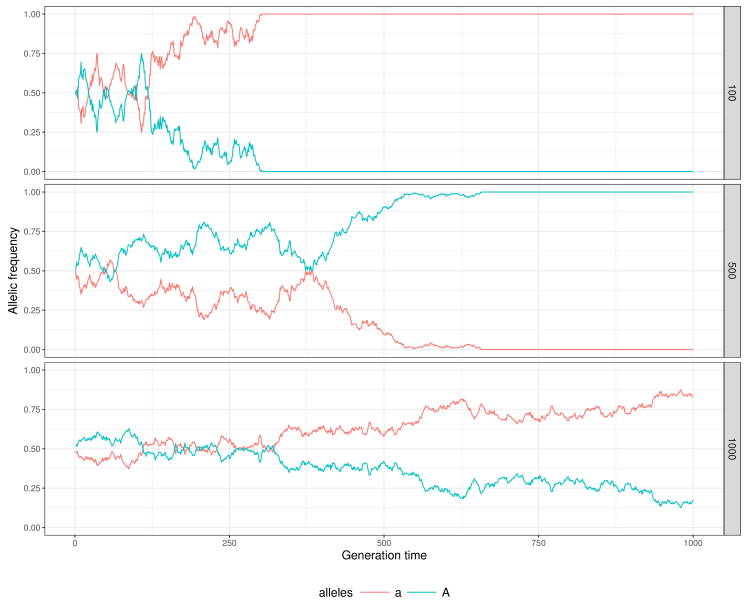


Figure 1.7: Illustration of genetic drift in finite, small, populations. The plots show the evolution of allelic frequencies for 2 alleles A and a at 1 locus over 1000 generations for 3 population sizes (100, 500 and 1000). The 2 alleles are set to have the same proportions in the 3 populations at generation 1 ($p = q = 1/2$) and the frequencies of both allele evolve to be either fixed or extinct more or less quickly depending on the size of the population.

variance that can be attributed to variance of genotypic values:

$$\mathrm{H}^2 = \frac{\sigma_G^2}{\sigma_P^2}.$$

The genetic variance σ_G^2 can further be partitioned into additive genetic effects (σ_A^2), dominance genetic effects (σ_D^2) and epistatic genetic effects (σ_I^2) and the narrow-sense heritability (h^2) is defined as:

$$\mathrm{h}^2 = \frac{\sigma_A^2}{\sigma_P^2}.$$

Heritabilities can be estimated from empirical data of the observed and expected resemblance between relatives. The expected resemblance between relatives depends on assumptions regarding its underlying environmental and genetic causes (**?**). To estimate the heritability from population sample rather from family studies, we can resort on the use of the generalized linear mixed model (GLMM[4]), this heritability is known as the genomic heritability (**?**).

1.6 Linkage disequilibrium

Every human genome has a unique DNA sequence, in part due to the few hundred novel mutations inherited from their parents, and by chromosomal segregation combined with crossovers that shuffle existing variation. However, although every human genome may be unique, certain combination of variants (e.g. SNP) may be shared by few individuals and sometimes by a large fraction of the population, resulting in allelic association also known as haplotype structure. The term linkage disequilibrium (LD) is broadly used to refer to the non-random association of combination of variants, therefore LD neither requires genetic linkage nor is particularly a disequilibrium.

Particular alleles at neighbouring loci tend to be co-inherited. For tightly linked loci, this might lead to associations between alleles in the population resulting in high LD between these loci. LD has recently become the focus of intense study in the hope that it might facilitate the mapping of complex disease loci through whole-genome association studies. This approach depends crucially on the patterns of LD in the human genome (**?**).

1.6.1 Definition

We consider two neighbouring biallelic loci A and B (A and a for locus A, and B and b for locus B) with allele frequencies f_A, f_a, f_B and f_b respectively.

[4]The generalized linear mixed model is an extension of the generalized linear model (**?**) in which the linear predictor contains random effects in addition to the usual fixed effects (see (**?**) for thorough introduction to GLMMs).

Table 1.2: Haplotype frequencies under linkage equilibrium

Haplotype	Expected frequency
AB	$f_{A} \times f_{B}$
Ab	$f_{A} \times f_{b}$
aB	$f_{a} \times f_{B}$
ab	$f_{a} \times f_{b}$

Table 1.3: Haplotype frequencies under linkage disequilibrium

Haplotype	Observed frequency	Positive LD	Negative
AB	$\hat{f}_{AB}$	$\hat{f}_{AB} > \hat{f}_{A} \times \hat{f}_{B}$	$\hat{f}_
Ab	$\hat{f}_{Ab}$	$\hat{f}_{Ab} > \hat{f}_{A} \times \hat{f}_{b}$	$\hat{f}_
aB	$\hat{f}_{aB}$	$\hat{f}_{aB} > \hat{f}_{a} \times \hat{f}_{B}$	$\hat{f}_
ab	$\hat{f}_{ab}$	$\hat{f}_{ab} > \hat{f}_{a} \times \hat{f}_{b}$	$\hat{f}_

Under linkage equilibrium, the four haplotypes formed by these loci have the frequencies shown in Table **??**. These frequencies are equal to the product of the component allele frequencies. These equalities are valid only when the alleles are independent, i.e. when the two loci are not genetically linked.

However, when the two loci are in linkage disequilibrium, the haplotypes are not observed at the frequencies expected if the alleles were independent.

Positive linkage disequilibrium exists when two alleles occur together on the same haplotype more often than expected, and negative LD exists when alleles occur together on the same haplotype less often than expected (Table **??**).

1.6.2 Measure of LD

The deviation of the observed from expected haplotype frequencies can be quantified by several linkage disequilibrium measures. The very first linkage disequilibrium measure was introduced by (**?**) and is defined as:

$$\begin{aligned} D_{AB} &= f_{AB} - f_A f_B \\ &= f_{AB} f_{ab} - f_{Ab} f_{aB}, \end{aligned} \tag{1.1}$$

where f_{AB} is the observed frequency of haplotypes carrying the A and B alleles and f_A, f_B are the marginal allele frequencies of alleles A and B. Any deviation from this expectation results in a non-zero value for D_{AB}, with a positive value indicating that the AB haplotype is found more often than expected assuming independence and a negative value indicating that it is found less frequently than expected.

Although this measure is easy to calculate, it has for disadvantage to be sensitive to allele frequencies at the extreme values of 0 to 1. Indeed, if we let D_{AB} be the population coefficient for LD, then the sample coefficient $\hat{D}_{AB}$ has the following properties (**?**):

$$D_{AB} = \hat{f}_{AB} - \hat{f}_A \hat{f}_B,$$
$$\mathbb{E}(\hat{D}_{AB}) = \frac{(n-1)}{n} D_{AB},$$
$$textVar(\hat{D}_{AB}) = \frac{1}{n}[f_A f_a f_B f_b + (f_A - f_a)(f_B - f_b)D_{AB} - D_{AB}^2], \quad (1.2)$$

Here $\hat{f}_{AB}$ is the estimate of f_{AB} (the population frequency) from the sample and is given, by n_{AB}/n where n_{AB} is the number of haplotype AB in the sample. Equation shows that the variance in the estimate is strongly influence by the allele frequencies at the two loci as for the range of values that $\hat{D}_{AB}$ can take. If we arbitrarily define A and B as minor allele at each locus and enforce $\hat{f}_B \leq \hat{f}_A$, then it follows that

$$-\hat{f}_A \hat{f}_B \leq \hat{D}_{AB} \leq \hat{f}_a \hat{f}_b.$$

The strong dependency on allele frequency of the standard measure of LD is an undesirable property because it makes comparison between pairs of alleles with different allele frequencies difficult. That is why methods less sensitive to marginal allele frequencies have been developed (**?**).

(**?**) suggested another measure

$$D' = \frac{D}{D_{max}},$$

where D_{max} is the theoretical maximum LD value for the observed allele frequencies.

D' thus ranges from -1 to 1 and reflects both positive and negative linkage disequilibrium. We can also use the absolute value of D' to measure the evidence of recombination between two loci.

$$|D'| = \begin{cases} \dfrac{-\hat{D}_{AB}}{\min(\hat{f}_A \hat{f}_b, \hat{f}_a \hat{f}_B)} & \hat{D}_{AB} < 0 \\ \dfrac{\hat{D}_{AB}}{\min(\hat{f}_A \hat{f}_b, \hat{f}_a \hat{f}_B)} & \hat{D}_{AB} > 0 \end{cases}$$

The greater the rate of recombination between loci, the more likely the alleles are to be in linkage equilibrium so a value of $|D'| = 1$ can be interpreted as evidence of no recombination while a value close to 0 can be viewed as evidence for strong recombination. However, even if all four haplotypes are present in

the population, it may be unlikely that all four haplotypes are observed in a finite sample if at least one allele is very rare (**??**) leading to an interpretation of $|D'| = 1$ dependent on the sample allele frequencies.

Due to the sensitivity of measurements D and D' to allele frequencies, another measure of LD is more commonly used which is the r^2 measure (**?**). If we assign an allelic value, X_A, to locus A as $X_A = 1$ for allele A and $X_A = 0$ for allele a and we assign an allelic value, X_B, to locus B with the same properties, then the quantity measured by (

$$eq : 1$$

) can be interpreted as the covariance in allelic value between the 2 loci. One way to transform the covariance is to measure the squared Pearson correlation coefficient:

$$r^2_{AB} = \frac{\mathrm{Cov}(X_A, X_B)^2}{\mathrm{Var}(X_A)\mathrm{Var}(X_B)} = \frac{D^2_{AB}}{f_A f_a f_B f_b}. \tag{1.3}$$

The r^2 measure has for advantage to be insensitive to how the two loci are labelled, as indicated by the lack of subscripts for D in . Moreover there is a direct relationship between the sample estimate $\hat{r}^2$ and the power to detect significant association, i.e. to reject the null hypothesis $H_0 : D = 0$ (**??**). As a proof, we can consider the contingency table test where, under the null hypothesis, the test statistic:

$$X^2 = \sum_{ij} = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1.4}$$

is χ^2 distributed with 1 degree of freedoms the sample size tends to infinity. Here O_{ij} and E_{ij} are the observed and expected counts, respectively, of the ij haplotype. The relation between and r^2 is therefore

$$X^2 = n\hat{r}^2.$$

Consequently, the null hypothesis of no association can be rejected at a specified level α if $n\hat{r}^2$ is greater than the critical value of the test statistic.

1.6.3 Estimation of linkage disequilibrium

Estimation of linkage disequilibrium between alleles at two loci requires observations of haplotype frequencies which is usually not the case. Therefore, haplotype frequencies are often estimated using statistical tools such as the expectation maximization (EM) algorithm (**?**). These methods take as input the

observed combined genotype frequencies at the two loci (for example, the distribution of the nine possible combinations of AA, Aa, and aa, with BB, Bb, and bb).

An example of pairwise linkage disequilibrium (r^2) plot for three different populations (as given by the software Haploview (**?**)) is illustrated in Figure **??**.

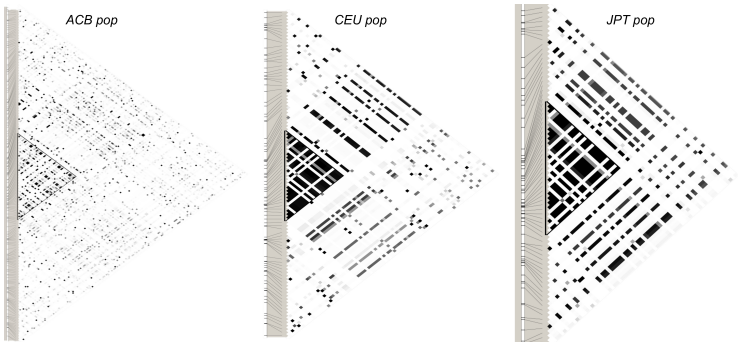


Figure 1.8: Plots of pairwise linkage disequilibrium for polymorphisms in the ACE (Angiotensin I Converting Enzyme) genomic region genotyped in three populations by the International HapMap Project. CEU, Utah residents with ancestry from northern and western Europe; ACB, African Caribbean in Barbados; JPT, Japanese in Tokyo; white, $r^2 = 0$; black, $r^2 = 1$; HapMap Release 22; chromosome 17 NCBI Build 37.

1.6.4 Origins of linkage disequilibrium

Founder mutations

Assume that a new mutation was introduced into the population at some point in the recent past. That mutation would have occurred on a single chromosome and would be transmitted with all alleles that are on the same chromosome, at least until recombination occurs. Thus, for many generations, the mutant allele would be associated with certain alleles at linked loci, and the strength of that association would diminish over time as a function of the recombination rate. If we look many generations later, the strength of the LD can be seen as an inverse measure of the distance between the loci. Of course, this presumes that the mutation was transmitted to an offspring and, through the process of genetic drift, expanded to sufficient prevalence to account for a significant burden of disease in present-day descendants of the affected founder.

Admixture

We consider a population that consists in a mixture of two subpopulations and two alleles, A and B, having the following frequencies $p_1 = q_1 = 0.9$ and $p_2 = q_2 = 0.1$. If the two loci were independently distributed within each subpopulation, then, in a 50-50 mixture of these two subpopulations, we would expect the following observed haplotypes distribution:

Through this example, we see an apparently very strong LD in the total population that is in fact spurious, leading to a complete artefact of population stratification. In statistical term, it is simply a reflection of Simpson's paradox (**?**) or confounding by ethnicity in epidemiologic term.

Others factors that influence LD

Mutation and recombination may have the most evident impact on linkage disequilibrium, but there exist other factors that influence the distribution of disequilibrium. Most of these involve demographic aspects of a population, and tend to sever the relationship between LD strength and the physical distance between loci:

- **Genetic drift:** Increased drift of small, stable populations tends to increase LD, as haplotypes are lost from the population.
- **Population growth:** Rapid population growth decreases LD by reducing genetic drift.
- **Gene flow:** LD can be created by gene flow (migration) between populations. Initially, LD is proportional to the allele frequency differences between the populations, and is unrelated to the distance between markers. In the next generations, the "artificial" LD between unlinked markers quickly fades, while LD between nearby markers is more slowly broken down by recombination.
- **Population structure:** Various aspects of population structure are thought to influence LD. Population subdivision is likely to have been an important factor in establishing the patterns of LD in humans (**?**).
- **Natural selection:** There are two principal ways by which selection can affect the level of LD. The first is an hitchhiking effect (genetic draft) (**?**), in which an entire haplotype that flanks an advantageous variant can highly increase in frequency or even be fixed. Although the effect is generally weak, selection against deleterious variants can also inflate LD, as the deleterious haplotypes are swept from the population (**?**). The second way in which selection can affect LD is through epistatic selection for combinations of alleles at two or more loci on the same chromosome. This form of selection leads to the association of particular alleles at different loci.

- **Variable recombination rate:** Recombination rates are known to vary by more than an order of magnitude across the genome. Because breakdown of LD is primarily driven by recombination, the extent of LD is expected to vary in inverse relation to the local recombination rate. It is even possible that recombination is largely confined to highly localized recombination hot spots, with little recombination elsewhere. According to this view, LD will be strong across the non-recombining regions and break down at hotspots.

1.7 Structure of haplotype blocks in the human genome

The distribution of linkage disequilibrium patterns along the genome can be seen as being noisy and unpredictable. For example, pairs of loci that are tens of kilobases apart might be in complete LD due to population structure or population size for instance, whereas nearby loci from the same region might be in weak LD if close to a recombination hotspot for instance (**?**).

It is often observed that LD in non-African populations extends over longer distances than in Africans, which might reflect a population bottleneck at the time when modern humans first left Africa (**??**). Similarly, there have been reports that certain isolated or admixed populations show LD over large distances (**??**).

However, despite the apparent complexity of observed patterns, some studies have proposed that the underlying structure of LD in the human genome can be described using a relatively simple framework in which the data are parsed into a series of discrete haplotype blocks (**??**), neighbouring blocks being separated by regions of numerous recombination events (**?**).

In response to these results, the United States National Human Genome Research Institute initiated a project, called the International HapMap Project, which aims to create a genome-wide map of LD and haplotype blocks. The HapMap Project seeks identify chromosomal regions where genetic variants are shared by comparing the DNA sequences among individuals. There are approximately ten million SNP estimated to be present in the human genome. Testing all of these SNP in chromosomes of individuals, however, can be extremely expensive and cost-inefficient. The development of the HapMap enables geneticists to take the advantage of how SNP and other genetic variants are organized on the same chromosome.

The number of tag SNP that capture most of the information of genetic variation patterns is estimated to be between 300,000 and 600,000, far fewer than the ten million common SNP.

Definition of haplotype blocks

A definition of haplotype block has been proposed by (**?**) where they focused on $|D'|$ measure of LD and defined haplotype blocks as sets of consecutive sites between which there is little or no evidence of historical recombination. For each pair of loci, the data are used to construct a confidence interval on the population value of $|D'|$ and the values of $|D'|$ are thus divided into three categories:

- **strong LD:** $|D'|$ near 1, which implies little or no evidence of historical recombination;
- **weak LD:** $|D'|$ significantly < 1, implying historical recombination;
- **intermediate/unknown LD:** The category includes pairs of sites with intermediate values of $|D'|$, as well as pairs for which the confidence intervals are relatively wide.

Two or more sites can be grouped together into a block if the outermost pair of sites is in strong LD, and if, for all pairwise comparisons in the block, the number of pairs in strong LD is at least 19-folds greater than the number of pairs in weak LD.

Patterns in human genome

To illustrate the patterns of LD in human genome, we refer to the results obtained by (**?**) where they characterized haplotype patterns across 51 autosomal regions (spanning 13 megabases of the human genome) in samples from Africa, Europe, and Asia and the analysis of these data by (**?**).

In Figure **??** are represented, for 4 different samples of population, the total proportions of sequence that was contained in haplotype blocks of various sizes. The results show that both the European-American and East Asian population samples have more extensive haplotype blocks than the African-American and sub-Saharan African samples and it is worth mentioning that in all four populations less than half of the total sequence is contained in identified haplotype blocks.

In Figure **??** are represented the values of $|D'|$ for all pairs of markers in a region. In this type of representation, the haplotype blocks appear as triangular regions of red (or light brown) squares that along the diagonal. These plots highlight the strong heterogeneity of LD within same regions: areas of strong LD that correspond well to the haplotype-block definition are often surrounded by equally large regions with little or no LD.

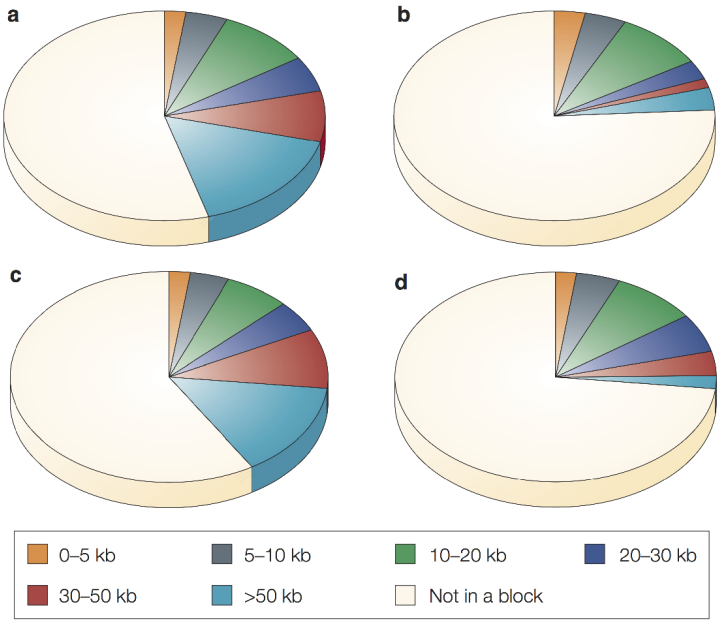


Figure 1.9: Proportion of sequence contained in haplotype blocks of various sizes from (**?**). (a) European-American sample; (b) African-American sample; (c) East Asian sample; (d) Sub-Saharan African sample.

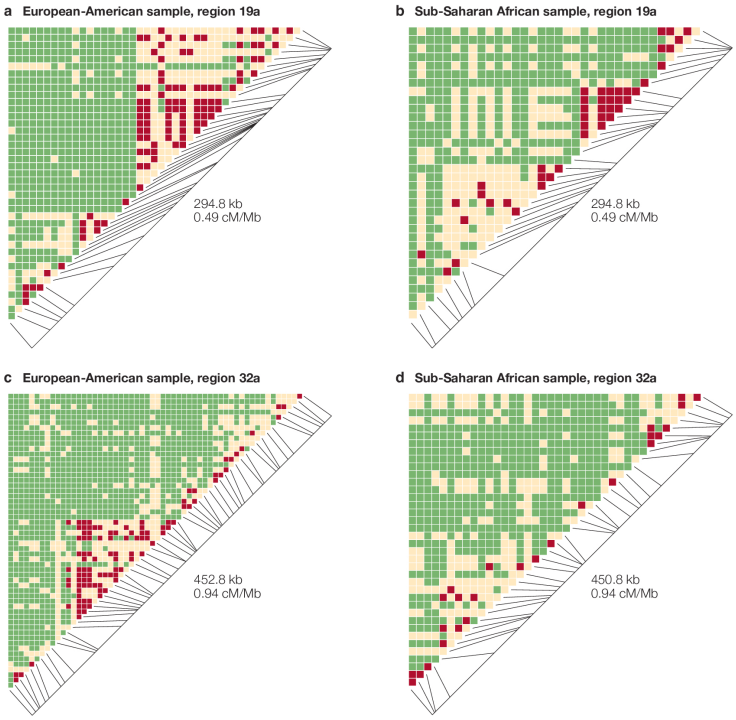


Figure 1.10: Pairwise $|\ D'\ |$ plots for representative regions from different population samples from (**?**). Each square in the triangle plots the level of linkage disequilibrium (LD) between a pair of sites in a region; comparisons between neighbouring sites lie along the diagonal. Red color indicates strong LD, green indicates weak LD and light brown indicates intermediate or uninformative LD. The long diagonal line indicates the physical length of the region, and the short black lines plot the position of each marker in this region.

Chapter 2

Statistical context

This chapter is intended to introduce statistical learning and hypothesis testing. We will present some state-of-the-art linear statistical methods and also a thorough introduction to splines and generalized additive models. The understanding of these methods is required to grasp the statistical concepts used in the methodology presented in Chapter **??** and **??**.

2.1 Notations

Let $\mathcal{T}$ be a training set consisting of n pairs of examples labelled on a space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$: $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{n}$, with $(x_i, y_i) \in \mathcal{Z}$, $\forall i$. Each couple (x_i, y_i) is the realization of an i^{th} independent copy of a random vector couple (X, Y) distributed according to $\mathcal{U}$, an unknown but fixed distribution on $\mathcal{Z}$.

Generally, we will represent the vectors in bold lowercase letters and the matrices in bold capital letters. Thus, when $\mathcal{X} \in \mathbb{R}^D$, the vector of the i^{th} component of $\mathcal{T}$ represented by D variables, will be designated by the column vector $x_i = (x_{i1}, ..., x_{id}, ..., x_{iD})^T \in \mathbb{R}^D$ and its associated matrix by $\mathbf{X} = (x_1^T, ..., x_i^T, ..., x_n^T)^T \in \mathbb{R}^{n \times D}$.

2.2 Concepts of statistical learning

Assuming that there is some relationship between an observed response vector $\mathbf{y} \in \mathbb{R}^n$ and D different predictors in $\mathbf{X} \in \mathbb{R}^{n \times D}$, we can write this relationship in the very general from

$$\mathbf{y} = f(\mathbf{X}) + \epsilon,$$

where $f : \mathcal{X} \to \mathcal{Y}$ is some fixed but unknown function of $\mathbf{X}$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$ is a random error term, independent of $\mathbf{X}$ and with $\mathbf{I} \in \mathbb{R}^{n \times D}$ being the identity matrix.

By definition, statistical learning refers to a set of approaches designed to estimate f for 2 main reasons: prediction and explanation.

2.2.1 Prediction

In the setting where we have a set of input variables $\mathbf{X}$ easily observable but where the output response $\mathbf{y}$ cannot be readily obtained, then, since the error term averages to zero, $\mathbf{y}$ can be predicted using

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}),$$

where $\hat{f}$ is the estimate of f and $\hat{\mathbf{y}}$ is the resulting prediction for $\mathbf{y}$. In this configuration, we are not especially concerned with the exact form of $\hat{f}$ as long as it yields to an accurate prediction for $\mathbf{y}$. In this context, we want to find a function $\hat{f}$ that approximate the true function f as well as possible by means of statistical learning method. We will make "as well as possible" in the sense of minimizing a particular cost function which must reflect how accurate we are in predicting $\mathbf{y}$. The most commonly used cost function in statistical regression is the mean-squared error (MSE) defined as:

$$\|\mathbf{y} - \hat{f}(\mathbf{X})\|_2^2$$

Most of statistical learning methods aim to minimize the MSE (also known as the quadratic loss function) to estimate $\hat{f}$ but other cost functions are also used in machine learning, depending on the task considered such as classification, regression or ranking (see for example the log loss, relative entropy, hinge loss, mean absolute error).

Bias-variance decomposition of mean squared error

Considering a couple of random variables (X, Y) defined on a training set $\mathcal{T}$, then it can be shown that the expected mean squared error $\mathbb{E}_{\mathcal{T}}[(\mathrm{Y} - \hat{f}(\mathrm{X}))^2]$, conditionally to $\mathcal{T}$ and a noise term ϵ, can be parsed into two errors terms, bias and variance:

$$\mathbb{E}_{\mathcal{T}}\{[\mathrm{Y} - \hat{f}(\mathrm{X})]^2\} = \underbrace{\mathbb{E}_{\mathcal{T}}[\hat{f}(\mathrm{X})^2] - \mathbb{E}_{\mathcal{T}}^2[\hat{f}(\mathrm{X})]}_{\mathrm{Variance}(\hat{f}(\mathrm{X}))} + \underbrace{(\mathbb{E}_{\mathcal{T}}[\hat{f}(\mathrm{X})] - \mathbb{E}[f(\mathrm{X})])^2}_{\mathrm{Bias}[\hat{f}(\mathrm{X})\}} + \underbrace{\sigma^2}_{\mathrm{Variance}(\epsilon)}.$$

Since the function $\hat{f}$ has been constructed on a training set $\mathcal{T}$, it is interesting to know how accurate it is on predicting $\mathbf{y}$ when applied to a new data set, this measure is represented by the variance term in the MSE. Estimates with high

variance will tend to perform poorly when seeing new data, in general more complex models tend to have a higher variance.

On the other hand, the bias error term refers to the error that is introduced by approximating the real, generally complex, function f. For instance, if we try to approximate a non-linear function using a learning method designed for linear models, there will be error in the estimate $\hat{f}$ due to this assumption. The more complex the model is, the lower the bias will be but at a cost of a higher variance.

That is why when we try to fit a model on some data, we always search for the best compromise between variance and bias (the so-called bias-variance trade-off) and therefore between complexity and simplicity. The fact that highly complex models have a lower bias but generalize poorly on new data is known as overfitting and occurs when the model fit too closely the data, going so far as to interpolate them in the most extreme case (see Figure **??** for an illustration of bias-variance trade-off).

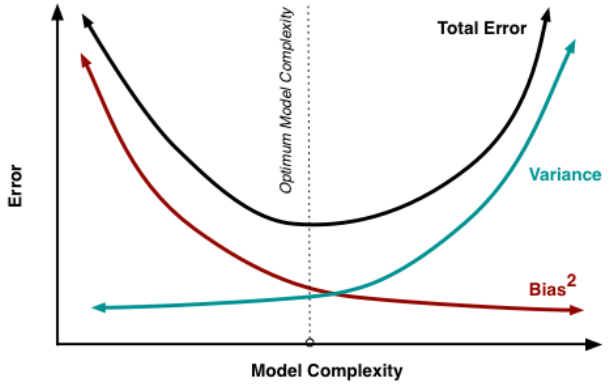


Figure 2.1: Bias and variance contribution to the total error. The bias (red curve) decreases as the model complexity increases unlike variance, which increases. The vertical dotted line shows the optimal model complexity, i.e. where the error criterion is minimized (image taken from http://scott.fortmann-roe.com/docs/BiasVariance.html).

The last term, σ^2, corresponds to the variance of the noise, also called the *irreducible error* because the response variable is also a function of ϵ which, by definition, cannot be predicted using the observations. Since all terms are non-negative, this error forms a lower bound on the expected error on unseen samples (**?**).

Explanation

We can also be interested in understanding the relationship between Y and X. In this situation we are more interested by the exact form of $\widehat{f}$ and we may want to answer the following questions: Which predictors are associated with the response? What is the relationship between the response and each predictor? Can the relationship be described using a linear equation or with a non-linear smoother? To answer these questions, we will tend to use more interpretable, i.e. simpler, models and to rely on the theory of hypothesis testing developed by (**?**). We will introduce the theory of hypothesis testing and the most common tests in Section **??**.

Estimation of f

All statistical learning methods can be roughly characterized as either parametric or non-parametric.

- **Parametric methods:** They are model-based approaches that reduce the problem of estimating f down to estimating a set of parameters. Assuming a parametric form for f simplifies the estimation problem because it is generally easier to estimate a set of parameters, as in the linear model, than to fit an entirely arbitrary function. The main drawback is that the chosen model is generally too far for the true form of f leading to a poor estimate. Even if more flexible models, such as polynomial models, can fit more closely the true form of f, they require in general to estimate a greater number of parameters which can lead to overfit the data, meaning that they follow the errors to closely and cannot be generalized to other data. We will present in Section **??** the linear model with its extension known as generalized linear models and some penalized approaches in Section **??**.

- **Non-parametric methods:** These approaches do not make explicit assumptions about the functional form of f but instead seek an estimate that fit closely the data to some degree to avoid overfitting. These methods have the advantage of being able to fit a wider range of form for f since no assumption about the functional form for f is made. However, they require a larger number of observations than is typically needed for a parametric approach to obtain an accurate estimate for f. We will present in Section **??** some non-parametric models such as the regression splines and the generalized additive model[1].

[1] To be more specific, we will present the semi-parametric forms of these models using a linear basis expansion.

2.3 Parametric methods

2.3.1 Linear models

Linear models are statistical models where a univariate response $\mathbf{y} \in \mathbb{R}^n$ is modelled as the sum of D linear predictor $\mathbf{X} \in \mathbb{R}^{n \times D}$ weighted by some unknown parameters, $\beta \in \mathbb{R}^D$, which have to be estimated, and a zero mean random error term, ϵ. A linear model is generally written in the following matrix form:

$$\mathbf{y} = \beta\mathbf{X} + \epsilon.$$

Statistical inference with such models is usually based on the assumption that the response variable has a normal distribution, i.e.

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2).$$

To estimate the unknown parameter, a sensible approach is to choose a value of β that makes the model fit closely the data. One possible way to proceed is to minimize a relevant cost function, defined by the residual sum of squares (RSS) of the model, with respect to β, known as the *least squares* method (**?**):

$$RSS(\beta) = ||\mathbf{y} - \mathbf{X}\beta\ ||_2^2.$$

The least squares estimator is obtained by minimizing $RSS(\beta)$. To that end, we set the derivative of equal to zero to obtain the normal equations:

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}.$$

Solving for β, we obtain the ordinary least squares estimate:

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

provided that the inverse of $\mathbf{X}^T\mathbf{X}$ exists, which means that the matrix $\mathbf{X}$ should have rank D. As $\mathbf{X}$ is an $n \times D$ matrix, this requires in particular that $n \geqslant D$, i.e. that the number of parameters is smaller than or equal to the number of observations.

2.3.2 Penalized linear regression

The Gauss-Markov theorem (**?**) asserts that the least squares estimates $\hat{\beta}^{OLS}$ have the smallest variance among all linear unbiased estimates. However, there may well exist biased estimators with smaller mean squared error that would trade a little bias for a larger reduction in variance. Subset selection, shrinkage methods (ridge regression, lasso regression, ...) or dimension reduction approaches such as Principal Components Regression or Partial least Squares are

useful approaches if we want to obtain such biased estimates with smaller variance. In this section we will only detailed the most commonly used shrinkage methods, as they are the ones used in association genetics.

Ridge regression

The least squares estimates are the best unbiased linear estimators but this estimation procedure is valid only if the correlation matrix $\mathbf{X}^T\mathbf{X}$ is close to a unit matrix or full-rank, i.e. when the predictors are not orthogonal. If not, (**?**) proposed to base the estimation of the regression parameters on the matrix $(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})$, $\lambda \geq 0$ rather than on $\mathbf{X}^T\mathbf{X}$ and have developed the method named *ridge regression* to estimate the biased coefficients $\hat{\beta}^{ridge}$. This method shrinks the coefficients of the regression towards zero by imposing a penalty on the sum of the squared coefficients. The ridge coefficients minimize a penalized residual sum of squares which can be written as follow:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_2^2 \right\},$$

or can be equivalently written as a constrained problem:

$$\underset{\beta}{\operatorname{argmin}} \left\{ ||\mathbf{y} - \mathbf{X}\beta||_2^2 \text{ subject to } \sum_{d=1}^{D} \beta_d \leq t \right\},$$

with $t \geqslant 0$ a size constraint and $\lambda \geqslant 0$ a penalty parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The ridge regression estimates can then be written as:

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

We can notice that the solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion, which makes the problem non-singular even if $\mathbf{X}^T\mathbf{X}$ is not full rank.

