



Contributions to biostatistics: categorical data analysis, data modeling and statistical inference

Mathieu Emily

► To cite this version:

Mathieu Emily. Contributions to biostatistics: categorical data analysis, data modeling and statistical inference. Mathematics [math]. Université de Rennes 1, 2016. tel-01439264

HAL Id: tel-01439264

<https://hal.science/tel-01439264>

Submitted on 18 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HABILITATION À DIRIGER DES RECHERCHES

Université de Rennes 1

**Contributions to biostatistics:
categorical data analysis, data modeling
and statistical inference**

Mathieu Emily

November 15th, 2016

Jury:

Christophe Ambroise	Professeur, Université d'Evry Val d'Essonne, France	Rapporteur
Gérard Biau	Professeur, Université Pierre et Marie Curie, France	Président
David Causeur	Professeur, Agrocampus Ouest, France	Examineur
Heather Cordell	Professor, University of Newcastle upon Tyne, United Kingdom	Rapporteur
Jean-Michel Marin	Professeur, Université de Montpellier, France	Rapporteur
Valérie Monbet	Professeur, Université de Rennes 1, France	Examineur
Korbinian Strimmer	Professor, Imperial College London, United Kingdom	Examineur
Jean-Philippe Vert	Directeur de recherche, Mines ParisTech, France	Examineur

Pour M.H.A.M.

Remerciements

En tout premier lieu je tiens à adresser mes remerciements à l'ensemble des membres du jury qui m'ont fait l'honneur d'évaluer mes travaux. Je remercie Hearther Cordell, Christophe Ambroise et Jean-Michel Marin pour le temps qu'ils ont consacré à la lecture de mon manuscrit. Merci également à Gérard Biau d'avoir présidé mon jury, à David Causeur, Valérie Monbet, Jean-Philippe Vert et Korbinian Strimmer pour leurs questions et commentaires précieux lors de la soutenance. Je vous exprime également toute ma gratitude pour votre disponibilité qui a permis de vous réunir tous physiquement le 15 novembre.

Les travaux présentés dans ce manuscrit représentent pour moi dix années de travail mais également, et surtout, dix années de rencontres et collaborations. Cette aventure a commencé un mois de novembre de 2006 où, juste après ma soutenance de thèse, j'ai pris la direction du Danemark pour y effectuer mon post-doctorat. Je remercie ici toutes les personnes qui ont fait de ces 18 mois à Aarhus une expérience unique tant sur le plan professionnel que personnel : merci à Mikkel, Thomas M., Søren et Bo pour m'avoir accompagné sur la voie des études d'association ainsi qu'à Thomas B., Carsten, Frank, Per et plus spécialement Julien pour le poker, les colons de catane et tant d'autres choses.

Pour mon retour en Bretagne, j'ai eu la chance d'avoir été accueilli au sein de l'équipe de Statistiques de l'IRMAR dont je remercie tous les membres passés et présents. J'adresse notamment un grand merci à Chantal pour son efficacité et sa disponibilité ainsi qu'à Hélène notamment pour la gestion de mes projets. Je remercie tout particulièrement les "statisticiens" de l'ex département MASS de l'université Rennes 2 qui m'ont fait confiance lors de mon recrutement en tant que maître de conférences. Merci tout particulièrement à Alain pour l'ensemble de son travail pour le Master, Arnaud pour son efficacité, Bruno pour avoir partagé mon bureau, Dominique pour sa gestion de la composante Rennes 2 de l'IRMAR, Eric pour son dynamisme hors du commun, Jacques pour sa diplomatie, Pierre-André pour son soutien, Laurent et Nicolas pour leur aide à mon arrivée et pour m'avoir fait rechausser les crampons de foot! Je tiens également à remercier Annabelle, Marie-Laure et Nelly qui m'ont assisté dans mes responsabilités administratives à Rennes 2 toujours dans la joie et la bonne humeur!

Depuis 2013, j'ai rejoint l'Unité Pédagogique de Mathématiques Appliquées du département Statistique et Informatique d'Agrocampus Ouest. Bien que la distance entre l'université Rennes 2 et Agrocampus Ouest soit petite, ce changement d'établissement de rattachement constitue indéniablement un tournant dans mon activité professionnelle. Je suis donc extrêmement reconnaissant au Professeur Pagès et à l'ensemble des personnes de l'UP pour leur accueil à mon arrivée. Merci également à Julie, Lucie, aux "anciens doctorants" (Tham, Vincent et Emeline) et aux "nouveaux" (Margot et Florian) d'avoir partagé de nombreux moments de mon quotidien (souvent agréments)

d'un café). Merci plus particulièrement à nos assistantes gestionnaires, Karine, Héléna et Elisabeth, pour leur efficacité et disponibilité ainsi que pour œuvrer à conserver cet esprit familial qui règne dans l'UP, et ce, toujours avec un grand sourire! C'est en totale confiance (ce qui est assez rare pour moi!) que je vous ai laissées mettre la touche finale à mon "pot déjeunatoire" qui fut, de mon point de vue, une grande réussite! Mille mercis à Magalie pour toutes ses petites choses qui rythment les relations entre collègues, notamment sa simplicité et son implication dans la vie de l'UP. Merci également à Sébastien pour nos échanges sur la pédagogie, la recherche et tout le reste! Merci à François pour son sens des responsabilités, son franc soutien et ses précieux conseils qui ont suivi ma pré-soutenance ainsi que pour son accessibilité. Merci enfin à David pour, d'une part, toute l'énergie qu'il peut mettre dans la direction du département et de l'UP et surtout d'autre part les nombreux échanges que nous avons pu avoir et qui ont levé mes doutes quant à ma capacité à obtenir ce diplôme. Durant ces derniers mois, j'ai senti une équipe autour de moi concernée par cet événement (pour autant si personnel) et avec qui je pouvais à tout moment échanger, ce qui est inestimable.

Je tiens également à remercier l'ensemble des étudiants que j'ai (co-)encadrés: Thomas, Anthony, Hillel, Floriane, Emeline, Florian K. et Florian H.. Je remercie également toutes les personnes avec qui j'ai eu la chance de collaborer ces dernières années. Je ne remercierai jamais assez Olivier d'avoir accepté d'encadrer ma thèse à la sortie de mon DEA et pour m'avoir fait découvrir le métier d'enseignant-chercheur. Merci à Mikkel pour son encadrement au cours de mon post-doc qui m'a permis de construire mon projet de recherche autour des études d'association. Un profond merci à Avner tout d'abord pour notre collaboration extrêmement riche qui m'a beaucoup fait progressé mais également pour tous ses conseils et son aide précieuse lors ma "leçon publique" à Agrocampus Ouest. Un merci particulier à Christophe pour nos nombreuses discussions sur, entre autre, le modèle canin et l'émergence de nouvelles technologies ainsi qu'à Thomas qui a naturellement intégré notre collaboration. Un très grand merci à Chloé notamment pour la réussite de nos travaux, les nombreuses JDS, l'organisation de StatLearn et également son soutien sans faille lors ma "leçon publique" à Agrocampus Ouest. Merci à Radu pour nos échanges sur les données spatialisées. Merci à Christian de m'avoir invité à partager ses recherches sur les protéines amyloïdes. Merci à Alain pour ce joli article à l'interface entre mathématiques et applications. Merci à Maud de partager son environnement (non virtuel) de recherche particulièrement innovant et tellement fascinant. Merci également à Amélie et Alexis de m'avoir permis d'ouvrir mon activité à d'autres domaines d'application comme la peptidomique et la physio-pathologie. Merci aussi à Magalie pour notre découverte commune de Bioconductor qui est loin de nous avoir livré tous ses secrets. Merci à David pour nos nombreux échanges scientifiques sur la modélisation de la dépendance et pour avoir partagé sa vision de la recherche.

La route de l'habilitation est parfois longue et sinueuse et il est important d'avoir autour de soi des personnes qui vous permettent de garder un cap à peu près fixe! Parmi ces personnes, je tiens à remercier tout particulièrement ma "belle-famille" (Isabelle, Philippe, Vincent et Olivier) notamment pour leur présence (même à distance pour Vincent) à ma soutenance. Merci également à mon frère, Floriane, Benjamin et Nathan qui suivent de plus loin ma route professionnelle. Merci bien sûr à mes parents qui sont toujours là quand il le faut. Leur soutien pendant toutes ces années est un moteur important pour avancer.

Enfin, mes derniers remerciements vont à ma "petite" famille qui n'a cessé de s'agrandir :). Avec toute mon affection, je pense tout d'abord aux trois jambes de mon triskell : Hugo, Arthur et Maxime. Trois jambes qui me font courir toujours plus vite sous leurs applaudissements, trois jambes pleines de vie qui m'apportent ce quelque chose d'indescriptible dans le fait de se faire appeler "papa", trois jambes sur lesquelles je peux toujours poser mes pensées en quête de stabilité.

Ces trois jambes rayonnent autour du soleil de mon existence : Maud. Il me faudrait beaucoup plus qu'un paragraphe de remerciements pour exprimer tout ce que je te dois. En quelques mots, tu m'as tout d'abord montré la voie vers l'HDR il y a deux ans. Depuis tu n'as cessé de me pousser dans cette direction notamment en participant activement à l'écriture de mes articles. Et enfin tu as donné à ce manuscrit toute la profondeur qu'il a aujourd'hui. A ce niveau, ce n'est plus d'une aide précieuse ni d'un soutien sans faille dont il faut parler mais d'un véritable travail d'équipe! En plus (et surtout), tu arrives à supporter mes humeurs souvent difficiles (comme lors des 6 derniers mois). Avec tout mon ♥, je veux te remercier simplement pour ce que tu es.

Contents

1	Introduction	1
1.1	Biological data feeds and needs the biostatistician	3
1.2	Statistical research challenges	5
1.3	Illustration of my research challenges	6
1.4	Research axes and contributions	10
2	Analysis of categorical data	13
2.1	Introduction and background	13
2.2	Statistical power for single testing in case-control association	15
2.3	Interaction in three-way contingency tables	21
2.4	Clustering in sparse two-way contingency tables	26
3	Statistical modeling of highly structured data	33
3.1	Introduction	33
3.2	Variable selection in the design of experiment	34
3.3	Multiple testing correction based on the interactome	40
3.4	Combining statistical tests	44
3.5	Meta-prediction	50
4	Probabilistic modeling and statistical inference	57
4.1	Introduction	57
4.2	Clustering in 2D spatial point process	58
4.3	Statistical inference in 2D spatial marked point process	65
4.4	Risk analysis in public health	70
5	Conclusion and perspectives	75
A	On the use of mixed models in life science	79
A.1	Introduction and motivation	79
A.2	Application to physiopathology, virtual reality and peptidomics	80
A.3	Perspectives	82
B	Scientific production	85
	Bibliography	89

Introduction


“We must be careful not to confuse data with the abstractions we use to analyze them.”
William James

I would like to start this thesis with a brief overview of my scientific career.

I began my scientific education in Brest in 1997 where I studied the main concepts of mathematics and physics. In 2000, I was admitted to the Ecole Nationale d’Informatique et de Mathématiques Appliquées de Grenoble (ENSIMAG) in order to pursue my education in applied mathematics and statistics. There, I graduated and obtained a master of engineering degree in applied mathematics and computer science in 2003. During that period, I was more and more attracted by studying problems arising in the “real” world with mathematical models. In 2002, when I started my last year at ENSIMAG, I was totally convinced that the application and the development of dedicated mathematical frameworks, and especially statistical frameworks, can help solving issues in other fields.

I therefore took the opportunity to consolidate my formation in applied statistics by following in parallel to my engineering degree a Master of Science in Applied Mathematics with a specialty in Statistics at the University of Grenoble (ex. Université Joseph Fourier). During that year, I developed a special feeling with problems and data encountered in biology and medical science. When I had to conclude my degree with an internship, I was naturally tempted by the most applied topic dealing with “classification of histological slides from breast cancer tissues” and spent 4 months in the LabSAD (Laboratoire de statistique et analyse de données) in Grenoble under the supervision of Etienne Bertin. During my internship, I had to find a quantitative indicator of the spatial organization of cells within a tissue that can discriminate between aggressive and non-aggressive breast tumors. It was an exciting experience during which I learnt a lot about the biological processes involved in tumorigenesis and tried to formalize the biological knowledge into a statistical framework. I developed a taste for giving a special care to the type and nature of available data. I also really appreciated working in close collaboration with biologists from our non-academic partner TriPath Imaging, who kindly shared the data. At the end of that year, I decided to start a PhD in order to engage my professional life in the research field.

I began my PhD in 2003, co-supervised by Olivier François (Professor at ENSIMAG - TIMC-IMAG) and Jean-Michel Billiot and Remy Drouilhet (Associated professors at the Université de Grenoble - LJK) on the statistical modeling of tumors. In my PhD, I considered two main statistical problems arising during the analysis of cancerous data. The first problem was to account for genomic data in the estimation of the age of the tumor and the second problem dealt with the modeling and the estimation of cell interaction in a living tissue. My PhD allowed me to acquire

skills in probabilistic modeling, estimation theory and implementation of statistical tools with the learning of the computer language and environment . My PhD confirmed my interest in (1) the knowledge of the biological mechanisms underlying the complexity of life, (2) the basic understanding of the biotechnological processes that generate observations to use an appropriate modeling of the data type and (3) the probabilistic modeling of such data in order to propose dedicated inference procedures.

The second part of my scientific career began just after my PhD and the research covered in this manuscript goes from this point to the present. After my PhD, I decided to take a post-doctoral position at the Bioinformatics Research Center (BiRC) belonging to the University of Aarhus in Denmark. I was motivated by the challenges of analyzing high-dimensional data and the opportunity to improve our knowledge on genes involved in breast and prostate cancer. I was part of an ambitious project that was granted by a European project where 4 main partners were involved: the University of Aarhus (Denmark), the University of Oxford (United Kingdom), the Radboud University of Nijmegen (the Netherlands) and DeCODE (Iceland). My goal was to design a statistical framework to test for an association between the development of a disease and the interaction between two genes in Genome-Wide Association Studies (GWAS). There I found in GWAS a thrilling interdisciplinary field of research where genetical, computational and statistical aspects, as well as their interactions, are crucial. In term of research theme, I was greatly introduced to genetical aspects by Mikkel Schierup (Professor of Bioinformatics at the University of Aarhus) and computational aspects by Thomas Mailund (Associate Professor of Bioinformatics at the University of Aarhus) which allowed me to develop an efficient method. I combined biological knowledge and computational techniques to propose a statistical procedure for interaction testing that is feasible at the genome scale and that accounts for multiple testing issues encountered in high-dimensional data.

After two years, I succeeded in obtaining a tenured position during autumn 2008 as an associate professor at the University of Rennes 2 and in the statistical team of the Institut de Recherche Mathématiques de Rennes (IRMAR). During my post-doc, I found in GWAS an extremely stimulating research area: the biological questioning and the nature of the data themselves raise new challenges regarding statistical modeling with fundamental applications in fields as diverse as agronomy or medicine. I therefore pursued my research in the analysis of GWAS through a statistical modeling approach with the supervision of several MSc. internships. I have also started new collaborations. With David Causeur (Professor at Agrocampus Ouest), we worked on the modeling of the dependence in high-dimensional data (with the co-supervision of MSc internships). I also worked with Chloé Friguet (Associate professor at the University of Bretagne-Sud) on the power of association tests. With Alain Mom (Associate Professor at the University of Rennes 2), we tackled the issue of classification in sparse contingency tables. I have also been concerned by the modeling of spatial data regarding the inference of interaction point processes with Radu Stoica (Associate Professor at the University of Lille 1) and the clustering of points with respect to covariates in collaboration with Avner Bar-Hen (Professor at the University of Paris Descartes).

When I arrived in Rennes, I have also diversified my fields of application by starting new collaborations with researchers in fields at the frontiers with statistics. I have been working with Christophe Hitte (Researcher at the University of Rennes 1) on the genetics of the Domestic Dogs by tackling issues regarding association studies, selection and regulation leading to the co-supervision of several MSc. internships. I have also been working with Christian Delamarche (Professor at the University of Rennes 1) on the prediction of amyloid fibers by supervising several MSc. internships and with Maud Marchal (Associate Professor at the Institut National des Sciences Appliquées in Rennes) on the statistical modeling, with mixed models, of the human perception in virtual reality. In 2013, I decided to apply for an associate professor position at Agrocampus Ouest. Although this change was not motivated by an evolution in my research top-

ics, it offered me the opportunity to start new collaborations in nutrition (with the co-supervision of a MSc. student with Amélie Deglaire (Associate Professor at Agrocampus Ouest) and in physiopathology with Alexis Le Faucheur (Associate Professor at the Ecole Nationale Supérieure in Rennes), where we designed appropriate linear mixed models. Furthermore, it has strengthened my collaborations with David Causeur (Professor at Agrocampus Ouest) thus leading to the co-supervision of a PhD thesis starting in autumn 2016.

To conclude, I define myself as a biostatistician with a consideration for all steps of a biological experiment, going from the design to the interpretation of the results through the modeling, the analysis and the inference of the collected data. In all my research projects, I used a similar methodology as a common denominator of my work. At first, I am investigating the applied question that I want to tackle by focusing on the nature and the type of data related to the problem. In a second step, I am focusing on the statistical or probabilistic modeling of such data by drawing statistical hypotheses that correspond (as close as possible) to the initial question. Then I am putting efforts on developing statistical procedures (inference, simulation, etc.) to test for statistical hypotheses. Such an “hypothesis and data driven” approach characterizes the way I enjoy doing research.

The remainder of this introductory chapter provides motivations for my research work in the field of biostatistics. First, I briefly describe the major role played by biological data in the birth and evolution of statistics. In the lights of the revolution in the nature of biological data, I exhibit some important statistical challenges associated with such “modern” data. I then present the main themes of my research and set them in the landscape of biostatistics.

1.1 Biological data feeds and needs the biostatistician

Although statistics interplay with all data-oriented fields, the relationship between statistics and biology is specific as biological data has always been a catalyzer for many advances in statistics.

1.1.1 The forward-backward algorithm of biostatistics

Biology is one of the most prolific field in generating data. Even, in the pre-era of statistics, the study of biological issues has lead researchers to collect, summarize and visualize data giving significant insights into the processes involved in the raised biological questions and helping researchers to solve them. For instance, some of the major historical advances in epidemiology are associated with (1) the **design of experiment** driven by statistical considerations [Lind, 1753], (2) the introduction of **time-to-event data** [Graunt, 1662] and (3) the characterization of **spatial data** [Snow, 1855]. These three issues are still generating an abundant literature in statistics and are tackled in this manuscript (Sections 2.2, 3.2, 4.2, 4.3 and 4.4)

At the end of the 19th century, the evolution of biological data, notably in terms of size and **heterogeneity in data type**, had lead to the birth of (bio)statistics as a rigorous mathematical discipline used for systematic analysis. Since, statistics and biology have shared a common lineage thoroughly illustrated by the contributions of R.A. Fisher [Fisher, 1922, Fisher, 1925, Fisher, 1935]. One the most important contribution of Fisher’s work was to begin a systematic approach of the analysis of real data as the springboard for the development of new statistical methods. Such a data-driven approach allowed Fisher to establish the foundation **statistical hypothesis testing** [Fisher, 1935] and to raise several statistical issues tackled throughout the manuscript (**Analysis of contingency tables** [Fisher, 1922], **meta-analysis** and **combination of tests** [Fisher, 1925]).

One of the biggest challenge of biostatistics concerns the design and the application of appropriate statistical methods that can handle the variety and the complexity of biological data types.

As such, I consider the relationship between statistics and biology as an iterative forward-backward algorithm. In a forward step, data are generated using biotechnology to tackle a specific biological question. During the backward step, statisticians develop statistical methods that take into account the biological constraints. Driven by biological applications, challenges for biostatisticians are, from my point-of-view, to unlock the scope for biological advancement and start with a new forward step. Research challenges in biostatistics are therefore associated with the characteristics of the biological data used to tackle modern biological issues.

1.1.2 The characteristics of modern biological data

The growth of knowledge in biostatistics is accompanied by the growth of biological data. The evolution of biological data, especially in epidemiology and public health, has been paved by two main technological revolutions: the emergence of computers and the advent of high-throughput technologies.

The impact of computers can indeed be observed at various steps of the statistical pipeline, going from the collection and storage of the data to the development of new statistical testing [Speed, 1985]. First, it can be remarked that data are now (almost) always collected automatically, thus favoring statistics to globally improve the quality of the data [Kettenring et al., 2003]. Next, the exponential storage capacity of computers, together with the increase of the speed, have allowed the analysis of large datasets. Finally, the advent of modern computer technology and relatively cheap computing resources have enabled computer-intensive biostatistical methods like bootstrapping and resampling methods.

