Variable selection for data aggregated from different sources with group of variable structure

Camilo Broc

DOCTOR OF PHILOSOPHY IN MATHEMATICS
operated within
University of Pau and Adour Countries
University of the Basque Country

École Doctorale N°211
École Doctorale sciences exactes et leurs applications

Spécialité de doctorat : Mathématiques

Camilo Broc

Variable selection for data aggregated from different sources with group of variable structure

Novel statistical methods for high dimension data

Devant le jury composé de :

Jacqmin-Gadda Hélène, Director of research, Bordeaux Population Health Research Center (France)	President
Ambroise Christophe, Pr., University of Évry Val d'Essonne (France)	Referee
Robin Stéphane, Director of research, AgroParisTech/INRA (France)	Referee
Jourdan Astrid, Associate Professor, EISTI (France)	Examinator
Liquet Benoît, Pr., University of Pau and Adour Countries (France) and Queensland University of Technology (Australia)	Supervisor
Calvo Borja, Pr., University of the Basque Country (Spain)	Co-supervisor

Résumé

Durant les dernières décennies, la quantité de données disponibles en génétique a considérablement augmenté. D'une part, une amélioration des technologies de séquençage de molécules a permis de réduire fortement le coût d'extraction du génome humain. D'autre part, des consortiums internationaux d'institutions ont permis la mise en commun de la collecte de données sur de larges populations. Cette quantité de données nous permet d'espérer mieux comprendre les mécanismes régissant le fonctionnement de nos cellules. Dans ce contexte, l'épidémiologie génétique est un domaine cherchant à déterminer la relation entre des caractéristiques génétiques et l'apparition d'une maladie. Des méthodes statistiques spécifiques à ce domaine ont dû être développées, en particulier à cause des dimensions que les données présentent : en génétique, l'information est contenue dans un nombre de variables grand par rapport au nombre d'observations.

Dans cette dissertation, deux contributions sont présentées. Le premier projet appelé PIGE (Pathway-Interaction Gene Environment) développe une méthode pour déterminer des interactions gène-environnement. Le second projet vise à développer une méthode de sélection de variables adaptée à l'analyse de données provenant de différentes études et présentant une structure de groupe de variables.

Le document est divisé en six parties. Le premier chapitre met en relief le contexte, d'un point de vue à la fois biologique et mathématique. Le deuxième chapitre présente les motivations de ce travail et la mise en œuvre d'études en épidémiologie génétique. Le troisième chapitre aborde les questions relatives à l'analyse d'interactions gène-environnement et la première contribution de la thèse y est présentée. Le quatrième chapitre traite des problématiques de méta-analyses. Le développement d'une nouvelle méthode de réduction de dimension répondant à ces questions y est présenté. Le cinquième chapitre met en avant la pertinence de la méthode dans des cas de pleiotropie. Enfin, le sixième et dernier chapitre dresse un bilan du travail présenté et dresse des perspectives pour le futur.

Abstract

During the last decades, the amount of available genetic data on populations has grown drastically. From one side, a refinement of chemical technologies have made possible the extraction of the human genome of individuals at an accessible cost. From the other side, consortia of institutions and laboratories around the world have permitted the collection of data on a variety of individuals and population. This amount of data raised hope on our ability to understand the deepest mechanisms involved in the functioning of our cells. Notably, genetic epidemiology is a field that studies the relation between the genetic features and the onset of a disease. Specific statistical methods have been necessary for those analyses, especially due to the dimensions of available data: in genetics, information is contained in a high number of variables compared to the number of observations.

In this dissertation, two contributions are presented. The first project called PIGE (Pathway-Interaction Gene Environment) deals with gene-environment interaction assessments. The second one aims at developing variable selection methods for data which has group structures in both the variables and the observations.

The document is divided into six chapters. The first chapter sets the background of this work, where both biological and mathematical notations and concepts are presented and gives a history of the motivation behind genetics and genetic epidemiology. The second chapter present an overview of the statistical methods currently in use for genetic epidemiology. The third chapter deals with the identification of gene-environment interactions. It includes a presentation of existing approaches for this problem and a contribution of the thesis. The fourth chapter brings off the problem of meta-analysis. A definition of the problem and an overview of the existing approaches are presented. Then, a new approach is introduced. The fifth chapter explains the pleiotropy studies and how the method presented in the previous chapter is suited for this kind of analysis. The last chapter compiles conclusions and research lines for the future.

Remerciements/Acknowledgements

It is with great pleasure that I can dedicate the following lines to acknowledge all the people that have supported and helped me during my work during those three years.

First, I acknowledge my supervisors Benoit Liquet and Borja Calvo for directing my thesis. I felt grateful and fortunate to have you directing my work. Thank you Benoit for the constant interest you put in me. Thank you Borja for your insight and your advices that always came at the right moment.

I want to thank a lot Stéphane Robin and Christophe Ambroise for reviewing my manuscript and all the interest you showed for my work. I would like to thank also Hélène Jacqmin-Gadda for supervising the defense of this thesis and Astrid Jourdan for participating to it.

I thank the LMAP (Laboratoire de Mathématiques et leurs Applications de Pau) for hosting me. I would like to thank all the people from the center of Anglet for all the daily moments at work. I would like to thank also the Intelligent System Group in Donostia for their warm welcome. I acknowledge the UPPA and the UPV/EHU for funding my thesis and hosting me. I would like also to thank the french association Ligue contre le Cancer for its financial support.

I want to give a special thank to my parents and friends that helped me by their presence in this work. I have a thought for all the people I met during this thesis that were part of this journey.

Contents

Background

Chapter 1

Molecular biology

The material in this dissertation deals with the statistical analysis of biological data. Before going into the details of the contributions of this thesis we need to review some fundamental concepts. In the first part of this chapter basic molecular biology concepts are presented. Then, computation biology and its vocabulary is introduced. Finally data used for analyses are presented.

1.1 Biological background

1.1.1 Behaviour of a cell

The cell is the structural and functional unit of all known living organisms (cell comes from the Latin cellula, meaning, a small room). Some organisms have only one cell (they are called unicellular), whereas, other organisms are composed of several ones (they are called multi-cellular). For a human, the number of cells is estimated to be 10^{14} and a typical cell size is 10 μm (10^{-5}m) with an average mass of 1 ng (10^{-9}g). Each cell is a small zone of water and chemicals wrapped by a membrane, which can interact with its close environment, and perform specialized functions. Two types of cell exist: eukaryotic and prokaryotic. Eukaryotic cells are often found in multi-cellular organisms. Prokaryotic cells are usually independent and they show the simplest structure. Eukaryotic cells present in general additional internal structures compared to prokaryotic cells (Figure 1.1). Notably, the major difference between those is that eukaryotic cells contains a nucleus in opposition to prokariotic ones: a membrane-delineated compartment that houses the eukaryotic cell's DNA (Deoxyribonucleic Acid) as presented in Figure 1.2) .

Eukariotic cells have the following notable internal entities that ensure its behaviour:

- The **mitochondrion** provides the energy needed by the cell through enzymes exchanging electrons. In a cell, mitochondria extract the energy contained in the nutrients used by the cell, as well as doing many other specialized tasks. Each human mitochondrion has a chromosome composed of mitochondrial DNA. This DNA is distinct from the DNA that is located in the cell's nucleus.
- The **ribosomes** are a large complex of RNA (Ribonucleic Acid) and protein molecules, and they are essential in the production of proteins. Those proteins will determine the behaviour of the cell.

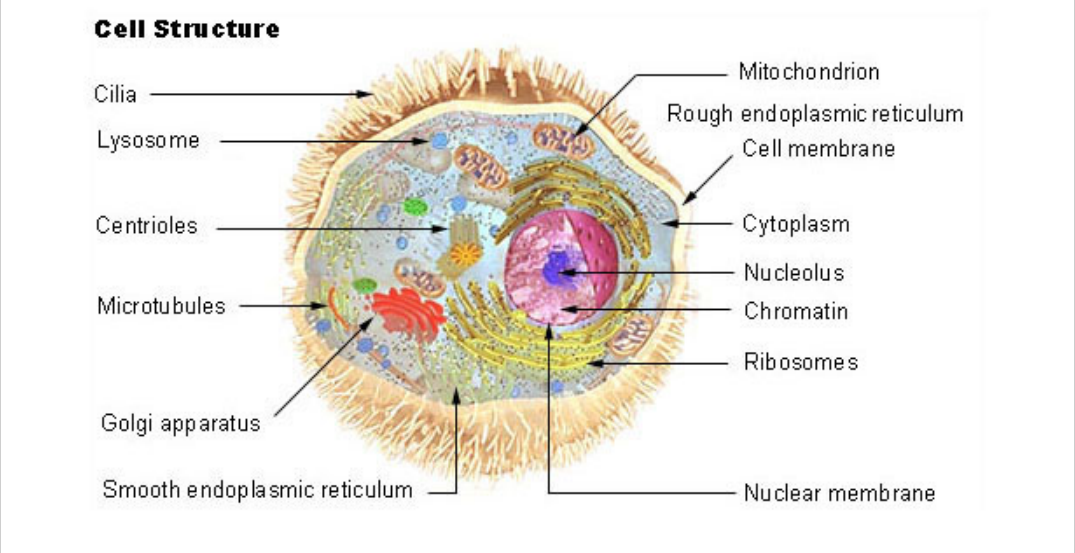


Figure 1.1 – Illustration of a the structure of a cell including: the nucleolus which is inside the nucleus, the ribosomes, a mitochondrion, the cilia, the lysosome, the centrioles, the microtubules, the golgi smooth endoplasmic reticulum, rough endoplamsic reticulum (cell membrane), the cytoplasm, the chromatin. Image taken from [1].

- The **cell nucleus** is the most noticeable structure in an eukaryotic cell. The nucleus has a spherical shape and is separated from the rest of the cell by a double membrane called the nuclear envelope. It is the location of the chromosomes (the structures that contain the genome), and it is responsible for maintaining them and controlling the activities of the cell by regulating the gene expression (i.e. the protein production). The nucleolus is a specialized region within the nucleus where ribosome are assembled. It is where the transcription process takes place (Section 1.1.3).

The nucleus of human cells (except for notable exceptions such as sperm cells) contains two sets of 23 chromosomes (22 regular ones and X or Y sex chromosome), each set inherited from each parent. Each chromosome is a very long DNA molecule that is wrapped tightly around proteins. Thus, chromosomes are structures of significant size (it can be seen under a light microscope). Differences in size and composition allow the 24 different chromosomes to be distinguished from each other through an analysis called a karyotype. Near the center of a chromosome is the centromere which is a narrow region that divides the chromosome into a long arm and a short arm. Chromosomes contain the information necessary to build a human being which is called the genome. It is often put in opposition with the phenotype which represents the external features of an organism.

1.1.2 DNA

DNA is made of a pair of molecules forming an helix structure. The molecules are held together by weak hydrogen bonds between base pairs of nucleotides. The double helix can be seen as an extremely long ladder twisted into a helix. The sides of the ladder are formed by a backbone of sugar and phosphate molecules, and the rungs consist of nucleotide bases weakly joined in the middle by the hydrogen bonds. The DNA chain is 2.2 to 2.6 nm wide, and one nucleotide unit is 0.33 nm long. Although each individual base is very small, the

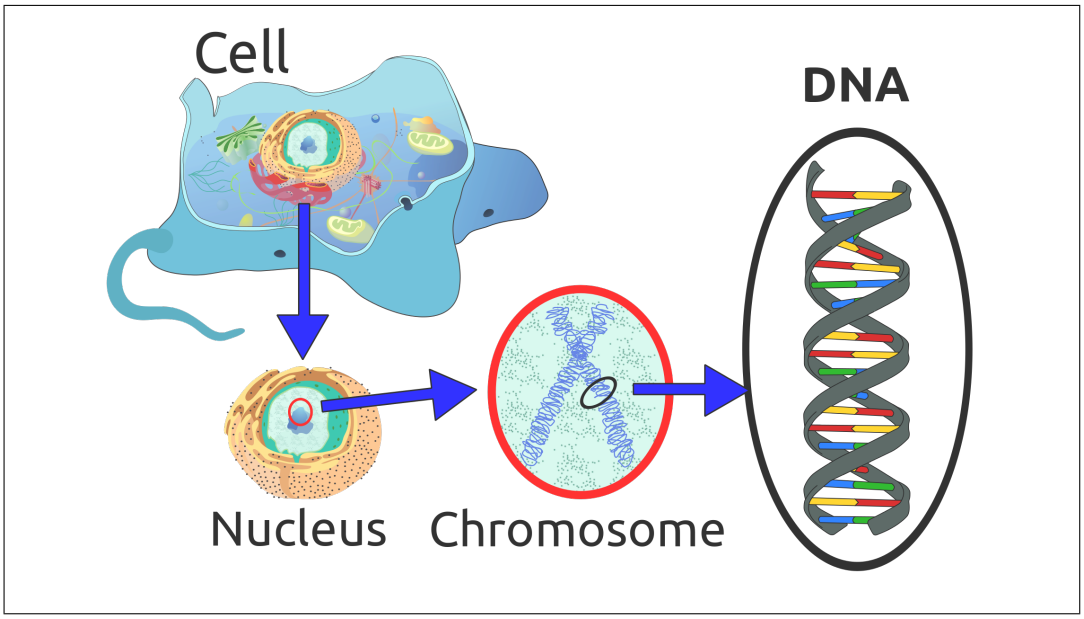


Figure 1.2 – Represenation of the DNA and its location in the cell. The cell (top left) contains the nucleus (bottom left) where chromosome are stored (middle). The chromosome is a long molecule with a double helix structure (right). Image taken from [1]

full molecule is a combination of millions of base units and is much larger. For instance, the largest human chromosome has around 300 million base pairs (bp).

DNA is coded in an alphabet with 4 basic units, the nucleotides. A DNA nucleotide is made of a five carbon sugar (a deoxyribose), a molecule of phosphoric acid, and a molecule called a base. The code letters for nucleotides are A, C, G, and T, which stand respectively for adenine, cytosine, guanine, and thymine. In DNA base pairing, adenine always pairs with thymine, and guanine always pairs with cytosine. Based on this, the information contained in one strand of the DNA can be inferred looking at the other strand. This information relies on the order of the nucleotides. DNA sequencing aims at retrieving this order. The genome of an organism is in fact made of the sequence of all the nucleotide bases for all the existing chromosomes.

Raw information in the genome must go through a certain chemical process before being able to be exploited by the cell. This genetic information must first be translated into proteins (amino acid sequences) by living cells. The canonical genetic code defines a mapping between DNA and the formation of the proteins.

1.1.3 From DNA to proteins

DNA is transcripted in RNA (Ribonucleic acid), a nucleic acid molecule similar to DNA. While DNA is only one molecule, the RNA is replicated in several copies in the cell which are then interpreted for creating proteins. The amount of copies of RNA molecules changes from one cell to another. This explains that even though every cell has the same DNA, their behaviour will be different depending on which part of the DNA is transcripted.

Contrarily to DNA, RNA is usually single-stranded and the sugar molecule of its nucleotides is a ribose rather than a deoxyribose. Furthermore, the nucleotide base thymine (T), that

is present in DNA, is replaced by uracil (U) in RNA. RNA play crucial roles in protein synthesis and other cell activities.

RNA is formed following DNA information. Synthesis of RNA is usually catalysed by the RNA polymerase which is an enzyme (a protein) that assembles the RNA from ribonucleotides. The transcription begins with the binding of the enzyme to a promoter sequence in the DNA. The DNA double helix is unwound by the enzyme and, then, the enzyme progresses along the template strand synthesizing a complementary RNA molecule. A DNA sequence indicates where RNA synthesis stops. There are more than thirty classes of RNA molecules. Among the most important ones we can cite the following three:

- **Messenger RNA (mRNA)** carries information from DNA to the ribosome where proteins are synthesized. The amino acid sequence in the protein is determined by the coding sequence of the mRNA.
- **Transfer RNA (tRNA)** is a short-chain type of RNA present in cells that carry a specific amino acid through the cell.
- **Ribosomal RNA (rRNA)** forms a site for the synthesis of the proteins.

The identification of the role of each part of the DNA (called genes) in the coding of specific proteins is then interesting. All humans have in common most of the genome, but the relatively few genomic differences between individuals account for most of the differences among individuals. There are estimated between 20 000 and 25 000 human protein-coding genes, although this number could drop. Human genes are distributed unevenly across the chromosomes: each chromosome contains various gene-rich and gene-poor regions. A gene contains both coding sequences that determine what the gene does, and non-coding sequences that determine when the gene is expressed (i.e. active) which have a crucial role in the control of the expression.

The most basic unit of genetic variation is the single nucleotide polymorphism or SNP. SNPs are single base-pair changes in the DNA sequence that occur with high frequency in the human genome [2]. Despite the high number of existing SNP, only some of them have an impact on the behaviour of cells through the transcription that leads to the production of proteins [3]. Genetic variants can also be called alleles.

SNPs exist in a variable proportion in human populations. The first information that is taken for a SNP is it's Minor Allele Frequency (MAF) that indicates whether the alleles know for that SNP are common or not. For instance, a SNP with a MAF of 0.30 implies that 30% of a population has the allele versus the more common allele (the major allele) which is found in 70% of the population (assuming that it has only 2 variants, which is the most common situation). An allele is called "rare" below 0.5% MAF, "low frequency" between 0.5% and 5% and "common" above 5% [4]. Depending on the categorization, the allele can be specific to small or large populations and the statistical tools used to assess an association can differ.

When a SNP is scarse in the population (i.e. rare variant), the SNP can be called a mutation. Cystic fibrosis is an example involving those rare variants. Results have shown that certain mutations show an association with a disease [5]. In those cases, the SNPs have a strong role in the biological process. Unfortunately, data on common variants can be easier to obtain and analyze, but SNP-disease association can be hard to find [6].

Studies on rare variants deal with small portions of the population being affected by a genetic trait. The idea of working with common variants aims at making the genetic studies beneficial for a larger portion of the population. It relies on the idea that common diseases have a different underlying genetic architecture than rare disorders which led to the development of the common disease/common variant (CD/CV) hypothesis [7].

The concept of linkage disequilibrium (LD) is commonly used for the correlation between SNPs. " Equilibrium " stands for the fact that, when a mutation appears in a population, corresponding data will be correlated in this population in opposition with other populations. Knowing the LD gives a priori information on the structure of the data.

Many measures of LD have been proposed, though all are ultimately related to the difference between the observed frequency of co-occurrence for two alleles and the frequency expected if the two markers were independent. The choice of measures for LD has been discussed in the litterature [8, 9].

1.2 Computational biology disciplines

With the increase in the computational capacity of computers and the amount of available biological data "bioinformatics" and "biostatistics" became key-words for various fields. Computational biology can be defined as a new interdisciplinary field that applies the techniques of computer science, applied mathematics and statistics to address biological problems. Being a very general definition, it encompasses many different fields. Here is a list of different approaches that have been taken.

- **Computational biomodeling** builds computational models of biological systems. These biomodels try to mathematically emulate the biological mechanisms involved in a particular system.
- **Computational genomics** is a field that studies the genomes of the organisms. It aims at understanding the genome and, particularly, the principles of how DNA controls the biology of any species at the molecular level. This is the axis studied in this document.
- **Molecular modeling** similarly to the computational biomodeling, models the behaviour of molecules (hence in a smaller scale). The fields of application can range from small chemical systems to large biological molecules. Typical examples are potential or energy functions that simulate biological behaviours by mathematics.
- **Protein structure prediction** aims at discovering the shape of the proteins, how it folds in different environments. The observation of the 3D structure of the proteins being extremely expensive, mathematical models are built to guess the possible shapes.
- **Computational biochemistry** makes extensive use of structural modeling and simulation methods in an attempt to elucidate the kinetics and thermodynamics of protein functions. Differential equations without a close solution and computational optimization are used to adjust a kinetic model corresponding to the real measures obtained by experimental methods.

1.3 Genetic epidemiology

Among the previously listed domain of computational biology, genomics focuses on DNA analysis. After decades of research, several biological mechanisms have been highlighted and different statistical approaches have been developed. Genetic epidemiology aims at understanding the role of genetic factors in the development of diseases. The issue consists in knowing which genetic factors are associated with the onset of a particular illness. For this purpose it is necessary to compare the genotype of people with the phenotype traits in question.

Since past decades, data analysis applied to high dimensional data has become essential, especially in biostatistics. Extracting information from ever larger data has become a trend in numerous fields. The key word "Big Data" has arisen to describe this kind of analysis which covers a variety of fields such as experimental data in physics (plasma data, astronomy data), or internet data (data from Facebook, emails). Medicine and biology are also involved in the increasing amount of data generated [10]. While the existence of the genome was discovered early in the 20th century, the ability to extract all the genome information on consequent amount of population have been quite recent. In the 70's extraction of sequences of the genome was performed with extremely expensive methods and huge means. Nowadays, a whole genome can be sequenced within the day at an affordable cost (less than 1 000 dollars); technological leaps have permitted to gather data at a much higher pace.

This amount of data have permited to explain more disease/trait heritability and have led to advances in genetic epidemiology [11].

Because of this need for managing ever-growing data the development of methods adapted for high-dimensional data is rising. The problem concerns not only the volume of data but also the variety of data.

1.3.1 Statistical focuses

From a statistical point of view, data are collected in a data set. Observations are performed on variables. Some variables are related to the genome and are predictors. Other variables (also called "traits") are related to the phenotype which is related to a disease in genetic epidemiology, they are outcome that must be explained from the predictors. Additional variables can exist and help building models and they are called covariates.

Variables can be either quantitative or qualitative. Qualitative variables represent types and are not countable, where as quantitative variables are the results of a measure. Models are inferred on the data by means of a given method or algorithm. They can either explain the data or being able to make predictions about the outcome when predictors are observed for a new case. When outcome data is qualitative, the prediction is called classification.

1.3.2 Association Studies

A first area of investigation consists in identifying the association between variants of SNPs. It permits to characterize a population by establishing correlation between variants. The study of this correlation leads to establishing of linkage disequilibrium information which represents the correlation between SNPs.

At the larger scale of association studies, Genome Wide Association Studies (GWAS) analyse DNA sequence variations at the genome level.

The first sequencing of a whole genome was achieved in 2003 [12]. Since then, the first Genome Wide Association Study, was conducted in 2005. 96 patients with age-related macular degeneration (ARMD) were compared with 50 healthy control individuals [13]. It was the first time the whole genome of patients was gathered on a population and the study identified two SNPs with significantly altered allele frequency from one group to the other. The establishing of a link from a statistical point of view was followed by the establishing of a link from a biological point of view.

A major motivation on GWAS is to highlight DNA sequences associated to a phenotype with the hope of revealing new biological mechanisms. Another motivation is providing models that help for the medication of patient after a genetic test. For instance, after the discovery of the influence of DNA sequences on warfarin dosing for blood treatments [14], a genetic test was proposed. Using genetic information from the patients to adapt and optimize their treatment is one of the main goals of personalized medicine.

1.3.3 Gene-Environment Interaction

Another approach consists in considering the gene-environment interaction: the exposition to a certain environment can act as a catalyst for the expression of a given genetic feature. Identifying such an interaction allow to highlight patient that would be more likely to have a certain disease under the exposition to a given environment. From a statistical point of view, the identification of interaction effect requires specific statistic tools. The problem is discussed in further details in Chapter 3.

1.3.4 Multi-level data analysis

As it will be explained in Section 1.3.7, data can be related to either the genome, the transcription, or the proteins. Most of the analyses in genetic focus on one type of data, but some approaches build models taking into account several of those type of data. In the context of this manuscript, only one type of data (genomic) is considered.

1.3.5 Meta-analysis

Meta-analysis is the study of data from a given omic data type but coming from different datasets. In this case biases between them cannot be quantified. In this case, separate models can be applied on each source, but the comparison of the results and the formulation of global model need a particular work. Some issues can appear like batch effects due to biases. The matter is explained in more details in Chapter 4.

1.3.6 Pleiotropy

Pleiotropy is the field that studies the case where one or more genetic factor have an effect on one or more phenotypes [15]. The problem can be seen either as a multivariate analysis or a meta-analysis problem and, thus, it can be tackled with several statistical approaches such as multivariate models, set based models or meta-analysis models. The subject is discussed in further detail in Chapter 5.

1.3.7 Data sources

A recent English neologism, omics, has been used to address biology fields that study cell molecules in their entirety. The first example is genomics, which stands for the study of the whole genome. In general, the term omics focuses on large scale and holistic data research to understand information contained in cells. Genomic data are the focus of this document and a list of the current database sources is given below. The study of genomic data is called genomics.

There are other types of omic worth of mentioning. Due to the success of high-throughput biological devices and new analytical tools, the suffix -om- has also been picked up by a wide array of other large-scale quantitative biology fields. Some of them are very recent terms that are not fully accepted by all the research community. Some of the most established ones are presented in the following.

Genomics

At first DNA data was sequenced pair by pair [16]. Since then, the methods of acquisition have been improved. Nowadays, two ways of acquisition of DNA data are mainly used: Microarrays and deep sequencing technologies (NGS). Microarrays are based on the hybridization of DNA [17]. Genetic information is retrieved through probes that correspond to predefined location in the genome. NGS (next-generation sequencing) relies on the fragmentation of DNA into pieces that are then sequenced and assembled to recontruct the complete sequence [18]. Microarrays can retrieve information faster to process but also need an a priori knowledge of locations in the genome while NGS is longer to process but can find new sequences.

The collection at genome wide level of DNA sequencing have been lead by consortia of institutions and laboratories. A list of the main ones is:

- The international HapMap Project established a catalog of SNPs mostly in European populations but also in Africa (the Yoruba population), in China (the Han population) and in Japan (Tokyo) [8], [19]. To increase the number of SNPs new projects have been lead and the project has been reaching 11 populations on its last version [4]. This collection of data has been made possible thanks to the edification of a nomenclature for the human genome project [20].
- The 1000 Genomes Project was concluded in 2015, with 2504 genomes sequenced from 26 populations [21].
- The UK 10K project propose an even larger number of observations[22].
- The 100k Genomes Project is one of the latest project [23].
- Precision Medicine Initiative is another of the latest project [24].

These databases allow to define the genetic variation in large populations. In genetic epidemiology, the genome of individuals showing given diseases are the main focus. The problem is more specific and hence, databases have less observations, and can show a reduced number of variables. Despite not being necessarily genome-wide studies, those genome studies are still subject to the same issues in the analysis.

Transcriptomic data

Transcriptomcis deals with the recollection of sequences of mRNA. Data on this kind of information can come from either microarrays or RNA sequencing and are analyzed through different steps. In microarrays, sequences of interest in the mRNA are a priori known. They are isolated through an array of probes, each probe corresponds to a feature that have been determined a priori. Data are compiled recollecting information from each probe. In RNA sequencing the full sequence of ribonucleic acid is obtained and is analysed. From one side, RNA sequencing allow to discover new RNA markers. From the other side, microarray can recollect data at a higher pace. Both approaches are used in a back and forth process where RNA sequencing help building a priori knowledge for making RNA sequencing.

We can note that genomic data and transcriptomic data have similarities: they have a large number of variables, each variable being related to a gene. This is the reason why statistical methods used for one of the type of data is also used for the other. They give an insight into the role of a gene in the behavior of a cell. Especially, it can give an apriori knowledge of the parts of the DNA that are most likely to have an impact on a cell (remember most of the DNA is not transcripted [25]). eQTLs (Expression quantitative trait loci) identify loci related to gene expressions. A large amount of loci has already been determined [26].

Proteomic data

Proteomic data are related to the proteins in the cell. The sequencing of the proteins on a basis of 20 amino acids is determined. Then, the shape of the molecule, i.e. how the molecule is folded, is inferred. This shape determines the function of the protein [27]. The comparison of proteins present within each cell permits to understand the dynamics in the cell. This information can be used for diagnostic but also for designing drugs with specific targets [27]. The field has its own databases like [28] and protein-protein interaction can be studied similarly to SNP data [29, 30].

1.4 Conclusion

Genetic studies aim at understanding the mechanisms ruling the behaviour of the cell. Genetic epidemiology raises hope for new medical approaches for treating diseases with preventive medicine or risk assessments in the application of drugs. The increase of results in the fields is due to the recent arrival of a deluge of new data. In terms of statistics presented, new challenges are raised by this data, advancing in these challenges is the motivation of this thesis. Especially, in the following chapters, studied data will be genomic data and the aim is to deal with genetic epidemiology.

Chapter 2

Statistical methods for genetic epidemiology

Genetic epidemiology aims at explaining the influence of the genome on the onset of a disease. It leads to studying the association between a genotype and a phenotype, i.e. predictors and an outcome. In this chapter statistical methods aiming at analyzing such associations are presented. First the mathematical notations for biological data are introduced. Secondly the challenges of the analysis are developed. Finally, statistical methods for finding simple effects in genomics are presented.

2.1 Mathematical notations

Data are represented by $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, two matrices, representing n observations of p predictors and q independent variables. For any matrix A of size (a, b), for $i \in \{1, \dots, a\}$, its rows are noted $A_{i,\cdot}$ and, for $j \in \{1, \dots, b\}$, its columns are noted $A_{\cdot,j}$. For subsets $\tilde{a} \subset \{1, \dots, a\}$ and $\tilde{b} \subset \{1, \dots, b\}$ resp. row and column sub-matrices are noted $A_{\tilde{a},\cdot}$ and $A_{\cdot,\tilde{b}}$. For any vector v of size a, for $i \in \{1, \dots, a\}$, its elements are noted v_i and for subsets $\tilde{a} \subset \{1, \dots, a\}$, $v_{\tilde{a}}$ represents the elements of the vector corresponding to the positions in the subset.

The Frobenius norm on matrices is denoted $\| \ \|_F$. We note X^T the transpose matrix of X and the cardinal of a set S is noted $\#S$. The positive value of a real number x is noted $(x)_+ = \frac{|x|+x}{2}$ and is equal to the number if it is positive and equal to zero otherwise.

The notations for submatrices permit to represent a data matrix considering sets of observations and/or group of variables. Let us consider M different sets of observations in the data. Noting, for $m \in \mathbb{N}$, $\mathbb{M}_m$ a subset of $\{1, \dots, n\}$, let $\mathbb{M} = (\mathbb{M}_m)_{m=1..M}$ be a partition of $\{1, ..., n\}$ corresponding to the observations sets. We note $\#\mathbb{M}_m = n_m$. Row blocks are defined by this partition and a representation is given in Figure 2.1 (observations are assumed to be ordered by observation set).

Let us consider that the variables are gathered in K groups. Let $\mathbb{P} = (\mathbb{P}_k)_{k=1..K}$ be a partition of $\{1, ..., p\}$ corresponding to this variable group structure. We note $\#\mathbb{P}_k = p_k$. We then we have $\sum_{k=1}^{K} p_k = p$. Column blocks are defined by this partitions as presented in Figure 2.2 (observations are assumed to be ordered by observation set).

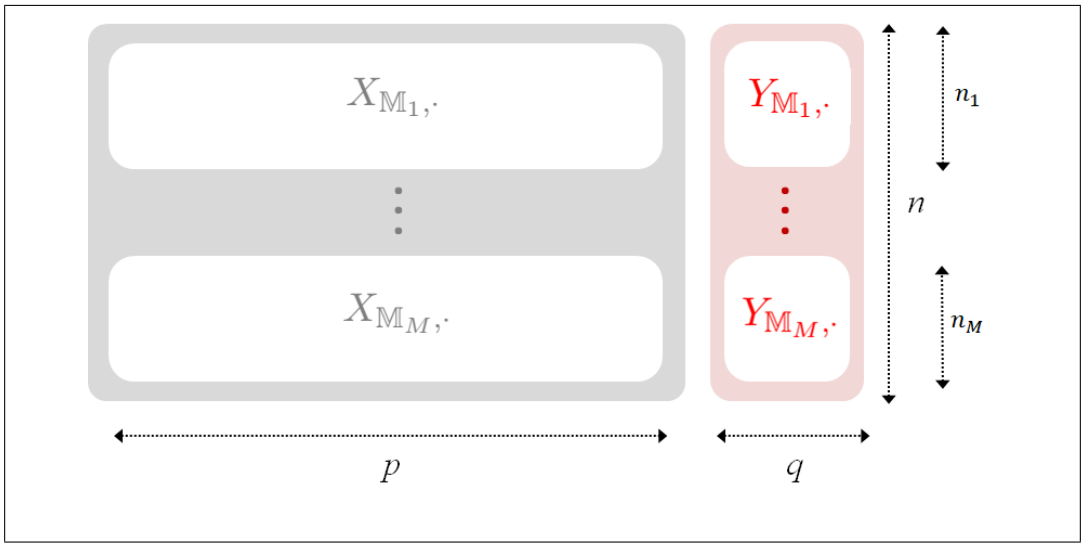


Figure 2.1 – Illustration of data structured by groups of observations. Observations are assumed to be ordered by observation set. p represents the number of variables of matrix X, q the number of variables of matrix Y and n is the number of observations. $n_1, \cdots, n_M$ are the resp. the number of observations of each observation set.

Both observation set structure and variable group structure can be defined at the same time as shown in Figure 2.3.

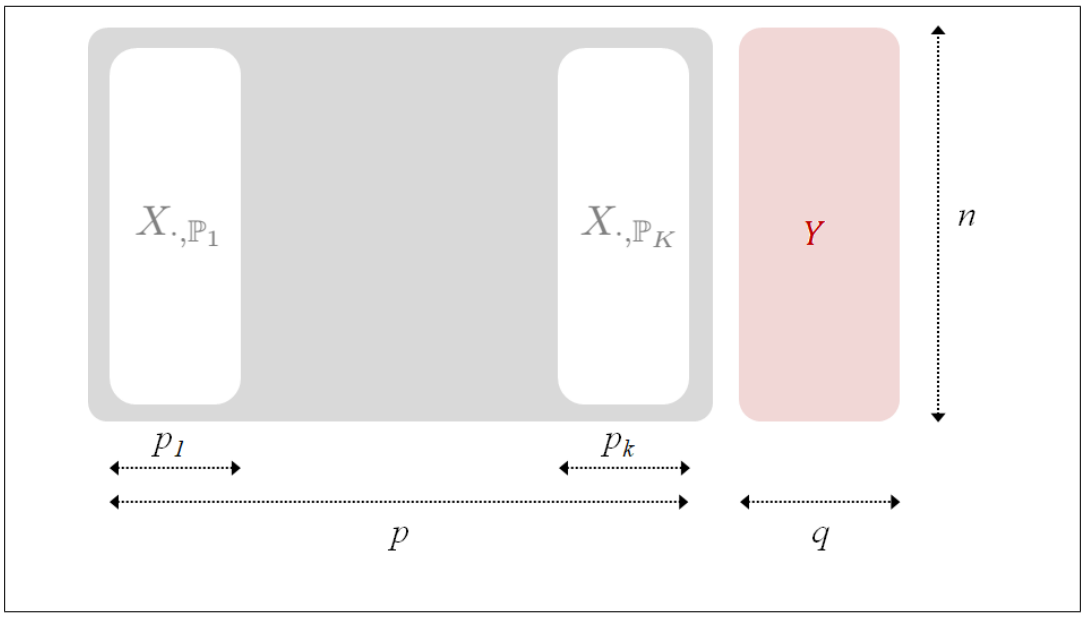


Figure 2.2 – Illustration of data structured by groups of variables. Variables are assumed to be ordered by observation set. p represents the number of variables of matrix X, q the number of variables of matrix Y and n is the number of observations. $p_1, \cdots, p_K$ are the resp. the number of variables in each group of variables.

In general, observation sets can represent the fact that different sets of observations come from different sources and must be analyzed accordingly. For instance, data coming from

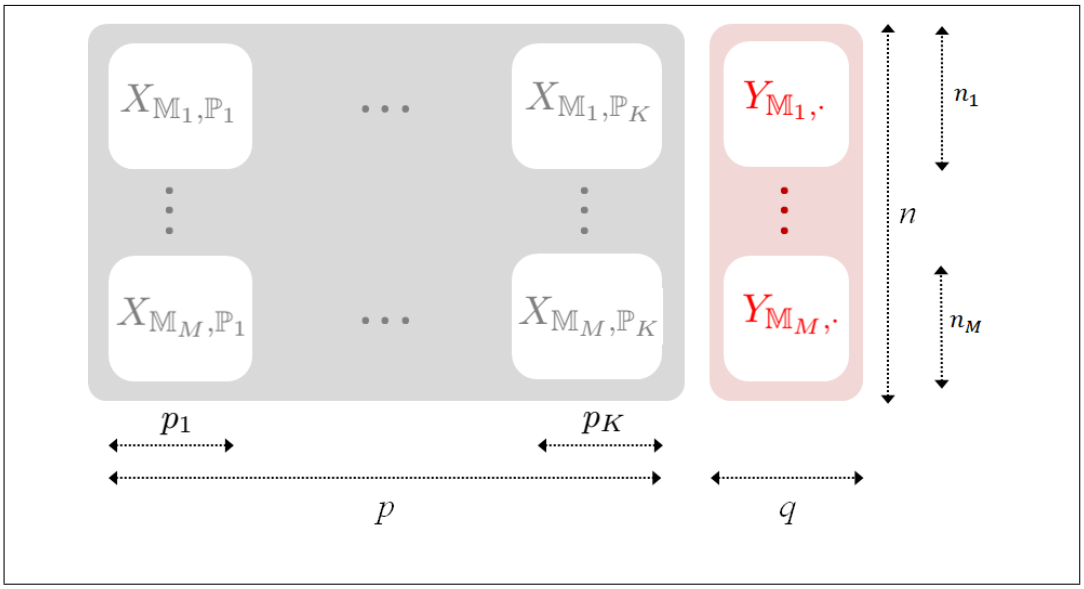


Figure 2.3 – Illustration of data structured by groups of variables and groups observations. Observations and variables are assumed to be ordered by resp. observations sets and variable groups. p represents the number of variables of matrix X, q the number of variables of matrix Y and n is the number of observations. $n_1, \cdots, n_M$ are resp. the number of observations of each observation set, $p_1, \cdots, p_K$, are the resp. the number of variables in each group of variables.

different studies may present biases. Group of variables can represent either a set of variables that is known to be quite correlated or a group of variables that must be treated together. For instance, in genetics, a locus defines a group of SNP variables.

2.2 Statistical challenges for genomics analysis

Due to the nature of the biological process and the method of acquisition of data, SNP data present particular statistical challenges. The main challenges are discussed in this section.