The penalty parameter λ can be chosen either by K-fold cross-validation, leave-one-out cross-validation or by using the generalized cross-validation (**?**). In generalized cross-validation, the estimate $\widehat{\lambda}$ is the minimizer of $V(\lambda)$ given by

$$V(\lambda) = \frac{1}{n} \frac{||(\mathbf{I} - \mathbf{A}(\lambda))\mathbf{y}||_2^2}{\left[1/n\text{Trace}(\mathbf{I} - \mathbf{A}(\lambda))\right]^2},$$

where $\mathbf{A}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^T$ and is known as the hat matrix.

Moreover (**?**) have shown that the total variance of the ridge coefficients decrease as λ increases while the squared bias decrease with λ and that there exists values λ for which the MSE is less for $\hat{\beta}^{ridge}$ than it is for $\hat{\beta}^{OLS}$. These properties

lead to the conclusion that it is advantageous to take a little bias to substantially reduce the variance and thereby improving the mean square error of estimation and prediction.

Lasso

The *lasso* (**?**) is also a shrinkage method but unlike ridge regression, it may set some coefficients to zero and thus perform variable selection. The lasso estimate is close to the ridge regression in the sense that it is a penalized linear regression with a penalty on the sum of the absolute value of the coefficients:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1 \right\},$$

which can be equivalently written as the constrained problem:

$$\underset{\beta}{\operatorname{argmin}} \left\{ ||\mathbf{y} - \mathbf{X}\beta||_2^2 \text{ subject to } \sum_{d=1}^{D} |\beta_d| \leq t \right\},$$

with $t \geqslant 0$ a size constraint and $\lambda \geqslant 0$ a penalty parameter.

Comparing and , we can see that the difference between lasso and ridge regression is found in the penalized term, the $||\beta||_2^2$ term (ℓ_2 squared norm) in ridge regression penalty has been replaced by $||\beta||_1$ (ℓ_1 norm) in the lasso penalty. The ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to 0 when the penalty parameter λ is sufficiently large. This leads to *sparse* models much easily interpretable than those produced by ridge regression. Figure **??** illustrates how the lasso procedure can achieve sparsity while the ridge coefficients are only shrinking to zero.

However, the constraint put on the ℓ_1 norm of the coefficients makes the solution of the lasso non-linear in $\mathbf{y}$ and therefore there is no closed form expression to calculate the solutions as in ridge regression. Efficient algorithms are available for computing the entire path of solutions as λ varied, with the same computational cost as for the ridge regression (see homotopy methods (**?**) such as LARS (**?**), or also proximal algorithms (**?**) for more details).

As for ridge regression, the tuning parameter λ needs to be chosen but with the lasso we cannot rely on the generalized cross-validation to calculate the best value for λ. However, it is possible to use an ordinary cross-validation where we choose a grid of λ values and compute the cross-validation error for each value of λ. We then select the tuning parameter for which the value of the cross-validation error is minimized and re-fit the model using all the available observations with the best λ.

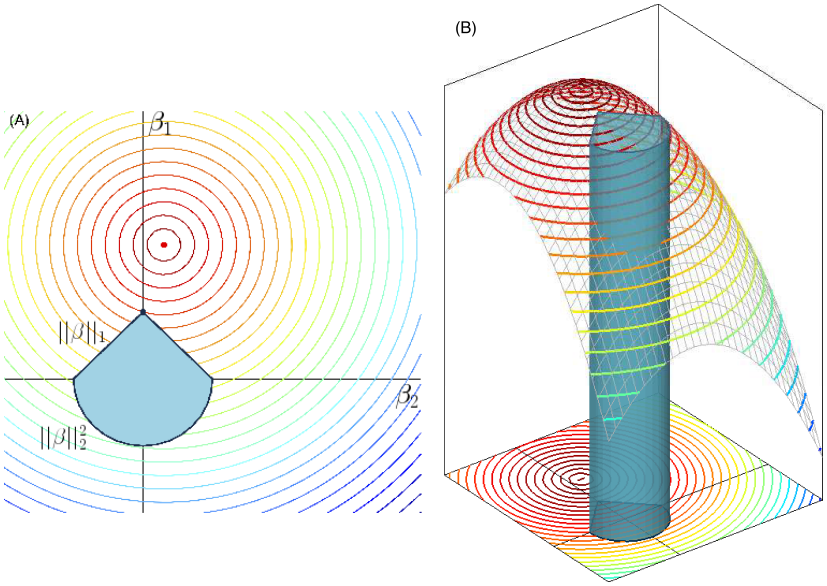


Figure 2.2: Geometrical representation of sparsity in penalized linear regression. (A) A 2-dimensional representation of space of the coefficients β_1 and β_2. The blue geometric form represents two types of constraints, $||\beta||_1$ and $||\beta||_2^2$, applied to the coefficients. The circular coloured lines represent the contour of the cost function and the red dotted point is the true parameter β we seek to reach. (B) 3-dimensional view of (A) where the constraints are represented as a tube in which the penalized methods are forced to stay to estimate the coefficients β_1 and β_2 (Image credit: Yves Grandvalet).

Group-Lasso

In some problems, the predictors belong to pre-identified groups; for instance genes that belong to the same biological pathway, SNP included in the same haplotype block or collections of indicator (dummy) variables for representing the levels of a categorical predictor. In this context it may be desirable to shrink and select the members of a group together. The group-lasso regression (**?**) is one way to achieve this.

If we suppose that D predictors are divided into G groups, with p_g the number of variables in the group g then the group-lasso solution minimizes the following penalized criterion:

$$\widehat{\beta}^{GL} = \underset{\beta}{\operatorname{argmin}} \left\{ ||\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \beta_g||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\beta_g||_2 \right\},$$

with $\mathbf{X}_g$ the matrix of predictors corresponding to the g^{th} group, $\sqrt{p_g}$ the terms accounting for the varying groups sizes and $||\beta_g||_2$ the ℓ_2-norm of the coefficients corresponding to group g. Since the Euclidean norm of a vector β_g is zero only if all of its components are zero, this model encourages sparsity at the group level.

Generalizations include more general ℓ_2 norms $||\nu^T||_K = (\nu^T K \nu)^{1/2}$ as well as overlapping groups of predictors (**??**).

2.3.3 Generalized linear models

Generalized linear models (GLMs) (**?**) are an extension of linear models where the strict linearity assumption of linear models is somewhat relaxed by allowing the expected value of the response to depend on a smooth monotonic function of the linear predictor and has the basic structure:

$$g(\theta) = \mathbf{X}\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D,$$

where $\theta \equiv \mathbb{E}(\mathrm{Y}|\mathrm{X})$, g is a smooth monotonic 'link function', $\mathbf{X}$ the $n \times D$ model matrix and β the unknown parameters. In addition, the assumption that the response should be normally distributed is also relaxed by allowing it to follow any distribution from the exponential family. The exponential family of distribution includes many distributions useful for practical modelling such as the Poisson, Binomial, Gamma and Normal distribution (see (**?**) for comprehensive reference on GLMs). A distribution belongs to the exponential family of distributions if its probability density function can be written as

$$g_\theta(\mathbf{y}) = \exp\left[\frac{\mathbf{y}\theta - b(\theta)}{a(\phi)} + c(\mathbf{y}, \phi) \right],$$

where a, b and c are arbitrary functions, ϕ the dispersion parameter and θ known as the canonical parameter of the distribution.

Furthermore, it can be shown that

$$\mathbb{E}(\mathrm{Y}) = b'(\theta) = \mu,$$

and

$$Var(\mathrm{y}) = b''(\theta)\phi.$$

Estimation and inference with GLMs are based on maximum likelihood estimation theory (**?**). The log-likelihood for the observed response $\mathbf{y}$ is given by

$$l(f_\theta(\mathbf{y})) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta)}{a(\phi)} + c(y_i, \phi).$$

The maximum-likelihood estimate of β are obtained by partially differentiating l with respect to each element of β, setting the resulting expression to 0 and solving for β:

$$\frac{\partial l}{\partial \beta_d} = \sum_{i=1}^{n} \frac{(y_i - b_i'(\theta_i))}{\phi b_i''(\theta_i)} \frac{\partial \mu_i}{\partial \beta_d} = 0.$$

Substituting and into this equation gives

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mathrm{Var}(\mu_i)} \frac{\partial \mu_i}{\partial \beta_d} = 0 \quad \forall d. \tag{2.1}$$

There are several iterative methods to solve the equation and estimate the maximum likelihood estimates $\widehat{\beta_d}$. One can use the well-known Newton-Raphson method (**?**), Fisher scoring method (**?**) which is a form of Newton's method or the Iteratively Reweighted Least Squares method developed by (**?**).

Logistic regression

The logistic regression model (**?**) is a generalized linear model where the logit function, defined as

$$\mathrm{logit}(t) = \log\left(\frac{t}{1-t}\right), \text{ with } t \in [0, 1],$$

is used as the 'link' function for g and is applied in the case where we want to model a qualitative random variable Y with K classes. The logit function allows to model the posterior probability $\mathbb{P}(\mathrm{Y} = k)$ via linear function of the observations while at the same ensuring that they sum to one and remain in $[0, 1]$. The model has the form:

$$\begin{aligned}
\log\left(\frac{\mathbb{P}(\mathrm{Y}=1|\mathrm{X}=x)}{1-\mathbb{P}(\mathrm{Y}=K|\mathrm{X}=x)}\right) &= \beta_{10}+\beta_1^T x,\\
\log\left(\frac{\mathbb{P}(\mathrm{Y}=2|\mathrm{X}=x)}{1-\mathbb{P}(\mathrm{Y}=K|\mathrm{X}=x)}\right) &= \beta_{20}+\beta_2^T x,\\
&\vdots\\
\log\left(\frac{\mathbb{P}(\mathrm{Y}=K-1|\mathrm{X}=x)}{1-\mathbb{P}(\mathrm{Y}=K|\mathrm{X}=x)}\right) &= \beta_{(K-1)0}+\beta_{K-1}^T x,
\end{aligned}$$

and equivalently

$$\mathbb{P}(\mathrm{Y}=k|\mathrm{X}=x)=\frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{k=1}^{K-1}\exp(\beta_{k0}+\beta_k^T x)},\ \text{with } k\in[1,\dots,K-1].$$

When $K=2$ the model becomes simple since there is only a single linear function. It is widely used in biostatistics when we want to classify an individual as being a case or a control in genome-wide association studies for instance.

Logistic regression models are usually fit by maximum likelihood using the conditional likelihood of the response given the observations. In the two class case where $\mathbf{y}$ is encoded as $0/1$, the log-likelihood of the estimator can be written as:

$$\begin{aligned}
l(\beta) &= \sum_{i=1}^{n}\left[y_i\log\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)+(1-y_i)\log\left(1-\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)\right]\\
&= \sum_{i=1}^{n}\left[y_i\beta^T x_i-\log(1+e^{\beta^T x_i})\right].
\end{aligned}$$

2.4 Splines and generalized additive models: Moving beyond linearit

2.4.1 Introduction

So far, we have been interested in estimating a function $\hat{f}$ linear in $\mathbf{X}$, but in reality, it is unlikely to be true. Linear models have the advantage of being easily interpretable and the approximation of f by a simple linear function can avoid overfitting. On the other hand, when the true function is highly non-linear, they are often limited if one wants to be able to model a complex phenomenon or to make accurate prediction.

In this section we will describe some methods that allow to take into account the non-linear form of f by working on a linear basis expansion of the initial features. The idea is to augment/replace the matrix of inputs $\mathbf{X}$ with additional

variables, which are transformations of $\mathbf{X}$, and then use linear models in this new space of derived input variables.

We define the linear basis expansion of $x \in \mathbb{R}$ by:

$$s(x) = \sum_{k=1}^{K} \beta_k h_k(x),$$

with $h_k(x) : \mathbb{R} \mapsto \mathbb{R}$ the k^{th} transformation of x, $k \in [1, \dots, K]$. The function $s(\mathbf{x})$ is also referred as a *smoother* since it produces an estimate of the trend that is less variables than the response variable $\mathbf{y}$ itself. We call the estimate produced by a smoother a *smooth*.

The linear basis expansion offers a wide range of possible transformations for x such as:

- Third order polynomial transformation: $h_1(x) = x, h_2(x) = x^2, h_3(x) = x^3$,
- non-linear transformation: $h_k(x) = \log(x), \sqrt{x}, ...$,
- Piecewise constant transformation: $h_1(x) = I(x < \xi_1), h_2(x) = I(\xi_1 \leq x \leq \xi_2), \dots, h_K(x) = I(x \geq \xi_{K-1})$.

In the following sections we will present some methods based on the linear basis expansion such as the regression splines (Section **??**, smoothing splines (Section **??** and generalized additive models (Section **??**). Note that the splines are methods applied to a univariate function x while the generalized additive models extend the uses of splines and other non-linear functions to the multivariate case.

2.4.2 Regression splines

Piecewise polynomials regression splines

Here the data are divided into different regions, each being defined by a polynomial function and separated by a sequence of knots, $\xi_1, \xi_2, \dots, \xi_K$ and each piece are smoothly joined at those knots. For example, with one knot ξ, dividing the data into two regions and with third-order polynomial pieces, we can write:

$$s(x) = \begin{cases} \beta_{01} + \beta_{11}x + \beta_{21}x^2 + \beta_{31}x^3 + \epsilon \text{ if } x < \xi, \\ \beta_{02} + \beta_{12}x + \beta_{22}x^2 + \beta_{32}x^3 + \epsilon \text{ if } x > \xi. \end{cases}$$

Piecewise cubic polynomials are generally used and constrained to be continuous and to have continuous first and second derivatives at the knots. For any given set of knots, the smooth is computed by multiple regression on an appropriate set of basis vectors. These vectors are the basis functions representing the family of piecewise cubic polynomials, evaluated at the observed values of the predictor x.

Cubic regression splines

A simple choice of basis functions for piecewise-cubic splines (truncated power series basis) derives from the parametric expression for the smooth

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \beta_k (x - \xi_k)_+^3,$$

which have the required properties:

- s is cubic polynomial in any subinterval $[\xi_k, \xi_{k+1}]$,
- s has two continuous derivatives,
- s has a third derivative that is a step function with jumps at $\xi_1, \ldots, \xi_K$.

The sequence of knots can be placed over the range of the data or at appropriate quantiles of the predictor variable (e.g., 3 interior knots at the three quartiles).

A cubic spline satisfies the following properties:

$$s(x) \in C^2[\xi_0, \xi_n] = \begin{cases} s_0(x), & \xi_0 \le x \le \xi_1, \\ s_1(x), & \xi_1 \le x \le \xi_2, \\ \ldots & \\ s_{n-1}(x), & \xi_{n-1} \le x \le \xi_n, \end{cases}$$

and

$$s(x) : \begin{cases} s_{k-1}(x_k) = s_k(x_k) \\ s'_{k-1}(x_k) = s'_k(x_k) \\ s''_{k-1}(x_k) = s''_k(x_k) \end{cases}, \text{ for } k = 1, 2, \ldots, (n-1).$$

The choice of a third-order polynomial allows the function $s(x)$ to be continuous at the knots.

Natural splines

A variant of polynomial splines are the natural splines: these are simply splines with an additional constraint that forces the function to be linear beyond the boundary knots. It is common to supply an additional knot at each extreme of the data and impose linearity beyond them. Then, with $K - 2$ interior knots (and two boundary knots), the dimension of the space of fits is K. The lesser flexibility at the boundaries of natural splines tends to decrease the variance we can get when fitting regular regression splines.

We add the following condition to get a natural cubic spline:

$$s''(\xi_0) = s''(\xi_n) = 0.$$

Figure **??** illustrates the use of natural cubic splines for the construction of an interpolating smooth curve.

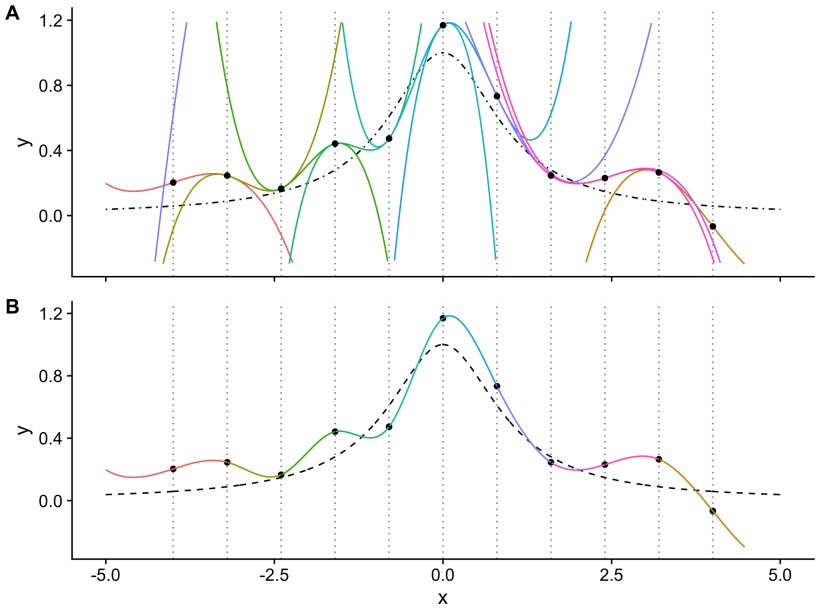


Figure 2.3: The black dashed line corresponds to the true distribution $y = \frac{1}{1+x^2}$ and each $n = 11$ black dots correspond to observations drawn from this distribution (with a little noise). In (A) we have represented the polynomial functions at each $K = 11$ knots, constituting the natural cubic splines basis and (B) the truncated polynomials to construct the smoother.

2.4.3 B-splines

The B-spline basis functions provide a numerically superior alternative basis to the truncated power series. Their main feature is that any given basis function $B_k(x)$ is non-zero over a span of at most five distinct knots which means that the resulting basis function matrix $\mathbf{B}$ is banded. The B_k are piecewise cubics and we need $K + 4$ of them ($K + 2$ for natural splines) if we want to span the entire space. The algebraic definition is detailed in (**?**).

With the B-spline basis, the functions are strictly local - each basis is only non-zero over the interval between $m + 3$ adjacent knots, where m is the order of the basis ($m = 2$ for cubic spline). To define a K parameters B-spline basis, we need to define $k + m + 1$ knots, $x_1 < x_2 < \cdots < x_{m+k+1}$, where the interval over which the spline is to be evaluated lies within $[x_{m+2}, x_k]$ (so that the first and last $m + 1$ knot locations are essentially arbitrary). An $(m + 1)^{th}$ order B-spline can be represented as

$$s(x) = \sum_{k=1}^{K} B_k^m(x)\beta_k,$$

where the B-spline basis functions are most conveniently defined recursively as follows:

$$B_k^m(x) = \frac{x - x_k}{x_{k+m+1} - x_k} B_k^{k-1}(x) + \frac{x_{k+m+2} - x}{x_{k+m+2} - x_{k+1}} B_{k+1}^{x-1}(x) \quad \text{for } k = 1, \ldots, K,$$

and

$$B_k^{-1}(x) = \begin{cases} 1 & x_k \leq x < x_{k+1} \\ 0 & \text{otherwise} \end{cases} .$$

For more detailed computational aspects see Annexe **??** and for a representation of B-spline functions see Figure **??**.

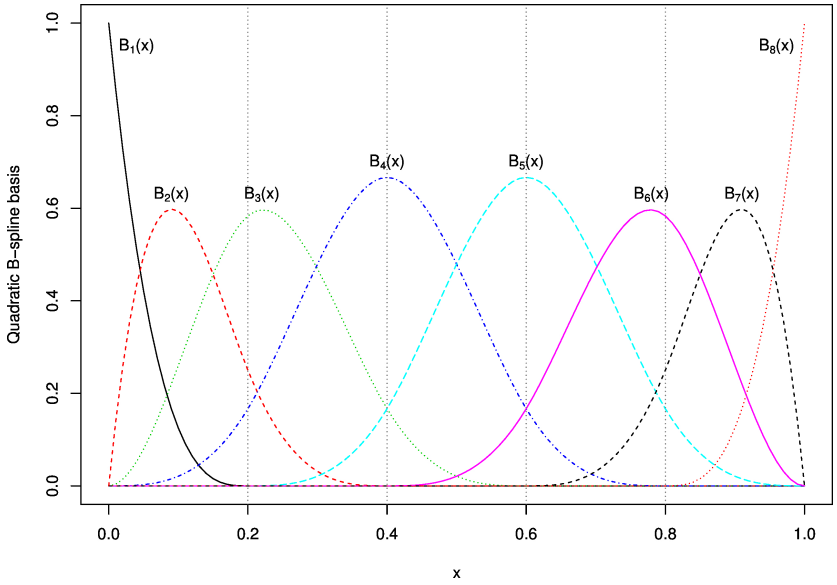


Figure 2.4: **Quadratic B-spline basis function representation** (for $m = 2$ and with $K = 4$ internal knots). Each $B_k(x)$ functions are piecewise cubic and $K + 4 = 8$ of them are need to span the entire space.

2.4.4 Cubic smoothing splines

This smoother is constructed as the solution to an optimization problem: among all function $f(x)$ with two continuous derivatives, find one that minimizes the

penalized residual sum of squares

$$\sum_{i=1}^{n} ||y_i - s(x_i)||_2^2 + \lambda \int_a^b s''(t)^2 dt,$$

where λ is a penalty factor, and $a \leq x_1 \leq ... \leq x_n \leq b$. The first term measures closeness to the data while the second penalizes curvature in the function, this criterion insuring a trade-off between bias and variance. The first term insures to fit as close as possible the data while the second penalizes the wiggliness of the smoothing curve to avoid interpolating the data. Large values of λ produce smoother curves while smaller values produce more wiggly curves.

As $\lambda \to \infty$, the penalty term dominates, forcing $s''(x) = 0$ everywhere and thus the solution is the least-squares line. On the contrary, as $\lambda \to 0$, the penalty term becomes unimportant and the solution tends to an interpolating twice-differentiable function.

Furthermore, it can be shown that this optimization problem has an explicit, unique minimizer which proves to be a natural cubic spline with knots at the unique value of x_i (see (**?**)).

We consider the smoothing function in the form:

$$s(x) = \sum_{k=1}^{K} N_k(x)\beta_k,$$

where the $N_k(x)$ are an (K)-dimensional set of basis functions for representing the family of natural splines. The natural cubic splines basis is computed as follow:

$$\begin{aligned} N_1(x) &= 1, \\ N_2(x) &= x, \\ N_{k+2}(x) &= d_k(x) - d_{k-1}(x), \end{aligned}$$

for $k \in [0, ..., K-1]$ and with

$$d_k = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k}$$

At first glance it would seems that the model is over-parametrized since there are as many as $K = n$ knots implying as many degrees of freedom. However, the penalty term converts into a penalty on the splines coefficients themselves, which are shrunk toward the linear fit.

Using this cubic spline basis for $s(x)$ means that can be written in the following minimization problem:

$$\underset{\beta}{\text{argmin}} ||\mathbf{y} - \mathbf{N}\beta||^2 + \lambda \beta^T \mathbf{W} \beta \tag{2.2}$$

where

$$N_{ik} = N_k(x_i),$$
$$W_{kk'} = \int_0^1 N_k''(x) N_{k'}''(x) dx,$$

with $\mathbf{W} \in \mathbb{R}^{n \times n}$ the penalty matrix and $\in \mathbb{R}^{n \times n}$ the matrix of basis functions.

Following (**?**), it can be shown that

$$W_{i+2,i'+2} = \frac{\left[\left(x_{i'} - \frac{1}{2}\right)^2 - \frac{1}{12}\right]\left[\left(x_i - \frac{1}{2}\right)^2 - \frac{1}{12}\right]}{4} - \frac{\left[\left(|x_i - x_{i'}| - \frac{1}{2}\right)^4 - \frac{1}{2}\left(|x_i - x_{i'}| - \frac{1}{2}\right)^2 + \frac{7}{240}\right]}{24},$$

for $i, i' \in [1, \ldots, K]$ with the first 2 rows and columns of $\mathbf{W}$ are equal to 0. For a given λ, the minimizer of , the penalized least squares estimator of β, is:

$$\widehat{\beta} = (\mathbf{N}^T\mathbf{N} + \lambda\mathbf{W})^{-1}\mathbf{N}^T\mathbf{y}.$$

It is interesting to note that this solution is similar to the ridge estimate , relating the smoothing splines to the shrinkage methods. Similarly the hat matrix, $\mathbf{A}$, for the model can be written as

$$\mathbf{A} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{W})^{-1}\mathbf{N}^T.$$

However, in spite of their apparent simplicity, these expressions are not the ones to use for computation. More computationally stable methods are preferred, i.e. the linear smoother described in (**?**), to estimate the smooth function $s(x)$ (see Annexe @ref(#linsmooth) for more details). For the choice of the smoothing parameter λ, see Annexe @ref(#lambda_smooth) and for an illustration of the cubic smoothing spline fit see Figure **??**.

2.4.5 Generalized additive models (GAM)

A generalized additive model (**?**) is a generalized linear model with a linear predictor involving a sum of smooth functions of D covariates.

$$g(\theta) = \beta_0 + \sum_{d=1}^{D} s_d(x_d) + \epsilon,$$

where $\theta \equiv \mathbb{E}(\mathrm{Y}|\mathbf{X})$, Y belongs to some exponential family distribution and g a known, monotonic, twice differentiable link function.

To estimate such model we can specify a set of basis functions for each smooth function $s_d(x)$.

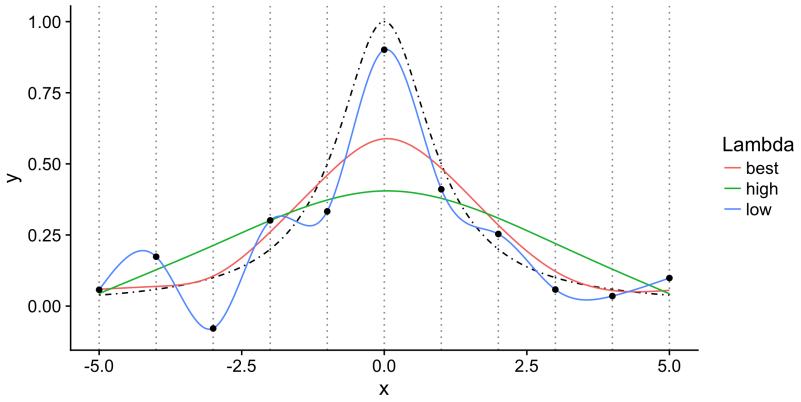


Figure 2.5: **Cubic smoothing splines** with different values of the regularization parameter λ. The black dashed line corresponds to the true distribution $y = \frac{1}{1+x^2}$ and each black dot corresponds to observations drawn from this distribution (with a little noise). In red is represented the fit at the best value of λ (chosen by GCV), in blue the fit with a value of lambda close to 0 and in green the fit with a high value for λ. We can see that, as λ increase, the fit pass from a 'wiggly' interpolating curve (as in Figure **??** to a very smoothed curve, which will eventually lead to a straight line as λ become very large.

For instance, with natural cubic splines, we get the following model:

$$g(\theta) = \beta_0 + \sum_{d=1}^{D} \sum_{k=1}^{K_d} \beta_{dk} N_{dk}(x_d) + \epsilon,$$

where K_d is the number of knots for variable d.

Furthermore, if we use cubic smoothing splines for each smooth function $s_d(x)$, we can define a penalized sum of squares problem of the form:

$$RSS(\beta_0, s_1, \dots, s_D) = \sum_{i=1}^{n} [y_i - \beta_0 - \sum_{d=1}^{D} s_d(x_{id})]^2 + \sum_{d=1}^{D} \lambda_d \int s_d''(t_d)^2 dt_d.$$

Each smoothing spline function $s_d(x)$ are then computed as described in Section **??** and the general model can be fitted with several methods such as backfitting or P-IRLS (Penalized-Iteratively Reweighted Least Squares) (**?**).

Fitting GAMs by backfitting

Backfitting is a simple procedure to fit generalized additive models which allow to use a large range of smooth function to represent the non-linear part of

the model. Each smooth component is estimate by iteratively smoothing partial residuals from the additive model, with respect to the covariates that the smooth relates to. The partial residuals relating to the d^{th} smooth term are the residuals resulting from subtracting all the current model term estimates from the response variable except for the estimate of d^{th} smooth.

Given the following additive model:

$$\mathbf{y} = \beta_0 + \sum_{d=1}^{D} s_d(x_d) + \epsilon.$$

Let $\hat{\mathbf{s}}_d$ denote the vector whose i^{th} element is the estimate of $s_d(x_{id})$. The backfitting algorithm is given in Algorithm 1.

Algorithm 1: Backfitting algorithm

1. Set $\hat{\beta}_0 = \bar{y}$ and $\hat{\mathbf{s}}_d = \mathbf{0}$ for $d = 1, \ldots, D$
2. **for** $d = 1, \ldots, D$ **do**
    1. Calculate partial residuals:
    $$\mathbf{e}^d = \mathbf{y} - \hat{\beta}_0 - \sum_{k \neq d} \hat{\mathbf{s}}_d.$$
    2. Set $\hat{\mathbf{s}}_d$ equal to the result of smoothing $\mathbf{e}^d$ with respect to $\boldsymbol{x}_d$.

   end
3. Repeat 2 until convergence.

2.4.6 High-dimensional generalized additive models (HGAM)

We consider an additive regression models in an high-dimensional setting with a continuous response $\mathbf{y} \in \mathbb{R}^n$ and $D \gg n$ covariates $x_1, \ldots, x_D \in \mathbb{R}^D$ connected through the model

$$\mathbf{y} = \beta_0 + \sum_{d=1}^{D} s_d(x_d) + \epsilon,$$

where β_0 is the intercept term, $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$ and $s_d : \mathbb{R} \to \mathbb{R}$ are smooth univariate functions. For identification purposes, we assume that all s_d are centered to have zero mean.

Sparsity-smoothness penalty

In order to get sparse and sufficiently smooth function estimates, (**?**), proposed the sparsity-smoothness penalty

$$J(s_d) = \lambda_1 \sqrt{||s_d||_n^2 + \lambda_2 \int [s_d''(x_d)]^2 dx}.$$

The two tuning parameters $\lambda_1, \lambda_2 \geq 0$ control the amount of penalization. The estimator is given by the following penalized least squares problem:

$$\hat{s}_1, \ldots, \hat{s}_D = \underset{s_1,\ldots,s_D \in \mathcal{F}}{\operatorname{argmin}} ||\mathbf{y} - \sum_{d=1}^{D} s_d||_n^2 + \sum_{d=1}^{D} J(s_d),$$

where $\mathcal{F}$ is a suitable class of functions and the same level of regularity for each function s_d is assumed.

Computational algorithm

For each functions s_d we can use a cubic B-spline parametrization with K interior knots placed at the empirical quantile of x_d.

$$s_d(x) = \sum_{k=1}^{K} \beta_{dk} b_{dk}(x_d),$$

where $b_{dk}(x)$ are the B-spline basis functions and $\beta_d = (\beta_{d1}, \ldots, \beta_{dK})^T \in \mathbb{R}^K$ is the parameter vector corresponding to s_d.

For twice differentiable functions, the optimization problem can be reformulated as

$$\widehat{\beta} = \underset{\beta=(\beta_1,\ldots,\beta_D)}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{B}\beta||_n^2 + \lambda_1 \sum_{d=1}^{D} \sqrt{\frac{1}{n}\beta_d^T \mathbf{B}_d^T \mathbf{B}_d \beta_d + \lambda_2 \beta_d^T \mathbf{W}_d \beta_d},$$

$$= \underset{\beta=(\beta_1,\ldots,\beta_D)}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{B}\beta||_n^2 + \lambda_1 \sum_{d=1}^{D} \sqrt{\beta_d^T \left(\frac{1}{n}\mathbf{B}_d^T \mathbf{B}_d + \lambda_2 \mathbf{W}_d\right) \beta_d},$$

where $\mathbf{B} = [\mathbf{B_1}|\mathbf{B_2}|\ldots|\mathbf{B}_D]$ with $\mathbf{B}_d$ is the $n \times K$ design matrix of the B-spline basis of the d^{th} predictor and where the $K \times K$ matrix $\mathbf{W}_d$ contains the inner products of the second derivative on the B-spline basis function.