The high-throughput technologies developed in the last decade have revolutionized the speed of data accumulation in the life sciences. Such a technology has been developed at various levels of living organisms with the design of biological chips (or arrays) as for example SNP arrays, DNA microarrays, protein and peptide arrays or tissue microarrays. These four examples of high-throughput technologies generate heterogeneous types of data that are under consideration in this manuscript. In more details, at the smallest scale of a single site in DNA, SNP arrays have been developed to detect polymorphisms within a population. Latest generation of SNP arrays allows the qualitative measure of genotypes of almost one million of Single Nucleotide Polymorphism (SNP) per individual. At the gene level, DNA microarrays allow the quantitative measure of expression levels of tens of thousands of genes simultaneously in a biological tissue. At the functional level of genes, protein arrays have been used to track the interactions and activities of a large number of proteins in parallel. The use of peptide chips can further be used to investigate the kinetics of protein-protein interactions. Mass spectrometry based-proteomics and peptidomics have also largely contributed to the huge expansion of data dealing with functional genes. Our final example are Tissue Microarrays (TMA) that have been proposed to study simultaneously the expression molecular targets at the DNA, mRNA, and protein levels. The use of paraffin in TMAs allows the observation of the spatial distribution of expression measures within a tissue. TMAs provides a practical and effective tool helping the identification of new diagnostic and prognostic markers and targets in human cancers for example. Figure 1.1 provides examples of the data collected with three different high-throughput technologies and reveals the **heterogeneity in the statistical nature of collected data**.

Therefore, the rapid evolution of biotechnologies has allowed the collection of massive amount of data in various fields of biology. The common motivation for collecting more data is a hope for a better understanding of the underlying processes that rule the observed biological systems. We indeed now have in hands very rich and complex data that hold great promises to solve many complex biological questions.

However, the data evolution has nowadays raised more questions than answers due to the ever

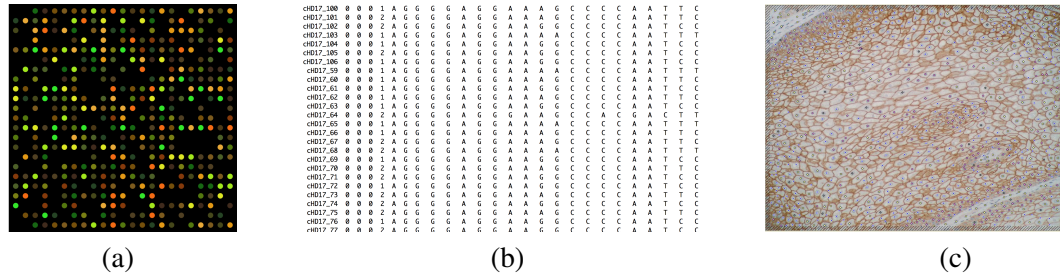


Figure 1.1: *Illustration of collected data with high throughput technologies: (a) gene expression levels using DNA microarray (b) genotypes using SNP array and (c) spatial organization of a tissue using a Tissue MicroArray (TMA).*

growing complexity of the typology of “modern” biological data [Wooley, 2006]. The complexity of the data can be observed either in the set of explanatory variables (also known as independent variables, predictors, regressors, exposure variables, risk factors, features or input variables in the various statistical contexts) or in the set of response variables (also known as dependent variables, predicted variables, measured variables, explained variables or output variables in other contexts) or even in the set of individuals (*i.e.* the objects described by the variables that can be people, animals things, etc.). Modern biological data are indeed characterized by variables (1) observed at various scale of the organism that are known to interact in complex biological systems, (2) collected through many different technologies thus requiring normalization and integration prior to the statistical analysis, (3) of many types such as: sequences data, graphs, patterns, images, spatial data, temporal data, etc. Another important aspect of the data is the set of individuals used as a statistical sample of the studied population. In some area, large public investments have contributed to the collection of large sample sizes that can contain 10,000 to 100,000 of individuals.

All these characteristics have raised many statistical challenges that are inherent to the nature of data, in terms of **statistical data type** and **correlation structure** of the data. In the next section, I will describe the main challenges I have tackled in my research and that are considered in this manuscript.

1.2 Statistical research challenges

In this manuscript, we define the research context of our work based on the four following challenges.

- **Designing powerful experiments for addressing biological challenges**

For a long time, experimental design is known to be the first step of a statistical analysis, however data revolution has deeply modified the practice of the **design of experiment** [Fisher, 1935]. In the current omics era, the collection of data is often not driven by a precise biological question. Important issues for the biostatistician are (1) to account for the experimental bias due to the gap between data collection and biological question and (2) to determine what is needed as variables (or measures) to statistically address the biological hypothesis.

A common and central question is the optimization of the **statistical power**. Sample size estimation is important at the design stage to ensure a sufficient statistical power to address the stated objective. Statistical power should also be included in the criteria used to choose the best statistical procedure. Furthermore, since biostatisticians aim at providing biologists

with tools for extracting meaning from those data, they should also guide them towards reasonable hypothesis testing.

- **Modeling data type**

The fundamental modification of biological data due technological advances has raised many statistical issues. These issues are also related to the nature of the data that come in many heterogeneous **data types**. In situations such as in microarray experiment, the expression levels of large numbers of genes is measured, thus corresponding to traditional *continuous* variables [Pease et al., 1994]. However in other situations, biostatistical analysis fall into the field of categorical data analysis. For instance, in epidemiology, collected data are usually *categorical* since scientists are interested in the analysis of the causes and effects of disease conditions compared to health conditions [Balding, 2006]. The “choice” of a data type plays a major role throughout the overall statistical pipeline of analysis and relies on biological assumptions. Interpreting the statistical results through the prism of the underlying biological assumptions is certainly the main issue of the modern biostatistics.

- **Formalizing biologically relevant statistical hypotheses**

Although data revolution has seen the emergence and/or the application of complementary statistical tools, such as Bayesian statistics, **hypothesis testing** is still widely used in many biological fields. As the complexity of various biological systems is still unknown, it is necessary to provide machine-readable representations of analytic and theoretical results as well as the inferential procedures that lead to various hypotheses [Wooley, 2006]. It is therefore crucial to address biological questions with well-defined statistical hypotheses. The accumulation of data has generated a tendency to let the data speak for themselves leading to an overfitting of the data in many situations and driving the issue of over-optimism in the biostatistical research [Boulesteix, 2010, Jelizarow et al., 2010].

- **Accounting for the structure of the data**

New biomedical and high-throughput technologies have generated enormous amounts of data that are known to be highly structured. **Data structure** is due to different observation scales, various technologies, plurality and complexity of the biological systems in place, etc. For example, the human genome can be parsed into haplotype blocks defined as sizable regions over which there is little evidence for historical recombination [Gabriel et al., 2002]. Otherwise, real biological entities, from cells to ecosystems, are not spatially homogeneous, and an interesting challenge can be found in understanding how one spatial region is different from another. Thus, spatial relationships must be captured in machine-readable form, and other biologically significant data must be overlaid on top of these relationships [Wooley, 2006]. Biological data are therefore highly structured, resulting in a deviation to the independence. Appropriate modeling of the biological systems have to be proposed to avoid the detection of confounding factors and to increase statistical power with efficient correction for multiple testing.

In the next section, these four research challenges are illustrated in the context of genome-wide associated studies.

1.3 Illustration of my research challenges through the example of genome-wide association studies (GWAS)

Genome-wide association studies (GWAS) have played a central role in my research work. Indeed, data that have arisen in GWAS bring together many of the characteristics shared by modern bio-

logical data and, therefore, I found in GWAS an exciting source of statistical research challenges. The purpose of this section is to provide illustrations of the research challenges I tackle in this manuscript. Therefore, for each challenge, I decided to indicate the corresponding references of my research work in the context of GWAS.

GWAS aim at investigating the genetic variations that are associated with a phenotypic trait. The most common approach of GWAS is the case-control setup, which compares two large groups of individuals, one healthy control group and one case group affected by a disease. Genetic variations can be assayed at the nucleotide level with SNP (Single Nucleotide Polymorphism) arrays, that monitor hundred of thousand of variants that cover the whole genome. The set of individuals is composed by n_0 controls and n_1 cases leading to a sample size of $n_0 + n_1 = n$. Each individual is characterized by one response variable that corresponds to the phenotype (*i.e* the disease status) and by a set of p explanatory variables coding for the genotypes observed at p sites. As displayed in Figure 1.2, data can be displayed by a single vector of phenotype and a $n \times p$ matrix of genotypes.

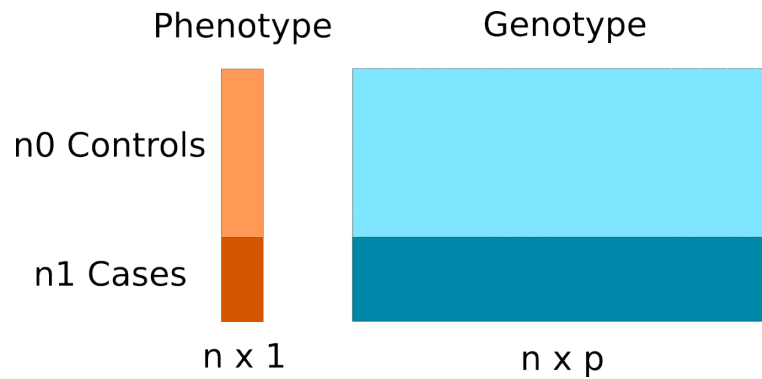


Figure 1.2: *Typical representation of a GWAS data set.*

In the remainder of this section, my four research challenges are detailed in the context of case-control genome-wide association studies and references are given regarding my corresponding publications.

Designing powerful experiment for addressing biological challenges

The design of a GWAS experiment requires the recruitment of the individuals as well as the choice of the explanatory variables (here the SNPs to be genotyped). The optimization of the power of a GWAS therefore involves appropriate choices for these two main characteristics.

Regarding the recruitment of the individuals, although financial constraints usually drive the sample size n , other parameters, such as the case-to-control ratio, n_1/n_0 , (or the balance of the design) is rarely fixed. One challenge is to understand the combined role of these design parameters in power functions in order to guide biologist toward the most powerful association test [JP4, PP3].

Concerning the choice of explanatory variables in a GWAS, it is noteworthy that the number of SNPs in the human genome is estimated to approximately 85 millions. For technological and cost reasons, genotyping all SNPs in a single chip is not feasible. However, as displayed in Figure 1.3(a), contiguous SNPs along the genome are likely to be correlated while the correlation vanishes with the physical distance between two SNPs. This structure of the human genome, known as blocks of Linkage Disequilibrium (LD) structure, raised the issue of selecting a subset of SNPs to be genotyped. Nevertheless, such a selection falls into the challenge of selecting the most informative set of explanatory variables among correlated variables [IC7, NC3].

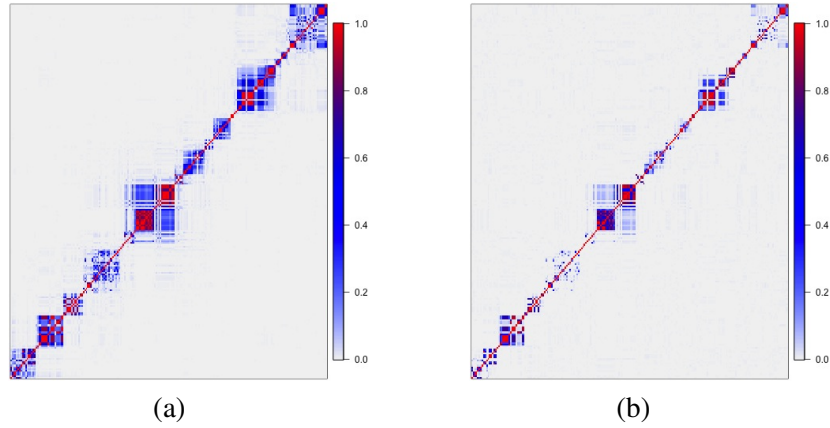


Figure 1.3: *Illustration of the 1-dimensional correlation structure along the human genome. Both figures are obtained with the same set of SNPs within a 1 Megabase region in Chromosome 6 and with the same genotype data from the WTCCC dataset [WTCCC, 2007]. Figure (a) displays the linkage disequilibrium pattern in SNPs and Figure (b) shows the empirical correlation between association tests computed with 1000 simulations of a phenotype under the null hypothesis.*

Modeling heterogeneous data type

In GWAS, explanatory variables are usually characterized by bi-allelic SNPs, *i.e.* variants nucleotide with only two possible nucleobases (or alleles) among the four canonical nucleobases (cytosine, guanine, adenine and thymine). Although such variable is classically modeled by a categorical random variable with 3 possible values in the set $\mathcal{S}_{Cat} = \{\text{Major homozygote, Heterozygote, Minor homozygote}\}$ (or $\mathcal{S}_{Cat} = \{AA, Aa, aa\}$), stating biological hypotheses can modify the probabilistic representation of the variable and thus the data type.

If \mathcal{S}_{Cat} is used as the set of possible values, data observed at a single SNP can be summarized into a 2×3 contingency table (see Table 1.1(a)). However if the assumption of independent pairing of the chromosome is made, it is very tempting to consider alleles rather than genotypes since the sample is doubled when counting alleles. In that case, data can be summarized in a 2×2 contingency table as displayed in Table 1.1(b). Genotypes can also be ordered according to the number of minor alleles so that the set of possible values for the SNP becomes $\mathcal{S}_{Ord} = \{aa < Aa < AA\}$. In that case, it is assumed that the more alleles are carried, the larger the effect size is. Given these three statistical modeling of the variables, one of the main challenge is the evaluation of the impact of the choice of the data type in the power association tests ([PP3]).

Several other statistical types can be used for analyzing GWAS data. If the relationship between number of alleles and effect size is assumed to be linear, genotype can be seen as a discrete random variable with a possible set of values given by $\mathcal{S}_{Dis} = \{0, 1, 2\}$. At last, genotype value is considered as a continuum, the set of possible values for the SNP can be $\mathcal{S}_{Con} = \mathbb{R}$. Another challenge in the choice of the data type is therefore to use specific biological hypotheses to allow the calculation and/or the computation of statistical procedure [JP3, JP8].

Formalizing biologically relevant statistical hypotheses

The goal of GWAS is the detection of an association between a set of variants (or SNPs) and the phenotype. However, from a statistical point-of-view, the term *association* is not well defined and can be interpreted in multiple ways. Testing for association can indeed be performed at different scales of the genome, as for example at the single SNP level, at the gene level or at the genome level. These different hypotheses can be formalized as follows. Let Y be the random variable coding for the phenotype and $\mathbb{X} = [X_1, \dots, X_p]$ for the random vector of genotype where X_i is the

Y \ X	AA	Aa	aa	Total
0	n_0^{AA}	n_0^{Aa}	n_0^{aa}	n_0
1	n_1^{AA}	n_1^{Aa}	n_1^{aa}	n_1
Total	n^{AA}	n^{Aa}	n^{aa}	n

(a)

Y \ X	A	a	Total
0	n_0^A	n_0^a	n_0
1	n_1^A	n_1^a	n_1
Total	n^A	n^a	n

(b)

Table 1.1: Tables (a) and (b) are the contingency tables crossing the phenotype and a single SNP at the genotype and the allele level respectively.

random variable that characterizes the genotype for the i^{th} SNP.

When testing for the association of a single variant, for example X_i , the statistical hypotheses can be written as:

$$\mathcal{H}_0 = Y \perp\!\!\!\perp X_i \text{ vs. } \mathcal{H}_1 = Y \not\perp\!\!\!\perp X_i$$

However, since a large number of variants is collected, epidemiologists are interested in testing all SNPs. From a statistical point-of-view, the statistical question is slightly modified and can be formulated as “Is there one or more SNP associated with the phenotype?”. Statistical hypotheses can therefore be formalized as follows:

$$\mathcal{H}_0 = \{\forall i \in [1 \dots p], Y \perp\!\!\!\perp X_i\} \text{ vs. } \mathcal{H}_1 = \{\exists i \in [1 \dots p] / Y \not\perp\!\!\!\perp X_i\}$$

If we consider that the main goal of GWAS is to detect associated regions, it is necessary to properly define what is meant by regions. Genomic regions can be haplotypes, LD block, exons, introns, genes, etc. In any cases, the set of variables is modified and we set $R_i = \{X_{i,1}, \dots, X_{i,n_i}\}$ the i^{th} genomic region composed by n_i SNPs. When considering a total of r regions, statistical hypotheses can be formulated as follows:

$$\mathcal{H}_0 = \{\forall i \in [1 \dots r], Y \perp\!\!\!\perp R_i\} \text{ vs. } \mathcal{H}_1 = \{\exists i \in [1 \dots r] / Y \not\perp\!\!\!\perp R_i\}$$

Testing these statistical hypotheses requires the application or the design of an appropriate statistical procedure. It is thus crucial to biologically interpret the results with respect to the biological hypotheses that underlies the performed statistical procedure. One of most important challenge for a (bio)statistician, at least when performing inference based on the hypothesis framework, is therefore to translate the biological hypotheses into a formal statistical hypotheses [JP3, JP7, JP8, PP1].

Accounting for the structure of the data

The complex architecture of SNP data at the genome scale raises several statistical issues. First, as shown in Figure 1.3(a), SNP data are correlated with a block structure, reflecting the LD block correlation observed at the level of the population. Such correlation induces a dependency between the statistical tests for association (see Figure 1.3(b)), thus raising a challenging issue in the correction for multiple testing for example [JP8].

Another issue related to the architecture of the SNP data is the possibility to test for complex hypotheses. Although testing each SNP independently in a single-marker approach has been successful [Hindorff et al., 2009], findings were of modest effect and a large proportion of the genetic heritability is still not covered for common complex diseases [Maher, 2008, Manolio et al., 2009]. Epistasis (that can be interpreted as the effect of the interaction between genes on the phenotype) is often cited as one of the main biological mechanism carrying the “missing heritability” in GWAS

[Moore, 2003, Phillips, 2008]. Since human complex diseases are generally caused by the combined effect of multiple genes, the detection of genetic interactions is thus essential to improve our knowledge of the etiology of complex diseases [Cordell, 2009, Hindorff et al., 2009]. To address the so-called missing heritability, it is therefore natural to test for interaction between SNPs in susceptibility with the disease. However, because of the block structure of the genome, testing for interaction at the gene scale remains challenging [JP3, PP1].

1.4 Research axes and contributions

In order to present our research work, we have focused our attention in this manuscript on three main research axes, corresponding to our contributions in the design of statistical methods for biomedical research. These three axes are displayed in Figure 1.4 as intermediate points between our research challenges and the application fields that have motivated and validated our work. These axes indeed represent three levels of complexity in the analyzed data: we first consider data with a simple probabilistic characterization (Axis 1), then we examine structure of data resulting from the complex combination of simple variables (Axis 2) and finally we study data with a complex probabilistic modeling (Axis 3).

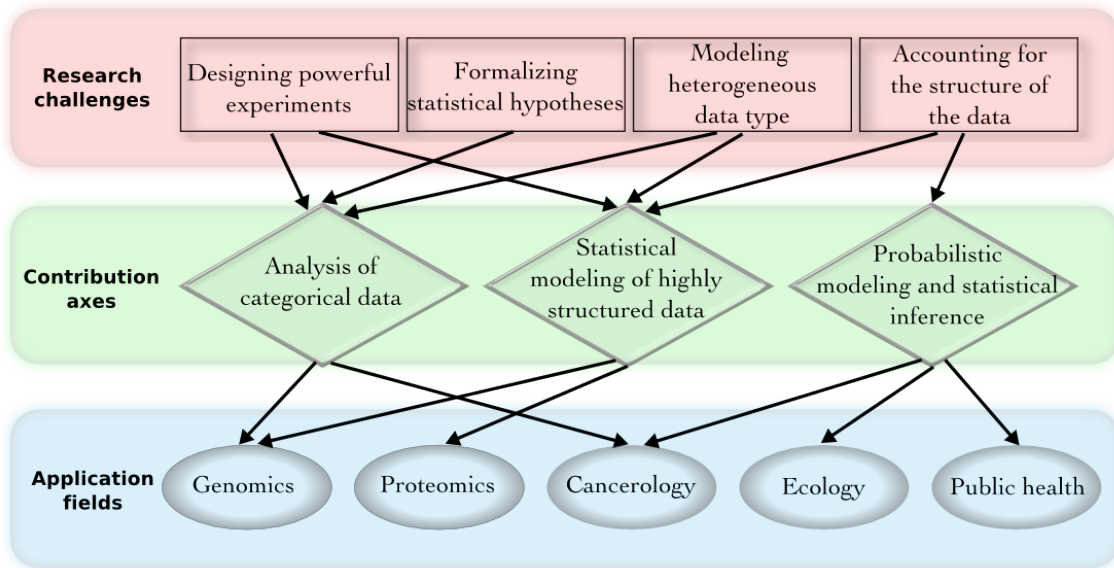


Figure 1.4: Summary of our contributions, drawing paths between research challenges and biomedical applications under consideration in this manuscript.