2.2.1 Ill-posed problems

In genomic data, the number of variables is often much bigger than the number of observations. Indeed, the number of variables is extremely large due to the length of the studied molecule while the number of observations remains smaller due to the cost of their acquisition. A large portion of the existing data analysis tools are not suited for this kind of data which is the reason why, in mathematics, such a case is called an "ill-posed" problem. As an example, the limitation of linear regression in this case is presented.

Let us consider the following linear model :

$$Y = X\beta + \varepsilon \tag{2.1}$$

where Y is a variable to explain, X are the predictors β are the coefficent of the linear regression and ε is the noise term which is assumed to follow a normal distribution.

The estimated coefficients of the linear regression $\hat{\beta}$ are chosen in order to minimize the residuals.

$$\hat{\beta} = \underset{\beta}{argmin} \|Y - X\beta\|_2 \tag{2.2}$$

The solution is

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.3}$$

and then the predictive outcome is

$$\hat{Y} = X(X^T X)^{-1} X^T Y \tag{2.4}$$

However such a model requires the inversion of the matrix $X^T X$ which cannot be calculated when the number of variables is larger than the number of observations.

In the following, each method used copes with the problem of $p \gg n$.

2.2.2 Correlated variables

In genomics, a large number of variables can be correlated. Notably, it is known that genes associated to a same gene will be highly correlated. In statistics, correlation can have a huge impact in the accuracy of the methods. Having correlated data imply that similar information belongs in several variables. Hence, it affects the performance of statistical methods for which the lesser variables exist, the better they behave. Furthermore, the addition of variables makes the results harder to interpret as several variables will have synonymous roles in the model. After years of analysis of the genome, sets of SNPs or genes have been highlighted. The location of a gene or a SNP of interest is often called locus (loci for plural). A set of genes that are known to be involved in a same biological process is called a pathway. This common representation leads to consider group of SNPs in general, as SNPs from a same gene are often highly correlated. We can gain power and reduce the number of tests by combining weak signals from SNP-level analysis. This is why in genomics, making inference at a group of variables is preferred. Over the recent years, numerous pathway analysis for GWAS data have been proposed in the literature for finding pathways associated with the studied disease [12, 31].

Two approaches compete for pathway analyses. From one side, self-contained GSA methods test whether genes within a pathway are collectively associated to the phenotype of interest [32]. Most widely used self-contained GSA methods combine either SNP-level or gene-level p-values into pathway test statistics, where their significance is typically assessed through permutation testing. From the other side pathway analysis has also been shown in enriching GWAS results by identifying SNPs that were missed by single-SNP analyses [33, 34, 35].

2.2.3 Variety of type of variables

While the first two challenges where related to the nature of the predictors, another challenge concerns the definition of the outcome. In genomic studies, phenotype of interest can either be quantitative or qualitative and statistical method exist for each case. When

a physical feature can be measured in an organism, the outcome is quantitative. When an expert performs a diagnostic on an organism, a trial is set and the variable is then qualitative. However, statistical methods must be adapted to the nature of the outcome.

Some tricks can allow to pass from one case to another. For instance, threshold can be set on a quantitative outcome to create categories, but a biological incentive is needed. For the contrary case, a qualitative variable with r different possible values can be seen as r dummy variables which are equal to 0 or 1 and that are then considered as quantitative variable between 0 and 1. However, a loss of power in the method can arise due to this transformation.

When a method is developed for genomics, it is important to be aware to which case it can be applied.

2.3 Methods for genetic epidemiology

In this section, a list of methods that have been developed for genetic epidemiology is presented.

2.3.1 Group of variables structure

As we have seen in Section 1.1.3, certain SNPs tend to be highly correlated due to their being transmitted together. Moreover, a single gene typically contains several SNPs and several genes take part in the same pathway. These facts can be used to reduce the complexity of the resulting models, incorporating the information in the methods. One way to achieve this is by considering a group structure for the variables, as shown in Figure 2.2.

The approaches dealing with this structure can be classified into two groups. A first group of approaches consist in summarizing the SNP data to a gene-level data or a pathway level data in a pre-process step. The obtained data are supposed to overcome the challenges listed bellow and a second step analysis can be performed on data showing less intrinsic correlation and the number of variables is reduced. A second group of approaches uses this group structure and integrates into the framework of a statistic tool.

In the first set of approaches requires, there is a loss of information in the pre-process and a well-established a priori biological knowledge is required. Conversely, the second set of approaches, keeps all information and is then less stringent on the a priori assumptions and less influenced by a pre-preocessing step. Its drawback is that integrating the a priori information to the framework needs investment in the mathematical analysis and in the coding of the implementation.

2.3.2 Pre-processing methods

As a first simple approach, a pre-processing can be performed. A first pre-processing method consists in summurizing closely related variables in a weighted sum. The idea is that single variables may have a low effect on a disease but the overall effect may be more noticeable. The use of the weighted sum principle has been proposed on linear model [36] and on Fourier transform analysis [37]. Other methods called screening methods, perform low time cost tests on variablse to identify variables with higher chances of being relevant for association. A pre-selection of variables is made which can be followed by a second

step analysis on data with a reduced number of variables [38, 39, 40, 41]. The methods presented in the following integrate the group structure to their framework.

2.3.3 Dimension reduction methods

Dimension reduction techniques is a set of popular methods that have been used in genetic studies, being the most common ones Canonical Correlation Analysis (CCA) [42], Principal Component Analysis (PCA) [43] and Partial Least Square (PLS) [44]. All these methods rely on the projection of the data into a subspace of lower dimension which represents most of the variation of the data, and they are often posed as an eigen value problem [45]. PLS and CCA are both supervised models and differ from the norm used whereas PCA is unsupervised. PCA was developed in the early days of statistics [46] and have been raised interest in last decades since the number of available variables has become larger and larger. It has been one of the first methods to be used to study GWAS . PLS was introduced later by H. Wold [47]. It has had succes in domains with a large number of variables to analyze, like Chemometrics [48], and GWAS [44]. CCA was proposed by Hotelling [49] and its application to GWAS came even later than the two previous methods[42]. All methods called "dimension reduction methods" derive from those ones.

Principal Component Analysis

In PCA we search for two linear regressions, u, of the variables of X that explains most of the covariance

$$u = \underset{\|u\|_2=1}{\operatorname{argmax}} \|Xu\|_2 \tag{2.5}$$

Partial Least Square

For PLS, we search for a linear combinations, u and v, of the variables of X and Y that explain most of the covariance of X and Y :

$$(u, v) = \underset{\|u\|_2=\|v\|_2=1}{\operatorname{argmax}} \left\|u^T X^T Y v\right\|_2 \tag{2.6}$$

Canonical Correlation Analysis

The CCA is similar to PLS, we search for linear combinations, u and v of the variable of X and Y that explain most of the correlation, instead of the covariance for the PLS, between X and Y.

$$(u, v) = \underset{\|u\|_2=\|v\|_2=1}{\operatorname{argmax}} \frac{\left\|u^T X^T Y v\right\|_2^2}{\|Xu\|_2 \|Yv\|_2} \tag{2.7}$$

Both PLS and CCA are closely related and are generally applied in fields with similar properties but they have difference statistical meaning. The study of the correlation (CCA) impose a normalization of the analyzed data. In practice, for most of the analysis, data are already standardized and hence PLS and CCA give similar results.

Sparse formulations

In these dimension reduction techniques, results are formulated with new variables that are linear combinations of the original ones. These combinations can be hard to interpret due

to the huge number of coefficient they represent. To cope with this problem, Lasso methods have been used [50]. Lasso's penalization shrinks to zero the participation to the model of the least relevant variables. Results highlight a smaller number of variables that are easier to explain. In addition, the noise in the signal is reduced and the power of the methods is boosted. In particular sparse Partial Least Square (sPLS), has shown encouraging results [51]. In the rest of the document the Lasso penalization is applied only to the weights associated to X but the Lasso penalization could also be applied to the weights associated to Y.

For PLS the sparse formulation with a Lasso penalization on u is as follows :

$$\begin{aligned} \min_{||u||=1,||v||=1} \left\| Z - uv^T \right\|_F^2 + \lambda P(u) \\ P(u) = \sum_{i=1}^{p} |u_i|_1 \\ Z = XY^T \end{aligned} \tag{2.8}$$

where λ is a parameter controlling the sparsity of the model. The larger the parameters, the less the number of non-null elements in the linear combination.

Group of variables

The Lasso penalization can be adapted when groups of variables are known a priori. In genetics, incorporation of this grouping structure is becoming increasingly common due to the success of gene set enrichment analysis approaches [52]. Using a model that takes into account this variable group structure allow to improve the performance and the readability of the results. To this purpose, the group Lasso has been proposed for the linear regression and used in genetics [53]. The sparse group Partial Least Square (sgPLS) adapts the penalization to the PLS [54] where two overlaid Lasso penalizations translate the group structure in the Partial Least Square formulation. A structure with group and sub-groups can also be handled by this generalization with three overlaid Lasso penalizations which is called sparse group sparse PLS (sgsPLS) [55].

The application of a group Lasso for the PLS is :

$$\begin{aligned} \{u_{opt}, v_{opt}\} = \underset{u,v}{\operatorname{argmin}} \left\| Z - uv^T \right\|_F^2 + \lambda P_{group}(u) \\ \text{with } P_{group}(u) = \sum_{k=1}^{K} \sqrt{p_k} \left\| u_{\mathbb{P}_k} \right\|_2 \text{ and } Z = X^T Y. \end{aligned}$$

where $\mathbb{P}_k$ refers to the notation for group of variables from section 2.1.

A group Lasso level penalization and a classical lasso penalization can be added up in a same model.

$$\begin{aligned} \{u_{opt}, v_{opt}\} = \underset{u,v}{\operatorname{argmin}} \left\| Z - uv^T \right\|_F^2 + \lambda \left(1 - \alpha\right) P_{group}(u) + \lambda \alpha P_{variable}(u) \\ \text{with } P_{group}(u) = \sum_{k=1}^{K} \sqrt{p_k} \left\| u_{\mathbb{P}_k} \right\|_2 \text{ , } P_{variable}(u) = \sum_{i=1}^{p} \left\| u_i \right\|_2 \\ \text{and } Z = X^T Y. \end{aligned}$$

Where α is a parameter between 0 and 1 controlling weather the group penalization or the individual variable penalization have more impact, the lower the parameter, the more group penalization is prevalent.

Case of the overlapping groups

The solution of the optimization problems presented above can be obtained through an SVD (Singular Value Decomposition) algorithm or a NIPALS algorithm [56]. However if the groups of variables are overlapped the solution is hard to obtain. Some methods proposed to integrate the overlap in the algorithm of the solution whereas Jacob [57] proposed to duplicate the variables related to the overlap. Pathways group Lasso with adaptive weights accounts for the presence of overlapping pathways [58]. It has been used to identify pathways associated with a trait of interest and its application [59]. Several other methods were also proposed for handling overlapping group structure and it is still a field of interest [60, 61]. In this dissertation, the method from Jacob is preferred as duplicating do not force a modification of the framework of an algorithm. It is the easiest approach to implement the possible increment of the number of variables.

2.3.4 P-value combination methods

Methods presented above have the common feature of trying to make an overall model of the data based on an underlying structure assumption. Other methods only aim at testing associations. It is the case of the methods presented in this section. While former approaches can provide a larger interpretability of the model and formulate predictions, latter need less stringent hypotheses on the data.

The idea behind p-value combination methods is to perform statistical tests for single variables and combine them to a group of variables result. In epidemiology genetics, groups of variables are typically genes (group of SNPs) or pathways (group of genes).

Fisher's method

Fisher's method is a well-established p-value combination method that combines the p-values from multiple statistical tests. The FM test statistic equals

$$\text{FM} = -2\sum_{\ell=1}^{L}\log\left(p_\ell\right) = -2\log\left(\prod_{\ell=1}^{L} p_\ell\right), \tag{2.9}$$

where L is the number of tests combined (for example the number of SNPs within a pathway). Under null hypothesis (that the p-values are independent) FM test statistic follows a χ^2 distribution with $2L$ degrees of freedom.

Adaptive rank truncated product (ARTP)

The idea behind the ARTP is to truncate the highest p-values in the FM method. The only p-values left are the most significant ones. We denote the ordered statistics of those p-values $p_{(1)} \leq \ldots \leq p_{(L)}$, with $p_{(\ell)}$ being the ℓ-th smallest p-value. The original RTP statistic given by

$$W_K = \sum_{k=1}^{K}\log\left(p_{(k)}\right) = \log\left(\prod_{k=1}^{K} p_{(k)}\right) \tag{2.10}$$

combines the K smallest p-values statistics of the tested pathway [62]. In the adaptive RTP method J different truncation $K_1 \leq \ldots \leq K_J$ are investigated. Let $\hat{s}(K_j)$ be the estimated p-value for W_{K_j}, $(1 \leq j \leq J)$. The ARTP statistic is then defined using the minimum p-value procedure

$$MinP = \min_{1 \leq j \leq J} \hat{s}(K_j). \tag{2.11}$$

Let $p_1^{(0)}, \ldots, p_L^{(0)}$ be the p-values for each interaction test on the null hypothesis based on the observed data. B datasets under the complete null hypothesis $H_0 = H_{0,1} \cap \ldots \cap H_{0,L}$ using appropriate resampling procedure are generated. Let $p_1^{(b)}, \ldots, p_L^{(b)}$ be the p-values for each interaction test on the null hypothesis based on the b-th generated data set, $1 \leq b \leq B$. The RTP statistic

$$W_j^{(b)} = \sum_{i=1}^{K_j} \log\left(p_{(i)}^{(b)}\right), \quad 0 \leq b \leq B,\ 1 \leq j \leq J \tag{2.12}$$

is calculated for each truncation point, for both the observed data set and each of the B simulated datasets. A p-value is computed

$$\hat{s}_j^{(b)} = \frac{\sum_{b^*=0}^{B} I\left(W_j^{(b^*)} \leq W_j^{(b)}\right)}{B+1}, \quad 0 \leq b \leq B,\ 1 \leq j \leq J \tag{2.13}$$

for each W_j. The p-value for the ARTP statistic $MinP^{(0)}$ of the group of variable is estimated as

$$\widehat{\mathrm{p}}_{ARTP} = \frac{\sum_{b^*=0}^{B} I\left(MinP^{(b)} \leq MinP^{(0)}\right)}{B+1}, \tag{2.14}$$

where

$$MinP^{(b)} = \min_{1 \leq j \leq J} \hat{s}_j^{(b)}, \quad 0 \leq b \leq B,\ 1 \leq j \leq J. \tag{2.15}$$

In order to compute this minimum, several p-values are needed for the same variable. A resampling of the data can provide those p-values, it can either be, permutations, bootstraps or parametric bootstraps.

Resampling methods

Both ARTP and FM rely on appropriate resampling strategy to generate data set under the null hypothesis considered.

Yu [63] proposed a permutation procedure to evaluate the significance level of the ARTP statistic. An alternative to the permutation procedure is to used a parametric bootstrap procedure [64, 65]. For permutation procedures, the number of valid permutation can be limited while their is not such a problem for parametric bootstrap [66].

Methods that perform a gene-based pathway analysis first combine the p-values of single-SNP analysis into gene-level test statistics (or p-values) that are subsequently summarized into pathway-level associations. The SNP-level p-values can also lead directly to a pathway-level analysis.

2.3.5 Variance component methods

Another group of methods has been popular in GWAS: the variance component methods. A joint test is propose at a group of variables level. The originality is to test the variance instead of means or weighting to capture information even with variables in different direction. This family of methods use commonly a variance component test [67] which has been used in genetic studies [68]. The burden test is a similar test that have also been used in genetics [36].

The most popular application of these types of methods to genetics is called SKAT [68]. It is especially suited for dealing with the rare variants case, but it can also be used for common variants. Indeed, a formulation combining rare and common variants has been formulated [69]. Improvements to its power and robustness have been also developed [70, 71].

A genotype can generally be related to the phenotype by a regression model. Let us consider a dichotomous phenotype y and a set of variables $i_1, \cdots, i_r$. The following logistic regression model can be proposed:

$$\text{logit}(\pi_i) = \alpha_0 + \sum_{l=1}^{r} \beta_{i_l} X_{i_l} \tag{2.16}$$

where logit is the function $x \mapsto \log\left(\frac{p}{1-p}\right)$ and π_i denotes the disease probability. X_i is the predictor and β_i, the regression parameter.

The regression model can be rewritten as

$$\text{logit}(\pi_i) = \alpha_0 + \beta_C \sum_{l=1}^{r} w_{i_l} X_{i_l} \tag{2.17}$$

where β_C iks a general coefficient parameter and w_{i_l} are weights that can depend on the minor allele frequency.

The burden score for testing the null hypothesis that $\beta_C = 0$ is:

$$Q_B = \left[\sum_{i=1}^{n} (y_i - \widehat{\pi}_i) \left(\sum_{j=1}^{m} w_j X_{i,}\right)\right]^2 \tag{2.18}$$

where y_j is the phenotype j.

The SKAT score is then:

$$Q_S = (y - \widehat{\pi})' K (y - \widehat{\pi}) \tag{2.19}$$

where $K = XWWX'$ is and $n \times n$ kernel matrix defined by the matrix of predictors X and a diagonal matrix of weights W.

The SKAT-O score combines burden and SKAT statistics in a weighted sum.

$$Q_\rho = \rho Q_B + (1 - \rho) Q_S \tag{2.20}$$

where ρ is a parameter that determine the participation of the burden score and the SKAT score.

The optimal Q_ρ is considered for SKAT-O

$$Q_{\text{optimal}} = \min_{0 \leq \rho \leq 1} Q_\rho \tag{2.21}$$

2.3.6 Other methods

The methods presented so far are the basis for the work in this dissertation, but are not the only approaches to the problems posed in genetic epidemiology. For frequentist side, methods are for instance c-alpha test [72], the group additive regression model [73], Tukey's model [74] and entropy-based methods [75]. Methods can also be Bayesian [76, 77] or been take from the machine learning domain [77]. We can also cite pairwise similarity based model [78] and U-statistic models [79, 80, 81].

2.4 Conclusion

Data analysis on omics data sets particular challenges in terms of statistics. Data have a low number of observations compared to the number of variables. Models must be able to extract information from this too wide variety of information and summarize it as readable results. A large number of variables are known to be correlated, which is a wide-spread problem in statistic in general and models must be resilient to it. Furthermore, variables are often gathered into groups of variables (like when different SNP data refer to the same gene). Models need to be able to give interpretations at group of variables results. In this manuscript, a special focus is given to dimension reduction methods and p-value combining statistics, and extensions are presented in both cases. In the rest of the document, the contributions of this dissertation to cope with these problems will be presented. The Chapter 3 is devoted to gene-environment interaction methods, while Chapter 4 and Chapter 5 deals with meta-analysis and pleiotropy. In both cases, after a brief summary of the work is presented and, then, the original papers where the contribution was published are reproduced.

Novel statistical methods for genetic epidemiology

Chapter 3

Gene-environment interaction methods

This part covers the study of gene-environment interactions (G×E) which accounts for situations in genetic epidemiology when the exposure to a certain environment catalyst the appearance of a disease. First, G×E interactions are presented. Then the statistical challenges raised by this kind of analyses are explained. After, current approaches used in the literature are presented. Finally, a novel method is presented. It is an extension of a combination test framework for a set based G×E interactions analysis, and it is the contribution of this thesis to G×E interactions studies. The methods have been highlighted by a piblication in the Journal of the French Society of Statistics [82]. The article is presented right after the chapter.

3.1 Definition of gene-environment interaction

A first definition of the phenomenom was given two decades ago [83]. In the case of dichotomous environments and phenotypes for instance, the analysis relies on having a population with individuals i) either exposed or not to an environment ii)with different disease diagnostic iii) with different genotype information. Observations need to span each possible combination. Then the question is to know if a population exposed to the environment and with specific genetic variants has disease appearance larger than the one that can be expected from a simple environment effect or a simple genetic effect. Formal genetic evidence for G×E interaction can consist in the observation that a certain exposure has different effects in different populations or ethnic groups or in people with different genetically determined phenotypes.

A first motivation, for these studies, is the idea of personalized recommendations in a public health perspective. Identifying G × E interaction helps identifying high-risk individuals if their environmental exposure is known. A second motivation is pharmacogenetics. The idea is to determine conditions enhancing the response of a patient to a treatment [84].

Interactions have been highlighted for a large variety of etiologies. Cigarette smoking has been identified as interacting with genetic markers for cancers [85, 86] and alcohol consumption interacts with genes in upper aerodigestive cancers [87]. Environment influence has also been noted for Parkinson disease [88], for example with coffee consumption [84].

3.2 Statistical challenges

From a statistical point of view interactions are often related to an interaction test. However, interaction effects are known to be harder to highlight. Especially, is has been settled that a four times larger database in terms of observations is needed for highlighting an interaction effect compared to a simple disease/trait effect [89]. This means that G×E questions are expected to need more extensive studies than usual.

The design of the recollection of clinical data takes a major role in this problem. For instance, in a first step, case-only studies were developed for G×E [90]. It appears that such studies rely heavily on strong hypothesis and the performances can be hindered when the hypothesis are violated. Then, methods using a case-control study were also used [91]. Unfortunately, case-control analysis suffers from low statistical power. In contrast, the case-only studies can be powerful in certain scenarios, although violation of the assumption of independence between the genetic and environmental factors can greatly bias the results. Some propositions to deal with the hypothesis reliance has been developed. A bayesian method trying to cope with this reliance was proposed [92] with a trade off between the hypothesis reliance and the method efficiency. Another model combines both case-only and case-control approaches in one method [93]. Nowadays the case-control approach is more common due to its freedom to assumptions.

3.3 Current methods

In general the global framework of statistical methods for detecting interaction can be quite similar to the simple effect case. If the methods rely on summary statistics at single variable levels, we can just take a summary statistic suited for interaction detection instead of simple effect detection, although interaction effects are harder to highlight.

3.3.1 iSKAT and GESAT

The SKAT method presented in Section 2.3.5 has been adapted to G×E interaction studies. The Gene-Environment Set Association Test (GESAT) [94] method adapts the SKAT framework for the analysis of common variants and the iSKAT method [95] adapt it for rare variants. The main difference with SKAT relies on the difference of regression models considered. For instance, for a dichotomous phenotype variable, a simple effects can be modelled by a regression model as

$$\text{logit}[P(Y=1)] = \alpha_i + \beta_i X_i \tag{3.1}$$

Where logit is the function $x \mapsto \log\left(\frac{p}{1-p}\right)$, X_i refers to the i-th SNP variable and Y is the outcome.

Under the null hypothesis, the model would be

$$\text{logit}[P(Y=1)] = \alpha_i \tag{3.2}$$

For G × E studies, the regression model would be

$$\text{logit}[P(Y=1|X_i,E)] = \alpha_i + \beta E + \beta_i X_i + \gamma X_i \times E \tag{3.3}$$

and under the null hypothesis, the model would be

$$\text{logit}[P(Y = 1|X_i, E)] = \alpha_i + \beta E + \beta_i X_i \tag{3.4}$$

3.3.2 P-value combination methods

Methods for combining p-values can be used for G×E interaction. In these methods, the p-value is computed for a G × E interaction like in iSKAT, as it has been described in Section 3.3.1.

A change in the resampling framework is also needed especially for permutations. Yu [63] use a permutation procedure to evaluate the significance level of the ARTP statistic in the context of disease-pathway association.

For the interaction model, the permutation procedure to generate data set under the complete null hypothesis fixes SNP and E and permutes Y for generating data. Although the procedure seems easy to implement, the number of a valid permutations can be restricted and alter the quality of the results. The alternative is a parametric bootstrap procedure [64] which implies less stringent assumptions [66, 65]. In our context, this procedure is chosen in the following.

3.4 Contribution to gene-environment interaction methods

In our contribution, we have decided to compare the Fisher's product test statistic and ARTP approach because they have been discussed in the literature to be the most powerful pathway analyses among combination tests [96, 97].

We choose to use an ARTP (Adaptative Rank Truncated Product) method. For SNP-environment interactions, p-values are computed and a result at the gene-environment and pathway environment is inferred. The chosen method uses a null hypothesis corresponding to a self contained approach [96]. In most of the p-value combination methods used in genetic epidemiology the null hypothesis is chosen to highlight simple effect. Instead, an interaction test is performed in the presented method. Also, the framework of the ARTP requires a resampling procedure. The permutation is widely used for this purpose. However, promising results with parametric bootstrap have been shown which is interesting as the method is more adaptable to cases where the number of observations is low [98]. The method still hasn't been used in a case of interactions. The proposed model aims at filling this gap.

The contribution has been used on real data.

The analysis is conducted in a population-based case-control study from France including 1 126 breast cancer cases and 1 174 controls. Data are composed of 23 genes gathered in one pathway: the circadian pathway. This is a case-control study conducted in Côte d'Or and Ille-et-Vilaine (France) [99]. Eligible cases were women aged between 25 and 75 years, resident in one of these two areas and diagnosed breast cancer between 2005 and 2007. A total of 1 232 incident breast cancer cases were included in the study. Controls were selected among women living in the same areas with no history of breast cancer. In total 1 317 controls were enrolled in the study. Genotyped data from a microarray for oncology (called oncoarray) are available on 1019 cases and 999 controls. The OncoArray chip targets

up to 500 000 variants with a genome-wide backbone of 250 000 tag SNPs. A preselection of SNP of interest can then be performed by epidemiologists.

The developed methods have been implemented in an R package [100].

3.5 Conclusion

This part presents the G $\times$ E interaction questions in general. We can note that there are not so many statistical methods dedicated to this problem and we can wonder how to use properties of existing statistical methods in genomics and apply them to G $\times$ E interaction. This approach have been followed in this chapter where the parametric bootstrap as a resampling is used while it has only been highlighted in the literature for simple effects but no G $\times$ E interaction effects.

Investigation Gene- and Pathway-environment Interaction analysis approaches

Camilo Broc, Marina Evangelou, Pascal Guenel, Therese Truong and Benoit Liquet

Published in Journal de la Société Française de Statistiques

Journal de la Société Française de Statistique
Vol. 159 No. 2 (2018)

Investigating Gene- and Pathway-environment Interaction analysis approaches

Camilo Broc[1], Marina Evangelou[2,3], Pascal Guenel[4], Therese Truong[4] and Benoit Liquet[1,5]

Abstract: Pathway analysis can increase power to detect associations with a gene or a pathway by combining several signals at the single nucleotide polymorphism (SNP)-level into a single test. In this work, we propose to extend two well-known self-contained methods, the Fisher's method (FM) and the Adaptive Rank Truncated Product (ARTP) method to the analysis of gene-environment (GxE) interaction at the gene and pathway-level. It has been previously suggested that the permutation procedures that are usually used to derive the significance of these tests are not appropriate for the analysis of GxE interaction and should be replaced by a bootstrap approach. We analyse and compare the performance of the extension of FM and ARTP using the permutation and the parametric bootstrap procedure in simulation studies. We illustrate its application by analysing the interaction between night work and circadian gene polymorphisms in the risk of breast cancer in a case-control study. The ARTP method, adapted for both gene- and pathway-environment interactions, gives promising results and has been wrapped to the R package PIGE available on the CRAN.

Résumé : Les analyses par pathway permettent d'augmenter la puissance statistique en combinant les signaux au niveau des SNPs pour définir des associations au niveau du gène et/ou du pathway. Dans cette étude, nous proposons d'adapter deux méthodes d'analyse par pathway, la méthode de Fisher (FM) et la méthode ARTP (Adaptive Rank Truncated Product), pour l'analyse des interactions gène-environnement (GxE) au niveau du gène et au niveau du pathway. Il a été précédemment suggéré que les procédures de permutations habituellement utilisées pour estimer la significativité de ces tests ne sont pas appropriées pour l'analyse des interactions GxE et devraient être remplacés par une approche Bootstrap. Ainsi, nous analysons et comparons dans une étude de simulation les performances de l'extension des méthodes FM et ARTP en utilisant une procédure de permutation et une méthode de Bootstrap paramétrique. Ces méthodes sont également appliquées aux données de l'étude cas-témoins CECILE sur les cancers du sein dans laquelle nous avons analysé l'interaction entre le travail de nuit et les polymorphismes des gènes circadiens dans le risque de cancer du sein. La méthode ARTP adaptée aux interactions GxE donne des résultats prometteurs. Un package R PIGE a été développé et est mis à disposition sur le CRAN.

Keywords: Gene-environment interactions, Generalized Linear Models, Pathway analysis, Resampling methods

Mots-clés : Interaction gène-environnement, Modèles linéaire généralisés, Analyse par pathway, Méthodes de ré-échantillonage

AMS 2000 subject classifications: 62F03, 62F40, 62P10

[1] Laboratoire de Mathématiques et de leur Applications, Université de Pau et Pays de l'Adour, UMR CNRS 5142, Pau, France. E-mail: camilo.broc@univ-pau.fr

[2] Department of Mathematics, Faculty of Natural Sciences, Imperial College London, UK

[3] Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, UK. E-mail: m.evangelou@imperial.ac.uk

[4] Cancer and Environment team, CESP (Center for Research in Epidemiology and Population Health), INSERM, University Paris-Saclay, University Paris-Sud, Villejuif, FRANCE. E-mail: therese.truong@inserm.fr. E-mail: pascal.guenel@inserm.fr

[5] ARC Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical, Sciences at Queensland University of Technology, Brisbane, Australia. E-mail: benoit.liquet@univ-pau.fr

 ISSN: 2102-6238

1. Introduction

During the last decade, genome-wide association studies (GWAS) have been successful in identifying several hundred single nucleotide polymorphisms (SNPs) associated with multiple cancer types (http://www.genome.gov/gwastudies/). However, such findings are not enough for explaining the genetic heritability of these cancers. Several reasons have been discussed that possibly explain the "missing heritability" in complex diseases such as the fact that most of these genetic associations were identified through single-SNP analyses (*each SNP tested independently*). It has been raised that polygenic effects, gene-gene and gene-environment (GxE) interactions are not fully explored in traditional methods (Manolio et al., 2009). Several approaches were developed in order to complete the agnostic GWAS in the discovery of additional genetic risk factors or to provide additional insights into the mechanisms involved in the studied disease.

One such approach is pathway analysis that consists of aggregating signals from SNPs (and/or genes) to pathways. Pathways are sets of genes that work together for the production of a specific biological outcome. Pathway analysis therefore incorporates the available biological knowledge of genes and SNPs for a better understanding of the genetic and biological mechanisms of the studied disease (Mooney et al., 2014, Pers (2016)). One of the main thrust of the statistical analyses will be to gain power and reduce the number of tests by combining weak signals from SNP-level analysis. Over the recent years, numerous pathway analysis for GWAS data have been proposed in the literature for finding pathways associated with the studied disease (a non exclusive list includes the methods proposed by Wang et al., 2007; Yu et al., 2009; Holmans et al., 2009; O'Dushlaine et al., 2009; Shahbaba et al., 2012; Carbonetto and Stephens, 2013; Evangelou et al., 2014a,b; Su et al., 2016). The challenges, properties and statistical methods for conducting pathway (and gene-set) analysis for GWAS data have been discussed and reviewed by Wang et al. (2011); Fridley and Biernacka (2011); de Leeuw et al. (2016).

These methods can be divided by the null hypothesis they test, namely the competitive (enrichment) or self-contained (association) null hypotheses. The self-contained null hypothesis states that no pathway genes are associated with the phenotype. On the other hand, the competitive hypothesis states that the statistics of the pathway genes are no more associated with the phenotype than the statistics of the genes outside of the pathway. A pathway where the competitive null hypothesis is rejected, is said to be an enriched one. The self-contained null hypothesis can be tested in both GWAS and candidate gene analysis, since only the statistics from a selection of genes is required. By contrast, competitive methods are usually used in GWAS data as all pathways are tested simultaneously. As discussed in the literature self-contained methods are generally more powerful than competitive methods (Evangelou et al., 2012).

Self-contained methods could also be classified into marginal approaches, which are based on the combination of p-values of individual SNPs (such as for instance Fisher's Method (FM), Adaptative Rank Truncated Product (ARTP)), or joint approaches that jointly model and test the effect of all the SNPs in the set (such as random and mixed effect models, Sequence Kernel association Test (SKAT) proposed by Ionita-Laza et al., 2013). p-value combination test statistics are usually combined with phenotype permutations for estimating their significance. Phenotype permutations avoid making any assumptions about the distribution of the effect of the genetic variants on the disease.

Another distinction factor is whether the pathway analysis is considered at the gene-level or

 ISSN: 2102-6238

at the SNP-level. Methods that perform a gene-based pathway analysis they first combine the p-values of single-SNP analysis into gene-level test statistics (or p-values) that are subsequently summarised into pathway-level associations. On the other hand, SNP level pathway methods skip the intermediate gene level and map SNPs directly to pathways.

Although a number of statistical approaches have been proposed to test for pathway association with disease, the literature has not been greatly extended for testing for GxE interactions at the pathway level. Lin et al. (2013) proposed a computationally efficient GxE set association test (GESAT), a variance component score test statistic is proposed that extends the SNP-set Kernel association test for GxE testing. The proposed method tests each set of SNPs independently from the other sets and it is a SNP-based pathway analysis approach.

In addition to this, Jiao et al. (2013) proposed the set based gene environment interaction (SBERIA) method and two more extensions that overcome the limitations of SBERIA (Jiao et al., 2015) for both rare and common variants. SBERIA firstly computes the correlation between the environmental factors and all SNPs in the set, where a z-score of correlation is obtained. These scores are translated into weights based on a preselected threshold that are included in a regression model that tests whether they are needed or not in the model. The first extension, named enhanced set-based GxE testing (eSBERIA), is composed of two steps: the first one tests the null hypothesis that the gene-environment weights are not associated with the response. The second step implements the SKAT statistic that accounts for any residual effects that might have been missed by the logistic regression model with the gene-environment interactions. As the two tests are independent, their p-values are combined using Fisher's product statistic. The third proposed approach is coSBERIA which combines SBERIA and SKAT tests for the case-only test. The case-only GxE test for a single SNP has been found to improve the power for testing for GxE under the assumption that G and E are independent (Albert et al., 2001).

In this work we were interested in extension of combination tests to the analysis of set based GxE interactions for which we have looked at replacing the phenotype permutation procedure for testing the significance of each pathway (and/or gene) by the bootstrap approach proposed by Buzkova et al. (2011). We have decided to compare the Fisher's product test statistic and ARTP approach as these two approaches have been discussed in the literature to be the most powerful pathway analyses among combination tests (Evangelou et al., 2012; Su et al., 2016). In contrast to the other proposed approaches GESAT and SBERIA, we are considering the case of gene-based pathway analyses over SNP-based ones. Further, Su et al. (2016) discussed the need for a fast algorithm to test for GxE interactions through pathway analysis and in this conducted work we aim to fill this gap.

A brief description of the context which has first motivated the development of our R package PIGE (Pathway Interaction Gene Environment, Liquet et al., 2017) is presented in Section 2. In Section 3, Fisher's Method (FM) and ARTP approaches are presented in context of gene and pathway-environment interaction. Both permutation and parametric bootstrap resampling methods are presented. In Section 4, a simulation study is presented to analyse the performances of FM and ARTP methods combined with both permutation and parametric bootstrap approaches. The methods are applied on genotype data from the CECILE case-control study in Section 5. Concluding remarks are presented in Section 6.

 ISSN: 2102-6238

2. Motivation

In a first step, we propose to test the performance of the proposed methods in simulated datasets that mimic the application dataset. In a second step, the methods will be illustrated on genotype data from the CECILE case-control study, in which we are interested in the interaction between nightwork, a binary environmental factor (defined as ever worked at night more than two years: yes/no) and polymorphisms from genes in the circadian rhythm pathway (Truong et al., 2014).

In the case of a binary response Y, the null hypothesis that there is no association between the response and the interaction term between SNP_ℓ and environment is evaluated through the following logistic model:

$$\text{logit}[P(Y=1|SNP_\ell,E)] = \alpha_\ell + \beta_\ell SNP_\ell + \beta_{E,\ell}E + \gamma_\ell E \times SNP_\ell. \quad (1)$$

where E presents the environmental factor. The likelihood ratio test (LRT) could be used to test the evidence of the interaction term ($H_{0,\ell}: \ \gamma_\ell = 0$ versus $H_{1,\ell}: \ \gamma_\ell \neq 0$), resulting to the ℓ-th p-value (p_ℓ).

As discussed earlier, there are usually multiple SNPs within each gene and multiple genes within each pathway. The questions that we will answer through our conducted work are: how to combine these results to get (i) association evidence between gene-environment interaction and the outcome, (ii) association evidence between pathway-environment interaction and the outcome?

In the context of a gene-based pathway analysis, a two-step procedure is needed. At the first level the association evidence between a gene and the response is found and at the second level these gene-level p-values are combined into a test statistic for the disease-pathway association.

Phenotype permutations are usually implemented for computing the null distribution of the test statistic that can be used for obtaining a p-value for the global null hypothesis of no association between the gene with the response. In this work, we are investigating the performance of two alternative resampling approaches one based on phenotype permutations and a second one on the bootstrap approach proposed by Buzkova et al. (2011) that has been proposed for interaction models. Both these resampling approaches are presented in Section 3.3.

3. Methods

In this section, we first present two frequentist approaches for combining p-values under investigation FM, and ARTP methods. We subsequently present the two resampling approaches. Finally, we shortly present an alternative frequentist approach iSKAT.

3.1. Fisher's method

Fisher's method is a well established association method that combines the results from multiple statistical tests. The FM test statistic equals

$$FM = -2\sum_{\ell=1}^{L}\log(p_\ell) = -2\log\left(\prod_{\ell=1}^{L}p_\ell\right), \quad (2)$$

 ISSN: 2102-6238

where L is for example the number of SNPs within a pathway. Under null hypothesis the FM test statistic follows a χ^2 distribution with $2L$ degrees of freedom when the p-values are independent. In the presence of linkage disequilibrium, the correlation between SNPs leads to dependent test-statistics. We have used the resampling approaches (presented in Section 3.3) to approximate the empirical distribution of the FM test statistics.