The term $(1/n)\mathbf{B}_d^T \mathbf{B}_d + \lambda_2 \mathbf{W}_j$ can be decomposed using the Choleski decomposition

$$(1/n)\mathbf{B}_d^T \mathbf{B}_d + \lambda_2 \mathbf{W}_d = \mathbf{R}_d^T \mathbf{R}_d$$

to some quadratic $K \times K$ matrix $\mathbf{R}_d$ and by defining

$$\tilde{\beta}_d = \mathbf{R}_d \beta_d \text{ and } \tilde{\mathbf{B}} = \mathbf{B}_d \mathbf{R}_d^{-1},$$

the optimization problem reduces to

$$\hat{\tilde{\beta}} = \underset{\beta=(\beta_1,\dots,\beta_D)}{\operatorname{argmin}} ||\mathbf{y} - \tilde{\mathbf{B}}\tilde{\beta}||_n^2 + \lambda_1 \sum_{d=1}^{D} ||\tilde{\beta_d}||,$$

where $||\tilde{\beta}_d|| = \sqrt{K}||\tilde{\beta}_d||_K$ is the Euclidean norm in $\mathbb{R}^K$. This is an ordinary group lasso problem for any fixed λ_2, and hence the existence of a solution is guaranteed. For λ_1 large enough, some of the coefficient groups $\beta_d \in \mathbb{R}^K$ will be estimated to be exactly zero. Hence, the corresponding function estimate will be zero. Moreover, there exists a value $\lambda_{1,max} < \infty$ such that $\hat{\tilde{\beta}}_1 = \cdots = \hat{\tilde{\beta}}_D = 0$ for $\lambda_1 \geqslant \lambda_{1,max}$. This is especially useful to construct a grid of λ_1 candidate values for cross-validation (usually on the log-scale). By rewriting the original problem in this last form, already existing algorithms can be used to compute the estimator. Coordinate-wise approaches as in (**?**) and (**?**) are efficient and have rigorous convergence properties.

2.5 Combining cluster analysis and variable selection

2.5.1 Hierarchical clustering

Hierarchical clustering is a method of cluster analysis which aims at building a hierarchy of clusters and result in a tree-based representation of the observations called a *dendrogram*. The term hierarchical refers to the fact that clusters obtained by cutting the dendrogram at a given height are necessarily nested within the clusters obtained by cutting the dendrogram at any greater height.

Strategies for hierarchical clustering generally fall into two types (**?**):

- Agglomerative: This is a "bottom up" approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Ω being the training set to classify and *dist* a measure of dissimilarity (metric) on this set, we define a distance LC (linkage criterion) between the parts of Ω. The agglomerative hierarchical clustering algorithm is described in Algorithm 2.

Algorithm 2: Agglomerative hierarchical clustering

1. Begin with n observations and a measure of all the $n(n-1)/2$ pairwise dissimilarities and treat each observations as its own cluster.
2. **for** $i = n, n-1, ..., 2$: **do**
    1. Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are the most similar.
    2. Fuse these 2 clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
    3. Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters

end

Metric

The choice of an appropriate metric will influence the shape of the clusters, as some clusters may be similar according to one distance or farther away according to another. Given two sets of observations $A \subset \Omega$ and $B \subset \Omega$ with i the index of the i^{th} observation, the most commonly used metrics are:

- Euclidean distance: $||A-B||_2 = \sqrt{(\sum_i (A_i - B_i)^2)}$
- Manhattan distance: $||A-B||_1 = \sum_i |A_i - B_i|$
- Maximum distance: $||A-B||_\infty = \max_i |A_i - B_i|$

Linkage criteria

The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations. Some commonly used linkage criteria between two sets of observations $A \subset \Omega$ and $B \subset \Omega$ are:

- Single linkage: The dissimilarity between two sets is measured as the minimum dissimilarity between the observations of the sets:

$$LC(A,B) = \min\{dist(i,i'), i \in A \text{ and } i' \in B\}$$

- Complete linkage: The dissimilarity between two clusters is measured as the maximum dissimilarity between the observations of the groups:

$$LC(A,B) = \max\{dist(i,i'), i \in A \text{ and } i' \in B\}$$

- Average linkage: The dissimilarity between two clusters is measured as the averaged dissimilarity between the observations of the groups:

$$LC(A,B) = \frac{\sum_{i \in A} \sum_{i' \in B} dist(i,i')}{\text{card}(A).\text{card}(B)}$$

Ward's method

When the set $\Omega \in \mathbb{R}^D$ to classify is measured by D variables and where each element of Ω is represented by a vector x, we could use the method developed by (**?**) to construct a hierarchy among these variables. We note $\mathcal{G} = \{\mathcal{G}^1, \dots, \mathcal{G}^s, \dots, \mathcal{G}^S\}$ the group partitions coming from the S levels of the hierarchical clustering performed on the matrix $\mathbf{X} \in \mathbb{R}^{n \times D}$.

Given $\mathcal{G}^s = (\mathcal{G}^s_1, \dots, \mathcal{G}^s_g, \dots, \mathcal{G}^s_{G_s})$ a partition of Ω in G_s groups at a particular level s of the hierarchy, the within-group inertia is defined as

$$I_W(\mathcal{G}^s) = \sum_{g=1}^{G_s} \sum_{x \in \mathcal{G}^s_g} dist^2(x, \bar{x}_g),$$

where $\bar{x}_g$ is the centroid of group $\mathcal{G}^s_g$.

Equivalently we define the inter-group inertia as

$$I_B(\mathcal{G}^s) = \sum_{g=1}^{G_s} \text{card}(\mathcal{G}^s_g) dist^2(\bar{x}, \bar{x}_g),$$

where $\bar{x}$ is the centroid of Ω.

It can be shown that the total inertia, at a given level s, can be decomposed as

$$I_s = I_W(\mathcal{G}^s) + I_B(\mathcal{G}^s).$$

A partition will then be all the more homogeneous as the within-group inertia will be close to 0 and it can be shown that the fusion of two groups necessarily increases the total inertia. It is then possible to propose an agglomerative hierarchical clustering algorithm that fuse, at each step, the two groups $\mathcal{G}^s_g \in \mathcal{G}^s$ and $\mathcal{G}^s_{g'} \in \mathcal{G}^s$ that minimize the Ward's minimum variance criterion:

$$LC(\mathcal{G}^s_g, \mathcal{G}^s_{g'}) = \frac{\text{card}(\mathcal{G}^s_g).\text{card}(\mathcal{G}^s_{g'})}{\text{card}(\mathcal{G}^s_g) + \text{card}(\mathcal{G}^s_{g'})} d^2(\bar{x}_g, \bar{x}_{g'}),$$

where $\bar{x}_g$ and $\bar{x}_{g'}$ are the centroids of groups $\mathcal{G}^s_g$ and $\mathcal{G}^s_{g'}$ respectively.

Estimation of the number of clusters

The choice of the number of groups in cluster analysis is often ambiguous and depends on many parameters of the dataset. Several model selection criteria have already been investigated to makes such a decision (**???**). These methods are based on the measure of within-group dispersion I_W.

The gap statistic was developed by (**?**) to find a way to compare the distribution of $\log I_W(\mathcal{G}^s)$, $\mathcal{G}^s = (\mathcal{G}^s_1, \dots, \mathcal{G}^s_g, \dots, \mathcal{G}^s_{G_s})$, with its expectation $\mathbb{E}^*[\log I_W(\mathcal{G}^s)]$

under a reference distribution, i.e. a distribution with no obvious clustering. The gap statistic for a given number of groups G_s is then defined as

$$\text{Gap}(G_s) = \mathbb{E}^*[\log I_W(\mathcal{G}^s)] - \log I_W(\mathcal{G}^s).$$

To obtain the estimate $\mathbb{E}^*[\log I_W(\mathcal{G}^s)]$, B copies of $\log I_W(\mathcal{G}^s)$ are generated with a Monte Carlo sample drawn from the reference distribution and averaged.

The gap statistic procedure to estimate the optimal number of groups $\widehat{G}_s^*$ can be summarized as follows.

Step 1 : Construct the hierarchy on $\mathbf{X} \in \mathbb{R}^{n \times D}$, varying the total number of clusters from $G = (G_1, \dots, G_S)$ and compute the within-group inertia $I_W(\mathcal{G})$ for each partition $\mathcal{G} = (\mathcal{G}^1, \dots, \mathcal{G}^s, \dots, \mathcal{G}^S)$.

Step 2 : Generate B reference data sets from a uniform distribution over the range of observed values and cluster each one giving $I_W^*(\mathcal{G}^b)$ for each bootstrapped partition $\mathcal{G}^b = (\mathcal{G}^{b1}, \dots, \mathcal{G}^{bs}, \dots, \mathcal{G}^{bS}), b = (1, \dots, B)$. Compute the estimated gap statistic

$$\text{Gap}(G_s) = \frac{1}{B} \sum_{b=1}^{B} \log I_W^*(\mathcal{G}^{bs}) - \log I_W(\mathcal{G}^s).$$

Step 3 : Compute the standard deviation

$$sd(\mathcal{G}^s) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} [\log I_W^*(\mathcal{G}^{bs}) - \bar{b}]^2},$$

where $\bar{b} = 1/B \sum_{b=1}^{B} \log I_W^*(\mathcal{G}^{bs})$, and define $SD_s = sd(\mathcal{G}^s)\sqrt{1 + 1/B}$.

Step 4 : Choose the estimated optimal number of clusters via

$$\widehat{G}_s^* = \text{smallest } G_s \text{ such that } \text{Gap}(G_s) \geq \text{Gap}(G_{s+1}) - SD_{s+1}.$$

2.5.2 Hierarchical Clustering and Averaging Regression

Hierarchical Clustering and Averaging Regression (HCAR) is a method developed by (**?**) that combines hierarchical clustering and penalized regression in the context of gene expression measurement.

The Algorithm 3 can be summarized as follows: At first a hierarchical clustering is applied to the gene expression data to obtain a dendrogram that reveals their nested correlation structure. At each level of the hierarchy, a unique set of genes and supergenes is created by computing the average expression of the current clusters. Then, the different sets of genes and supergenes are used as inputs for a Lasso regression.

Algorithm 3: Hierarchical Clustering and Averaging Regression

1. Apply hierarchical clustering of the genes to yield the nested correlation structure. We define $\mathcal{G}_s = (\mathcal{G}_1^s, \ldots, \mathcal{G}_g^s, \ldots, \mathcal{G}_{G_s}^s)$ a group partition in G_s groups and $\mathbf{X}_{\mathcal{G}^s}^s = \left[\mathbf{X}_{\mathcal{G}_1^s}^s, \ldots, \mathbf{X}_{\mathcal{G}_{G_s}^s}^s\right]$ the concatenated matrix of variables for the partition $\mathcal{G}^s$.
2. **for** $s = 1, \ldots, S$ **do**
   Create supergenes matrix $\tilde{\mathbf{X}}^s = \tilde{\mathbf{X}}_{\mathcal{G}_1^s}^s, \ldots, \tilde{\mathbf{X}}_{\mathcal{G}_{G_s}^s}^s$ by averaging the gene expressions at each cluster of the current group partition $\mathcal{G}^s$:

$$\tilde{x}_{ig}^s = \frac{1}{G_s} \sum_{g \in \mathcal{G}_s} \mathbf{x}_{i\mathcal{G}_g^s}^s \qquad \text{with } i = 1, \ldots, n$$

   Fit Lasso, using the supergenes matrix $\tilde{\mathbf{X}}^s$ as input and retrieve the set of solution paths of the coefficients

end

3. Using cross-validation, find the optimal degree of shrinkage and level of the hierarchy.

Hierarchical clustering proved to be especially adapted in this context because it provides multiple levels at which the supergenes can be formed. Due to the fact that the Euclidean distance measure among the genes is a monotone function of their correlation (when the genes are properly standardized), hierarchical clustering provides flexibility in model selection in such a way that the genes are merged into supergenes in order of their correlation.

(**?**) proved that, in the presence of strong collinearity among the predictors, an averaged predictor yields to an estimate of the OLS coefficients with lower expected squared error than the raw predictors. The authors claimed that this theorem could easily be generalized to a block-diagonal correlation structure. The average features within each block may yield a more accurate fit than the individual predictors.

2.5.3 Multi-Layer Group-Lasso (MLGL)

(**?**) define the Multi-layer Group-Lasso (MLGL) as a two-step procedure that combines a hierarchical clustering with a Group-Lasso regression. It is a weighted version of the overlapping Group-Lasso (**?**) which performs variable selection on multiple group partitions defined by the hierarchical clustering. A weight is attributed to each possible group identified at all levels of the hierarchy. Such weighting scheme favours groups creating at the origin of large gaps in the hierarchy.

We note $\mathcal{G} = \{\mathcal{G}^1, \ldots, \mathcal{G}^s, \ldots, \mathcal{G}^S\}$ the group partition coming from the $s = 1, \ldots, S$ levels of the hierarchical clustering performed on the matrix $\mathbf{X} \in \mathbb{R}^{n \times D}$. $\mathcal{G}^s = (\mathcal{G}_1^s, \ldots, \mathcal{G}_{G_s}^s)$ is the group partition at the level s of the hierarchy and G_s

the total number of groups at the current level.

A group-lasso procedure is then fitted on the concatenated matrix of all group partition at all levels of the hierarchy

$$\mathbf{X}_{\mathcal{G}} = \left[\mathbf{X}^1_{\mathcal{G}^1}, \dots, \mathbf{X}^s_{\mathcal{G}^s}, \dots, \mathbf{X}^S_{\mathcal{G}^S}\right] \text{ where } \mathbf{X}^s_{\mathcal{G}^s} = \left[\mathbf{X}^s_{\mathcal{G}^s_1}, \dots, \mathbf{X}^s_{\mathcal{G}^s_{Gs}}\right].$$

The Multi-Layer Group-Lasso solution is defined by:

$$\hat{\beta}^{MLGL} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2}||\mathbf{y} - \mathbf{X}_{\mathcal{G}}\beta||_2^2 + \lambda \sum_{s=1}^{S} \rho_s \sum_{g=1}^{G_s} \sqrt{\text{Card}(\mathcal{G}^s_g)}||\beta_{\mathcal{G}^s_g}||_2 \right\},$$

with $\lambda \geqslant 0$ the penalty parameter, $\mathcal{G}^s_g \in \mathcal{G}^s$ the g^{th} cluster coming from level s of the hierarchy. The parameter ρ_s is a weight attributed to each group $\mathcal{G}^s_g$ and its purpose is to quantify the level of confidence in each level of the hierarchy. This weight is defined by:

$$\rho_s = \frac{1}{\sqrt{l_s}}$$

with $l_s = h_{s-1} - h_s$ the length of the gap between two successive levels of the hierarchy. Thus, the weight ρ_s is minimal when the length of the gap is maximal with the consequence of less penalizing in the groups at the origin of large gaps in the hierarchy.

2.6 Statistical testing of significance

2.6.1 Introduction

In statistical hypothesis testing, statistical significance refers to the acceptance or reject of the null hypothesis and corresponds to the likelihood that the difference between a given variation and the baseline is not due to random chance. For a given study, the defined level of significance α is the probability to reject the true null hypothesis and the p-value, p, is the probability of obtaining a result at least as extreme given that H_0 is true. We can therefore state that the result is statistically significant, by the standard of the study, if $p < \alpha$.

Ronald Fisher first advanced the idea of statistical hypothesis testing in his famous publication *Statistical Methods for Research Workers* (**?**). He suggested a probability of 5% has an acceptable threshold level to reject the null hypothesis and this cut-off was later taken over by Jezzy Neyman and Egon Pearson in (**?**) where they named it the significance level α.

They proposed the following hypothesis testing procedure:

(a) Before getting the experimental measures:

Table 2.1: Confusion matrix

	H_0 true	H_1 true
H_0 accepted	True Positive	False Positive
H_1 accepted	False Negative	True Negative

- Define the null hypothesis H_0 and the alternative hypothesis H_1.
- Choose a level α.
- Choose a test statistic, T, which is larger under H_1 than under H_0:

$$\text{Reject } H_0 \Leftrightarrow T \geq u.$$

- Study the distribution of T under H_0 and set the following condition:

$$\mathbb{P}(T \geq u) \leq \alpha.$$

- Deduce the threshold u.
- Give the test with the value retained for u and the real level:

$$\text{Reject } H_0 \Leftrightarrow T \geq u.$$

(b) Once the measures are done:

- Perform the numerical application and conclude if we accept or reject H_0 based on the p-value $= \mathbb{P}(T \geq t_{obs})$.

with

- Type I error: $\alpha = \mathbb{P}$ (accept H_1, H_0 is true),
- Type II error: $\beta = \mathbb{P}$ (accept H_0, H_1 is true),
- Power of the test: $1 - \beta = \mathbb{P}$ (accept H_1, H_1 is true).

and the confusion matrix defined in Table **??**.

2.6.2 χ^2 test

The chi-squared test, also written as χ^2 test, is a statistical hypothesis test developed by Karl Pearson and first published in (**?**). It is used when the sampling distribution of the test statistic under the null hypothesis follows a chi-squared distribution.

The χ^2 distribution with k degrees of freedom is the distribution of a sum of the squares of D independent standard normal random variables. If $\mathrm{X}_1, ..., \mathrm{X}_D$

are independent, normally distributed random variables, then the sum of their squares:

$$Z = \sum_{d=1}^{D} \mathrm{X}_d^2,$$

is distributed according to the χ^2 distribution with D degrees of freedom. This is usually denoted as $Z \sim \chi^2(D)$ or $Z \sim \chi^2_D$. The chi-squared distribution has one parameter: D — a positive integer that specifies the number of degrees of freedom.

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem.

2.6.3 Likelihood ratio test

The likelihood ratio test is used for comparing the goodness of fit of two statistical models, a null model against an alternative model. The log-likelihood ratio statistic is generally used to compute a p-value to decide whether or not to reject the null model.

Given the null $H_0 : \theta = \theta_0$ and the alternative hypothesis $H_1 = \theta = \theta_1$ for a statistical model $f(x|\theta)$, the likelihood ratio is defined as

$$\Lambda(x) = \frac{l(\theta_0|x)}{l(\theta_1|x)},$$

where $\theta \mapsto l(\theta|x)$ is the likelihood function and with $\alpha = \mathbb{P}(\Lambda(x) \leq u|H_0)$ the significance level at a threshold u.

In practice we define the test statistic as

$$\begin{aligned} T &= -2\log\left(\frac{l(\theta_0|x)}{l(\theta_1|x)}\right) \\ &= 2 \times [\log(l(\theta_1|x)) - \log(l(\theta_0|x))] \end{aligned}$$

The Neyman-Pearson lemma introduced in (**?**) states that the likelihood ratio test is the most powerful test at a significance level α.

2.6.4 Calculation of *p*-values in GAM

Let $\beta^j \in \mathbb{R}^K$ be the coefficients vector of the k covariates for a single smooth term j and $\mathbf{V}_{\beta_j}$ the covariance matrix of β_j. In the context of generalized additive models, if the covariates of the smooth are uncorrelated with other

smooth terms in the model, then $\mathbb{E}(\hat{\beta}_j) = 0$, otherwise there is little bias and $\mathbb{E}(\hat{\beta}_j) \simeq 0$.

Under the null hypothesis $H_0 : \beta_j = 0$ we have

$$\hat{\beta}_j \sim \mathcal{N}(0, \mathbf{V}_{\beta_j}).$$

It follows that if $\mathbf{V}_{\beta_j}$ is of full rank, then under the null hypothesis

$$\hat{\beta}_j{}^T \mathbf{V}_{\beta_j}^{-1} \hat{\beta}_j \sim \chi^2_k.$$

However, applying a penalty on the coefficients of the smooth, as it is the case with smoothing splines, often suppress some dimensions of the parameter space and consequently the covariance matrix $\mathbf{V}_{\beta_j}$ is not of full rank. If so, the test is performed using the rank $r = rank(\mathbf{V}_{\beta_j})$ pseudo-inverse of the covariance matrix $\mathbf{V}_{\beta_j}^{r-}$ and under the null,

$$\hat{\beta}_j{}^T \mathbf{V}_{\beta_j}^{-r} \hat{\beta}_j \sim \chi^2_r.$$

As stated in (**?**), as long as the p-values give a clear cut result it is usually safe to rely on them, but when they are close to the threshold of accepting or rejecting the null, they must be carefully treated. Indeed, as the uncertainty on the smoothing parameter estimation has been neglected in the reference distribution used for testing, these distributions are typically too narrow and attribute too low a probability to moderately high values in the test statistics. In that case, to obtain more accurate p-values, it may be preferable to perform test on overspecified unpenalized models even if it induces a cost in terms of statistical power.

2.6.5 Multiple testing comparison

In some context, as it is the case with the analysis of genes expression data or in Genome-Wide Association Studies (GWASs) for instance, we may need to perform simultaneously a very large number, $d \in [1, \dots, D]$, of tests and therefore the same large number of p-value. If we reject, for the d^{th} tests, the null hypothesis $H_{0,d}$ when its associated p-value $\hat{p}_d$ is not larger than α, then for each tests d, the probability to reject wrongly $H_{0,d}$ is at most α. Nevertheless, if we consider the D tests simultaneously the number of hypothesis $H_{0,d}$ wrongly rejected (false positive or type I error) can be very large. Actually, the expectation of the number of false positives in given by:

$$\mathbb{E}[\text{False Positives}] = \sum_{d:H0,d}^{D} \mathbb{P}_{H0,d}(T_d \geq u_\alpha) = \text{card}\{d : H_{0,d} \text{ is true}\} \times \alpha,$$

if the threshold u_α is such that $\mathbb{P}_{H0,d} = \alpha$for every d. For instance, for a typical value of $\alpha = 5\%$ and card $\{d : H_{0,d}$ is true$\} = 1000$, then we obtain on average 500 false positives. It is therefore necessary to adjust the threshold u_α at which we reject the null hypothesis in order to control for the number of false positives while not losing too much power.

Controlling the Family-Wise Error Rate

There exist many adjustments methods for multiple testing, including controls of the Family-Wise Error Rate (FWER), i.e. the probability of rejecting H_0 when it is true at least one time, noted as

$$\text{FWER} = \mathbb{P}(\text{card}(\text{False Positives}) \geq 1).$$

- Bonferroni procedure:

The most commonly used method for controlling the FWER is the Bonferroni method (**?**). The test of each H_d is controlled so that the probability of a Type I error is less than or equal to α/D, ensuring that the overall FWER is less than to a given α.

- Šidák method:

The method of (**?**) is closely related to Bonferroni's procedure where the p-value are adjusted as:

$$p_d^{adj} = 1 - (1 - p_d)^D,$$

where p_d is the unadjusted p-value for the d^{th} test.

- Holm method:

A less conservative adjustment method is the (**?**) method that orders the p-values and makes successively smaller adjustments. Let the ordered p-values be denoted by $p_1 \leq p_2 \leq ... \leq p_D$. Then, the Holm method calculates the adjusted p-values by

$$\begin{aligned} p_1^{adj} &= D \times p_1, \\ p_1^{adj} &= \max\{p_{d-1}, (D - d + 1) \times p_d\} \; 1 \leq d \leq D. \end{aligned}$$

The principal issue with these approaches is that they control the probability of at least one false positive regardless of the number of hypothesis being tested. They reduce the number of type I error but tends to be very conservative in the sense that the number of type II error is increased resulting in a loss of power. That is why less conservative methods are preferred in high-dimensional settings.

Controlling the False Discovery Rate

The False Discovery Proportion (FDP) corresponds to the proportion of false positives among the positive FP/(FP+TP). The False Discovery Rate, introduced in the seminal paper of (BH, **?**), is defined as the expected value of the FDP:

$$\text{FDR} = \mathbb{E}\left[\frac{\text{FP}}{\text{FP+TP}}\mathbb{1}_{\text{FP+TP}\geq 1}\right].$$

Controlling the FDR quantity offers a less conservative multiple-testing criterion than the FWER control. (**?**) proved that their approach, referred as the BH procedure, control the FDR at level α under the condition that the p-values following the null distribution are independent and uniformly distributed.

The BH procedure can be described as follow: Step 1 : Let $p_1 \leq p_2 \leq ... \leq p_D$ be the observed p-values.

Step 2 : Calculate

$$\widehat{k} = \underset{1\leq k\leq D}{\text{argmax}}\{k : p_k \leq \alpha k/D\}.$$

Step 3 : If $\widehat{k}$ exists, then reject the null hypothesis corresponding to $p_1 \leq ... \leq p_k$. If not, accept the null hypothesis for all tests.

(**?**) have shown that the FDR is upper-bounded by:

$$\text{FDR} \leq \alpha d_0/D,$$

with d_0 the number of true null hypothesis and have shown that this upper bounding is also true for positively dependent test statistics, i.e. when the distribution of p-values fulfils the Weak Positive Regression Dependency Property (WPRDS).

Since the BH procedure controls the FDR at a level of $\alpha d_0/D$ instead of α, a lot of work has been done in order to achieve a better level, mainly by trying to estimate d_0 (see (**?**) and references therein for more details).

Chapter 3

Genome-Wide Association Studies

This chapter focuses on Genome-Wide Associations Studies. It aims at explaining the principles and limitations of such studies. Section **??** exposes the critical points to consider in terms of genotyping quality control to avoid false positives. Section **??** introduces the concepts of disease penetrances and odds ratio generally used in genetic epidemiology. Section **??** places emphasis on the problem of population structure in GWAS. In section **??** is explained the classical single marker approach used in GWAS while Section **??** focuses on multi-marker methods to which we will refer in Chapter **??** and **??**.

3.1 Introduction

Linkage analysis (Section **??** was the traditional approach for disease gene mapping, where the co-segregation of marker alleles with disease within large pedigrees or smaller family is studied. This approach is efficient for locating genes contributing to simple Mendelian disorders where there is a strong relationship between phenotype and genotypes at the underlying functional polymorphisms. However, it proved to be less reliable regarding mapping of complex diseases as there may be multiple interacting genes underlying these phenotypes and that the effects of these genes may vary according to exposure to environmental and other non-genetic risk factors.

Whole Genome Association studies (WGA) focus on identifying genetic markers that occur with different frequencies between samples of unrelated affected individuals and unaffected controls, exploiting the fact that it is easier to establish large cohorts of affected individuals sharing a genetic risk factor for a complex disease across the whole population than within individual families,

as it is required for traditional linkage analysis. WGA rely in two types of association study: *direct association* and *indirect association*. On one hand, direct association focus on directly genotyping and studying functional polymorphisms which have relatively high prior probability of functional relevance such as non-synonymous polymorphisms[1], splice-site variants[2], and copy number polymorphisms (CNP[3]). One the other hand, indirect association, also referred as Genome-Wide Association Study (GWAS), focuses on both functional SNP, such as non-synonymous SNP, and those flanking them. Even if the flanking SNP are themselves unlikely to be directly associated with the phenotype, at sufficiently high density one or more is likely to be correlated (i.e. in linkage disequilibrium, see Section **??** with the underlying causal variants.

Furthermore, recent breakthroughs in micro-array technology have meant that hundreds of thousands of SNP can now be densely genotyped at moderate cost. As a result, it has become possible to characterize the genome of an individual with up to a million genetic markers. These rapid advances in DNA sequencing technologies have also made it possible to carry out exome and whole-genome sequencing studies of complex diseases. In this context, Genome-Wide Association Studies have been widely used to identify causal genomic variants[4] implied in the expression of different human diseases (rare, Mendelian or multifactorial diseases). Thanks to the Next Generation Sequencing techniques, it is now possible to genotype the complete DNA sequence of an individual at a moderate cost, around 1000 $ in 2016 (**?**), and in a very short time. Consequently, it is reasonable to think that the SNP will be abandoned in favour of a complete genotype and it is therefore necessary to develop statistical methods that can handle this kind of massive data.

3.2 Genotype quality control

In GWAS, the data filtering step used to identify genotyping mistakes is of primary importance since it can determine whether real discoveries are made or just false positives wrongly interpreted. With such large numbers of SNP being studied at the same time and with relatively moderate sample sizes, even small genotyping error rates can have a significant impact on the results.

As stated in (**?**), genetic effects on most multifactorial phenotypes follow an L-shaped distribution, with a few alleles having large effects and many alleles

[1] A non-synonymous SNP is a SNP that modifies the protein sequence in opposition to a synonymous SNP.

[2] A genetic alteration in the DNA sequence that occurs at the boundary of an exon and an intron (splice site). This change can disrupt RNA splicing, resulting in the loss of exons or the inclusion of introns leading to an altered protein-coding sequence.

[3] A CNP is a normal variation in DNA due to the varying number of copies of a sequence within the DNA. Large-scale copy number polymorphisms are common and widely distributed throughout the genome.

[4] In the remainder of the paper, the terms variant, marker, locus, SNP or polymorphism will equivalently refer to the variable studied in GWAS.

with a small effect size. This means that GWAS principally aim to identify small differences in allele frequencies between case and control, therefore even small experimental error can have strong effects on the results, particularly in the presence of rare alleles (**??**). The following paragraphs describe some of the filtering procedures designed to identify issues on specific SNP.

3.2.1 Deviation from HWE.

Neutral genetic variants in a large random-mating population are expected to display Hardy–Weinberg Equilibrium (see Section **??**). However, observed frequencies might be modified by genotyping error, leading to a deviation from HWE. A traditional approach for detecting genotyping errors is to test such deviation using the Pearson goodness-of-fit statistic (Section **??** and to look for significant deviations from the HWE (**?**). We usually perform this test only in the control sample since a deviation from HWE may also indicate an association with the disease. This test is insensitive to small deviations that are most often observed and, in a setting where there is a huge number of SNP to test for HWE deviation, an appropriate threshold of significance is therefore difficult to determine. Taking in account these considerations, the most prudent use of HWE tests for genotyping error may be only to exclude the most important deviations by setting an extreme significance threshold such as 1.10^{-7} or less, and using exact tests for rare alleles (**?**).

3.2.2 Missing data.

In case-control studies, markers having large differences in missing data rates between cases and controls often yield false positives (**?**). One can use the normal approximation to the binomial distribution to test for significant differences in missing data rates between cases and controls:

$$z = \frac{m_c - m_t}{\sqrt{m(1-m)(1/n_0 + 1/n_1)}},$$

with m_c and m_t the proportion of missing genotypes among cases and controls respectively, n_0 and n_1 the samples sizes of missing and non-missing data and m the overall missing genotype rate at the marker.

3.2.3 Distribution of test statistics.

When there are many significant loci coming out of a particular study it may more likely reflect systematic genotype error in some of those markers than reflect real discoveries. Indeed, remembering the L-shaped distribution of effect sizes, a study with $10^5 - 10^6$ genetic markers genotyped on one or two thousand cases and equal numbers of controls should reveal few genuine loci with single

locus p-values below 1.10^{-6} (**?**). Quantile-Quantile plot (Q-Q plot) is an efficient graphical way to examine the distribution of p-value and to evaluate whether there are too many data points in the tail. Q-Q plots are constructed by ordering test statistics and plotting them against the corresponding ordered expected values (see Figure **??** for an example).

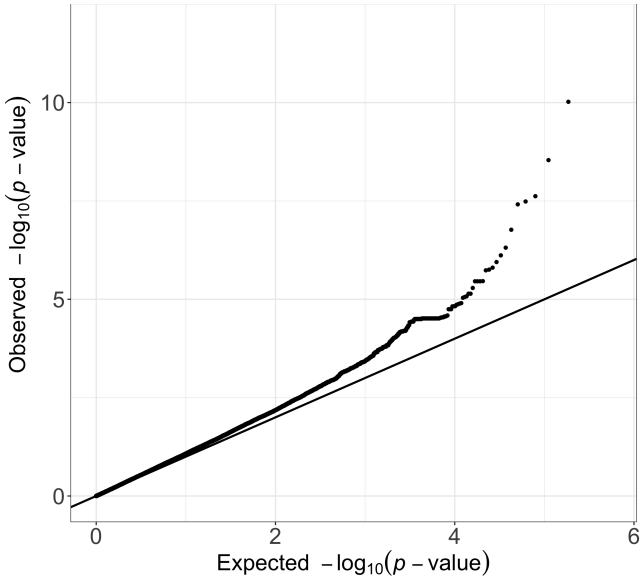


Figure 3.1: Example of Quantile-Quantile plot (Q-Q plot) representing the distribution of the test statistic for a classical GWAS study (results from GWAS analysis on Bipolar disorder data coming from the Welcome Trust Case-Control Consortium (**?**)). In this example we can see that the smallest p-value is equal to $4.5.10^{-5}$ and there are relatively few data points in the tail. Thus, based solely on the study of this distribution, there is therefore no reason to suspect genotyping errors.

3.3 Disease penetrance and odds ratio

Considering a biallelic locus with alleles A and a, the possible genotypes are then \$A/A \$, A/a and a/a. The *disease penetrance* associated with a given genotype is the risk of disease in individuals carrying this genotype. Assuming a genetic penetrance parameter $\gamma > 1$, the main disease penetrance models in association genetics can be summarized as:

- Multiplicative model: The risk of disease is increased by a factor of γ with each additional a allele

- Additive model: The risk of disease is increased by a factor of γ for genotype A/a and by a factor of 2γ for genotype a/a.
- Recessive model: The risk of disease is increased by a factor of γ for genotype a/a only.
- Dominant model: The risk of disease is increased by a factor of γ both for genotype A/a and a/a.

A commonly used measure of the strength of an association between phenotype and genotype is the *relative risk* (RR), which compares the disease penetrances between individuals carrying different genotypes (Table **??**).

	Penetrance			*Relative risk*	
Disease model	A/A	A/a	a/a	A/a	a/a
Multiplicative	f_0	$f_0\gamma$	$f_0\gamma^2$	γ	γ^2
Additive	f_0	$f_0\gamma$	$2f_0\gamma$	γ	2γ
Recessive	f_0	f_0	$f_0\gamma$	1	γ
Dominant	f_0	$f_0\gamma$	$f_0\gamma$	γ	γ

Disease penetrances for genotype A/A, A/a and a/a and the associated relative risks for genotypes A/a and a/a with f_0 the disease penetrance of baseline genotype A/A and γ the genetic penetrance parameter.