Axis 1 - Analysis of categorical data

In a first axis, we focus our research work on the analysis of data composed by variables with a relatively simple probabilistic modeling of marginal and joint distributions. We indeed consider the study of the relationships between 2 or 3 categorical variables through the analysis of two or three-ways contingency tables for two main reasons. First, categorical variables are particularly encountered in biomedical sciences, epidemiology, public health, genetics, etc. Next, compared to statistical models with continuous variables where the normal distribution plays the central role, models with categorical variables rely on several different distributions such as binomial, multinomial, Poisson distributions. One major issue with categorical distributions is that the size of parameter sets is increasing with the number of categories so that the degrees-of-freedom in a

contingency tables can be very important. Therefore statistical methods for studying relationships between categorical data were late gaining the level of sophistication achieved by continuous variables and many challenging questions remain.

In our work, we addressed the question of the **design of experiment** in association studies by proposing a formal comparison of power functions obtained with widely used association tests. In the context of genomics and cancerology we proposed recommendations to optimize statistical power with respect to experimental parameters and biological hypotheses. We further tackled the issue of **formalizing biologically-relevant statistical hypotheses** in genomics. Using specific **modeling of the data**, we derived a fast and efficient statistical procedure to detect an association between a phenotype and the interaction between two biological markers. Finally, we used a **biologically-oriented statistical hypothesis** to account for the **sparse type of data** used to detect genomic regions under selection.

Axis 2 - Statistical modeling of highly structured data

In most situations, because of the emergence of high-throughput technologies, it is of interest to investigate the relationship between one categorical response variable and a large number of explanatory variables falling into the paradigm of “high-dimensional” data. Fortunately, due to the (high) correlation between variables, the dimension of the underlying biological processes is often much more low dimensional than expected. Although the probabilistic modeling of individual variable is relatively simple, the combination of these data raises statistical issues, notably regarding the probabilistic characterization of the joint probability of such dataset. In our second axis, we therefore focus on the development of methodologies to account for these correlation structures in order to circumvent the issues raised by the high-dimensionality.

In our work, we first address the issue of variable selection in the **design of experiment** in genome-wide association studies. Using the argument that variable selection is driven by an objective function, we stress that the practice of GWAS is biased towards the identification of single-marker association. We then focus on the multiple testing issue by using **modeling of data type** oriented by biological scales and **correlation structure** among them. By mixing biological knowledge and statistical arguments we detect potential interaction between genes. To account for the correlation between variables, we also address the issue of the aggregation of tests. By using an appropriate **modeling of the data**, we therefore develop a region-based test for detecting interaction in the association studies. In the context of proteomics, where the number and the heterogeneity of features is very important, we propose a statistical framework to build meta-predictors. The application of such a methodology in the context of amyloidogenesis allows the detection of proteins involved in neurodegenerative diseases.

Axis 3 - Probabilistic modeling and related statistical inference

In our third axis, we focus on other complex data type where complexity is inherent to the statistical nature of the data. In many fields of biostatistics, such as ecology and public health, data are measured in complex mathematical support spaces. Such complexity requires a specific probabilistic characterization that can be guided by biological hypotheses. From a statistical point-of-view, the development of inference procedures therefore relies on the probabilistic modeling of the data. Thus, one major challenge for a biostatistician is to propose and develop relevant probabilistic models on which statistical inference can be based.

In our work, we tackle the issue of clustering data in a 2-dimensional space by using a point processing modeling framework. Such a framework allows us to propose a methodology that account for spatial covariates by integrating the **structure of the data** which plays a major role in ecology. In the context of cancerology, we propose a dedicated model of the spatial organization of tumor cells configurations. Based on such a model that formalizes the interaction **structure**

of the data points, we propose an inference procedure to discriminate between aggressive and non-aggressive tumors. Finally, by drawing a parallel between temporal **structure of data** in insurance and in hospital-acquired-disease, we propose an inference procedure to estimate the risk of developing a disease during a stay in a hospital.

Organization of the manuscript

This series of three research axes gives a natural organization of this manuscript in three chapters that can be read almost independently. Chapter 2 describes our contributions on the analysis of the relationship between categorical variables. Based on the contingency table framework, we propose several statistical models motivated by the characteristics of data in genetics and molecular biology. Chapter 3 presents novel methods that aim at accounting for the structure of data. Our contributions focus on methods for reducing the dimension of omics data (specifically genomics and proteomics data). In chapter 4, we describe our contributions on the probabilistic modeling of data observed in ecology and public health. Based on such a modeling, we develop adapted statistical inference procedures. Chapter 5 concludes the summary of our contributions and gives some perspectives on our future research activities.

Finally, in appendix A, I briefly introduce a more applied series of works in various fields such as physiopathology, virtual reality and peptidomics that I have been involved in. All these contributions share the common characteristics of relying on human-based experiments where the number of subjects is low and the individual variability is high. In all these works, we therefore used linear mixed models to analyze the data. However since the associated publications do not involve any significantly new statistical methodology, we will not reach level of details achieved in the three main chapters 2, 3 and 4.

All the contributions described in this manuscript have been achieved during my post-doctoral position (2007-2008 at the University of Aarhus) and my two associate professor positions (2008-2013 at the University of Rennes 2 and since 2013 at Agrocampus Ouest). During this period, I have been particularly honored to supervise Master students that have contributed to my research activities: Dr. Aida Eslami, Dr. Anthony Talvas, Hillel Jean-Baptiste Adolphe, Floriane Ethis de Corny, Emeline Geoffroy, Florian Kroell and Florian Hébert. I have also collaborated with different researchers and their names will be mentioned in the appropriate sections of this manuscript.

2.1 Introduction and background

In this chapter we present our contributions on the modeling and the statistical analysis of the relationship between categorical variables. The study of such relationships is usually performed through the analysis of contingency tables that are widely used in many fields of bioscience (epidemiology, public health, medicine, etc.) [Agresti, 2013]. The study of specific structures in contingency tables allows us to address the main statistical issues of **power** in association testing, **interaction** and **clustering**.

Let first introduce the main notations used in this chapter. Let consider a set of categorical variables displayed in a contingency table. We assume that the first categorical variable, denoted by Y with I categories, corresponds to the division of a population of individuals into I subgroups. For example, in a case-control study, two groups of individuals are compared (for example a group of healthy patients and a group of patients affected by a disease) so that $I = 2$ and Y is a binary variable. In our research context, Y can be considered as the response variable thus justifying the choice of the notation.

Let first consider the relationship between Y and a second categorical response variable, denoted by X with J categories. Observed data are displayed in a $I \times J$ table where the cell n_{ij} (for $i = 1 \dots I$ and $j = 1 \dots J$) contains the number of times the i^{th} category of Y and j^{th} category for X are jointly observed (see figure 2.1(a)). When investigating the relationship between Y and two categorical variables, X_1 with J categories and X_2 with K categories, observed data are summarized into a $I \times J \times K$ table as shown in figure 2.1(b).

One of the main questions addressed when handling contingency tables is the independence between the categorical variables. In a two-way contingency table, when Y and X are independent, the conditional distribution of Y (or X) is identical to the marginal distribution of Y (or X). The marginal distribution of a categorical variable can be seen as its unconditional distribution and is observed through the marginal counts defined as in Equation 2.1 for Y :

$$\forall i \in [1, I], \quad n_{i.} = \sum_{j=1}^J n_{ij} \quad (2.1)$$

Throughout this chapter, the categories of Y are considered as subgroups of a population. We further assume that the number of individuals in each subgroup is fixed so that data come from a specific design of experiment, called one-margin fixed design [Lydersen et al., 2009]. In other words, $\forall i \in [1, I], n_{i.}$ is not random, as for example in case-control study where the numbers of diseased and healthy patients are fixed according to financial constraints.

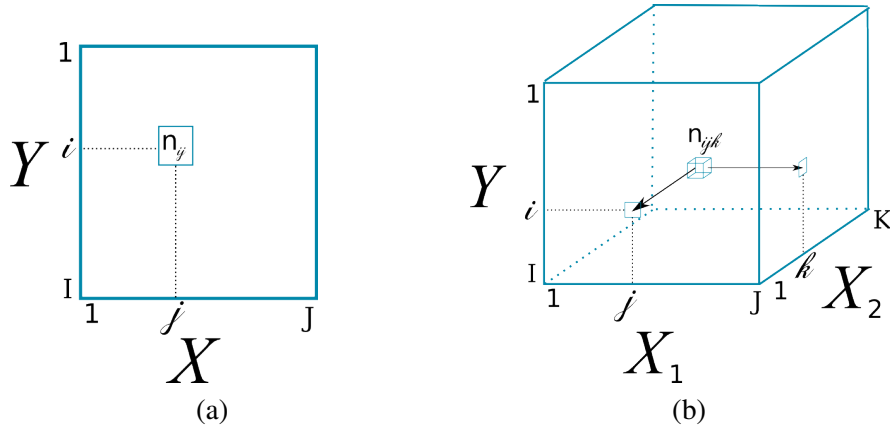


Figure 2.1: Scheme of observed contingency tables crossing (a) two categorical variables Y and X and (b) three categorical variables Y , X_1 and X_2 .

If the independence between the categorical variables is rejected, it is natural to investigate the cause of such a deviation from independence. In the most simple situation, where we consider two categorical variables each with 2 or 3 categories, there is only very few possible families of dependent structure. Statistical **power** is therefore the most appropriate measure to compare existing tests. If the number of categories for each of the two variables increases, dependence between variables can be summarized and interpreted by **clustering** similar categories. However, when considering the association between 3 (or more) variables, the number of structure of dependence is becoming so important that the design of statistical hypotheses, and especially the modeling of the pairwise **interaction** between variables, plays a major role in the detection of deviation from independence. Our contributions in contingency table analysis tackle the general statistical issues of power comparison (Section 2.2), interaction detection (Section 2.3) and cluster detection (Section 2.4).

In Section 2.2, we address the issue of comparing the statistical power of widely used association tests. In the literature, such comparisons have been performed using computations of the power functions that are obtained through approximations. However, computation-based comparisons hardly account for the multidimensionality of the set of features involved in power functions, thus highlighting the need for an analytical comparison. In [JP4] and [PP3], we proposed a general framework to compare χ^2 distributed association statistical based on the comparison of non-centrality parameters.

In Section 2.3, we tackle the issue of detecting an association between a binary variable and the interaction between two other categorical variables. In [JP7], we proposed a formal interpretation of the biological meaning of interaction between genetic markers. Based on odds-ratio modeling we further derived an appropriate statistical test.

In Section 2.4, we aim at clustering individuals in categories population that are characterized by different sets of categories observed from another variable. Since cluster detection depends on a precise definition of a cluster, we introduced in [JP2] a characterization of a cluster in the context of genomic selection that specifically accounts for the sparsity of the contingency table. We further proposed a statistical framework to cluster individuals under selection based on the definition of a novel dissimilarity and elements of graph theory.

2.2 Statistical power for single testing in case-control association

Our research work presented in this section proposes a general framework to compare the power of χ^2 distributed statistics. Part of this work has been conducted in collaboration with Chloé Friguet (Université Bretagne-Sud, Vannes, France).

2.2.1 Context and issue

Single testing in case-control association consists in testing the association between a binary variable Y , such as presence ($Y = 1$) or absence ($Y = 0$) of a disease, and a categorical explanatory variable X . Motivated by GWAS, our research work focuses on a categorical variable X with two or three categories since X represents a bi-allelic SNP. At the allele level, the two categories of a SNP are A and a , while at the genotypic level, the SNP is characterized by three categories in $[AA, Aa, aa]$ (see paragraph **data type** in section 1.3 for more details). As shown in Table 1.1, data can therefore be displayed in a 2×2 or a 2×3 contingency table whether X is considered at the allelic or the genotypic level.

Association testing can be performed by a large collection of statistical tests. However some statistics are very popular in practice such as Pearson's χ^2 , likelihood ratio, odds ratio or Cochran-Armitage [Agresti, 2013]. We start by setting the formal expression for each popular statistic and their associated statistical hypotheses, thus corresponding to specific biological assumption. The general statistical hypotheses in single testing in case-control association can be written as:

$$\mathcal{H}_0 = Y \perp\!\!\!\perp X \quad \text{vs.} \quad \mathcal{H}_1 = Y \not\perp\!\!\!\perp X \quad (2.2)$$

However, in case of a 2×2 contingency, association testing consists in comparing two binomial proportions and hypotheses in Equation 2.2 can be rewritten as:

$$\mathcal{H}_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad \mathcal{H}_1 : \pi_1 \neq \pi_2, \quad (2.3)$$

where $\pi_i = \mathbb{P}[X = a | Y = i]$ for $i = 0, 1$. We can easily see that \mathcal{H}_0 is similar to the non-association between X and Y , as conditioning on Y does not modify the distribution of X . Testing the equality of two binomial proportions can be performed by widely-used Pearson's chi-squared test or the likelihood-ratio test. Furthermore, \mathcal{H}_0 and \mathcal{H}_1 are equivalent to the following statistical hypotheses:

$$\mathcal{H}_0 : \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) \neq 0, \quad (2.4)$$

where $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ is the so-called odds ratio (OR). Given the links between the odds ratio and the coefficients of a logistic regression model, the statistical hypotheses proposed in Equations (2.3) and (2.4) can be rewritten as:

$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \beta_1 \neq 0.$$

β_1 is a coefficient of the following logistic regression model:

$$\text{logit}(\mathbb{P}[X = 2 | Y = y]) = \beta_0 + \beta_1 \mathbb{I}_{\{1\}}(y),$$

where $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ and \mathbb{I} is the indicator function. Therefore, a classical statistical inference regarding the odds ratio can be applied to compare two binomial proportions.

In practice, Pearson's χ^2 and likelihood-ratio (LR) tests are very popular. Both tests are based on the comparison between observed counts and the estimated expected counts under \mathcal{H}_0 , denoted m_{ij} and defined as:

$$m_i^j = \frac{n_i n_{\cdot j}}{N} \quad \text{for } i = 0, 1 \quad \text{and } j \in [a, A]$$

Pearson's χ^2 test statistic is given by:

$$P_A = \sum_{i \in \{0,1\}} \sum_{j \in [a,A]} \frac{(n_i^j - m_i^j)^2}{m_i^j}$$

and the LR test statistic is defined as:

$$LR_A = 2(\ell_1 - \ell_0) = 2 \sum_{i \in \{0,1\}} \sum_{j \in [a,A]} n_i^j \log \left(\frac{n_i^j}{m_i^j} \right)$$

where ℓ_1 (resp. ℓ_0) is the maximized log-likelihood under \mathcal{H}_1 (resp. \mathcal{H}_0). Inference of the odds ratio is usually performed by testing the nullity of the log odds ratio. Such test is based on a test statistic, hereafter denoted z , which is defined as:

$$z^2 = \frac{t^2}{\sigma^2} = \frac{\left(\log \left(\frac{n_0^A n_1^a}{n_0^a n_1^A} \right) \right)^2}{\frac{1}{n_0^A} + \frac{1}{n_0^a} + \frac{1}{n_1^A} + \frac{1}{n_1^a}}$$

When considering 2×3 contingency tables, both Pearson's χ^2 and LR tests can be easily extended to handle with genotypic counts. Similar to the allele case, the expected genotypic counts under \mathcal{H}_0 are obtained by $m_i^j = \frac{n_i n^j}{n} \quad \forall i \in \{aa, aA, AA\}$ and $j = 0, 1$. Pearson's χ^2 test statistic (P_G) and LR test statistic (LR_G) can thus be written as:

$$P_G = \sum_{i \in \{0,1\}} \sum_{j \in \{AA, Aa, aa\}} \frac{(n_i^j - m_i^j)^2}{m_i^j} \quad \text{and} \quad LR_G = 2 \sum_{i \in \{0,1\}} \sum_{j \in \{AA, Aa, aa\}} n_i^j \log \left(\frac{n_i^j}{m_i^j} \right).$$

In the context of genetic association studies, the Cochran-Armitage Test of linear Trend (CA) has also been widely used [Armitage, 1955]. CA modifies the Pearson chi-squared test to incorporate a suspected ordering in the effects of the categories for X . The most widely used version of CA aims at testing a linear effect of the number of copies of the minor allele and is defined as follows:

$$CA = \frac{n_2 \left((n_0^{AA} n_1 - n_1^{AA} n_0) + 2(n_0^{Aa} n_1 - n_1^{Aa} n_0) \right)^2}{n_0 n_1 \left(n^{AA} (n - n^{Aa}) + 4n^{Aa} (n - n^{aa}) - 4n^{Aa} n^{aa} \right)}.$$

Under \mathcal{H}_0 , P_A , LR_A , z^2 , as well as CA follow asymptotically a central chi-squared distribution with one degree-of-freedom. The statistics P_G and LR_G also follow asymptotically a central chi-squared distribution but with two degrees-of-freedom. In the following, the χ_A^2 test, the LR_A test and the z^2 test refer to Pearson's chi-squared test, likelihood-ratio test based on allele counts and to odds ratio based test. Furthermore, χ_G^2 and the LR_G tests respectively correspond to Pearson's chi-squared test and likelihood-ratio test based on genotype counts. Finally, CA is used for the Cochran-Armitage test of trend.

A substantial literature is dedicated to the analysis of association tests especially for the comparison of two binomial proportions in 2×2 contingency tables [Lin and Yang, 2009, Fagerland et al., 2015, Hirji et al., 1991, Agresti, 2013]. Many studies have focused on comparing the behavior of Pearson's χ^2 and log-likelihood-ratio tests. The study of the power divergence family demonstrates that Pearson's χ^2 and LR are asymptotically equivalent under the null hypothesis [Cressie and Read, 1984, Cressie and Read, 1989].

Besides, comparing the behavior of statistics under the alternative hypothesis is also a crucial issue since it allows us to choose the most powerful test. Various studies have focused on comparing the power of asymptotic tests based on Pearson's χ^2 and LR statistics [Lydersen et al., 2009,

[Ruxton and Neuhauser, 2010]. However, although differences have been observed in the behavior of both tests, neither of them is uniformly more powerful than the other. Indeed, Zar [Zar, 2008] suggests that the two tests generally yield similar predictions and cites studies that express preferences for either one test or the other. For instance, the LR statistic is preferred due to some theoretical and computational advantages [Sokal and Rohlf, 1995], while it has been suggested that Pearson’s χ^2 behaves better than LR when sample size is low [Quinn and Keough, 2002]. Such comparisons have been performed using computations of the power functions that are obtained through approximations, such as in the method proposed by [Drost et al., 1989]. However, computation-based comparisons hardly account for the multidimensionality of the set of features involved in power functions, thus highlighting the need for an analytical comparison. Nevertheless, association test statistics based on categorical variables are usually χ^2 distributed, for which power functions are intractable.

2.2.2 Approach

Main idea

In [JP4] we introduced a general framework to evaluate and compare the power of one degree-of-freedom χ^2 distributed tests of association. We extended such a framework to compare two degrees-of-freedom tests in the context of genetic association studies [PP3]. Beside previous power studies, our methodology is based on the analytical comparison of power functions. Compared to computation-based and simulation-based methods, the formal derivation of power functions offers the advantage of simultaneously investigating the whole multidimensional space of parameters.

Under the alternative hypothesis, association test statistics follow asymptotically non-central χ^2 distributions with one degree-of-freedom for P_A , LR_A , z^2 and CA and two degrees-of-freedom for P_G and LR_G . For both tests, the power function thus increases with the corresponding non-centrality parameter. Instead of studying the power function, for which no explicit formula is available, we therefore focus our effort on a comparison of the non-centrality parameters of each test. Let us introduce the non-centrality parameters, called λ_{P_A} , λ_{LR_A} , λ_{z^2} , λ_{P_G} , λ_{LR_G} and λ_{CA} used to define the distributions of the test statistics as shown in Table 2.1. Given a significant level α , the power of each test is defined as the probability under \mathcal{H}_1 , that the statistic is greater than the $(1 - \alpha)$ -quantile of the corresponding central chi-squared distribution, denoted $q_{1-\alpha}^i$ where $i = 1, 2$ is the number of degree-of-freedom. Table 2.1 summarizes the power functions for each test and shows that these functions can be compared directly by evaluating their non-centrality parameters.