3.2. *Adaptive rank truncated product (ARTP)*

The idea behind the ARTP is to truncate the highest p-values in the FM method. The only p-values left are the most significant ones. To simplify the presentation of the ARTP proposed by Yu et al. (2009), we consider a pathway consisting of L SNPs and we want to test the null hypothesis that there is no pathway-environment interaction associated to the disease phenotype. Using model (1), we can perform a LRT test on individual interaction E×SNPs within the considered pathway. We denote the ordered statistics of those p-values $p_{(1)} \leq \ldots \leq p_{(L)}$, with $p_{(\ell)}$ being the ℓ-th smallest p-value. The original RTP statistic given by

$$W_K = \sum_{k=1}^{K} \log\left(p_{(k)}\right) = \log\left(\prod_{k=1}^{K} p_{(k)}\right) \tag{3}$$

combines the K smallest p-values E×SNP statistics of the tested pathway (Dudbridge and Koeleman, 2003). In the adaptive RTP method J different truncation $K_1 \leq \ldots \leq K_J$ are investigated. Let $\hat{s}(K_j)$ be the estimated p-value for W_{K_j}, $(1 \leq j \leq J)$. The ARTP statistic is then defined using the minimum p-value procedure

$$MinP = \min_{1 \leq j \leq J} \hat{s}(K_j). \tag{4}$$

Note that for a single truncation point ($J = 1$), the ARTP method is the RTP method and the RTP statistic simplifies to the FM test statistic when the truncation point K is fixed to L. Two levels of resampling approach are required to get the adjusted p-value for $MinP$: (1) for estimating $\hat{s}(K_j)$, (2) for the adjustment for multiple testing over different truncation points. To avoid this computational issue specially when the number of test L is large, Yu et al. (2009) uses the Ge et al. (2003)'s algorithm which reduces the multiple-level resampling procedure into a single level resampling procedure. In this work, we use the same algorithm.

Let $p_1^{(0)}, \ldots, p_L^{(0)}$ be the p-values for each interaction test on the null hypothesis based on the observed data. We generate B datasets under the complete null hypothesis $H_0 = H_{0,1} \cap \ldots H_{0,L}$ using appropriate resampling procedure (see section 3.3). Let $p_1^{(b)}, \ldots, p_L^{(b)}$ be the p-values for each interaction test on the null hypothesis based on the b-th generated dataset, $1 \leq b \leq B$. The RTP statistic

$$W_j^{(b)} = \sum_{i=1}^{K_j} \log\left(p_{(i)}^{(b)}\right), \quad 0 \leq b \leq B,\ 1 \leq j \leq J \tag{5}$$

is calculated for each truncation point, for both the observed data-set and each of the B simulated datasets. Then Ge's algorithm is used to estimate the p-value

$$\hat{s}_j^{(b)} = \frac{\sum_{b^*=0}^{B} I\left(W_j^{(b^*)} \leq W_j^{(b)}\right)}{B+1}, \quad 0 \leq b \leq B,\ 1 \leq j \leq J \tag{6}$$

 ISSN: 2102-6238

for each W_j. The p-value for the ARTP statistic $MinP^{(0)}$ of the pathway is estimated as

$$\widehat{\mathrm{p}}_{ARTP} = \frac{\sum_{b^*=0}^{B} I\left(MinP^{(b)} \leq MinP^{(0)}\right)}{B+1}, \tag{7}$$

where

$$MinP^{(b)} = \min_{1\leq j\leq J} \hat{s}_j^{(b)}, \quad 0 \leq b \leq B,\ 1 \leq j \leq J. \tag{8}$$

Remark. The adjusted p-value for $MinP^{(b)}$, the ARTP statistic from the b-th dataset, can also be estimated similarly using $\frac{\sum_{b^*=0}^{B} I\left(MinP^{(b^*)} \leq MinP^{(b)}\right)}{B+1}$.

Thus this procedure can give an evidence of association between a pathway-environment interaction and the disease outcome. It is called a SNP-based strategy. We describe in the following the gene-based strategy consisting to used the ARTP method for both derive the gene-environment interaction level summary and to combine gene-environment interaction level p-values across all genes within a pathway. This procedure adapted for interaction investigation is the one described in Yu et al. (2009).

Consider a pathway composed of L genes, with the ℓ-th composed of n_ℓ SNPs, $1 \leq \ell \leq L$. Let $p_{\ell,i}^{(0)}$ be the p-value for the association test on the i-th interaction SNP×environment of the ℓ-th gene based on the observed dataset. We then generate using resampling approach B datasets under the null hypothesis, and define $p_{\ell,j}^{(b)}$ the p-value for the test on the ith interaction SNP×environment of the ℓ-th gene based on the b-th generated dataset, $1 \leq b \leq B$. The ARTP is then applied (with a predetermined set of candidate truncation points, which could be varied from gene to gene) to combine interaction SNPs×environment-level evidence of association within a gene. For the ℓ-th gene, we apply the minimum p-value procedure ($MinP$) given earlier on, $1 \leq i \leq n_\ell$, $0 \leq b \leq B$, to obtain $p_\ell^{*(0)}$, the interaction gene×environment-level p-value for the observed data, and $p_\ell^{*(b)}$, the interaction GxE level p-value for the b-th permuted dataset. Finally in order to get a evidence of interaction pathway×environment the ARTP statistic is used to combine the gene×environment-level p-values for the observed and the resampling "null" data sets. We use the $MinP$ procedure one more time to obtain the adjusted p-value for the pathway×environment-level ARTP statistic. Note that the same set of generated "null" datasets are exploited each time for the $MinP$ procedure to derive interaction gene×environment-level and interaction pathway×environment-level evidence. Thus the full procedure overcomes the expensive computational multi-layer resampling issue. The same procedure is used for FM method.

3.3. Resampling methods

Both ARTP and FM rely on appropriate resampling strategy to generate data set under the null hypothesis considered. For gene- and pathways- environment interaction, we consider the global null hypothesis:

$$H_0 = H_{0,1} \cap \ldots H_{0,\ell} \ldots \cap H_{0,L}, \quad \text{with } H_{0,\ell} :\ \gamma_\ell = 0 \text{ (see equation (1))} \tag{9}$$

where L is the number of SNPs within a considered pathway.

 ISSN: 2102-6238

Yu et al. (2009) use a permutation procedure to evaluate the significance level of the ARTP statistic in the context of disease-pathway association which for example corresponds to the situation of the simplified model:

$$\text{logit}[P(Y=1|SNP_\ell,E)] = \alpha_\ell + \beta_\ell SNP_\ell \tag{10}$$

with $H_{0,\ell}: \ \beta_\ell = 0$. In this situation, there is no difficulty to define the permutation procedure for the complete null hypothesis. One just need to permute the phenotype Y. However, for the interaction model (1), a valid permutation procedure to generate data set under the complete null hypothesis (9) is complex to define. As noted by Buzkova et al. (2011), fixing SNP and E and permuting Y generates data in which the generated phenotype Y^* is independent of SNP and E. This procedure fails to generate data set under the null hypothesis since in model (1) the phenotype Y is not independent of SNP and E. Indeed, for the logistic model there is no permutation procedure which can be used to generate data set for the complete null hypothesis (Edgington, 1987). An alternative to the permutation procedure is to used a parametric bootstrap procedure (Efron and Tibshirani, 1994; Liquet and Riou, 2013) which implies less stringent assumptions (Good, 2000). In our context, the procedure could be defined in the following.

For each SNP ($\ell = 1,\ldots,L$):

1. Fit the model under the null hypothesis $H_{0,\ell}$, using the observed data, and obtain $\hat{\alpha}_\ell$, $\hat{\beta}_\ell$, $\hat{\beta}_{E,\ell}$, the maximum likelihood estimate (MLE) of respectively α_ℓ, β_ℓ and $\beta_{E,\ell}$
2. Generate a new outcome $Y^*_{i,\ell}$ for each subject from the probability measure defined under $H_{0,\ell}$. For example, for model (1), we generate $Y^*_{i,\ell}$ according to:

$$P(Y^*_{i,\ell}=1|SNP_\ell,E) = \frac{\exp\left(\hat{\alpha}_\ell + \hat{\beta}_\ell SNP_\ell + \hat{\beta}_{E,\ell}E\right)}{1+\exp\left(\hat{\alpha}_\ell + \hat{\beta}_\ell SNP_\ell + \hat{\beta}_{E,\ell}E\right)}.$$

   Repeat this for all the subjects to obtain a sample noted $s^*_\ell = \{Y^*_{i,\ell}, SNP_{i,\ell}, E_i\}$ which is related to the ℓ-th SNP.
3. Generate B new datasets $s^*_{b,\ell}$, $b = 1,\ldots,B$ by repeating B times the steps 1, 2 and 3.

Remark: In case of marginal association of both SNP and environmental factor, step 2 might generate unbalanced data which could affect the statistical power of the resampling methods. A screening investigation on the marginal association might be used before using the booststrap method.

3.4. iSKAT

The other frequentist approach is iSKAT proposed by Lin et al. (2016). The method uses the spirit of SKAT-O methods (Wu et al. (2011)) and apply it to an interaction test context. From one side burden tests are know to be an efficient test in many cases but they struggle when rare variants are involved in the data. From the other side kernel test can handle those rare variants. The idea behind the algorithm is to separate from the data the rare variants from the rest and to take advantage of both burden tests and kernel tests. Furthermore, iSKAT offer the possibility of weighting the covariates to take into account extra information. However, no weight have been added in the use of iSKAT in this article. The method GESAT is a particular case of the iSKAT method.

 ISSN: 2102-6238

4. Simulation Study

In this section, we compare the FM and ARTP methods through a simulation study investigating their control of Type-I error and FWER and their power performance. Two resampling approaches (permutation and bootstrap) are compared for a range of sample sizes ($n = 200,\ 500, 1000$). The combination methods FM and ARTP are compared to iSKAT and to the popular frequentist approach MinP which combines p-value by considering only the most signifactive p-value:

$$\text{MinP} = \min_{\ell \in \{1,\dots,L\}} p_\ell.$$

Let's note that MinP method doesn't have the same meaning than the quantity $MinP$ used in the intermediary steps of ARTP (see equation (4)).

4.1. Data Generated

We work on generated data which are supposed to mimic experimental data. The parameters of the generation are inspired from Buzkova et al. (2011). The genetic structure simulated is composed by one pathway containing I genes (genes are called $G_1, G_2, \dots, G_I$). Each gene contains several SNPs. The SNPs are binary variables. In order to generate the i-th gene, composed by k_i SNPs, $SNP_1^i \dots SNP_{k_i}^i$, we use the following procedure:

$$\begin{aligned} S_i &\sim Bern(0.2) \\ \text{logit}(p_j) &= \text{logit}(0.2) + S_i \text{ for } j \in \{1,\dots,k_i\} \\ SNP_j^i | S_i &\sim Bern(p_j) \text{ for } j \in \{1,\dots,k_i\}. \end{aligned}$$

Hence, conditionally on the latent polymorphism S_i, for a given gene i, the individual SNP_j^i are independent and identically distributed, but they are marginally dependent.

A binary environment variable is also simulated that is marginally dependent with one gene i_E and generated with the following procedure:

$$\begin{aligned} \text{logit}(p_E) &= a + bS_{i_E} \\ P[E = 1] &= p_E \end{aligned}$$

Finally, a binary outcome variable is simulated. It is generated from a logistic model using SNPs from gene i_Y. Among those s' SNPs only s SNPs are associated to the response variable Y as specified in the following equations:

$$\lambda_1, \dots, \lambda_{s'} \in \{0,1\} \text{ and } \sum_{l=1}^{s'} \lambda_l = s \tag{11}$$

$$\text{logit}[P(Y = 1 | G_{i_Y}, E)] = \alpha + \beta_E E + \sum_{j=1}^{k_{i_Y}} \lambda_j \beta_{SNP_j^{i_Y}} SNP_j^{i_Y} + \sum_{j=1}^{k_{i_Y}} \lambda_j \gamma_{SNP_j^{i_Y}} SNP_j^{i_Y} \times E \tag{12}$$

 ISSN: 2102-6238

The parameters λ_l control the choice of the SNP involved in the generation of Y and the parameter s controls the number of those SNPs. Different choices for parameters β_j and γ_j are chosen in order to highlight different results. For Type I error results the parameters γ_j are set to 0 whereas for power results parameters β_j and γ_j are chosen in order to evidence the different rejection of the null hypothesis for different parameters and methods used.

4.2. Simulation design

We present nine different simulation models. The first four ones are used to investigate the Type-I error and the Family Wise Error Rate (FWER) results while the others are used to explore the power of the different methods.

As discussed above, different resampling methods are used: (i) permutation that permutes the outcome Y and (ii) parametric bootstrap. We set to 1,000 the number of permutations and bootstrap resampling. The sample size n of the simulation datasets are 200, 500 and 1000. The results we look for are the p-values at gene level that are based on SNP-level information. Then a pathway-level p-values is computed from this information.

We use cases 1, 2, 3 and 4 (defined in the following) to investigate the control of the Type-I error rate for each gene. We also investigate the control of the FWER at the gene level and at the pathway level. An empirical FWER for gene-environment interaction is defined by the number of times the procedure detects wrongly at least one significant gene-environment interaction (from all the genes within the Pathway) over the $N = 500$ simulation replications. We also add the empirical FWER computed using a Bonferroni correction (i.e., each gene-environment interaction p-value is divided by the number of investigated genes). Indeed, the output of each method is a set of p-values (one for each investigating gene-environment interaction). This set of p-values is associated to a set of null hypotheses which define our family of hypotheses. Then it is important to control an overall error for these hypotheses. The empirical FWER for pathway-environment interaction is defined by the number of times the procedure detects wrongly a significant pathway-environment interaction over the 500 replications. As only one pathway is considered in our simulation study, the control of the FWER at the pathway level is similar to the Type-I error rate control of the pathway investigated.

Simulation cases for investigating Type-I error rate and FWER

— **Case 1**: Data is composed of 5 genes with 10 SNPs each. True model is based on the main effect of E and the main effect of 5 randomly selected SNP from the first gene. The environment is marginally correlated with the first gene but not with the other genes. The outcome is therefore generated from the following model:

$$\lambda_1, \ldots, \lambda_{10} \in \{0,1\} \text{ and } \sum_{j=1}^{10} \lambda_j = 5$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{10} \lambda_j \beta_{SNP_j^1} SNP_j^1$$

 ISSN: 2102-6238

— **Case 2**: Data is composed of 5 genes with 10 SNPs The true model is based on the main effect of E and the main effect of all SNPs from the second gene. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\text{logit}[P(Y=1|G_2,E)] = \alpha + \beta_E E + \sum_{j=1}^{10} \beta_{SNP_j^2} SNP_j^2$$

— **Case 3**: Data is composed of 5 genes, 4 of them with 5 SNPs and one with 50 SNPs (the last one).True model is based on the main effect of E and 2 randomly selected SNPs from the first gene. The environment is marginally correlated with the first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1,\ldots,\lambda_5 \in \{0,1\} \text{ and } \sum_{j=1}^{5} \lambda_j = 2$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{5} \lambda_j \beta_{SNP_j^1} SNP_j^1$$

This model enables us to see how the methods perform when the pathway gene members have different sizes.

— **Case 4**: Data is composed of 20 genes with 20 SNPS for the first gene and 10 SNPs for the others. True model is based on: the main effect of E; the main effect of 10 randomly selected SNPs from the first gene and 5 from the second genes. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1^1,\ldots,\lambda_{20}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{20} \lambda_j^1 = 10 \text{ and } \lambda_1^2,\ldots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10} \lambda_j^2 = 5$$

$$\text{logit}[P(Y=1|G_1\,,G_2,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j^1 \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{10} \lambda_j^2 \beta_{SNP_j^2} SNP_j^2$$

Simulation cases for power performance

— **Case 5**: Data is composed of 5 genes with 20, 10, 10, 10, 10 SNPs. True model is based on the main effect of E and the main effect of 10 randomly selected SNPs from the first gene and the interaction between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1,\ldots,\lambda_{20} \in \{0,1\} \text{ and } \sum_{j=1}^{20} \lambda_j = 10$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{20} \lambda_j \gamma_{SNP_j^1} SNP_j^1 \times E$$

 ISSN: 2102-6238

— **Case 6**: Data is composed of 5 genes with 20, 10, 10, 10, 10 SNPs. True model is based on the main effect of E and the main effect of 2 randomly selected SNPs from the first gene and the interaction between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1,\dots,\lambda_{20} \in \{0,1\} \text{ and } \sum_{j=1}^{20} \lambda_j = 2$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{20} \lambda_j \gamma_{SNP_j^1} SNP_j^1 \times E$$

— **Case 7**: Data is composed of 20 genes with 20 SNPS for the first gene and 10 SNPs for the others. True model is based on: the main effect of E; the main effect of 10 randomly selected SNPs from the first gene and 5 from the second genes; the interactions between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1^1,\dots,\lambda_{20}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{20} \lambda_j^1 = 10 \text{ and } \lambda_1^2,\dots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10} \lambda_j^2 = 5$$

$$\begin{aligned}\text{logit}[P(Y=1|G_1,G_2,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j^1 \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{10} \lambda_j^2 \beta_{SNP_j^2} SNP_j^2 \\ + \sum_{j=1}^{20} \lambda_j^1 \gamma_{SNP_j^1} SNP_j^1 \times E + \sum_{j=1}^{10} \lambda_j^2 \gamma_{SNP_j^2} SNP_j^2 \times E\end{aligned}$$

— **Case 8**: Data is composed of 20 genes with 20 SNPS for the first gene and 10 SNPs for the others. True model is based on: the main effect of E; the main effect of 2 randomly selected SNPs from the first gene and 2 from the second genes; the interactions between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1^1,\dots,\lambda_{20}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{20} \lambda_j^1 = 2 \text{ and } \lambda_1^2,\dots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10} \lambda_j^2 = 2$$

$$\begin{aligned}\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j^1 \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{10} \lambda_j^2 \beta_{SNP_j^2} SNP_j^2 \\ + \sum_{j=1}^{20} \lambda_j^1 \gamma_{SNP_j^1} SNP_j^1 \times E + \sum_{j=1}^{10} \lambda_j^2 \gamma_{SNP_j^2} SNP_j^2 \times E\end{aligned}$$

— **Case 9**: Data is composed of 2 pathways with 10 genes in each pathways. Each genes includes 10 SNPS. True model is based on: the main effect of E; the main effect of 2 randomly selected SNPs from the first gene and 2 from the second genes of each pathways;

 ISSN: 2102-6238

the interactions between the environment with each of the selected SNPs. The outcome is therefore generated from the following model:

$$\lambda_1^1,\dots,\lambda_{10}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^1 = 2 \text{ and } \lambda_1^2,\dots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^2 = 2$$

$$\lambda_1^{11},\dots,\lambda_{10}^{11} \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^{11} = 2 \text{ and } \lambda_1^{12},\dots,\lambda_{10}^{12} \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^{12} = 2$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{k\in K}\sum_{j=1}^{10}\lambda_j^k \beta_{SNP_j^k} SNP_j^k + \sum_{k\in K}\sum_{j=1}^{10}\lambda_j^k \gamma_{SNP_j^k} SNP_j^k \times E$$

where $K = \{1,2,11,12\}$.

Simulation parameters

The different coefficient used in our cases are gathered in table 1. The notation used refer to part 4.1. For each simulation case, FM, and ARTP methods are applied for the 9 simulation cases to investigate the presence of interaction effects of gene- and pathway- environment based on a gene-based strategy (see end of Section 3.2).

For all of the 9 cases the truncation points of the ARTP parameters are optimized like in previous ARTP results Yu et al. (2009). Let m be an integer; k_G and k_{SNP} be real numbers; k_i be the number of SNP in the i-th gene; I the number of genes. Let $\lfloor k_G \times I \rfloor, \lfloor 2 \times k_G \times I \rfloor, \dots, \lfloor m \times k_G \times I \rfloor$ be a set of truncation points for genes and, for each gene i, let $\lfloor k_{SNP} \times k_i \rfloor, \lfloor 2 \times k_{SNP} \times k_i \rfloor, \dots, \lfloor m \times k_{SNP} \times k_i \rfloor$ be a set of truncation points for the SNPs of this gene. The notation $\lfloor x \rfloor$ gives the largest integer that does not exceed x (if $\lfloor x \rfloor = 0$ we set the value to 1) . We define $p_{k_{SNP},k_G,m}$ the p-value of the ARTP computed with this set of truncation points. The optimal p-value of the ARTP is defined as:

$$\min_{k_{SNP}\in\mathscr{A},k_G\in\mathscr{A}} p_{k_{SNP},k_G,m} \quad \text{with} \quad \mathscr{A} = \{2\%,4\%,\dots,20\%\}.$$

The parameter m is fixed to 5 in our study.

4.3. Type I error rate and FWER

Type I error rate and FWER of the methods are computed in cases 1, 2, 3 and 4. The data are generated under the null hypothesis (i.e. no interaction). The SNP-level tests are performed under the interaction assumption with a significance level of 0.05. Hence, the expected value of all p-values at gene-level and pathway level are 0.05. In the section, we study the behavior of the different methods for this case. Tables 2, 3, 4 , 5 , 6 and 7 present the results. The expected value of the average p-values is 0.05. Computing a binomial model, the average of the p-values on the 500 iterations should be between 3% and 7%.

 ISSN: 2102-6238

TABLE 1. *Generating parameters for cases 1 to 9. The notation used refer to part 4.1.*

	p_E	a	b	α	β_E	$\beta_{SNP_j^{i_Y}}$ $i_Y \in \{1,\dots,k_{i_Y}\}$	$\gamma_{SNP_j^{i_Y}}$ $i_Y \in \{1,\dots,k_{i_Y}\}$
case 1	0.2	logit(2)	log(2)	-2	2	$\in \{3,2,1\}$	$=0$
case 2	0.2	logit(2)	log(2)	-2	2	$\in \{1.5,1.0,0.5\}$	$=0$
case 3	0.2	logit(2)	log(2)	-2	2	$\in \{3,2,1\}$	$=0$
case 4	0.5	logit(2)	2	-1	1	$\in \{0.3,0.2,0.1\}$	$=0$
case 5	0.2	logit(2)	log(2)	-2	0.4	$\in \{0.06,0.04,0.02\}$	$=0.5$
case 6	0.2	logit(2)	log(2)	-2	2	$\in \{0.3,0.2,0.1\}$	$=1.5$
case 7	0.5	logit(2)	2	-1	0.1	$\in \{0.03,0.02,0.01\}$	$=0.3$
case 8	0.5	logit(2)	2	-1	0.1	$\in \{0.075,0.050,0.025\}$	$=0.7$
case 9	0.5	logit(2)	2	-1	0	$=0$	$=0.5$

The permutation approach obtained very low error rate for both approaches (FM, MinP, ARTP). The bootstrap approach gives good results for controlling the Type-I error rate for both FM and ARTP methods. As expected the FWER at the gene level is not controlled. The FWER computed can then be corrected using the Bonferroni method which is known to be conservative and more trustable. Finally, a pathway p-value is given by each combining method (FM, MinP, ARTP). When the number of genes is low (cases 1 to 3), the Type-I error rate at the pathway level is well controlled using the bootstrap approach for both FM, MinP and ARTP methods while permutation approach give conservative results. When the number of genes is higher (case 4), the type-I error rate at the pathway level of the ARTP and iSKAT is slightly inflated.

4.4. Power performance

Tables 8, 9, 10, 11 and 12 present the results for the power of the methods. As expected for all methods power performances increase with larger sample size.

In those results the proportion of significant SNP in the true model have a huge importance on the performances. In cases 5 and 7 the proportions of significant SNPs are high whereas in cases 6, 8 and 9 they are low. For higher proportions bootstrap is slightly but consistently more powerful than the permutation. For lower proportions permutation and bootstrap results are equivalent. We can see that when the proportion is high FM and ARTP have equivalent results and MinP has abysmal results. This is due to the fact that MinP truncates too much of the information contained in the data. When the proportion are lower, ARTP and then MinP have a good performance but FM have lower ones. This is due to the fact that FM results take into account too much irrelevant SNPs in its combination. The ARTP have the merit of having good power whatever is the proportion of significant SNP in the true model. FM and MinP seems to detect different patterns but ARTP can detect both. When we compare the combining methods with iSKAT we can see that the level of performance of ARTP and iSKAT is similar. We notice that in general, ARTP is more powerful on small sample sizes (200 and 500).

 ISSN: 2102-6238

TABLE 2. *Simulation of case 1 with 1000 permutations and 1000 bootstrap resampling with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0	0.03	0.036	0.04	0.024	0.038	0.024
Gene2	10	0.016	0.08	0.04	0.04	0.038	0.076	0.052
Gene3	10	0.022	0.074	0.048	0.064	0.042	0.078	0.054
Gene4	10	0.008	0.056	0.04	0.052	0.028	0.05	0.05
Gene5	10	0.02	0.07	0.04	0.05	0.042	0.066	0.038
$FWER_{BF}$		0.014	0.066	0.046	0.052	0.042	0.068	0.038
$FWER$		0.066	0.282	0.188	0.23	0.166	0.29	0.2
Type-I Error: Pathway		0.014	0.066	0.048	0.052	0.028	0.078	0.052

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.002	0.016	0.008	0.01	0.01	0.02	0.03
Gene2	10	0.028	0.062	0.052	0.062	0.044	0.064	0.044
Gene3	10	0.028	0.046	0.028	0.044	0.028	0.05	0.038
Gene4	10	0.018	0.034	0.03	0.044	0.028	0.038	0.03
Gene5	10	0.028	0.054	0.036	0.042	0.042	0.054	0.066
$FWER_{BF}$		0.028	0.052	0.038	0.042	0.042	0.064	0.042
$FWER$		0.098	0.188	0.148	0.188	0.142	0.208	0.186
Type-I Error: Pathway		0.03	0.052	0.038	0.042	0.042	0.078	0.04

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.008	0.032	0.026	0.028	0.018	0.028	0.032
Gene2	10	0.032	0.04	0.03	0.048	0.036	0.052	0.042
Gene3	10	0.022	0.032	0.036	0.044	0.04	0.05	0.038
Gene4	10	0.044	0.054	0.046	0.064	0.06	0.076	0.054
Gene5	10	0.036	0.036	0.032	0.046	0.036	0.042	0.042
$FWER_{BF}$		0.026	0.04	0.036	0.04	0.04	0.048	0.03
$FWER$		0.138	0.178	0.16	0.21	0.178	0.228	0.196
Type-I Error: Pathway		0.026	0.04	0.036	0.04	0.042	0.048	0.044

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 3. *Simulation of case 2 with 1000 permutations and 1000 bootstrap resampling with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.018	0.09	0.032	0.046	0.03	0.094	0.046
Gene2	10	0.002	0.028	0.054	0.044	0.036	0.032	0.04
Gene3	10	0.008	0.062	0.03	0.046	0.02	0.054	0.038
Gene4	10	0.012	0.084	0.042	0.056	0.04	0.104	0.062
Gene5	10	0.016	0.076	0.03	0.038	0.032	0.07	0.038
$FWER_{BF}$		0.008	0.084	0.03	0.034	0.04	0.076	0.038
$FWER$		0.056	0.304	0.172	0.212	0.152	0.31	0.204
Type-I Error: Pathway		0.008	0.086	0.03	0.036	0.03	0.09	0.028

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.02	0.046	0.028	0.05	0.022	0.05	0.032
Gene2	10	0.002	0.032	0.036	0.038	0.03	0.03	0.058
Gene3	10	0.028	0.07	0.038	0.046	0.034	0.066	0.054
Gene4	10	0.028	0.062	0.042	0.054	0.044	0.076	0.044
Gene5	10	0.018	0.05	0.036	0.052	0.026	0.046	0.026
$FWER_{BF}$		0.016	0.046	0.042	0.042	0.042	0.06	0.036
$FWER$		0.09	0.232	0.164	0.214	0.146	0.238	0.196
Type-I Error: Pathway		0.016	0.046	0.048	0.044	0.032	0.066	0.036

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.056	0.07	0.044	0.064	0.056	0.078	0.046
Gene2	10	0.014	0.03	0.022	0.03	0.024	0.026	0.04
Gene3	10	0.038	0.054	0.03	0.042	0.042	0.06	0.038
Gene4	10	0.032	0.042	0.028	0.036	0.036	0.046	0.062
Gene5	10	0.042	0.048	0.028	0.038	0.038	0.046	0.038
$FWER_{BF}$		0.032	0.048	0.032	0.038	0.034	0.07	0.038
$FWER$		0.164	0.216	0.144	0.194	0.178	0.226	0.204
Type-I Error: Pathway		0.032	0.048	0.032	0.04	0.042	0.076	0.044

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 4. *Simulation of case 3 with 1000 permutations and 1000 bootstrap resampling with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	5	0.02	0.064	0.042	0.05	0.034	0.054	0.044
Gene2	5	0.026	0.07	0.048	0.068	0.038	0.07	0.07
Gene3	5	0.03	0.068	0.04	0.06	0.044	0.074	0.054
Gene4	5	0.018	0.04	0.026	0.036	0.02	0.034	0.04
Gene5	50	0.01	0.144	0.066	0.066	0.062	0.16	0.192
$FWER_{BF}$		0.02	0.144	0.044	0.046	0.052	0.11	0.072
$FWER$		0.096	0.32	0.19	0.242	0.176	0.326	0.348
Type-I Error: Pathway		0.022	0.146	0.046	0.048	0.058	0.132	0.008

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	5	0.026	0.048	0.036	0.05	0.032	0.048	0.058
Gene2	5	0.028	0.042	0.038	0.052	0.036	0.056	0.042
Gene3	5	0.038	0.056	0.048	0.056	0.044	0.068	0.06
Gene4	5	0.042	0.062	0.032	0.062	0.04	0.06	0.052
Gene5	50	0.032	0.098	0.032	0.038	0.06	0.11	0.088
$FWER_{BF}$		0.038	0.08	0.042	0.05	0.044	0.092	0.088
$FWER$		0.156	0.278	0.172	0.232	0.198	0.304	0.278
Type-I Error: Pathway		0.038	0.084	0.044	0.052	0.034	0.096	0.02

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	5	0.016	0.03	0.034	0.042	0.022	0.028	0.04
Gene2	5	0.046	0.05	0.04	0.054	0.046	0.054	0.046
Gene3	5	0.04	0.05	0.042	0.056	0.046	0.054	0.052
Gene4	5	0.03	0.042	0.038	0.046	0.032	0.04	0.04
Gene5	50	0.042	0.05	0.036	0.05	0.06	0.098	0.05
$FWER_{BF}$		0.024	0.044	0.044	0.054	0.044	0.056	0.038
$FWER$		0.16	0.208	0.176	0.222	0.188	0.248	0.21
Type-I Error: Pathway		0.024	0.044	0.046	0.056	0.04	0.06	0.012

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 5. *Simulation of case 4 for sample size of 200 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200								
Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.046	0.062	0.036	0.05	0.058	0.076	0.028
Gene2	10	0.05	0.062	0.044	0.056	0.05	0.064	0.034
Gene3	10	0.038	0.042	0.034	0.036	0.046	0.058	0.024
Gene4	10	0.034	0.044	0.046	0.052	0.04	0.046	0.032
Gene5	10	0.048	0.066	0.038	0.042	0.048	0.062	0.042
Gene6	10	0.058	0.07	0.062	0.064	0.06	0.068	0.056
Gene7	10	0.044	0.042	0.026	0.034	0.036	0.038	0.036
Gene8	10	0.054	0.068	0.044	0.04	0.044	0.06	0.04
Gene9	10	0.046	0.064	0.052	0.062	0.06	0.066	0.044
Gene10	10	0.032	0.036	0.026	0.042	0.026	0.04	0.032
Gene11	10	0.042	0.056	0.038	0.044	0.046	0.058	0.034
Gene12	10	0.054	0.072	0.054	0.062	0.06	0.072	0.046
Gene13	10	0.038	0.052	0.036	0.052	0.048	0.058	0.026
Gene14	10	0.062	0.078	0.044	0.048	0.066	0.076	0.056
Gene15	10	0.032	0.044	0.05	0.06	0.04	0.048	0.038
Gene16	10	0.038	0.044	0.042	0.05	0.038	0.048	0.034
Gene17	10	0.062	0.076	0.028	0.038	0.048	0.07	0.056
Gene18	10	0.058	0.07	0.032	0.042	0.044	0.054	0.05
Gene19	10	0.038	0.042	0.032	0.048	0.042	0.046	0.028
Gene20	10	0.06	0.076	0.04	0.05	0.072	0.086	0.048
$FWER_{BF}$		0.02	0.056	0.024	0.026	0.054	0.08	0.018
$FWER$		0.612	0.704	0.564	0.642	0.646	0.732	0.556
Type-I Error: Pathway		0.028	0.07	0.028	0.04	0.084	0.152	0.004

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 6. *Simulation of case 4 for sample size of 500 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 500								
Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.054	0.060	0.046	0.056	0.068	0.073	0.038
Gene2	10	0.048	0.053	0.048	0.053	0.064	0.070	0.056
Gene3	10	0.044	0.043	0.046	0.049	0.052	0.051	0.032
Gene4	10	0.056	0.062	0.042	0.043	0.052	0.060	0.052
Gene5	10	0.048	0.043	0.052	0.049	0.052	0.053	0.038
Gene6	10	0.046	0.043	0.046	0.049	0.04	0.047	0.028
Gene7	10	0.054	0.058	0.042	0.053	0.05	0.049	0.042
Gene8	10	0.038	0.030	0.044	0.041	0.052	0.041	0.02
Gene9	10	0.066	0.073	0.052	0.053	0.064	0.071	0.054
Gene10	10	0.038	0.039	0.052	0.053	0.052	0.058	0.036
Gene11	10	0.036	0.045	0.038	0.043	0.054	0.058	0.04
Gene12	10	0.054	0.062	0.054	0.064	0.064	0.066	0.056
Gene13	10	0.05	0.058	0.04	0.047	0.06	0.075	0.042
Gene14	10	0.038	0.036	0.05	0.049	0.04	0.051	0.026
Gene15	10	0.048	0.053	0.06	0.062	0.056	0.058	0.044
Gene16	10	0.06	0.053	0.058	0.062	0.06	0.066	0.048
Gene17	10	0.056	0.060	0.074	0.068	0.058	0.056	0.044
Gene18	10	0.052	0.056	0.042	0.045	0.056	0.068	0.048
Gene19	10	0.046	0.043	0.07	0.071	0.07	0.079	0.036
Gene20	10	0.032	0.032	0.03	0.041	0.04	0.047	0.028
$FWER_{BF}$		0.046	0.056	0.042	0.058	0.078	0.111	0.032
$FWER$		0.626	0.637	0.628	0.658	0.676	0.712	0.544
Type-I Error: Pathway		0.056	0.068	0.056	0.071	0.122	0.156	0.02

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 7. *Simulation of case 4 for sample size of 1000 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 1000								
Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.038	0.038	0.05	0.052	0.052	0.054	0.04
Gene2	10	0.066	0.074	0.058	0.054	0.076	0.07	0.054
Gene3	10	0.052	0.054	0.042	0.044	0.046	0.056	0.04
Gene4	10	0.044	0.054	0.038	0.038	0.042	0.044	0.03
Gene5	10	0.058	0.06	0.038	0.038	0.044	0.052	0.046
Gene6	10	0.052	0.058	0.066	0.068	0.06	0.06	0.036
Gene7	10	0.058	0.06	0.058	0.054	0.064	0.066	0.046
Gene8	10	0.048	0.052	0.044	0.044	0.046	0.046	0.034
Gene9	10	0.036	0.036	0.038	0.038	0.032	0.036	0.028
Gene10	10	0.062	0.06	0.064	0.064	0.07	0.072	0.052
Gene11	10	0.052	0.054	0.042	0.042	0.046	0.056	0.046
Gene12	10	0.044	0.048	0.032	0.036	0.038	0.03	0.034
Gene13	10	0.046	0.044	0.062	0.07	0.052	0.058	0.036
Gene14	10	0.05	0.048	0.054	0.046	0.05	0.058	0.036
Gene15	10	0.06	0.054	0.054	0.056	0.062	0.06	0.052
Gene16	10	0.034	0.036	0.052	0.048	0.054	0.05	0.028
Gene17	10	0.034	0.032	0.048	0.05	0.042	0.044	0.028
Gene18	10	0.056	0.06	0.038	0.032	0.052	0.048	0.048
Gene19	10	0.048	0.048	0.052	0.046	0.044	0.046	0.038
Gene20	10	0.044	0.048	0.06	0.066	0.056	0.06	0.036
$FWER_{BF}$		0.044	0.056	0.044	0.05	0.076	0.092	0.046
$FWER$		0.614	0.61	0.63	0.622	0.634	0.67	0.536
Type-I Error: Pathway		0.058	0.058	0.054	0.066	0.12	0.13	0.104

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 8. *Simulation of case 5 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.1	0.122	0.054	0.052	0.1	0.122	0.21
Power: Pathway		0.062	0.096	0.044	0.042	0.084	0.098	0.012

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.512	0.536	0.162	0.17	0.496	0.526	0.532
Power: Pathway		0.308	0.326	0.078	0.078	0.328	0.35	0.356

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene1	20	0.876	0.884	0.432	0.42	0.86	0.872	0.872
Power: Pathway		0.722	0.744	0.184	0.196	0.722	0.726	0.836

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 9. *Simulation of case 6 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.076	0.09	0.046	0.052	0.094	0.086	0.186
Power: Pathway		0.056	0.068	0.054	0.04	0.1	0.098	0.048

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.264	0.254	0.222	0.19	0.364	0.348	0.382
Power: Pathway		0.162	0.152	0.11	0.094	0.226	0.192	0.212

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.62	0.616	0.73	0.7	0.798	0.782	0.752
Power: Pathway		0.406	0.426	0.448	0.408	0.616	0.598	0.684

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 10. *Simulation of case 7 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.108	0.118	0.060	0.080	0.108	0.134	0.096
Gene2	10	0.072	0.084	0.054	0.072	0.062	0.088	0.072
Power: Pathway		0.060	0.062	0.040	0.034	0.086	0.128	0.004

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.272	0.278	0.144	0.148	0.270	0.272	0.268
Gene2	10	0.158	0.170	0.100	0.100	0.138	0.144	0.132
Power: Pathway		0.112	0.132	0.074	0.070	0.214	0.240	0.168

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.572	0.580	0.262	0.268	0.562	0.578	0.592
Gene2	10	0.318	0.324	0.180	0.186	0.290	0.288	0.3
Power: Pathway		0.298	0.334	0.080	0.090	0.416	0.444	0.464

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 11. *Simulation of case 8 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.102	0.114	0.108	0.110	0.134	0.153	0.124
Gene2	10	0.092	0.094	0.104	0.102	0.128	0.125	0.120
Power: Pathway		0.044	0.060	0.050	0.048	0.142	0.157	0.016

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.224	0.230	0.160	0.160	0.256	0.250	0.302
Gene2	10	0.310	0.318	0.274	0.272	0.340	0.332	0.398
Power: Pathway		0.108	0.116	0.096	0.100	0.262	0.280	0.168

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.376	0.372	0.436	0.438	0.476	0.467	0.550
Gene2	10	0.538	0.548	0.544	0.544	0.602	0.618	0.700
Power: Pathway		0.266	0.284	0.292	0.288	0.516	0.526	0.568

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 12. *Simulation of case 9 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.070	0.080	0.040	0.045	0.070	0.065	0.090
Gene2	10	0.095	0.105	0.070	0.075	0.095	0.095	0.080
Power: Pathway1		0.055	0.065	0.04	0.030	0.110	0.130	0.012
Gene3	10	0.110	0.125	0.095	0.090	0.100	0.120	0.120
Gene4	10	0.085	0.100	0.085	0.100	0.110	0.115	0.118
Power: Pathway2		0.065	0.095	0.04	0.050	0.105	0.105	0.016
Power: Pathway all		0.055	0.070	0.05	0.025	0.115	0.155	0.012

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.150	0.145	0.130	0.135	0.150	0.170	0.194
Gene2	10	0.155	0.160	0.125	0.125	0.160	0.160	0.190
Power: Pathway1		0.085	0.10	0.075	0.075	0.145	0.150	0.132
Gene3	10	0.145	0.150	0.100	0.105	0.150	0.145	0.204
Gene4	10	0.135	0.145	0.125	0.110	0.140	0.145	0.166
Power: Pathway2		0.100	0.09	0.080	0.085	0.155	0.165	0.084
Pathway all		0.065	0.10	0.075	0.070	0.260	0.230	0.148

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.245	0.245	0.260	0.260	0.295	0.290	0.366
Gene2	10	0.235	0.260	0.225	0.220	0.275	0.270	0.364
Power: Pathway1		0.145	0.150	0.135	0.12	0.240	0.255	0.324
Gene3	10	0.230	0.235	0.240	0.240	0.270	0.285	0.364
Gene4	10	0.295	0.305	0.270	0.265	0.345	0.335	0.392
Power: Pathway2		0.160	0.195	0.150	0.14	0.250	0.265	0.304
Pathway all		0.160	0.220	0.170	0.18	0.425	0.465	0.484

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

5. Application: Breast cancer and night work

Circadian rhythm is a roughly 24 hours cycle of biological processes that are synchronized by external cues such as light or temperature, and regulated endogenously by periodic transcription of a set of genes that form a network of self-regulated feedback loop. The circadian rhythm pathway plays a key role in the maintenance of various endocrine, physiological factors and behavioral functions including cell cycle regulation, hormone secretion, body temperature and sleep/wake cycle. Shift work that involves circadian disruption was classified as probably carcinogenic to humans (group 2A) by the International Agency for Research on Cancer in 2007 (Straif et al., 2007). An increased risk of breast cancer was reported in women working at night by several studies (Hansen and Lassen, 2012; Menegaux et al., 2013) and it was hypothesized that this association could be modulated by polymorphisms in the circadian pathway genes. As the circadian pacemaker requires multiple molecular interactions to generate the circadian rhythms, single-SNP analyses may not be sufficient to analyze the association between circadian genes and breast cancer. Therefore, we have investigated the role of circadian clock gene polymorphisms and their interaction with nightwork in breast cancer risk using a pathway analysis. This work was previously described in more details using only the ARTP method with a modified permutation procedure that permutes the outcome, the environmental factor and the adjustment variables together (Truong et al., 2014). Here, we present the results using FM, MinP and ARTP methods using permutation and Bootstrap resampling procedures as well as iSKAT method for comparison.