To estimate the RR it is therefore necessary to assess the disease penetrances which can only be derived directly from prospective cohort studies. In these studies, a group of exposed and unexposed individuals from the same population are followed up to evaluate who develop the disease of interest. However, in a case-control study, in which the case-control ratio is controlled by the investigator, it is not possible to make direct estimates of disease penetrance, and hence of RRs. In this type of study, the strength of an association is measured by the *odds ratio* (OR) (**?**).

In a case-control study, the odds of disease are defined as the probability that the disease is present compared with the probability that it is absent in exposed versus non-exposed individuals. Because of selected sampling, odds of disease are not directly measurable. However, conveniently, the disease OR is mathematically equivalent to the exposure OR (the odds of exposure in cases versus controls), which can be calculated directly from exposure frequencies (**?**). Two types of OR can be calculated:

- Allelic OR: It is estimated by comparing the odds of disease in an individual carrying allele A to the odds of disease in an individual carrying allele a.
- Genotypic OR: It represents the association between disease and genotype by comparing the odds of disease in an individual carrying one genotype

to the odds of disease in an individual carrying another genotype.

The risk factor for case versus control status is the genotype or allele at a specific marker. For each SNP with minor allele a and major allele A in case and control groups comprising n individuals, it is possible to represent the data as a $2 \times k$ contingency table of disease status by either allele ($k = 2$) or genotype ($k = 3$) count (Table **??**).

	Genotype count				*Allelic count*		
	A/A	A/a	a/a	Total	A	a	Total
Cases	n_{01}	n_{11}	n_{21}	$n_{.1}$	m_{01}	m_{11}	$m_{.1}$
Controls	n_{00}	n_{10}	n_{20}	$n_{.0}$	m_{00}	m_{10}	$m_{.0}$
Total	$n_{0.}$	$n_{1.}$	$n_{2.}$	N	$m_{0.}$	$m_{1.}$	N

2×3 contingency table of genotype counts and 2×2 contingency table of allelic counts for a single locus with alleles A and a. The genotype count n_{ij} corresponds to the observed frequency of individuals carrying i copies of the minor allele a with phenotype $j = 1$ for cases and $j = 0$ for controls. The allelic count m_{ij} can be summarized in different ways according to the disease penetrance models: for the dominant model, $i = 0$ if an individual is A/A and $i = 1$ otherwise, for a recessive model $i = 0$ if an individual is A/A or A/a and $i = 1$ otherwise.

Using the genotype count and allelic count exposed in Table **??**, we define the allelic odds ratio (OR_A), the allelic relative risk (RR_A) as:

$$OR_A = \frac{m_{01}m_{10}}{m_{00}m_{11}}, \quad RR_A = \frac{OR_A}{1 - p_0 + p_0 OR_A},$$

with p_0 is the estimated disease prevalence.

The genotypic odds ratio for genotype a/a relative to genotype A/A and for genotype A/a relative to genotype A/A is estimated by:

$$OR_{aa} = \frac{n_{21}n_{00}}{n_{01}n_{20}}, \quad OR_{Aa} = \frac{n_{11}n_{00}}{n_{01}n_{10}}.$$

Given a disease prevalence p_0, the relative risk of disease in individuals carrying a genotype a/a compared with an A/A genotype is:

$$RR_{AA} = \frac{OR_{AA}}{1 - p_0 + p_0 OR_{AA}}.$$

Figure **??** illustrates the relationship between allele frequency and disease penetrance in terms of disease representation. Low-frequency alleles which also have a low penetrance are very difficult to identify with common approaches while high-frequency alleles are those most commonly identified.

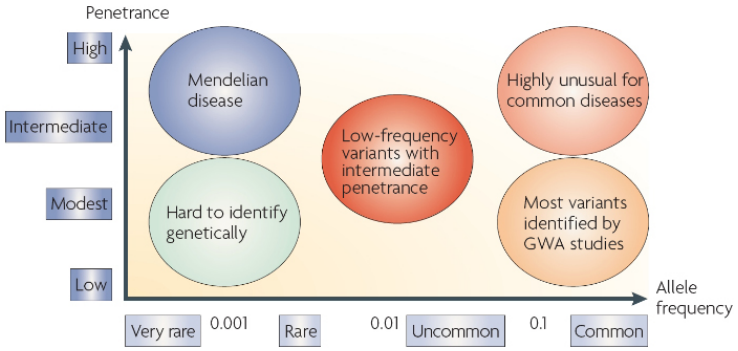


Figure 3.2: Relationship between allele frequency and penetrance on disease representation.

3.4 Single Marker Analysis

The standard statistical method to identify variants associated with a disease is to test the effect of each SNP one at a time using standard hypothesis testing methods. The goal is to identify genetic variants statistically associated with the phenotype, these variants being themselves in linkage disequilibrium with a potential causal polymorphism. Here we will review some of the most commonly used tests.

3.4.1 Pearson's χ^2 statistic

The expected value under the independence hypothesis of the genotype count n_{ij}, as defined in Table **??**, is noted as:

$$\mathbb{E}(n_{ij}) = \frac{n_{i.}n_{.j}}{N}.$$

It is thus possible to construct a genotypic association test by testing the independence between the rows and columns of the contingency table using the standard Pearson's χ^2 statistic for independence given by:

$$\chi^2_{genotypic} = \sum_{i=[0,1,2]} \sum_{j=[0,1]} \frac{(n_{ij} - \mathbb{E}(n_{ij}))^2}{\mathbb{E}(n_{ij})}.$$

This genotypic association test statistic has an approximate χ^2 distribution with 2 degrees of freedom (d.f.) under the null hypothesis H_0 of independence between the rows and columns of the contingency table.

As shown in Table **??**, it is also possible to consider alternative models of penetrance by focusing on allele count rather than genotype count. In this situation

the allelic association test is performed using a 2×2 contingency table and its associated χ^2 statistic is defined as:

$$\chi^2_{allelic} = \sum_{i=[0,1]} \sum_{j=[0,1]} \frac{(m_{ij} - \mathbb{E}(m_{ij}))^2}{\mathbb{E}(m_{ij})}.$$

This allelic association test, which have 1 d.f., will be more powerful than the genotypic test with 2 d.f., as long as the penetrance of the heterozygote genotype is intermediate compared to those of the two homozygous genotypes (**?**).

3.4.2 Cochran-Armitage trend test

Any penetrance model specifying a trend in risk with increasing numbers of alleles a can be examined using the Cochran-Armitage trend test (**??**) given by:

$$\chi^2_{CA} = \frac{\left[\sum_{i=0}^{2} w_i (n_{.1} n_{2.} - n_{.2} n_{1.})\right]^2}{\frac{n_1 . n_2 .}{n} \left[\sum_{i=0}^{2} w_i^2 n_{.i} (n - n_{.i}) - 2 \sum_{j=0}^{1} \sum_{i=j+1}^{2} w_j w_i n_{.j} n_{.i}\right]},$$

where $w = (w_0, w_1, w_2)$ are weights chosen to detect particular types of association. For instance, with a dominant model $w = (0, 1, 1)$ is optimal while for a recessive model weights $ w=(0,0,1) $ are rather chosen.

Under the null hypothesis of no association between the SNP and disease (H_0 : independence between rows and columns of the contingency table), χ^2_{CA} has an approximate χ^2 distribution with 1 d.f.. The power of this test is often improved as long as the disease risks associated with the A/a genotype are intermediate to those associated with the a/a and A/A genotypes. In GWAS, in which the underlying genetic model is unknown, the additive version of this test, i.e. with $w = (0, 1, 2)$, is most commonly used.

3.4.3 Logistic regression and likelihood ratio test

Another possible framework for modelling the relationship between a case-control phenotype and SNP genotype is to use the logistic regression model, as described in Section **??**. The logistic regression model is parametrised in terms of the log-odds of disease for each SNP genotype, denoted by β. The log-likelihood of observed phenotype data, $\mathbf{y}$ and genotype data $\mathbf{G}$, is given by:

$$l(\mathbf{y}|\mathbf{G}, \beta) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right) + (1 - y_i) \log \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right) \right],$$

where the linear predictor $\eta_i = \beta_0 + \beta g_i$.

Under the null hypothesis of no association $H_0 : \beta = 0$, we expect each genotype to have equal odds of disease, so that $\eta_i = \beta_0$. Under the additive model and treating allele A as baseline, the linear predictor becomes:

$$\eta_i = \beta_0 + \beta_A z_{(A)i},$$

where β_A corresponds to the additive effect of allele a and $z_{(A)i}$ is a variable representing the additive component of the ith genotype (see Table **??** for the SNP coding in different disease penetrance models).

Genotype	*Additive component* $\mathbf{z}_{(A)i}$	*Dominance component* $\mathbf{z}_{(D)i}$	*Recessive component* $\mathbf{z}_{(R)i}$
AA	0	0	0
Aa	1	1	0
aa	2	1	1

Coding of additive, dominance and recessive components of SNP genotypes.

In this framework, tests of association can be conducted with likelihood ratio (LR) methods in which inference is based on the likelihood of the genotyped data given disease status. The likelihood of the observed data under the proposed model of disease association is compared with the likelihood of the observed data under the null model of no association. For example, the log-likelihood ratio statistics,

$$\Lambda_{Gen} = l(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \hat{\beta}_A) - 2l(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \hat{\beta}_A = 0),$$

provides a genotype-based test of association which have an approximate χ^2 distribution with 2 d.f. under the null hypothesis. In large samples, it can be shown that χ^2 and LR methods are equivalent under the null hypothesis (**?**).

Furthermore, by using the flexible logistic regression framework, it is straightforward to incorporate additional covariates in the linear component, to allow the modelisation of environmental effects or to correct for population structure as we will see in the following section. The linear predictor η_i can thus be extend to:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \alpha_j x_{ij} + \beta_A z_{(A)i},$$

where x_{ij} is the response of the i^{th} individual to the j^{th} covariate and α_j its corresponding coefficient. Covariate adjustment reduces spurious associations due to sampling artefacts or biases in study design, but adjustment comes at the price of using additional degrees of freedom which may impact statistical power.

3.5 Limitations

The classical Single Marker Analysis approach is subject to false positives (i.e. SNP that are falsely identified as significant variables) due to the number of tests performed at the same time. One way around this problem is to apply a correction for multiple comparisons as described in Section **??**. Unfortunately, this increases the risk of missing true associations that have only a small effect on the phenotype, which is usually the case in GWAS. Indeed, simultaneously testing 1.10^5 SNP with single marker analysis would require that the associated p-value reach a threshold of at least 5.10^{-5}, using a Bonferroni correction, to be consider as significant and a little higher with FDR control method.

Furthermore, another commonly used approach for multiple testing comparisons in GWAS relies on the concept of genome-wide significance. It is based on the distribution of LD in the genome for a specific population and consider that there are an "effective" number of independent genomic regions, and thus an effective number of statistical tests that should be corrected for. For European-descent populations, this threshold has been estimated at $7.2.10^{-8}$ (**?**). This approach should however be used with caution since the only scenario where this correction is appropriate is when hypotheses are tested on the genome scale. Candidate gene studies or replication studies with a focused hypothesis do not require correction to this level, as the number of effective, independent statistical tests is much lower than what is assumed for genome-wide significance (**?**).

Furthermore, as stated in (**?**), these approaches face other limitations:

- It does not directly account for correlations among the predictors, whereas these correlations can be very strong as a result of linkage disequilibrium (LD). SNP can be correlated even where they are not physically linked, because of population structure or epistasis (gene by gene interactions).
- It does not account for epistasis, i.e. causal effects that are only observed when certain combinations of mutations are present in the genome.
- It does not directly provide predictive models for estimating the genetic risk of the disease.
- It focuses on identifying common markers with minor allele frequency (MAF) above 5%, although it is likely that analysing low-frequency ($0.5\% < \text{MAF} < 5\%$) and rare ($\text{MAF} < 0.5\%$) variants would be able to explain additional disease risks or trait variability (**?**).

Uncovering some of the missing heritability can sometimes be achieved by taking into account correlations among variables, interaction with the environment, and epistasis, but this is rarely feasible in the context of GWAS because of the multiple testing burden and the high computational cost of such analyses (**?**). That is why, knowing these limitations, we propose in Chapter 4 a new approach that take benefit of the correlation structure among SNP to improve statistical

power in GWAS.

3.6 Population structure

One of the most important covariate to consider in GWAS is the measure of population structure which, if not accounted for, can inflate the false positive error rate. As stated in Section @ref(#originLD), we know that population stratification as an important impact on patterns of LD and allele frequencies are highly variable across human subpopulations, meaning that in a sample with multiple strata, strata-specific SNP will likely be associated to the trait due to population structure. As a result, SNP with allele frequency differences between the strata will appear to be associated with disease, even if there is no association within each stratum. Several methods to identify and adjust for population stratification have been developed of which the most commonly used are genomic control, structured association and principle components correction (**?**).

3.6.1 Genomic control

Under the null hypothesis of no disease association, the distribution of Cochran–Armitage test statistics is χ^2_{CA} with 1 d.f. However, in a stratified population, we expect different allele frequency at many SNP and hence an excess of false positive signals of association. As a result, the observed distribution of association statistics will be inflated by a genomic inflation factor λ (**?**). The genomic inflation factor λ is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median:

$$\lambda = \text{median}(\chi^2_{CA})/0.456.$$

The genomic control method takes account of structure by a linear rescaling of observed test statistics to approximately restore the χ^2_{CA} with 1 d.f null distribution:

$$\chi^2_{CA-adj} = \chi^2_{CA}/\lambda.$$

3.6.2 Structured association

The method known as structured association, implemented in the STRUCTURE software (**?**), uses an admixture model[5] where the proportion of an individual's genome into K specific ancestral strata is treated as unknown. The posterior distribution of ancestry for each individual is then approximated using bayesian Markov Chain Monte Carlo (MCMC) methods based on genotype information

[5]An admixture model is a statistical model taking in account the phenomenon known as population admixture (see Section @ref(#originLD)).

from several hundred genome-wide SNP and the estimated structure is then included as covariates in a logistic regression framework. The main drawback of this approach is that the number of ancestral subpopulations must be inferred using an ad hoc estimation procedure and the computational load of the MCMC algorithm is such that it cannot accommodate for the numbers of markers commonly used in GWAS.

3.6.3 Principle components correction

This method makes use of the Principal Component Analysis (PCA) to detect and correct for population structure. In PCA, the few first principal components, calculated using the eigen-decomposition of a matrix, explain the greatest amount of variation in the data and has long been used to study population structure in genetic data (**?**). In GWAS, PCA has been used to explicitly model ancestry differences between cases and controls along continuous axes of variation and the first principle components may be used as covariates in a logistic regression model to adjust for the population structure effect. PCA being a computationally efficient algorithm, this approach has the advantage that it can be applied to datasets with more than 1.10^5 SNP.

The software EIGENSTRAT (**?**) use this approach by computing an adjusted test statistic defined as follow:

$$\chi^2_{eigen} = (n-k-1)r^2(\mathbf{z}_m^{adj}, \mathbf{y}^{adj}),$$

where $\mathbf{z}_m^{adj}$ is the adjusted genotype at marker m, defined as the residuals after regressing genotypes on the top k principal components. The adjusted phenotype $\mathbf{y}^{adj}$ is similarly defined. The test statistic χ^2_{eigen} approximately follows a χ^2 distribution with 1 d.f under the null hypothesis of no association. It has been shown that the EIGENSTRAT method has a higher power than genomic control because the correction in EIGENSTRAT is specific to a variation in frequency of a candidate marker across ancestral populations, which will minimize spurious associations as well as maximize power to detect true associations (**?**) .

3.7 Multi-locus analysis

As previously mentioned, in GWAS it is necessary for the SNP to be correlated to the causal polymorphisms in order to have an efficient disease mapping and, in complex disease, each single SNP have small effects on the phenotype. In this section we will show that we can take benefit from performing joint association tests of multiple SNP flanking a causal polymorphism to increase power in the case of rare-variant analysis or when the genetic effects are too small to be detected by single-locus approaches.

Several ways of grouping SNP together for multi-locus analysis are possible; we may consider to group SNP that fall within an established biological context such as a biochemical pathway, protein family, or gene. We can also consider working at haplotypes level rather than the genotypes and used the haplotype structure of the genome to define relevant groups (Section **??** and **??**).

Performing multi-locus analysis is not as straightforward as single marker analysis and presents some computational and statistical challenges. The most commonly used model to regress multiple SNP is the multiple linear regression with which we can simultaneously fit all SNP in the same gene or small genomic region. To reduce the problem of collinearity and overfitting that may arise, we can resort to penalized approaches, as described in Section **??**, such as ridge, lasso or group-lasso regression models. Furthermore, with such models, it is also possible to examine statistical interactions among genetic variants and so to investigate epistatic effects as in (**?**).

Other methods using multiple linear regression take into account the linkage disequilibrium within the genes to improve power (**?**) or cluster variants with weak association around known loci to increase the percentage of variance explained in complex traits (**?**). Finally, other approaches will focus on the aggregation of summary statistics of single SNP within a same gene with for instance the data-driven aggregation of summary statistics described in (**?**) or the procedures of p-value combination in (**?**).

3.7.1 Haplotype-based approaches

One approach to multi-locus analysis is to focus on haplotype effects. As seen in Section **??**, the human genome can be partitioned into haplotype blocks where most of the intra-block variability is imputable to mutation rather than recombination. As a result, much of common genetic variation can also be structured into haplotypes within blocks of LD that are rarely disturbed by meiosis.

It is common to assume that each of the pair of haplotypes, H_{i1} and H_{i2}, labelled according to their relative frequency in the population and forming the diplotype H_i of the i^{th} individual, contributes independently to the disease risk. Under this assumption the logistic regression model can be parametrised in terms of the log odds of disease for each haplotype (**?**). The linear predictor η_i of the i^{th} individual, as defined in Section **??**, can thus be defined as:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \alpha_j x_{ij} + \beta_{H_{i1}} + \beta_{H_{i2}},$$

where x_{ij} is the response of the i^{th} individual to the j^{th} covariate and α_j its corresponding coefficient. β_k denotes the log-OR of the k^{th} most frequent

haplotype, relative to the baseline haplotype, usually the most common, so that $\beta_1 = 0$.

One major issue with this approach is that we do not directly observe the diplotype H from the unphased genotype data. One solution is to take a point estimate of the diplotype for each individual, using statistical methodology, such as PHASE (**?**) or by maximum likelihood using the expectation-maximisation algorithm (**?**). However, due to the uncertainty in the haplotype reconstruction process, the variances of the model parameters are under-estimated leading to an inflation of type I error (**?**).

3.7.2 Rare-variant association analysis

In the context of rare-variant association analysis, a number of region- or gene-based multimarker tests have been proposed as burden tests (**?**), variance-component tests (**?**) or combined burden and variance-component tests (**?**). Instead of testing each variant individually, these methods evaluate the cumulative effects of multiple genetic variants in a gene or a region, increasing power when multiple variants in the group are associated with a given disease or trait.

We first introduce the statistical model used in various rare-variant tests and that is again based on the logistic regression framework defined in Section **??**. We assume that n individuals have been genotyped in a region comprising M genetic markers and defined the linear predictor η_i of the i^{th} individual as:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \alpha_j x_{ij} + \sum_{m=1}^{M} \beta_m z_{im},$$

where $z_{im} = z_{(A)im}$ is the variable representing the additive component of the i^{th} individual for the m^{th} marker and x_{ij} the response of the i^{th} individual to the j^{th} covariate. We define the score statistic of the model for variant m as

$$S_m = \sum_{i=1}^{n} z_{im}(y_i - \eta_i).$$

Note that S_m is positive when marker m is associated with increased disease risk or trait values and negative when marker m is associated with decreased risk or trait values.

Burden tests

Burden tests (**??**) compute a single genetic score from multiple genetic markers and test for association between this score and a phenotype of interest. A simple

approach summarizes genotype information by counting the number of minor alleles across all variants in the set. The summary genetic score is then:

$$C_i = \sum_{m=1}^{M} \omega_m z_{im},$$

where w_m is a weight attributed to marker m. The linear predictor can thus be written as:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \alpha_j x_{ij} + \beta_1 C_i.$$

To compute a p-value for a set of M markers, the specific test statistic Q_{burden} is calculated and compared to a χ^2 distribution with 1 d.f.:

$$Q_{burden} = \left[\sum_{m=1}^{M} \omega_m S_m \right]^2.$$

This framework is flexible in the sense that we can attribute different weights to the markers or define the genetic score C_i to accommodate for different assumptions about disease mechanism. For instance, the cohort allelic sums test (CAST, **?**) assumes that the presence of any rare variant increases disease risk and sets the genetic score $C_i = 0$ if there are no minor alleles in a region and $C_i = 1$ otherwise. Furthermore, to focus on rarer variants, we can assign $w_m = 1$ when the MAF of variant m is smaller than a prespecified threshold and $w_m = 0$ otherwise. Alternatively, a continuous weight function can be used to upweight rare variants with for instance $\omega_m = 1/\sqrt{\text{MAF}_m(1 - \text{MAF}_m)}$ as proposed by (**?**).

The burden methods make a strong assumption that all rare variants in a set are causal and associated with a trait with the same direction and magnitude of effect (after adjustment for the weights) which may in a substantial loss of power if these assumptions prove to be false (**?**).

Sequence Kernel Association Test (SKAT)

In (**?**), the authors proposed to group SNP into sets on the basis of their proximity to genomic features such as genes or haplotype blocks and then to identify the joint effect of each set via a logistic kernel-machine-based test. This approach lays the foundation for the Sequence Kernel Association Test method (SKAT, **?**).

SKAT uses the same logistic regression framework and the linear predictor as with burden tests but instead of testing the null hypothesis $H_0 : \beta_1, \ldots, \beta_M = 0$, it assumes that each β_m follows an arbitrary distribution with a mean of zero and variance of $\omega_m \tau$ where τ is a variance component and ω_m the weight attributed to marker m. With this assumption, we can see that $H_0 : \beta_1, \ldots, \beta_M =$

0 is equivalent to test $H_0 : \tau = 0$ which can be efficiently tested with a variance-component score test as used in generalized linear mixed model (GLMM) and is known to be a locally most powerful test (**?**). An advantage of this score test is that it requires to fit only the null model and to compute the following variance-component score statistic:

$$Q_{SKAT} = \sum_{i=1}^{n} \sum_{i'=1}^{n} (y_i - \eta_i) K(\mathbf{z}_i, \mathbf{z}_{i'}) (y_{i'} - \eta_{i'})$$

where $\eta_i = \beta_0 + \sum_{j=1}^{p} \alpha_j x_{ij}$ is the linear predictor of the null model including only the p covariates for individual i, and where $K(\mathbf{z}_i, \mathbf{z}_{i'}) = \sum_{m=1}^{M} \omega_m z_{im} z_{i'm}$. $K(.,.)$ is called the kernel function and measures the genetic similarity between individuals i and i', weighted by a factor ω_m, via the M genetic markers in the region of interest. This particular form of $K(.,.)$ is called the weighted linear kernel function and can take several forms to accommodate for epistatic effects for instance. In fact, any positive semi-definite function can be used as a kernel function and in their paper, (**?**) tailored the following commonly used kernels specifically for the purpose of rare-variant analysis:

- The weighted linear kernel:

  $$K(\mathbf{z}_i, \mathbf{z}_{i'}) = \sum_{m=1}^{M} \omega_m z_{im} z_{i'm}$$

  implies a linear relationship between the trait of interest and the genetic variants and is equivalent to the classical linear and logistic model described in Section **??**.

- The weighted quadratic kernel:

  $$K(\mathbf{z}_i, \mathbf{z}_{i'}) = (1 + \sum_{m=1}^{M} \omega_m z_{im} z_{i'm})^2$$

  assumes that the model depends on the main effects and quadratic terms for the gene variants and the first-order variant by variant interactions.

- The weighted identity by state (IBS[6]) kernel:

  $$K(\mathbf{z}_i, \mathbf{z}_{i'}) = \sum_{m=1}^{M} \omega_m IBS(z_{im}, z_{i'm})$$

  defines similarity between individuals as the number of alleles that share IBS.

Under the null hypothesis, Q_{SKAT} follows a mixture of chi-square distributions, which can be closely approximated with the computationally efficient Davies method (**?**).

[6] A DNA segment is identical by state (IBS) in two or more individuals if they have identical nucleotide sequences in this segment.

3.7.3 LD based approach to variable selection in GWAS

Region-based multi-marker analysis necessarily need that we define a group structure among SNP either by using the gene definition or biochemical pathway. However, these approaches limit the search for association to coding region only and therefore potential interesting associations located in non-coding region[7] are set aside.

One way to circumvent this issue is to use non-supervised clustering techniques such as hierarchical clustering described in Section **??**. In their paper, (**?**) proposed an approach where they used a modified version of the hierarchical clustering combined with a group-lasso regression to select groups of markers associated with phenotype of interest. The clustering method used is a spatially constrained agglomerative hierarchical clustering based on Ward's criterion in which the measure of dissimilarity is not based on the Euclidean distance but rather on the linkage disequilibrium level between two markers: $1 - r^2(m, m')$. The algorithm also makes use of the fact that the LD matrix can be modelled as block-diagonal by allowing only groups of variables that are adjacent on the genome to be merged, which significantly reduces the computation cost (*adjclust*, **?**).

The number of groups is then determined using a modified version of the gap statistic defined in Section **??**:

$$Gap(g) = \frac{1}{B}\sum_{b=1}^{B}(I^b_{W_g} - I_{W_g}),$$

where for $b = 1, \ldots, B$, $I^b_{W_g}$ denotes the within-cluster dispersion of clustering the reference dataset b in g groups. They decided to use the I_{W_g} instead of $\log(I_{W_g})$ in estimation since they noticed that it led to better estimation of the number of groups in the simulation studies, which were performed under a variety of parameters and on several data sets.

Finally, once the LD-defined groups structure have been determined, a group-lasso regression is performed in order to select groups of SNP associated with the phenotype. Given a phenotype vector $\mathbf{y}$ and a scaled matrix $\mathbf{Z}$ of additively coded SNP, the group-lasso estimate is defined as:

$$\hat{\beta}^{GL} = \underset{\beta}{\operatorname{argmin}} \left[||\mathbf{y} - \mathbf{Z}\beta||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g}||\beta^g||_2 \right],$$

[7] Non-coding DNA (formerly referred as 'junk' DNA) represents 99% of the genome and does not provide instructions for making proteins. However, recent studies have shown that non-coding DNA sequences can act as regulatory elements like sites for transcription factors implied in the control of gene transcription (source: https://ghr.nlm.nih.gov/primer/basics/noncodingdna).

where $||.||$ denotes the euclidean norm, $\lambda > 0$ is a penalty factor and β_g the vector of regression coefficients corresponding to the g^{th} group, so that $\beta = (\beta^1, ..., \beta^G)$.

Chapter 4

Learning the Optimal in GWAS through hierarchical SNP aggregation

The present chapter proposes a block-wise approach for GWAS analysis which leverages the LD structure among the genomic variants to reduce the number of hypothesis testing. We named this method **LEOS** for **LE**arning the **O**ptimal **S**cale in GWAS. Section **??** introduces some related works that have been studied to develop our methodology. The method is presented in Section **??**. In Section **??**, we compare our method in different scenarios with the baseline approach, i.e. univariate hypothesis testing (**?**) and with the logistic kernel machine method presented in Section **??** on both synthetic and real datasets from the Wellcome Trust Case Control Consortium (**?**) and on ankylosing spondylitis data (**?**). Finally, an example of an application using the generalized additive models in the context of GWAS is exposed in Section **??**.

4.1 Related work

Although classical GWAS have limitations that prevent a full understanding of the heritability of genetic and/or multifactorial diseases, there are nevertheless ways of overcoming these limitations to some degree. For instance, it is possible to take into account the structure of the data in the hypothesis testing procedure. As an illustration, (**?**) proposed a hierarchical testing approach which considers the influence of clusters of highly correlated variables rather than individual variables. The statistical power of this method to detect relevant variables at single SNP level was comparable to that of the Bonferroni-Holm procedure (**?**), but the detection rate was much higher for small clusters, and it increased

further at coarser levels of resolution.

In the broad family of linear models, (**?**) introduced a likelihood ratio-based set test that accounts for confounding structure. The model is based on the linear mixed model and uses two random effects, one to capture the set association signal and one to capture confounders. They demonstrate a control of type I error as well as an improved power over more traditionally used score test.

Other methods focus on multiple linear regression either by taking into account the linkage disequilibrium within the genes to improve power (**?**) or by clustering variants with weak association around known loci to increase the percentage of variance explained in complex traits (**?**).

Finally, other approaches will focus on the aggregation of summary statistics of single SNP within a same gene with for instance the data-driven aggregation of summary statistics described in (**?**) or the procedures of p-value combination in (**?**). In the cited articles, the methods are used on SNP located in coding region (or extended intronic region in (**?**)) but can be extended to any set of SNP as long as we pre-specified a set of variants within a region. However, the power for each test remains dependent of the true disease model. Furthermore, this kind of approaches may also lose statistical power in comparison to single-variant-based tests when only a very small number of the variants in a gene are associated with the trait, or when many variants have no effect or causal variants are low-frequency variants (**?**).

4.2 Method

In this section we describe a new method for performing GWAS using a four-step method that combines unsupervised and supervised learning techniques. This method improves the detection power of genomic regions implied in a disease while maintaining a good interpretability.

This method consists in:

- **Step 1**: Performing a spatially constrained Hierarchical Agglomerative Clustering of the additively coded SNP matrix $\mathbf{Z} \in \mathbb{R}^{n \times D}$ using the algorithm **??** developed by (**?**).
- **Step 2**: Applying a function to reduce the dimension of $\mathbf{Z}$ using the group definition from the constrained-HAC. This step is described and illustrated in Figure **??**.
- **Step 3**: Estimating the optimal number of groups using a supervised learning approach to find the best cut into the hierarchical tree (cut level algorithm). This algorithm combines Steps 1 and 2 into an iterative process.
- **Step 4**: Applying the function defined in Step 2 to each group identified in Step 3 to construct a new covariate matrix and perform multiple hy-

potheses testing on each new covariate to find significant associations with a disease phenotype **y**.

We entitled this method **LEOS** for **LE**arning the **O**ptimale **S**cale in GWAS, implemented in a web server too available at http://stat.genopole.cnrs.fr/leos.

4.2.1 Step 1. Constrained-HAC

To take into account the structure of the genome in haplotype blocks, we group the predictors (SNP) according to their LD in order to create a new predictor matrix which reflects the structure of the genome. We use the algorithm *adjclust* developed by (**?**) which consists in only allowing adjacent clusters to be merged, as described in Section **??**. This algorithm is available via the `R` package at https://cran.r-project.org/web/packages/adjclust.

A similar adjacency-constrained hierarchical clustering using Ward's linkage have already been proposed in (**?**), together with an algorithm called CONISS for Constrained Incremental Sums of Squares. However, the quadratic complexity of its implementation prevents it from being used on large genomic data sets.

In the context of GWAS, it is nevertheless possible to circumvent this problem by assuming that the similarity between physically distant SNP is small due to the particular LD structure of the genome, as seen in Section **??**.

More specifically, we assume that the $D \times D$ matrix of pairwise similarities defined as $\mathbf{S} = dist(i,j)_{1 \leq i,j \leq D}$ is a band matrix of bandwidth $h+1$, where $h \in [1, \dots, D] : dist(i,j) = 0$ for $|i-j| \geq h$ and D the number of naturally ordered objects (SNP) to classify. This assumption is not restrictive, as taking $h = D$ always works. However, considering the large dimension of genomic data, we are mostly interested in the case where $h \ll D$.

Adjclust is an algorithm that uses this band similarity assumption to improve time and space complexity in the context of a genome-wide hierarchical clustering. The main features of this algorithm are the constant-time calculation of each of the Ward's linkage involved in the spatially-constrained HAC and the storage of the candidate merges in a min-heap.

4.2.1.1 Ward's linkage as a function of pre-calculated sums.

To decrease the complexity in the calculation of each of the Ward's linkage, the trick is to note the sum of all similarities in any cluster $K = \{u, \dots, v-1\}$ of size $k = v - u$ as a sum of elements in the first $\min(h,k)$ subdiagonals of $\mathbf{S}$.

To see this, we define, for $1 \leq r, l \leq D$, the sum of all elements of $\mathbf{S}$ in the first l subdiagonals of the upper-right $r \times r$ block of S as

$$P(r,l) = \sum_{1 \leq i,j \leq r, |i-j| < l} dist(i,j),$$

and symmetrically, $\bar{P}(r,l) = P(p+1-r,l)$. Because P and $\bar{P}$ are sums of elements in pencil-shaped areas, they are called *forward pencil* and *backward pencil*, as illustrated in Figure **??**.