Test statistic	Distribution under \mathcal{H}_1	Power function
P_A	$\chi_1^2(\lambda_{P_A})$	$\mathbb{P}\left(P_A > q_{1-\alpha}^1\right)$
LR_A	$\chi_1^2(\lambda_{LR_A})$	$\mathbb{P}\left(LR_A > q_{1-\alpha}^1\right)$
z_2	$\chi_1^2(\lambda_{z^2})$	$\mathbb{P}\left(z^2 > q_{1-\alpha}^1\right)$
CA	$\chi_1^2(\lambda_{CA})$	$\mathbb{P}\left(CA > q_{1-\alpha}^1\right)$
P_G	$\chi_2^2(\lambda_{P_G})$	$\mathbb{P}\left(P_G > q_{1-\alpha}^2\right)$
LR_G	$\chi_2^2(\lambda_{LR_G})$	$\mathbb{P}\left(LR_G > q_{1-\alpha}^2\right)$

Table 2.1: *Distribution and power function for each test statistic.*

Estimation of the non-centrality parameters

To be compared, non-centrality parameters for association tests have to be estimated since they do not have a general closed form. In [JP4], we proposed to estimate non-centrality parameters by using the expected observed counts under a predefined alternative hypothesis. We therefore consider that conditional probabilities, $\mathbb{P}[Y = j|X = i]$, for $j = 0, 1$ and $i \in \{A, a\}$ or $i \in \{AA, Aa, aa\}$, are fixed. Expected observed counts can thus be defined, for $i \in \{A, a\}$ or $i \in \{AA, Aa, aa\}$, by :

$$ne_i^1 = n\varphi\mathbb{P}[Y = 1|X = i] \quad \text{and} \quad ne_i^0 = n(1 - \varphi)\mathbb{P}[Y = 0|X = i],$$

where $\varphi = \frac{n_1}{n_0}$ is the balance of the design (or case-to-control ratio) that is not random in our one-margin fixed design. Non-centrality parameters are thus estimated by comparing the expected observed counts ne_i^j to $me_i^j = \frac{ne_i^j ne_i^j}{n}$ the expected counts assuming that \mathcal{H}_0 is true which leads to:

$$\begin{aligned} \widehat{\lambda_{P_A}} &= \sum_{i \in \{A, a\}} \sum_{j \in \{0, 1\}} \frac{(ne_i^j - me_i^j)^2}{me_i^j} \\ \widehat{\lambda_{LR_A}} &= 2 \sum_{i \in \{A, a\}} \sum_{j \in \{0, 1\}} ne_i^j \log \left(\frac{ne_i^j}{me_i^j} \right) \\ \widehat{\lambda_{z^2}} &= \frac{\left(\log \left(\frac{ne_0^A ne_1^a}{ne_0^a ne_1^A} \right) \right)^2}{\frac{1}{ne_0^A} + \frac{1}{ne_0^a} + \frac{1}{ne_1^A} + \frac{1}{ne_1^a}} \\ \widehat{\lambda_{P_G}} &= \sum_{i \in \{AA, Aa, aa\}} \sum_{j \in \{0, 1\}} \frac{(ne_i^j - me_i^j)^2}{me_i^j} \\ \widehat{\lambda_{LR_G}} &= 2 \sum_{i \in \{AA, Aa, aa\}} \sum_{j \in \{0, 1\}} ne_i^j \log \left(\frac{ne_i^j}{me_i^j} \right) \\ \widehat{\lambda_{CATT}} &= \frac{n^2 \left((ne_0^{Aa} n_1 - ne_1^{Aa} n_0) + 2(ne_0^{aa} n_1 - ne_1^{aa} n_0) \right)^2}{n_0 n_1 (ne^{Aa}(n - ne^{Aa}) + 4ne^{aa}(n - ne^{aa}) - 4ne^{Aa} ne^{aa})} \end{aligned}$$

Parameters that influence power

In [JP4] and [PP3], we formalized a set of 5 parameters (φ , n , π_a , h and β_1) that interplay in power functions of association tests. First, we focused on φ , the balance of the design and n the total number of observations, that are controlled beforehand since they characterize the experimental design of the study. We also introduced π_a and h , two parameters related to the distribution of the X variable. When X is assumed to be a categorical variable with two categories, we set $\pi_a = \mathbb{P}[X = a]$ as the parameter of binomial law followed by X . If X has three categories, X has a multinomial law and we further introduced h as a measure of the deviation from Hardy-Weinberg Equilibrium as follows:

$$\mathbb{P}[X = AA] = (1 - \pi_a)^2 \left(1 + \frac{\pi_a}{1 - \pi_a} (1 - h) \right), \quad \mathbb{P}[X = Aa] = 2h(1 - \pi_a)\pi_a \quad \text{and} \quad \mathbb{P}[X = aa] = \pi_a^2 \left(1 + \frac{1 - \pi_a}{\pi_a} (1 - h) \right)$$

The final parameter, denoted by β_1 , is used to parameterize the size and the nature of the association effect computed under \mathcal{H}_1 .

2.2.3 Main Results and application to GWAS

Analytical comparison under the recessive alternative hypothesis

Let consider a class of association models parameterized by β_1 and defined as follows:

$$\mathcal{M}_{Rec} : \quad \text{logit}(\mathbb{P}[Y = 1|X = x]) = \beta_0 + \beta_1 \mathbb{I}_{x=aa}$$

Model \mathcal{M}_{Rec} corresponds to the so-called recessive model of inheritance commonly used in genetic epidemiology. We performed the estimation of non-centrality parameters by using a Taylor decomposition on β_1 . As proposed in [PP3], focusing on the first order coefficients of the Taylor decompositions provides an interpretable comparison of the power. For example, Equation 2.5 displays the first order coefficient for CA test.

$$\widehat{\lambda_{CA}} = \beta_1^2 \times \frac{2n\varphi(1-\varphi)\pi_a(1-\pi_a)(1-h(1-\pi_a))^2}{2-h} + o(\beta_1^2) \quad (2.5)$$

In more details, Equation 2.5 shows that $\widehat{\lambda_{CA}}$ is increasing with n and is maximum when the design is well balanced ($\varphi = 1/2$). Furthermore, $\widehat{\lambda_{CA}}$ is increasing with π_a when $\pi_a \in [0, 1/2]$. Moreover, the impact of a deviation from HWE on the power of detection is governed by the term $(1-h(1-\pi_a))/(2-h)$. It can be remarked that such a term is decreasing with h , for all $\pi_a \in [0, 1/2]$. As a consequence, the statistical power of CA is maximum when the allele is common ($\pi_a \approx 1/2$) and HWE is not satisfied ($h \approx 0$). To support our conclusions, the evolution of $\widehat{\lambda_{CA}}$ according to π_a and h is displayed in Figure 2.2. Our results confirm that $\widehat{\lambda_{CA}}$, and hence power, is increasing with π_a and decreasing with h .

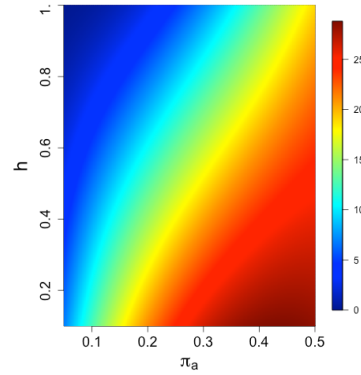


Figure 2.2: Evolution of the λ_{CA} with respect to π_A and h .

Recommendations

In [JP4] and [PP3], we compared all non-centrality parameters and provided a table of recommendations regarding the most powerful test with respect to multidimensional set of parameters. In Table 2.2, it can be remarked that for the recessive disease model, parameters h and φ play a major role in the comparison of the statistical association tests.

Extensions

In [JP4], we further investigated situations where Y and X are not compared directly. In such cases, X is tagged by another binomial (or multinomial) variable T , and the true difference between the Y and X can only be detected by comparing the difference between the Y and T . Such a situation, called indirect association, is encountered in GWAS where the SNP (Single Nucleotide

π_a	$h < 0.1$			$h = 0.5$			$h > 0.9$		
	0.05	0.25	0.45	0.05	0.25	0.45	0.05	0.25	0.45
$\varphi = 0.05$	CA	CA	CA	CA, χ_G^2	CA, χ_G^2	CA, χ_G^2	χ_G^2	χ_G^2	χ_G^2
$\varphi = 0.5$	CA	CA	CA	CA, χ_G^2, LR_G	CA, χ_G^2, LR_G	CA, χ_G^2, LR_G	χ_G^2, LR_G	χ_G^2, LR_G	χ_G^2, LR_G
$\varphi = 0.95$	CA	CA	CA	CA, LR_G	CA, LR_G	CA, LR_G	LR_G	LR_G	LR_G

Table 2.2: *Most powerful test(s) under \mathcal{M}_{Rec} .*

Polymorphism) array does not genotype all SNPs but a selection of informative SNPs, called tag-SNPs [Carlson et al., 2004]. Although it is widely assumed that testing for indirect association tends to decrease the power of detection [Moskvina and O'Donovan, 2007], our methodology allows us to investigate in more detail the impact of the indirect association on the power of each test. We showed that power is also influenced by r the correlation between X and T , as well as π_T , the parameter of the tag binomial. We first showed that, in the case of indirect association, the sample size has to be increased by a factor of $1/r^2$ to reach the same power as we obtained in the case of direct association case. Conclusions drawn in the indirect case are very similar to those obtained in the direct case. However, it is noteworthy that differences between tests are less marked with the increasing of the parameter of the tag binomial, π_T .

In [PP3], we also investigated other disease models, such as the dominant model (\mathcal{M}_{Dom}) and the multiplicative model (\mathcal{M}_{Mult}) defined as:

$$\begin{aligned} \mathcal{M}_{Dom} : \quad \text{logit}(\mathbb{P}[Y = 1|X = x]) &= \beta_0 + \beta_1 \mathbb{I}_{x=Aa} + \beta_1 \mathbb{I}_{x=aa} \\ \mathcal{M}_{Mult} : \quad \text{logit}(\mathbb{P}[Y = 1|X = x]) &= \beta_0 + \beta_1 \mathbb{I}_{x=Aa} + 2\beta_1 \mathbb{I}_{x=aa} \end{aligned}$$

Recommendations obtained were very different from one alternative hypothesis to the others.

Applications to GWAS

In [JP4], we compared the behavior of χ_A^2 , LR_A , z^2 , χ_G^2 , LR_G and $CATT$ tests in a real situation by performing genome-wide scans for single-locus associations on a publicly available dataset from the Wellcome Trust Case Control Consortium (WTCCC) [WTCCC, 2007]. In our analysis, we focused on Crohn's disease, a chronic inflammatory disease, which causes inflammation of the gastrointestinal tract, as it has been proved to be heritable [Franke et al., 2010]. The number of cases were approximately 2,000 while the number of patients not affected by Crohn's Disease was equal to $\approx 15,000$, thus leading a case-to-control ratio of $\varphi \approx 2,000/15,000 \approx 0.13$.

Figure 2.3 shows the “ $-\log_{10}$ p-values” for SNPs that are significant (after a Bonferroni correction) for at least one association test. Our results confirmed that the behavior of the tests can be very different when the case-to-control ratio φ and π_a are low.

2.2.4 Concluding remarks

In this work, we proposed a methodology to compare the power of association tests that have a χ^2 distribution. We provided analytical estimation of the power functions under different alternative hypotheses that correspond to classical disease models in genetic epidemiology. Our main results, confirmed by the analysis of real situations, show that none of the tests is uniformly the most powerful. The choice of the most powerful test depends on the **design of experiment**, on the **data type** and on the **alternative hypothesis**, *i.e.* the nature of the investigated signal.

Such a result confirmed that each test is based on an analytical hypothesis that can be translated into biological assumptions, thus enhancing the importance of putting the biological interpretation of a statistical analysis in the perspective of modeling assumptions.

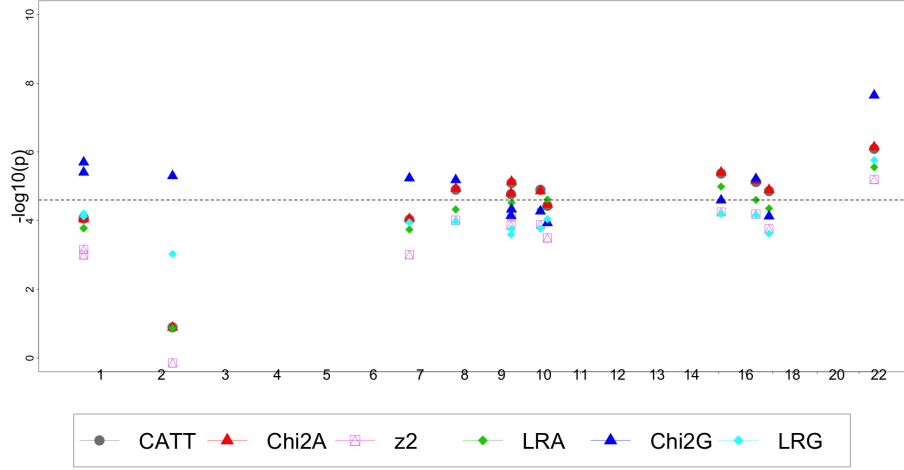


Figure 2.3: *Genome-wide scans obtained with the 6 association tests in the WTCCC dataset for Crohn's disease.*

2.3 Interaction in three-way contingency tables

Our research work presented in this section proposes a methodology to detect an association between one binomial variable and the interaction between two categorical variables.

2.3.1 Context and issue

In life science, it is very common that the interaction between a set of explanatory variables $\{X_1, \dots, X_p\}$ impacts the distribution of a response variable Y , such as the interaction between two genes in susceptibility with a phenotypic trait. Even in this simple situation, where $p = 2$, and considering that each explanatory variable has only three categories, such as SNP genotypes, the number of interaction models is enormous [Li and Reich, 2000, Hallgrimsdottir and Yuster, 2008]. In that context, an interaction model corresponds to a typical structure in a three-way contingency table that summarizes the relationship between three categorical variables: Y with 2 categories $\{0, 1\}$, X_1 with three categories $\{AA, Aa, aa\}$ and X_2 with three categories $\{BB, Bb, bb\}$. Such a table can be displayed as in Figure 2.4 where $n_i^{j,k}$ is the number of times the event “ $Y = i \cap X_1 = j \cap X_2 = k$ ” is observed.

From a statistical point-of-view, an interaction is defined as a deviation from the additivity of the marginal effects of X_1 and X_2 on Y . Testing for interaction can therefore be formalized through the following statistical hypotheses:

$$\mathcal{H}_0 : [\beta_5, \beta_6, \beta_7, \beta_8] = [0, 0, 0, 0] \text{ vs. } \mathcal{H}_1 : [\beta_5, \beta_6, \beta_7, \beta_8] \neq [0, 0, 0, 0]$$

where $[\beta_5, \beta_6, \beta_7, \beta_8]$ is a subset of regression coefficients of a logistic model:

$$\begin{aligned} \text{logit}(\mathbb{P}[Y = 1 | X_1 = x_1, X_2 = x_2]) &= \beta_0 + \beta_1 \mathbb{I}_{x_1=Aa} + \beta_2 \mathbb{I}_{x_1=aa} + \beta_3 \mathbb{I}_{x_2=Bb} + \beta_4 \mathbb{I}_{x_2=bb} \\ &\quad + \beta_5 \mathbb{I}_{x_1=Aa} \mathbb{I}_{x_2=Bb} + \beta_6 \mathbb{I}_{x_1=aa} \mathbb{I}_{x_2=Bb} + \beta_7 \mathbb{I}_{x_1=Aa} \mathbb{I}_{x_2=bb} + \beta_8 \mathbb{I}_{x_1=aa} \mathbb{I}_{x_2=bb} \end{aligned}$$

Testing for \mathcal{H}_0 can be performed using likelihood ratio tests.

However, such a definition in the context of genetic epidemiology has raised a controversy regarding the ability of statistical interaction to detect biological interaction. Although biological

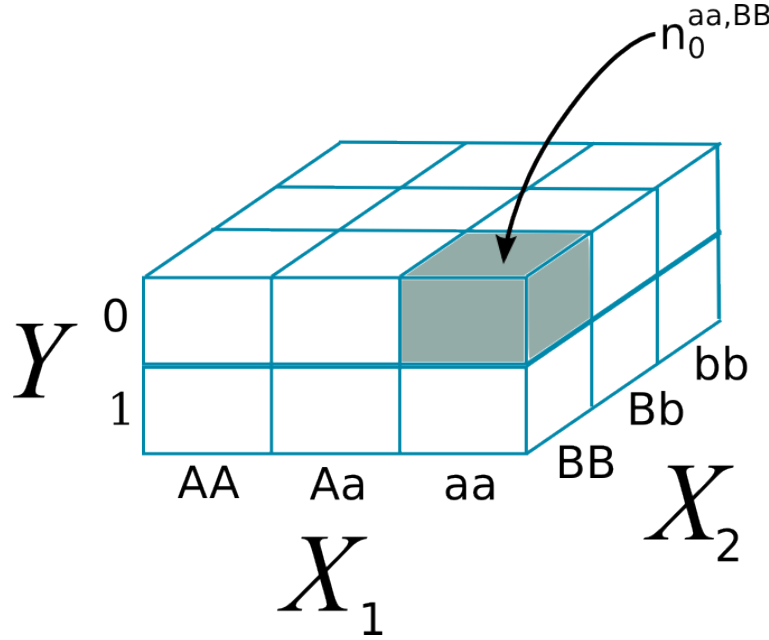


Figure 2.4: *Three-way contingency table summarizing the cross tabulation between a case-control phenotype (Y) and two bi-allelic SNPs (X_1 and X_2).*

interaction does not have a formal definition, many studies provided evidence of discrepancies between statistical and biological interpretation of interaction [Cordell, 2009]. Following pioneered definition given by Bateson [Bateson, 1909], it has been argued that biological interaction between two genes, also called epistasis, is a departure from independence between the two genes. To fill the gap between biological and statistical definitions of interaction, it is therefore crucial to propose an appropriate hypothesis testing framework.

2.3.2 Our approach: IndOR for Independent Odds-Ratio

In [JP7], we designed the statistical procedure IndOR (for Independent Odds-Ratio) in order to detect a variation in the dependency between two SNPs in the affected and in the unaffected populations. The novelty of our approach is that the independence is assumed to be the statistical independence so that under the null hypothesis \mathcal{H}_0 , cases and controls share the same amount of dependency.