Briefly, the analyses are conducted in a population-based case-control study from France including 1126 breast cancer cases and 1174 controls.

We considered the circadian pathway as defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database that included 23 genes (*CLOCK*, *ARNTL*, *NPAS2*, *CRY1*, *CRY2*, *PER1*, *PER2*, *PER3*, *RORA*, *RORB*, *RORC*, *BHLHE40*, *BHLHE41*, *SKP1*, *FBXW11*, *CUL1*, *TIMELESS*, *FBXL3*, *NR1D1*, *CSNK1D*, *CSNK1E*, *RBX1*, and *BTRC*). These genes constitute a complex regulatory network with multiple negative and positive feedback loops. A selection of tag SNPs from these genes were selected in order to capture SNPs within 5 kb of each genes (pairwise approach with a square of correlation coefficient $r^2 > 0.8$) with a minimum minor allele frequency of 0.05 in the CEU population from HapMap project. After quality controls, we have included 577 SNPs from the 23 genes. The circadian pathway was additionnally divided into two subpathways: the core circadian genes which are involved in the same transcriptional feedback loop (*CLOCK*, *ARNTL*, *NPAS2*, *CRY1*, *CRY2*, *PER1*, *PER2*, *PER3*,*CSNK1E*) and the other genes that are involved in other auxiliary loops.

Odds ratios (OR) and corresponding 95% confidence intervals (CI) were calculated using unconditional logistic regression models adjusted for the matching factors (age, area of residence) and for established risk factors of breast cancer (age at menarche, age at first full-term pregnancy, parity, current use of menopausal hormone therapy, body mass index, alcohol consumption and tobacco consumption). An OR of 1.42 (95% CI: 1.08-1.88) (p=0.01) was observed in women that have a lifetime duration of nightwork greater than 2 years compared to less. The interaction between the polymorphisms in circadian genes and nightwork were first analysed using a SNP by SNP approach and no interaction term was statistically significant after correction for multiple testing (results not shown). Gene-level and pathway-level interaction p-values obtained by

 ISSN: 2102-6238

the FM, MinP, ARTP and iSKAT are shown in Table 13 for 1000 resampling. The parameters of the ARTP are calibrated in the same way than in the simulation part (see section 4.2).

TABLE 13. *Results of the investigation of gene-environment interaction of Circadian Pathway using 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP.*

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	
ARNTL	24	**0.04**	**0.001**	0.1359	0.1578	**0.0539**	**0.001**	**0.0078**
PER1	5	**0.0849**	**0.035**	**0.032**	**0.04**	**0.049**	**0.0619**	0.1132
NPAS2	62	0.2647	0.1249	0.0879	0.1309	0.2957	0.1608	0.4338
CSNK1E	9	0.3337	0.3427	0.4166	0.4965	0.4655	0.5265	0.6011
CRY1	7	0.4935	0.5235	0.4985	0.5504	0.5295	0.5694	0.6081
CRY2	9	0.5425	0.7722	0.6603	0.9111	0.6374	0.8761	0.3475
PER2	11	0.9071	0.967	0.9401	0.983	0.8721	0.9491	0.734
PER3	15	0.8901	0.984	0.8571	0.976	0.8941	0.995	0.5695
CLOCK	11	0.9181	0.99	0.982	1	0.964	0.999	0.8104
subpathway		0.2967	**0.007**	0.2537	0.2997	0.2957	**0.0020**	0.2095
FBXL3	7	0.1658	0.0769	0.1239	0.1848	0.1628	0.1918	0.2986
SKP1	4	0.4046	0.4236	0.2128	0.2318	0.2827	0.3057	0.4056
CSNK1D	3	0.3706	0.3906	0.4256	0.4655	0.3736	0.4056	0.4412
RBX1	2	0.4146	0.3876	0.5005	0.4825	0.4505	0.4296	0.4977
BHLHE40	9	0.3277	0.3237	0.5864	0.7113	0.4126	0.4396	0.7204
RORA	288	0.3836	0.3027	0.6184	0.7612	0.4515	0.4456	0.4576
NR1D1	8	0.4476	0.4985	0.3906	0.4945	0.4206	0.4825	0.5439
RORC	14	0.1958	0.1139	0.4226	0.6434	0.2687	0.5055	0.4815
CUL1	23	0.4585	0.5814	0.1009	0.1578	0.3916	0.5105	0.1035
TIMELESS	7	0.6643	0.7632	0.4406	0.5504	0.5395	0.6513	0.7471
BTRC	13	0.977	0.998	0.4236	0.6064	0.7153	0.7013	0.8507
RORB	34	0.5974	0.7163	0.5994	0.7972	0.5614	0.7293	0.7152
FBXW11	8	0.8741	0.952	0.8711	0.9361	0.9121	0.962	0.7715
BHLHE41	4	0.959	0.983	0.8891	0.957	0.9211	0.97	0.8227
subpathway		0.9231	0.6474	0.7682	0.9101	0.8042	0.8951	0.5552
circadien		0.6054	**0.009**	0.5085	0.6114	0.6374	**0.02**	0.4166

At the gene level, we observed that both methods FM and ARTP highlight the same two genes *PER1* and *ARNTL* in the interaction analysis with nightwork, while only *PER1* is significant with the MinP method and only *ARNTL* is significant with the iSKAT method. Bootstrap resampling method tends to give lower p-values than permutations for these two genes in particular. This is in accordance with the simulation section in which we shown that the parametric bootstrap method is more powerful for large sample size.

At the pathway level, a significant interaction p-value (see Table 13) was observed for the overall circadian pathway for both FM and ARTP methods when parametric bootstrap is used while no association is observed using permutation resampling approach. This association is observed only for the core circadian genes subpathway that includes the genes *PER1* and *ARNTL*. No significant association was observed while using the methods MinP and iSKAT.

To summarize, FM and ARTP gave similar results in our data. Significant interaction p-values were observed at the gene and pathway levels using the boostrap resampling method, while only

 ISSN: 2102-6238

significant results at the gene level were observed using the permutation resampling method. MinP and iSKAT methods highlighted only part of the genes that were found significant by FM and ARTP methods and reported non-significant interaction at the pathway level.

PER1 and *ARNTL* which are highlighted in the gene level analysis, are important components of the circadian system which is regulated by molecular feedback loops. Heterodimers composed of ARNTL and either of the two related proteins CLOCK or NPAS2 are transcriptional factors that induce the expression of *PER* and *CRY* genes by binding to their promoters, which in turn will act on the ARNTL-CLOCK/NPAS2 complex to repress their own transcription.

Variants in both genes has been previously associated to breast cancer risk (Hansen and Lassen, 2012; Zienolddiny et al., 2013). The finding with *PER1* from the interaction analysis may be of particular interest, as a variant in *PER1* (rs2735611) was previously associated with an extreme morning preference (Carpen et al., 2006), a condition that was associated with an increased breast cancer risk among Danish military women working in night shifts (Hansen and Lassen, 2012).

5.1. Running time performance

The most demanding part of the p-value algorithms in terms of time computation is the resampling part. All p-value combining methods have been ran with the same resampling samples. We focus on the mesure of the running time related to this part of the algorithm. Tables 14 and 15 presents the running time performances of permutation and bootstrap approaches. The results given are computed on one standard core, and results are running times on the application data (see table 14) and on 500 iterations of simulation case 5 (see table 15). We can see that bootstrap and permutation have similar running times. The running time of iSKAT is added in comparison. p-value combining methods have a much higher running time than iSKAT. Hopefully it can be computed in parallel, whereas iSKAT cannot.

TABLE 14. *Running time: permutation and bootstrap performances using 1000 resampling related to the application. Results are in seconds.*

Running time		
permutation	bootstrap	iSKAT
20187.5	20264.8	486.3

TABLE 15. *Running time: permutation and bootstrap performances using 1000 resampling related to the simulation case 5. Results are the average time in seconds over 500. replications.*

Running time								
size 200			size 500			size 1000		
Permutation	Bootstrap	iSKAT	Permutation	Bootstrap	iSKAT	Permutation	Bootstrap	iSKAT
488	491	2.37	611	709	2.81	1170	1376	3.58

 ISSN: 2102-6238

6. Concluding Remark

Based on the work of Yu et al. (2009), we have proposed an efficient practical tool for investigating gene- and pathway-environment interaction. Both FM and ARTP methods are extended in this context and available through our R package `PIGE` (Liquet et al., 2017). Permutation and parametric bootstrap approaches have been implemented. Our simulation study suggests sightly better results from bootstrap compared to permutation, especially when the number of significant SNP is high. Furthermore we have shown that our proposed methods can be competitive and even slightly more powerful then the cutting edge methods like iSKAT.

The cornerstone of the implemented approaches are the running time of the resampling approaches which could be problematic in presence of large data set (i.e., large sample size and large number of genetic information). To overcome this issue, PIGE offers a parallel implementation of these approaches. As an example, our application on interaction between circadian genes and night work in breast cancer risk which includes $n = 2300$ subjects and $p = 577$ SNPs took 45 minutes with the permutation procedure and 1 hours 5 minutes using 4 cores and 1000 resampling.

In this application study, using ARTP method with the parametric bootstrap approach, we highlighted significant interactions at the pathway-level which were missed when using the permutation procedures. Our results suggest that polymorphisms in the circadian rhythms pathway could modulate the association between night work and breast cancer. This association seems to be driven mostly by the genes *PER1* and *ARNTL*.

Note that our approaches can deal in the context of $p > n$ as the methods are based on combining individual p-values. Finally, our proposed approaches are not restricted to a binary case-control outcome. In this study, we focus the presentation on an binary environment variable which was motivated by binary environment data of our application. The method is not restricted to binary environment variable and has been extended and implemented in our R package PIGE (Liquet et al., 2017) to any quantitative environment variable. Further, our package also offers the possibility to deal with survival outcome variable or quantitative outcome in general. It is also possible to investigate gene- and pathway-environment interaction for more than one pathway during the same analysis.

Acknowledgement

Most of the computations presented in this paper were performed thanks to the computing facilities the Direction du Numï¿½rique of the Universitï¿½ de Pau et des Pays de l'Adour provided us.

References

Albert, P., Ratnasinghe, D., Tangrea, J., and S., W. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154:687–693.

Buzkova, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *American Journal of Human Genetics*, 75(1):36–45.

Carbonetto, P. and Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central for for il-2 signaling genes in type 1 diabetes and cytokine signaling genes in crohn's disease. *PLoS Genetics*, 9:e1003770.

 ISSN: 2102-6238

Carpen, J., von Schantz, M., Smits, M., Skene, D., and Archer, S. (2006). A silent polymorphism in the per1 gene associates with extreme diurnal preference in humans. *Journal of Human Genetics*, 51:1122–1125.

de Leeuw, C., Neale, B., Heskes, T., and Posthuma, D. (2016). The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17:353–364.

Dudbridge, F. and Koeleman, B. (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology*, 25:360–366.

Edgington, E. (1987). *Randomization tests*. Statistics, textbooks and monographs. M. Dekker.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, London.

Evangelou, M., Dudbridge, F., and Wernisch, L. (2014a). Two novel pathway analysis methods based on a hierarchical model. *Bioinformatics*, 30(5):690–697.

Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L., and Dudbridge, F. (2012). Comparison of methods for competitive tests of pathway analysis. *PloS one*, 7(7):e41018.

Evangelou, M., Smyth, D., Fortune, M., Burren, O., Walker, N., Guo, H., Onengut-Gumuscu, S., Chen, W., Concannon, P., Rich, S., Todd, J., and Wallace, C. (2014b). A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genetic Epidemiology*, 38(8):661–670.

Fridley, B. and Biernacka, J. (2011). Gene set analysis of snp data: benefits, challenges and future directions. *European Journal of Human Genetics*, 19:837–843.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77.

Good, P. (2000). *Permutation Tests: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New-York.

Hansen, J. and Lassen, C. (2012). Nested case control study shift work and breast cancer risk among women in the danish military. *Occupational and Environmental Medicine*, 69:551–556.

Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., Consortium, W. T. C.-C., Owen, M. J., MC'Donovan, O., and Carddock, N. (2009). Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics*, 85:13–24.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92:841–853.

Jiao, S., Hsu, L., Bezleau, S., Brenner, H., Chan, A., Chang-Claude, J., Le Marchand, L., Lemire, M., Newcomb, P., Slattery, M., and Peters, U. (2013). Sberia: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genetic Epidemiology*, 37(5):452–464.

Jiao, S., Peters, U., Berndt, S., Bezleau, S., Brenner, H., Campbell, P., Chan, A., Chang-Claude, J., Lemire, M., Newcomb, P., Potter, J., Slattery, M., Woods, M., and Hsu, L. (2015). Powerful set-based gene-environment interaction testing framework for complex diseases. *Genetic Epidemiology*, 39:609–618.

Lin, X., Lee, S., Christiani, D., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72(1):156–164.

Liquet, B. and Riou, J. (2013). Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 13(1):75.

Liquet, B., Truong, T., and Broc, C. (2017). *PIGE: Self Contained Gene Set Analysis for Gene- And Pathway-Environment Interaction Analysis*. R package version 1.1.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., DJ, H., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A., Cho, J., Guttmacher, A., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C., Slatkin, M., Valle, D., Whittemore, A., Boehnke, M., Clark, A., Eichler, E., Gibson, G., Haines, J., Mackay, T., McCarroll, S., and Visscher, P. (2009). Functional and genomic context in pathway analysis of gwas data. *Nature*, 461(7265):747–53.

Menegaux, F., Truong, T., Anger, A., Cordina-Duverger, E., Lamkarkach, F., Arveux, P., Kerbrat, P., Fevotte, J., and Guenel, P. (2013). Night work and breast cancer: a population-based case control study in france (the cecile study). *International Journal of Cancer*, 132:924–931.

Mooney, M., Nigg, J., McWeeney, S., and Wilmot, B. (2014). Functional and genomic context in pathway analysis of gwas data. *Trends in Genetics*, 30(9):390–400.

O'Dushlaine, C., Heron, E. A., Segurado, R., Gill, M., Morris, D. W., and Corvin, A. (2009). The snp ratio test:

 ISSN: 2102-6238

pathway analysis of genome-wide association datasets. *Bioinformatics, Application Note*, 25:2762–2763.

Pers, T. (2016). Gene set analysis for interpreting genetic studies. *Human molecular genetics*, 25(R2):R133–R140.

Shahbaba, B., Shachaf, C. M., and Yu, Z. (2012). A pathway analysis method for genome-wide association studies. *Statistics in Medicine*, 31:988–1000.

Straif, K., Baan, R., Grosse, Y., Secretan, B., El Ghissassi, F., Bouvard, V., Altieri, A., Benbrahim-Tallaa, L., and Cogliano, V. (2007). Carcinogenicity of shift-work, painting, and fire-fighting. *Lancet Oncology*, 8:1065–1066.

Su, Y., Gauderman, W. J., Berhane, K., and Lewinger, J. P. (2016). Adaptive set-based methods for association testing. *Genetic Epidemiology*, 40:113–122.

Truong, T., Liquet, B., Menegaux, M., Plancoulaine, S., Laurent-Puig, P., Mulot, C., Cordina-Duverger, E., Sanchez, M., Arveux, P., Kerbrat, P., Richardson, S., and Guenel, P. (2014). Breast cancer risk, nightwork and circadian clock gene polymorphisms. *Endocrine-related cancer*, 21(4):629–38.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, 81:1278–1283.

Wang, L., Jia, P., Wolfinger, R., Chen, X., and Zhao, Z. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1):1–8.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89:82–93.

Yu, K., Li, Q., andrew W., B., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology*, 33(8):700–709.

Zienolddiny, S., Haugen, A., Lie, J., Kjuus, H., K.H., A., and KjÃ¦rheim, K. (2013). Analysis of polymorphisms in the circadian-related genes and breast cancer risk in norwegian nurses working night shifts. *Breast Cancer Research*, 15:R53.

 ISSN: 2102-6238

Chapter 4

Meta-analysis methods for genomics

This chapter deals with meta-analysis in the context of genetic epidemiology. Meta-analysis consists in comparing results between different studies. In genetic epidemiology, gathering data coming from different studies is an efficient way to increase the amount of data available which can enhance the power of statistical methods. In a first section, the batch effect problem due to the meta-analysis is explained. In a second section, methods used for meta-analysis in pleiotropy context are presented. Finally a novel method for tackling this issue is presented. The method has been presented in an article in the Arabian Journal of Mathematics which is reproduced after this chapter.

4.1 The batch effect

For a meta analysis, data are gathered difference sources, i.e., from different institutions on different populations and with similar but potentially different technologies and protocols. In multi study data sets the accuracy of the results is impacted by the different experimental bias between each study. Those differences in the recollection of data can lead to involuntary bias, known as batch effect. In general, a proper step of harmonization of the different chunks of data is necessary. Several methods have been proposed to "standardize" (or normalize) data according to the sources [101, 102]. In general, before the application of a method each column of X and Y is centered according to its mean and scaled according to its standard deviation. As a basic approach for meta-analysis, the normalization has to be changed to take this diversity into account. For instance, the normalization can be performed study by study [103]. This decomposition can even be used to retrieve a study by study model from dimension reduction models like in MINT method [103].

Another approach consists in integrating the knowledge in the framework of the method [104] but this approach is less used than standardization methods. In general a mean-centering method is commonly used. It is easy to implement and need few hypothesis on data [105]. More advanced methods have been discussed extensively in notable articles [106].

In the following, we consider that study by study standardization is applied to data.

4.2 Methods for meta-analysis in genetic epidemiology

In this section three approaches to the meta-analysis question are presented. The first one is ASSET which is one of the most used methods in this field for meta-analysis. The method relies on the exploration of subsets of studies for the presence of true association signals that are in either the same direction or possibly opposite directions. CPBayes is a bayesian equivalent of the method. The second one is an extension of SKAT (see Section 2.3.5) named meta-SKAT. Finally, extensions of the Lasso methods are discussed.

A large number of other methods also exists, but are not developed in this dissertation. ASSET and meta-SKAT are presented because they are the most established methods in the field, while the last approach sets the background for the introduction of the novel method "sparse group Partial Least Square for structured data" which is one the contribution of this thesis.

Note that the mathematical notation are introduced in Section 2.1. Especially, for this chapter, the structure of data presented in Figure 2.3 is considered.

4.3 ASSET and CPBayes

ASSET and CPBayes are methods outputting statistics for meta-analysis. Both frameworks rely on common ideas, except that ASSET is a frequentist method whereas CPBayes is Bayesian.

ASSET

ASSET is a method suited for meta-analysis providing a p-value across studies [107]. The input of the method are single variables summary statistics which are combined by the method. ASSET exhaustively explores subsets of studies for the presence of true association signals that are in either the same direction or possibly opposite directions.

For a given variable i, the considered summary statistics are $\beta_{i,m}$ and $s_{i,m}$ resp. the regression parameter and standard error of a model for variable i and study m. The standard fixed effect of the variable i is calculated as a weighted sum of Z statistics:

$$Z_i = \sum_{m=1}^{M} w_{i,m} Z_{i,m} \tag{4.1}$$

with $Z_{i,m} = \frac{\beta_{i,m}}{s_{i,m}}$ and where $w_{i,m}$ is a weight calculated as

$$w_{i,m} = \frac{1/s_{i,k}}{\sqrt{\sum_{m=1}^{M} 1/w_{i,m}^2}} \tag{4.2}$$

With those weights, the statistic is known to be asymptotically equivalent to a pooled analysis of the studies:

$$Z_i = \frac{\sum_{m=1}^{M} Z_{i,m}/s_{i,m}}{\sqrt{1/s_{i,k}}} \tag{4.3}$$

In this framework, the hypothesis that some studies do not contribute to the pleiotropy effect is considered. Hence for a given subset of study $S \subset \{1, \cdots, M\}$, the pooled Z statistic associated is

$$Z_i(S) = \sum_{m \in S} \sqrt{\pi_m(S)} Z_{i,m} \tag{4.4}$$

in which $\pi_m(S) = n_m / \sum_{m \in S} n_m$ denotes the sample size of study m compared to the subset S.

$$Z_i^{meta-analysis} = \max_{\text{possible S}} |Z_i(S)| \tag{4.5}$$

The number of possible subsets S is high: 2^{K-1} but all the combinations are based on the same statistics $Z_{i,m}$ which reduces the computational cost.

In order to address the possibility of effect with opposite directions a "two sided test" is proposed (whereas the statistic $Z_{meta-analysis}$ is called "one sided test"). Positive and negative effect are detected separately through two statistics $Z_{meta-analysis,+}$ and $Z_{meta-analysis,-}$.

CPBayes

CPBayes [108] uses summary statistics combinations similar to ASSET but is based on a Bayesian framework with and an approach called spike and slab priors and a MCMC Gibbs sampling. The evidence of pleiotropy is measured by the local false discovery rate (locFDR) and with Bayes factors (BF). CPBayes also estimates the posterior probability of association (PPA) and coefficient for each phenotype.

We can note that neither ASSET nor CPBayes take into account a group structure of the variables. And, thus, they do not take advantage of any SNP-gene-pathway structure knowledge.

4.3.1 Meta-SKAT

SKAT is a method to detect association between rare variants in a region and a phenotype (continuous or binary). It is a supervised test for joint effects of multiple variants in a region on a phenotype. Meta-SKAT can do the same but aggregating several studies. This method outputs a p-value corresponding to a set of variables, for instance a gene or a pathway. The method is based on a weighted sum of SKAT statistics of the different studies [109].

4.3.2 Lasso penalization for meta analysis on dimension reduction methods

Different methods have been proposed for using dimension reduction methods in the case of different sets of observations. They rely on simultatenous selections of variables for each study. The fuse Lasso proposed by Tibshirani [110] encourages the selected variables to be similar from one study to another. A joint group Lasso that selects the same variables for each study has also been proposed [111] .

For each study m, u_m and v_m optimal weight vectors are computed to maximize the covariance $cov(X_m u_m, Y_m v_m)$. Let us gather in two matrices those weight vectors, U a $p \times M$ matrix and V a $q \times M$ matrix in which columns are respectively the vectors u_m and v_m. The rows correspond to all the weights related to a same variable. Then, we want

to set to zero all the weights related to a same variable at the same time. This is why a L_2 penalization is introduced on the columns $(U_{\cdot,i})$ of U.

Application of those penalization to the sparse PLS are presented and they can be generalized to sparse group PLS. Data have the structured presented in Figure 2.3 from Section 2.1. The fuse Lasso for a sparse PLS (penalizing X) is:

$$\begin{aligned}\{U_{opt}, U_{opt}\} = \underset{U,V}{\operatorname{argmin}} \sum_{m=1}^{M} \left\| Z^{(m)} - U_{\cdot,m} V_{\cdot,m}^T \right\|_F^2 + \lambda_1 P_{\text{Lasso}}(U) + \lambda_2 P_{\text{Fused}}(U) \\ \text{with } P_{\text{Lasso}}(U) = \sum_{i=1}^{p} \|U_{i,\cdot}\|_1 \\ \text{with } P_{\text{Fused}}(U) = \sum_{i=1}^{p-1} \|U_{i,\cdot} - U_{i+1,\cdot}\|_1 \\ \text{and } Z^{(m)} = X_{\mathbb{M},\cdot}^T Y_{\mathbb{M},\cdot}\end{aligned} \tag{4.6}$$

where λ_1 and λ_2 are parameters driving the fuse Lasso penalization P. P_{Lasso} is the classical Lasso penalization whereas P_{Fused} is the fused term.

The common Lasso introduced by Obozinski is:

$$\begin{aligned}\{U_{opt}, U_{opt}\} = \underset{U,V}{\operatorname{argmin}} \sum_{m=1}^{M} \left\| Z^{(m)} - U_{m,\cdot}^T V_{m,\cdot} \right\|_F^2 + \lambda P(U) \\ \text{with } P(U) = \sum_{i=1}^{p} \|U_{i,\cdot}\|_2 \\ \text{and } Z^{(m)} = X_{\mathbb{M},\cdot}^T Y_{\mathbb{M},\cdot}\end{aligned} \tag{4.7}$$

where λ a parameter driving the penalization P.

The fuse Lasso forces the coefficients corresponding to different observation sets to be similar whereas the second penalization forces variables shrunk to zero to be shrunk to zero for all observation sets at the same time.

This second penalization can then be seen as a variable selection method. This is why we are interested in studying this kind of penalization.

4.4 Sparse group Partial Least Square for structured data

This part introduces the contributions developed during this PhD: the sgPLS for structured data. The theoretical background of the method has been presented in [112]. The method has also been presented in two conferences at the French Society of Statistics [113] [114]. The content of those presentations is not included in this section as they do not add additionnal material to this dissertation and the presented journal articles. The contribution is a Partial Least Square to which the Lasso penalization presented by Obozinski [111] is adapted.

In this part, we consider that data that are composed of independent observation sets and group of variables and this structure is knwon a priori. The presented methods allow

us to use the information about the edification of the data set in order to improve the performance of the analysis. Although this theory has been developed with the aim to answer a problem occurring in genomic public data sets, it can be applied to any field where a certain observation set structure exists. The method called “penalized PLS for structured data" is defined where separate PLS model are linked together with a common-Lasso penalization similar to the one developed in [111]. Variables selected by the model are the same for all observation sets but the underlying model computes separated models for each observation set, giving both readability and flexibility to the model. The theoretical background for this method is presented.

The implementation have been performed through an R package [115].

4.4.1 Framework of proposed method

The method is based on the following key results. In the same way that for Section 2.3.3 Lasso penalization is applied to X in the formulas but can be also applied to Y. Taking the notations from Section 4.3.2, we can introduce the following sparse group PLS formulation:

$$\begin{aligned} &\min_{U,V} \sum_{m=1}^{M} \left\| Z^{(m)} - U_{\cdot,m} {V_{\cdot,m}}^T \right\|_F^2 + P_\lambda(U) \\ &\text{with } P_\lambda(U) = \lambda \sum_{k=1}^{K} \sqrt{(}p_k) \left\| U_{\mathbb{P}_k,\cdot} \right\|_F \\ &\text{with } Z^{(m)} = X_{\mathbb{M}_m,\cdot}^T Y_{\mathbb{M}_m,\cdot} \end{aligned} \tag{4.8}$$

The biconvex resolution use the following formulas. Fixing each $\left\|V_{\cdot,m}\right\|_2 = 1$ the optimal $U_{(k,m)}$ is

$$U_{(\cdot,m)} = \left(1 - \frac{\lambda}{2\sqrt{\sum_{m=1}^{M} \left\| Z_{\mathbb{P}_k,\cdot}^{(m)} V_{\cdot,m} \right\|_F^2}} \right)_+ Z_{\mathbb{P}_k,\cdot}^{(m)} V_{\cdot,m} \tag{4.9}$$

and fixing $\left\|U\right\|_2 = 1$ the optimal V is

$$\begin{aligned} V_{(\cdot,m)}, &= Z^{(m)T} V_{(\cdot,m)} \\ \text{with } Z^{(m)} &= X_{\mathbb{M}_m,\cdot}^T Y_{\mathbb{M}_m,\cdot}. \end{aligned} \tag{4.10}$$

where u_m implies another thresholding function. The thresholding function sets to zero all the weights of a same gene at the same time. Gene selection is performed across all the study in this way.

Group Lasso and single variable Lasso can be combined in a Sparse Group Sparse PLS for

structured data. The minimization problem becomes:

$$\begin{aligned}\min_{U,V} \sum_{m=1}^{M} \left\| Z^{(m)} - U_{\cdot,m} V_{\cdot,m}^T \right\|_F^2 + (\alpha)\, P_\lambda^{(single)}(U) + (1-\alpha)\, P_\lambda^{(group)}(U) \\ \text{with } P_\lambda^{(single)}(U) = \lambda \sum_{i=1}^{p} \|U_{i,\cdot}\|_2 \\ P_\lambda^{(group)}(U) = \lambda \sum_{k=1}^{K} \sqrt{(p_k)}\, \|U_{\mathbb{P}_k,\cdot}\|_F \\ \text{with } Z^{(m)} = X_{\mathbb{M}_{\gg},\cdot}^T Y_{\mathbb{M}_{\gg},\cdot}\end{aligned} \tag{4.11}$$

The biconvex resolution use the following formulas. Fixing each $\left\|V_{(\cdot,m)}\right\|_2 = 1$ the optimal $U_{(k,m)}$ is

$$\begin{aligned} U_{\mathbb{P}_k,\cdot} &= \left(1 - \frac{\lambda\,(1-\alpha)}{2\left\|\tilde{U}_{\mathbb{P}_k,\cdot}\right\|_F^2}\right)_+ \tilde{U}_{\mathbb{P}_k,\cdot} \\ \text{with } \tilde{U}_{i,\cdot} &= \left(1 - \frac{\lambda\alpha}{2\left\|Z_{i,\cdot}V\right\|_2}\right)_+ Z_{i,\cdot}V_{i,\cdot} \end{aligned} \tag{4.12}$$

and fixing $\|U\| = 1$ the optimal V is

$$\begin{aligned} V_{\cdot,m}, &= Z^{(m)T} V_{\cdot,m} \\ \text{with } \ Z^{(m)} &= X_{\mathbb{M}_m,\cdot}^T Y_{\mathbb{M}_m,\cdot} \end{aligned} \tag{4.13}$$

The Joint-Lasso needs to set two penalization parameters. The larger the first parameter is the lesser variables are selected in the model. The parameters can be optimized for instance by minimizing the error prediction of the model under a cross-validation procedure. The minimum zone being quite flat we allow the parameters to be in a neighborhood of the minimum.

Remark: If the number of observations is different from one observation set to another, the observation sets with the largest number of observation can prevail in the model. In order to cope with this, we propose to divide data from a an observation set m by $\sqrt{n_m}$.

4.5 Conclusions

This chapter presents the problems specific to meta-analysis. An extension of the PLS is proposed which has the merit of taking into account both the group of variables structure and the study structure. We will see in next chapter that the method can be used for a particular case of meta-analysis where few method are well established: pleiotropy studies. After this chapter, the original work where the method was proposed is reproduced.

Penalized Partial Least Square applied to structured data

Camilo Broc, Borja Calvo and Benoit Liquet

Published in Arabian Journal of Mathematics

Arabian Journal of Mathematics manuscript No.
(will be inserted by the editor)

Penalized Partial Least Square applied to structured data

Camilo Broc · Borja Calvo · Benoit Liquet

Received: date / Accepted: date

Abstract Nowadays, data analysis applied to high dimension has arisen. The edification of high dimensional data can be achieved by the gathering of different independent data. However each independant set can introduce its own bias. We can cope with this bias introducing the observation set structure into our model. The goal of this article is to build theoretical background for the dimension reduction method sparse Partial Least Square (sPLS) in the context of data presenting such an observation set structure. The innovation consist in building different sPLS models and linking them through a common-Lasso penalization. This theory could be applied to any field where observation present this kind of structure and therefore improve the sparse Partial Least Square in domains where it is competitive. Furthermore it can be extended to the particular case where variables can be gathered in given a priori groups, where sparse Partial Least Square is defined as a sparse group Partial Least Square.

Keywords Batch Effect · High dimensional data · Partial Least Square · Sparse methods

1 Introduction

Since past years data analysis applied to high dimension in all domains has arisen [1]. Extracting information from ever larger data has become a trend in numerous fields and a large number of observation need to be gathered in order to evaluate statistical models. When data are hard to retrieve, gathering existing data sets is an efficient way for assembling data of high dimension. However this technique have its drawbacks : existing independent data sets can present intrinsic bias which can decrease the performance of the models used.

Camilo Broc
Laboratoire De Mathématiques et de leurs Applications de PAU Fédération MIRA, UMR5142 64000, Pau, France
E-mail: camilo.broc@univ-pau.fr

Borja Calvo
Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia, 20018, Spain

Benoit Liquet
Laboratoire De Mathématiques et de leurs Applications de PAU Fédération MIRA, UMR5142 64000, Pau, France And Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical, Sciences at Queensland University of Technology, Brisbane, Australia.

Those biases imply an unwanted underlying structure that will interfere with the signal we want to find. Bias can come from a difference in the source of information or the process used during the recollection of the data. This set structure has to be taken into account in order to improve the efficiency of the models. For instance, in genomics, data can be gathered from different studies because of the cost of the experimentation. Each clinical study may have been performed with its own chemistry protocol, with its own experimental material and on its specific populations, and bias can arise among the different data sets obtained. This "batch effect" is known and can significantly decrease the power of the analysis [2]. Another bias can occur in particular analysis where different "dynamics" exist between the studies : a predictor can be highly correlated with independent variable, but the direction of the correlation depends on the study. For instance, pleiotropy [3] is a field of genetics where a gene (predictor) can have a particular effect on different phenotypes (independent variables). Data can be gathered from different studies where the nature of the phenotype differs. Therefore, a gene can be highly correlated with each phenotype but an overall model struggles to catch the particularity of those effects.

In the article we tackle the problem of "batch effect" for dimension reduction such as Partial Least Square (PLS) method introduced by Wold [4]. Common dimension reduction techniques are Canonical Correlation Analysis (CCA) [5], Principal Component Analysis (PCA) [6] and PLS [7]. All these methods rely on the projection of the data into a subspace of lower dimension which represents most of the variation of the data. They are often posed as an eigen value problem [8]. PLS and CCA are both analysis two blocks of data and differ from the norm used whereas PCA analyses one block. Aiming to apply our method to supervised analysis, PLS approach is considered in this article.

In these dimension reduction techniques, results are formulated with new variables that are linear combination of the original ones. These combination can be hard to interpret due to the huge number of coefficient they represent. To answer this problem, Lasso methods have been used. Introducing this penalization shrink to zero the participation to the model of the least relevant variables. Results highlight a smaller number of variable that are easier to explain. In addition, noise of the signal is reduced and the power of the methods is boosted. These are called sparse method and have been developed for linear regression [9] [10], CCA [11], PCA [12] and Partial Least Square (sPLS). The sparse PLS (sPLS) has shown encouraging results [13] [14] and is the object of analysis in this article. The PLS and sPLS methods have also been used to control "the batch effect" when related studies are combining to increase sample size combining independent but related studies ([15],[16]). In particular combining sPLS separating models and linking them can be an option like in the Multivariate INTegrative method (MINT) proposed in [16]. However, this approach cannot identify the true signal in presence of different dynamics.