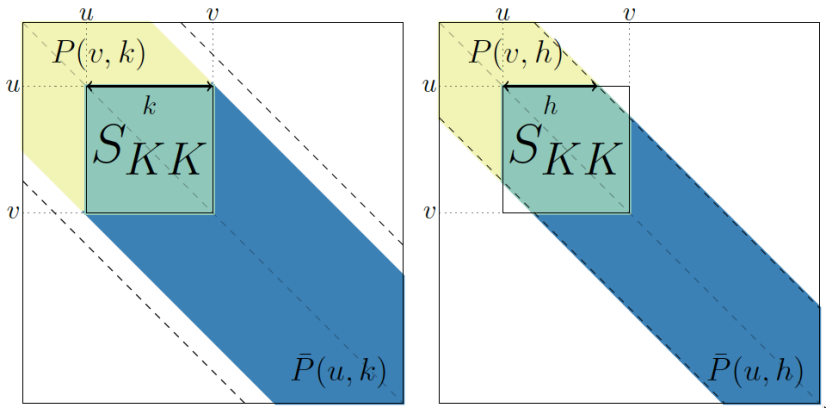


Figure 4.1: Example of forward pencils (in yellow and green) and backward pencils (in green and blue), and illustration of Equation (**??**) for cluster $K = \{u, \ldots, v-1\}$. Left: cluster smaller than bandwidth ($k \leq h$); right: cluster larger than bandwidth $k \geq h$.

The advantage of computing the sums P and $\bar{P}$ is that they can be used to calculate the sum S_{KK} of all similarities in cluster K following the identity:

$$P(v, h_k) + \bar{P}(u, h_k) = S_{KK} + P(p, h_k) \tag{4.1}$$

where $h_k := \min(h, k)$ and $P(p, h_k)$ is the "full" pencil of bandwidth h_k (which also corresponds to $\bar{P}(1, h_k)$). By construction, all the bandwidths of the pencils involved are less than h. Therefore, only pencils $P(u,k)$ and $\bar{P}(u,k)$ with $1 \leq u \leq p$ and $1 \leq k \leq h$ have to be pre-calculated, so that the total number of pencils to calculate and stored is less than $2ph$. By calculating these pencils recursively using cumulative sums, the time complexity of the pre-calculation step is ph (see proof in (**?**)).

4.2.1.2 Storing candidate fusions in a min-heap.

Each iteration i of the hierarchical agglomerative clustering (Algorithm 1, Section **??**, consists in finding the minimum of $D - i$ elements, corresponding to the candidate fusions between the $D - i + 1$ clusters, stored in a sorted list, and merging the corresponding clusters. However, as the cost of deleting and inserting an element in a sorted list is linear in D, adjclust choose to reduce the complexity by storing the candidate fusions in a partially-ordered data structure called a *min-heap* (**?**).

A min-heap is a binary tree structure constructed such that the value of each node is smaller than the value of its two children. The advantage of such structure is the cost trade-off they achieve between maintaining the structure and finding the minimum element at each iteration. More specifically, at the beginning of the clustering, the heap is initialized with $D-1$ candidate fusions in $\mathcal{O}(D\log(D))$. Then, each of the D iteration involves at most $\mathcal{O}(\log(D))$ operations as:

- finding the best candidate fusion (root of the min heap) in $\mathcal{O}(1)$,
- creating a new cluster corresponding to this fusion in $\mathcal{O}(1)$,
- deleting the root of the min heap in $\mathcal{O}(\log(D))$,
- inserting two possible fusions in the min heap in $\mathcal{O}(\log(D))$.

Globally, with a space complexity of $\mathcal{O}(Dh)$, corresponding to the $2Dh$ pre-calculated pencils, and a time complexity of $\mathcal{O}(D(h+log(D))$, where $\mathcal{O}(Dh)$ comes from the pre-calculation of pencils and $\mathcal{O}(D\log(D))$ from the D iterations of the algorithm, adjclust achieves a quasi-linear time complexity and linear space complexity when $h \ll D$.

In a GWAS application, the choice of h will mainly depends on the genotyping density and on the strength of the LD structure in the studied population. In the evaluation of our method in both numerical simulations **??** and real data application (Section **??**), we set the value at $h = 100$, having observed that higher values had no impact on the performance of the method.

4.2.2 Step 2. Dimension reduction function

One way of addressing issues related to high-dimensional statistics (and in particular the multiple testing burden that we mentioned in Section **??**) is to reduce the dimensionality of the predictor matrix $\mathbf{Z} \in \mathbb{R}^{N\times D}$ by creating a reduced matrix $\tilde{\mathbf{X}}$ with new covariates that nevertheless remain representative of the initial matrix. This means reducing the number of predictors D to $G \ll D$, with row $\tilde{\mathbf{S}x}_i$ the G-dimensional vector of new predictors for observation i. In this study we use a blockwise approach to construct a matrix of new uncorrelated predictors $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times G}$, with G the number of groups in linkage disequilibrium identified via the constrained agglomerative hierarchical clustering described in Step 1.

While classical methods use the initial set of covariates to predict a phenotype, we propose combining a clustering model with a dimension reduction approach in order to predict $\mathbf{y}$. For each group identified with the constrained-HAC, we apply a function to obtain a single variable defined as the number of minor alleles present in the group. For each observation i and in each cluster $g \in [1, ..., G]$, the variable is defined as:

$$\tilde{x}_{ig} \;=\; \sum_{d\in g} z_{id}.$$

We note that this function is close to the function used in the burden tests (Section **??**) where we attribute a weight $\omega_d = 1$ to each SNP since we do not particularly focus on rare variants but rather on variants having a $MAF \geq 5\%$. In order that the values for the different groups are comparable, we eliminate the effect of group size by centering and scaling the matrix $\tilde{\mathbf{X}}$ to unit variance. In the remainder of the paper we will refer to the covariates in $\tilde{\mathbf{X}}$ as *aggregated-SNP* variables.

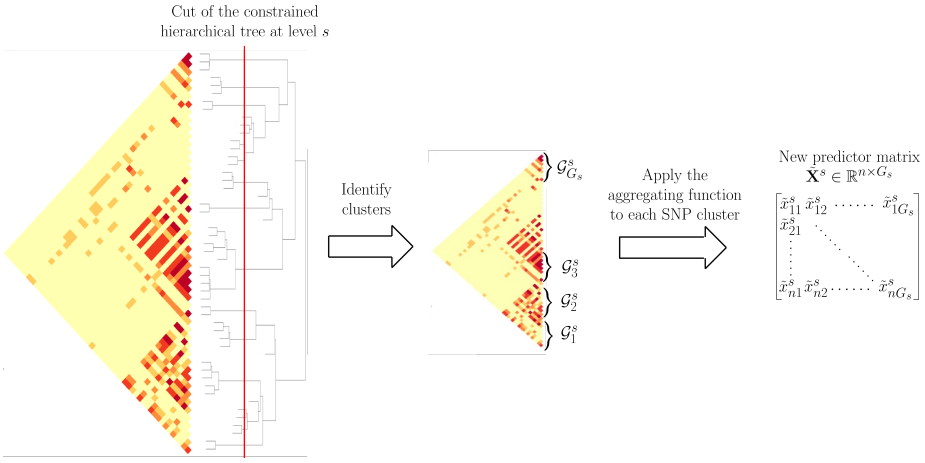


Figure 4.2: Schematic view of Step 2 of the algorithm to calculate the matrix of predictors $\tilde{\mathbf{X}}^s$ at a given level s of the hierarchy.

4.2.3 Step 3. Optimal number of groups estimation

Estimating the optimal number of groups to select, i.e. the level at which the hierarchical clustering tree should be cut, is a fundamental matter which impacts the relevance of the association analysis. As we have seen in Section **??**, it is known that the human genome is structured into haplotype blocks with little or no within-block recombination, but it is not easy to determine how these blocks are allocated throughout the genome for a given set of SNP.

In the literature, in an unsupervised learning context, a number of models have been proposed for determining the optimal number of groups in a hierarchical clustering (see Section **??**). However, since GWAS consist in evaluating the likelihood of the disease from genetic markers, we propose an algorithm that makes use of the phenotype $\mathbf{y}$ to determine the optimal number of clusters.

We propose here a supervised validation set approach to find this optimum. Since this algorithm aims to identify phenotype-related SNP clusters, it is necessary to split the dataset into two subsets to avoid an inflation of type I errors in the testing procedure. One subset, $[\mathbf{y}_{S1}, \mathbf{Z}_{S1}]$ with sample size $t_1 = n/2$ is

used to choose the optimal cut and the second one, $[\mathbf{y}_{S2}, \mathbf{Z}_{S2}]$ of sample size $t_2 = n/2$, to perform the hypothesis testing in Step 4.

The algorithm we propose can be summarized as follows:

- Apply the constrained-HAC described in Step 1 on a training set $\mathbf{T} = \mathbf{X}_{S1}^{train} \subset \mathbf{X}_{S1}$, and for a given level s of the hierarchy we apply the dimension reduction function defined above (Step 2) to each of the G_s clusters to construct the matrix $\tilde{\mathbf{T}}^s = \left\{\tilde{\mathbf{T}}_g^s\right\}_{g=\mathcal{G}_1^s}^{\mathcal{G}_{G_s}^s}$.
- Fit a ridge regression model to estimate the coefficients of the predictors in $\tilde{\mathbf{T}}^s$. We chose to resort on the ridge regression model because, as we explained in Section **??**, it is known to have a better stability in comparison to other penalized-regression models such as lasso regression (**?**).
- Once the ridge coefficients are estimated, we predict the phenotypic values on the test set using the matrix $\mathbf{U} = \mathbf{X}_{S2}^{test}$ and calculate either the mean test set error when the phenotype is quantitative or the Area Under the ROC curve (AUC-ROC) when it is binary.
- Repeat with procedure for different levels in the hierarchy and defined the optimal cut level s^* (or equivalently the optimal number of groups G^{s^*}) as the level which maximizes the prediction accuracy criterion.

Algorithm 4: Supervised learning cut level algorithm

input : Training set $\mathbf{T} = \mathbf{Z}_{S1}^{train}$ and test set $\mathbf{U} = \mathbf{Z}_{S1}^{test}$
output: Matrix $\tilde{\mathbf{X}}^{(s^*)}$ of aggregated-SNP at best cut level s^*

```
hierarchy ← Constrained-HAC on T
cutlevel ← Initialize levels where to cut hierarchy
for s ← Sequence(cutlevel) do
    T̃^s ← Aggregating(T, hierarchy, cutlevel[s]);
    Ũ^s ← Aggregating(U, hierarchy, cutlevel[s]);
    ridgecoef ← RidgeRegression(y_S1^train ~ T̃);
    y_S1^pred ← Predict(Ũ,ridgecoef);
    AUC[s] ← ROC(y_S1^test, y_S1^pred);
end
s* ← Which(cutlevel, Max(AUC));
X̃^(s*) ← Aggregating(Z, hierarchy, bestlevel);
```

At last, once the optimal number of groups G^* has been determined, we apply the function to each selected group and construct the matrix $\tilde{\mathbf{X}}^{(s^*)}$.

4.2.4 Step 4. Multiple testing on aggregated-SNP variables

Here we use a standard Single Marker Analysis, has described in Section **??**, to find associations with the phenotype, but instead of calculating p-value for each

SNP in $\mathbf{Z}$, we calculate p-value for each aggregated-SNP variable in $\tilde{\mathbf{X}}_{S2}^{(s^*)} \subset \tilde{\mathbf{X}}^{(s^*)}$.

For each single-predictor model, we perform a Likelihood Ratio Test (Section **??**) where we compare the intercept-only model against the single-predictor model and get for each predictor a p-value using the $\tilde{\chi}^2$ distribution.

As seen in Section **??**, we need to compute an appropriate significance threshold to control either the Family-Wise Error Rate or the False Discovery Rate. However, as the FWER control methods reduce the significance level according to the number of tests carried out in the study, it is preferable, in this context, to control for the FDR to be less stringent on the significance threshold. We therefore chose to use the Benjamini-Hochberg procedure described in Section **??** to adjust the significance threshold.

4.3 Numerical simulations

The performance evaluation described below was designed to assess the ability of our method to retrieve causal SNP or causal clusters of SNP under different simulation scenarios. For each scenario, we use a matrix $\mathbf{Z}_{\text{HAPGEN}}$ of SNP generated by the software (**?**) with a sample size of 1000 individuals. This software allows to simulate an entire chromosome conditionally on a reference set of population haplotypes (from HapMap3) and an estimate of the fine-scale recombination rate across the region, so that the simulated data share similar patterns with the reference data. We generate the chromosome 1 (103 457 SNP) using the haplotype structure of CEU population (Utah residents with Northern and Western European ancestry from the CEPH collection) as reference set. The software allows to generate a controls-only matrix of SNP (no disease allele). We filtered this matrix according to the minor allele frequency to only keep SNP with a MAF greater than 5% thus reducing the size of $\mathbf{Z}_{\text{HAPGEN}}$ to 60 179 SNP.

We generate a posteriori the phenotype using the logit model with a given set of causal SNP or cluster of SNP. The main difference between the different scenarios is to be found in the way that the case-control phenotype $\mathbf{y}$ is simulated.

4.3.1 Simulation of the case-control phenotype

For each scenario, we simulated a case-control phenotype $\mathbf{y}$ under a logistic regression model. The case-control phenotype is generated following a Bernoulli distribution function, following the conditional probability $\mathbb{P}(\mathrm{Y} = 1|\mathbf{H})$ with $\mathbf{H} \in \mathbb{R}^{n \times \ell}$ a matrix constructed by sampling ℓ causal variables from $\mathbf{Z}_{\text{HAPGEN}}$.

The conditional probability is calculated using the logit model:

$$\mathbb{P}(\mathrm{Y} = 1|\mathbf{H}) = \frac{\exp(\beta_0 + \beta^T\mathbf{H})}{1 + \exp(\beta_0 + \beta^T\mathbf{H})},$$

where $\beta = [\beta_1, ..., \beta_\ell]$ is the vector of coefficients corresponding to the ℓ predictors and β_0 is the intercept defined as $\log\left(\frac{\pi}{(1-\pi)}\right)$, with π the true prevalence of the disease in the population. The predictors are centered to have zero-mean before generating the vector of probability.

One way to have an association between the response and the predictors strong enough to be detected is to set large β coefficients on the predictors. Indeed, there is a direct relationship between the odd ratio of a covariate and its corresponding coefficients in the logistic regression model given by $OR_i = e^{(\beta_i)}$ (**?**). In our simulations, the difficulty of the problem, i.e. the power to detect an association, is linked to the number of causal predictors used to generate $\mathbf{y}$ and the OR set to each predictor.

To simulate different scenarios, we considered the following parameters:

1. Nature of the causal predictors:

   - **Clusters of SNP:** For each replicate, $\ell = \{1, 2, 3\}$ genomic regions have been identified to be causal. These regions have been chosen among the matrix $\mathbf{Z}_{\text{HAPGEN}}$ to have different levels of LD among the SNP that compose them. The average correlation coefficient among the SNP in these regions varies from $r^2 = 0.6$ to $r^2 = 0.85$ and the size of the region varies from 20 SNP to 60 SNP. Once identified, the causal regions were aggregated using the function described in Step **??** to construct a matrix $\tilde{\mathbf{H}}$ of aggregated-SNP predictors. This matrix was then used to generate the case-control phenotype following $\mathbb{P}(\text{Y} = 1|\tilde{\mathbf{H}})$. We will refer to this scenario as the *SNPclus* scenario.
   - **Single SNP:** In this scenario the phenotype was simulated by directly sampling SNP from the same causal regions identified in the *SNPclus* scenario. For each replicate, we chose 10 individuals SNP among each of these regions to construct a matrix $\mathbf{H}$ with $\ell = \{10, 20, 30\}$ single SNP predictors, depending on the number of causal regions . This matrix is then used to generate the case-control phenotype. The chosen SNP have a MAF varying from 10% to 30%. We will refer to this scenario as the *singleSNP* scenario.

2. Number of causal predictors ℓ and number of replicates:

   We performed 5 replicates for each combination $\ell \times 2$ scenarios and we evaluate the average performance over these 5 replicates. For each scenario we considered from 1 to 3 causal genomic regions, thus, for *SNPclus* scenario, we used up to 3 causal *aggregated*-SNP predictors, and for the *singleSNP* scenario, up to $10 \times 3 = 30$ causal single-SNP predictors to generate the phenotype.

3. Odds ratio (β coefficients) of the causal predictors:

   For the *SNPclus* scenario we chose an equal OR of 2.7 for each causal aggregated predictor, corresponding to a β coefficient equal to 1. For the

singleSNP scenario we chose an equal OR of 1.1 for each causal predictor, corresponding to a β coefficient equal to 0.1. The rationale behind these coefficients arises from the hypothesis that the combined effect of several low-effect SNP on the phenotype is stronger than the effects of each individual SNP.

As previously mentioned, we generated the phenotype using causal SNP simulated with the software. However, as commercial micro-arrays such as Affymetrix and Illumina arrays do not genotype the full sequence of the genome, some SNP are thereby unmapped and the marker density is in general lower than the HapMap marker density. That is why we chose, in our numerical simulation, to generate the phenotype with causal variables chosen from $\mathbf{Z}_{\text{HAPGEN}}$ and to assess the performance of the methods using only those SNP which are mapped on a standard Affymetrix micro-array (about 23 823 mapped SNP in our case). By doing so, some causal SNP are not mapped on the commercial SNP set and thus simulations are more similar to real genome-wide analysis conditions.

4.3.2 Performance evaluation

4.3.2.1 Competitors

The objective of our method being to identify the optimal scale at which to perform association studies, we compared our proposal with several methods working at different genomic scales. The purpose is to assess the ability of each method to retrieve true causal genomic regions in the different simulation scenarios.

For each scenario, four approaches have been considered:

- SKAT*tree*, a SKAT model, as described in Section **??**rare-variant), which use our group definition,
- SKAT*notree*, a SKAT model using an alternative group definition produced by successive chunks of 20 SNP,
- SMA, the classical Single Marker Analysis described in Section **??**,
- SASA (Single Aggregated-SNP Analysis) a method close to SMA, where instead of testing the genotype-phenotype association using each single SNP, we are testing it using aggregated-SNP variables.

We chose to consider two different group definitions for SKAT in order to evaluate the impact of the group structure on the association findings. The comparison with SMA allows to highlight the advantage of working at a group scale. We hypothesize that grouping low-effect SNP should have a better statistical power than testing the main effects at single-SNP level.

For all methods, we compare the results using 2 types of multiple testing corrections: the methods of Holm-Bonferroni (**?**) and (**?**).

4.3.2.2 True and False Positive definitions

The problem of retrieving true causal associations can be represented as a binary decision problem where the compared methods are considered as classifiers. The decision made by a binary classifier can be summarized using four numbers: True Positives (TP), False Positive (FP), True Negatives (TN) and False Negatives (FN). We represent True Positive Rate (Recall or Power $= TP/(FN + TP)$) versus Precision (Precision $= TP/(FP + TP)$). In this context, a true positive corresponds to a true causal genomic region associated to significant p-value. The definition of what can be considered as the true causal genomic region may nevertheless be subject to some ambiguity. In GWAS, the presence of LD between SNP often leads to consider the signal associated to multiple neighbouring SNP as indicating the existence of a single genomic locus with possible influence on the phenotype.

In our simulations, a causal genomic region is defined *a priori* as a causal predictor in the logit model. However, since the clusters of SNP identified by our algorithm are not totally independent, some residual correlation may remain between clusters. This leads to question the notion of relevant variable when the variables are structured into strongly correlated groups. Should all the variables of the block be considered as explanatory, or should we define as only true positives the causal variables used to generate the phenotype? In order to circumvent this issue, we chose to relax the definition of a false positive joining the work of (**?**) and (**?**) where they propose to control the FDR in GWAS by considering significant SNP correlated to the true causal variables as true positives. For the simulation of the phenotype, we hypothesize an underlying multivariate regression model, but test for univariate model as it is the usual practice, which leads to reconsider the definition of true positive. As in (**?**) we consider the set of true positive as the union of the causal true positive and the linked true positive, which are regions adjacent to the causal regions and correlated with them at a level of at least 0.5. Regarding the single-marker analysis approach, since it works at the single SNP level, we compare it with the others in the *singleSNP* scenario only.

4.4 Results

4.4.1 Results and discussions of the numerical simulations

4.4.1.1 Area Under the ROC Curve

For each simulation, the cut level algorithm was applied. We recall that this algorithm calculates a prediction error on a test set for several levels in a constrained-HAC tree with a ridge regression model and chooses the level for which this error is the smallest. The AUC-ROC is plotted for the different levels, and the best

cut level corresponds to the level for which AUC-ROC is the greatest. The results from the simulation scenario *clusSNP* and *singleSNP* described in Section **??** are shown in Figure **??**.

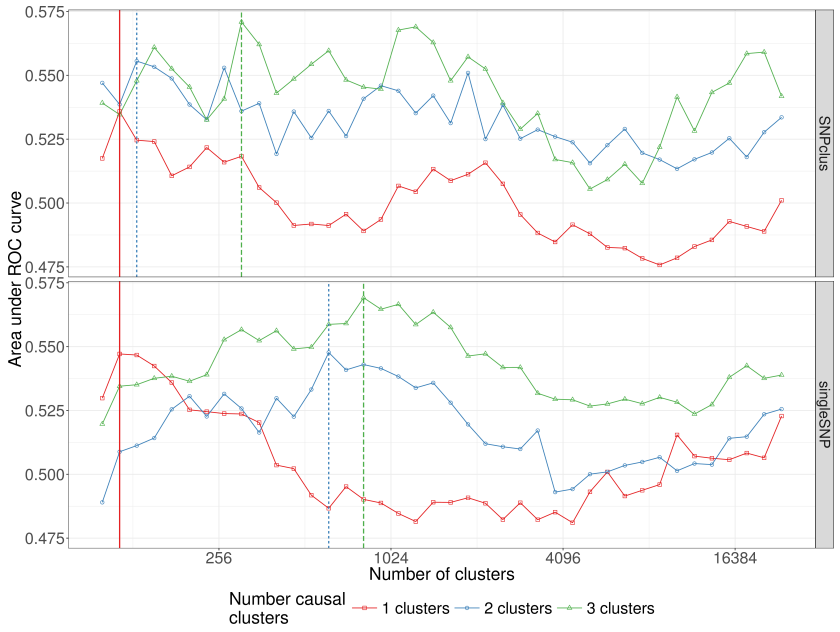


Figure 4.3: **Area under ROC curves according to the number of clusters in the *clusSNP* and *singleSNP* scenarios**: the vertical lines indicate the number of aggregated-SNP (clusters) obtained with Algorithm 4, i.e. the level where the prediction error is minimized (AUC-ROC at its maximum).

Our algorithm cuts the hierarchy either at a fairly high level (few large clusters) or at a low level (many small clusters), depending on the number of causal variables we used to generate the phenotype. The more the number of causal regions decreases, the higher the algorithm cuts in the hierarchical tree. In either case our algorithm is able to increase the predictive power by aggregating SNP with the function . We are thus able build a matrix of uncorrelated aggregated-SNP predictors that are representative of the initial SNP matrix and strongly linked to the phenotype.

4.4.2 Performance results for simulated data.

As previously described, we evaluate and compare the methods using two metrics, namely *Recall* and *Precision*.

Here the precision metric is somewhat relaxed compared to its true definition since we adapted the definition of a true positive and false positive to the GWAS

context. It is important to note that for all the methods, we compare the Benjamini-Hochberg method to control FDR with the Bonferroni correction to control FWER at a threshold of 5%. However, since there are residual correlations between SNP clusters and that the replication of numerous samples per combination of parameters is difficult in this realistic setting of simulations, the observed Type I error rate may be greater than 5%. What we think is important to put forward to in these simulations is the ability of our algorithm to define groups of relevant clusters that will be detected on average with more precision and more power (SASA and SKAT*tree*) than using an arbitrary group definition (SKAT*notree*) or no definition of groups at all (SMA).

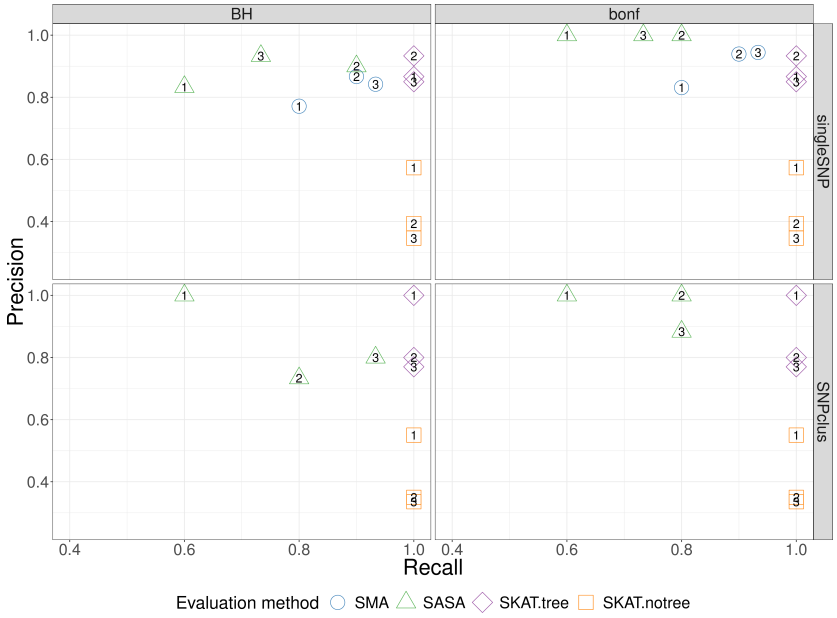


Figure 4.4: **Recall vs Precision for each method** (shape and colours in plot). In rows are the simulation scenarios. In columns, we evaluate performance using Benjamini-Hochberg threshold (left) and bonferroni correction threshold (right). The second row illustrates the performance to retrieve the true causal genomic region under the *SNPclus* scenario, thus only group-based approaches are considered (SASA, SKAT*tree* and SKAT*notree*). The numbers inside the points correspond to the number of causal predictors and each point is the average value of 5 replicates.

The results represented in Figure **??** show that the methods using our algorithm for the cluster definition (SASA and SKAT*tree*) have in average a better precision than the two other methods. The approach SASA, which combine our clustering algorithm and the aggregating function to test the association of aggregated-SNP with the phenotype, perform poorly in term of Recall but is far better in term of Precision compared to SMA and SKAT*notree*. These results

suggest that it is better to combine our algorithm with the SKAT method than with the SASA method. We also note that applying the SKAT approach on an arbitrary group definition (SKAT*notree*) lead to a good recall but a very poor precision, showing the benefit of using our custom group definition in this context. Regarding the SMA approach in the *singleSNP* scenario, we can observe a loss in term of Recall compare to the SKAT*tree* and SKAT*notree* method suggesting that we can take benefit of grouping low effect SNP to improve the power to detect causal genomic regions.

In GWAS, having a method with a good precision is as important, or even more important, than having a good recall. It is better to spot a few significant associations with a high certainty than to spot numerous significant associations but with only a low level of certainty for most of them. For this reason, we believe that our method represents an improvement in terms of precision without loss of power insofar as SKAT*tree* seems able to detect significant genomic regions associated with the phenotype with a higher degree of certainty than standard approaches.

4.4.3 Application in Wellcome Trust Case Control Consortium(WTCCC) and Ankylosing Spondylitis (AS) studies

To evaluate the performance of our method on real data, we performed GWAS analysis on datasets made available by (**?**). The WTCCC data collection contains 17000 genotypes, composed of 3000 shared controls and 14000 cases representing 7 common diseases of major public health concern: inflammatory bowel disease (IBD), bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), and Type I (T1D) and Type II (T2D) diabetes. Individuals were genotyped with the Affymetrix 500K Mapping Array Set and are represented by about 500,000 SNP (before the application of quality control filters).

In parallel to the analysis of the WTTCC data, we decided to assess our method on another dataset from a different study. The ankylosing spondylitis (AS) dataset consists of the French subset of the large study of the International Genetics of Ankylosing Spondylitis (IGAS) study (**?**). For this subset, unrelated cases were recruited through the Rheumatology clinic of Ambroise Paré Hospital (Boulogne-Billancourt, France) or through the national self-help patients' association: "Association Française des Spondylarthritiques". Population-matched unrelated controls were obtained from the "Centre d'Etude du Polymorphisme Humain", or were recruited as healthy spouses of cases. The dataset contains 408 cases and 358 controls, and each individual was genotyped for 116,513 SNP with Immunochip technology.

To remove the bias induced by population stratification in Genome-Wide analysis, we added the first 5 genomic principal components into the regression model

as described in Section **??**. Since the methods evaluated here do not deal with missing values, we chose to impute the missing genotypes with the most frequent genotypic value observed for each SNP.

For each dataset, we filtered the values to keep only those SNP having a MAF greater than 5%. The minor allele frequencies of each dataset are represented in Figure **??**.

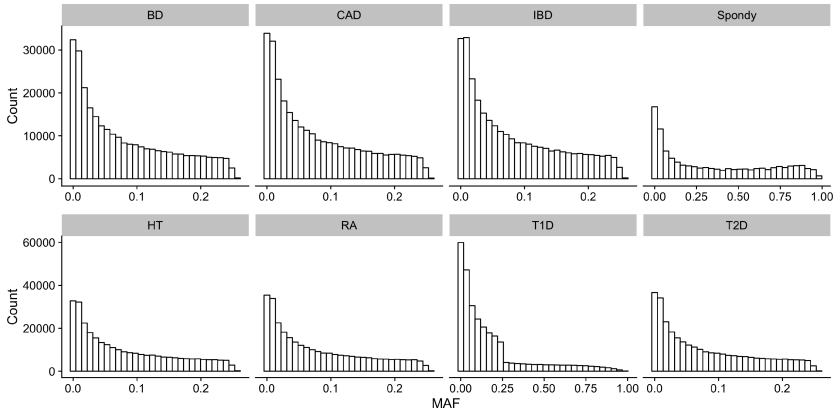


Figure 4.5: Histograms of Minor Allele Frequencies (MAF) distribution in each datasets. (BD) Bipolar disorders; (CAD) Coronary artery disease; (IBD) Inflammatory bowel disease; (HT) Hypertension; (RA) Rheumatoid arthritis; (T1D) Type I diabetes; (T2D) Type II diabetes.

We applied our cut level algorithm to find relevant clusters of SNP and we performed single marker analysis on single SNP (SMA) and on groups of SNP (SASA, SKAT*tree*, SKAT*notree*).

4.4.4 Results in WTCCC and AS studies

4.4.4.1 AUC-ROC curves

In this section, we compare the AUC-ROC curves generated by our cut level algorithm for each disease (WTCCC and AS data).

Concerning the WTCCC diseases, given that patients were all genotyped using the same micro-array, their genotypes have the same LD structure, and therefore the shapes of the AUC-ROC curves should be very similar between the different diseases. As can be observed in Figure **??** (WTCCC diseases), the shapes of the AUC-ROC curves are closely similar, with a chosen cut level located around 100 000 clusters of SNP, suggesting a shared LD pattern among patients.

In contrast, the AUC-ROC from the AS data (Figure **??**) behaves differently from the WTCCC data. Predictive power is substantially improved if

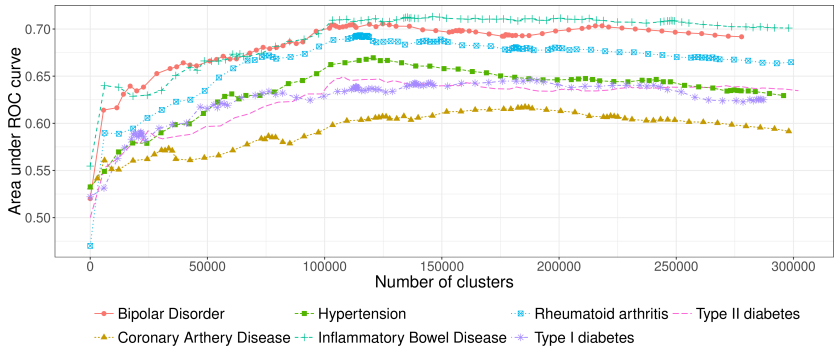


Figure 4.6: AUC-ROC for different cut levels in a HAC-tree of 7 WTCCC diseases after quality control filters. Each point corresponds to an AUC value computed on a test set from a logistic ridge regression model for a given level in the constrained-HAC tree.

aggregated-SNP predictors are used at a fairly high level in the hierarchical tree (7478 optimal clusters identified by the cut level algorithm). It is relevant to note that the pattern we observe on this real dataset is similar to the pattern we observed in the numerical simulations, especially under the *clusSNP* scenario.

Figure 4.7: AUC-ROC for different cut levels in a HAC-tree of the spondylitis arthritis disease (Immunochip micro-array). Each point corresponds to an AUC value computed on a test set from a logistic ridge regression model for a given level in the constrained-HAC tree.

As we remarked concerning the WTCCC results, the algorithm identifies a relatively high number of clusters in relation to AS and simulated data. This difference is certainly due to the LD level among the genetic markers in the Affymetrix array. The correlation levels among SNP for a given bandwidth are similar between the simulated and the AS data, but greater than for the WTCCC data (Table **??** and Figure **??**). This suggests that there is a stronger

LD pattern between blocks of SNP in AS and simulated data, implying that the optimal number of clusters identified by the algorithm is dependent on the LD level among variables.