Definition of the statistic IndOR

Our null hypothesis can be detailed in the following way:

$\forall (x_1, x_2) \in [AA, Aa, aa] \times [BB, Bb, bb]$:

$$\mathcal{H}_0 : \frac{\mathbb{P}[X_1 = x_1, X_2 = x_2 | Y = 1]}{\mathbb{P}[X_1 = x_1 | Y = 1] \mathbb{P}[X_2 = x_2 | Y = 1]} = \frac{\mathbb{P}[X_1 = x_1, x_b | Y = 0]}{\mathbb{P}[X_1 = x_1 | Y = 0] \mathbb{P}[X_2 = x_2 | Y = 0]} \quad (2.6)$$

Using Bayes formula, Equation 2.6 is equivalent to :

$$\frac{\mathbb{P}[Y = 1 | X_1 = x_1, X_2 = x_2] \mathbb{P}[Y = 1]}{\mathbb{P}[Y = 1 | X_1 = x_1] \mathbb{P}[Y = 1 | X_2 = x_2]} = \frac{\mathbb{P}[Y = 0 | X_1 = x_1, X_2 = x_2] \mathbb{P}[Y = 0]}{\mathbb{P}[Y = 0 | X_1 = x_1] \mathbb{P}[Y = 0 | X_2 = x_2]} \quad (2.7)$$

By considering AA, BB as the baseline genotype, odds-ratios (OR) can be defined as following:

$$OR(x_1, x_2) = \frac{odds(x_1, x_2)}{odds(AA, BB)} = \frac{\frac{\mathbb{P}[Y=1|X_1=x_1, X_2=x_2]}{\mathbb{P}[Y=0|X_1=x_1, X_2=x_2]}}{\frac{\mathbb{P}[Y=1|X_1=AA, X_2=BB]}{\mathbb{P}[Y=0|X_1=AA, X_2=BB]}}$$

$$OR(x_1) = \frac{odds(x_1)}{odds(AA)} = \frac{\frac{\mathbb{P}[Y=1|X_1=x_1]}{\mathbb{P}[Y=0|X_1=x_1]}}{\frac{\mathbb{P}[Y=1|X_1=AA]}{\mathbb{P}[Y=0|X_1=AA]}} \quad \text{and} \quad OR(x_2) = \frac{odds(x_2)}{odds(BB)} = \frac{\frac{\mathbb{P}[Y=1|X_2=x_2]}{\mathbb{P}[Y=0|X_2=x_2]}}{\frac{\mathbb{P}[Y=1|X_2=BB]}{\mathbb{P}[Y=0|X_2=BB]}}$$

Thus, according to equation 2.7, under \mathcal{H}_0 we have:

$$\frac{OR(x_1, x_2)}{OR(x_1)OR(x_2)} = 1 \quad (2.8)$$

The joint effect between X_1 and X_2 has four degrees of freedom. As AA, BB was considered as the baseline genotype pair, we measured in [JP7] the variation of dependency over the four following pairs: $AaBb, aaBb, Aabb$ and $aabb$. We used equation 2.8 to propose the four-dimensional vector $\Phi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$ where :

$$\begin{aligned} \varphi_1 &= \log\left(\frac{OR(Aa, Bb)}{OR(Aa)OR(Bb)}\right) ; & \varphi_2 &= \log\left(\frac{OR(aa, Bb)}{OR(aa)OR(Bb)}\right) ; \\ \varphi_3 &= \log\left(\frac{OR(Aa, bb)}{OR(Aa)OR(bb)}\right) ; & \varphi_4 &= \log\left(\frac{OR(aa, bb)}{OR(aa)OR(bb)}\right) ; \end{aligned}$$

The null and alternative hypotheses can then be defined as:

$$\mathcal{H}_0 : \Phi = [0; 0; 0; 0] \text{ and } \mathcal{H}_1 : \Phi \neq [0; 0; 0; 0] \quad (2.9)$$

In [JP7], to test for \mathcal{H}_0 , we defined our Wald statistic, IndOR, as follows:

$$\text{IndOR} = \Phi V_{\Phi}^{-1} \Phi^t \quad (2.10)$$

where V_{Φ}^{-1} is the inverse of variance-covariance matrix for Φ and Φ^t is the transposed vector of Φ . Under the null hypothesis of the same amount of dependence between cases and controls, the score IndOR follows a central χ^2 distribution with four degrees of freedom.

Estimation of the multidimensional vector Φ

Φ is defined in terms of odds-ratio. Odds-ratio are well-studied coefficients in epidemiology and Maximum Likelihood Estimator (MLE) can be easily computed [Thomas, 2004]. In our context, MLE for the four coefficients can be estimated by:

$$\begin{aligned} \widehat{\varphi}_1 &= \log(n_1^{AaBb}) + \log(n_1^{AA}) + \log(n_1^{BB}) - \log(n_1^{AABB}) - \log(n_1^{Aa}) - \log(n_1^{Bb}) \\ &\quad - \log(n_u^{AaBb}) - \log(n_u^{AA}) - \log(n_u^{BB}) + \log(n_u^{AABB}) + \log(n_u^{Aa}) + \log(n_u^{Bb}) \\ \widehat{\varphi}_2 &= \log(n_1^{aaBb}) + \log(n_1^{AA}) + \log(n_1^{BB}) - \log(n_1^{AABB}) - \log(n_1^{aa}) - \log(n_1^{Bb}) \\ &\quad - \log(n_u^{aaBb}) - \log(n_u^{AA}) - \log(n_u^{BB}) + \log(n_u^{AABB}) + \log(n_u^{aa}) + \log(n_u^{Bb}) \\ \widehat{\varphi}_3 &= \log(n_1^{Aabb}) + \log(n_1^{AA}) + \log(n_1^{BB}) - \log(n_1^{AABB}) - \log(n_1^{Aa}) - \log(n_1^{bb}) \\ &\quad - \log(n_u^{Aabb}) - \log(n_u^{AA}) - \log(n_u^{BB}) + \log(n_u^{AABB}) + \log(n_u^{Aa}) + \log(n_u^{bb}) \\ \widehat{\varphi}_4 &= \log(n_1^{aabb}) + \log(n_1^{AA}) + \log(n_1^{BB}) - \log(n_1^{AABB}) - \log(n_1^{aa}) - \log(n_1^{bb}) \\ &\quad - \log(n_u^{aabb}) - \log(n_u^{AA}) - \log(n_u^{BB}) + \log(n_u^{AABB}) + \log(n_u^{aa}) + \log(n_u^{bb}) \end{aligned}$$

Estimation of the covariance matrix V_{Φ}

In [JP7], we used the δ method to estimate the variance-covariance matrix V_{Φ} . The δ method is particularly adapted to the estimation of variance-covariance matrices of multidimensional odds-ratio vectors [Casella and Berger, 1990, Freedman, 2000, Lui, 2004, Chen and Chatterjee, 2007].

We assumed counts for cases and controls to follow independent multinomial distributions such as:

$$[N_1^{AABB}, \dots, N_1^{aabb}] \sim \text{Mult}(p_1^{AABB}, \dots, p_1^{aabb})$$

and

$$[N_u^{AABB}, \dots, N_u^{aabb}] \sim \text{Mult}(p_u^{AABB}, \dots, p_u^{aabb})$$

where $N_1^{x_a x_b}$ (resp. $N_u^{x_a x_b}$) is the random variable modeling the number of diseased (resp. healthy) patients with x_a, x_b genotype. $p_1^{x_a x_b}$ (resp. $p_u^{x_a x_b}$) is the probability of having AA, BB genotype for a diseased (resp. healthy) patient.

Thus, counts for all genotype (x_a, x_b) in the affected population are asymptotically equivalent to the following expression:

$$N_1^{x_a x_b} = n_1 p_1^{x_a x_b} \left(1 + \sqrt{\frac{(1 - p_1^{x_a x_b})}{n_1 p_1^{x_a x_b}}} \delta_1^{x_a x_b} \right)$$

where the δ coefficients have the following correlation structure:

$$\delta_1^{x_a x_b} \sim \mathcal{N}(0, 1) \quad (2.11)$$

$$\text{Cov}(\delta_1^{x_a x_b}, \delta_1^{x'_a x'_b}) = -\sqrt{\frac{p_1^{x_a x_b} p_1^{x'_a x'_b}}{(1 - p_1^{x_a x_b})(1 - p_1^{x'_a x'_b})}} \quad \text{when } (x_a, x_b) \neq (x'_a, x'_b) \quad (2.12)$$

Furthermore, when $n_1 p_1^{x_a x_b}$ is high enough, the following approximation can be used:

$$\log(N_1^{x_a x_b}) \approx \log(n_1 p_1^{x_a x_b}) + \sqrt{\frac{(1 - p_1^{x_a x_b})}{n_1 p_1^{x_a x_b}}} \delta_1^{x_a x_b} \quad (2.13)$$

Similar approximation is obtained for counts in unaffected population by replacing the '1' indice by '0'. We then used Equation 2.13 to derive the four odds-ratio estimators ($\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3$ and $\widehat{\varphi}_4$) with respect to the δ coefficients. The estimation of V_Φ is achieved by using δ correlation structure given by Equations 2.11 and 2.12.

2.3.3 Main results and application to GWAS

In [JP7], we evaluated the statistical properties of IndOR and compared its power against widely used statistical procedures.

Control of the type-1 error and computational cost

Given the assumptions made in the Maximum Likelihood estimation of Φ and in the use of the δ method in the estimation of the covariance matrix V_Φ , the control of the Type-1 error rate is straightforward. However, the two above mentioned assumptions rely on the asymptotic theory and are hardly verified in practice. To evaluate the robustness to these assumptions, we performed simulations under various scenarios where the null hypothesis is assumed. Results obtained in [JP7] showed an appropriate control of the Type-1 error.

Furthermore, since we derived a closed form for the estimation of Φ and V_Φ , the computational cost of IndOR is very low. As computational cost is indeed a burden to perform large-scale association studies in genetic epidemiology, the use of IndOR statistic might be very helpful in routine.

Power simulation study

The evaluation was first conducted using extensive simulations through a large power study. We focused on the effect of three main factors (the underlying interaction model, the correlation between X_1 and X_2 and the control-to-case ratio) in a total of 45 scenarios.

The results obtained in our power study demonstrate the efficiency of our statistical procedure IndOR. IndOR is indeed the most powerful test in a majority of scenarios and is comparable to the best test in 76% of the tested scenarios. The efficiency of IndOR is remarkable when the two SNPs of interest are correlated and/or when control-to-case ratios are higher than one. Furthermore, IndOR has proven its ability to detect a wider range of epistatic interactions by accounting for non-linear effects.

Application to GWAS

In [JP7], we used IndOR on the publicly available dataset from the WTCCC [WTCCC, 2007] to perform two genome-wide scans. Cases for the first and the second genome scan were individuals affected by Crohn's disease. Controls population for the first genome scan was made by the 3,000 shared controls in the original WTCCC study. For the second genome scan, control individuals were combined as individuals not affected by Crohn's disease, thus leading to a total of 15,000 controls.

Figure 2.5 displays the quantile-quantile plots obtained from our genome-wide scans. The shaded regions in the plots correspond to the 95% concentration band obtained from the null hypothesis of non-interaction (corresponding to a χ^2 test with four degrees of freedom). We can observe an excess of points outside the 95% concentration band at the tail of the distribution revealing a significant deviation from the expected distribution for IndOR. The deviation for IndOR was mainly due to a single interaction between two genomic regions. This interaction involved gene Adenomatous Polyposis Coli (*APC*) and the IQ-domain GTPase-activating protein 1 (*IQ-GAP1*). The most interacting SNP pair was made by rs6496669 on chromosome 15 and SNP rs434157 on chromosome 5. It can be remarked that IndOR showed a remarkable increase in the significance for the pair rs6496669-rs434157 (corrected p-value of 2.83×10^{-8}).

Removing all *APC-IQGAP1* SNP pairs from the analysis completely eliminated any significant p-values for IndOR with shared controls. However when the set of controls was the combination of individuals not diagnosed for Crohn's disease, one SNP pair was still globally significant for IndOR. The significant SNP pair involved SNP rs9009 on chromosome 8 and SNP rs2830075 on chromosome 21. These two SNPs belong to two genomic regions where the cathepsin B (*CTSB*) and the amyloid beta (*APP*) are located. The p-values obtained for the pair rs9009-rs2830075 are summarized in Table 2.3.

Control set	SNP1	Chr1 (Position)	SNP2	Chr2 (Position)	p-value	corr. p-value
Shared	rs6496669	15 (88696269)	rs434157	5 (112219541)	4.44×10^{-9}	1.33×10^{-3}
Combined	rs6496669	15 (88696269)	rs434157	5 (112219541)	9.42×10^{-14}	2.83×10^{-8}
Shared	rs9009	8 (11739415)	rs2830075	21 (26424313)	1.36×10^{-6}	0.40
Combined	rs9009	8 (11739415)	rs2830075	21 (26424313)	1.42×10^{-7}	0.042

Table 2.3: *P-values and corrected p-values for the two SNP pairs rs6496669-rs434157 and rs9009-rs2830075 for IndOR, when controls are shared and combined. The two SNP pairs rs6496669-rs434157 and rs9009-rs2830075 are the two most significant SNP pairs associated with Crohn's disease in the WTCCC data set.*

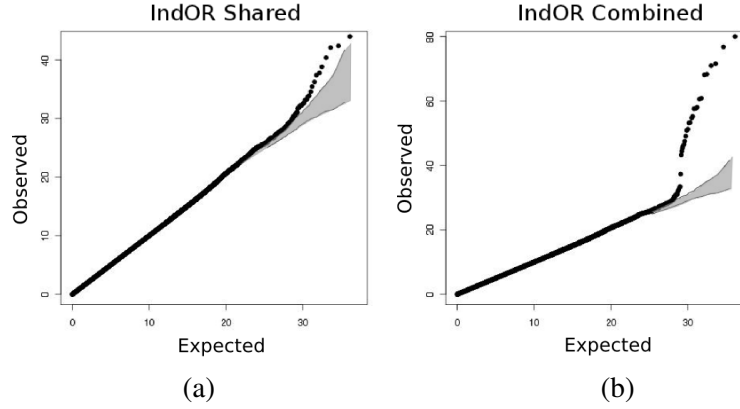


Figure 2.5: **Quantile-Quantile plots.** Figure (a) displays quantile-quantile plots obtained with the set of shared controls. Figure (b) reports QQ plots with the combined set of controls. The shaded region is the 95% concentration band, calculated assuming independent SNP pairs.

2.3.4 Concluding remarks

In this work we proposed a new statistical procedure to test for the association between a binary response variable and the interaction between two categorical variables. The method is based on an original formal expression of “association” and our results show its complementarity with other methods. Several other methods have since been proposed in the literature that are based on other definitions of association. These methods can therefore be seen as complementary methods rather than concurrent methods. However, in practice, people are not interested in testing one particular formal interpretation of association and thus, an interesting perspective is to find an appropriate combination of all these methods to produce a “meta” method.

From a practical point-of-view, our method is based on closed-form estimators which fasten its computation. In [JP7], the illustration of our method on a large scale dataset proved the feasibility of a genome-wide scan in a reasonable computational time. The method has been implemented in an R package that is available at <https://github.com/MathieuEmily/IndOR>.

2.4 Clustering in sparse two-way contingency tables

In this section, we present our work on the clustering of individuals in sparse contingency tables. This work was conducted within a collaboration with Alain Mom (Université Rennes 2 and Institut de Recherche Mathématiques de Rennes) for the statistical modeling and with Christophe Hitte (Université Rennes 1 and Institut de Génétique de Rennes) for the illustrative example.

2.4.1 Context and issue

The study of the link(s) between two categorical variables, Y with I categories and X with J categories, starts by testing the independence between X and Y [Agresti, 2013]. When independence is rejected and considering that the categories of Y correspond to subpopulations from a global population, investigating the nature of the relationship between Y and X can be performed by clustering the subpopulations (*i.e.* the Y categories) [Greenacre, 1988]. For that purpose, subpopulations are usually characterized by their conditional profile that corresponds to the distribution, for one subpopulation, of the observations over the J categories of X [Hirotzu, 2009].

Clustering subpopulations is usually based on detecting categories for X (also called features in machine learning) that are specific to sets of subpopulations. However, with the advent of high-

throughput technologies, situations where J (the number of categories for X) is of the order of n (the sample size) are becoming very common. As the dimensionality increases, the sparseness is therefore likely to be observed even when the total sample size is large. Since classical statistical frameworks are known to loose optimality as tables become sparse [Agresti and Yang, 1987], it is therefore challenging to detect conditional profiles by accounting for sparsity and specificity simultaneously. Such profiles, called *sparse-specific* profiles are characterized by two main features. Firstly, the sparse profiles are those profiles for which only very few categories have non-zero counts. Secondly, specific profiles are those profiles presenting specific categories, i.e. categories that are (almost) never observed in the other subpopulations.

The detection of *sparse-specific* profiles can be performed by using hierarchical clustering techniques. The quality of the clustering depends on the choice of (1) a dissimilarity measure between individuals and (2) a linkage criterion for the hierarchical clustering. As quoted in [Hastie et al., 2009] (p.506), “Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm”. For that reason, attention was first focused on the choice of an appropriate dissimilarity to detect *sparse-specific* profiles. An abundant literature has been dedicated to improving the measure of similarity between individuals in sparse contingency tables, either by applying dimension reduction techniques or by proposing dissimilarity measures [Aggarwal and Zhai, 2012]. Dimension reduction techniques, such as Latent Semantic Indexing (LSI) [Landauer et al., 1998] and Non-negative Matrix Factorization (NMF) [Lee and Seung, 2001] aim at transforming a high dimension space of features to a space of fewer dimensions using linear or non-linear combinations [Hastie et al., 2009]. Applying such techniques can help selecting the most relevant categories, thus improving the quality of the clustering [Witten and Tibshirani, 2010]. However, these techniques do not explicitly account for *sparse-specific* profiles in the reduction of the dimensionality of the feature space. As a consequence, power to detect *sparse-specific* profiles for dimension reduction techniques is likely to be limited. On the one hand, the similarity between individuals can be measured with many different functions developed to deal with sparse contingency tables. In the text domain, the most well known and commonly used similarity function is the cosine similarity function [Singhal, 2001]. In ecology, dedicated dissimilarities are the Bray-Curtis dissimilarity, the Jaccard dissimilarity, the d_1^2 (or Manhattan) distance, the Hellinger distance or the Gower dissimilarity [Oksanen et al., 2015].

However, these methods usually work directly with counts which might not be appropriate to detect *sparse-specific* profiles. Indeed, heterogeneity in the marginal counts of individuals gives different weights to individuals, thus leading to inappropriate conclusions. A natural way to control weights given to individuals is to focus on the conditional distribution of the categories, also known as conditional profiles. The analysis of conditional profiles is classically performed by using either the χ^2 distance or the d_2^2 distance (also known as L^2 norm). Nevertheless, capturing *sparse-specific* profile with the χ^2 distance raises some limitations since χ^2 is sensitive to profiles specificities. On the other hand, d_2^2 distance between two profiles is more influenced by the sparsity than the specificity.

2.4.2 Approach

In [JP2], we tackle the issue of detecting *sparse-specific* profiles by introducing a novel dissimilarity called d_s^2 adapted to the detection of *sparse-specific* profiles. d_s^2 is based on the comparison of conditional profiles and gives equal influence to sparsity and specificity of profiles, compared to other dissimilarities. To identify *sparse-specific* profiles, we propose a procedure called SMILE, for Statistical Method to detect *sparse-specific* profileLEs, which consists in a single-linkage hierarchical clustering [Jardine and Sibson, 1971] constructed using the d_s^2 dissimilarity. Selected profiles with the SMILE procedure correspond to the smallest subset of conditional profiles that

coalesce at the final step of the hierarchical clustering.

The dissimilarity d_s^2

Conditional profile for $Y = i$ is defined by a J -dimensional vector $y_i = [p_1^i, \dots, p_J^i]'$ where $p_j^i = n_{ij}/n_{i\cdot}$. In the following, E is used to denote the set of those I conditional profiles. Our dissimilarity d_s^2 is defined by:

Definition 2.4.1 $\forall x, y \in E$:

$$d_s^2(x, y) = \|x\|_2 \|y\|_2 d_\theta^2(x, y) \quad (2.14)$$

where:

$$d_\theta^2(x, y) = 2(1 - \cos(\widehat{xy})) = 2 \left(1 - \frac{\langle x, y \rangle_2}{\|x\|_2 \|y\|_2} \right) \quad (2.15)$$

is the square of the angular distance between the lines spanned by x and y , \langle, \rangle_2 is the L_2 scalar product and $\|\cdot\|_2$ its corresponding norm.

The sparsity of a given profile x is indeed measured by $\|x\|_2$ since the sparser x is, the higher $\|x\|_2$ is and the specificity between two profiles is measured by the angular distance d_θ . Thus, from Equation 2.14, it can be remarked that d_s^2 gives the same importance to the sparsity and the specificity since $0 \leq \|x\|_2 \|y\|_2 \leq 1$, $0 \leq d_\theta^2(x, y)/2 \leq 1$ and multiplying the dissimilarities by the same scalar does not modify the clustering.

According to Equation 2.15, d_s^2 can further be reformulated as:

$$\forall x, y \in E, d_s^2(x, y) = 2(\|x\|_2 \|y\|_2 - \langle x, y \rangle_2). \quad (2.16)$$

It can easily be remarked that d_s^2 is symmetric. Moreover, according to the Cauchy-Schwarz inequality, $\forall x, y \in E$ $d_s^2(x, y) \geq 0$, thus proving that d_s^2 is actually a dissimilarity.

Single-linkage detection

The SMILE procedure is based on single-linkage detection, where single-linkage detection corresponds to the selection of the smaller of the two subsets linked at the final step of a single-linkage hierarchical clustering constructed with some dissimilarity d . In the single-linkage hierarchical clustering, the linkage criterion between two clusters C_i and C_j is defined by $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$ [Jardine and Sibson, 1971]. The clustering is then obtained by iteratively merging the pair of clusters that minimizes the single linkage criterion.