For high dimensional regression problems, using problem-specific prior information improves the accuracy of the prediction and the interpretability of the model[17]. For example, in genomics, genes within the same pathway have similar functions and act together in regulating a biological system. Incorporation of this grouping structure is becoming increasingly common due to the success of gene set enrichment analysis approaches [18]. Using a model taking into account this variable group structure allow to improve the performance and the readability of the results. To this end sparse group Partial Least Square (sgPLS) have been developed [19] where two overlaid Lasso penalizations translate the group structure in the Partial Least Square formulation. A structure with group and sub-groups can also be handle by its generalization with three overlaid Lasso penalizations (sgsPLS [20]). Methods such as MINT don't take into account this kind of group structure.

In this article we consider data that are composed of independent observation sets. The observation sets are assumed to be known and are expected to introduce bias in the data. The presented methods allow us to use the information about the edification of the data set in order to improve the performance of the analysis. Although this theory have been developed with the aim to answer a problem occurring in genomic public data sets, it can be applied to any field where a certain observation set structure exists. Different method using Lasso penalization on data structured toward observation sets are discussed. In particular a "penalized PLS for structured data" is defined where separate PLS model are linked together with a common-Lasso penalization. In the end variables selected by the model are the same for all observation sets but the underlying model computes separated models for each observation set, giving both readability and flexibility to the model. We present the theoretical background for this method. Especially, we can show that the common-Lasso constraint that is used (i.e. a penalization across studies) can be be written as a standard Lasso with an overlaid group structure in an equivalent formulation of the PLS problems. We extend also this idea of common-Lasso constraint to a case where an a priori structure is known, where the variables are gathered into groups.

2 Notations

Before going into further details, the notation used in this article are introduced. Data are represented by $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{qn}$, two matrices, representing n observations of p predictors and q independent variables. Then X is a (n,p) matrix and Y a (n,q) matrix. For any matrix A of size (a,b) , for $i \in \{1,\dots,a\}$ its rows are noted $A^{(i,\cdot)}$ and for $j \in \{1,\dots,b\}$ its columns are noted $A^{(\cdot,j)}$ and for subsets $\tilde{a} \subset \{1,\dots,a\}$ and $\tilde{b} \subset \{1,\dots,b\}$ resp. row and column sub-matrices are noted $A^{(\tilde{a},\cdot)}$ and $A^{(\cdot,\tilde{b})}$. For any vector ω of size a , for $i \in \{1,\dots,a\}$ its elements are noted $\omega^{(i)}$ and for subsets $\tilde{a} \subset \{1,\dots,a\}$ $\omega^{(\tilde{a})}$ represents the elements of the vector corresponding to the positions in the subset. Matrices will always be in uppercase letters and vectors in lowercase letters to avoid any confusion.

The Frobenius norm on matrices is denoted $\|\ \|_F$. We note X^T the transpose matrix of X. The cardinal of a set S is noted $\#S$. The positive value of a real number x is noted $(x)_+ = \frac{|x|+x}{2}$.

2.1 Data with observation sets

Some data may present a structure among the observations gathered around groups of observations. For instance data can be composed of different studies, each one presenting its own mechanisms and bias. Let us consider M different sets in the data. Noting, for $m \in \mathbb{N}$, $\mathbb{M}_m$ a subset of $\{1,\dots,n\}$, let $\mathbb{M} = (\mathbb{M}_m)_{m=1..M}$ be a partition of $\{1,...,n\}$ corresponding to the observation sets. We note $\#\mathbb{M}_m = n_m$. Row blocks are defined by this partitions in Figure 1 (observations are assumed to be ordered by observation set).

2.2 Data with group of variables

Some data may present a structure among the variable gathered around groups. Let us consider that the variables are gather in K groups. Let $\mathbb{P} = (\mathbb{P}_k)_{k=1..K}$ be a partition of $\{1,...,p\}$ corresponding to this variable group structure. We note $\#\mathbb{P}_k = p_k$. We then have $\sum_{k=1}^{K} p_k = p$. This partition can define column blocks among the variables if

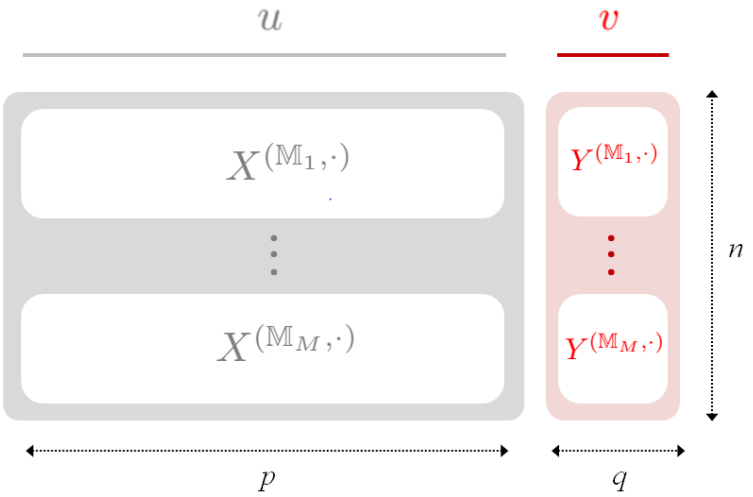


Fig. 1 Illustration of data structured by group of observation. Observations are assumed to be ordered by observation set.

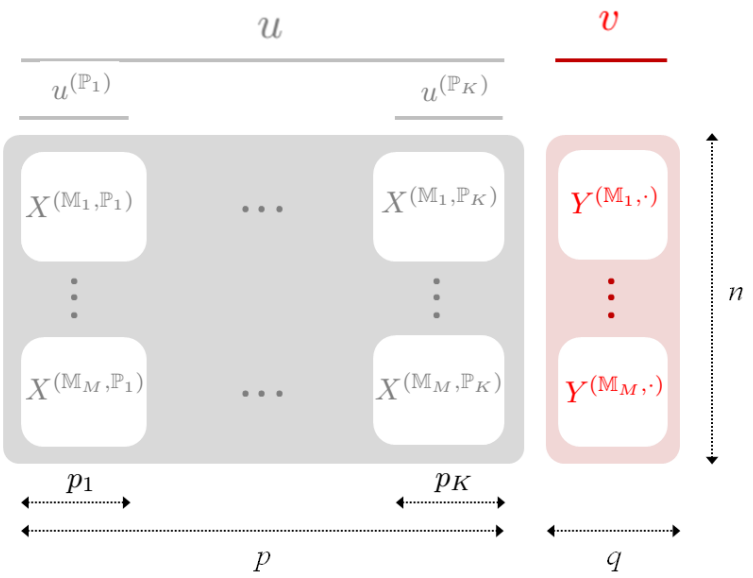


Fig. 2 Illustration of data structured by group of variables and group observation. Variables are assumed to be ordered by variable group.

variables are assumed to be ordered by variable group. Both observation set structure and variable group structure can be defined at the same time like in Figure 2.

3 Formulation of the sparse Partial Least Square

In the literature, two formulation of the Partial Least Square exist, some extension of the PLS follow a first one usually called PLS1 [21] and other extensions follow a second one called "PLS2" [14]. In the context of the article we study exclusively the first one.

3.1 PLS and sPLS

Let X be a predictor matrix of size (n, p) and Y a matrix of independent variables of size (n, q). PLS finds successively couples of vector $\{u_1, v_1\}, \ldots, \{u_r, v_r\}$ for $r < \min(p, q)$ where the couples are composed of vectors of length resp. p and q, maximizing $Cov(Xu_i, Yv_i)$ for any $i \in \{1, \ldots, r\}$, under the constraint that the family of vectors $u_1, \ldots, u_r$ and $v_1, \ldots, \ldots, v_r$ are both of them orthogonal families [4]. It can be solved considering successive maximization problems [22], for $h \in \{1, \ldots, r\}$

$$\max_{||u_h||_2=||v_h||_2=1} Cov(X_{h-1}u_h, Y_{h-1}v_h),$$

where $X_0 = X$, $Y_0 = Y$ and X_{h-1}, Y_{h-1} are deflated matrices computed from u_{h-1},v_{h-1}, X_{h-2}, Y_{h-2} for $h \in \{2, \ldots, r\}$. The deflation depends on the PLS mode that is chosen ([23],[4]). In this article we focus on the enhancement of the optimization problem and its Lasso formulation in its h-th step. According to [22] this step can be written as

$$\{u_{opt}, v_{opt}\} = \underset{||u||_2=||v||_2=1}{\operatorname{argmin}} \left\|X^TY - uv^T\right\|_F^2 + \underbrace{\lambda P(u)}_{\substack{\text{Lasso Penalty term} \\ \text{for sparse PLS}}} . \quad (1)$$

where the notation h is removed in order to simplify the formulation because we are interested in only one of the r steps of the PLS.

The sparse PLS introduces a penalization in this formulation of the problem. The penalty $P(\cdot)$ forces lowest values of u to be set to zero. The parameter controlling the degree of sparsity in the model is λ. In the presented formula the sparsity is applied only to the vector u, but a similar penalization can be define for v. In the context of this article we treat only the penalization of u but all the results stand also for a v penalization. The following sections compare different ways of writing the sPLS optimization problem presented in Equation (1) taking into account an observation or/and variable set structure.

Remark: Before analysis, the X and Y matrices are transformed by subtracting their column averages. Scaling each column by their mean and standard deviation is also often recommended [24]. Thus, the cross-product matrix X^TY is proportional to the empirical covariances between X- and Y -variables when the columns of X and Y are centered. When the columns are standardized, X^TY is proportional to the empirical correlations between X- and Y -variables. In this article the standardization is an important step to overcome the issue of the "batch effect" or to aggregate observations from different studies.

3.2 Formulation of the penalized PLS

Six different formulations of the sPLS are presented in this article. The first four correspond to data presenting an observation set structure like in Figure 1. The two last correspond to data presenting an observation set structure and a variable group structure like in Figure 2 which correspond to sgPLS models (see [19]). We can note that Problem 5 is a particular case of Figure 2 where there is only one observation set ($M = 1$). Loading vectors introduced in those figures refer to vectors formulated in the following problems. The study of Problems 4 and 6 are the main contribution of the article.

– Problem 1 (standard sPLS): This approach consists in simply considering all the observation set as one set. Data are standardized across all the sets, i.e. X and Y are standardized. The formulation is a standard sPLS problem

$$\{u_{opt}, v_{opt}\} = \underset{||u||_2=||v||_2=1}{\operatorname{argmin}} \left\|X^T Y - uv^T\right\|_F^2 + \lambda P(u). \tag{2}$$

In the model the loading u is composed of p elements and the loading v is composed of q elements. The sparsity of u is controlled by the parameter λ : for a given λ, s_λ elements of u will be non-zero.

– Problem 2 (MINT): Introduced in [16], this approach consists in considering M different sPLS problems corresponding to each of the M observation sets. Data are standardized within each observation set, i.e. for every $m \in \{1, \ldots, M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. The sPLS problem is the same than in previous problem in Equation (2).
In the model, the loading u is composed of p elements and the loading v is composed of q elements. The sparsity of u is controlled by the parameter λ : for a given λ, s_λ elements of u will be non-zero.

– Problem 3 (multiple sPLS): This approach consists in considering all the observation set as one set, i.e. $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized. Data are standardized within each observation set, i.e. for every $m \in \{1, \ldots, M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. Formulation is a classic sPLS problem

$$\{u_{m,opt}, v_{m,opt}\} = \underset{||u_m||_2=||v_m||_2=1}{\operatorname{argmin}} \left\|X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)} - u_m v_m^T\right\|_F^2 + \lambda_m P(u_m). \tag{3}$$

In the model the set of loading $\{u_m\}_{m\in\{1,\ldots,M\}}$ is composed of $p \times m$ elements (p elements per u_m). The set of loading $\{u_m\}_{m\in\{1,\ldots,M\}}$ is composed of $q \times m$ elements (q elements per v_m). The sparsity of u_m is controlled by the parameter λ_m : for a given λ_m, s_{m,λ_m} elements of u_m will be non-zero. Therefore, variables concerned by the shrinkage to zero will depend on the observation set m.

– Problem 4 (" sparse PLS for structured data"): This approach consists in considering M different sPLS problems corresponding to each of the M observation sets. Data are standardized within each observation set, i.e. for every $m \in \{1, \ldots, M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. All problems are solved at the same time with a common-Lasso.
The formulation of the problem is

$$\begin{aligned} \{U_{opt}, V_{opt}\} &= \underset{U,V}{\operatorname{argmin}} \sum_{m=1}^{M} \left\|Z_m - U^{(\cdot,m)} V^{(\cdot,m)^T}\right\|_F^2 + \lambda P(U) \\ \text{with } P(U) &= \sum_{i=1}^{p} \left\|U^{(i,\cdot)}\right\|_2 \text{ and } Z_m = X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)}. \end{aligned} \tag{4}$$

In the model the set of loading U is composed of $p \times m$ elements (p elements per $U^{(\cdot,m)}$). The set of loading V is composed of $q \times m$ elements (q elements per $V^{(\cdot,m)}$). The sparsity of all $U^{(\cdot,m)}$ is controlled by the parameter λ : for a given λ, the same s_λ elements of each $U^{(\cdot,m)}$ will be non-zero.

– Problem 5 (classical sgPLS): When variables can be gathered in groups (Figure 2), the sgPLS propose to add a group-Lasso penalization to the classical PLS. Data are standardized within each observation set, i.e. for every $m \in \{1, \ldots, M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and

$Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. The formulation of the problem is

$$\begin{gathered}\{u_{opt}, v_{opt}\} = \underset{u,v}{\operatorname{argmin}} \left\|Z - uv^T\right\|_F^2 + \lambda\,(1-\alpha)\,P_{group}(u) + \lambda\alpha P_{variable}(u) \\ \text{with } P_{group}(u) = \sum_{k=1}^{K} \sqrt{p_k} \left\|u^{(\mathbb{P}_k)}\right\|_2 \;,\; P_{variable}(u) = \sum_{i=1}^{p} \left\|u^{(i)}\right\|_2 \\ \text{and } Z = X^T Y.\end{gathered} \tag{5}$$

In the model the loading vectors u and v is composed of resp. p and q elements. The penalization $P_{variable}$ forces single variables to be set to zero whereas the penalization P_{group} forces sets of variables to be set to zero. The degree of sparsity in general in the model is λ whereas the parameter controlling the balance between both kind of sparsity is α. In this model elements of u corresponding to least relevant variables and least relevant group of variables are set to zero.

– Problem 6 (" sgPLS for structured data"): In the same spirit of adapting problem 2 into problem 4, problem 5 can be adapted with a common-Lasso penalization. Data are standardized within each observation set, i.e. for every $m \in \{1,\dots,M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. The formulation of the problem is

$$\begin{gathered}\{U_{opt}, V_{opt}\} = \underset{U,V}{\operatorname{argmin}} \left\|Z_m - U^{(\cdot,m)} V^{(\cdot,m)^T}\right\|_F^2 + \lambda\,(1-\alpha)\,P_{group}(U) + \lambda\alpha P_{variable}(U) \\ \text{with } P_{group}(U) = \sum_{k=1}^{K} \sqrt{p_k} \left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F \;,\; P_{variable}(U) = \sum_{i=1}^{p} \left\|U^{(i,\cdot)}\right\|_2 \\ \text{and } Z_m = X^{(\mathbb{M}_m,\cdot)^T} Y^{(\mathbb{M}_m,\cdot)}.\end{gathered} \tag{6}$$

In the model the set of loading U is composed of $p \times m$ elements (p elements per $U^{(\cdot,m)}$). The set of loading V is composed of $q \times m$ elements (q elements per $V^{(\cdot,m)}$). In this model elements of U corresponding to least relevant variables and least relevant group of variables are set to zero. In this model the same variables and variable groups corresponding to least significant variables are set to zero for all $U^{(\cdot,m)}$, $m \in \{1,\dots,M\}$.

4 Solutions of the penalized PLS

The classical sPLS can be seen as a biconvex optimization problem. It can be solved by successively optimizing the loading u and v [22]. For a given v an optimized $\tilde{u}$ is computed and the value of u is updated. Then the same is performed permuting the roles of u and v. This optimization process relies on solving the problems

$$\begin{aligned} u_{opt} &= \underset{||u||_2=1}{\operatorname{argmin}} \left\|X^T Y - uv^T\right\|_F^2 + \lambda P(u) \\ v_{opt} &= \underset{||v||_2=1}{\operatorname{argmin}} \left\|X^T Y - uv^T\right\|_F^2 . \end{aligned} \tag{7}$$

The solution of problems 1 to 3 (composed of standard sPLS methods) is given by the following theorem :

Theorem 1 *The marginal optima in $\tilde{u}$ and $\tilde{v}$ in the sPLS (Equation (1)) are : Fixing v, the optimal u_{opt} for (7) is*

$$u_{opt}^{(i)} = u_0^{(i)} \left(1 - \frac{\lambda}{2\left\|u_0^{(i)}\right\|_2}\right)_+ , \; u_0 = X^T Y v. \tag{8}$$

Fixing u, the optimal v_{opt} for (7) is

$$v_{opt} = Y^T X u. \tag{9}$$

In this formula a soft thresholding sets down to zero loadings corresponding to variables whose scores are too low. Setting λ equal to zero we find the formulation of the PLS problem without Lasso constraint. A proof can be find in [13].
For problems 4, 5 and 6 the solution is more complex. Problem 4 introduces a common-Lasso penalization, problem 5 introduces a variable group structure and the problem 6 introduces both common-Lasso penalization and variable group structure. We can note that problem 4 is a particular case of problem 6 where there is no group penalty, i.e. $\alpha = 1$. Problem 5 is a particular case of problem 6 where there is only one observation set, i.e. $M = 1$. The solution of problem 6 is given in theorem 2 (presented in the following) whereas solutions of problem 4 and 5 are corollaries of this theorem and can be found after the proof (corollary 1 and 2).

Theorem 2 *The marginal optima in U and V in the "sparse group PLS for structured data" (Equation (6)) are :*

Fixing V, the optimal U_{opt} for (6) is :

$$U_{opt}^{(\mathbb{P}_k,\cdot)} = U_1^{(\mathbb{P}_k,\cdot)} \left(1 - \frac{\lambda(1-\alpha)}{2\sqrt{\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2}}\right)_+ = U_1^{(\mathbb{P}_k,\cdot)} \left(1 - \frac{\lambda(1-\alpha)}{2\left\|U_1^{(\mathbb{P}_k,\cdot)}\right\|_F}\right)_+$$

$$\text{With } U_1^{(i,\cdot)} = U_0^{(i,\cdot)} \left(1 - \frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+ , \; U_0^{(\cdot,m)} = Z_m V^{(\cdot,m)} \text{ and } Z_m = X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)} \tag{10}$$

Fixing U, the optimal V_{opt} for (6) is :

$$V_{opt}^{(\cdot,m)} = Z_m^T U^{(\cdot,m)} \tag{11}$$

Proof. The proof is composed of three steps. In step 1 we settle the sub-gradient equation corresponding to the minimization problem. In step 2, we make the sPLS thresholding emerge in the equation. In step 3, we make emerge the group thresholding and prove the theorem.
Let's settle the sub-gradient equation. The optimal U for a given V is

$$\min_U \sum_{m=1}^{M} \left\|Z_m - U^{(\cdot,m)} V^{(\cdot,m)T}\right\|_F^2 + \lambda(1-\alpha)\sum_{k=1}^{K}\sqrt{p_k}\left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F + \lambda\sum_{i=1}^{p}\left\|U^{(i,\cdot)}\right\|_2 .$$

We note that the problem can be formulated making appearing the column blocks corresponding to the variable groups. A second formulation of the problem would be

$$\min_U \sum_{k=1}^{K}\sum_{m=1}^{M} \left\|Z_m^{(\mathbb{P}_k,\cdot)} - U^{(\mathbb{P}_k,m)} V^{(\cdot,m)^T}\right\|_F^2 + \lambda(1-\alpha)\sum_{k=1}^{K}\sqrt{p_k}\left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F + \lambda\sum_{k\in 1}^{K}\sum_{i\in\mathbb{P}_k}\left\|U^{(i,\cdot)}\right\|_2 .$$

We can see that the problem can be separated in K distinct problems for every $k \in \{1, \ldots, K\}$

$$\min_{U^{(\mathbb{P}_k,\cdot)}} \sum_{m=1}^{M} \left\| Z_m^{(\mathbb{P}_k,\cdot)} - U^{(\mathbb{P}_k,m)} V^{(\cdot,m))^T} \right\|_F^2 + \lambda (1-\alpha) \sqrt{p_k} \left\| U^{(\mathbb{P}_k,\cdot)} \right\|_F + \lambda \sum_{i \in \mathbb{P}_k} \left\| U^{(i,\cdot)} \right\|_2 .$$

In order to solve this problem, let's consider the k-th problem developing the Frobenius norm

$$\min_{U^{(\mathbb{P}_k,\cdot)}} \sum_{m=1}^{M} \left[\text{Trace}\left(Z_m^{(\mathbb{P}_k,\cdot)} Z_m^{(\mathbb{P}_k,\cdot)T} \right) - 2\text{Trace}\left(Z_m^{(\mathbb{P}_k,\cdot)} V^{(\cdot,m)} U^{(\mathbb{P}_k,m)T} \right) + \text{Trace}\left(U^{(\mathbb{P}_k,m)} U^{(\mathbb{P}_k,m)T} \right) \right]$$
$$+\lambda (1-\alpha) \sqrt{p_k} \left\| U^{(\mathbb{P}_k,\cdot)} \right\|_F + \lambda\alpha \sum_{i \in \mathbb{P}_k} \left\| U^{(i,\cdot)} \right\|_2 .$$

Taking the sub-gradient, the optimal U_{opt} verify for $m \in \{1, \ldots, M\}$

$$-U_{opt}^{(\mathbb{P}_k,m)} + U_0^{(\mathbb{P}_k,m)} = \frac{\lambda (1-\alpha) \sqrt{p_k}}{2} \Theta_g^{(\mathbb{P}_k,m)} + \frac{\lambda\alpha}{2} \Theta_v^{(\mathbb{P}_k,m)}$$
$$\text{with the } (p \times M) \text{ matrix } U_0 \text{ such that } U_0^{(\mathbb{P}_k,m)} = Z_m^{(\mathbb{P}_k,\cdot)} V^{(\cdot,m)}$$
$$\text{with the } (p \times M) \text{ matrix } \Theta_g \text{ such that } \Theta_g^{(\mathbb{P}_k,m)} = \begin{cases} \frac{U_{opt}^{(\mathbb{P}_k,\cdot)}}{\left\| U_{opt}^{(\mathbb{P}_k,\cdot)} \right\|_F} & \text{if } U^{(\mathbb{P}_k,\cdot)} \neq 0 \\ \Theta_g \in \{\Theta_g, \|\Theta_g\|_F \leq 1\} \text{ if } U_{opt}^{(\mathbb{P}_k,\cdot)} = 0 \end{cases}$$
$$\text{and with the } (p \times m) \text{ matrix } \Theta_v \ \Theta_v^{(i,\cdot)} = \begin{cases} \frac{U_{opt}^{(i,\cdot)}}{\left\| U_{opt}^{(i,\cdot)} \right\|_2} & \text{if } U_{opt}^{(i,\cdot)} \neq 0 \\ \Theta_g \in \{\Theta_v, \|\Theta_v\|_2 \leq 1\} \text{ if } U_{opt}^{(i,\cdot)} = 0 \end{cases} . \tag{12}$$

We can note that when there is no penalty (i.e. $\lambda = 0$,) $U_{opt} = U_0$ is the solution of the non sparse problem.

The sub-gradient equation is settle (step 1). Let's now make emerge the thresholding of sPLS.
We investigate in which case $U_{opt}^{(\mathbb{P}_k,\cdot)} = 0$, i.e. when loading corresponding to a group of variables is set to zero. If $U_{opt}^{(\mathbb{P}_k,\cdot)} = 0$ then $U_{opt}^{(i,\cdot)} = 0$ for every $i \in \mathbb{P}_k$. Hence we have

$$U_0^{(\mathbb{P}_k,m)} = \frac{\lambda (1-\alpha) \sqrt{p_k}}{2} \Theta_g^{(\mathbb{P}_k,m)} + \frac{\lambda\alpha \Theta_v^{(\mathbb{P}_k,m)}}{2}$$
$$\text{with } \left\| \Theta_g^{(\mathbb{P}_k,\cdot)} \right\|_2^2 \leq 1 \tag{13}$$
$$\text{and with } \left\| \Theta_v^{(i,\cdot)} \right\|_2^2 \leq 1.$$

and for $i \in \mathbb{P}_k$ we have also

$$U_0^{(i,\cdot)} - \frac{\lambda\alpha \Theta_v^{(i,\cdot)}}{2} = \frac{\lambda (1-\alpha) \sqrt{p_k}}{2} \Theta_g^{(i,\cdot)}$$
$$\text{with } \left\| \Theta_g^{(i,\cdot)} \right\|_2^2 \leq 1$$
$$\text{and with } \left\| \Theta_v^{(i,\cdot)} \right\|_2^2 \leq 1.$$

Let's define

$$U_1^{(i,\cdot)} = \left(1 - \frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+ U_0^{(i,\cdot)}. \tag{14}$$

We can establish in the following lemma, which makes emerge the variable thresholding term of sPLS like in (1) in Equation (12).

Lemma 1

$$\left\|U_1^{(i,\cdot)}\right\|_2 \leq \left\|U_0^{(i,\cdot)} - \frac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2}\right\|_2 \tag{15}$$

and there is a Θ_v such that

$$U_1^{(i,\cdot)} = U_0^{(i,\cdot)} - \frac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2}. \tag{16}$$

Proof. if $\left\|U_0^{(i,\cdot)}\right\|_2 \leq \left\|\frac{\lambda\alpha}{2}\right\|_2$ then $\left(1 - \frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+ = 0$ and then $U_0^{(i,\cdot)} = 0$. The inequality is then true. Furthermore, there is a $\Theta_v^{(i,\cdot)} = U_0^{(i,\cdot)}$ reach the equality (16). Otherwise, $U_1^{(i,\cdot)} \neq 0$ and

$$U_0^{(i,\cdot)} - \frac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2} = U_0^{(i,\cdot)} - \frac{\lambda\alpha U_0^{(i,\cdot)}}{2\left\|U_0^{(i,\cdot)}\right\|_2} = U_1^{(i,\cdot)}$$

The inequality (15) is true because the equality (16) is reached. In any case the lemma 1 is proved. □

From lemma 1 we can infer that in (12)

$$\left\|U_1^{(i,\cdot)}\right\|_2 \leq \left\|\frac{\lambda\left(1-\alpha\right)\sqrt{p_k}}{2}\Theta_g^{(i,\cdot)}\right\|_2$$

and the inequality can be reached as an equality.
We have

$$\sum_{i\in\mathbb{P}_k}\left\|\Theta_g^{(i,\cdot)}\right\|_2^2 = \left\|\Theta_g^{(\mathbb{P}_k,\cdot)}\right\|_2^2$$

and we have also

$$\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2 \leq \left\|\Theta_g^{(\mathbb{P}_k,\cdot)}\right\|_2^2$$

and the inequality can be reached.

Therefore

$$\left\|\Theta_g^{(\mathbb{P}_k,\cdot)}\right\|_2^2 \leq 1$$

stand if and only if

$$\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2 \leq 1.$$

In the end we have $U^{(i,\cdot)} = 0$ if

$$\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2 \leq 1.$$

Let's now consider that $U_{opt}^{(\mathbb{P}_k,\cdot)} \neq 0$ or $U_{opt}^{(i,\cdot)} \neq 0$ for at least one $i \in \mathbb{P}_k$ then

$$U_{opt}^{(i,\cdot)} = U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} - \frac{\lambda\,(1-\alpha)}{2}\frac{U_{opt}^{(i,\cdot)}}{\left\|U_{opt}^{(\mathbb{P}_k,\cdot)}\right\|_F}.$$

If $\left\|U_0^{(i,\cdot)}\right\|_2 \leq \frac{\lambda\alpha}{2}\|$ then we can set $\Theta_v^{(i,\cdot)}$ such that $U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} = 0$. Otherwise $U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} = U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\frac{U^{(i,\cdot)}}{\left\|U^{(i,\cdot)}\right\|_2}$. In both cases we can consider that $U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} = U_1^{(i,\cdot)}$.

From this point we find successively

$$U_{opt}^{(i,\cdot)} = U_1^{(i,\cdot)} - \frac{\lambda\,(1-\alpha)\,U_{opt}^{(i,\cdot)}}{2\left\|U_{opt}^{(\mathbb{P}_k,\cdot)}\right\|_F},$$

$$U_{opt}^{(i,\cdot)}\left(1 + \frac{\lambda\,(1-\alpha)}{2\left\|U_{opt}^{(i,\cdot)}\right\|_2}\right) = U_1^{(i,\cdot)}$$

and

$$U_{opt}^{(i,\cdot)}\left(1 + \frac{\lambda\,(1-\alpha)}{2\left\|U_{opt}^{(i,\cdot)}\right\|_2}\right) = U_1^{(i,\cdot)}.$$

Summing the square for every element of $\mathbb{P}_k$ we have

$\sum_{i\in\mathbb{P}_k}\left\|U_{opt}^{(i,\cdot)}\right\|_2^2\left(1+\frac{\lambda(1-\alpha)}{2\left\|U_{opt}^{(i,\cdot)}\right\|_2}\right)^2 = \sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2$.

and hence $\left\|U_{opt}^{(i,\cdot)}\right\|_F^2\left(1+\frac{\lambda(1-\alpha)}{2\left\|U_{opt}^{(i,\cdot)}\right\|_2}\right)^2 = \sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_F^2 = \left\|U_1^{(\mathbb{P}_k,\cdot)}\right\|_F^2$.

After extracting the value of $\left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F$ from this equation we finally find that

$$U_{opt}^{(\mathbb{P}_k,\cdot)} = U_1^{(\mathbb{P}_k,\cdot)}\left(1 - \frac{\lambda\,(1-\alpha)}{2\left\|U_1^{(\mathbb{P}_k,\cdot)}\right\|_F}\right).$$

□

Corollary 1 *The solution to Equation (4) can be seen as a biconvex optimization problem.*

Fixing V, the optimal U_{opt} for (4) is :

$$U_{opt}^{(i,\cdot)} = U_0^{(i,\cdot)}\left(1 - \frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+, \; U_0^{(\cdot,m)} = Z_m V^{(\cdot,m)} \tag{17}$$

$$\text{and } Z_m = X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)}$$

Fixing U, the optimal V_{opt} for (4) is :

$$V_{opt}^{(\cdot,m)} = Z_m^T U^{(\cdot,m)} \tag{18}$$

Corollary 2 *The solution to Equation (5) can be seen as a biconvex optimization problem is :*

Fixing v, the optimal u_{opt} for (5) is

$$u^{(\mathbb{P}_k)} = u_1^{(\mathbb{P}_k)} \left(1 - \frac{\lambda(1-\alpha)}{2\sqrt{\sum_{i\in\mathbb{P}_k} \left\|u_1^{(i)}\right\|_2^2}}\right)_+ = u_1^{(\mathbb{P}_k)} \left(1 - \frac{\lambda(1-\alpha)}{2\left\|u_1^{(\mathbb{P}_k)}\right\|_F}\right)_+$$
$$With\ u_1^{(i)} = u_0^{(i)} \left(1 - \frac{\lambda\alpha}{2\left\|u_0^{(i)}\right\|_2}\right)_+,\ u_0 = Zv \tag{19}$$
$$and\ Z = XY^T.$$

Fixing u, the optimal v_{opt} for (5) is

$$v_{opt} = Z^T u. \tag{20}$$

5 Discussion

Problems 1 to 4 are discussed in this part, but we consider that the following remarks can be transposed to Problems 5 and 6, Problem 5 and 6 being resp. the equivalents of Problems 2 and 4 for data with a variable groups structure.

5.1 Size of the data

The larger data (in term of number of observations) are, the better models are supposed to perform. We can see that Problem 1 and 2 have the merit of performing an sPLS on data containing n observations whereas Problem 3 and 4 performs M different sPLS methods on data with resp. $\mathbb{M}_m$ observations for $m \in \{1, \ldots, M\}$. For some observation set the number of observations can be significantly smaller than the size of the hole data which can have a negative impact on the result.

5.2 Number of loading elements in the model

Number of loading elements is an important parameter to control. From one side, the bigger the number is, the more information can be stored by the model, from the other side having too much loading elements can give results harder to interpret and there is higher risk of over-fitting. Problems 1 and 2 have only p loadings for u whereas Problems 3 and 4 have $M \times p$ ones. For Problem 4 the number of loadings is important but the number of non-zero variables will vary between 1 and p in the same way than in Problem 1 and Problem 2. The Problem 4 gives readable results while keeping the flexiblity of a model with higher number of loading elements. For Problem 3 the non-zero variables can be different from one study to another, we cannot control weather a variable will be null for all studies and the number of non-zero variable will be significantly higher than for other problems.

5.3 Sensibility to batch effect

Batch effect can arise when data provided from different source present a bias. This effect can happen when observation sets are expected to introduce its intrinsic error. Therefore the cross-product matrices could be represented by a model like :

$$Z_m = Z + E_m \text{ for } m \in \{1, \dots, M\}$$

Where Z follows a given law and E_m are Gaussian noise with parameters depending on m. Under this hypothesis the standardization within studies can bypass this bias. Therefore Problem 3 to 4 can correct this kind of batch effect.

However, more complex bias can exist. For instance what happens if different observation sets have different dynamics? Let us consider a variable that is positively correlated in some observation sets and negatively correlated in others. In Problems 1 and 2 and their overall sPLS, the variable will have a small corresponding loading because positive and negative effect compensate each other and the variable will be cut because of the sparsity heuristic. In Problem 3, the distinction between all dynamics will be highlighted by the model. Finally, Problem 4 will select the same variable because it has a significant loading on every observation set. In the end, Problem 4 can handle more cases where the observation sets introduce bias.

5.4 Relation between Problem 4 and classic sgPLS

We can establish also that Problem 4 (sPLS method with a common-Lasso penalization) applied to matrices X and Y of size resp. $(n, p)$and (n, q) can be equivalent to a classical sgPLS without a standard Lasso on well chosen matrices $\tilde{X}$ and $\tilde{Y}$ of size resp. $(n, p \times M)$ and $(n, q \times M)$. Those matrices are constructed by shifting the row blocks of X and Y : they are diagonal bloc matrices whose blocs are resp. $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ for $m \in \{1, \dots, M)\}$. The corresponding loading vectors of size resp. $p \times M$and $q \times M$ are called here resp. u_e and v_e. The representation of those objects is shown in Figure 3.
On those basis the formulation of the sgPLS problem searching for optimal $u_{e,opt}$ and $u_{e,opt}$ would be :

$$\begin{aligned}\{u_{e,opt}, v_{e,opt}\} = \underset{u_e, v_e}{\operatorname{argmin}} \left\| \tilde{Z} - u_e {v_e}^T \right\|_F^2 + \lambda (1-\alpha) P_{group}(u_e) + \lambda\alpha P_{variable}(u_e) \\ \text{with } P_{group}(u_e) = \sum_{k=1}^{K} \sqrt{p_k} \left\| {u_e}^{(\mathbb{P}_k)} \right\|_2 , \\ P_{variable}(u_e) = \sum_{i=1}^{p} \left\| {u_e}^{(i)} \right\|_2 \text{ and } \tilde{Z} = \tilde{X}^T \tilde{Y}\end{aligned} \tag{21}$$

In this formulation loading vectors u_e and v_e can be seen as the concatenation of the rows of resp. U and V in a unique uni-dimensional vector. This notation is interesting from a theoretical point of view because it ensure that Problem 4 can inherit properties from sPLS. However this notation is not wise for computational efficiency because the matrices $\tilde{X}$ and $\tilde{Y}$ are M times bigger than X and Y, where M is the number of observation set. For implementation, computing directly the solution from equations (10) and (11) seems wiser.

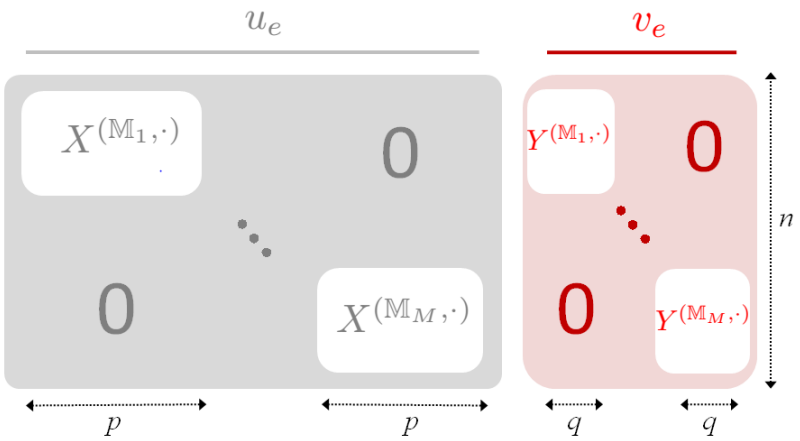


Fig. 3 Notation of $\tilde{X}$ (grey rectangle) and $\tilde{Y}$ (red rectangle) to write the sPLS for structured data as a sgPLS.

6 Application on simulated data

Presented methods are illustrated on simulated data. A first simulation case presents data where a batch effect exists and a second simulation case presents data where different dynamics exist among different observation sets. For each case different noise levels are considered. Every simulation is performed 50 times. The code can be found at `https://github.com/camilobroc/sgPLS_for_structured_data`.

6.1 Design of the simulated data

In the following, a training data set of 900 observations gathered in 3 observation sets of 300 observations are generated and then a test data set of 300 observations gathered in 3 observation sets of 100 observations (for the training data $M = 3$ and $n_1 = n_2 = n_3 = 300$, for the test data $n_1 = n_2 = n_3 = 100$).

6.1.1 Batch effect cases

In first the simulation case, applications of the methods with data presenting a batch effect are performed. The simulation are performed with different noise levels. Data have an observation set structure and a group of variable structure as shown in figure 2. A batch effect implies that one same physical process is observed but the methods of measurement vary among the different group of observation. We represent this difference of measurements by a bias in depending on the observation set.