Dataset	*SNP/kb*	*Median*	*Mean*
Simulated data	$1.3.10^{-27}$	1.10^{-2}	0.11
WTCCC data	7.10^{-32}	9.10^{-4}	0.03
AS data	9.10−9	3.10^{-2}	0.27

Comparison of marker density and averaged LD level between markers in a region of 300 SNP for the different datasets.

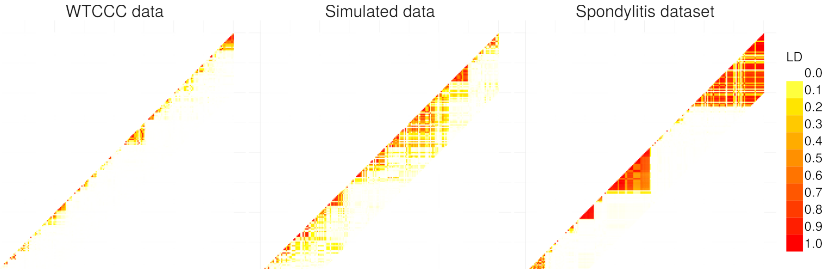


Figure 4.8: Comparison of linkage disequilibrium level among SNP for 3 different types of dataset: WTCCC, simulated and ankylosing spondylitis datasets. LD computation is based on R^2 between SNP.

4.4.4.2 GWAS analysis on AS and WTCCC datasets

To evaluate the ability of our procedure to discover new associations between SNP and ankylosing spondylitis, we compare our procedure with the univariate approach (SMA) and SKAT model with our group definition and arbitrary group definition (20 SNP). For SASA, we perform multiple hypotheses testing on the aggregated-SNP predictors in order to unravel significant associations with the phenotype. Figure **??** presents the results of the association analysis. For each method the logarithm of the p-value of the different predictors is plotted along their position on the genome.

Either method highlight a region on chromosome 6 strongly associated with the phenotype. This region corresponds to the Major Histocompatibility Complex (MHC), and Human Leukocyte Antigen (HLA) class I molecules HLA B27 belonging to this region have been identified as a genetic risk factor associated with ankylosing spondylitis (**?**). The approach SASA succeeds in detecting this risk locus with a good precision, 64 aggregated-SNP variables are significantly

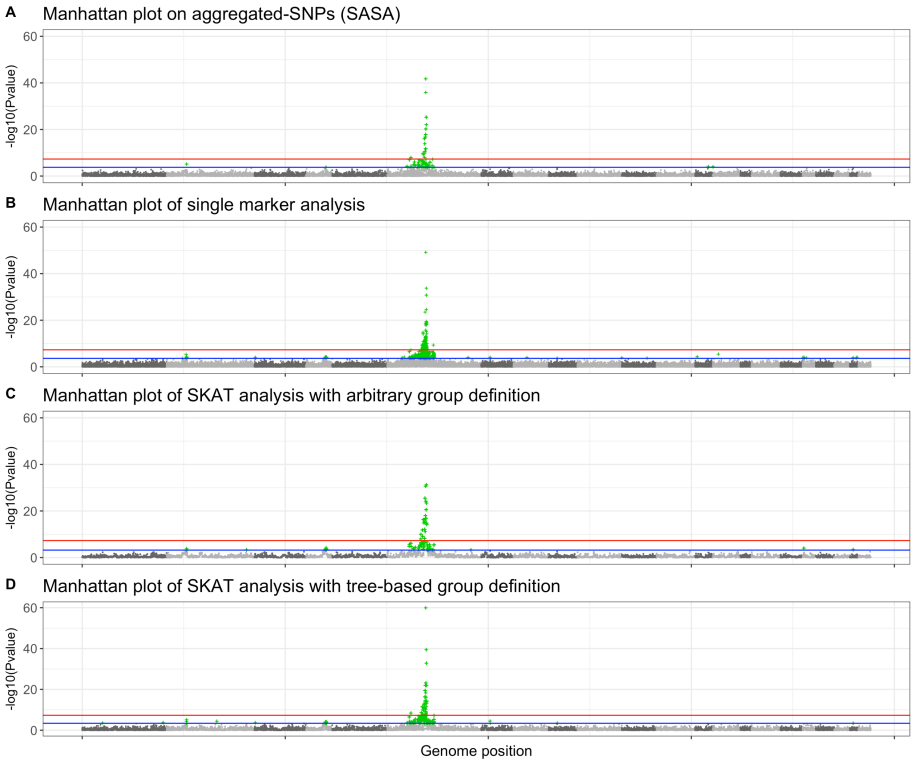


Figure 4.9: Manhattan plots showing results of GWAS analysis on ankylosing spondylitis data. For each Manhattan plot, the Benjamini-Hochberg (BH) threshold is represented by the blue line and the Bonferroni threshold by the red line. According to the BH threshold, there are: (A) 64 significantly associated aggregated-SNP; (B) 602 significantly associated single SNP; (C) 80 significantly associated groups of SNP and (D) 138 significantly associated groups of SNP.

associated with the phenotype compared to 602 significantly associated SNP with the standard SMA.

For the analysis of the WTCCC datasets, we represent the results, in Figure **??**, by plotting the expected p-value against the observed p-value. We perform the analysis using the approach SASA only.

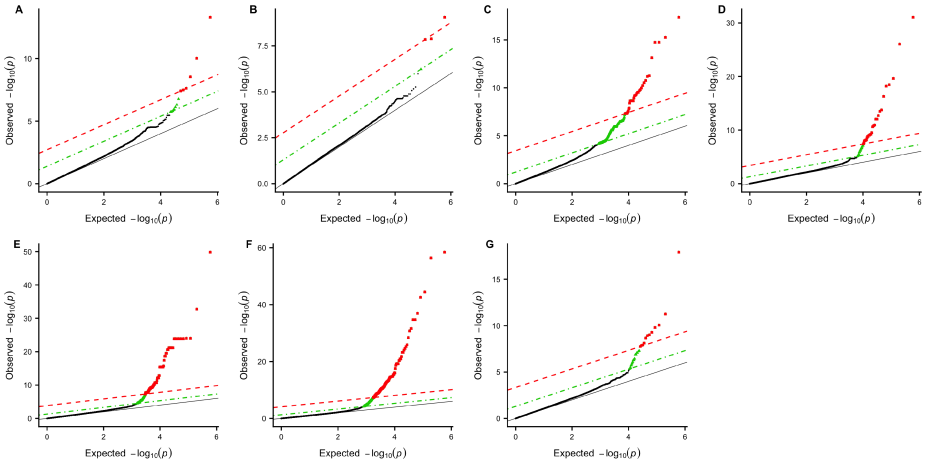


Figure 4.10: Q-Q plots of group-based genome-wide analysis on WTCCC data using the SASA approach. For each Manhattan plot, the Benjamini-Hochberg (BH) threshold is represented by the green dotted line and the Bonferroni threshold by the red dashed line. (A) Bipolar disorder - 13 significant clusters of SNP; (B) Coronary artery disease - 4 significant clusters of SNP; (C) Inflammatory bowel disease - 356 significant clusters of SNP ; (D) Hypertension - 47 significant clusters of SNP ; (E) Rheumatoid arthritis - 202 significant clusters of SNP ; (F) Type I diabetes - 358 significant clusters of SNP ; (G) Type II diabetes - 28 significant clusters of SNP.

4.5 Generalized additive models in GWAS

So far, we have modelled the phenotype as a linear function of the predictors using the logistic linear regression framework but we could also be interested to put forward non-linear relationships. One way to achieve this would be to use smoothing splines and generalized additive models (Section **??** and **??**). Implementation of such methods in classical GWAS is not straightforward because the predictors are ordinal variables which take at most three different values ($\{0, 1, 2\}$ with additive coding), making the use of smoothing splines irrelevant. However, in our context where we average the values of the SNP in each group, the use of smoothing splines becomes appropriate and may lead to the identification of non-linear otherwise undetectable with classical linear regression

framework. We can nevertheless qualify this statement by mentioning that the SKAT model may also be able to identify non-linear relationships with an appropriate choice of kernel function, e.g. the weighted quadratic kernel.

We chose to focus on smoothing splines rather than other functions because they are, in our opinion, more easily interpretable. More specifically we seek at first to investigate what benefit we could take from replacing the ridge regression model in Algorithm 4 by the high-dimensional additive models (HGAM) described in Section **??** to estimate the optimal number of groups. Secondly, we would like to highlight non-linear behaviour between groups of SNP and the phenotype by fitting cubic smoothing splines on each of the *aggregated*-SNP predictors constructed at best level in the hierarchy.

4.5.1 Comparison of predictive power

To evaluate the contribution of generalized additive models in our methodology, we compare the results in term of predictive power of four different regression model used to estimate an optimal number of clusters for the ankylosing spondylitis dataset. Specifically, we compare the AUC-ROC curves obtained from Algorithm 4 when using respectively lasso, group-lasso, HGAM and ridge regression as the learning method. The results are presented in Figure **??**. Note that for the group-lasso, the algorithm was applied at the single-SNP level rather than the *aggregated*-SNP.

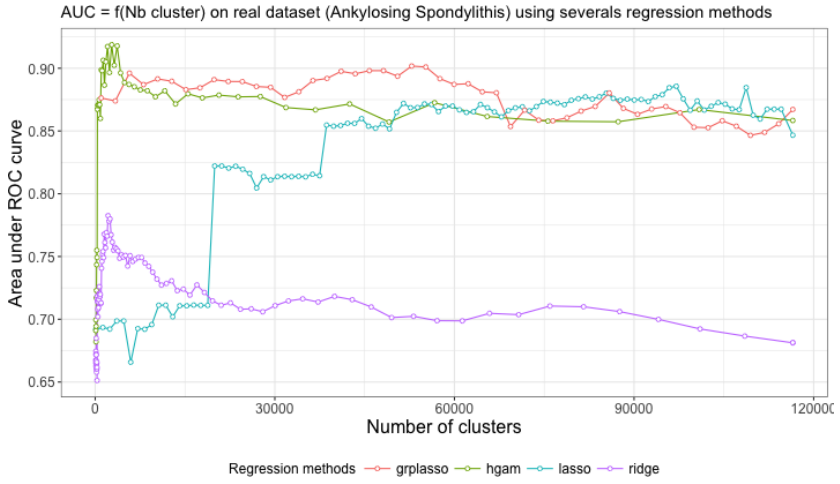


Figure 4.11: AUC-ROC plot illustrating the predictive power of four statistical learning approaches for several levels of a hierarchical clustering applied on the ankylosing spondylitis dataset.

We observe a similar pattern between the ridge regression curve and the HGAM curve, with about the same optimal number of clusters identified. The use

of cubic smoothing splines in the model greatly increase the predictive power compare to the others regression model. The group-lasso regression has a good predictive performance but is outperform by HGAM when we fit the model on the aggregated-SNP predictors at the best cut-level.

4.5.2 Results of univariate smoothing splines on aggregated-SNP

4.5.2.1 Manhattan plot

The best cut-level identified using high-dimensional additive model is set to 2750 aggregated-SNP. Firstly, to each of these variables, a univariate additive model using cubic smoothing splines with knots at each unique value is fitted. Secondly, we calculate p-value for each smooth as described in Section **??**. The results are shown in Figure **??**, where 23 significant aggregated-SNP have been identified.

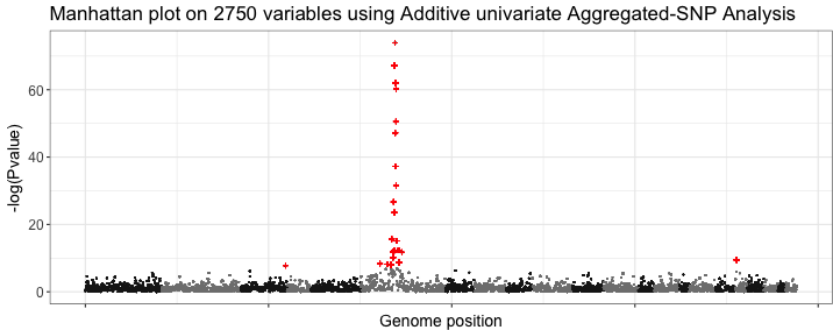


Figure 4.12: Manhattan plot of p-value calculated for 2750 aggregated-SNP using cubic smoothing splines.

4.5.2.2 Fitted values of the most significant aggregated-SNP

In this section, the plots in Figure **??** represent the fitted value of the 23 most significant aggregated-SNP variables previously identified. These aggregated-SNP are almost all located on the same region, on chromosome 6, region having already been identified as genetic risk factor for the disease (see Section **??**. We only observe a new signal on chromosome 18 which could be interesting to investigate.

We observe that the significant regions identified on chromosome 6 have a non-linear behaviour. However, since these regions have also been identified with a classical linear regression approach, we cannot conclude that we have detected

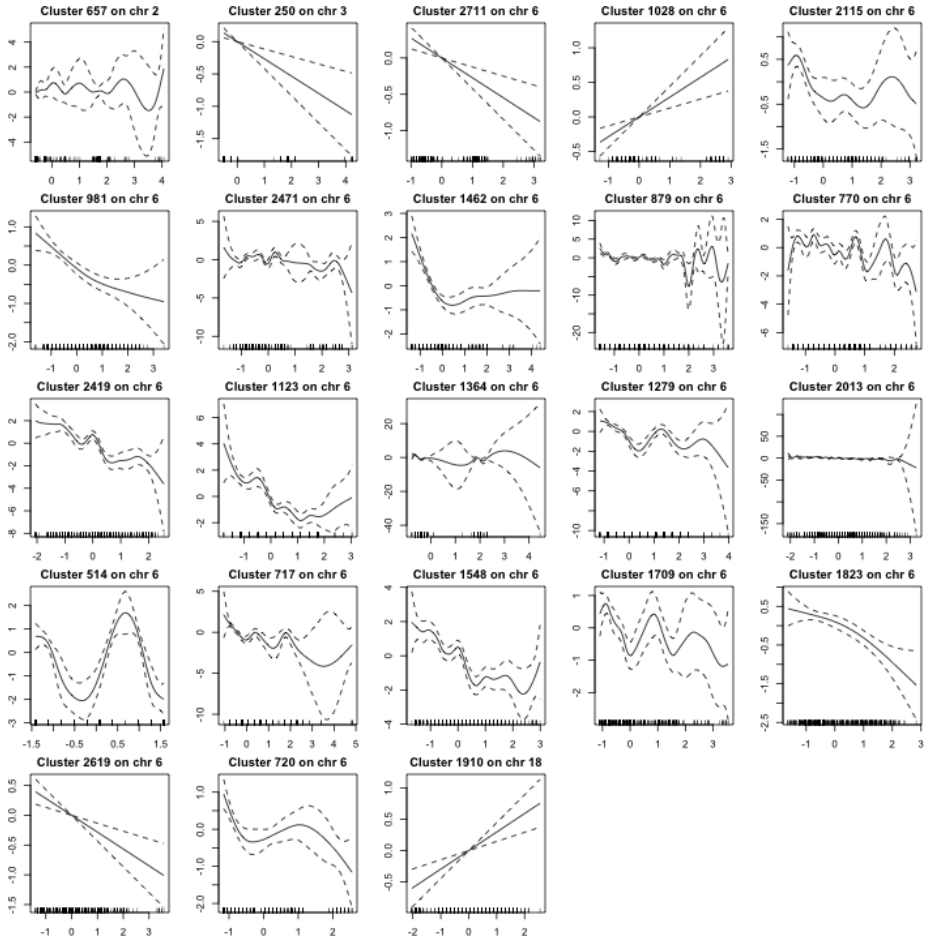


Figure 4.13: Representation of the smooth fits for the 23 most significant *aggregated*-SNP using cubic smoothing splines.

these regions thanks to the smoothing splines. Whether or not there is a non-linear behaviour, this region on chromosome 6 is always identified as associated with the disease. However, we could have thought that the new signal on chromosome 18 could be due to the non-linear nature of the association with the phenotype, but it is not the case. Indeed, if we look at the plot of the fitted value for this cluster, we can see that it is a straight line, leading to the conclusion that this signal might be a false positive.

4.6 Discussion

Overall, accounting for the linkage disequilibrium structure of the genome and aggregating highly-correlated SNP is seen to be a powerful alternative to standard marker analysis in the context of GWAS. In terms of risk prediction, our algorithm proves to be very effective at classifying individuals given their genotype, while in terms of the identification of loci, it shows its ability to identify genomic regions associated with a disease with a higher precision than standard methods.

It is also worth mentioning that our algorithm can also accommodate imputed variables as imputation in GWAS uses the linkage disequilibrium between variables to improve the coverage of variants. Our method being based on LD to define groups of common variants, we expect the group structure not to be impacted by imputation.

In this work we propose a four-step method explicitly designed to utilize the linkage disequilibrium in GWAS data. Our method combines, on the one hand, unsupervised learning methods that cluster correlated-SNP, and on the other hand, supervised learning techniques that identify the optimal number of clusters and reduce the dimension of the predictor matrix. We evaluated the method on numerical simulations and real datasets and compared the results with standard single-marker analysis and group-based approaches (SKAT*tree* and SKAT*notree*). We remarked that the combination of our aggregating function with a ridge regression model leads to a major improvement in terms of predictive power when the linkage disequilibrium structure is strong enough, hence suggesting the existence of multivariate effects due to the combination of several SNP. These results remained consistent across two applications involving several binary traits (WTCCC and ankylosing spondylitis datasets).

In terms of the identification of associated loci in different simulation scenarios, our method demonstrates its ability to retrieve true causal SNP and/or clusters of SNP with substantially higher precision coupled with a good power. On real GWAS data, our method has been able to recover a genomic region associated with ankylosing spondylitis (HLA region on chromosome 6) with a higher precision than standard single-marker analysis.

By making use of the continuous nature of aggregated-SNP variables (in contrast

to the ordinal nature of single SNP variables), we were able to further improve our method using generalized additive models and natural cubic splines. In terms of predictive power, the implementation of such models to the analysis of the AS data proved to be more efficient compared to linear regression models such as group-lasso, lasso and ridge regression. As for the detection of non-linear behaviour, the results obtained on the AS dataset show interesting non-linear patterns between some aggregated-SNP in the specific HLA region of chromosome 6 and the phenotype. However, the use of cubic splines has not been able to identify chromosome regions different from those previously identified with a classical linear regression model. It could thus be interesting to analyse other datasets with this methodology to see if we are able to detect any relevant associations ever identified before.

Chapter 5

Selection of interaction effects in compressed multiple omics representation

This chapter is organized as follows. Section **??** introduces the setting related to linear models of interactions and proposes a framework to learn with complementary datasets. Section **??** describes the method, entitled **SICOMORE** for **S**election of **I**nteraction effects in **CO**mpressed **M**ultiple **O**mics **RE**presentation, which combines hierarchical clustering, data compression, variable selection with a lasso procedure and model testing for recovering relevant interactions. Our approach is illustrated with numerical simulations in Section **??** and with an application to study the interactions between the genome of the species *Medicago truncatula* and the microbial community of its rhizosphere in Section **??**.

5.1 Introduction

5.1.1 Background

GWAS are a powerful tool for investigating the genetic architecture of complex diseases and have been successful in identifying hundreds of associated variants. However, they have been able to explain only a small proportion of the disease heritability calculated from classical family studies. As previously stated in Section **??**, it is nonetheless possible to uncover some of the missing heritability by

taking into account correlations among variables, interaction with the environment and epistasis, but not without some difficulties due to the multiple testing burden.

Other avenues to explain the variability in some traits of interest have yet to be explored, for instance an interesting lead would be to consider the contribution of microbial communities on the expression of a phenotype. Indeed, there is growing evidences of the role of gut microbiota in basic biological processes and in the development and progression of major human diseases such as infectious diseases, gastrointestinal cancers, metabolic diseases…(**?**). In plants, the role of rhizosphere[1] microflora on plant growth is well known and has been widely studied (**??**).

Analysis equivalent to GWAS have been conducted using the metagenome[2] rather than the genome of an individual and are known as Metagenome Wide Association Study (MWAS) (**??**). Those metagenome association analyses may often explain larger variation of the phenotype than classical GWAS and have been successful in finding relevant association for complex pathologies such as obesity, Crohn's disease, colorectal cancer…

5.1.2 Combining genome and metagenome analyses.

One possible way to relate genetic and metagenomic data consists in considering the metagenome as phenotype and thus performing quantitative trait locus (QTL) mapping. This kind of metagenome QTL analysis demonstrates the role of host genetics in shaping metagenomic diversity between individuals (**??**).

Another possibility for taking into account both type of variables consists in including metagenomic variables as environmental variables in GWAS. In that case interactions may naturally be modelled using a classical generalized linear model with interactions terms (**?**).

The main drawback of the later idea lies in the number of interactions to test, both datasets having a large number of variables. In order to reduce the dimension of the problem, variable selection or variable compression may be of use.

5.1.3 Taking structures into account in association studies.

Data compression for dimension reduction may be achieved in various ways. A usual distinction is often established between feature selection and feature extraction. Feature selection consists in selecting few relevant variables among the

[1] The rhizosphere is the term used to describe the zone of intense activity around the roots of leguminacea (Fabaceae) which contains a considerable diversity of microbial and mycorrhizal species.

[2] The metagenome corresponds to all the genetic material present in an environmental sample, consisting of the genomes of many individual organisms.

original ones, while feature extraction consists in computing new representative variables.

In our problem of association study, feature selection is often preferred to feature extraction for interpretative purposes. In this chapter, we advocate for a mixed approach which combines feature extraction and feature selection. The basic idea relies in grouping close variables via an unsupervised approach. Supervariables are computed to summarize the information of each cluster of variables and eventually the best supervariables are selected using a penalized regression approach.

We already investigate the idea of considering groups of variables in Chapter **??**. It also has already been suggested in the context of MWAS in (**?**). In the context of prediction from gene expression regression, the method HCAR developed by (**?**) described in Section **??** show that regressing over supergenes improves the precision if the correlation structure is strong enough. Moreover, (**?**) proposed a strategy to deal with large-dimension datasets in classification, called aggregation. It consists in a clustering step of redundant variables, using kNN or Classification and Regression Tree (CART) algorithms, and a group-compression step. They develop a statistical framework to define tailored aggregation methods that can be combined with selection methods to build reliable classifiers with possible applications on microarray data.

The method SICOMORE presented in this chapter can be summarized as follows: (1) it uses a hierarchical clustering algorithm to identify a group structure within the data; (2) it compresses the hierarchical structure by averaging the groups as in HCAR; (3) it performs a lasso procedure on the compressed variables as in HCAR with a penalty factor weighted by the length of the gap between two successive levels of the hierarchy as in MLGL; (4) it performs multiple hypothesis testing in a linear model with interactions.

5.2 Learning with complementary datasets

This section introduces the setting with the notations. It also sketches the approach to define a compact model of interactions between complementary datasets.

5.2.1 Setting and notations

Let us consider observations stemming from two complementary views, G (for Genomic data) and M (for Metagenomic data), which are gathered into a training set $\mathcal{T} = \{(\mathbf{x}_i^G, \mathbf{x}_i^M, y_i)\}_{i=1}^N$, where $(\mathbf{x}_i^G, \mathbf{x}_i^M, y_i) \in \mathbb{R}^{D_G} \times \mathbb{R}^{D_M} \times \mathbb{R}$.

We assume an underlying biological information on G and M encoded as groups. The group structure over G is defined by N_G groups of variables $\mathcal{G} = \{\mathcal{G}_g\}_{g=1}^{N_G}$. We denote $\mathbf{x}_i^g \in \mathbb{R}^{D_g}$, the sample i restricted to the variables of G from group

$\mathcal{G}_g$. Similarly, the group structure over M is defined by N_M groups of variables $\mathcal{M} = \{\mathcal{M}_m\}_{m=1}^{N_M}$ and $\mathbf{x}_i^m \in \mathbb{R}^{D_m}$ is the sample i restricted to the variables of M from group $\mathcal{M}_m$.

We also introduce $D_I = D_G \cdot D_M$ and $N_I = N_G \cdot N_M$, the number of variables and the number of groups that may interact.

Finally, we use the following convention: vectors of observations indexed with i, such as $\mathbf{x}_i$, will usually be row vectors[3] while vectors of coefficients, such as β, will usually be column vectors.

5.2.2 Interactions in linear models

Interactions between data stemming from views G and M may be captured in the model

$$y_i = \mathbf{x}_i^G \gamma_G + \mathbf{x}_i^M \gamma_M + \mathbf{x}_i^G \Delta_{GM} (\mathbf{x}_i^M)^T + \epsilon_i \,,$$

where the vectors $\gamma_G \in \mathbb{R}^{D_G}$ and $\gamma_M \in \mathbb{R}^{D_M}$ respectively denote the linear effects related to G and M, the matrix $\Delta_{GM} \in \mathbb{R}^{D_G \times D_M}$ contains the interactions between all pairs of variables of G and M and $\epsilon_i \in \mathbb{R}$ is a residual error.

Underlying notions in models of interactions are the one of *strong dependency* (SD) and *weak dependency* (WD), the first one being more common (see for instance (**?**) and the discussion therein). Under the hypothesis of *strong dependency*, an interaction is effective if and only if the corresponding single effects are also effective while the hypothesis of *weak dependency* implies that an interaction is effective if one of the main effect is also effective. Formally, for all variables $j \in \mathbf{x}^G$ and for all variables $j' \in \mathbf{x}^M$, if γ_j, $\gamma_{j'}$ and $\delta_{jj'}$ are the coefficients related to γ_G, γ_M and Δ_{GM}, then

$$\begin{array}{lllll} (SD) & \delta_{jj'} \neq 0 & \Rightarrow & \gamma_j \neq 0 \quad \text{and} & \gamma_{j'} \neq 0\,, \\ (WD) & \delta_{jj'} \neq 0 & \Rightarrow & \gamma_j \neq 0 \quad \text{or} & \gamma_{j'} \neq 0\,. \end{array}$$

In this context, (**?**) have proposed a sparse model of interactions which faces computational limitations for large dimensional problems according to (**?**) and (**?**). While (**?**) introduce a method for learning pairwise interactions in a regression model by solving a constrained overlapping Group-Lasso (**?**) in a manner that satisfies strong dependencies, (**?**) propose a formulation with an overlapping regularization that fits both kind of hypotheses and provide theoretical insights on the resulting estimators[4].

[3] For the sake of clarity, we use these lightened notations which slightly differ from those used in the previous chapters.

[4] To our knowledge, their implementation based on alternating direction method of multipliers is not publicly available.

Yet, the dimension $D_G + D_M + D_I$ involved in Problem to estimate γ_G, γ_M and Δ_{GM} may be large especially for applications with an important number of variables such as in biology with genomic and metagenomic data. To reduce the dimension, we propose to compress the data according to an underlying structure which may be defined thanks to a prior knowledge or be uncovered with clustering algorithms.

5.2.3 Compact model

Assuming we are given a compression function for each group of G and M, we can shape Problem into a compact form

$$y_i = \sum_{g\in\mathcal{G}} \tilde{x}_i^g \beta_g + \sum_{m\in\mathcal{M}} \tilde{x}_i^m \beta_m + \sum_{g\in\mathcal{G}}\sum_{m\in\mathcal{M}} \underbrace{(\tilde{x}_i^g \cdot \tilde{x}_i^m)}_{\phi_i^{gm}} \theta_{gm} + \epsilon_i \,,$$

where $\tilde{x}_i^g \in \mathbb{R}$ is the i^{th} compressed sample of the variables that belong to the group g for the view G and $\beta_g \in \mathbb{R}$ is its corresponding coefficient. The counterparts on the group m for the view M are $\tilde{x}_i^m \in \mathbb{R}$ and $\beta_m \in \mathbb{R}$. Finally, $\theta_{gm} \in \mathbb{R}$ is the interaction between groups g and m.

We can reformulate Problem in a vector form. Let $\tilde{\mathbf{x}}_i^G \in \mathbb{R}^{N_G}$, $\beta_G \in \mathbb{R}^{N_G}$, $\tilde{\mathbf{x}}_i^M \in \mathbb{R}^{N_M}$ and $\beta_M \in \mathbb{R}^{N_M}$ be

$$\begin{aligned}
\tilde{\mathbf{x}}_i^G &= (\tilde{x}_i^1 \cdots \tilde{x}_i^g \cdots \tilde{x}_i^{N_G}) \,, \qquad \beta_G = (\beta_1 \cdots \beta_g \cdots \beta_{N_G})^T \,, \\
\tilde{\mathbf{x}}_i^M &= (\tilde{x}_i^1 \cdots \tilde{x}_i^m \cdots \tilde{x}_i^{N_M}) \,, \quad \beta_M = (\beta_1 \cdots \beta_m \cdots \beta_{N_M})^T \,.
\end{aligned}$$

We denote by $\phi_i \in \mathbb{R}^{N_I}$, the vector whose general component is given by ϕ_i^{gm} in Equation , that is

$$\phi_i = \left(\phi_i^{11} \cdots \phi_i^{1N_M} \cdots \phi_i^{gm} \cdots \phi_i^{N_G 1} \cdots \phi_i^{N_G N_M}\right) \,,$$

and $\theta \in \mathbb{R}^{N_I}$, the corresponding vector of coefficients, by

$$\theta = \left(\theta_{11} \cdots \theta_{1N_M} \cdots \theta_{gm} \cdots \theta_{N_G 1} \cdots \theta_{N_G N_M}\right)^T \,.$$

Finally, Problem reads as a classical linear regression problem

$$y_i = \tilde{\mathbf{x}}_i^G \beta_G + \tilde{\mathbf{x}}_i^M \beta_M + \phi_i \theta + \epsilon_i \,,$$

of dimension $N_G + N_M + N_I$.

5.2.4 Recovering relevant interactions

Compared to Problem and provided that N_G and N_M are reasonably lower than D_G and D_M, the dimension of Problem decreases drastically so that

it might be solved thanks to an appropriate optimization algorithm coupled with effective computational facilities. For instance, (**?**) give an overview of ℓ_1 regularized algorithms to solve sparse problems like Lasso, which in our case could take the form:

$$\underset{\beta_G,\beta_M,\theta}{\operatorname{argmin}}, \quad \sum_{i=1}^{n} \left(y_i - \tilde{\mathbf{x}}_i^G \beta_G - \tilde{\mathbf{x}}_i^M \beta_M - \phi_i \theta\right)^2 + \lambda_G \sum_{g=1}^{N_G} |\beta_g| + \lambda_M \sum_{m=1}^{N_M} |\beta_m| + \lambda_I \sum_{g,m=1}^{N_I} |\theta_{gm}|,$$

with λ_G, λ_M and λ_I being the positive hyperparameters that respectively control the amount of sparsity related to coefficients β_G, β_M and θ. Still, the dimension may remain large regarding the dimension $N_G + N_M + N_I$ compared to the number of observations N. Also, note that without additional constraints, such a formulation would not induce the dependences hypothesis (SD) and (WD). For that purpose, one could adapt the works of (**??**) or (**?**) mentioned above. We present in the next Section another way to reduce further the dimension and fulfil the strong dependency hypothesis.

5.3 Method

In this section, we provide some elements to enhance Problem for biological problems involving metagenomic and genomic data. After a brief discussion related to the preprocessing of the data, we explain how to obtain the group structure on G and M using hierarchical clustering strategies and describe how to efficiently take into account the different scales of the groups defined by each level of the hierarchies. We then present some compressions that may be used to summarize the groups. Finally, we propose a linear model testing to recover the relevant interactions.

We entitled the proposed approach SICOMORE for Selection of Interaction effects in COmpressed Multiple Omics REpresentations, implemented in the R package `SICoMORe` available at https://github.com/fguinot/sicomore-pkg.

5.3.1 Preprocessing of the data

To tackle problems that involve genomic and metagenomic interactions, some prior transformations are mandatory. Also, a first attempt to reduce the dimension may be achieved at this step.

5.3.2 Preprocessing of metagenomic data

5.3.2.1 Normalization.

In shotgun metagenomics,[5] microorganisms are studied by sequencing DNA fragments directly from samples, without the need for cultivation of individual isolates (**?**).

Shotgun metagenomic sequencing data are often produced by analysing the presence of genes and their abundances in and between samples from different experimental conditions. The gene abundances are then estimated by matching each generated sequence read against a comprehensive and annotated reference database and by counting the number of reads matching each gene in the reference database (**?**).

Gene abundance data generated by such analysis are however affected by systematic variability that significantly increases the variation between samples and thereby decrease the ability to identify genes that differ in abundance (**???**). The process known as normalisation therefore referred to the methods designed to remove such systematic variability.

A wide range of different methods has been applied to normalize shotgun metagenomic data. The majority of these normalization methods are based on scaling, where a sample-specific factor is estimated and then used to correct the gene abundances. For instance, one can simply calculates the scaling factor ψ_i as the sum of all reads counts in a sample i:

$$\psi_i = \sum_{j=1}^{D_M} x_{ij}^M.$$

A more robust method to estimate the scaling factor is the Trimmed Mean of M-values (TMM) (**?**) which compares the gene abundances in the samples against a reference, typically set as one of the samples in the study. We note t_i the scaling normalization factors for raw library sizes calculated using the TMM normalization method for sample i; $l_i = \mathbf{x}_i^M t_i$ is then the corresponding normalized library size for sample i and

$$\psi_i = \frac{l_i}{\sum_{t=1}^{n} l_t/n},$$

is the associated normalization scaling factor.