The single-linkage method has a tendency to form long and straggly clusters. This phenomenon, often known as “chaining phenomenon”, refers to the gradual growth of a cluster as one element at a time gets added to it [Calinski and Corsten, 1985]. Considered as a potential drawback in some practical situations, the chaining effect is an advantage in our situation by allowing the separation of two highly distinct groups. Another advantage of using the single-linkage criterion is that the dissimilarities between clusters are the original dissimilarities between individuals: dissimilarities remain unchanged during the clustering, preserving their good properties, if any, all through the study. Thus, the single-linkage criterion is adapted to separate individuals with *sparse-specific* profiles from the rest of the population.

2.4.3 Main results and application to the detection of selection

Benefit of the d_s^2 in single-linkage detection

In [JP2], we first proposed an original characterization of the structure of the individual subset selected by the single-linkage detection by considering the parallel between single-linkage clustering trees and Minimum Spanning Trees (MST) introduced in graph theory [Gower and Ross, 1969].

For that purpose, we introduced a restricted version of the Kruskal algorithm that is widely used to compute MST [Kruskal, 1956] and proved theorem 2.4.1 that provides a necessary and sufficient conditions, for a set of subpopulations A to be selected at the final step of the single-linkage detection.

Theorem 2.4.1 *Let A be a set of subpopulations of E and d be a dissimilarity. A is linked to its complementary \bar{A} , at the final step of the hierarchical clustering based on d and using the single-linkage criterion if and only if A and \bar{A} are the two last connected components at the final step of the Kruskal algorithm. A is thus selected by single-linkage detection if and only if:*

$$\max(d^{MST}(A), d^{MST}(\bar{A})) < \min_{x \in A, y \in \bar{A}} d(x, y) \quad (2.17)$$

where $d^{MST}(A)$ is the length of the final edge of a restricted version of the Kruskal algorithm.

Illustrations of Theorem 2.4.1 as well as the calculation of $d^{MST}(A)$ are provided in Figure 2.6. Theorem 2.4.1 helps understanding the roles played of the within structure of A and \bar{A} in single-linkage clustering.

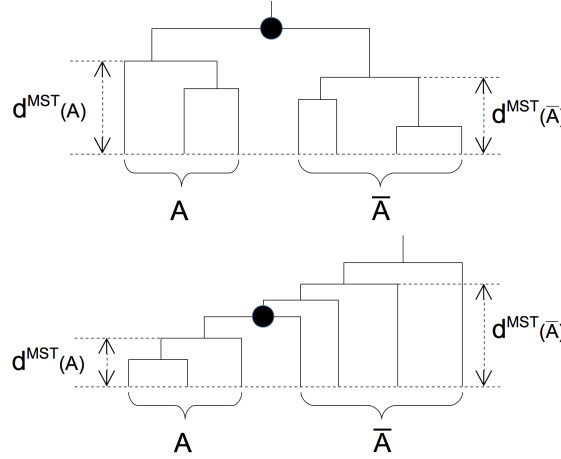


Figure 2.6: Two examples of hierarchical clustering with $n = 7$ and $n_A = 3$. The black dots represent the first level of clustering between an individual in A and an individual in \bar{A} , i.e. $\min_{x \in A, y \in \bar{A}} d(x, y)$. The two examples illustrate the calculation of $d^{MST}(A)$ and $d^{MST}(\bar{A})$. In the first example, subset A is selected by the SMILE procedure while in the second example, subset A is not selected by the SMILE procedure.

The benefits of using d_s^2 is then stated by comparing d_s^2 to the L^2 norm, called d_2^2 . We indeed proved that

$$\forall x, y \in E \quad d_2^2(x, y) = d_s^2(x, y) + (\|x\|_2 - \|y\|_2)^2. \quad (2.18)$$

It is noteworthy that compared to d_s^2 , the d_2^2 dissimilarity gives more weight to variation in sparsity between the 2 compared profiles by adding the term $(\|x\|_2 - \|y\|_2)^2$. Thus, compared to d_s^2 , d_2^2 is likely to be more sensitive to the heterogeneity, and especially heterogeneity in sparsity, in a given subset. Therefore, the detection of a targeted subset A with single-linkage detection is less influenced by the structure of \bar{A} when using d_s^2 compared to d_2^2 . By focusing on the case where A is a singleton, situations for which our dissimilarity, d_s^2 , is more powerful than d_2^2 are described in Theorem 2.4.3.

Before stating Theorem 2.4.3, further definitions are needed. Let first introduce the sparsest and the least sparse conditional profiles as follows:

Definition 2.4.2 The sparsest conditional profile, x_s , and the least sparse conditional profile, x_0 , are defined, for all $x \in E$ such that $x \neq x_s$ and $x \neq x_0$, by:

$$\|x_0\|_2 < \|x\|_2 < \|x_s\|_2.$$

Let also formalize a hierarchy in the specificity of a profile by defining a totally specific conditional profile (see Definition 2.4.3) and a nearly totally specific conditional profile (see Definition 2.4.4).

Definition 2.4.3 $x \in E$ is said to be totally specific if and only if: $\forall y \neq x \in E, \langle x, y \rangle_2 = 0$.

Definition 2.4.4 $x \in E$ is said to be nearly totally specific for the dissimilarity d if and only if $\langle x, x_0 \rangle_2 = 0$ and $d(x, x_0) = \min_{y \neq x} d(x, y)$.

We then proved the two following theorems:

Theorem 2.4.2 If x_s is nearly totally specific for the dissimilarity $d \in \{d_s^2, d_2^2\}$ then x_s is selected by single-linkage detection.

Theorem 2.4.3 If x_s is nearly totally specific for d_2^2 then x_s is nearly totally specific for d_s^2 .

If x_s is nearly totally specific for d_2^2 , then it is selected by single-linkage detection using d_2^2 . Theorems 2.4.2 and 2.4.3 thus ensure that x_s is selected by single-linkage detection using d_s^2 . It is noteworthy that the reciprocity of Theorem 2.4.3 is false. Conditional profiles selected by d_s^2 might be missed by d_2^2 , thus proving the benefit of using d_s^2 instead of d_2^2 . This advantage for d_s^2 is further highlighted by the analysis of our illustrative example.

Simulation study

We evaluated the performance of the SMILE procedure by comparing our novel dissimilarity d_s^2 to 11 other measures in single-linkage detection: 5 dissimilarities (Bray-Curtis, d_1^2 (or Manhattan), Jaccard, Gower and Hellinger), 3 distances (cosine, χ^2 and d_2^2) and 3 reduction dimension techniques dedicated to the analysis of sparse contingency tables (Sparse clustering, Latent Semantic Analysis or LSA and non-negative matrix factorization).

We proposed an algorithm for the simulation of contingency tables designed to simulate the structure of the set of selected individuals, A , and its complementary \bar{A} . Since the structure of \bar{A} plays a major role in the detection of A (see Theorem 2.4.1), simulated scenarios that focus on the impact of the structure of \bar{A} on single-linkage detection are designed. On the one hand, a simulated scenario, Scenario #1, is considered where specific and non-sparse profiles are observed in \bar{A} . On the other hand, the impact of heterogeneity in sparsity is analyzed for profiles in \bar{A} thanks to two simulated scenarios, Scenarios #2 and #3.

Our results proved that our procedure is the only method (1) able to detect *sparse-specific* profiles when specific and non-sparse profiles are observed in \bar{A} (2) not impacted by the structure in \bar{A} , coming from the fact that d_s^2 gives equal influence to sparsity and specificity of profiles.

Application to the detection of selection in domestic dogs

We used the SMILE procedure on genomic data from the European consortium LUPA [Lequarré et al., 2011] where $I = 30$ dog breeds are considered as subgroups of the population of domestic dogs. Attention is focused to six genomic regions, Regions #1, #2, ..., #6, defined as small parts of the genome. For each of the six regions of interest, the set of categories are defined as the set of observed DNA sequences, also called the set of haplotypes. The six regions of interest have been chosen as being

previously reported causative for the following morphological traits: brachicephaly, furnishings, wrinkled skin, periodic fever syndrome, chondrodysplasia and curly hair. For each region, genetic studies have shown that the presence or absence of a particular DNA sequence (or category) is associated with the observation of the trait.

These six genomic regions were considered as test regions to compare the ability for the novel dissimilarity d_s^2 and the 11 compared methods to detect known signals with single-linkage detection. For each of the six regions of interest, the true signal, *i.e.* the breed(s) that is(are) under selection, is(are) known. Thus, for the novel dissimilarity d_s^2 and the 11 competitive methods, we first evaluated the set of breeds detected in each region and then compare it with the known set of breeds that should have been detected according to biological knowledge. The performance of each method, given in Table 2.4, showed that the proposed dissimilarity d_s^2 was the only method able to correctly detect 5 regions. All other methods also failed at finding the only missed region by d_s^2 , (Region #2), thus demonstrating the strength of d_s^2 in a real situation.

Method	Regions						Total
	#1	#2	#3	#4	#5	#6	
d_s^2	x		x	x	x	x	5
d_2^2	x		x	x	x		4
Sparcl	x		x	x	x		4
NMF	x		x	x			3
LSA	x			x			2
Jaccard			x				1
Bray-Curtis			x				1
d_1^2			x				1
χ^2			x				1
Hellinger			x				1
Gower							0
Cosine							0

Table 2.4: *Summary of the results obtained for d_s^2 and the 11 compared methods on the six genomic regions used as test regions to validate the method. A x means a correct detection of the signal known from biological experiments. The last column gives the total number of signals correctly found by the corresponding method.*

We drew a parallel between results obtained on the real data set analysis and more general results on the SMILE procedure. The low power observed for several dissimilarities was first explained. Regions with only one selected breed were then focused on. Finally, results were discussed regarding regions with more than one targeted breed.

2.4.4 Final comments

In this contribution, we proposed a statistical framework to detect a particular conditional profile in data summarized by a two-contingency. The main originality of our work consists in defining the set of targeted profiles that is the formal interpretation of series of biological assumptions. Based on such assumptions we provided an appropriate metric to distinguish between profiles. Using a mixture of theoretical results, simulations and a real data example, we provided evidence that (1) our method is efficient to detect the statistical hypotheses designed to raise the biological question and (2) the statistical hypotheses characterize well the biological assumption underlying the biological question.

The computation of the method described in this section is integrated in an R package available at <https://github.com/MathieuEmily/SMILE>.

Statistical modeling of highly structured data

3.1 Introduction

The following sections propose an overview of our research work dedicated to the analysis of large-scale genomic data. Due to the emergence of high-throughput data, the increasing complexity of genomic data has raised new statistical challenges. Compared to our contributions in chapter 2, where we developed statistical methodologies for situations where only 2 or 3 variables are considered, our aim in this chapter is to consider genomic data as whole (such as whole genome sequencing data), where the number p of variables ranges from a few hundred to a few billion. Therefore, genomic data falls into the paradigm of “high-dimensional data” which is known to be a major issue for the statistical analysis [Bühlmann and van de Geer, 2011].

The impact of high-dimensionality on statistics is multiple and often refers to the “curse of dimensionality” [Donoho, 2000, Sammut and Webb, 2011]. It relates to the fact that the convergence of any estimator to the true value is very slow and that an enormous amount of observations is needed to obtain a good estimate. However, since the past few years has not seen a significant increase in sample sizes, the number of variables may exceed the sample size by several orders of magnitude. Circumventing the curse of dimensionality seems therefore to be hopeless in the analysis of whole genome sequencing data.

Fortunately, high-dimensional data are often much more low dimensional than they seemed to be and are usually concentrated around low-dimensional structures [Giraud, 2014]. These structures are due to the intrinsic low complexity of the systems producing the data and we can hope to extract useful information from them. Taking into account these structures may indeed be sufficient to overcome the issues raised by the high-dimensionality.

However, a major issue with genomic data is that these structures are, most of the time, only very partially known and must be guessed from the data themselves. The main task is therefore to identify, at least approximately, these structures. Nevertheless, one of the main consequence of the completion of the human genome is a better understanding of the global structure of the genome and many studies have showed that genomes are highly structured at various scales [Little, 2005]. At the smallest possible unit of measure, *i.e.* the nucleobase, it can be first considered that genomes have a 1-dimensional structure along the chromosome [Gabriel et al., 2002]. As shown in Figure 1.3 (see page 8), the correlation between Single Nucleotide Polymorphisms (SNPs) is local thus conferring a block structure to the genome. The ordered sequences of nucleobases also play a major in functional genomics. By reading three nucleotides at a time (the so-called codon), the transcription-translation machinery can convert DNA sequences into proteins. Knowledge of the structure and function of the proteome is central to the exploitation of the wealth of biological information available in the post-genome era [Adams, 2008]. This knowledge provides funda-

mental understanding of biological processes and can inform the systematic development of novel pharmaceuticals [Fleming et al., 2006].

In this chapter, we summarized our contributions on the statistical modeling of the complexity of genomic data that focuses on the two above levels of structure: the nucleobase level and the sequence of amino-acid level. Our contributions were motivated by (1) the test of an association between a binary variable and the interaction between 2 categorical variables (at the lowest possible level of the nucleobase) and (2) the prediction of a functional trait (amyloidogenesis) in proteins (at the functional level of a sequence of amino-acid). First, improving statistical procedures to detect interaction is crucial since interaction is commonly assumed to be one of main factor contributing to heritability. Next, amyloid proteins are associated with the pathology encountered in a range of diseases including Alzheimer's, Parkinson's and type II diabetes, all of which are progressive disorders with associated high morbidity and mortality. Our contributions address the general statistical issues of the **design of experiment** (Section 3.2), the **multiple testing correction** (Section 3.3), the **aggregation of statistical tests** (Section 3.4) and the design of **meta-predictor** (Section 3.5).

In Section 3.2, we tackle the issue of selecting the most informative set of SNPs, when searching for associations between a binary phenotype and the interaction between two-SNPs, by accounting for the 1-dimensional structure of the genome. Although such an approach has already been used to help the **design of experiment**, these studies aimed at optimizing the power of single-association testing. They may not be appropriate to the interaction issue and, in [NC3, IC7], we proposed a specific formulation and resolution of the optimization based on information theory.

In Section 3.3, we focus on the multiple testing correction issue when a large number of interaction tests are performed. Due to the correlation observed between variables, there exists a potentially important correlation between statistical tests for interaction thus complicating the correction for multiple testing. In [JP8], we proposed a statistical framework to perform multiple testing correction that is based on combining biological knowledge integration with information theory.

In Section 3.4, we address the issue of combining statistical interaction tests in order to test for interaction at the level of the gene (or region) rather than at the nucleobase level, where a gene is considered as a sequence of several nucleobases. Due to the correlation between variables within each set of nucleobases, the dimensionality of the whole set of interaction tests is much lower than the number of pairwise interaction tests that can be performed. In [JP3], we proposed a novel procedure based on the minimum p-value statistic that integrates over the covariance structure between the individual tests.

In Section 3.5, our aim is to define a statistical framework to propose a (meta-)predictor that combines predictions obtained from a set of individual predictors. When the trait of interest, such as amyloidogenesis, is an intricate phenomenon in which many features interplay, individual predictors usually account for a very few number of features, thus reducing the overall predictive capacity of each method. However since the overall set of features involved in the trait is highly correlated, individual predictors also share a large part of common information. In [JP6], we introduced a statistical procedure to select and combine individual methods into a meta-predictor in order to improve the overall prediction.

3.2 Variable selection in the design of experiment

Our research work presented in this section proposes a statistical method to select a set of informative variable to detect an association between a binary variable and the interaction between two variables. It is the results of a collaboration with Chloé Friguet.

3.2.1 Context

In the human genome, the number of SNPs reported in the famous database dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi) is now larger than 80,000,000. When considering a group of highly correlated SNPs, the information carried by an individual SNP tends to be redundant with the other SNPs. The selection of a tag SNP as a representative of these groups reduces the amount of redundancy when analyzing parts of the genome associated with traits/diseases. Figure 3.1 shows an example of a selection of tag SNPs in a small region. Therefore tag SNPs selection is a crucial step in the design of experiment.

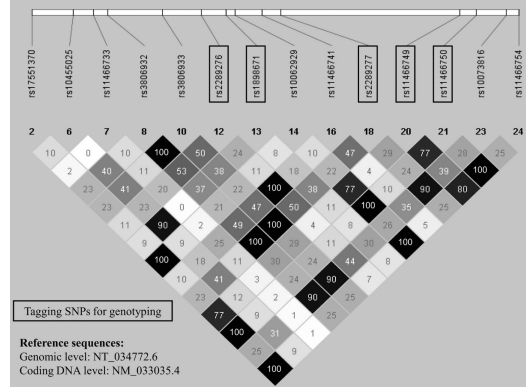


Figure 3.1: *Pairwise correlation plot (or LD plot) of a set of SNPs in TLSP gene. The level of grey indicates the pairwise r^2 values (%) in the lozenges. Selected tag SNPs are framed.*

Let consider a set of p categorical variables, each with 3 categories to mimic bi-allelic SNPs, denoted as X_1, \dots, X_p . The statistical question raised by tag SNPs selection is the identification of a maximum informative subset of variables, called $X_{t(1)}, \dots, X_{t(k)}$, where $\{t(1), \dots, t(k)\} \subseteq \{1, \dots, p\}$. Various measures have been used to quantify and maximize the information retrieved by tagging and existing methods proposed in the literature can be classified in different groups whether a partition of the p initial SNPs into contiguous blocks is assumed and whether the number of k tag-SNPs is fixed or random.

The wide majority of these measures is based on the pairwise correlation between SNPs (the r^2 linkage disequilibrium statistic). The use of r^2 is very popular because r^2 is related to statistical power to detect single association. If the true signal of association is carried by variable X_c , then the power to detect association at X_t with a sample of size n is approximately the power attained with a sample size of $r^2 \times n$ at X_c where r^2 is the correlation between X_c and X_t [Pritchard and Przeworski, 2001]. In other words, to achieve the same power with indirect association as achieved in the case of direct association, the sample size must be increased by a factor of $1/r^2$.

Furthermore, to reduce the computational cost, the search for tags is performed using a system of contiguous set of SNPs defined through the 1-dimensional block structure. Such a set of SNPs, also called neighborhood, can be based on either *a priori* biological knowledge, such as the identification of recombination hotspots, or contiguous region of observed correlated SNPs or sliding windows. Figure 3.2 provides a scheme of the cutting of a chromosome in B blocks. Once blocks are defined, the selection of tag-SNPs is performed within each block. Let consider the ℓ^{th} block ($\ell = 1, \dots, B$), denoted by \mathbb{X}_ℓ , that is composed of set of p_ℓ variables:

$$\mathbb{X}_\ell = [X_{1,\ell}, \dots, X_{p_\ell,\ell}]$$

The purpose of tagging is to find a subset of \mathbb{X}_ℓ , called \mathbb{X}_ℓ^t , that maximizes the information contains in \mathbb{X}_ℓ , where:

$$\mathbb{X}_\ell^t = [X_{t_\ell(1),\ell}, \dots, X_{t_\ell(p_\ell),\ell}]$$

Figure 3.2 provides a overview of the main notations.

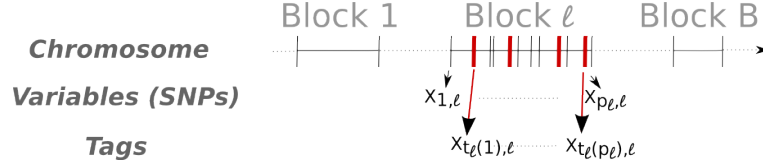


Figure 3.2: Block representation and tag-SNP selection where tag SNPs are in red.

In practice, the issue of tag selection is to find \mathbb{X}_ℓ^t that satisfies the following criterion [Carlson et al., 2004]:

$$\forall i \in [1, \dots, p_\ell], \exists j \in [t_\ell(1), \dots, t_\ell(p_\ell)] \quad / \quad r^2(X_{i,\ell}, X_{j,\ell}) \geq 0.8 \quad (3.1)$$

According to Equation 3.1, each variable of a block should be tagged with at least one tag variable with a correlation threshold of $r^2 \geq 0.8$. Solution for satisfying Equation 3.1 is not unique and additional constraints, such as minimizing $t_\ell(p_\ell)$, the dimensionality of tags or maximizing the correlation threshold, can be used to reach the “best” solution [Stram, 2004]. From a computational point-of-view, computing a maximum informative \mathbb{X}_ℓ^t is a NP-hard problem in its full general [Halldorsson et al., 2004] form and an iterative greedy algorithm is used [de Bakker et al., 2005]. Although such a strategy has proved its efficiency in single-marker association studies, it may be not adapted to the detection of interaction between SNPs for two main reasons. First, in situations where the signals is carried out by a pair of interacting variables, each variable is likely to be tagged by a tag-SNP. The correlation threshold in Equation 3.1 does not account for this “double tagging” situation that generates a “double” loss in statistical power. Next, when considering that blocks of variables are not independent, a pair of variables is not necessarily best tagged by a pair of tagged variables. Since interaction between variables is one of the most common biological phenomenon used to explain the lack of power of single-variable association, it is therefore important to account for interaction in the design of SNP arrays.