A matrix X with 1000 variables gathered in 50 groups of 20 variables ($K = 50$ and $p_1 = \cdots = p_K = 20$) and a matrix Y with 3 variables ($q = 3$) are generated. In order to mimic a batch effect, the generation procedure of the matrices resp. X and Y has different parameters depending on different sub-matrices of resp. X and Y. Those matrices are composed of a signal and a noise. The signal corresponds to a PLS model with one latent variable. For $m \in \{1, \ldots, M\}$:

$$\begin{aligned} X^{(\mathbb{M}_m,\cdot)} &= \underbrace{H^{(\mathbb{M}_m)}C^T}_{\text{Signal}}\underbrace{\lambda_m^{X,B}+\mu_m^{X,B}\mathbb{1}_{n_m\times p}}_{\text{Batch}}+\underbrace{E_X^{(\mathbb{M}_m,\cdot)}}_{\text{Noise}} \\ Y^{(\mathbb{M}_m,\cdot)} &= \underbrace{H^{(\mathbb{M}_m)}D^T}_{\text{Signal}}\underbrace{\lambda_m^{Y,B}+\mu_m^{Y,B}\mathbb{1}_{n_m\times q}}_{\text{Batch}}+\underbrace{E_Y^{(\mathbb{M}_m,\cdot)}}_{\text{Noise}} \end{aligned} \tag{22}$$

Signal

The latent variable H is a $(n \times 1)$ column vector where each element follows a normal distribution of mean 0 and standard deviation 1. The loadings associated to this latent variable are resp. C a $p \times 1$ column vector and D a $q \times 1$ column vector corresponding resp. to X and Y.

Batch effect

The signal is blurred by a batch effect. The parameters $\lambda_m^{(X,B)}$, $\mu_m^{(X,B)}$, $\mu_m^{(Y,B)}$ and $\lambda_m^{(Y,B)}$ are real numbers depending on the observation set m. They control the shape of the batch effect. The notations $\mathbb{1}_{n_m,p}$ and $\mathbb{1}_{n_m,q}$ correspond to the matrices which elements are equal to 1 and of respective size $n_m \times p$ and $n_m \times q$.

Noise

The noise is represented by E_X, a $(n \times p)$ matrix, and E_Y, a $(n \times q)$ matrix. The matrix E_X is constructed by group of variables : for $k \in \{1, \ldots, K\}$, the rows of $E_X^{(\cdot,\mathbb{P}_k)}$ follow a multivariate normal distribution $N_{p_k}(\mathbb{0}_{p_k}, \lambda^{X,E}\Sigma_{p_k,\rho})$ where ρ and $\lambda^{X,E}$ are real parameters and $\Sigma_{p_k,\rho}$ is a $(p_k \times p_k)$ matrix which diagonal elements are equal to $1-\rho$ and non-diagonal elements are equal to ρ. The notations $_{p_k}$ stands fro the vector of size p_k and which elements are all equal to 0. The rows of the matrix E_Y follow a multivariate normal distribution $N_q(\mathbb{1}_q \times 0, \lambda^{Y,E}\Sigma_{q,\rho})$ where $\lambda^{Y,E}$ is a real parameter and $\Sigma_{q,\rho}$ is a $(q \times q)$ matrix which diagonal elements are equal $1-\rho$ and non-diagonal elements are equal to ρ. The notations $_q$ to the vector of size q and which elements are all equal to 1. The parameter ρ represent a correlation between variables of a same group and $\lambda^{Y,E}$ and $\lambda^{Y,E}$ represent the noise levels.

The non null parameters of C are the 15 first variables of the first 4 group of variables. Among those elements resp. 15, 30 and 15 are equal to resp. 1, -1, 1.5 and the values are randomly distributed. Other parameters are given in table 1 and the noise levels are indicated in table 2.

6.1.2 Effects of different magnitudes among group of observations

This simulation case mimic data presenting different dynamics among observation sets. The generation process follow the same formulas as the previous one but with different parameters. The main difference with the previous cases is that the parameters $\lambda_m^{X,B}$ for $m \in \{1, \ldots, M\}$ can have opposite signs. While in the first case a batch could be represented by a difference of magnitude, the effects can have here opposite directions. In this simulation case we are not interested in a bias concerning $\mu_m^{(X,B)}$ or $\mu_m^{(Y,B)}$, the parameters are set to zero. The non null parameters of C are the 15 first variables of the first 4 group of variables. Among those elements resp. 15, 30 and 15 are equal to resp. 1, -1, 1.5 and the values are randomly distributed. Other parameters are given in table 1 and the noise levels are indicated in table 2.

Table 1 Table of the parameters used in first and second simulation cases.

	First simulation case	Second simulation case
ρ	0.05	0.05
$\mu_1^{X,B}$	2	0
$\mu_2^{X,B}$	-1	0
$\mu_3^{X,B}$	-1	0
$\lambda_1^{X,B}$	1	1
$\lambda_2^{X,B}$	0.8	-0.8
$\lambda_3^{X,B}$	1.5	1.05
$\mu_1^{Y,B}$	2	0
$\mu_2^{Y,B}$	0	0
$\mu_3^{Y,B}$	-2	0
$\lambda_1^{Y,B}$	0.6	0.6
$\lambda_2^{Y,B}$	1.4	1.4
$\lambda_3^{Y,B}$	1	1
D	$\{1, -1, 1.5\}$	$\{1, -1, 1.5\}$

6.2 Compared methods

In the first simulation case , methods corresponding to problems 1, 2 and 5 are compared whereas in the second simulation case, methods corresponding to problems 1,2, 4 and 6 are compared. For the methods corresponding to problems 1, 2 and 4 the penalization parameter λ is set such that the number of variables is equal to the true number of variables having an effect (in this case 60). For the method corresponding to problems 5 and 6, the penalization parameter λ is set such that the number of groups of variable is equal to the true one (in this case 4) while α is chosen from the set $\{0.1, \ldots, 0.9\}$ by cross validation : the value giving the best Mean Square Error Prediction is kept.

6.3 results

The performances of the method are measured through the True Positive Rate (TPR), the Total Discordance (TD) and the Mean Square Error Prediction. The TPR is defined as :

$$TPR = \frac{\text{True Positives}}{\text{True Positive} + \text{False Negatives}}$$

and TN is defined as :

$$TD = \text{False Postives} + \text{False Negatives}$$

Results of first and second simulation are given in table 2.

In the first simulation case, data present a bias that depends on the observation set. Different noise levels are generated ($\lambda^{X,E} = 2, 20, 30$). We can see that when noise is small ($\lambda^{X,E} = 2$), MINT and "sparse group PLS for structured data" can retrieve the true variables whereas a classic sPLS cannot. We can also see that the MSEP is better for the "sparse group PLS for structured data" than for MINT. We can note that "sparse group PLS for structured data" misses a few true variables which gives a non null TD. This is due to the fact that the calibration of the method doesn't seek for a selection of the true number of variables, hence, a small number of true variables can be missed. When noise is greater ($\lambda^{X,E} = 20$), a difference in terms of detection of true variables is observed. The classical PLS have a much worse TPR and TD while "sparse group PLS

Table 2 Results for first and second simulation case. Results in terms of MSEP, TPR and TD are presented for each noise level.

Simulation case 1									
Noise level	$\lambda_1^{X,E} = \lambda_2^{X,E} = 2$			$\lambda_1^{X,E} = \lambda_2^{X,E} = 20$			$\lambda_1^{X,E} = \lambda_2^{X,E} = 30$		
	MSEP	TPR	TD	MSEP	TPR	TD	MSEP	TPR	TD
sPLS	3.99	0.5	60	22.85	0.27	87.76	33.35	0.18	98.92
MINT	2.20	1.0	0.0	20.97	0.76	28.88	31.59	0.54	55.00
sgPLS	2.20	1.0	7.4	20.83	0.99	15.44	31.16	0.98	17.88
Simulation case 2									
Noise level	$\lambda_1^{X,E} = \lambda_2^{X,E} = 2$			$\lambda_1^{X,E} = \lambda_2^{X,E} = 10$			$\lambda_1^{X,E} = \lambda_2^{X,E} = 20$		
	MSEP	TPR	TD	MSEP	TPR	TD	MSEP	TPR	TD
sPLS	3.58	0.61	46.84	12.38	0.15	101.68	23.19	0.08	110.08
MINT	3.67	0.88	14.12	12.40	0.17	99.84	23.16	0.08	110.12
sPLS for structured data	2.85	1.00	0.00	11.11	0.79	24.84	21.76	0.43	67.96
sgPLS for structured data	2.85	1.00	5.56	11.06	0.98	13.60	21.39	0.93	20.28

for structured data" is above MINT. When noise is even greater ($\lambda^{X,E} = 30$), "sparse group PLS for structured data" clearly outperforms MINT.

In the second simulation case, data present a magnitude in the latent variables that depends on the observation set. Different noise levels are generated ($\lambda^{X,E} = 2, 10, 20$). We can see that when noise is small ($\lambda^{X,E} = 2$) we can see that in order to retrieve the true variables "s(g)PLS for structured data" is better than MINT which is better than sPLS. In the same way that in the first simulation case, "sgPLS for structured data" miss a few number of the true variables whereas "sPLS for structured data" do not because of the specificity of the calibration. We can see at noise level $\lambda^{X,E} = 10$, that only the methods calibrated "for structure data" are able to retrieve the true variables. At the highest noise level ($\lambda^{X,E} = 20$) we can see that sgPLS stands clearly above "sPLS for structured data", and those two methods outperform the existing ones.

7 Conclusion

In the end different ways of formulating an sPLS problem on data presenting an observation set structure have been discussed. The MINT formulation have the merit of being easy to implement and correct the batch effect. The novel method "sparse PLS for structured data" can also correct it. Further more it allows to take into account a lot of different bias, especially when the different observation set don't have the same dynamics. Despite its high number of parameters, the common-Lasso penalization ensures that the result is readable with a small number of selected variables in the overall analysis.

This article proved it's ability to inherit properties of sPLS. It's adaptation to variable groups developed in this article, called "sparse group PLS for structured data", is a notable example of " sparse PLS for structured data" benefiting from an extension of the sPLS. We can note also that it can be applied on either quantitative or qualitative variables as any sPLS can.

A simulation shows that the new methods can outperform existing methods for detecting a small signal in a large noise. Because its requirements on the nature of the data are very general we are confident that the method can be applied to the wide area of domains where sPLS is competitive.

References

1. Saint John Walker. Big data: A revolution that will transform how we live, work, and think. *International Journal of Advertising*, 33(1):181–183, 2014.

2. Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.
3. Annalise B Paaby and Matthew V Rockman. The many faces of pleiotropy. *Trends in Genetics*, 29(2):66–73, 2013.
4. Herman Wold. Path models with latent variables: The nipals approach. In *Quantitative sociology*, pages 307–357. Elsevier, 1975.
5. Jose A Seoane, Colin Campbell, Ian NM Day, Juan P Casas, and Tom R Gaunt. Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS computational biology*, 10(10):e1003876, 2014.
6. Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.
7. Anne-Laure Boulesteix and Korbinian Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44, 2006.
8. Tijl De Bie, Nello Cristianini, and Roman Rosipal. Eigenproblems in pattern recognition. In *Handbook of Geometric Computing*, pages 129–167. Springer, 2005.
9. Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
10. Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
11. Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, 2014.
12. Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.
13. Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.
14. Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.
15. Aida Eslami, El Mostafa Qannari, Achim Kohler, and Stéphanie Bougeard. Algorithms for multi-group pls. *Journal of Chemometrics*, 28(3):192–201, 2014.
16. Florian Rohart, Aida Eslami, Nicholas Matigian, Stéphanie Bougeard, and Kim-Anh Le Cao. Mint: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC bioinformatics*, 18(1):128, 2017.
17. Benoit Liquet, K Mengersen, AN Pettitt, M Sutton, et al. Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis*, 12(4):1039–1067, 2017.
18. Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
19. Benoît Liquet, Pierre Lafaye de Micheaux, Boris P Hejblum, and Rodolphe Thiébaut. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1):35–42, 2015.
20. Matthew Sutton, Rodolphe Thiébaut, and Benoit Liquet. Sparse partial least squares with group and subgroup structure. *Statistics in Medicine (in press)*, 2018.
21. Tao Wang, Gloria Ho, Kenny Ye, Howard Strickler, and Robert C Elston. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genetic epidemiology*, 33(1):6–15, 2009.
22. Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.
23. Vincenzo Esposito Vinzi, Laura Trinchera, and Silvano Amato. Pls path modeling from foundations to recent developments and open issues for model assessment and improvement. In *Handbook of partial least squares*, pages 47–82. Springer, 2010.
24. Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

Chapter 5

Meta-analysis methods applied to pleiotropy

This chapter tackles the problem of pleiotropy in genetic epidemiology. At first, an overview of the definitions covered by the name pleiotropy are presented. Especially, its formulation as a meta-analysis method is explained. Then the application of the method introduced in Section 4.4 to pleiotropy studies is presented. The method was applied to a thyroid and breast cancer study and results have been submitted to BMC bioinformatics journal. The current version of this paper (draft) is included after this chapter.

5.1 Pleiotropy defintions

Pleiotropy defines cases in genetic epidemiology with complex relations between genome and phenotype, especially when one or more genetic traits have an effect on several phenotype traits. It occurs in genetics when a genetic factor influences different phenotypes. The definition is quite large and encompasses different statistical approaches [15].

• It can be seen as a multivariate analysis where one genetic trait have an influence on several phenotype traits.

• If phenotypes come from different sources, it can lead to a meta-analysis problem where several data sets have common genetic predictors but different phenotypes. The goal is to find genetic markers related to several phenotype across different studies.

Pleiotropy enriches our insights into the mechanisms involved in the appearance of the phenotypes. Gene set analysis with pleiotropy is expected to improve the chances of revealing the underlying genetic architecture of complex phenotypes. Highlighting pleiotropy provides opportunities for understanding the shared genetic underpinnings among associated diseases.

The scope of application in genetic epidemiology is wide as it can cover a large number of different traits at the same time. For instance, 42 traits or diseases have been studied in a single analysis approach recently [116]. In this case, the goal is to search new relations between diseases. On the opposite side, other pleiotropy analysis can take into account a fewer number of disease that are already expected to have common biological mechanisms. For instance, the large-scale Collaborative Oncological Gene-environment Study (COGS)

studies breast, ovarian and prostate cancers and several cancer showed peliotropy effects [117]. Common mechanisms in psychotic disorder disease has also been investigated [118].

In this dissertation, we are interested in cases with phenotypes coming from different sources. The motivation comes from oncology where the onset of a disease is rare and finding individuals presenting different diseases (type of cancers) can be very hard. Hence, different data sets with different phenotypes have te be used.

For the rest of the chapter we are interested in pleiotropy tackled as a problem of meta-analysis where the goal is the identification of predictors common to two or more data sets.

5.2 Challenges specific to pleiotropy

Meta-analysis is often necessary

Genetic epidemiology aims at studying populations presenting given disease, which reduces the number of individuals available. For pleiotropy, the population of interest must present several illness traits which constraints even more the population sample and, thus, finding patients presenting targeted symptoms of disease may be impossible. In this case, an alternate is to use patients coming from different studies and each patient having only one disease.

An example of such issues exist in oncology which is the field studying cancer. In oncology study, patients rarely present several cancer. In general, data related to different cancer are gather in distinct data bases that are usual analysed separately. To overcome this limitation, a meta-analysis is needed. This is the main motivation for studying pleiotropy as a meta-analysis in this dissertation.

Opposite effects

Sometimes a variable can have a positive effect for one phenotype and a negative one for another. In this case, some methods can have problems identifying an effect because a mean procedure will compensate the positive and the negative contribution of the model. This effect can be problematic especially in pleiotropy. Indeed, antagnostic pleiotropy has been observed in some studies, and hypothesis regarding their origins from an evolutionary point of view has been debated [119]. Hence, statistical methods need to cope with this aspect.

Interest of sparse group PLS for structured data

Regarding pleiotropy, many meta-analysis statistical tools can be used. So far, this method was used to investigate candidate genes across different cancer types and allowed to identify novel regions with pleiotropic effects [120]. Although a number of statistical approaches have been proposed in the literature for identifying the existence of pleiotropy at SNP-level, few work covers pleiotropy at gene- and pathway-levels. For instance, ASSET does not take into account a group structure of variables while SKAT only considers homogeneous effect across studies which does not take into account effect of opposite directions. The method proposed aims at filling this gap.

5.2.1 Application to thyroid and breast cancer

In this contribution, the method presented in the previous chapter/paper is applied to the study of thyroid and breast cancer data. The results of the study are included in a paper that has been submitted to BMC Bioinfomatics.

The developed statistical approaches will be applied for enriching our insights about the genetic mechanisms of thyroid and breast cancer types. We are interested in exploring gene level as well as pathway level associations between the individual cancer types as well as at both cancer types together, as breast and thyroid cancers share common hormonal risk factors [121]. Both cancers are hormonally mediated and occur more frequently in women than in men and have been found to be influenced by reproductive and hormonal factors. Breast cancer is the most common form of cancer in women and thyroid cancer is a less common cancer and was associated with early age at menarche, late menopause, late age at first full-term pregnancy, or low parity. Hormonal treatments, particularly those combining estrogens and synthetic progestagens can be a cause. Obesity is also a recognized risk factors for breast cancer in post-menopausal women [122]. On the other hand, thyroid cancer risk has been associated with late age at menarche and high parity. Conversely, breastfeeding and oral contraceptive use were associated with a lower risk of thyroid cancer [99, 123]. Anthropometric factors such as tall height and large body size have been also consistently associated with risk of thyroid cancer [124]. Regarding breast cancer risk, established genetic variants currently explain ∼49% of heredity risk, with the 90 variants identified in GWAS accounting for about 14% [124]. Thyroid cancer has been shown to be the only cancer for which the contribution of inherited genetic factors exceeds that of environmental factors [125]. However, recent GWAS have identified a few susceptibility loci explaining about 4% of familial heredity [126].

Presentation of the data

The data used for this contribution and its expected structure is the one depicted in Chapter 1 and 2, that is, n observations are available with p variables for the genotype (SNP data) and q variables for the phenotype (traits), and the number of observations, n is inferior to p (see Figure 2.3). Data has been provided by the INSERM institute. SNP variables can be organized in groups corresponding either to genes or pathways. We suppose that one SNP is in exactly one gene and one pathway. Observations can also be gathered by groups corresponding to different studies.

We use Beluhca data set (see Table 5.1) which includes data from CECILE, a french case-control study on breast cancer (1 125 cases, 1 172 controls) and from CATHY and Young-thyr, two french case-control studies on thyroid cancer (463 female cases and 482 female controls). All these individuals were genotyped using a customized microarray including 8716 genetic variants from 28 candidate pathways (648 genes) selected from KEGG database and from a litterature review. After quality controls, we retained 6 677 SNPs available for both type of cancers.

We went through the following test to pre-process the datasets:

- missing values were imputed using median among case/control and data centered to $\mu = 0$. Then, we studied the correlation between SNPs to remove those with a coefficient equal to 1. This step is necessary because some methods don't allow correlated data.

Cancer	SNPs	Observations	Missing values	Pairs of SNPs removed
Breast	6 677	2 297	34 869	42
Thyroid	6 677	945	13 215	20

Table 5.1 – Data from Beluhca data set after pre-processing.

- Couple of extremely correlated SNPs belonging to same genes were eliminated.
- As some methods do not deal with overlapping groups, 10 non-overlapping pathways were selected and only the 3766 SNPs related to those groups were kept in the final database. After all this steps, the new dataset is composed of 3766 SNPs, grouped in 337 genes and 10 non-overlapping pathways.

After, this pre-processing, data analyzed in this contribution are composed of 3766 SNP data corresponding to 337 genes that participate in 10 pathways.

5.3 Conclusion

As we have seen, pleiotropy is deals with meta-analysis in genetics and raises statistical challenges. From our knowledge, statistic methods for meta-analysis tackling both the opposite effect possibility and the group of variable structure hasn't been quite investigated. The proposed method joint-sgPLS aims at filling this gap. It leads to new perspectives for modeling pleiotropy in genetic epidemiology.

Penalized Partial Least Square for pleiotropy

Camilo Broc, Thérèse Truong and Benoit Liquet

Submitted to BMC Bioinformatics (2020)

RESEARCH

Penalized Partial Least Squares for pleiotropy

Camilo Broc[1,2*]
, Therese Truong[3,4] and Benoit Liquet[2,5]

*Correspondence:
camilo.broc@univ-pau.fr
[1]Laboratory for Data Sciences and Decision (Digiteo), LIST, CEA, , Gif-sur-Yvette, France
[2]Laboratoire De Mathématiques et de leurs Applications de PAU E2S UPPA, CNRS, , Pau, France
Full list of author information is available at the end of the article

Abstract

Background
The increasing number of genome-wide association studies (GWAS) have revealed several loci that are associated to multiple distinct phenotypes, suggesting the wide existence of pleiotropic effects. Highlighting these cross-phenotype genetic associations could help to identify and understand common biological mechanisms underlying some diseases. Common approaches test the association between genetic variants and multiple traits at the SNP-level. In this paper, we propose a novel gene- and a pathway-level approach in the case where several independent GWAS on independent traits are available. The method is based on a generalization of the sparse group Partial Least Squares (sgPLS) to take into account variable groups, and a Lasso penalization that links all independent data sets. This method, called joint-sgPLS, is able to convincingly detect signal at the variable level and at the group level.
Results
Our method has the merit of proposing a global readable model while coping with the structure of data. It can outperform traditional methods and provides a wider insight in terms of a priori information. Results are also provided on synthetic data and on an application to real data. Genetic markers common to breast cancer and thyroid cancer are investigated in a real data application.
Conclusion
The joint-sgPLS shows interesting properties for detecting a signal. As an extension of the PLS, the method is suited for data with a large number of variables. The choice of Lasso penalization copes with structures of variables groups and observations sets. Furthermore, although the method has been applied to a genetic study, its formulation is adapted to any data with high number of variables and an exposed a priori structure in other application fields.

Keywords: Genetic epidemiology; High dimensional data; Lasso Penalization; Meta-analysis; Oncology; Partial Least Square; Pathway analysis; Pleiotropy; Sparse methods; Variable selection

Background

Genome-wide association studies (GWAS) have identified numerous genetic markers linked to multiple phenotypes, suggesting the existence of pleiotropy that occurs when a single variant or gene can influence several phenotype traits [1, 2, 3, 4]. Highlighting pleiotropy provides opportunities for understanding the shared genetic underpinnings among associated diseases. As the genetic effects for most complex traits are small, several methods were proposed to combine results across studies of different phenotypes in order to improve the power of detecting cross-phenotype or pleiotropic associations.

We propose to focus on the case where several GWAS on different traits are available from different independent sources. One way to analyze pleiotropy is to analyze each trait separately in each independent datasets. In order to gain statistical power to detect pleiotropy, we propose a method that can combine the different traits into a meta-analysis taking into account that effect in opposite directions may exist across the different traits. We also propose a gene-based approach that combines the association signals from the single nucleotide polymorphisms (SNP) into a signal at the gene level or at the pathway-level. Genes will be defined by their coordinates and pathways are group of genes involved in a common biological mechanism. Independent data sets give an observation set structure to the data while the gene and the pathways give a group of variables structure. The challenge of pleiotropy is then to take advantage of this structure. In addition, possible biases between observation sets can be introduced, which is a common problem for meta-analyses, and methods for pleiotropy must take it into account [5]. Furthermore, such methods must cope with the case where a genetic variable have a positive effect on one disease and a negative effect on an other disease. Those opposite effects, may be hard to highlight statistically.

In this article, an extension of the Partial Least Square (PLS) method suited for meta-analysis for pleiotropy is proposed. It deals with observation sets and group of variables information while taking into account the possibility of opposite effects.

PLS is a dimension reduction developed by Wold [6] and that have been widely used for the analysis of data with large number of variables [7]. Unlike, its cousin method, the Principal Component Analysis (PCA) [8], the PLS deals with two blocks of data and is then widely used for genotype-phenotype analyses. Moreover its sparse extension using Lasso penalization have been successful for providing readable models[9]. Especially sparse group Partial Least Square can take into account group of variables as a priori information [10] [11]. For different group of studies an alternative Lasso penalization has been proposed by Obozinski [12] for a linear regression to deal with data made of different sets of observations. An adaptation of the Lasso penalization, the joint-sgPLS, has recently been proposed for the PLS [13], answering the specific structure of both groups of variables and sets of observations. In this article, we exploit the same idea to leverage pleiotropy effects, especially because the method copes with the challenge of detecting small possible opposite effects. The method is compared to two well established statistical methods in genetic studies. The first one, ASSET [14] extends standard fixed-effects meta-analysis to allow for potential opposite directions of the same SNP on different traits. However the method does not take into account the group level information such as pathways. The second one metaSKAT [15] permits to carry out gene-based meta-analysis, but all effects are in the same direction.

The developed statistical approaches will be applied to real dataset for enriching our insights about the genetic mechanisms of thyroid and breast cancer types. We are interested into exploring gene level as well as pathway level associations between the individual cancer types as well as at both cancer types together.

Notations

Data are represented by $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$, two matrices, representing n observations of p predictors and q independent variables. The Frobenius norm

on matrices is denoted $\| \ \|_F$. We note X^T the transpose matrix of X and the cardinal of a set S is noted $\#S$. The positive value of a real number x is noted $(x)_+ = \frac{|x|+x}{2}$ and is equal to the number if the number is positive and equal to zero otherwise. In general, observation sets can represent the fact that different sets of observations come from different sources and must be analyzed accordingly. For instance, data coming from different studies may present biases. Variables groups can represent either a set of variables that is known to be quite correlated or a group of variables that must be treated together. For instance, in genetics a gene defines an established group of SNP variables and pathway define established group of genes. From one side, let us consider M different sets of observations in the data. Noting, for $m \in \mathbb{N}$, $\mathbb{M}_m$ a subset of $\{1, \ldots, n\}$, let $\mathbb{M} = (\mathbb{M}_m)_{m=1..M}$ be a partition of $\{1, ..., n\}$ corresponding to the observation sets. We note $\#\mathbb{M}_m = n_m$. Row blocks are defined by this partition. From the other side, let us consider that variables are gathered in K groups. Let $\mathbb{P} = (\mathbb{P}_k)_{k=1..K}$ be a partition of $\{1, ..., p\}$ corresponding to this variable group structure. We note $\#\mathbb{P}_k = p_k$. We then we have $\sum_{k=1}^{K} p_k = p$. Column blocks are defined by this partitions. Both observation set structure and variable group structure can be defined at the same time as shown in Figure 1. For matrices, the notation $\cdot$ is used to refer to blocks of matrices. For instance $X_{\cdot,\mathbb{P}_k}$ is the block of matrix of X corresponding the columns of the k-th group of variables and $X_{\mathbb{M}_m,\cdot}$ is the block of matrix of X corresponding the columns of the m-th set of observations.

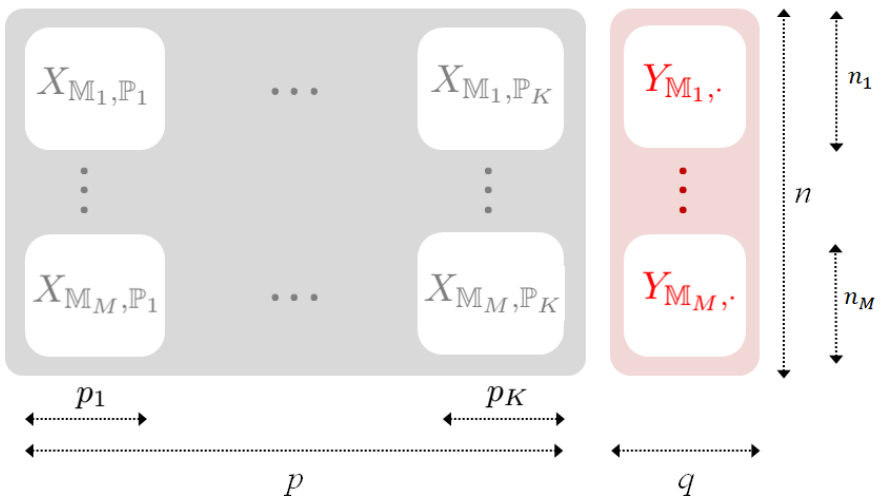


Figure 1: Illustration of data structured by groups of variables and groups observations. Observations and variables are assumed to be ordered by resp. observations sets and variable groups. The notation p represents the number of variables of matrix X, q the number of variables of matrix Y, n is the number of observations. $n_1, \cdots, n_M$ are the resp. number of observations of each observation set. $p_1, \cdots, p_K$ are the resp. number of variables in each group of variables.

Sparse Partial Least Square for structured data

In the literature, several formulations of the PLS exist [16]. In the scope of this article, only one of them, the called "PLS1" is considered [17].

PLS finds successively couples of vector $\{u_1, v_1\}, \ldots, \{u_r, v_r\}$ for $r < \min(p, q)$ where the couples are composed of vectors of length resp. p and q, maximizing $Cov(Xu_i, Yv_i)$ for any $i \in \{1, \ldots, r\}$, under the constraint that the family of vectors $u_1, \ldots, u_r$ and $v_1, \ldots, \ldots, v_r$ are both of them orthogonal families [6]. It can be solved considering successive maximization problems [18], for $h \in \{1, \ldots, r\}$

$$\max_{||u_h||_2=||v_h||_2=1} Cov(X^{(h-1)}u_h, Y^{(h-1)}v_h), \tag{1}$$

where $X_0 = X$, $Y_0 = Y$ and $X^{(h-1)}$, $Y^{(h-1)}$ are deflated matrices computed from $u^{(h-1)}$,$v^{(h-1)}$, $X^{(h-2)}$, $Y^{(h-2)}$ for $h \in \{2, \ldots, r\}$. The deflation depends on the PLS mode that is chosen ([6, 19]). In the following, the notation h is removed in order to simplify the formulation because we are interested in only one of the r steps of the PLS.

The sparse PLS (sPLS) propose to add a penalization to the loading vectors u and v. The following equivalence is used:

$$\underset{||u||_2=||v||_2=1, u\in\mathbb{R}^p, v\in\mathbb{R}^q}{\operatorname{argmax}} Cov(Xu, Yv) = \underset{||u||_2=||v||_2=1, u\in\mathbb{R}^p, v\in\mathbb{R}^q}{\operatorname{argmin}} \left\|X^TY - uv^T\right\|_F^2 \tag{2}$$

and the proof can be found in [10].

The sPLS [18] can be written as

$$\{u_{opt}, v_{opt}\} = \underset{||u||_2=||v||_2=1, u\in\mathbb{R}^p, v\in\mathbb{R}^q}{\operatorname{argmin}} \left\|X^TY - uv^T\right\|_F^2 + \underbrace{\lambda P(u)}_{\substack{\text{Lasso Penalty term} \\ \text{for sparse PLS}}}. \tag{3}$$

The sparse PLS introduces a penalization in this formulation of the problem. The penalty $P(\cdot)$ forces lowest values of u to be set to zero. The parameter controlling the degree of sparsity in the model is λ. In the presented formula the sparsity is applied only to the vector u, but a similar penalization can be define for v. In the context of this article we treat only the penalization of u but all the results stand also for a v penalization.

Remark : Before analysis, the X and Y matrices are transformed by subtracting their column averages. Scaling each column by their mean and standard deviation is also often recommended [20]. Thus, the cross-product matrix X^TY is proportional to the empirical covariances between X- and Y-variables when the columns of X and Y are centered. When the columns are standardized, X^TY is proportional to the empirical correlations between X- and Y-variables. In this article the standardization is an important step to overcome the issue of the "batch effect" or to aggregate observations from different studies. The subject has been discussed in [13].

Extensions of the sparse Partial Least Square

In the following, extension of the sPLS take into account an observation or/and variable set structure are presented. The last method has been recently developed [13] and deals with both kind of structures.

In order to cope with the group structure, sgPLS have been proposed [10]:

$$\begin{aligned}\{u_{opt}, v_{opt}\} = \underset{||u||_2=||v||_2=1, u\in\mathbb{R}^p, v\in\mathbb{R}^q}{\operatorname{argmin}} & \left\|Z - uv^T\right\|_F^2 + \lambda\left(1-\alpha\right)P_{group}(u) + \\ & \lambda\alpha P_{variable}(u) \\ \text{with } P_{group}(u) = \sum_{k=1}^{K}\sqrt{p_k}\left\|u_{\mathbb{P}_k}\right\|_2 \ , \ & P_{variable}(u) = \sum_{i=1}^{p}\left\|u_i\right\|_2 \\ & \text{and } Z = X^TY.\end{aligned} \tag{4}$$

where the loading vectors u and v is composed of resp. p and q elements. The penalization $P_{variable}$ forces single variables to be set to zero whereas the penalization P_{group} forces sets of variables to be set to zero. The degree of sparsity in general in the model is λ whereas the parameter controlling the balance between both kind of sparsity is α. In this model elements of u corresponding to least relevant variables and least relevant group of variables are set to zero.

An extension using the joint Lasso penalization from Obozinski [12] has been proposed. Its formulation for the sgPLS is:

$$\begin{aligned}\{U_{opt}, V_{opt}\} = \underset{\substack{U\in\mathbb{R}^{p\times M} \text{and } V\in\mathbb{R}^{q\times M} \\ ||U_{\cdot,m}||_2=||V_{\cdot,m}||_2=1 \text{ for } m\in\{1,\cdots,M\}}}{\operatorname{argmin}} & \sum_{m=1}^{M}\left\|Z^{(m)} - U_{\cdot,m}{V_{\cdot,m}}^T\right\|_F^2 \\ & +\lambda\left(1-\alpha\right)P_{group}(U) + \lambda\alpha P_{variable}(U) \\ \text{with } P_{group}(U) = \sum_{k=1}^{K}\sqrt{p_k}\left\|U_{\mathbb{P}_k,\cdot}\right\|_F \ , \ & P_{variable}(U) = \sum_{i=1}^{p}\left\|U_{i,\cdot}\right\|_2 \\ & \text{and } Z^{(m)} = X^T_{\mathbb{M}_m,\cdot}Y_{\mathbb{M}_m,\cdot},\end{aligned} \tag{5}$$

where the set of loading U is composed of $p \times m$ elements (p elements per $U_{\cdot,m}$). The set of loading V is composed of $q \times m$ elements (q elements per $V_{\cdot,m}$). In this model elements of U corresponding to least relevant variables and least relevant group of variables are set to zero. In this model the same variables and variable groups corresponding to least significant variables are set to zero for all $U_{\cdot,m}$, $m \in \{1,\ldots,M\}$.

For all sparse methods, parameters driving the penalization (λ and α) must be chosen. In general, the choice is made through a K-fold cross validation, take an error in prediction as criteria. An example of the procedure can be found in the implementation of many extension of the sPLS [10] [21] [22].

Remark : The proposed joint penalization is biconvex, and thus multiple local minima may exist. In this work we implement an algorithm based on the alternating

convex search [23]. It is known to be an effective method for biconvex optimisation and recent work has shown alternating methods for minimisation of biconvex functions converge with high probability to estimates with similar statistical accuracy to a global optimum [24].

Benchmark methods

Both ASSET and metaSKAT are considered as benchmark methods.

ASSET is a method suited for meta-analysis providing a p-value across studies [14]. The input of the method are single variables summary statistics which are combined by the method. ASSET exhaustively explores subsets of studies for the presence of true association signals that are in either the same direction or possibly opposite directions. However ASSET does not exploit the group information such as genes or pathways. Further, the current version of ASSET provides pleiotropy result for each variant which should be corrected using a FDR correction in order ro control possible false positive pleiotropy effect.

SKAT is a method to detect association between rare variants in a region and a phenotype (continuous or binary). It is a supervised test for joint effects of multiple variants in a region on a phenotype. The metaSKAT method can do the same but aggregating several studies. This method outputs a p-value corresponding to a set of variables, for instance a gene or a pathway. The method is based on a weighted sum of SKAT statistics of the different studies [15].

Results

The code used for running the methods is available on github (https://github.com/camilobroc/BMC_joint_sgPLS).

Simulation

Presented methods are illustrated on simulated data presenting the structure in Figure 1. From one side, SNP genotypes are coded as minor allele counting $\{0, 1, 2\}$ and a certain correlation is expected within a group of SNP from the same linkage disequilibrium block. From the other side, phenotype data are binary and have a true effect from one or more genetic markers. In order to simulate the correlation between SNPs, for a group of variables $\mathbb{P}_k$, a multivariate normal distribution with n observations $\boldsymbol{x}_k^{(continuous)} \sim \mathcal{N}_{p_k}(\mu_k, \Sigma_k)$ is simulated where μ_k is a null vector of size p_k and Σ_k is a $p_k \times p_k$ matrix with 1 on the diagonal and ρ_k, coefficients controlling the correlation between SNPs within a group, outside of the diagonal. A simulation of this variable gives a matrix which represents simulated observations for group of variables k. Those blocks are concatenated in a $n \times p$ matrix, $X^{continuous}$ that represents the whole data.

In order to have $\{0, 1, 2\}$ genotype data, a discretization is performed. For a given variable $j \in \mathbb{P}_k$, we aim at simulating a SNP variable with a Minor Allele Frequency (MAF), which we note MAF_j. This MAF means that:

$$
\begin{aligned}
P(x_j = 0) &= (1 - \text{MAF}_j)^2 \\
P(x_j = 1) &= 2\text{MAF}_j(1 - \text{MAF}_j) \\
P(x_j = 2) &= \text{MAF}_j^2.
\end{aligned}
$$

To this aim, for a given MAF_j, quantiles $q_1^{(j)}$ and $q_2^{(j)}$ are chosen such as $P(x_j \leq q_1) = (1 - MAF_j)^2$ and $P(x_j \leq q_2) = (1 - MAF_j)^2 + 2MAF_j(1 - MAF_j)$

A discrete genotype, $X^{(discrete)}$, is computed such that

$$X_{i,j}^{(discrete)} = \begin{cases} 0 & \text{if } X_{i,j}^{(discrete)} \leq q_1^{(j)} \\ 1 & \text{if } X_{i,j}^{(discrete)} \leq q_2^{(j)} \\ 2 & \text{if } X_{i,j}^{(discrete)} > q_2^{(j)}, \end{cases}$$

where $i \in \{1, \cdots, n\}$ are simulated observations and $j \in \mathbb{P}_k$ is a variable of the k-th group of variables.

For each observation i, a binary phenotype y_i is simulated with a logit model

$$\text{logit}(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}) = \alpha + \sum_{j=1}^{p} X_{i,j}^{(discrete)} \beta_j,$$

where $\pi_i = P(y_i = 1|\text{data})$, β_j for $j \in \{1, \cdots, p\}$ is a regression parameter.