The raw counts x_{ij}^M, with $j \in [1, \dots, D_M]$, are then divided by the scaling factor to obtain the normalized counts:

$$\tilde{x}_{ij}^M = x_{ij}^M / \psi_i.$$

[5] see Section **??** for details on shotgun sequencing.

5.3.2.2 Transformation.

Metagenomic shotgun sequencing results in features which take the form of proportions in different samples, referred in the statistical literature as compositional data (**?**). These data are known to be subject to negative correlation bias (**??**)) and the assumption of conditional independence among samples is unlikely to be true for the vast majority of metagenomic datasets.

Several data transformations have been suggested for RNA-seq data, most often in the context of exploratory or differential analyses. These transformations include log transformation (where a small constant is typically added to read counts to avoid 0's), a variance-stabilizing transformation (**??**), moderated log counts per million (**?**), and a regularized log-transformation (**?**).

(**?**) also proposed to calculate normalized expression *profiles* for each feature, that is, the proportion of normalized reads observed for gene j with respect to the sum of all samples in gene j:

$$p_{ij} = \frac{\tilde{x}_{ij}^{M} + 1}{\sum_{t=1}^{n} \tilde{x}_{tj}^{M} + 1},$$

where a constant of 1 is added to the numerator and denominator due to the presence of 0 counts.

However, the vector of values $\mathbf{p}_j$ are linearly dependent, which imposes constraints on the covariance matrices that can be problematic for most standard statistical approaches (**?**). One solution is to apply the commonly used centered log ratio (CLR) transformation for compositional data (**?**). It is defined as:

$$\mathrm{CLR}(\mathbf{p}_j) = \left[\ln\left(\frac{p_{1j}}{g(\mathbf{p}_j)}\right), \ldots, \ln\left(\frac{p_{nj}}{g(\mathbf{p}_j)}\right) \right],$$

where $g(\mathbf{p}_j)$ is the geometric mean of $\mathbf{p}_j$.

5.3.2.3 A first selection of variables

As seen in Section **??**, we assume strong dependencies on interactions, which means that an interaction can be effective only if the two simple effects making up the interaction are involved in the problem. Then, it may be clever to apply a first process of selection to discard the inoperative single effects on G and M respectively. Different approaches may be envisioned to proceed this selection. Among them, screening rules can eliminate variables that will not contribute to the optimal solution of a sparse problem sweeping all the variables upstream to the optimization. When such a screening is appropriate, we may use the work of (**?**) focused on Lasso problems, which present a recent overview of these techniques together with a screening rule ensemble. Once the screening is done, the optimization of a Lasso problem gives the final set of variables.

5.3.3 Structuring the data

Once the data are preprocessed, we can resort to hierarchical clustering using Ward criterion with appropriate distances to uncover the tree structures.

5.3.3.1 Clustering of metagenomic data

A common approach to analyse metagenomic data is to group sequences into taxonomic units. The features stemming from metagenome sequencing are often modelled as Operational Taxonomic Units (OTU), each OTU representing a biological species according to some degree of taxonomic similarity. (**?**) propose a comparison of methods to identify OTU that includes hierarchical clustering. While the structure on microbial species could be defined according to the underlying phylogenetic tree, it also makes sense to use more classical distances, such as the Ward criterion, to define a hierarchy based on the abundances of OTU. In our application we use an agglomerative hierarchical clustering with the Ward criterion (see Section **??**).

5.3.3.2 Clustering of genomic markers

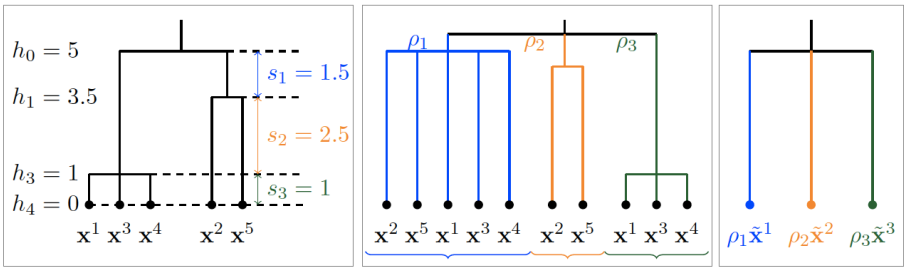
On the other hand, when the genomic information is available through SNP, the tree structure on G will be defined using a Ward spatially constrained hierarchical clustering algorithm which integrates the linkage disequilibrium as the measure of dissimilarity using the *adjclust* algorithm (**?**).

5.3.4 Using the structure efficiently

Different approaches related to the problem of finding an optimal number of clusters may be envisioned to find the optimal cut in a tree structure obtained with hierarchical clustering (see for instance (**?**) or (**?**)). Whatever the chosen approach, a systematic exploration of different levels of the hierarchy is mandatory to find this optimal cut. We define an alternative strategy to bypass this expensive exploration which consists in:

1. (a) Expanding the hierarchy considering all possible groups at a single level;
2. (b) Assigning a weight to each group based on gap distances between two consecutive groups in the hierarchy;
3. (b) Compressing each group into a supervariable.

The different steps of this strategy are illustrated in Figure **??**, from the original tree structure in Figure **??**(a) to a final flatten, weighted and compressed representation in Figure **??**(c).



(a) Original hierarchy (b) Expanded representation (c) Compressed representation

Figure 5.1: Dimension reduction strategy. (a) Original hierarchical tree with an example for 5 variables. (b) Expanded representation of the tree with all possible weighted groups derived from the original hierarchy. (c) Compressed representation of the tree after construction of the supervariables.

Expanding the hierarchy

To reduce the dimension involved in Problem , the first step consists in flattening the respective tree structures obtained on views G and M so that only a group structure remains. Thus, each group of variables defined at the deepest level may be included in other groups of larger scales, as shown in Figure **??**(b).

Assigning weights to the groups

To keep track of the tree structure, we may integrate an additional measure quantifying the quality of the groups on two successive levels. More specifically, for a tree structure of height H and for $1 \leq h \leq H-1$, we define s_h as the gap between heights s_h and s_{h-1}. Following the lines of (**?**) for the Multi-Layer Group-Lasso described in Section **??**, we define this quantity as $\rho_h = 1/\sqrt{s_h}$. The process is shown in Figure **??**(a) and **??**(c).

Compressing the data

To summarize each group of variables, the mean, the median or other quantiles may be used as well as more sophisticated representations based on eigen values decompositions such as the first factor obtained with a PCA. This step is similar to the dimension reduction step of the method presented in Section **??**.

5.3.5 Identification of relevant supervariables

With this compressed representation at hand, we can recover relevant interactions with a multiple testing strategy.

Selection of supervariables

The compression is a key ingredient to reduce significantly the dimension involved in Problem . Yet, we are going a step further with an additional feature selection process applied to the compressed variables, as suggested at the begin of this section to preprocess the data, using screening rules and / or applying a Lasso optimization on each view G and M:

$$\underset{\beta_G}{argmin} \sum_{i=1}^{n} \left(y_i - \tilde{\mathbf{x}}_i^G \beta_G\right)^2 + \lambda_G \sum_{g=1}^{N_G} \rho_g |\beta_g|,$$

and

$$\underset{\beta_M}{argmin} \sum_{i=1}^{n} \left(y_i - \tilde{\mathbf{x}}_i^M \beta_M\right)^2 + \lambda_M \sum_{m=1}^{N_M} \rho_m |\beta_m|,$$

with penalty factors being defined by $\rho_g = 1/\sqrt{s_g}$ and $\rho_m = 1/\sqrt{s_m}$ as explained in Section **??**.

Linear model testing

In a feature selection perspective, the relevant interactions may be recovered separately considering each selected group $g \in \mathcal{G}$ coupled with each selected group $m \in \mathcal{M}$ in a linear model of interaction and by performing an hypothesis test on the interaction parameter:

$$y_i = \tilde{x}_i^g \beta_g + \tilde{x}_i^m \beta_m + (\tilde{x}_i^g \cdot \tilde{x}_i^m)\, \theta_{gm} + \epsilon_i \,.$$

This strategy has the advantage of highlighting all the potential interactions between the selected simple effects in an exploratory rather than predictive analysis perspective. Also, it may be regarded as an alternative shortcut to Problem in that it involves N_I problems of dimension 3 instead of a potentially large problem of dimension $N_G + N_M + N_I$. Finally, this scheme of selection preserves strong dependencies by construction.

5.4 Numerical simulations

We provide here numerical simulations to assess the ability of SICOMORE to recover relevant interactions against three other methods. We also show that our method is computationally competitive compared to the others.

5.4.1 Data generation

Generation of metagenomic and genomic data matrices

5.4.1.0.1 Genomic data

In order to get a matrix $\mathbf{x}^G$ close to real genomic data, we used the software software (**?**). This software allows to simulate an entire chromosome conditionally on a reference set of population haplotypes (from HapMap3) and an estimate of the fine-scale recombination rate across the region, so that the simulated data share similar patterns with the reference data. We generate the chromosome 1 using the haplotype structure of CEU population (Utah residents with Northern and Western European ancestry from the CEPH collection) as reference set and we selected $D_G = 200$ variables from this matrix to obtain our simulated dataset. An example of the linkage disequilibrium structure among the simulated SNP is illustrated in Figure **??**(a).

5.4.1.0.2 Metagenomic data

The data matrix $\mathbf{x}^M$, with $D_M = 100$ variables, has been generated using a multivariate Poisson-log normal distribution (**?**) with block structure dependencies.

The Poisson-log normal model is a latent gaussian model where latent vectors $\mathcal{L}_i \in \mathbb{R}^{D_M}$ are drawn from a multivariate normal distribution

$$\mathcal{L}_i \sim \mathcal{N}_{D_M}(0, \Sigma),$$

where Σ is a covariance matrix that allows to obtain a correlation structure among the variables.

The centered phenotypic count data Y_i are then drawn from a Poisson distribution conditionally on $\mathcal{L}_i$

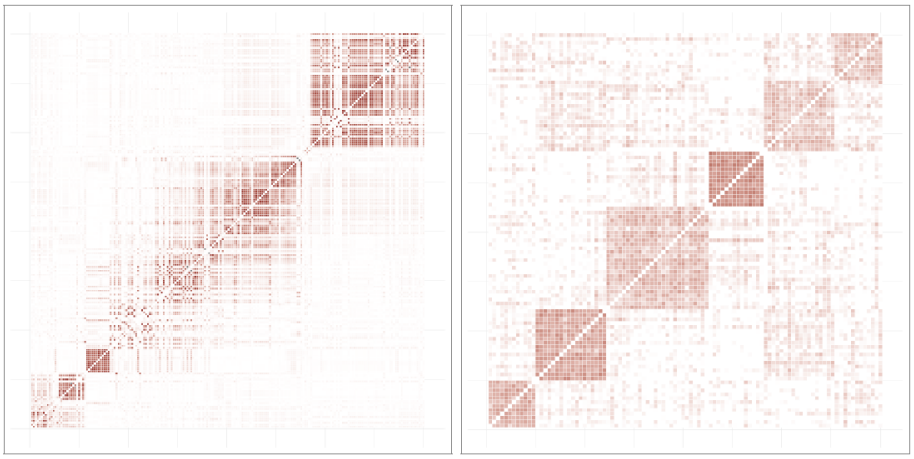
$$\mathrm{Y}_{ij} | \mathcal{L}_{ij} \sim \mathcal{P}\left(e^{\mu_j + \mathcal{L}_{ij}}\right),$$

with $\mu_j = 0$.

The block structure, pictured in Figure **??**(b), has been obtained by drawing a latent multivariate normal vector using a covariance matrix Σ such that the correlation level between the latent variables of a group are between 0.5 and 0.95. By simulating this way, we obtain a matrix of count data with a covariance structure close to what is observed with metagenomic data. As stated in Section **??**, we calculated the proportions in each random variable and transformed them using centered log-ratios.

5.4.2 Generation of the phenotype

For all simulations, we used a fixed value of $N_M = 6$ groups for the matrix $\mathbf{x}^M$ and for the case of the matrix $\mathbf{x}^G$, since we cannot exactly control the



(a) Correlation matrix of $\mathbf{X}^G$ (b) Correlation matrix of $\mathbf{X}^M$

Figure 5.2: Examples of hierarchical structures}: correlations observed on (a) genomic data $\mathbf{x}^G$ and (b) metagenomic data $\mathbf{x}^M$.

block structure with , we used the Gap Statistic (see Section **??**) to identify a number of groups in the hierarchy. For instance, in Figure **??**(a), the Gap Statistic identified $N_G = 16$ groups. The supervariables were then calculated using averaged groups of variables to obtain the two matrices of supervariables, $\tilde{\mathbf{x}}^G$ and $\tilde{\mathbf{x}}^M$.

To generate the phenotype, we considered a data structure for which the data to regress has been generated using supervariables according a linear model with interactions of the form:

$$y_i = \sum_{g \in \mathcal{S}^G} \tilde{x}_i^g \beta_g + \sum_{m \in \mathcal{S}^M} \tilde{x}_i^m \beta_m + \sum_{g \in \mathcal{S}^G} \sum_{m \in \mathcal{S}^M} \underbrace{(\tilde{x}_i^g \cdot \tilde{x}_i^m)}_{\phi_i^{gm}} \theta_{gm} + \epsilon_i \,,$$

where $\mathcal{S}^G$ and $\mathcal{S}^M$ are subsets of randomly chosen effects from the matrices $\tilde{\mathbf{x}}^G$ and $\tilde{\mathbf{x}}^M$ respectively, $\tilde{x}_i^g$ is the i^{th} sample of the g effect and β_g its corresponding coefficient, $\tilde{x}_i^m$ is the i^{th} sample of the m effect and β_m its corresponding coefficient. Finally, θ_{gm} is the interaction between variables $\tilde{x}_i^g$ and $\tilde{x}_i^m$.

We considered $I \in \{1, 3, 5, 7, 10\}$ true interactions between the supervariables to generate the phenotype so that I blocks of the coefficients of θ_{gm} have non-zero values. The process was repeated 30 times for each couple of parameters in $N = \{50, 100, 200\} \times mean(\epsilon) = \{0.5, 1, 2\}$.

5.4.3 Comparison of methods

To evaluate the performance of our method, SICOMORE, to retrieve the true causal interaction, we compared it with three other methods, namely **HCAR** (**?**), **MLGL** (**?**) and **glinternet** (**?**). It is worth mentioning that, as we already stated, SICOMORE is an approach that borrow from HCAR and MLGL and that is designed to detect interactions. We had then to adapt these approaches to our problematic, as we will describe it in the following sections, they are therefore not evaluated in the context they were meant to be used. Thus, the purpose of this evaluation is to know if SICOMORE is capable of improving the individual performance of these methods by combining them to detect statistical interactions.

Hierarchical Clustering and Averaging Regression (HCAR)

This methodology can be simply adapted to our problematic by performing two hierarchical clustering on each data matrix $\mathbf{x}^G$ and $\mathbf{x}^M$ and then compute the unweighted compressed representations of those hierarchies as explained in Section **??** and illustrated in Figure **??**(c). We can then fit a Lasso regression model on both compressed representations with interactions between all possible groups. We consider that HCAR is able to retrieve a true causal interaction if the Lasso procedure selects the interaction term at the correct levels of the two hierarchies.

Multi-Layer Group-Lasso (MLGL)

The model is fitted with weights on the groups defined by the expanded representation of the two hierarchies as illustrated in Figure **??**(b). This method does not work on the compressed supervariables but on the initial variables. Our evaluation considers that the method is able to retrieve real interactions if it selects the correct interaction terms between two groups of variables at the right level in both hierarchies.

5.4.3.1 Group-Lasso interaction network (glinternet)

glinternet (**?**) is a procedure that considers pairwise interactions in a linear model in a manner that satisfies strong dependencies between main and interaction effects: whenever an interaction is estimated to be non-zero, both its associated main effects are also included in the model. This method uses a Group-Lasso model to accommodate with categorical variables and apply constraints on the main effects and interactions to result in interpretable interaction models.

The glinternet model fits a hierarchical group-lasso model with constraints on the main and interactions effects as specified in the equation whilst accommodating for the strong dependence hypothesis by adding an appropriate penalty to the loss function (we refer the reader to (**?**) for more details on the form of the penalty). For very large problems (with a number of variables $\geq 1.10^5$), the group-lasso procedure is preceded by a screening step that gives a candidate set of main effects and interactions. They use an adaptive procedure that is based on the strong rules (**?**) for discarding predictors in lasso-type problems.

Since this method can only work at the level of variables, it was necessary to include a group structure into the analysis. Therefore, we decided to fit the glinternet model on the compressed variables and to constraint the model to only fit the interaction terms between the supervariables of the two matrices $\tilde{\mathbf{x}}^G$ and $\tilde{\mathbf{x}}^M$. We explicitly removed all interaction terms between supervariables belonging to the same data matrix.

For a fair comparison with the other methods, we considered two options namely *GLtree* and *GLgap*. On one hand, option GLtree works on the unweighted compressed representations of the two hierarchies (Figure **??**(c)) thus considering all the possible interactions between the supervariables of the two datasets. On the other hand, option GLgap considers only the interactions between the compressed variables constructed at a specific level in the hierarchies, chosen by the Gap Statistic.

Given that D^G and D^M are the number of variables in $\mathbf{x}^G$ and $\mathbf{x}^M$, the dimension of the compressed matrices $\tilde{\mathbf{x}}^G$ and $\tilde{\mathbf{x}}^M$ are respectively $\tilde{D}^G = D^G + (D^G - 1)$ and $\tilde{D}^M = D^M + (D^M - 1)$. Thus, for GLtree the number of interactions to investigate are $\tilde{D}^G \times \tilde{D}^M$ while for GLgap this number will depend on the level chosen by the Gap statistic but will be either way smaller since we consider only a specific level of the hierarchy in this option. In the numerical simulations, given that $D^G = 200$ and $D^M = 100$, the use of strong rules to discard variables is therefore not necessary as (**?**) argued that glinternet can handle very large problems without any screening (360M candidate interactions were fitted when evaluating the method on real data examples).

5.4.4 Evaluation metrics

For each run, we evaluated the quality of the variable selection using Precision and Recall. More precisely, we compared the true interaction matrix θ that we used to generate the phenotype with the estimated interaction matrix $\hat{\theta}$ compute for each model. For all possible interactions $\{gm\}$, we then determined the following confusion matrix:

	$\hat{\theta}_{gm} = 0$	$\hat{\theta}_{gm} \neq 0$
$\theta_{gm} = 0$	True Negative	False Positive
$\theta_{gm} \neq 0$	False Negative	True Positive

Confusion matrix for the hypothesis test on interaction parameter θ

and hence compute Precision $= \frac{TP}{(FP+TP)}$ and Recall $= \frac{TP}{FN+TP}$. In this context, a true positive corresponds to a significant p-value on a true causal interaction, a false positive to a significant p-value on a noise interaction, and a false negative to a non-significant p-value on a true causal interaction. An example of the interaction matrix θ is given in Figure **??** for $I = 5$ blocks in interaction.

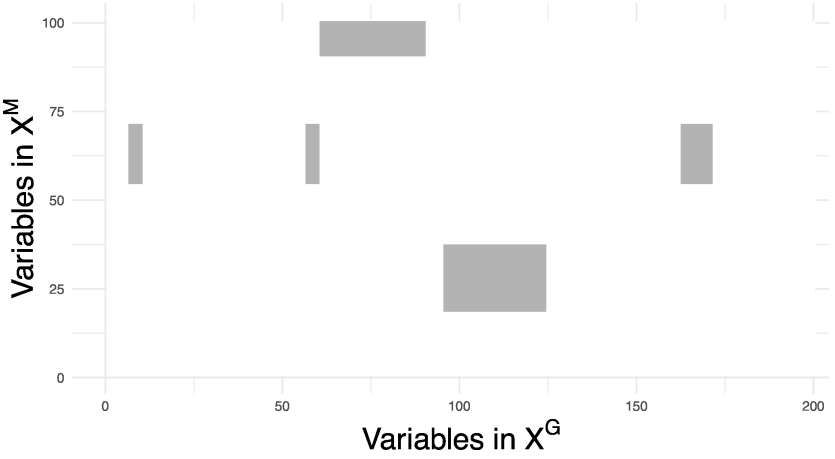


Figure 5.3: Illustration of the true interaction matrix θ with $I = 5$, $\sigma = 0.5$ and $n = 100$. Each non-zero value in this matrix is considered as a true interaction between two variables.

For all methods, we correct for multiple testing by controlling the Family Wise Error Rate using the method of Holm-Bonferroni. Even though it is known to be stringent, we chose to rely on Holm-Bonferroni method to adjust for multiple testing because the number hypothesis tests performed in our simulation context is not that high. In a high-dimensional context such as with the analysis of real DNA chip data, we would rather use the Benjamini-Hochberg method for the control of the false discovery rate.

5.4.5 Performance results

The performances of each method to retrieve the true causal interactions are illustrated in Figure **??** for precision and Figure **??** for recall. For the sake of clarity we only show the results for $I = 7$ blocks of variables in interaction.

The results in terms of recall reveal good abilities of MLGL and SICOMORE to retrieve True Positive interactions, with an overall advantage for our method. HCAR achieves a lower performance due to the fact that it favours the selection of small groups which are only partly contained in the groups that generate the interactions showing that the weighting scheme of MLGL and SICOMORE is efficient. GLgap is not able to retrieve relevant interactions but the way to define the structure among variables, using the Gap Statistic, is also quite different than for the three other methods.

In terms of precision, all methods perform poorly with a significant number of false positive interactions. MLGL and SICOMORE tend to select groups of variables and supervariables too high in the tree structure, inducing false positives which are spatially closed to the true interactions. HCAR, which favours small groups as explained above, is less subject to that. The behaviour of GLgap may vary according to the selected cut with the Gap statistic into the tree structure while option GLtree exhibit slightly better precision. Still, the method glinternet is globally not able to retrieve correctly the true interactions whether or not it uses the compressed or original representation. The plots in Figure **??** represent the recovered confusion matrices of interaction θ_{gm} for each compared algorithm for one particular set of simulation parameters ($I = 5$, $\sigma = 0.5$, $n = 100$).

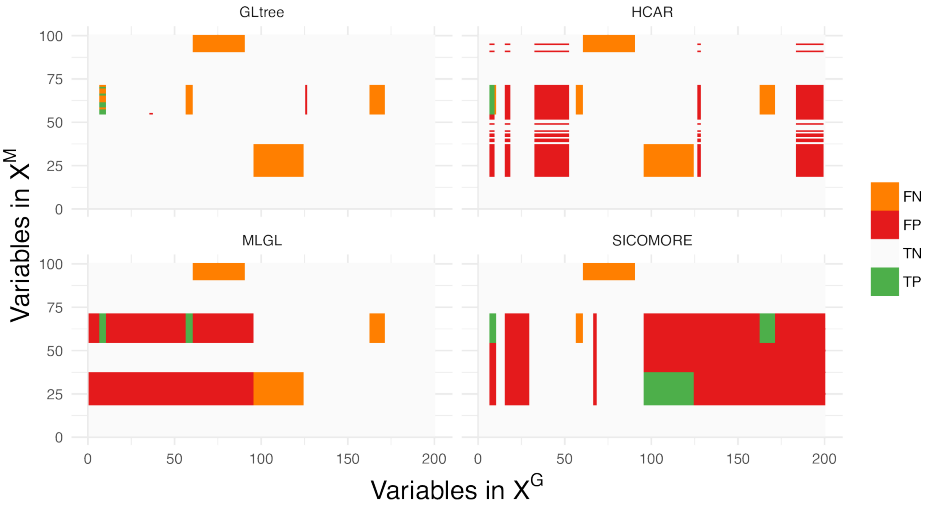


Figure 5.4: Confusion matrices of interactions θ_{gm} for each compared algorithm with the following simulation parameters: $I = 5$, $\sigma = 0.5$, $n = 100$. We can see in this example that MLGL and SICOMORE behaves similarly with very large genomic regions identified while HCAR tends to work with smaller genomic and metagenomic regions.

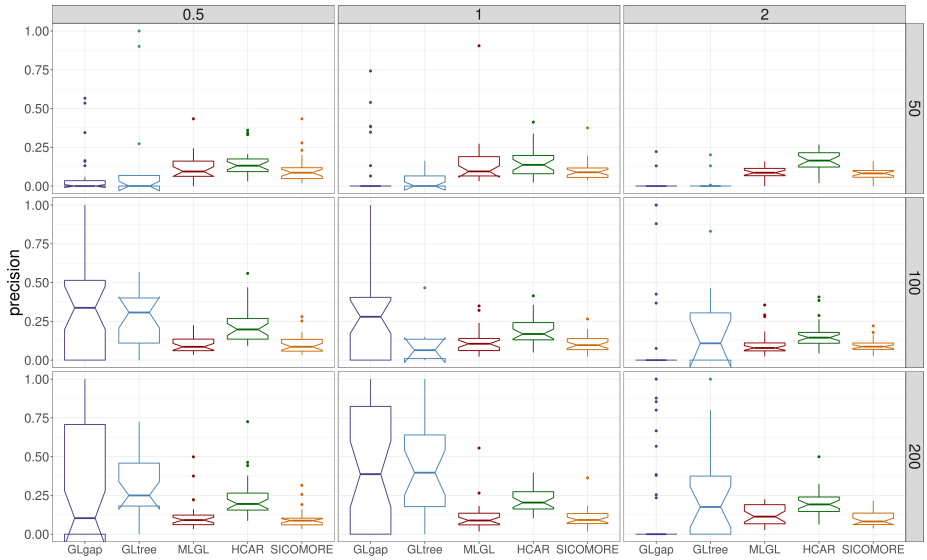


Figure 5.5: Boxplots of precision for each couple of parameters N (number of examples in rows) and ϵ (difficulty of the problem in columns) for $I = 7$.

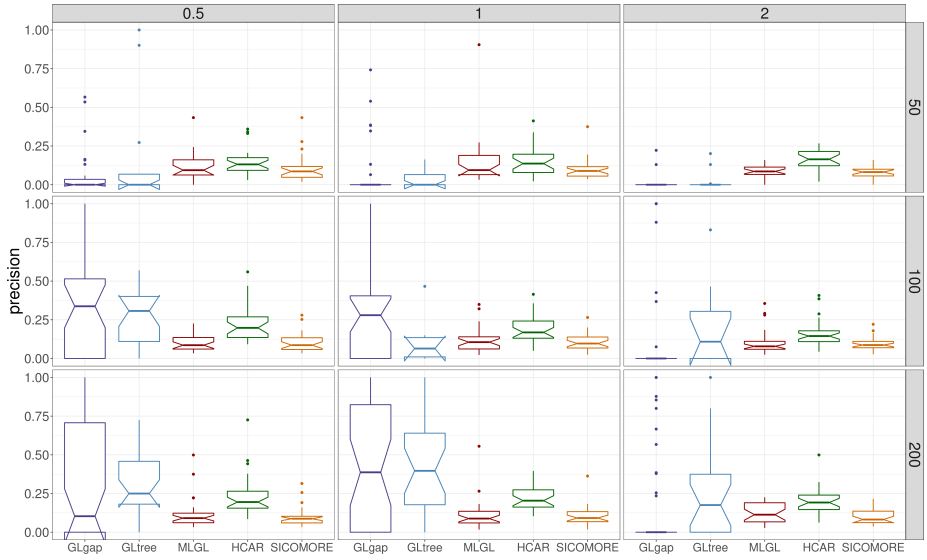


Figure 5.6: Boxplots of recall for each couple of parameters N (in rows) and ϵ (in columns) for $I = 7$.

5.4.6 Computational time

In order to decrease the calculation time in our algorithm, we chose to restrain the search space in the tree to a certain amount, depending on the number of initial features. We can choose to limit the search in the area of the tree where the jumps in the hierarchy are the highest and arbitrarily set the number of groups to evaluate at five times the number of initial features. By doing so, we are reducing the number of variables to be fitted in the Lasso regression without affecting the performance in terms of Recall or Precision.

We compared the computational performance of our method with the three others by varying D^G (we fixed $D^M = 100$ and $n = 200$). We repeated the number of evaluation five times for each D^G and averaged the calculation time.

	$\hat{\theta}_{gm} = 0$	$\hat{\theta}_{gm} \neq 0$
$\theta_{gm} = 0$	True Negative	False Positive
$\theta_{gm} \neq 0$	False Negative	True Positive

Results of averaged calculation time (in minutes) over 5 replicates for varying D^G

We can conclude from the results presented in table **??** that two methods, MLGL and glinternet, are not suitable for large-scale analysis of genomic data since the calculation time increase drastically as soon as the dimension of the problem exceed a few thousand variables. HCAR and SICOMORE behave similarly. That being said, remember that HCAR is tuned with an unweighted compressed representation avoiding having to choose the optimal cut in the tree, as in SICOMORE. With its original strategy based on a K cross validation, there is no doubt that the gap between HCAR and SICOMORE would have been much larger. Indeed, the computational cost of an additional exploration to find the optimal cut in HCAR grows with the number of variables and therefore with h_T, the height of the tree. HCAR has to evaluate $h_T \times K$ compressed models while SICOMORE only has to compress $h_T - 1$ groups to evaluate the final model.

5.5 Application on real data: rhizosphere of *Medicago truncatula*

5.5.1 Material

In order to study the interactions between *Medicago truncatula* and the microbial community of its rhizosphere, a core collection of 154 accessions have been analysed. The purpose of the study is to identify significant interactions between

the plant genome and the microbial metagenome to better understand the effect of the microbial community on the growth of the plant.

Each accession was grown in a controlled environment and phenotyped for several traits related to the growth and nutritional strategy:

- Measure of total biomass (BMtot).
- Root Shoot Ratio (RTR).
- Specific Nitrogen Uptake (SNU). It is the correlation between the total amount of nitrogen and below-ground biomass

The distributions of the phenotypic values are shown in Figure **??**.

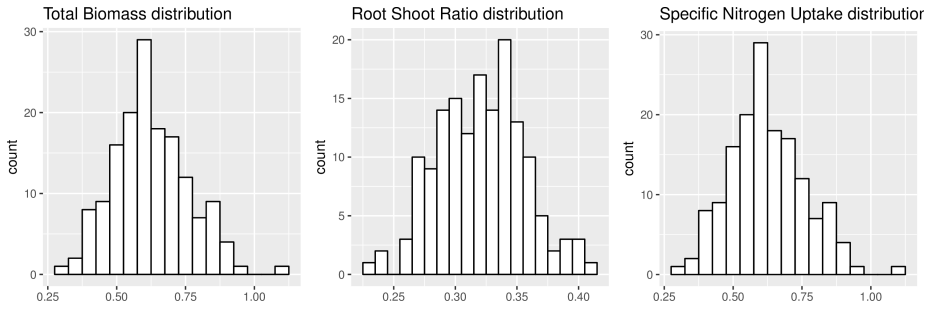


Figure 5.7: Distribution of the phenotypic values for the BMtot, RTR and SNU traits.

In addition to the phenotypic measurement, the rhizosphere of each accession was also analysed to determine the microbial diversity in terms of number of species and abundance of each species. The metagenomic composition of the rhizosphere has been analysed by DNA extraction and shotgun sequencing. A total of 848 different species were found in the rhizosphere of the plants (repartition shown in Figure **??**).

Finally, 154 accession were genotyped with a DNA microarray chip for a total number of 6 372 968 SNP. The missing values were imputed using the 'snp.imputation' function from the `SNPstats` R package. Given two set of SNP typed in the same subjects, this function calculates rules which can be used to impute one set from the other in a subsequent sample.

Some SNP having too many missing values to be imputed at 100%, we only kept the SNP which have been completely imputed, thus reducing the size of the data to 2 148 505 SNP. We also looked at the linkage disequilibrium level among some SNP to get an overview of the genome structure (Figure **??**).

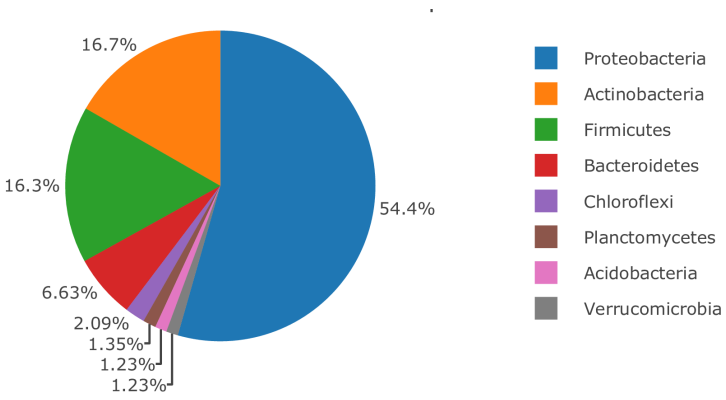


Figure 5.8: Distribution of the phenotypic values for the BMtot, RTR and SNU traits.