3.2.2 Our approach: EpiTag

In [NC3, IC7], we proposed a statistical method to select subset of tag variables that maximally retrieved the information of all pairs of variables between two sets of variables. Our method is based on information theory measures.

Let first define two sets of variables, \mathbb{X}_{ℓ_1} with p_{ℓ_1} variables and \mathbb{X}_{ℓ_2} with p_{ℓ_2} variables, as follows:

$$\mathbb{X}_{\ell_1} = [X_{1,\ell_1}, \dots, X_{p_{\ell_1},\ell_1}] \quad \text{and} \quad \mathbb{X}_{\ell_2} = [X_{1,\ell_2}, \dots, X_{p_{\ell_2},\ell_2}]$$

To select the most informative subset of variables, we consider \mathcal{X} , the set of all possible pairs of variables between \mathbb{X}_{ℓ_1} and \mathbb{X}_{ℓ_2} :

$$\mathcal{X} = \{(X_{i,\ell_1}, X_{j,\ell_2}), \forall i \in [1, \dots, p_{\ell_1}], \forall j \in [1, \dots, p_{\ell_2}]\}.$$

Our aim is to select a subset of \mathcal{X} , called \mathcal{T} , so that each element of \mathcal{X} is tagged by at least one element of \mathcal{T} at a given threshold τ . We further assume that \mathcal{T} is a cross product between two sets of tag variables (one for \mathbb{X}_{ℓ_1} and one for \mathbb{X}_{ℓ_2}) so that \mathcal{T} can be written as follows:

$$\mathcal{T} = \{(X_{i,\ell_1}, X_{j,\ell_2}), \forall i \in [t_{\ell_1}(1), \dots, t_{\ell_1}(p_{\ell_1})], \forall j \in [t_{\ell_2}(1), \dots, t_{\ell_2}(p_{\ell_2})]\}.$$

To measure the information of a couple of variables retrieved by another couple of variables, we used the Normalized Mutual Information. The objective of our method EpiTag can therefore be defined as follows:

$$\forall (i, j) \in \{1, \dots, p_{\ell_1}\} \times \{1, \dots, p_{\ell_2}\}, \quad \exists (r, s) \in \{p_{\ell_1}(1), \dots, p_{\ell_1}(p_{\ell_1})\} \times \{p_{\ell_2}(1), \dots, p_{\ell_2}(p_{\ell_2})\} /$$

$$NMI((X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})) \geq \tau \quad (3.2)$$

where:

$$NMI[(X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})] = \frac{I[(X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})]}{\sqrt{H(X_{i,\ell_1}, X_{j,\ell_2})H(X_{r,\ell_1}, X_{s,\ell_2})}}$$

and

$$I[(X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})] = \sum_{(x_i, x_j, x_r, x_s) \in \{0,1,2\}^4} p(i, j, r, s) \log \left(\frac{P(i, j, r, s)}{\mathbb{P}[(X_{i,\ell_1}, X_{j,\ell_2}) = (x_i, x_j)] \mathbb{P}[(X_{r,\ell_1}, X_{s,\ell_2}) = (x_r, x_s)]} \right)$$

$$H(X_{i,\ell_1}, X_{j,\ell_2}) = I[(X_{i,\ell_1}, X_{j,\ell_2}), (X_{i,\ell_1}, X_{j,\ell_2})]$$

$$p(i, j, r, s) = \mathbb{P}[(X_{i,\ell_1}, X_{j,\ell_2}, X_{r,\ell_1}, X_{s,\ell_2}) = (x_i, x_j, x_r, x_s)].$$

Thus, NMI is the Normalized Mutual Information between the pair of variables $(X_{i,\ell_1}, X_{j,\ell_2})$ and the other pair $(X_{r,\ell_1}, X_{s,\ell_2})$ while I and H respectively corresponds to the mutual information and the entropy measure.

To implement EpiTag, we used a greedy algorithm that starts by selecting the pair of variables that cover the maximum number of pairs. Then the tag and its covered variables are excluded from the dataset. At each iteration, the most informative pair of variable is therefore the one that covers the most couple of variables that have not been selected. Given the following f function:

$$\forall (i, j) \in U_1 \times U_2, \quad f(X_i, X_j) = \sum_{(k,\ell) \in U_1 \times U_2} \mathbb{1}_{NMI[(X_{i,\ell_1}, X_{j,\ell_2}), (X_{k,\ell_1}, X_{\ell,\ell_2})] \geq \tau}$$

where $U_1 \subseteq \{1, \dots, p_{\ell_1}\}$ is the subset of unselected variables X_{i,ℓ_1} and $U_2 \subseteq \{1, \dots, p_{\ell_2}\}$ the subset of unselected variables X_{j,ℓ_2} at the current step. At the current iteration, the selected couple of variables is therefore defined as:

$$(T_1, T_2) = \underset{(X_i, X_j) \in U_1 \times U_2}{\operatorname{argmax}} f(X_i, X_j)$$

Iterations stop when all variables are selected.

3.2.3 Results

In [NC3, IC7], we compared our procedure EpiTag to two other methods of variable selection: NoTag and Tagger [de Bakker et al., 2005]. In NoTag, we consider that no variable selection is performed so that all variables are tested for association in the analysis. Tagger is the classical method used for tagging SNP and aims at proposing a solution for Equation 3.1. Tagger is therefore a selection method only based on the 1-dimensional structure of the genome and may thus be appropriate for detecting interaction.

Our comparison focuses on the statistical power to detect the interaction between two variables. In the following paragraphs we considered two sets of variables \mathbb{X}_{ℓ_1} and \mathbb{X}_{ℓ_2} , and assumed a dominant-dominant model for simulating Y . Let $(X_{i,\ell_1}, X_{j,\ell_2})$ be the associated pair (or causal pair) of variables with Y as follows:

$$\operatorname{logit}(\mathbb{P}(Y = 1 | (X_{i,\ell_1}, X_{j,\ell_2}) = (x_i, x_j))) = \alpha + \beta(\mathbb{1}_{x \in \{Aa, aa\}}(x_i) \mathbb{1}_{x \in \{Aa, aa\}}(x_j)) \quad (3.3)$$

We then performed all pairwise testing between a variable in $\mathbb{X}_{\ell_1}^t$ and a variable in $\mathbb{X}_{\ell_2}^t$, where $\mathbb{X}_{\ell_1}^t$ (resp. $\mathbb{X}_{\ell_2}^t$) is the subset of tag variables for \mathbb{X}_{ℓ_1} (resp. \mathbb{X}_{ℓ_2}). It is noteworthy that the subset $\mathbb{X}_{\ell_1}^t$ and $\mathbb{X}_{\ell_2}^t$ depend on the selection method (NoTag, Tagger or EpiTag). To test the interaction between $X_{r,\ell_1} \in \mathbb{X}_{\ell_1}^t$ and $X_{s,\ell_2} \in \mathbb{X}_{\ell_2}^t$ we used the following likelihood ratio test:

$$\text{LRT}(X_{r,\ell_1}, X_{s,\ell_2}) = D(\mathcal{M}_0(X_{r,\ell_1}, X_{s,\ell_2})) - D(\mathcal{M}_1(X_{r,\ell_1}, X_{s,\ell_2})) \sim_{\mathcal{H}_0} \chi^2(4)$$

where D is the deviance computed for the two models \mathcal{M}_0 et \mathcal{M}_1 :

$$\mathcal{M}_0 : \text{logit} \left[P(Y = 1 | (X_{r,\ell_1}, X_{s,\ell_2}) = (x_i, x_j)) \right] = \beta_0 + \beta_1 \mathbb{1}_{x=Aa}(x_i) + \beta_2 \mathbb{1}_{x=aa}(x_i) + \beta_3 \mathbb{1}_{x=Aa}(x_j) + \beta_4 \mathbb{1}_{x=aa}(x_j)$$

$$\begin{aligned} \mathcal{M}_1 : \text{logit} \left[P(Y = 1 | (X_{r,\ell_1}, X_{s,\ell_2}) = (x_i, x_j)) \right] = & \beta_0 + \beta_1 \mathbb{1}_{x=Aa}(x_i) + \beta_2 \mathbb{1}_{x=aa}(x_i) + \beta_3 \mathbb{1}_{x=Aa}(x_j) + \beta_4 \mathbb{1}_{x=aa}(x_j) \\ & + \beta_5 \mathbb{1}_{x=Aa}(x_i) \mathbb{1}_{x=Aa}(x_j) + \beta_6 \mathbb{1}_{x=aa}(x_i) \mathbb{1}_{x=Aa}(x_j) \\ & + \beta_7 \mathbb{1}_{x=Aa}(x_i) \mathbb{1}_{x=aa}(x_j) + \beta_8 \mathbb{1}_{x=aa}(x_i) \mathbb{1}_{x=aa}(x_j) \end{aligned}$$

We declared that a method successfully detected the causal signal if at least one pair is significant after a Benjamini-Hochberg correction for multiple testing.

Simulation

In this paragraph we simulated two regions, each composed of 6 variables following the scheme displayed in Figure 3.3. Variables are simulated according to a complete joint probability distribution where the marginal probabilities are fixed to $(0.6^2, 2 \times 0.6 \times 0.4, 0.4^2)$ for each variable, thus corresponding to SNPs in Hardy-Weinberg Equilibrium with a Minor Allele Frequency of 0.4. Furthermore, pairwise correlation between SNPs within a region is fixed to $r^2 = 0.8$.

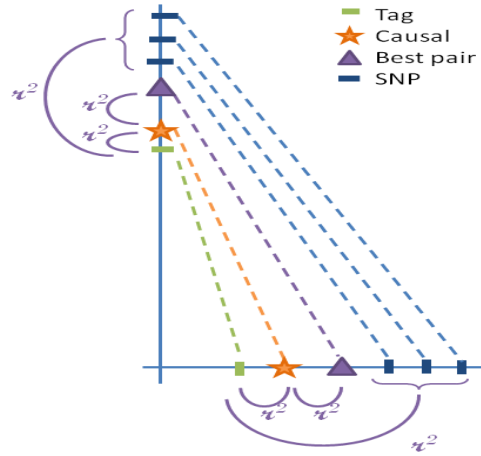


Figure 3.3: *Simulation scheme. For each region, one variable is causal (in orange), one variable is the Tag obtained with Tagger (in green), one variable is the tag obtained with EpiTag (purple triangle) and 3 others are linked to the tag (in blue).*

We performed simulations with values for $\beta \in [0, 0.5]$, thus investigating different effect sizes. As shown in Figure 3.4, EpiTag is the most powerful method for all value of β .

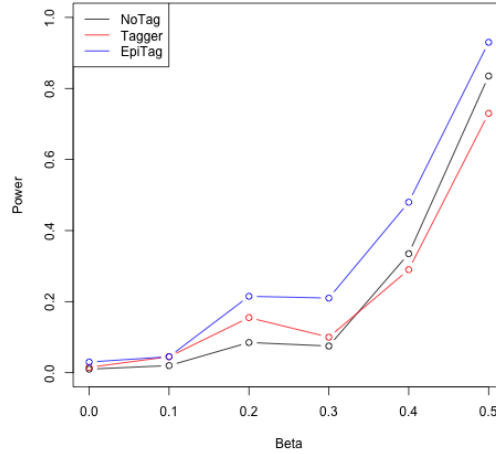


Figure 3.4: *Power for NoTag, Tagger and EpiTag with respect to the effect size β .*

Power based on a reference panel

In [NC3, IC7], we performed a second simulation study based on a realistic pattern of correlation between SNPs. For that purpose, we used the GWASimulator [Li and Li, 2008] to simulate 1,000 datasets that mimic the structure of the genome (observed in the CEU population of HapMap Phase 3 [Consortium, 2003]) in two regions of 1 megabase. For each of the 1,000 simulated datasets, 2,000 individuals (1,000 cases and 1,000 controls) have been simulated, and one pair have been randomly selected to be directly associated with Y . We used Equation 3.3, to simulate Y and tuned the parameters so that the power for detecting the association is 0.5.

In our comparison, we used different thresholds for Tagger ($r^2 \geq 0.8$, $r^2 \geq 0.9$ and $r^2 \geq 0.95$) and EpiTag ($NMI \geq 0.7$ and $NMI \geq 0.8$) for selecting subsets of variables. Results obtained in Table 3.1 show that EpiTag is the best trade-off between power and false discovery proportion. Although the NoTag strategy is the most powerful method, it suffers from a very high false discovery proportion.

It can further be remarked that power is very low for Tagger when using a threshold lower than 0.9. Such a result suggests that accounting only for the 1-dimensional structure in the selection of variable is not appropriate to the detection of interaction.

Method	Threshold	Number of tests	Power	False discovery proportion
NoTag		484	0.47	0.45
Tagger	$r^2 \geq 0.95$	252	0.40	0.35
Tagger	$r^2 \geq 0.90$	165	0.0	0.28
Tagger	$r^2 \geq 0.80$	126	0.0	0.27
EpiTag	$NMI \geq 0.8$	121	0.47	0.38
EpiTag	$NMI \geq 0.7$	74	0.45	0.29

Table 3.1: *Power and false discovery proportion estimated from 1,000 simulations for the 6 compared methods.*

3.2.4 Concluding remarks

In this work, we proposed an information theory based method to select the most informative set of SNPs that optimizes the power of detecting interaction between two categorical variables. Our results demonstrate the importance, when dealing with heterogeneous variables where several scales of structure interplay, of accounting for the various levels of correlation. In variable selection, driving the selection by the 1-dimensional structure may optimize power for single association testing to the detriment of higher order of association.

It is therefore ideal to **design the experiment**, here selecting variables, with respect to **hypothesis testing**. EpiTag has been implemented in an R package that is available at <https://github.com/MathieuEmily/EpiTag>.

3.3 Multiple testing correction based on the interactome

Our research work presented in this section proposes a statistical method to use biological knowledge in the search for interacting variables and validate a procedure for correcting multiple comparisons. It is the results of a collaboration with Mikkel Schierup (University of Aarhus, Denmark), Thomas Mailund (University of Aarhus, Denmark), Leif Schauser (University of Aarhus, Denmark) and Jotun Hein (University of Oxford, United Kingdom).

3.3.1 Context

In the previous section, we demonstrated that the experimental design of SNP arrays used in genome-wide association studies may not be optimal to detect an association between a binary variable Y and the interaction between two variables. However, experimental design is not the only factor limiting the power of detecting such an association. Considering the variable selection performed in commercial SNP array, such as the Affymetrix GeneChip 500k Mapping Array Set used in the WTCCC data set [WTCCC, 2007], the number of variables is of the order of 500,000. Testing exhaustively all pairs of variables leads to a total of $\approx 1.25 \times 10^{11}$ tests.

Such a large number of tests is a challenge both statistically and computationally. Statistically, it implies that significant tests after a trivial Bonferroni correction for multiple testing should have p-values lower than 4×10^{-13} . Due the stringency of the correction, such a p-value is very unlikely and the large majority of the true signals of interaction are missed. Furthermore, although it is computationally possible to perform 125 billion tests, these tests have to be very simple to be run in a reasonable time even on large CPU clusters.

However, as commonly suggested in high-dimensional design, the dimensionality of the data is likely to be much lower than the dimensionality of the observed data [Giraud, 2014]. To reduce the dimensionality of the data, statistical methods can be used to learn and account for the correlation structure [Bühlmann and van de Geer, 2011]. In the context of interaction, the correlation structure is even more complex than the 1-dimensional correlation of the genome and guessing such a structure remains highly challenging. As an alternative to the statistical inference of the structure of the data, dimension reduction can also be performed by using *a priori* knowledge.

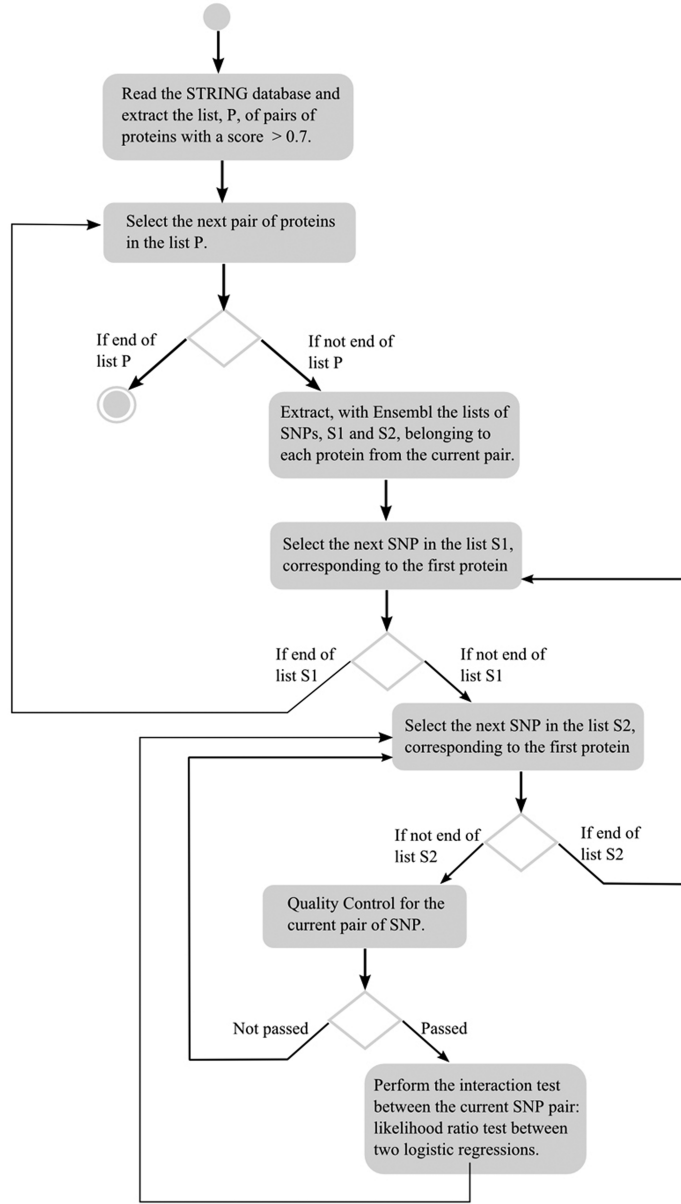
For example, in molecular biology, interaction between molecules belonging to different biochemical families (proteins, nucleic acids, lipids, carbohydrates, etc.) are particularly investigated to define interactomes [Bonetta, 2010]. Most commonly, interactome refers to protein-protein interaction, where physical interactions among proteins are studied. Prioritizing the search for interacting variables can therefore help reducing the complexity of the data. However, even after prioritization, data are still highly structured and issues regarding correction for multiple remains.

3.3.2 Approach

In [JP8], we proposed an approach to combine the biological and statistical perspectives of detecting an association between Y and the interaction between two variables. We first postulate that two genes that biologically interact are good candidates to a statistical analysis. For that purpose we have reduced the search to variables belonging to gene pairs known to interact and referenced in protein databases. We first described the bioinformatic pipeline used to extract variables of interest. Next we developed the statistical analysis performed in our analysis.

Bioinformatic pipeline

Before the statistical testing of association, a series of bioinformatic steps were performed to extract tested variables. The bioinformatic pipeline is summarized in a flowchart shown in Figure 3.5. In a first step, we used the STRING database to select approximately 71,000 potential protein-protein interactions that we wanted to test [von Mering et al., 2007]. For each pair of protein, we selected the two lists of SNPs located within the genes coded for two proteins. A quality control is then performed to remove variables with a poor genotyping quality. Finally we tested all pairs of good variables for the association of their interaction with Y .

Figure 3.5: *Flowchart of the main steps of our statistical pipeline.*

Statistical procedure

The statistical procedure is composed of two main steps. First, all pairwise testing between SNPs are performed sequentially. In a second step, we developed a multiple testing correction to provide a genome-wide significant level of testing.