Then different simulations of the process can be performed successively in order to simulated several studies.

In this article, 8 combinations of parameters are considered. Data composed of two studies. The simulated genotype has 25 groups of 20 variables. The number of observations, n can be either 200 (cases 1,3,5,7) or 400 (cases 2,4,6,8). The intra-group correlation parameters ρ_k are equal to 0.5 and the MAF is equal to 0.3 for each variable. The first 5 groups have an effect in the model of the simulation. Either all the SNPs of the group have an effect (cases 1,2,5,6) or half of the SNPs have an effect (cases 3,4,7,8). For each group, half of the non-null regression parameters are positives (taken at random) while the other half is negative. In cases 1,2,5 and 6, the absolute value of those parameters is set to $exp(0.1)$ whereas in cases 3,4,7 and 8, the absolute value of those parameters is set to $exp(0.5)$. Effect are either in the same direction from one study to another (cases 1,2,3,4) or in opposite direction (cases 5,6,7,8).

For all methods 50 replications of the data are performed. For the implementation of the sgPLS and joint-sgPLS, penalisation parameters must be chosen similarly to [10]. The penalization parameter λ and α are optimized through a K-fold penalization procedure with an error of prediction as criteria. Choosing a parameter λ is equivalent to set a number of selected groups [10]. In this simulation the grid of number of selected groups $\{1, \cdots, 25\}$ is used and the grid for α is $\{0.1, 0.5, 0.9\}$. Figures 2 and 3 show the error of prediction performances through a cross-validation procedure of the sgPLS and joint-sgPLS in a simulation of case 1, for different levels for α and different levels of group selection. The observed mean and the variance of the error rate over 50 replication are presented. In the framework of the method the set of parameters corresponding to the lowest error of prediction rate is kept for the model.

For ASSET, sgPLS and joint-sgPLS, the variables selected by the models are compared to the variable having an effect on the true model. For metaSKAT, sgPLS

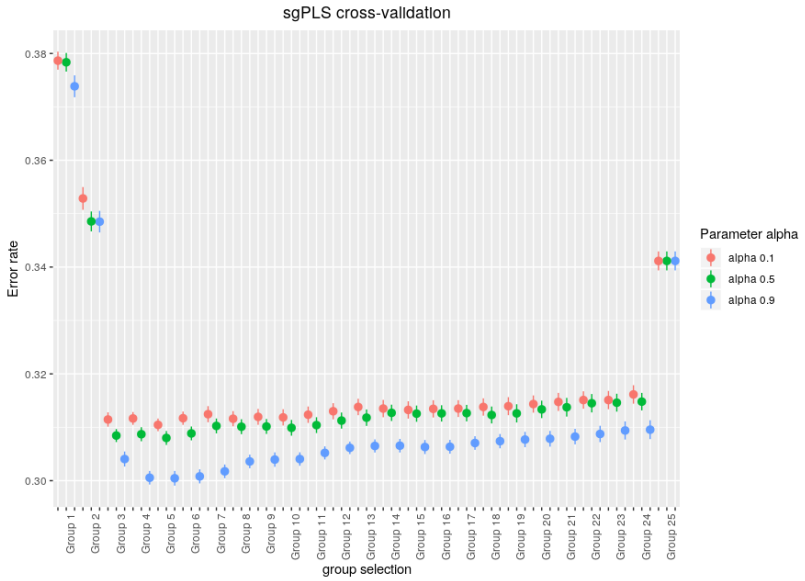


Figure 2: Mean and variance of the error of prediction in cross-validation of sgPLS, for one simulation of case 1 of the simulations. The cross-validation is performed for $\alpha \in \{0.1, 0.5, 0.9\}$ and for levels of group selection corresponding to $\{1, \cdots, 25\}$.

and joint-sgPLS, the group of variables selected by the models are compared to the group of variables having an effect on the true model.

Results of the simulations are presented in tables 1, 2, 3, 4 for sgPLS, joint-sgPLS, ASSET and metaSKAT. The measures of performance are the True Positives (TP), False positives (FP), False Negatives (FN) and True Negatives (TN).

n=200									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	46.78	29.24	53.22	370.76	sgPLS	3.54	3.82	1.46	16.18
joint-sgPLS	40.84	27.16	59.16	372.84	joint-sgPLS	3.44	3.78	1.56	16.22
ASSET	29.98	22.14	70.02	377.86	metaSKAT	2.22	1.08	2.78	18.92
n=400									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	75.76	139.34	24.24	260.66	sgPLS	4.74	11.9	0.26	8.1
joint-sgPLS	66.12	76.44	33.88	323.56	joint-sgPLS	4.48	7.62	0.52	12.38
ASSET	47.74	25.4	52.26	374.6	metaSKAT	3.14	1.26	1.86	18.74

Table 1: Perfomances in terms of mean number of TP, FP, FN and TN over 50 replications for methods sgPLS, joint-sgPLS, ASSET and metaSKAT in simulations cases 1 and 2 (when all SNPs of the 5 groups are involved in the true model and the effects are in the same directions from one study to another). The number of observations per study is n.

We can see that at the SNP level sgPLS and joint-sgPLS have a higher TP than ASSET and have a better TP than metaSKAT at the group level (tables 1, 2, 3 and 4). The number of true positives is high for sgPLS, joint-sgPLS and ASSET however. We can remark that when effects in the true model are in opposite directions the joint-sgPLS stands out as its TP is clearly higher than other methods (tables 3 and 4). We can also see that the number of FP for sgPLS and joint-sgPLS are inflated

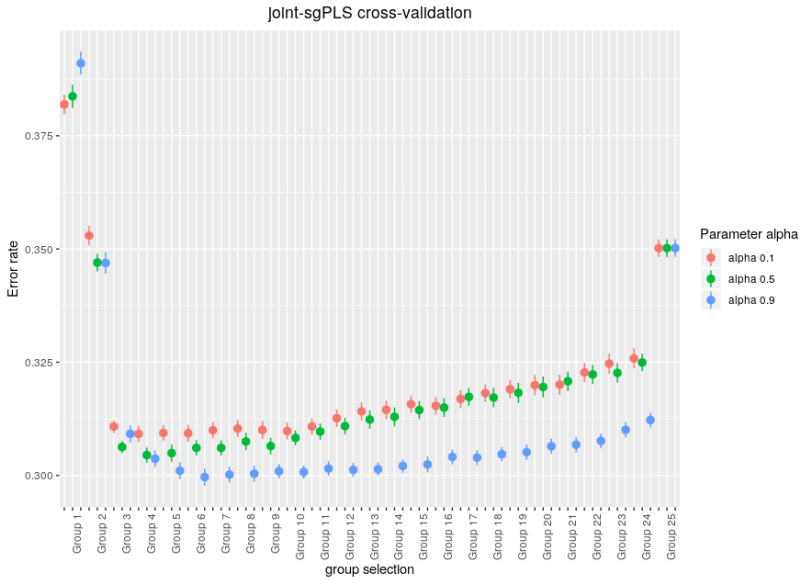


Figure 3: Mean and variance of the error of prediction in cross-validation of joint-sgPLS, for one simulation of case 1 of the simulations. The cross-validation is performed for $\alpha \in \{0.1, 0.5, 0.9\}$ and for levels of group selection corresponding to $\{1, \cdots, 25\}$.

when only 10 SNPs per true-groups are involved in the model of simulation (tables 2 and 4). This is probably due to the fact that the method selects too many variables with a selected group.

Overall, we can see that sgPLS and joint-sgPLS performs better for detecting effect in the same direction while joint-sgPLS is the method with the best performance for detecting opposite effects. MetaSKAT has a lower true positive rate than the two previous methods while having less total discordance which means it is more conservative on this simulation. A good compromise would be considering joint-sgPLS as both opposite and same direction effects are well selected.

Pleiotropy investigation on breast and thyroid cancer

The developed statistical approaches will be applied for enriching our insights about the genetic mechanisms of thyroid and breast cancer types. Thyroid and breast cancers share a lot of similarities in their biology: both are more frequent in women, are influenced by hormonal and reproductive factors and are hormonally-mediated. Moreover, individuals diagnosed with breast cancer are more likely to develop thyroid cancer as a secondary malignancy than patient diagnosed with other cancer types, and vice-versa [25]. Genetic factor contributing to the incidence of breast cancer have been extensively studied, and it is known that genetic variants explain approximately 49 percent of the familial risk to develop this disease. Using GWAS, 313 risk variants were identified for breast cancer [26]. On the other hand, GWAS studies on thyroid cancer have been scarce, due to the lesser incidence of this disease as well as the lack of data. However, we know that thyroid cancer is the only cancer for which genetic factors contribute more than environmental factors [27].

n=200									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	36.58	63.56	13.42	386.44	sgPLS	4.78	4.02	0.22	15.98
joint-sgPLS	31.3	48.04	18.7	401.96	joint-sgPLS	4.5	3.06	0.5	16.94
ASSET	29.46	46.88	20.54	403.12	metaSKAT	3.62	0.96	1.38	19.04
n=400									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	42.68	148.78	7.32	301.22	sgPLS	4.88	11.1	0.12	8.9
joint-sgPLS	40.42	115.96	9.58	334.04	joint-sgPLS	4.92	8.54	0.08	11.46
ASSET	35.26	54.68	14.74	395.32	metaSKAT	4.18	1.02	0.82	18.98

Table 2: Perfomances in terms of mean number of TP, FP, FN and TN over 50 replications for methods sgPLS, joint-sgPLS, ASSET and metaSKAT in simulations cases 3 and 4 (when 10 SNPs of the 5 groups are involved in the true model and the effects are in the same directions from one study to another). The number of observations per study is n.

n=200									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	17.96	36.28	82.04	363.72	sgPLS	1.46	4.32	3.54	15.68
joint-sgPLS	43.44	25.2	56.56	374.8	joint-sgPLS	3.46	3.5	1.54	16.5
ASSET	30.58	22.48	69.42	377.52	metaSKAT	2.2	0.98	2.8	19.02
n=400									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	75.76	139.34	24.24	260.66	sgPLS	4.74	11.9	0.26	8.1
joint-sgPLS	66.12	76.44	33.88	323.56	joint-sgPLS	4.48	7.62	0.52	12.38
ASSET	47.74	25.4	52.26	374.6	metaSKAT	3.14	1.26	1.86	18.74

Table 3: Perfomances in terms of mean number of TP, FP, FN and TN over 50 replications for methods sgPLS, joint-sgPLS, ASSET and metaSKAT in simulations cases 5 and 6(when all SNPs of the 5 groups are involved in the true model and the effects are in opposite directions from one study to another). The number of observations per study is n.

Only 4 loci have been associated with thyroid cancer risk and have been replicated in other studies [28]. One of them, 2q35, was also previously reported to increase risk of breast cancer [29]. To date, no study has been conducted to identify common genetic factors between breast and thyroid cancer. Exploring the genetic relationship between the two cancers would help to elucidate the common mechanisms between both disease and could permit to improve their diagnostic and therapeutic management.

We propose to illustrate the methods on real datasets, by investigating the pleiotropic effect of genetic variants from candidate pathways in breast and thyroid cancers.

Beluhca dataset includes data from CECILE, a french case-control study on breast cancer (1 125 cases, 1 172 controls) and from CATHY a french case-control study on thyroid cancer (463 female cases and 482 female controls). All these individuals were genotyped using a customized microarray including 8 716 genetic variants from 28 candidate pathways (648 genes) selected from KEGG database and from a literature review. After quality controls, we retained 6 677 SNPs available for both type of cancers. Missing values were imputed using the median among cases or controls and

n=200									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	13.2	74.96	36.8	375.04	sgPLS	2.04	6.5	2.96	13.5
joint-sgPLS	35.62	94.24	14.38	355.76	joint-sgPLS	4.58	7.02	0.42	12.98
ASSET	29.18	45.5	20.82	404.5	metaSKAT	3.54	0.92	1.46	19.08
n=400									
Variable level performances					Group level performances				
	TP	FP	FN	TN		TP	FP	FN	TN
sgPLS	14.3	72.28	35.7	377.72	sgPLS	1.88	5.76	3.12	14.24
joint-sgPLS	39.12	99.06	10.88	350.94	joint-sgPLS	4.9	7.18	0.1	12.82
ASSET	34.04	56.14	15.96	393.86	metaSKAT	4.22	0.86	0.78	19.14

Table 4: Perfomances in terms of mean number of TP, FP, FN and TN over 50 replications for methods sgPLS, joint-sgPLS, ASSET and metaSKAT in simulations cases 7 and 8 (when 10 SNPs of the 5 groups are involved in the true model and the effects are in the opposite directions from one study to another). The number of obervations per study is n.

data were centered to $\mu = 0$. When 2 SNPs were correlated at $r^2 = 1$, one of the SNP was removed and couple of extremely correlated ($r^2 > 0.98$) SNPs belonging to same genes were eliminated.

As dimension reduction methods such as sgPLS and joint-sgPLS need to be extended in case of overlapping groups of variables [30], 10 non-overlapping pathways were selected and only the 3766 SNPs related to those groups were kept in the final database. After all this steps, the new dataset is composed of 3766 SNPs, grouped in 337 non-overlapping genes and 10 non-overlapping pathways. The list of the pathways and there genes is displayed in Tables 6 and 7 in Appendix A.

The methods implemented in this article are : ASSET, metaSKAT, sgPLS and joint-sgPLS. For metaSKAT, sgPLS and joint-PLS, SNP-level, gene-level and pathway-level results are given by the methods whereas ASSET gives only SNP-level results. Hence, in the case of ASSET, genes corresponding to selected SNPs are considered. For each SNP i, an univariate logistic model for gene-disease association can be considered separately for thyroid data and breast data (thyroid and breast cancer, Figure 4).

As it has been presented before, for sgPLS and joint-sgPLS, a calibration of the parameters is generally performed through a cross-validation process. This process relies on the definition of a measure of performance: the error of prediction of the model. However, in genetic studies, the effects are small and the prediction performances based on genetic units are usually very low. The prediction performance results of sgPLS and joint-sgPLS shows no significant difference from different sets of penalization parameters (i.e., numbers of genes and pathways). In order to facilitate the interpretation, we present the results for calibration parameters set to 20 genes and 3 pathways and $\alpha = 0.5$. We explore the stability of the methods using a bootstrap strategy.

The method sgPLS and joint-sgPLS are computed with those parameters on the data. A 100 bootstraps procedure is performed on the data. The methods sgPLS and joint-sgPLS are then implemented on each bootstrap. We are interested in knowing whether genes selected in the computation on original data are still selected in the bootstrap. Genes and pathways selected by the methods applied on the original

data are considered as preselected. Then, the rates of selection of selected genes and pathways and non selected genes and pathways over the 100 bootstraps are compared. Figures 5 and 6 present this rate for preselected and non preselected features. A gene and resp. a pathway is kept in the final selection if and only if it is preselected and its rate of selection among the bootstraps is higher than any other gene (resp. pathway) that is not preselected. We can see that for joint-sgPLS less genes are selected than for other methods (4 against resp. 20 and 18 for metaSKAT and sgPLS on both data).

Results of the selection are presented in Table 5 where the name of genes and pathways is presented. "sgPLS single" stand for the use of the sgPLS on thyroid and breast data separately while "sgPLS both" stands for the use of the method on a concatenation of both data set standardizing by study.

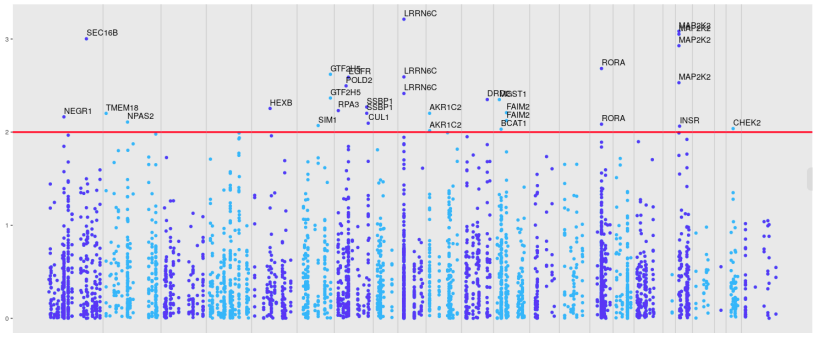
From one hand we can see that for non meta-analysis methods (SKAT and sgPLS), except for *INSR*, genes selected for thyroid cancer are distinct from genes selected for breast cancer. From the other hand, meta-analysis methods, which are metaSKAT, sgPLS, joint-sgPLS, select gene and pathways for both data sets.

We can note that *LRRN6C* is selected for thyroid data with the univariate model, but the gene is selected only by ASSET and not by the others. A similar remark can be said about *RORA* which is selected for breast cancer with the univariate model and ASSET and not with the other methods. The fact that those genes are selected only by SNP based methods would mean that only one SNP of the gene has an effect on the phenotype, but the rest of the gene do not have much effect.Longer genes has more chance to be selected by these SNP based methods.

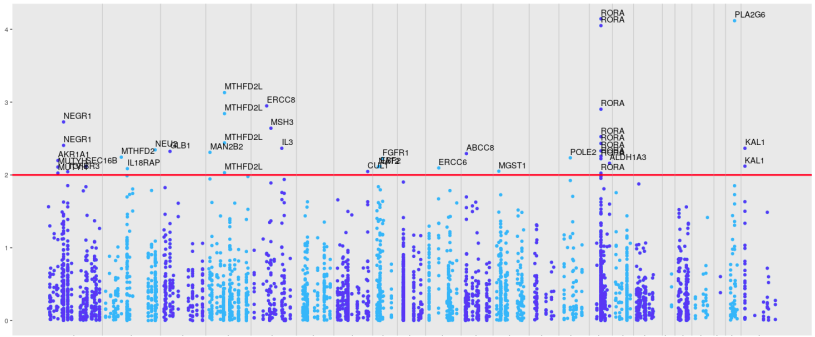
The gene *MAP2K2* is selected by all methods for thyroid and by joint-sgPLS as well. The gene seems to be associated with thyroid cancer, but the fact that it is selected also by sgPLS means it could be a candidate for pleiotropy. The same can be said for *PLA2G6, MTHFD2, ERCC3* that is initally associated to breast cancer but could be pleiotropic.

We can see that joint-sgPLS selects the pathways of inflammatory response and obesity and obesity-related phenotypes that are not selected by other methods.These results are interesting as obesity and inflammation are known to play a role in both cancers. We can wonder if there is an association at the pathway-level that is not detected by other methods.

Remark : Results based on different choice of calibration parameters for sgPLS and joint-sgPLS (50, 100 genes and 5 pathways) showed similar patterns.



(a) Results for univariate model on thyroid data.



(b) Results for univariate model on breast data.

Figure 4: Score for association of SNPs with the outcome for univariate model. The score is computed as $-\log_{10}(p)$

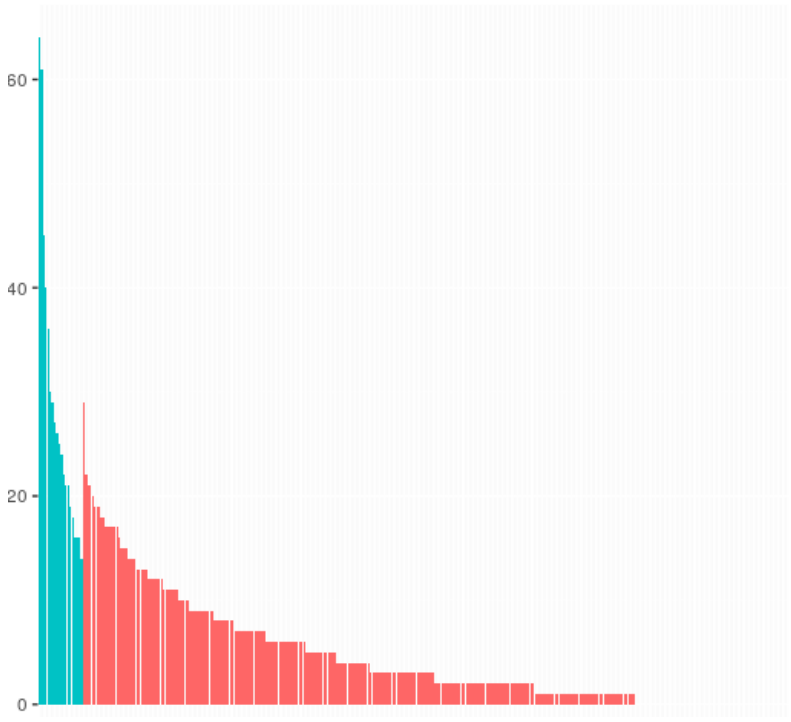
where p is the p-value. The red line corresponds to the threshold 0.01.

gene	ASSET	SKAT	metaSKAT	sgPLS single	sgPLS Both	joint-sgPLS
Cell cycle						
All pathway				thyroid	both	
EGFR	thyroid					
MAP2K2	both	thyroid		thyroid		both
PLA2G6	both	breast		breast	both	both
PTEN	both					
Circadian Rythm						
All pathway				breast		
RORA	both					
DNA repair						
All pathway		thyroid				
ERCC3	both	breast	both	breast	both	both
ERCC6		breast	both	breast		
ERCC8	both	breast		breast		
FEN1					both	
GTF2H1		thyroid			both	

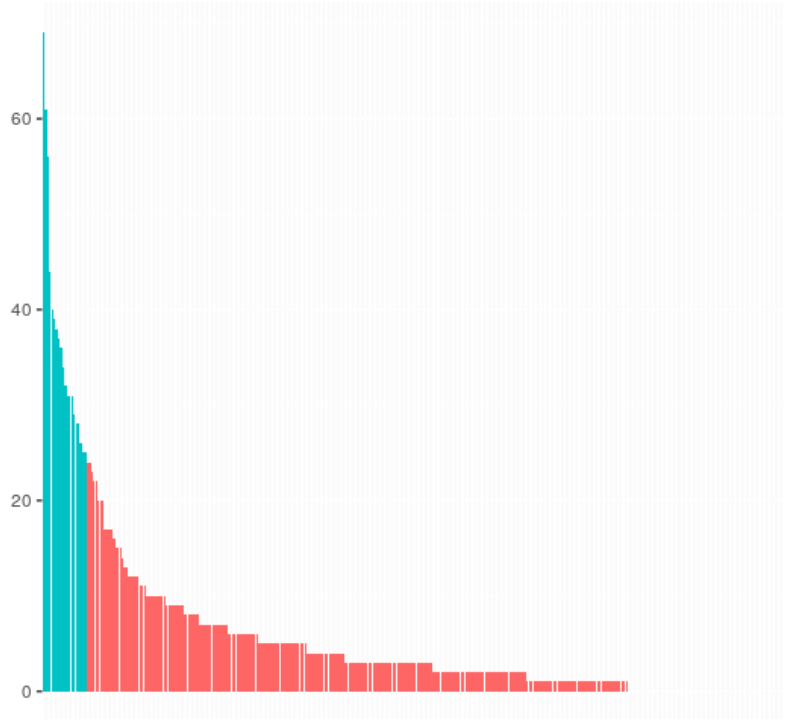
gene	ASSET	SKAT	metaSKAT	sgPLS single	sgPLS Both	joint-sgPLS
GTF2H5	thyroid	thyroid		thyroid		
MAPK8				breast		
MSH3	both					
MUTYH		breast	both	breast	both	
OGG1		breast		breast		
POLD2	thyroid					
POLE2		breast		breast	both	
RPA2				breast	both	
SSBP1		thyroid		thyroid		
Folate metabolism						
All pathway				breast	both	both
MTHFD2		breast		breast	both	both
MTHFD2L	both	breast		breast	both	
Inflammatory response						
All pathway				breast		both
CYP4F11			both			
IL13		breast	both	breast	both	
IL15		thyroid				
IL18RAP	both					
IL1A				breast		
IL3		breast		breast		
MMP25				breast	both	
Obesity and obesity-related phenotypes						
All pathway						
DRD2		thyroid				
FAIM2		thyroid				
GNPDA2			both			
INSR	both	both	both		both	
LRRN6C	both					
NEGR1	both					
NR3C1			both			
SEC16B	both					
Other glycan degradation						
All pathway					both	both
HEXA			both			
HEXB			both			
MAN2B2	both		both		both	
NEU2	both		both			
Precocious or delayed puberty						
All pathway			both			
FGFR1		breast		breast		
KAL1		breast	both			

gene	ASSET	SKAT	metaSKAT	sgPLS single	sgPLS Both	joint-sgPLS
TGFBR3	both		both			
Nicotinate and nicotinamide metabolism						
All pathway						
ENPP3			both			
NADK				breast	both	
NMNAT2			both			
NMNAT3			both			
NT5C		thyroid		thyroid		
PNP					both	
ADH1A		thyroid				
AKR1A1		breast	both	breast		
AKR1C2		thyroid				
ALDH1A3		breast		breast	both	
CYP2C18			both			
CYP2C19			both			
CYP2E1		thyroid				
CYP2F1	both	thyroid		thyroid		
GSTA2				thyroid		
MGST1	both				both	
NAT2	both	breast		breast	both	

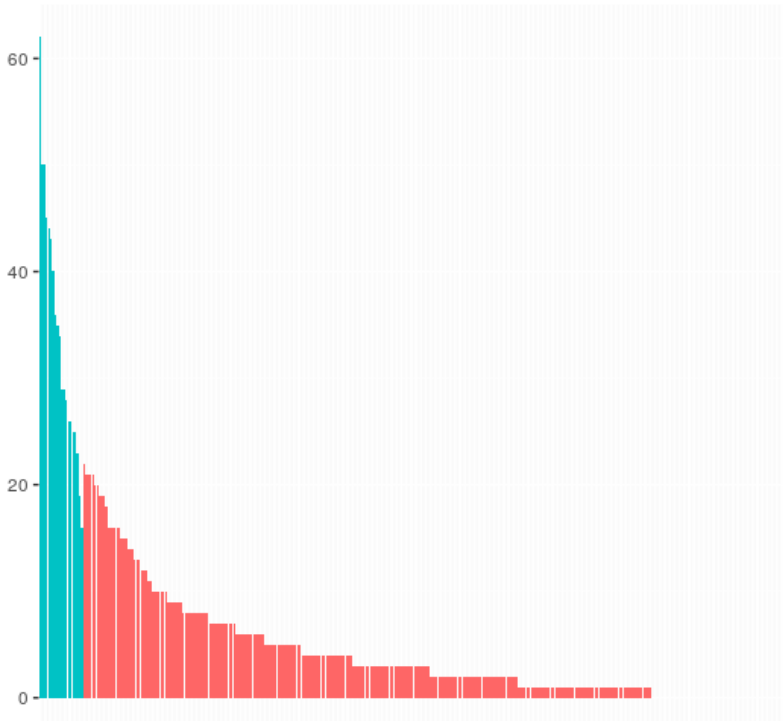
Table 5: Selected genes and pathways



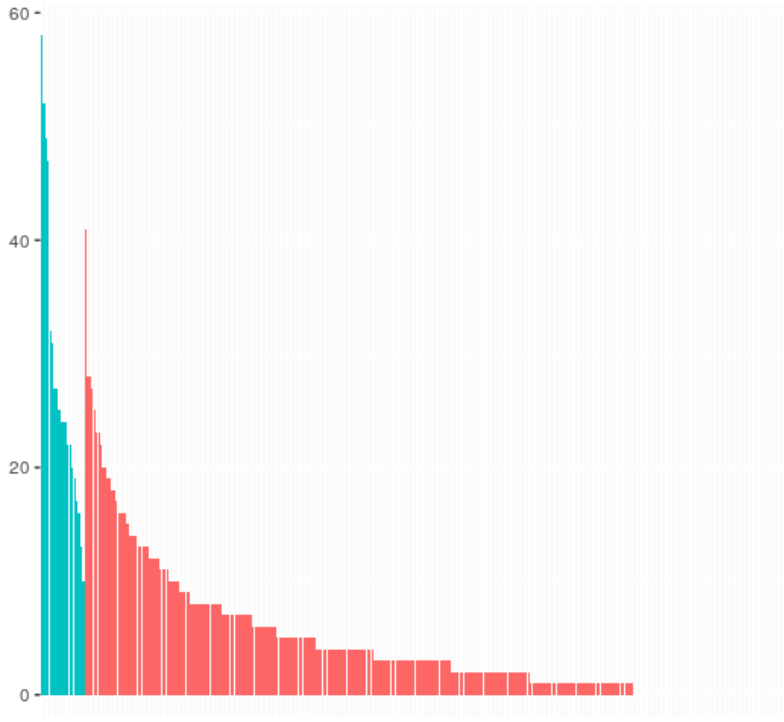
(a) Percent of selection of genes for sgPLS on thyroid data, 100 bootstraps



(b) Percent of selection of genes for sgPLS on breast data, 100 bootstraps

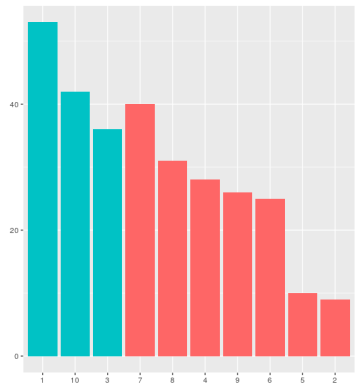


(c) Percent of selection of genes for sgPLS on all data, 100 bootstraps

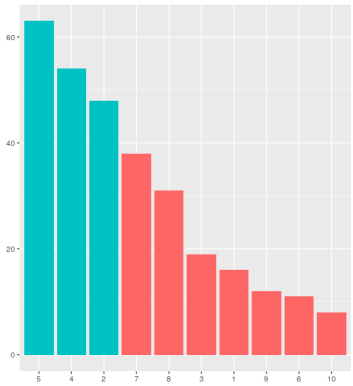


(d) Percent of selection of genes for joint-sgPLS on all data, 100 bootstraps

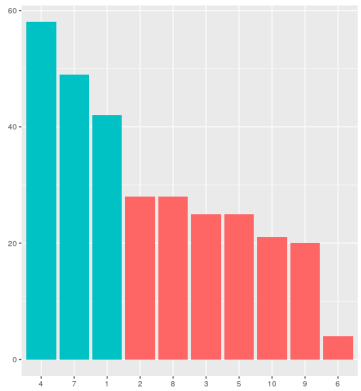
Figure 5: Percent of selection of genes for sgPLS and joints sgPLS on 100 bootstraps. (a) sgPLS on thyroid data (b) sgPLS on breast data (c) sgPLS on both data (d) joint-sgPLS. Genes selected on original data (preselected ones) are in blue while other genes (non preselected ones) are in red.



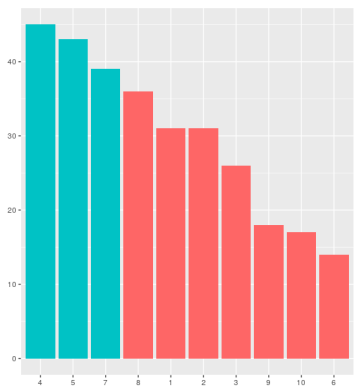
(a) Percent of selection of pathways for sgPLS on thyroid data, 100 bootstraps



(b) Percent of selection of pathways for sgPLS on breast data, 100 bootstraps



(c) Percent of selection of pathways for sgPLS on all data, 100 bootstraps



(d) Percent of selection of pathways for joint-sgPLS on all data, 100 bootstraps

Figure 6: Percent of selection of pathways for sgPLS and joints sgPLS on 100 bootstraps. (a) sgPLS on thyroid data (b) sgPLS on breast data (c) sgPLS on both data (d) joint-sgPLS. Pathways selected on original data (preselected ones) are in blue while other pathways (non preselected ones) are in red.

The pathways are noted: 1) Cell cycle 2) Circadian rhythm 3) Folate metabolism 4) Other glycan degradation 5) Obesity and obesity-related phenotypes 6) DNA repair 7) Metabolism of xenobiotics 9) Precocious or delayed puberty 10) Inflammatory response.

Discussion

In this article, the properties of the joint-sgPLS is presented and is compared to the classical sgPLS, the ASSET method and metaSKAT. From one hand, joint-sgPLS and ASSET are design to retrieve effect of opposite direction. From the other hand, joint-sgPLS, sgPLS and metaSKAT can give results at a gene-level instead of a SNP-level whereas ASSET can not. We have seen through a simulation that joint-sgPLS can be competitive against any of the other methods considered. Hence, joint-sgPLS seems perfectly suited for meta-analysis where effects in opposite directions can exist which invite us to pursue further investigation with it in complex studies for genetic epidemiology such as pleiotropy.

Conclusion

We do believe that further investigation can be done on the same subject. In this article, sgPLS and joint-sgPLS have been applied with one component, but several components could be considered. This could lead to the selection of variables that are orthogonal to the selection of the first component but that have still a large participation to the covariance matrix.

We acknowledge that on the application the stability of the method is an important point due to the fact that the cross-validation procedure is not satisfying for choosing the parameters of penalization. One improvement could consist in exploiting different the criteria of the procedure (the error prediction) with, for instance, stability measures [31]. Another improvement could consist in adaptating the adaptative Lasso [32] for our method which could bypass the stability questions.

In order to advance on the application, this study should be replicated on a larger data base. Particularly, thyroid cancer has been less studied than breast cancer, and data for thyroid are still scarce in this application. Other cases of pleiotropy could be investigated, for instance for the case where the phenotype is multivariate for each subject. The joint-sgPLS is suitable for any kind of phenotype, continuous or qualitative. R code is available from the author to reproduce the results and is available on github (`https://github.com/camilobroc/BMC_joint_sgPLS`).

Declarations

Ethics approval and consent to participate
Written informed consent for the present study was obtained from all participants. Study protocols were approved by the French ethic committees (CNIL, CCPPRB) (reference numbers 05-3144 for CATHY study and 04-53 for CECILE study).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Funding
This study was supported by the "Ligue contre le Cancer" for it's Cross Cancer Genomic Investigation of Pleiotropy project.

Authors' contributions
CB and BL designed the novel method. CB implemented the analysis. TT perfomed interpretation on real data. CB, TT and BL wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The authors acknowledge Pascal Guénel for providing the breast and thyroid cancer data. The authors acknowledge also the calculus center MCIA (Mésocentre de Calcul Intensif Aquitain) for providing its facilities. The "Ligue contre le Cancer" is acknowledged as well for its support for Cross Cancer Genomic Investigation of Pleiotropy project.

Author details

[1]Laboratory for Data Sciences and Decision (Digiteo), LIST, CEA, , Gif-sur-Yvette, France. [2]Laboratoire De Mathématiques et de leurs Applications de PAU E2S UPPA, CNRS, , Pau, France. [3] Université Paris-Saclay, UVSQ, Inserm, CESP, 94807, , Villejuif, France. [4]Gustave Roussy, 94805, , Villejuif, France. [5]Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical Sciences at Queensland University of Technology, , Brisbane, Australia.