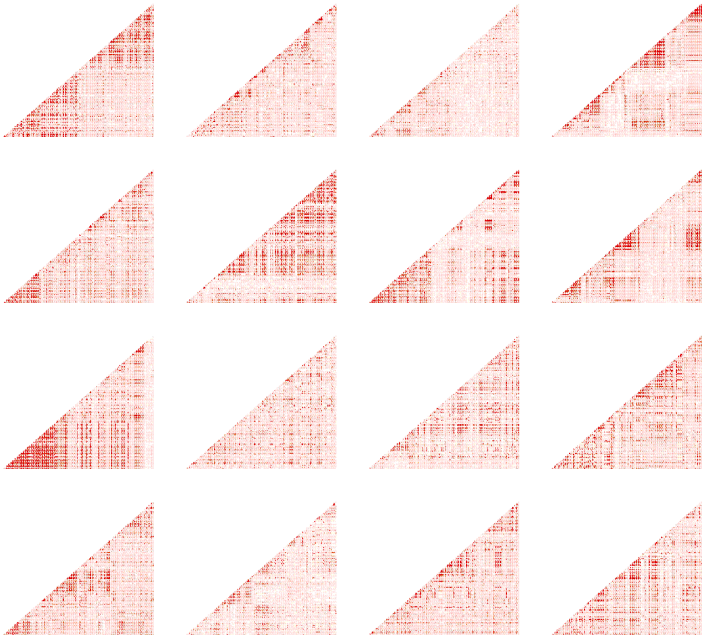


Figure 5.9: Heatmap of LD level among SNP of chromosome 4 (position 17448921 to 22706884).

5.5.2 Analysis

The algorithm SICOMORE requires that we choose several hyper-parameters in order to run properly:

- **Aggregating function:** For both metagenomic and genomic data we define the mean value of the group as supervariable.

- **Clustering algorithm:** For the metagenomic data we used 2 types of clustering: a hierarchical clustering using Ward's distance as the measure of similarity and a phylogenetic clustering using the taxonomic information to construct a tree. The first method does not use information a priori while the second uses phylogenetic information to build a tree. For the genomic data, we used spatially constrained hierarchical clustering algorithm which integrates the linkage disequilibrium as the measure of dissimilarity. It is also possible not to specify any hierarchy for one the 2 datasets, in that case we are looking for interaction between groups of variables in one dataset and single variables in the second dataset.

- **Search space**: For computational reasons, we searched at first for interaction between a subset of the SNP data and the metagenomic data. We chose arbitrarily a subset of 10% of the initial data matrix (214 851 SNP). We also chose to divide the analysis chromosome by chromosome.

To summarize we performed an exhaustive search for interaction by setting different parameters:

Option 1 : Hierarchical clustering on metagenomic data + spatially constrained hierarchical clustering on subset of genomic data (214 851 SNP) + chromosome by chromosome.

Option 2 : Hierarchical clustering on metagenomic data + spatially constrained hierarchical clustering on subset of genomic data (214 851 SNP) + all chromosomes combined.

Option 3 : Hierarchical clustering on metagenomic data + spatially constrained hierarchical clustering on all genomic data (2 148 505 SNP) + chromosome by chromosome.

Option 4 : Phylogenetic clustering using taxonomic information on metagenomic data + spatially constrained hierarchical clustering on subset of genomic data (214 851 SNP) + chromosome by chromosome.

5.5.3 Results

5.5.3.1 Results on Total Biomass

	selected groups in Metagenome	*selected groups in Genome*	*signif interactions chr by chr*	*signif interactions on all chr*
Option 1	14	157	1	0
Option 2	12	4	0	0
Option 3	14	84	0	0
Option 4	39	96	0	0

Results from SICOMORE analysis on total biomass

One significant interaction was found for the phenotype Total Biomass when we applied a BH correction chromosome by chromosome (column 3) instead of correcting on the all set of p-value, all chromosome confounded:

Metagenomic group	*Chromosome*	*Genomic position (pb)*	*number of SNP in genomic region*	*p-value*
'Hyalangium'	5	30921278	1	$2.5e.10^{-4}$

Results from SICOMORE analysis on total biomass

5.5.4 Results on Root Shoot Ratio

	selected groups in Metagenome	*selected groups in Genome*	*signif interactions chr by chr*	*signif interactions on all chr*
Option 1	9	16	7	0
Option 2	9	4	0	0
Option 3	4	38	0	0
Option 4	6	16	10	0

Results from SICOMORE analysis on Root Shoot Ratio.

We found some significant interactions when we applying a BH correction chromosome by chromosome (column 3). We observe different results according to the clustering applied on the metagenomic data. We found 7 significant interactions for the phenotype Root Shoot Ratio when we applied a hierarchical clustering and 10 significant interactions when we applied a phylogenetic clustering.

1. Option 1 (hierarchical clustering on Metagenomic data and subset of SNP data):

Metagenomic group	*Chromosome*	*Genomic position (pb)*	*number of SNP in genomic region*	*p-value*
'Ramlibacter'	1	41164000:52989926	5414	$2.1.10^{-3}$
'Ramlibacter'	2	37760407:45726447	3699	5.10^{-3}
'Ramlibacter'	5	38102314:40264975	1184	1.10^{-2}
39 species	6	13605:2142611	1016	$1.2.10^{-3}$
'Ramlibacter'	6	13605:2142611	1016	$4.8.10^{-3}$
'Ramlibacter'	6	2142998:35275274	19060	$1.9.10^{-3}$
'Ramlibacter'	7	1383	23535	$4.6.10^{-3}$

2. Option 4 (phylogenetic clustering on Metagenomic data and subset of SNP data):

Metagenomic group	*Chromosome*	*Genomic position (pb)*	*number of SNP in genomic region*	*p-value*
'Ramlibacter'	1	823:33141394	19400	$8.4.10^{-3}$
'Ramlibacter'	1	41164000:52989926	5414	$2.1.10^{-3}$
'Ramlibacter'	2	531:37759102	21196	$2.5.10^{-2}$
'Ramlibacter'	2	37760407:45726447	3699	5.10^{-3}
'Ramlibacter'	3	38120792:41521643	1737	$6.9.10^{-3}$
'Ramlibacter'	3	54877971:55514282	286	$2.9.10^{-3}$
5 species ('Sphingobacteriia')	5	38102314:40264975	1184	$8.8.10^{-2}$
'Ramlibacter'	5	38102314:40264975	1184	1.10^{-2}
'Ramlibacter'	6	13605:2142611	1016	$4.8.10^{-3}$
'Ramlibacter'	6	2142998:35275274	19060	$1.9.10^{-3}$

5.5.5 Results on Specific Nitrogen Uptake

	selected groups in Metagenome	*selected groups in Genome*	*signif interactions chr by chr*	*signif interactions on all chr*
Option 1	4	15	0	0
Option 2	4	2	0	0
Option 3	4	32	4	2
Option 4	4	34	0	0

Results from SICOMORE analysis on Specific Nitrogen Uptake.

2 significant interactions (on a total 128 potential interactions) were found between 2 groups of microbial species and 2 groups of SNP when assessing all the genomic data (Option 4) and applying the BH correction on the all set of p-value:

Metagenomic group	*Chromosome*	*Genomic position (pb)*	*number of SNP in genomic region*	*p-value*
140 species	2	15784171:15825022	462	$5.4.10^{-5}$
140 species	2	37906624:37914488	53	$1.7.10^{-3}$
613 species	5	15072986:15100311	251	$3.4.10^{-3}$
140 species	5	15072986:15100311	251	$3.5.10^{-4}$

The 140 microbial species found in the interactions with chromosome 2 and 5 are the same species. The repartition in phylum of these species are illustrated in Figure **??**, there is a total of 13 phylum represented with a vast majority of Proteobacteria.

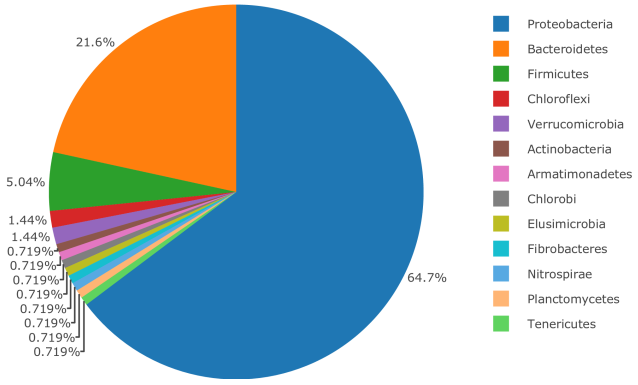


Figure 5.10: Microbial phylum found in interaction with chromosome 2 and chromosome 5 of *Medicago truncatula* for the Specific Nitrogen Uptake phenotype.

5.6 Discussions

Although the detection of interaction effects in a high-dimensional remain a difficult problem, on one hand due to the multiple testing burden and on the other hand to the small effect sizes in term of significance, our approach has demonstrated the ability to recover interaction effects with a high statistical power. In our simulations, whether we varied the sample sizes, noise or number of true interactions, SICOMORE always exhibited the strongest recall compared to MLGL, HCAR or glinternet. This can be explained mainly by the fact that we advantageously use the strengths of different methods to combine them in a powerful single algorithm.

Regarding the results in terms of precision, we can see that all methods exhibit weak performance mainly due to the fact that the algorithms select groups which are too high in the hierarchy, i.e. that the selected supervariables, or

groups of single variables for MLGL, contain too many variables. This results in the detection of interactions between the complementary datasets with a good power but a weak resolution. One solution would be to constrain the algorithm to work only on the lowest levels of the two hierarchies at a potential cost in terms of recall.

As for the application of our method to the *Medicago truncatula* dataset, we were able to find significant interactions between genomic and metagenomic features in relation with 3 phenotypes. Particularly we notice than one particular microbial species, 'Ramlibacter', seems to highly interact with the genome of the plant. We detected a lot of interactions for the RTR phenotype with potentially interesting genomic regions to look at in more details. The results on the phenotype SNU are more difficult to interpret because it is a very large group of microbial species which interact with the genome. Furthermore, we can notice in these results that the variable selection step suffers from instability. Indeed, as we used the same metagenomic data across the different options, the number of selected groups should also remains the same, but it is not the case. This instability could be due to the cross-validation step necessary to estimate the hyper-parameters and would need some adjustments to be corrected.

To conclude we can state that SICOMORE is able to find significant metagenomic-genomic interactions in a high dimensional context within a reasonable computational time. Indeed, the algorithm is able to work very fast even with large genomic dataset, an analysis of the full genomic data only takes a few hours to run and only a few tens of minutes if we work on a small subset of the data.

Conclusions

Since the last decade, the rapid advances in genotyping technologies have changed the way genes involved in mendelian disorders and complex diseases are mapped, moving from candidate genes approaches to linkage disequilibrium mapping, of which GWAS is a large-scale example. In the mid-1990s, some researchers already foresaw the coming of the GWAS era and the crucial contribution of high-throughput genotyping technologies in the field of genetic epidemiology. Indeed, (**?**) noted that small genetic effects could be detected with greater power by association analyses and proposed that genome-wide LD mapping (GWAS) could be applied if technologies were developed to study SNP frequencies in all genes, contrasting in ill cases vs. control subjects. On another side, (**?**) suggested the common disease common variant (CDCV) hypothesis and proposed cataloguing common SNP (with MAF $\geq$ 5%) and studying their association to disease in large samples. GWAS strategy under the CDCV hypothesis assumed that many different common SNP have small effects on each disease, and that some could be found by testing enough SNP in enough people.

Since 2005 ((**?**)), GWAS have produced strongly significant evidence that specific common DNA sequence differences among people influence their genetic susceptibility to many different common diseases (**?**). However, they are also subject to several limitations intrinsic to the types of data but also to the statistical methods used. On one side the strong correlations between genetic variants, population structure, epistasis or effect size of rare-variant are partly responsible for the missing heritability. But on the other hand, although the single marker method remains the most widely used approach in GWAS, its relevance may be called into question in the context of complex diseases.

The new methodologies developed during this PhD are therefore part of this context. We try with this manuscript to provide a thorough introduction to GWAS by reminding in a first time the genetic precepts fundamental to the understanding of our works but also by introducing the concept of statistical learning. We chose not only to detail several state-of-the-art methods used in GWAS but also to put a particular emphasis on statistical learning by devoting an entire chapter to it. This choice was motivated by the conviction that a multidisciplinary approach combining both biological and statistical learning knowledges

can help to understand the limits of traditional methods used in GWAS but also to imagine potential levers for improvement in terms of methodology.

Discussions on LEOS algorithm

Based on the observation that baseline single-marker analysis in GWAS is strongly affected by the multiple testing burden due to the high dimensionality of the data leading to the inability to identify variants having small effect on phenotype, we first came up with the idea of aggregating SNP within a same LD block for a dimension-reduction purpose. This reasoning led to the development of the method LEOS, described in Chapter **??**. In this work we proposed a four-step algorithm explicitly designed to take benefit of the linkage disequilibrium structure in GWAS data. LEOS combines, on the one hand, unsupervised learning methods that cluster correlated-SNP, and on the other hand, supervised learning techniques that identify the optimal number of clusters and reduce the dimension of the predictor matrix.

The evaluation of the method was carried out from both a predictive and explanatory point of view. One part of the method consist in finding the optimal group structure to construct a matrix on new aggregated-SNP variables using supervised learning techniques. We noticed, in the assessment of the method on simulated and real datasets, that the combination of our aggregating function with a ridge regression model leads to a major improvement in terms of predictive power when the linkage disequilibrium structure is strong enough, hence suggesting the existence of multivariate effects due to the combination of several SNP. Furthermore, when using high-dimensional generalized additive model (HGAM) in place of linear models, we remarked that we were able to further increase the predictive accuracy. These results suggest a first interesting feature of our method if one wants to predict a phenotype based solely on genetic markers, with possible application in personalized medicine. However, these preliminary results, although encouraging, must be subjected to additional tests such as a comparative analysis with other machine learning algorithms specialized in the predictive aspect. It also seems important to confirm the robustness of these results on other data sets and on replicative studies.

Although the predictive aspect of the algorithm is of crucial importance, the main objective we had in mind while developing the method was to find a way to increase statistical power and precision in GWAS. Regarding this matter, accounting for the linkage disequilibrium structure of the genome and aggregating highly-correlated SNP is seen to be a powerful alternative to standard marker analysis. Indeed, LEOS demonstrates its ability, in different simulation scenarios, to retrieve true causal SNP and/or clusters of SNP with substantially higher precision coupled with a good power than standard approaches. Even though it has been able to recover a genomic region known to be associated with ankylosing spondylitis, we have not been able to detect new genomic regions significantly associated with the disease, certainly suggesting that some

effects might still be too small to be detected or that there are other causes that cannot be detected with this type of approach, such as effects of interactions with the environment or epistasis. We also investigated, using HGAM on the aggregated-SNP matrix, the possibility to detect non-linear relationship with the phenotype. Albeit the regions identified did not differ from those previously identified with a classical linear regression model, the results obtained on the AS dataset still point interesting non-linear patterns between some aggregated-SNP in the specific HLA region of chromosome 6 and the phenotype. Nevertheless, we remain convinced that generalized additive models could be of great benefit in GWAS, particularly in terms of predictive power but also in the identification of non-linear behaviour.

Discussions on SICOMORE algorithm

One possible way to understand the expression of certain diseases is to consider gene-environment interactions. Sensitivity to environmental risk factors for a disease may be inherited, leading to cases where individuals exposed to the same environment but with different genotypes can be affected differently, resulting in different disease phenotypes. In the context of medical genetics and epidemiology, the study of gene-environment interactions is of great importance. Indeed, if we estimate only the separate contributions of genes and environment to a disease, and ignore their interactions, we will incorrectly estimate the fraction of phenotypic variance attributable to genes, environment, and their joint effect. Restricting analysis of environmental factors in epidemiological studies to individuals who are genetically susceptible to the exposure should increase the magnitude of relative risks, increasing the confidence that the observed associations are not due to chance (**?**).

A possible lead to investigate gene-environment interactions is take into account the contribution of microbial communities on the expression of a phenotype. As previously stated, there is growing evidences of the role of microbiome in basic biological processes whether in progression of major human diseases or in plant growth. These facts motivated the development of a new statistical method to tackle the detection of such interactions in a GWAS context. This topic offers many statistical challenges, among which the way to deal with the multiple testing burden. That is why we choose to use the idea to compress the data, as with the LEOS method, and to combine several statistical learning methods to develop an algorithm dedicated to the search for statistical interactions, with a focus on genomic and metagenomic data.

The SICOMORE method, described in Chapter **??**, advantageously uses the strengths of different existing methods to combine them in a powerful single algorithm. First of all, we constructed the hierarchy of the genetic data with a well-proven spatially-constrained hierarchical clustering adapted to SNP data developed by (**?**). Secondly, taking the average values of strongly correlated predictors, such as SNP within the same LD-block, and use them into a predic-

tive model has already proved by (**?**) to be a powerful approach. Finally, we took benefit of the weighting scheme proposed by (**?**) for the selection of the supervariables in the lasso procedure where we used a penalty factor defined by the length of the gap in the hierarchical tree, as explained in Section **??**.

We evaluated and compared the performance SICOMORE with others methods in terms of power and precision. The results have put forward that, in terms of precision, all methods exhibit weak performances mainly due to the fact that the algorithms select groups which contain too many variables. As for the statistical power, SICOMORE always exhibited in the numerical simulations the strongest recall compared to the other methods. The application of our method to the *Medicago truncatula* dataset highlighted some significant interactions between genomic and metagenomic features in relation with three different phenotypes. However, although promising, these results need to be confirmed by a relevant biological interpretation that will be carried out by a discussion with our collaborators from INRA who have gracefully provided us these data. This should allow to append a biological interpretation to these results in the paper to come (currently in a preprint state).

Despite these interesting results, SICOMORE is nonetheless subject to some limitations that need to be addressed in future works. First of all, although the lasso procedure to select the supervariables in both complementary datasets is relevant for a dimension-reduction purpose, it may induce some biases in the multiple testing procedure we use afterwards because we perform a variable selection step before adjusting the p-values. One way around this problem could be to use post-hoc inference for multiple comparisons (**?**).

Secondly, as observed in the analysis of the *Medicago truncatula* dataset, the stability of the variable selection step is problematic. The use of a variable selection model other than the lasso may circumvent this issue, with for instance the *Bolasso* model (**?**) where the author proposed to intersect the supports of replicated bootstrapped Lasso estimates for consistent model selection. In the same fashion, (**?**) introduced the stability selection based on subsampling in combination with high-dimensional selection algorithms.

Perspectives

The works presented in this thesis are the result of a reflection on ways to improve GWAS studies through the creation of new data-driven methodologies. Still, the possible contributions to the field of GWAS brought by the development of new statistical methods are not limited to those mentioned in this manuscript and can fall into a number of categories depending on their objectives. To conclude, we will therefore suggest some avenues of research not mentioned so far but worthwhile to be explored in future works.

At first, we can mention methods designed to better modelled population struc-

ture and relatedness between individuals in a sample during association analyses such as the works on linear mixed models in (**???**) or the methods for estimating and partitioning genetic (co)variance (**??**).

In another fashion, methods combining classical statistical approaches with Machine Learning are of interest for exploratory purposes as in (**?**) where multiple hypothesis tests are combined with support vector machine (SVM) to increase statistical power. Similarly, for purely predictive purposes, several machine learning methods such as random forest (**?**), classification-regression trees (CRT) (**?**) or even Deep Learning (Neural Network) (**?**) are also worthwhile considering in GWAS.

At last, the discovery of causal pathways between genomes and molecular traits such as gene expression, DNA methylation, or metabolites is of great importance to unravel cause and consequence in genetic epidemiology. The combination of sequence variation with molecular phenotypes, disease data and environmental covariates with novel analytical methods such as Mendelian randomization (**??**) or causal Bayesian networks as in (**?**) have great potential in this respect.

Appendix A

Derivation of the MSE bias-variance decomposition

For the sake of brevity, we will abbreviate $f = f(\mathbf{x})$ and $\hat{f} = \hat{f}(\mathbf{x})$ estimated on a training set $\mathcal{T}$.

$$\begin{aligned}\mathbb{E}_{\mathcal{T}}[(\mathrm{Y} - \hat{f})^2] &= \mathbb{E}_{\mathcal{T}}[\mathrm{Y}^2 + 2\mathrm{Y}\hat{f} + \hat{f}^2] \\ &= \mathbb{E}_{\mathcal{T}}(\mathrm{Y}^2) + \mathbb{E}_{\mathcal{T}}[\hat{f}^2] - 2\mathbb{E}_{\mathcal{T}}[\mathrm{Y}\hat{f}].\end{aligned}$$

Remembering the following properties of the variance and expectation:

$$\begin{aligned}&Var(\mathrm{X}) = \mathbb{E}(\mathrm{X}^2) - \mathbb{E}^2(\mathrm{X}), \\ &\mathbb{E}(\mathrm{XY}) = \mathbb{E}(\mathrm{X})\mathbb{E}(\mathrm{Y}) + Cov(\mathrm{X}, \mathrm{Y}), \\ &Var(\mathrm{X} + \mathrm{Y}) = Var(\mathrm{X}) + Var(\mathrm{Y}) + 2Cov(\mathrm{X}, \mathrm{Y}), \\ &Var(\mathrm{X} - \mathrm{Y}) = Var(\mathrm{X}) + Var(\mathrm{Y}) - 2Cov(\mathrm{X}, \mathrm{Y}), \\ &Cov(\mathrm{X}, \mathrm{Y}) = 0 \text{ if X and Y are independent,}\end{aligned}$$

and using them in we get:

$$\mathbb{E}_{\mathcal{T}}[(\mathrm{Y} - \hat{f})^2] = Var(\mathrm{Y}) + \mathbb{E}^2_{\mathcal{T}}(\mathrm{Y}) + Var(\hat{f}) + \mathbb{E}^2_{\mathcal{T}}(\hat{f}) - 2\mathbb{E}_{\mathcal{T}}[(f + \epsilon)\hat{f}]. \quad \text{(A.1)}$$

Developing the expression:

$$\begin{aligned}2\mathbb{E}_{\mathcal{T}}[(f + \epsilon)\hat{f}] &= 2\mathbb{E}_{\mathcal{T}}(f\hat{f}) + 2\mathbb{E}_{\mathcal{T}}(\hat{f}\epsilon) \\ &= 2\mathbb{E}_{\mathcal{T}}(f\hat{f}) + 2[\mathbb{E}_{\mathcal{T}}(\hat{f}) \underbrace{\mathbb{E}_{\mathcal{T}}(\epsilon)}_{=0} + \underbrace{cov(\hat{f}, \epsilon)}_{=0}] \\ &= 2[\mathbb{E}_{\mathcal{T}}(f)\mathbb{E}_{\mathcal{T}}(\hat{f}) + Cov(f, \hat{f})],\end{aligned}$$

and remplacing $Var(\mathrm{Y}) = \underbrace{Var(f)}_{=0} + Var(\epsilon) + \underbrace{2Cov(f,\epsilon)}_{=0} = \sigma^2$ in (**??**), we get:

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}[(\mathrm{Y} - \hat{f})^2] &= Var(\mathrm{Y}) + \mathbb{E}^2_{\mathcal{T}}(\mathrm{Y}) + Var(\hat{f}) + \mathbb{E}^2_{\mathcal{T}}(\hat{f}) - 2[\mathbb{E}_{\mathcal{T}}(f)\mathbb{E}_{\mathcal{T}}(\hat{f}) + Cov(f,\hat{f})] \\
&= Var(\hat{f}) + \sigma^2 + \mathbb{E}^2_{\mathcal{T}}(\mathrm{Y}) + \mathbb{E}^2_{\mathcal{T}}(\hat{f}) - 2\mathbb{E}(f)\mathbb{E}_{\mathcal{T}}(\hat{f}) \\
&= Var(\hat{f}) + \sigma^2 + \mathbb{E}^2_{\mathcal{T}}(\mathrm{Y}) + \mathbb{E}^2(\hat{f}) - 2\mathbb{E}(f)\mathbb{E}_{\mathcal{T}}(\hat{f})
\end{aligned}$$

Knowing that $\mathbb{E}^2_{\mathcal{T}}(\mathrm{Y}) = \mathbb{E}^2_{\mathcal{T}}(f+\epsilon) = \mathbb{E}^2(f)$ and replacing in (**??**), we finally obtain:

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}[(\mathrm{Y} - \hat{f})^2] &= Var(\hat{f}) + \sigma^2 + \mathbb{E}^2(f) + \mathbb{E}^2_{\mathcal{T}}(\hat{f}) - 2\mathbb{E}(f)\mathbb{E}_{\mathcal{T}}(\hat{f}) \\
&= \underbrace{Var(\hat{f})}_{Variance} + \underbrace{[\mathbb{E}(f) - \mathbb{E}_{\mathcal{T}}(\hat{f})]^2}_{Bias^2} + \underbrace{\sigma^2}_{noise}
\end{aligned}$$

Appendix B

Linear smoother (?)

One can also show that the smoothing spline is a linear smoother, and hence we can write down a smoother matrix. The following is taken from (**?**).

Let $h_i = x_{i+1} - x_i$, $i = 1, 2, \ldots, n-1$, Δ be a tridiagonal $(n-2) \times n$ matrix such as

$$\Delta_{ii} = \frac{1}{h_i}, \quad \Delta_{i,i+1} = -(\frac{1}{h_i} + \frac{1}{h_{i+1}}), \quad \Delta_{i,i+2} = \frac{1}{h_{i+1}},$$

and let $\mathbf{C}$ be a symnetric tridiagonal matrix of order $n-2$ with :

$$C_{i-1,i} = C_{i,i-1} = \frac{h_i}{6}, \quad C_{ii} = \frac{h_i + h_{i+1}}{3},$$

Then is can be showned that solving

$$\sum_{i=1}^{n} ||y_i - s(x_i)||^2 + \lambda \int_a^b s''(t)^2 dt,$$

is equivalent to minimizing

$$||\mathbf{y} - s(\mathbf{x})||^2 + \lambda s(\mathbf{x}) \mathbf{K} s(\mathbf{x}) \quad \text{where } \mathbf{K} = \Delta^T \mathbf{C}^{-1} \Delta$$

with solution $\hat{\mathbf{y}} = \hat{s}(\mathbf{x}) = \mathbf{S}\mathbf{y}$, where $\mathbf{S} = (\mathbf{I} + \lambda \mathbf{K})^{-1}$.

Appendix C

Smoothing parameter λ for smoothing splines

A suitable criterion to choose λ can be the mean-square error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - s_i)^2,$$

However s is unknown so the MSE cannot be used directly but it is possible to derive an estimate of $\mathbb{E}(MSE) + \sigma^2$, which is the expected squared error in predicting a new variable. We define the ordinary cross validation score as

$$CV_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}^{[-i]} - y_i)^2$$

Substituting $y_i = s_i + \epsilon_i$,

$$\begin{aligned} CV_o &= \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i^{[-i]} - s_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i^{[-i]} - s_i)^2 - (\hat{s}_i^{[-i]} - s_i)\epsilon_i + \epsilon_i^2. \end{aligned}$$

Since $\mathbb{E}(\epsilon_i) = 0$, and that ϵ_i and $\widehat{f}^{[-i]}$ are independent, the second term in the summation vanishes if expectations are taken:

$$\mathbb{E}(CV_o) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(\hat{s}_i^{[-i]} - s_i)^2\right) + \sigma^2.$$

$\hat{s}^{[-i]} \approx \hat{s}$ with equality in the large sample limit, so $\mathbb{E}(CV_o) \approx \mathbb{E}(MSE) + \sigma^2$ also with equality in the large sample limit. Choosing λ in order to minimize CV_o is known as ordinary cross validation.

It can be shown that ordinary leave-one-out cross validation is defined as follow

$$LVOCV_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{s}_i)^2/(1 - A_{ii})^2,$$

where $\hat{s}$ is the estimate from fitting all the data and $\mathbf{A}$ is the corresponding influence matrix. In practice the weights, $1 - A_{ii}$, are often replaced the mean weight, $\text{trace}(\mathbf{I} - \mathbf{A})/n$ in order to get the generalized cross validation score

$$GCV = \frac{n\sum_{i=1}^{n}(y_i - \hat{s}_i)^2}{\text{trace}(\mathbf{I} - \mathbf{A})^2}.$$

Appendix D

B-spline basis

Given that the solution of the optimization problem is a natural cubic spline with $n-2$ interior knots, we can represent it in terms of B-spline basis functions.

We can write $s(\mathbf{x}) = \sum_1^{n+2} \gamma_d B_d(\mathbf{x})$, where γ_j are coefficients and the B_d are the cubic B-spline basis functions. We define the $n \times (n+2)$ matrix $\mathbf{B}$ and the $(n+2) \times (n+2)$ matrix $\mathbf{\Omega}$ by

$$B_{id} = B_d(x_i)$$

and

$$\Omega_{ii'} = \int B_i''(\mathbf{x}) B_{i'}''(\mathbf{x}) dx$$

The optimization criterion

$$\sum_{i=1}^{n} ||y_i - s(x_i)||^2 + \lambda \int_a^b s''(t)^2 dt,$$

can be rewrite as:

$$(\mathbf{y} - \mathbf{B}\gamma)^T(\mathbf{y} - \mathbf{B}\gamma) + \lambda \gamma^T \Omega \gamma$$

We set the derivative of to 0 with respect to γ to get the solution:

$$\frac{\partial[(\mathbf{y} - \mathbf{B}\gamma)^T(\mathbf{y} - \mathbf{B}\gamma) + \lambda \gamma^T \mathbf{\Omega} \gamma]}{\partial \gamma} = 0$$

$$\frac{\partial[\mathbf{y}^T\mathbf{y} - 2\mathbf{B}^T\mathbf{y}\gamma + \mathbf{B}^T\gamma^T\mathbf{B}\gamma + \lambda \gamma^T \Omega \gamma]}{\partial \gamma} = 0$$

$$-2\mathbf{B}^T\mathbf{y} + 2\mathbf{B}^T\mathbf{B} + 2\lambda\Omega\gamma = 0$$

$$(\mathbf{B}^T\mathbf{B} + \lambda\Omega)\gamma = \mathbf{B}^T\mathbf{y}$$

$$\hat{\gamma} = (\mathbf{B}^T\mathbf{B} + \lambda\Omega)^{-1}\mathbf{B}^T\mathbf{y}$$

For computational purpose, it can be shown (**?**) that the B-spline basis function of size $n \times (K+4)$ can be expressed as a $n \times (K+2)$ basis matrix $\mathbf{N}$ for the natural cubic splines with the same interior and boundary knots at the extreme of $\mathbf{x}$.

The solution vector $\hat{s}$ can be write as

$$\hat{\mathbf{S}} = \mathbf{N}\mathbf{B}\hat{e}ta = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{y} = (\mathbf{I} + \lambda\mathbf{K})^{-1}\mathbf{y}$$

where $\mathbf{K} = \mathbf{N}^{-T}\mathbf{\Omega}\mathbf{N}^{-1}$ and $\mathbf{B}\hat{e}ta$ the transformed version of $\hat{\gamma}$ corresponding to the change in basis. In terms of the candidate fitted vector $\mathbf{f}$ and $\mathbf{K}$, the cubic smoothing spline $\hat{f}$ minimizes

$$(\mathbf{y} - \mathbf{s})^T(\mathbf{y} - \mathbf{s}) + \lambda\mathbf{s}^T\mathbf{K}\mathbf{s}$$

over all vectors $\mathbf{s}$.

To compute all of this efficiently, the natural spline basis functions should be chosen so that $\mathbf{N}$ and (4) are band limited which thereby allow the fitted values to be computed in $O(n)$ calculations (**?**). Specific ways to obtain such band limited structures are given in (**?**). In our case, where the natural spline basis is a cubic spline basis, we can use the piecewise polynomial representation for the estimator describe in (**?**) to show that

$$\hat{\mathbf{S}} = \mathbf{y} - \lambda\mathbf{Q}(\lambda\mathbf{Q}^T\mathbf{Q} + \Delta)^{-1}\mathbf{Q}^T\mathbf{y},$$

where $\mathbf{Q}^T$ is an $(n-2) \times n$ tridiagonal matrix with ith row

$$(\underbrace{0, \dots, 0}_{i-1}, \frac{1}{t_{i+1} - t_i}, -\frac{1}{t_{i+2} - t_{i+1}} - \frac{1}{t_{i+1} - t_i}, \frac{1}{t_{i+2} - t_{i+1}}, \underbrace{0, \dots, 0}_{n-i-2})$$

and Δ is symmetric, $(n-2) \times (n-2)$, tridiagonal matrix having first and last rows $(t_2 - t_1, t_3 - t_2, \underbrace{0, \dots, 0}_{n-4})$ and $(\underbrace{0, \dots, 0}_{n-4}, t_{n-1} - t_{n-2}, t_n - t_{n_1})$, and with i^{th} row

$$(\underbrace{0, \dots, 0}_{i-2}, t_{i+1} - t_i, 2(t_{i+2} - t_i), t_{i+2} - t_{i+1}, \underbrace{0, \dots, 0}_{n-i-3})$$

for $i = 2, \dots, n-3$.

The fitted values for the cubic smoothing splines can therefore be obtained in $O(n)$ operations by first solving the 5-banded system

$$(\lambda \mathbf{Q}^T \mathbf{Q} + \Delta)\gamma = \mathbf{Q}^T \mathbf{y}$$

and then using $\hat{\mathbf{S}} = \mathbf{y} - \lambda \mathbf{Q}\gamma$.

References