References

1. Paaby, A.B., Rockman, M.V.: The many faces of pleiotropy. Trends in Genetics **29**(2), 66–73 (2013)
2. Gratten, J., Visscher, P.M.: Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome medicine **8**(1), 78 (2016)
3. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., Smoller, J.W.: Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics **14**(7), 483 (2013)
4. Yang, C., Li, C., Wang, Q., Chung, D., Zhao, H.: Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. Frontiers in genetics **6**, 229 (2015)
5. Gagnon-Bartsch, J.A., Speed, T.P.: Using control genes to correct for unwanted variation in microarray data. Biostatistics **13**(3), 539–552 (2012)
6. Wold, H.: Path models with latent variables: The nipals approach -, 307–357 (1975)
7. Boulesteix, A.-L., Strimmer, K.: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in bioinformatics **8**(1), 32–44 (2006)
8. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901)
9. Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., Besse, P.: A sparse pls for variable selection when integrating omics data. Statistical applications in genetics and molecular biology **7**(1) (2008)
10. Liquet, B., de Micheaux, P.L., Hejblum, B.P., Thiébaut, R.: Group and sparse group partial least square approaches applied in genomics context. Bioinformatics **32**(1), 35–42 (2015)
11. Sutton, M., Thiébaut, R., Liquet, B.: Sparse partial least squares with group and subgroup structure. Statistics in Medicine (in press) (2018)
12. Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing **20**(2), 231–252 (2010)
13. Broc, C., Calvo, B., Liquet, B.: Penalized partial least square applied to structured data. Arabian Journal of Mathematics, 1–16 (2019)
14. Bhattacharjee, S., Rajaraman, P., Jacobs, K.B., Wheeler, W.A., Melin, B.S., Hartge, P., Yeager, M., Chung, C.C., Chanock, S.J., Chatterjee, N., *et al.*: A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. The American Journal of Human Genetics **90**(5), 821–835 (2012)
15. Lee, S., Teslovich, T.M., Boehnke, M., Lin, X.: General framework for meta-analysis of rare variants in sequencing association studies. The American Journal of Human Genetics **93**(1), 42–53 (2013)
16. Lafaye de Micheaux, P., Liquet, B., Sutton, M., *et al.*: Pls for big data: A unified parallel algorithm for regularised group pls. Statistics Surveys **13**, 119–149 (2019)
17. Wang, T., Ho, G., Ye, K., Strickler, H., Elston, R.C.: A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genetic epidemiology **33**(1), 6–15 (2009)
18. Shen, H., Huang, J.Z.: Sparse principal component analysis via regularized low rank matrix approximation. Journal of multivariate analysis **99**(6), 1015–1034 (2008)
19. Vinzi, V.E., Trinchera, L., Amato, S.: Pls path modeling from foundations to recent developments and open issues for model assessment and improvement, 47–82 (2010)
20. Geladi, P., Kowalski, B.R.: Partial least-squares regression: a tutorial. Analytica chimica acta **185**, 1–17 (1986)
21. Eslami, A., Qannari, E.M., Kohler, A., Bougeard, S.: Algorithms for multi-group pls. Journal of Chemometrics **28**(3), 192–201 (2014)
22. Sutton, M., Thiébaut, R., Liquet, B.: Sparse partial least squares with group and subgroup structure. Statistics in medicine **37**(23), 3338–3356 (2018)
23. Tseng, P., et al.: Coordinate ascent for maximizing nondifferentiable concave functions (1988)
24. Netrapalli, P., Jain, P., Sanghavi, S.: Phase retrieval using alternating minimization. In: Advances in Neural Information Processing Systems, pp. 2796–2804 (2013)
25. Nielsen, S.M., White, M.G., Hong, S., Aschebrook-Kilfoy, B., Kaplan, E.L., Angelos, P., Kulkarni, S.A., Olopade, O.I., Grogan, R.H.: The breast–thyroid cancer link: a systematic review and meta-analysis. Cancer Epidemiology and Prevention Biomarkers **25**(2), 231–238 (2016)
26. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H., Wang, Q., Bolla, M.K., *et al.*: Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. The American Journal of Human Genetics **104**(1), 21–34 (2019)
27. Czene, K., Lichtenstein, P., Hemminki, K.: Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. International journal of cancer **99**(2), 260–266 (2002)
28. Gudmundsson, J., Thorleifsson, G., Sigurdsson, J.K., Stefansdottir, L., Jonasson, J.G., Gudjonsson, S.A., Gudbjartsson, D.F., Masson, G., Johannsdottir, H., Halldorsson, G.H., *et al.*: A genome-wide association study yields five novel thyroid cancer risk loci. Nature communications **8**, 14517 (2017)
29. Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., *et al.*: Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer. Nature genetics **39**(7), 865 (2007)
30. Jacob, L., Obozinski, G., Vert, J.-P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 433–440 (2009). ACM

31. Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. Journal of Machine Learning Research **18**, 174–1
32. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American statistical association **101**(476), 1418–1429 (2006)

Appendix A: Tables of genes and pathways

1)		2)		3)		4)	
gene	count	gene	count	gene	count	gene	count
ENPP1	27	RORA	282	MTHFS	19	GLB1	22
NMNAT3	25	NPAS2	60	MTHFD1	11	MAN2B2	18
AOX1	24	RORB	34	MTHFR	11	ENGASE	13
NMNAT2	21	ARNTL	23	MTHFD2L	9	HEXB	12
ENPP3	15	CUL1	22	MTHFD2	1	MANBA	9
BST1	14	BTRC	13			AGA	8
CD38	11	RORC	13			FUCA2	8
NT5M	10	PER3	12			NEU2	5
NT5C1A	9	CLOCK	11			FUCA1	3
NT5C2	8	PER2	11			HEXA	3
PNP	8	CRY2	9			NEU3	3
NAMPT	7	CSNK1E	9			GBA	2
NNMT	7	BHLHE40	8			MAN2B1	2
NT5C3	7	FBXW11	8			NEU1	2
NADSYN1	6	CRY1	7			NEU4	1
NNT	6	FBXL3	7				
NT5E	5	NR1D1	7				
NMNAT1	4	TIMELESS	7				
NUDT12	4	PER1	5				
QPRT	4	SKP1	4				
NT5C1B	3	BHLHE41	3				
NADK	2	CSNK1D	3				
NT5C	2	RBX1	2				

5)		5)		6)		6)	
gene	count	gene	count	gene	count	gene	count
LRRN6C	197	CRHR1	8	RPA3	25	TDG	8
FTO	122	ETV5	8	NEIL3	24	XRCC1	8
NEGR1	93	IL1RN	8	XRCC5	21	DDB2	7
SCARB1	33	LDLR	8	EXO1	19	ERCC8	7
ABCC8	31	HTR2C	7	MSH3	19	MSH2	7
SEC16B	25	MCHR1	7	NEIL2	17	PMS2	7
LEPR	23	TMEM18	7	RPA1	17	POLD1	7
MAP2K5	23	BDNF	6	BRCA2	16	POLE	7
NR3C1	18	KCTD15	6	RAD23B	16	RFC4	7
FAIM2	17	UCP2	6	RFC3	14	XRCC2	7
DRD2	16	ACE	5	PARP4	13	XRCC3	7
PPARG	16	ADRB2	4	POLD3	12	CDK2	6
SIM1	14	IL6	4	RFC5	12	GTF2H1	6
FANCL	12	LEP	4	CHEK1	11	LIG1	6
GHRL	12	MC4R	4	CHEK2	11	MUTYH	6
ADIPOQ	11	PLIN	4	MSH6	11	PARP2	6
CRHR2	11	PTPN11	4	TERT	10	POLD2	6
GNB3	10	UCP3	4	CASP7	9	POLE2	6
INSR	10	GNPDA2	3	MNAT1	9	TP53	6
GPRC5B	9	NR0B2	3	PARP1	9	XPA	6
MC3R	9	RETN	3	RFC1	9	BRCA1	5
PCSK1	9	CCL5	2	XPC	9	CDK7	5
TFAP2B	9	LEPROTL1	2	CASP3	8	CHRNA4	5
TNF	9	SH2B1	1	ERCC5	8	CUL4B	5
UCP1	9			ERCC6	8	ERCC3	5

Table 6: First pathways and their corresponding genes. The number of SNPs for each gene is presented. The pathways are 1) Nicotinate and nicotinamide metabolism 2) Circadian rhythm 3) Folate metabolism 4) Other glycan degradation 5) Obesity and obesity-related phenotypes 6) DNA repair

6)		6)		7)		7)	
gene	count	gene	count	gene	count	gene	count
GTF2H5	5	POLL	3	UGT1A8	58	CYP1B1	8
LIG3	5	RPA2	3	UGT2A1	29	UGT2B4	8
MAPK8	5	SSBP1	3	MGST2	22	ADH4	7
NTHL1	5	CCNH	2	CYP2C8	17	ADH5	7
OGG1	5	ERCC4	2	AKR1C2	16	AKR1C1	7
RAD50	5	GTF2H3	2	CYP2C9	15	ALDH3B1	7
RPA4	5	HMGB1	2	CYP2B6	14	CYP2E1	7
APEX1	4			EPHX1	14	NQO1	7
CDKN1A	4			MGST3	14	ALDH1A3	6
ERCC2	4			AKR1C4	13	CYP2F1	6
FEN1	4			COMT	12	DHDH	6
GTF2H4	4			CYP2C19	12	SOD2	6
MBD4	4			CYP2S1	12	GSTA2	5
POLE3	4			CYP2C18	11	GSTM3	5
RAD51	4			ADH1B	10	GSTO2	5
RFC2	4			ADH7	10	GSTP1	5
UNG	4			GSTA4	10	CYP1A1	4
CDKN2D	3			GSTZ1	10	CYP1A2	4
CETN2	3			MGST1	10	CYP3A43	4
CUL4A	3			NAT2	10	GSTA3	4
ERCC1	3			ADH1C	9	GSTM4	4
MLH1	3			AHR	9	UGT2B7	4
MPG	3			NAT1	9	AKR1A1	3
POLB	3			ADH6	8	ALDH3B2	3
POLE4	3			AKR1C3	8	CYP2A6	3

7)		8)		9)		10)	
gene	count	gene	count	gene	count	gene	count
CYP3A4	3	TGFBR3	77	TGFB2	22	EGFR	103
CYP3A7	3	EBF2	60	IL18RAP	13	CCND3	29
GSTM2	3	BCAT1	51	CYP4F11	12	MAPT	19
GSTM5	3	VDR	34	EPHX2	12	MAP2K4	15
UGT2A3	3	KAL1	26	IL7	11	EGF	11
UGT2B11	3	TM7SF3	19	IGFBP1	9	MAP2K2	11
ADH1A	2	CASC1	17	IGFBP3	9	GSK3B	8
CYP2D6	2	FGFR1OP2	12	IL17A	9	PLA2G6	8
		FGFR1	11	IL10	8	PTEN	8
		KRAS2	10	IGFBP4	7	MAP2K1	6
		CCR3	8	IL15	7	MYBL2	6
		KISS1	8	MMP25	7	AKT1	5
		PROK2	7	IL16	6	MAP2K3	5
		LIF	5	IL12A	5	FGFR3	3
		PROKR2	5	IL13	5	MAP2K7	3
		PTH1R	4	IL18	5	MAPK12	3
		NKX2-1	2	IL19	5	TP53I3	3
				IL2	5	CCNA2	2
				IL9	5	MAPK7	2
				PLA2G4B	5		
				IL3	4		
				TGFB1	4		
				IL1B	3		
				IL4	3		
				IL1A	2		
				IL23A	1		

Table 7: Last pathways and their corresponding genes. The number of SNPs for each gene is presented. The pathways are 6) DNA repair 7) Metabolism of xenobiotics 8) Precocious or delayed puberty 9) Inflammatory response 10) Cell cycle.

Conclusions and futur work

Genetic epidemiology aims at understanding better the influence of the genome on the onset of a disease. It has promising perspectives it terms of personalized medicine and for pharmacology. The rise of high dimensional data in the field, which are called genomics, has raised the stakes even further in our expectations in terms of findings. The use of the term “big data” for this trend must be very cautious. Indeed, genomic data are massive but are structure in an atypical way compared to other data analysis domains because the number of variables is extremely high while the number of observations is moderate. Furthermore, the domain needs to cope carefully with the biological meaning of its models. The treatment in an appropriate way of groups of variables, groups of observations and covariates must translate medical knowledge.

5.4 Contributions

This dissertation deals with the development of statistical methods for genomic data where two genetic problems has been considered.

Both contributions answer a general framework specific to genetic studies. The idea consists in finding algorithms with mathematical properties which are adapted to the field of application. It leads to the development of methods suited for the study of “ill-posed” problems. It also leads to treat a group of variables structure as an a priori information. For this purpose, truncation, sparsity, weighting are different terms referring to distinct statistical method that follow a general philosophy: concentrate the models on sets of variables that have the highest contribution to the signal. However it leads generally to the choice of hyper-parameters to which results are heavily dependent. A proper calibration needs generally repeated measure and resampling techniques are a common answer to this. The present dissertation deals with all those aspects at the different levels of the contributions.

The first one, gene-environment interaction, studies the influence of a gene on the appearance of a disease. It aims at target specific environmental factors for predicting risks in a population regarding their genome or at determining conditions for the application of a medical treatment. This work was published in the Journal of the French Society the implementation is available in the PIGE package for R (in CRAN).

The second one, pleiotropy, studies complex links when several genetic traits and several

phenotypes are involved in a same biological mechanisms. The goal is, for instance, to highlight processes implying several diseases. In this thesis, the method has been applied to oncology data. This method has published in the Arabian Journal of Mathematics and its application to oncology will be submitted to BMC Bioinformatics. the implementation is available in the sgPLS package for R (in CRAN).

5.5 Discussion

Some aspects presented in this dissertation leave some open question and provide perspectives for further investigations.

From a statistical point of view, we do believe that the development of statistical methods that have been done in this thesis can be extended to an even larger number of frameworks.

- The first contribution introduces the parametric bootstrap as a resampling method in the case a framework for interaction studies. The choice of resampling methods can be overlooked when developing a new method and putting an emphasis in the choice of the resampling methods, in the same spirit that in this contribution, is an interesting path of investigation.
- The second contribution introduces an extension of the sgPLS in order to take into account effect in opposite directions in a meta-analysis. Especially, its use with a group-Lasso is proposed. From our knowledge tackling both opposite effects and a group of variable structure in a same method hasn't been quite investigated. In the same spirit, other perpendicular statistical techniques for PLS could be integrated to the presented framework. For instance, an elastic net can deal with stability of the results and adaptive Lasso and its oracle properties can solve the choice of hyper-parameters. Altough, this material hasn't been discussed in this thesis, it can be a promising development for future works.

From an epidemiologic point of view, the statistical methods could be applied to a larger number of biological problems.

- The first contribution proposes to investigate $G \times E$ interaction while the second one propose to investigate pleiotropy. The next step would be to study $G \times E$ interaction with pleiotropy. The method could be applied to the data provided by the french health institute INSERM. From a statistical point of view, the method of the first contribution could be adapted using a meta-analysis framework like presented in this dissertation. This could open new biological interpretations.
- Diseases studied in this dissertation are mainly cancers. However, pleiotropy can be related to other kind of diseases. The developed method could be applied to other fields than oncology.
- Other fields of research have data with features similar to genetics. Mainly, chemometrics or neuroscience handle data with a large number of variables compared to the number of observations and results at a group of variables level can be beneficial. For instance in neuroscience, data are collected through probes that captures the

neurons signal in given areas of the brain overtime. Observations corresponding to related time frames and related brain areas can be correlated and can be used as an a priori structure of the variables.

List of Symbols, Notations and concepts

Biology and bio-statistics vocabulary

	Definition
DNA	Molecule in the kernel of human cells. The genetic information is the same in each cell of a human.
Gene	Identified section of the DNA. A gene can exist in different variations in a population and has an effect on the functioning of a cell.
SNP	The single nucleotide polymorphism or SNP is the smallest measured genetic variations : single base-pair changes in the DNA sequence that occur with high frequency in the human genome. Several SNP can generally be found in one gene.
Variants, Alleles	The different versions of a SNP are called variants or alleles.
MAF	The Minor Allele Frequency is the proportion of the most rare variant of a SNP in the population. For instance, a SNP with a MAF of 0.30 implies that 30% of a population has the allele versus the more common allele (the major allele) which is found in 70% of the population.
Rare variant	Below 0.5% MAF, the variant is called rare. It can also be called a mutation.
Low-frequency variant	between 0.5% and 5% the allele is called low frequency.
Common variant	and above 5% the variant is called common variant.
Locus (pl. loci)	A position on a chromosome, either a gene or a SNP.
Pathway	A group of gene involved in a same biological process.
Data	A set of variables and observations. For each observation and each variable a value is collected. Data can be represented in a matrix where the rows represent the observations and the columns represent the variables.
Predictors,explanatory variables	Variables that are assumed to have an effect on an outcome

Outcome,dependent variable, response	The variable that we want to explain.
Covariate	Variables that gives an additional information and can be used.
Statistic	A statistic is a value computed with a method that characterize a data set.
p-value	A p-value is a value between 0 and 1 associated to a statistic. It indicates weather a statistic is abnormal under a certain hypothesis. The p-value has classically been at the center of discussions in statistics. It is considered by many statisticians as the corner stone of statistics, eventhough some opposite positions are arising.
Biomarker	Measurable indicator on an organisms. They are the variable of interest we look for in bio-statistics. Notably, Genes, SNPs, loci can be biomarkers.

Mathematical notations

Each notation refer to the example introduced above it.

Object	Example	Definition
Usual mathematical sets	$\mathbb{R}$, $\mathbb{N}$	$\mathbb{R}$, $\mathbb{N}$ represent resp. the set of the real numbers and the set of the positive integers.
Element of a set	$i, j, n, p, q \in \mathbb{N}$	$\in$ indicates that i,j,n, p, q are elements of a set.
Subset	$\mathbb{S}, \mathbb{S}_i, \mathbb{S}_j \subset \mathbb{R}$	$\subset$ indicates a subset of a set.
Vector	$v \in \mathbb{R}^p$	this notation indicates that v is a p-vector of $\mathbb{R}$, i.e., a finite ordered list of p elements of $\mathbb{R}$.
Element of a vectors	v_i	v_i is the i-th element of v
Subvector	$v_{\mathbb{S}}$	$v_{\mathbb{S}}$ is the vector made of the elements of a vector $v \in \mathbb{R}^p$ with indices corresponding to a subset $\mathbb{S} \in \mathbb{R}$
Matrix	$M \in \mathbb{R}^{p \times q}$	M is a matrix of size $p \times q$, i.e. with p rows and q columns. In general, in the rest of the document, Matrices are noted in capital letters.
Element of a matrix	$M_{i,j}$	$M_{i,j}$ is the element of $M \in \mathbb{R}^{p \times q}$ corresponding to the i-th row and j-th column.
Submatrix	$M_{\mathbb{S}_i,\mathbb{S}_j}$	$M_{\mathbb{S}_i,\mathbb{S}_j}$ is a matrix corresponding to the rows of M indicated by the subset $\mathbb{S}_i \subset \{1, \cdots, n\}$ and the columns of M indicated by the subset $\mathbb{S}_j \subset \{, \cdots, p\}$.
Submatrix with whole column or line	$M_{.,\mathbb{S}_i}$, $M_{i,.}$	The $\cdot$ notation, as presented, stands for the whole set of rows or columns.
Transpose	M^T	M^T is the transpose matrix of M. $M^T \in \mathbb{R}^{q \times p}$ and $M_{i,j} = M^T_{i,j}$.

Cardinal	$\#S$	stands for the cardinal of a set S.
1-Norm	$\|v\|_1$	is the 1-norm on vectors
2-Norm	$\|v\|_2$	is the 2-norm on vectors
Frobeius norm	$\|M\|_F$	is the Frobenius norm for matrices, i.e., $\|M\|_F = \sqrt{MM^T}$
Positive value	$(x)_+$	stands for the positive value of a real $x \in \mathbb{R}$. It is equal to the number if the number is positive and equal to zero otherwise. In other words $(x)_+ = \frac{\|x\|+x}{2}$

Bibliography

[1] Cell biology. https://en.wikipedia.org/wiki/Cell_Biology. Accessed: 2019-09-01. Citations: § 4, 5, and 133

[2] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010. Citations: § 6

[3] Obi L Griffith, Stephen B Montgomery, Bridget Bernier, Bryan Chu, Katayoon Kasaian, Stein Aerts, Shaun Mahony, Monica C Sleumer, Mikhail Bilenky, Maximilian Haeussler, et al. Oreganno: an open-access community-driven resource for regulatory annotation. *Nucleic acids research*, 36(suppl_1):D107–D113, 2007. Citations: § 6

[4] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010. Citations: § 6 and 10

[5] Bat-sheva Kerem, Johanna M Rommens, Janet A Buchanan, Danuta Markiewicz, Tara K Cox, Aravinda Chakravarti, Manuel Buchwald, and Lap-Chee Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080, 1989. Citations: § 6

[6] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012. Citations: § 6

[7] David E Reich and Eric S Lander. On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510, 2001. Citations: § 7

[8] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005. Citations: § 7 and 10

[9] B Devlin and Neil Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322, 1995. Citations: § 7

[10] Vivien Marx. Biology: The big challenges of big data, 2013. Citations: § 8

[11] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009. Citations: § 8

[12] Vijay K Ramanan, Li Shen, Jason H Moore, and Andrew J Saykin. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *TRENDS in Genetics*, 28(7):323–332, 2012. Citations: § 9 and 16

[13] Jonathan L Haines, Michael A Hauser, Silke Schmidt, William K Scott, Lana M Olson, Paul Gallins, Kylee L Spencer, Shu Ying Kwan, Maher Noureddine, John R Gilbert, et al. Complement factor h variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421, 2005. Citations: § 9

[14] Gregory M Cooper, Julie A Johnson, Taimour Y Langaee, Hua Feng, Ian B Stanaway, Ute I Schwarz, Marylyn D Ritchie, C Michael Stein, Dan M Roden, Joshua D Smith, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*, 112(4):1022–1027, 2008. Citations: § 9

[15] Annalise B Paaby and Matthew V Rockman. The many faces of pleiotropy. *Trends in Genetics*, 29(2):66–73, 2013. Citations: § 9 and 87

[16] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977. Citations: § 10

[17] Roger Bumgarner. Overview of dna microarrays: types, applications, and their future. *Current protocols in molecular biology*, 101(1):22–1, 2013. Citations: § 10

[18] Erwin L Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014. Citations: § 10

[19] Marylyn D Ritchie, Joshua C Denny, Dana C Crawford, Andrea H Ramirez, Justin B Weiner, Jill M Pulley, Melissa A Basford, Kristin Brown-Gentry, Jeffrey R Balser, Daniel R Masys, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *The American Journal of Human Genetics*, 86(4):560–572, 2010. Citations: § 10

[20] Hester M Wain, Elspeth A Bruford, Ruth C Lovering, Michael J Lush, Mathew W Wright, and Sue Povey. Guidelines for human gene nomenclature. *Genomics*, 79(4):464–470, 2002. Citations: § 10

[21] Data Coordination Centre at EMBL-EBI. Pop g, 2008. Citations: § 10

[22] UK10K consortium et al. The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571):82, 2015. Citations: § 10

[23] Vivien Marx. The dna of a nation, 2015. Citations: § 10

[24] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015. Citations: § 10

[25] Chen Yao, Brian H Chen, Roby Joehanes, Burcak Otlu, Xiaoling Zhang, Chunyu Liu, Tianxiao Huan, Oznur Tastan, L Adrienne Cupples, James B Meigs, et al. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation*, 131(6):536–549, 2015. Citations: § 11

[26] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015. Citations: § 11

[27] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302, 2016. Citations: § 11

[28] Juan Antonio Vizcaíno, Attila Csordas, Noemi Del-Toro, José A Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, Yasset Perez-Riverol, Florian Reisinger, Tobias Ternent, et al. 2016 update of the pride database and its related tools. *Nucleic acids research*, 44(D1):D447–D456, 2015. Citations: § 11

[29] Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-Francois Rual, Heather Borick, Pascal Braun, Matija Dreze, et al. Literature-curated protein interaction datasets. *Nature methods*, 6(1):39, 2009. Citations: § 11

[30] Gavin CKW Koh, Pablo Porras, Bruno Aranda, Henning Hermjakob, and Sandra E Orchard. Analyzing protein–protein interaction networks. *Journal of proteome research*, 11(4):2014–2031, 2012. Citations: § 11

[31] Miguel A Garcia-Campos, Jesus Espinal-Enriquez, and Enrique Hernandez-Lemus. Pathway analysis: state of the art. *Frontiers in physiology*, 6:383, 2015. Citations: § 16

[32] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007. Citations: § 16

[33] M. Evangelou, D.J. Smyth, M.D. Fortune, O.S. Burren, N.M. Walker, H. Guo, S. Onengut-Gumuscu, W. Chen, P. Concannon, S.S. Rich, J.A. Todd, and C. Wallace. A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genetic Epidemiology*, 38(8):661–670, 2014. Citations: § 16

[34] Peter Carbonetto and Matthew Stephens. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. *PLoS genetics*, 9(10):e1003770, 2013. Citations: § 16

[35] Hariklia Eleftherohorinou, Victoria Wright, Clive Hoggart, Anna-Liisa Hartikainen, Marjo-Riitta Jarvelin, David Balding, Lachlan Coin, and Michael Levin. Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases. *PloS one*, 4(11):e8068, 2009. Citations: § 16

[36] Mingyao Li, Kai Wang, Struan FA Grant, Hakon Hakonarson, and Chun Li. Atom: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*, 25(4):497–503, 2008. Citations: § 17 and 22

[37] Tao Wang and Robert C Elston. Improved power by use of a weighted score test for linkage disequilibrium mapping. *The american journal of human genetics*, 80(2):353–360, 2007. Citations: § 17

[38] Cassandra E Murcray, Juan Pablo Lewinger, and W James Gauderman. Gene-environment interaction in genome-wide association studies. *American journal of epidemiology*, 169(2):219–226, 2008. Citations: § 18

[39] Charles Kooperberg and Michael LeBlanc. Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(3):255–263, 2008. Citations: § 18

[40] Cassandra E Murcray, Juan Pablo Lewinger, David V Conti, Duncan C Thomas, and W James Gauderman. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genetic epidemiology*, 35(3):201–210, 2011. Citations: § 18

[41] Li Hsu, Shuo Jiao, James Y Dai, Carolyn Hutter, Ulrike Peters, and Charles Kooperberg. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic epidemiology*, 36(3):183–194, 2012. Citations: § 18

[42] Jose A Seoane, Colin Campbell, Ian NM Day, Juan P Casas, and Tom R Gaunt. Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS computational biology*, 10(10):e1003876, 2014. Citations: § 18

[43] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006. Citations: § 18

[44] Tao Wang, Gloria Ho, Kenny Ye, Howard Strickler, and Robert C Elston. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genetic epidemiology*, 33(1):6–15, 2009. Citations: § 18

[45] Tijl De Bie, Nello Cristianini, and Roman Rosipal. Eigenproblems in pattern recognition. In *Handbook of Geometric Computing*, pages 129–167. Springer, 2005. Citations: § 18

[46] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. Citations: § 18

[47] Herman Wold. Path models with latent variables: The nipals approach. In *Quantitative sociology*, pages 307–357. Elsevier, 1975. Citations: § 18

[48] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. Citations: § 18

[49] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992. Citations: § 18

[50] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. Citations: § 19

[51] Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008. Citations: § 19

[52] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. Citations: § 19

[53] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013. Citations: § 19

[54] Benoît Liquet, Pierre Lafaye de Micheaux, Boris P Hejblum, and Rodolphe Thiébaut. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1):35–42, 2015. Citations: § 19

[55] Matthew Sutton, Rodolphe Thiébaut, and Benoît Liquet. Sparse partial least squares with group and subgroup structure. *Statistics in medicine*, 37(23):3338–3356, 2018. Citations: § 19

[56] Herman Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142, 1975. Citations: § 20

[57] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009. Citations: § 20

[58] Matt Silver, Giovanni Montana, null Alzheimer's Disease Neuroimaging Initiative, et al. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Statistical applications in genetics and molecular biology*, 11(1):1–43. Citations: § 20

[59] Matt Silver, Peng Chen, Ruoying Li, Ching-Yu Cheng, Tien-Yin Wong, E-Shyong Tai, Yik-Ying Teo, and Giovanni Montana. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two asian cohorts. *PLoS genetics*, 9(11):e1003939, 2013. Citations: § 20

[60] Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011. Citations: § 20

[61] Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pages 352–360, 2011. Citations: § 20

[62] F. Dudbridge and B.P. Koeleman. Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology*, 25:360–366, 2003. Citations: § 21

[63] Kai Yu, Qizhai Li, Bergen andrew W., Ruth M. Pfeiffer, Philip S. Rosenberg, Neil Caporaso, Peter Kraft, and Nilanjan Chatterjee. Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology*, 33(8):700–709, 2009. Citations: § 21 and 29

[64] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, London, 1994. Citations: § 21 and 29

[65] Benoit Liquet and Jérémie Riou. Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 13(1):75, 2013. Citations: § 21 and 29

[66] P. Good. *Permutation Tests: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New-York, 2000. Citations: § 21 and 29

[67] Xihong Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997. Citations: § 22

[68] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011. Citations: § 22

[69] I. Ionita-Laza, S. Lee, V. Makarov, J. Buxbaum, and X. Lin. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92:841–853, 2013. Citations: § 22

[70] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012. Citations: § 22

[71] Andriy Derkach, Jerry F Lawless, and Lei Sun. Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic epidemiology*, 37(1):110–121, 2013. Citations: § 22

[72] Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly. Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322, 2011. Citations: § 23

[73] Yihui Luan and Hongzhe Li. Group additive regression models for genomic data analysis. *Biostatistics*, 9(1):100–113, 2007. Citations: § 23

[74] Nilanjan Chatterjee, Zeynep Kalaylioglu, Roxana Moslehi, Ulrike Peters, and Sholom Wacholder. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *The American Journal of Human Genetics*, 79(6):1002–1016, 2006. Citations: § 23

[75] Jinying Zhao, Eric Boerwinkle, and Momiao Xiong. An entropy-based statistic for genomewide association studies. *The American Journal of Human Genetics*, 77(1):27–40, 2005. Citations: § 23

[76] B Shahbaba, C M Shachaf, and Z Yu. A pathway analysis method for genome-wide association studies. *Statistics in Medicine*, 31:988–1000, 2012. Citations: § 23

[77] M. Evangelou, F. Dudbridge, and L. Wernisch. Two novel pathway analysis methods based on a hierarchical model. *Bioinformatics*, 30(5):690–697, 2014. Citations: § 23

[78] Indranil Mukhopadhyay, Eleanor Feingold, Daniel E Weeks, and Anbupalam Thalamuthu. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic epidemiology*, 34(3):213–221, 2010. Citations: § 23

[79] Jung-Ying Tzeng, B Devlin, Larry Wasserman, and Kathryn Roeder. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *The American Journal of Human Genetics*, 72(4):891–902, 2003. Citations: § 23

[80] Daniel J Schaid, Shannon K McDonnell, Scott J Hebbring, Julie M Cunningham, and Stephen N Thibodeau. Nonparametric tests of association of multiple genes with human disease. *The American Journal of Human Genetics*, 76(5):780–793, 2005. Citations: § 23

[81] Jennifer Wessel and Nicholas J Schork. Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79(5):792–806, 2006. Citations: § 23

[82] Camilo Broc, Marina Evangelou, Thérèse Truong, and Jérémie Riou. Investigating gene- and pathway-environment interaction analysis approaches. *Journal de la Société Française de Statistique*, 2018. Citations: § 27

[83] Ruth Ottman. Gene–environment interaction: definitions and study design. *Preventive medicine*, 25(6):764–770, 1996. Citations: § 27

[84] Taye H Hamza, Honglei Chen, Erin M Hill-Burns, Shannon L Rhodes, Jennifer Montimurro, Denise M Kay, Albert Tenesa, Victoria I Kusel, Patricia Sheehan, Muthukrishnan Eaaswarkhanth, et al. Genome-wide gene-environment study identifies glutamate receptor gene grin2a as a parkinson's disease modifier gene via interaction with coffee. *PLoS genetics*, 7(8):e1002237, 2011. Citations: § 27

[85] Montserrat García-Closas, Núria Malats, Debra Silverman, Mustafa Dosemeci, Manolis Kogevinas, David W Hein, Adonina Tardón, Consol Serra, Alfredo Carrato, Reina García-Closas, et al. Nat2 slow acetylation, gstm1 null genotype, and risk of bladder cancer: results from the spanish bladder cancer study and meta-analyses. *The Lancet*, 366(9486):649–659, 2005. Citations: § 27

[86] Nathaniel Rothman, Montserrat Garcia-Closas, Nilanjan Chatterjee, Nuria Malats, Xifeng Wu, Jonine D Figueroa, Francisco X Real, David Van Den Berg, Giuseppe Matullo, Dalsu Baris, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature genetics*, 42(11):978, 2010. Citations: § 27

[87] Mia Hashibe, James D McKay, Maria Paula Curado, Jose Carlos Oliveira, Sergio Koifman, Rosalina Koifman, David Zaridze, Oxana Shangina, Victor Wünsch-Filho, Jose Eluf-Neto, et al. Multiple adh genes are associated with upper aerodigestive cancers. *Nature genetics*, 40(6):707, 2008. Citations: § 27

[88] Natalie E Allen, Colleen G Canning, Catherine Sherrington, Stephen R Lord, Mark D Latt, Jacqueline CT Close, Sandra D O'Rourke, Susan M Murray, and Victor SC Fung. The effects of an exercise program on fall risk factors in people with parkinson's disease: a randomized controlled trial. *Movement Disorders*, 25(9):1217–1225, 2010. Citations: § 27

[89] PG Smith and NE Day. The design of case-control studies: the influence of confounding and interaction effects. *International journal of epidemiology*, 13(3):356–365, 1984. Citations: § 28

[90] P.S. Albert, D. Ratnasinghe, J. Tangrea, and Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154:687–693, 2001. Citations: § 28

[91] Nilanjan Chatterjee and Raymond J Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92(2):399–418, 2005. Citations: § 28

[92] Bhramar Mukherjee and Nilanjan Chatterjee. Exploiting gene-environment independence for analysis of case–control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–694, 2008. Citations: § 28

[93] Dalin Li and David V Conti. Detecting gene-environment interactions using a combined case-only and case-control approach. *American journal of epidemiology*, 169(4):497–504, 2008. Citations: § 28

[94] Xinyi Lin, Seunggeun Lee, David C Christiani, and Xihong Lin. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681, 2013. Citations: § 28

[95] Peng-Lin Lin, Ya-Wen Yu, and Ren-Hua Chung. Pathway analysis incorporating protein-protein interaction networks identified candidate pathways for the seven common diseases. *PloS one*, 11(9):e0162910, 2016. Citations: § 28

[96] Marina Evangelou, Augusto Rendon, Willem H Ouwehand, Lorenz Wernisch, and Frank Dudbridge. Comparison of methods for competitive tests of pathway analysis. *PloS one*, 7(7):e41018, 2012. Citations: § 29

[97] Y. Su, W. J. Gauderman, K Berhane, and J. P. Lewinger. Adaptive set-based methods for association testing. *Genetic Epidemiology*, 40:113–122, 2016. Citations: § 29

[98] Petra Buzkova, Thomas Lumley, and Kenneth Rice. Permutation and parametric bootstrap tests for gene–gene and gene–environment interactions. *Annals of human genetics*, 75(1):36–45, 2011. Citations: § 29

[99] Thérèse Truong, Benoît Liquet, Florence Menegaux, Sabine Plancoulaine, Pierre Laurent-Puig, Claire Mulot, Emilie Cordina-Duverger, Marie Sanchez, Patrick Arveux, Pierre Kerbrat, et al. Breast cancer risk, nightwork, and circadian clock gene polymorphisms. *Endocrine-related cancer*, 21(4):629–638, 2014. Citations: § 29 and 89

[100] Benoit Liquet, Therese Truong, and Camilo Broc. *PIGE: Self Contained Gene Set Analysis for Gene- And Pathway-Environment Interaction Analysis*, 2017. R package version 1.1. Citations: § 30

[101] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007. Citations: § 61

[102] Roman Hornung, Anne-Laure Boulesteix, and David Causeur. Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. *BMC bioinformatics*, 17(1):27, 2016. Citations: § 61

[103] Florian Rohart, Aida Eslami, Nicholas Matigian, Stéphanie Bougeard, and Kim-Anh Le Cao. Mint: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC bioinformatics*, 18(1):128, 2017. Citations: § 61

[104] Jennifer Listgarten, Carl Kadie, Eric E Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470, 2010. Citations: § 61

[105] Andrew H Sims, Graeme J Smethurst, Yvonne Hey, Michal J Okoniewski, Stuart D Pepper, Anthony Howell, Crispin J Miller, and Robert B Clarke. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1(1):42, 2008. Citations: § 61

[106] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012. Citations: § 61

[107] Samsiddhi Bhattacharjee, Preetha Rajaraman, Kevin B Jacobs, William A Wheeler, Beatrice S Melin, Patricia Hartge, Meredith Yeager, Charles C Chung, Stephen J Chanock, Nilanjan Chatterjee, et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90(5):821–835, 2012. Citations: § 62

[108] Arunabha Majumdar, Tanushree Haldar, Sourabh Bhattacharya, and John S Witte. An efficient bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS genetics*, 14(2):e1007139, 2018. Citations: § 63

[109] Seunggeun Lee, Tanya M Teslovich, Michael Boehnke, and Xihong Lin. General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics*, 93(1):42–53, 2013. Citations: § 63

[110] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. Citations: § 63

[111] Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010. Citations: § 63, 64, and 65

[112] Camilo Broc, Borja Calvo, and Benoit Liquet. Penalized partial least square applied to structured data. *Arabian Journal of Mathematics*, pages 1–16, 2019. Citations: § 64

[113] Camilo Broc, Benoit Liquet, Thérèse Truong, and Calvo Borja. Sparse group pls on data gathered from different studies : application to experimental bias correction and pleiotropy analysis. In *Journée de la SFDS*. Société Française de Statiques, 2017. Citations: § 64

[114] Camilo Broc, Benoit Liquet, Thérèse Truong, and Calvo Borja. Joint-lasso applied to sparse group partial least square and application to pleiotropy. In *Journée de la SFDS*. Société Française de Statiques, 2019. Citations: § 64

[115] Benoit Liquet, Pierre Lafaye de Micheaux, and Camilo Broc. *Sparse Group Partial Least Square Methods*, 2017. R package version 1.7. Citations: § 65

[116] Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics*, 48(7):709, 2016. Citations: § 87

[117] Lori C Sakoda, Eric Jorgenson, and John S Witte. Turning of cogs moves forward findings for hormonally mediated cancers. *Nature genetics*, 45(4):345, 2013. Citations: § 88

[118] S Hong Lee, Stephan Ripke, Benjamin M Neale, Stephen V Faraone, Shaun M Purcell, Roy H Perlis, Bryan J Mowry, Anita Thapar, Michael E Goddard, John S Witte, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics*, 45(9):984, 2013. Citations: § 88

[119] George C Williams. Pleiotropy, natural selection, and the evolution of senescence. *evolution*, pages 398–411, 1957. Citations: § 88

[120] Rayjean J Hung, Cornelia M Ulrich, Ellen L Goode, Yonathan Brhane, Kenneth Muir, Andrew T Chan, Loic Le Marchand, Joellen Schildkraut, John S Witte, Rosalind Eeles, et al. Cross cancer genomic investigation of inflammation pathway for five common cancers: lung, ovary, prostate, breast, and colorectal cancer. *Journal of the National Cancer Institute*, 107(11):djv246, 2015. Citations: § 88

[121] Sarah M Nielsen, Michael G White, Susan Hong, Briseis Aschebrook-Kilfoy, Edwin L Kaplan, Peter Angelos, Swati A Kulkarni, Olufunmilayo I Olopade, and Raymon H Grogan. The breast–thyroid cancer link: a systematic review and meta-analysis. *Cancer Epidemiology and Prevention Biomarkers*, 25(2):231–238, 2016. Citations: § 89

[122] Marie Lof, Leena Hilakivi-Clarke, Sven Sandin, Sonia De Assis, Wei Yu, and Elisabete Weiderpass. Dietary fat intake and gestational weight gain in relation to estradiol and progesterone plasma levels during pregnancy: a longitudinal study in swedish women. *BMC women's health*, 9(1):10, 2009. Citations: § 89

[123] Constance Xhaard, Carole Rubino, Enora Cléro, Stéphane Maillard, Yan Ren, Françoise Borson-Chazot, Geneviève Sassolas, Claire Schvartz, Marc Colonna, Brigitte Lacour, et al. Menstrual and reproductive factors in the risk of differentiated thyroid carcinoma in young women in france: a population-based case-control study. *American journal of epidemiology*, 180(10):1007–1017, 2014. Citations: § 89

[124] Laura Fachal and Alison M Dunning. From candidate gene studies to gwas and post-gwas analyses in breast cancer. *Current opinion in genetics & development*, 30:32–41, 2015. Citations: § 89

[125] Kari Hemminki and Xinjun Li. Familial risk of cancer by site and histopathology. *International journal of cancer*, 103(1):105–109, 2003. Citations: § 89

[126] Gisella Figlioli, Bowang Chen, Rossella Elisei, Cristina Romei, Chiara Campo, Monica Cipollini, Alfonso Cristaudo, Franco Bambi, Elisa Paolicchi, Per Hoffmann, et al. Novel genetic variants in differentiated thyroid cancer and assessment of the cumulative risk. *Scientific reports*, 5:8922, 2015. Citations: § 89

List of Figures

List of Tables