Let first consider a pair of variables, $(X_{r,\ell_1}, X_{s,\ell_2})$, where X_{r,ℓ_1} (resp. X_{s,ℓ_2}) is a variable of the first (second) protein of the ℓ^{th} protein pair. To test the interaction between X_{r,ℓ_1} and X_{s,ℓ_2} we used the following likelihood ratio test:

$$\text{LRT}(X_{r,\ell_1}, X_{s,\ell_2}) = D(\mathcal{M}_0(X_{r,\ell_1}, X_{s,\ell_2})) - D(\mathcal{M}_1(X_{r,\ell_1}, X_{s,\ell_2})) \sim_{\mathcal{H}_0} \chi^2(4)$$

where D is the deviance computed for the two models \mathcal{M}_0 et \mathcal{M}_1 :

$$\begin{aligned}\mathcal{M}_0 : \text{logit} \left[P(Y = 1 | (X_{r,\ell_1}, X_{s,\ell_2}) = (x_i, x_j)) \right] &= \beta_0 + \beta_1 \mathbb{1}_{x=Aa}(x_i) + \beta_2 \mathbb{1}_{x=aa}(x_i) + \beta_3 \mathbb{1}_{x=Aa}(x_j) + \beta_4 \mathbb{1}_{x=aa}(x_j) \\ \mathcal{M}_1 : \text{logit} \left[P(Y = 1 | (X_{r,\ell_1}, X_{s,\ell_2}) = (x_i, x_j)) \right] &= \beta_0 + \beta_1 \mathbb{1}_{x=Aa}(x_i) + \beta_2 \mathbb{1}_{x=aa}(x_i) + \beta_3 \mathbb{1}_{x=Aa}(x_j) + \beta_4 \mathbb{1}_{x=aa}(x_j) \\ &\quad + \beta_5 \mathbb{1}_{x=Aa}(x_i) \mathbb{1}_{x=Aa}(x_j) + \beta_6 \mathbb{1}_{x=aa}(x_i) \mathbb{1}_{x=Aa}(x_j) \\ &\quad + \beta_7 \mathbb{1}_{x=Aa}(x_i) \mathbb{1}_{x=aa}(x_j) + \beta_8 \mathbb{1}_{x=aa}(x_i) \mathbb{1}_{x=aa}(x_j)\end{aligned}$$

In the second step, we adjusted p-values for multiple comparisons, to estimate the significance level of interaction. In [JP8], we proposed to apply a Bonferroni-like correction based on the effective number of SNP pairs. In our network-based approach, there are two levels of dependencies. First, for a particular pair of genes, each SNP from the first gene is tested against each SNP from the second gene. Second, gene pairs are not independent, as one gene can belong to more than one gene pair. Although the second source of dependency might have an impact on the significance level, we accounted only for the multiple comparisons arising in a single gene pair test. If n_{GG} is the number of gene pairs and $p_{\ell_1}^i$ and $p_{\ell_2}^i$ are respectively the number of variables in the first and second gene of the pair i , then the total number of tests, denoted by N , is given by:

$$N = \sum_{i=1}^{n_{GG}} p_{\ell_1}^i \times p_{\ell_2}^i.$$

To account for the dependency between tests, we propose to estimate the number of effective tests in a single gene pair i , denoted by n_{eff}^i and to use it in the above formula in place of $p_{\ell_1}^i \times p_{\ell_2}^i$. The effective number of tests was calculated using the eigen-values of a correlation matrix, where the correlation between two pairs of SNPs can be measured with the entropy and the mutual information. If we considered two pairs of variables $(X_{i,\ell_1}, X_{j,\ell_2})$ and $(X_{r,\ell_1}, X_{s,\ell_2})$ the correlation is given by:

$$\text{Cor}\left((X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})\right) = \frac{\text{Cov}\left((X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})\right)}{\sqrt{\text{Var}(X_{i,\ell_1}, X_{j,\ell_2}) \times \text{Var}(X_{r,\ell_1}, X_{s,\ell_2})}}$$

where

$$\begin{aligned}\text{Var}(X_{i,\ell_1}, X_{j,\ell_2}) &= H(X_{i,\ell_1}, X_{j,\ell_2}) \\ \text{Cov}\left((X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})\right) &= I\left((X_{i,\ell_1}, X_{j,\ell_2}), (X_{r,\ell_1}, X_{s,\ell_2})\right) \\ &= H(X_{i,\ell_1}, X_{j,\ell_2}) + H(X_{r,\ell_1}, X_{s,\ell_2}) - H(X_{i,\ell_1}, X_{j,\ell_2}, X_{r,\ell_1}, X_{s,\ell_2})\end{aligned}$$

with H the entropy measure and I the mutual information measure. As described by Li and Ji [Li and Ji, 2005], letting λ_k ($k = 1, \dots, p_{\ell_1}^i \times p_{\ell_2}^i$) the eigen-values of the correlation matrix of the variables, the number of effective tests in the gene pair i , n_{eff}^i is given by:

$$n_{\text{eff}}^i = \sum_{k=1}^{p_{\ell_1}^i \times p_{\ell_2}^i} f(|\lambda_k|)$$

with $f(x) = \mathbb{1}_{x>1}(x) + (x - \lfloor x \rfloor)$, where $\mathbb{1}$ is the indicator function and $\lfloor x \rfloor$ is the floor of x . In [JP8], we estimated the number of effective pairs of variables as follows:

$$N_{\text{eff}} = \sum_{i=1}^{n_{GG}} n_{\text{eff}}^i \quad (3.4)$$

An effective correction for multiple testing consists in multiplying the p-values by N_{eff} .

3.3.3 Results and applications to the WTCCC dataset

Performance of the multiple testing correction procedure

In [JP8], we tested the efficiency of our correction for multiple comparisons on simulated data sets based on the WTCCC data. Ten thousand gene pairs were randomly generated in the set of genes from the STRING database. For each of the 10,000 gene pairs, the number of effective pairs was calculated with the procedure described in the previous section and compared with the total number of pairs that is used in the conventional Bonferroni correction. Type I error rate at the 5% level showed that a Bonferroni correction is overly conservative: we estimated that the probability of rejecting the null hypothesis of non-interaction to be 0.8%. It proved that LD structure within genes induces dependency between SNP pairs, lowering the power to detect epistasis. The use of the effective number of pairs gave a better correction, improving the power to detect interaction, and we estimated that the probability of rejecting the null hypothesis at a 5% level was 4.5%; still conservative but much less than the Bonferroni correction.

Search for interacting variable in susceptibility with Crohn's Disease (CD)

After applying some data quality filter, we were left with approximately 3,500,000 tests for Crohn's disease (CD) in the WTCCC data. Figure 3.6 shows the quantile-quantile plots for the interaction tests using the 71,000 well-established protein-protein interactions in the STRING database. The shaded region in the plots corresponds to the 95% concentration band obtained from the null hypothesis of non-interaction. The computation of the number of effective tests gave $N_{eff} \approx 580,000$ thus providing a genome-wide significant level of 8.62×10^{-8} . Consistent with the quantile-quantile plots, CD showed a strong interaction with a p-value of $1,13 \times 10^{-9}$, yielding an overall p-value of 6×10^{-4} after correction.

We indeed observed an excess of points outside the 95% concentration band at the tail of the distribution. In total, eight SNP pairs showed a significant interaction that belongs to the same putative biological interaction. This interaction involves genes Adenomatous Polyposis Coli (APC) and the IQ-domain GTPase-activating protein 1 (IQGAP1). Removing all APC-IQGAP1 SNP pairs from the analysis completely eliminates the deviation from the expected Q-Q plot (see the blue points in Figure 3.6).

3.3.4 Concluding remarks

In this work we developed a method that drives the search for testing an association between a categorical variable using biological networks. Using elements of information theory we proposed an original method for multiple testing correction. Our results showed that an appropriate **design of experiment** may help circumventing the burden of **multiple testing** in high-dimensional testing.

3.4 Combining statistical tests

Our research work presented in this section proposes a statistical method to aggregate interaction tests of an association between a binary variable and the interaction between two variables.

3.4.1 Context

In the previous section, we focused on the multiple testing correction of a complete set of p-values obtained from a large number of interaction tests. However, as seen previously, the genome is structured through a multilevel scheme where the 1-dimensional structure allows defining contiguous blocks of correlated variables. In single association testing, such a 1-dimensional has been used to propose association tests at the level of the block. In contrast to single variable approach,

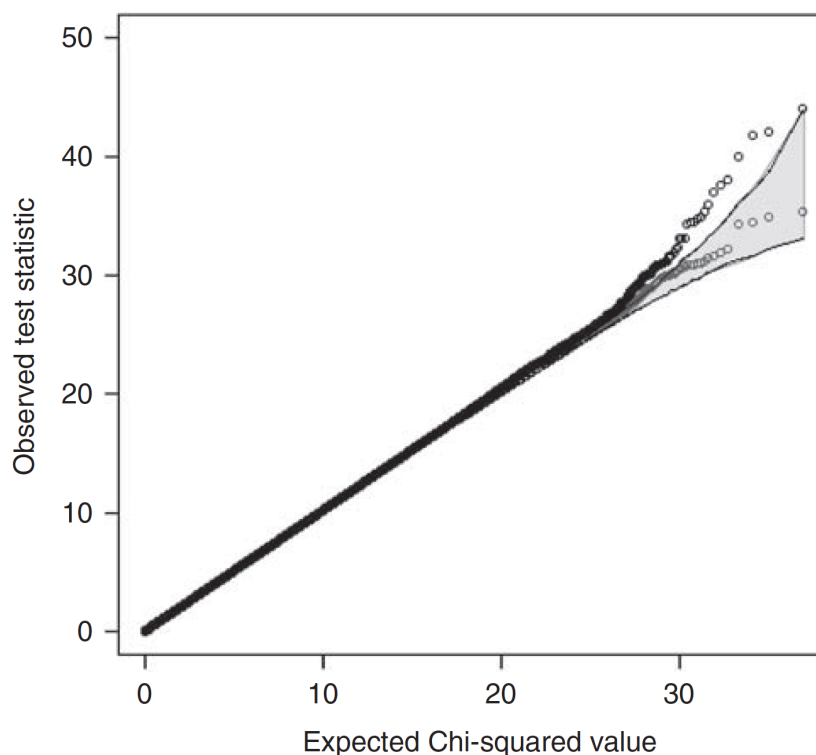


Figure 3.6: *Quantile-quantile plots of the test statistic observed for the CD. The black dots correspond to the entire data set. The blue dots result from the removal of SNP pairs included in APC-IQGAP1 regions. The shaded region shows the 95% concentration band for the non-interaction hypothesis.*

block-level testing can help characterizing functional, compositional and statistical interactions [Phillips, 2008]. Such tests allow for all the variables within the block to be jointly modeled as a set and can take into account the correlation structure within a block [Huang et al., 2011]. Single block-based strategy has been successfully applied in many applications and more specifically in the detection of an association between Y and X with low probabilities. Furthermore the use of the block (or gene) as the statistical unit can greatly facilitate the biological interpretation of findings [Jorgenson and Witte, 2006, Neale and Sham, 2004]. In the context of interaction, by aggregating signals across variables in a block, statistical power is likely to be increased in situations when multiple interactions are associated with Y [Wu et al., 2010]. Furthermore, if the interacting variables are only tagged, rather than directly observed, block-based tests can aggregate signals from different tag variables. Therefore, block-based block-block interaction methods have recently grown in popularity.

In case-control studies, where Y is dichotomous, principal component analysis (PCA) has first been used to test the association between synthetic variables (*i.e.* principal components) from each block [Li et al., 2009]. In another approach, Peng *et al.* proposed a U-statistic, called CCU, to measure the difference of correlation between two blocks in cases and controls [Peng et al., 2010]. In CCU, correlations in cases and controls is based on canonical correlation analysis in order to detect block-block co-association [Peng et al., 2010]. Although CCU has good performances, it is limited to detect linear correlation, which may be unsuitable for finding nonlinear signals. To overcome this limitation, CCU has been extended to KCCU, where correlation is estimated by kernel canonical correlation [Yuan et al., 2012, Larson et al., 2014]. Kernel-based methods have also been successfully adapted to block-block interaction via kernel regression [Larson and Schaid, 2013].

Partial Least Squares Path Modeling (PLSPM) has also been proposed as an alternative measure of correlation between two blocks and proved its efficiency when the two blocks are linked [Zhang et al., 2013]. Rather than focusing on a single measure of correlation between blocks, Rajapakse *et al.* proposed a test to compare the whole covariance structure between two blocks in cases and controls [Rajapakse et al., 2012]. More recently, a non-parametric statistic, called GBIGM and based on information theory, has been introduced as an attractive option to detect non-linear relationship between two blocks [Li et al., 2015]. All these methods assume that the modeling of the joint distribution of SNPs within and between the two blocks has to be the initial step of the statistical procedure.

3.4.2 Our approach: AGGrEGATOR

In [JP3], rather than considering multiple variables in both block as part of a joint model, we proposed an alternative strategy, called AGGrEGATOR for A Gene-based GEne-GEne interActiOn test for case-control association studies, that aims at aggregating p-values obtained at the SNP level into a test at the gene level. We start by applying a logistic regression model to test all pairs of SNPs between the two blocks. We then used a minP procedure to combine the p-values into a single test at the block level. Since the distribution of the minP statistic depends on the correlation between the combined statistics, we proposed an estimation of the correlation between variable-variable interaction statistics. The various steps of the AGGrEGATOR framework are illustrated in Figure 3.7.

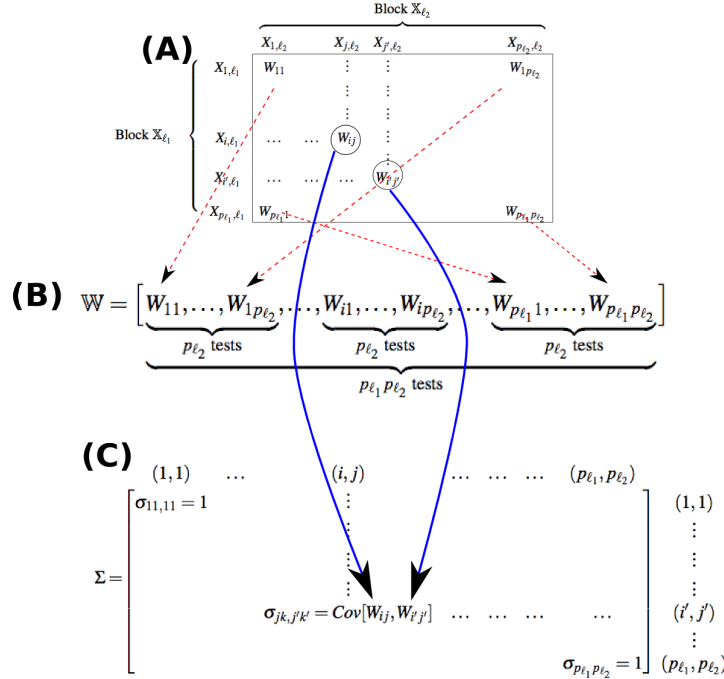


Figure 3.7: Overview of the different steps of the AGGrEGATOR procedure. In step (A) W_{ij} statistics are computed for all pairs of SNPs between the 2 genes (X_1 and X_2). Step (B) consists in unfolding the matrix of statistics into a vector \mathbb{W} of $p_{\ell_1} \times p_{\ell_2}$ statistics. In step (C), the covariance matrix of vector \mathbb{W} is computed.

Let consider two blocks of variables \mathbb{X}_{ℓ_1} and \mathbb{X}_{ℓ_2} such as:

$$\mathbb{X}_{\ell_1} = [X_{1,\ell_1}, \dots, X_{p_{\ell_1},\ell_1}] \text{ and } \mathbb{X}_{\ell_2} = [X_{1,\ell_2}, \dots, X_{p_{\ell_2},\ell_2}],$$

where p_{ℓ_1} and p_{ℓ_2} are the number of variables in each block. We further assumed that each $X_{i,j}$ is a discrete random variable with values in $\{0; 1; 2\}$ that corresponds to the number of copies of the minor allele.

Pairwise interaction test

In [JP3], we used a standard logistic regression to model the association between the two variables, X_{i,ℓ_1} and X_{j,ℓ_2} , and the response variable Y .

$$\text{logit}[\mathbb{P}(Y = 1 | X_{i,\ell_1} = x_1, X_{j,\ell_2} = x_2)] = \beta_0^{i,j} + \beta_1^{i,j} x_1 + \beta_2^{i,j} x_2 + \beta_3^{i,j} x_1 x_2 \quad (3.5)$$

where $[\beta_0^{i,j}, \beta_1^{i,j}, \beta_2^{i,j}, \beta_3^{i,j}]$ are the regression coefficients and $\beta_3^{i,j}$ is interpreted as the weight of the interaction between the two variables. The interaction between the two variables is then tested by means of the following statistical null and alternative hypotheses:

$$\mathcal{H}_0^s : \beta_3^{i,j} = 0 \quad \text{and} \quad \mathcal{H}_1^s : \beta_3^{i,j} \neq 0.$$

To test \mathcal{H}_0^s against \mathcal{H}_1^s , we used the following Wald statistic:

$$W_{ij} = \frac{\widehat{\beta}_3^{i,j}}{\sigma(\widehat{\beta}_3^{i,j})} \quad (3.6)$$

where $\widehat{\beta}_3^{i,j}$ is an estimate of $\beta_3^{i,j}$ and $\sigma(\widehat{\beta}_3^{i,j})$ an estimate of its standard deviation.

Our block-based interaction test: minP

When applying pairwise interaction tests to all variable pairs between two blocks with p_{ℓ_1} and p_{ℓ_2} variables, we obtained $p_{\ell_1} \times p_{\ell_2}$ p-values (see Figure 3.7 (A)). The main goal of our method is to combine the $p_{\ell_1} \times p_{\ell_2}$ p-values into a single block-block interaction test. To do so, we proposed to test the null hypothesis \mathcal{H}_0 against the alternative \mathcal{H}_1 where:

$$\begin{aligned} \mathcal{H}_0 : \quad & \forall 1 \leq i \leq p_{\ell_1} \text{ and } \forall 1 \leq j \leq p_{\ell_2}, \quad \beta_3^{i,j} = 0, \\ \mathcal{H}_1 : \quad & \exists (i, j) \text{ where } 1 \leq i \leq p_{\ell_1} \text{ and } 1 \leq j \leq p_{\ell_2}, \quad \beta_3^{i,j} \neq 0. \end{aligned}$$

As described in Equation 3.6, the $p_{\ell_1} \times p_{\ell_2}$ p-values are related to $p_{\ell_1} \times p_{\ell_2}$ Wald statistics $W_{ij}(i = 1 \dots p_{\ell_1}, j = 1 \dots p_{\ell_2})$. It is well known that under \mathcal{H}_0 :

$$\mathbb{W} = [W_{11}, \dots, W_{p_{\ell_1} p_{\ell_2}}] \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where $\mathcal{N}(\mathbf{0}, \Sigma)$ is the multivariate normal density with mean $\mathbf{0}$, the $p_{\ell_1} \times p_{\ell_2}$ null vector, and covariance matrix Σ . $\Sigma = [\sigma_{(i,j),(i',j')}]_{\substack{i=1\dots p_{\ell_1}; j=1\dots p_{\ell_2} \\ i'=1\dots p_{\ell_1}; j'=1\dots p_{\ell_2}}}$ is a $(p_{\ell_1} \times p_{\ell_2}) \times (p_{\ell_1} \times p_{\ell_2})$ symmetric matrix

where $\sigma_{(i,j),(i',j')} = \text{Cov}(W_{ij}, W_{i',j'})$.

To combine a set of p-values, we compared the maximum of the absolute values for the observed Wald statistics to the asymptotic distribution expected under \mathcal{H}_0 . More precisely we compute the probability minP that at least one absolute value for Wald statistics is as large as the

maximum of the observed absolute values under the null hypothesis. Let $\mathbb{Z} = [Z_1, \dots, Z_{p_{\ell_1 p_{\ell_2}}}]$ be a multivariate Gaussian random vector with the following distribution $\mathbb{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and $W_{\max} = \max\{|W_{11}|, \dots, |W_{p_{\ell_1 p_{\ell_2}}}| \}$ be the maximum of the absolute values for the observed Wald statistics. Thus, the minP probability is obtained by the following formula:

$$\text{minP} = 1 - \mathbb{P}\left[\max(|Z_1|, |Z_2|, \dots, |Z_{p_{\ell_1 p_{\ell_2}}}|) < W_{\max}\right]. \quad (3.7)$$

Since our pairwise interaction test is two-sided, one can remark that $W_{\max} = \Phi^{-1}(1 - P_{\min}/2)$, where Φ is the standard normal distribution function and P_{\min} the minimum of the observed p-values. Equation (3.7) is then equivalent to the one proposed by Conneely and Boehnke in [Conneely and Boehnke, 2007].

Estimation of the variance-covariance matrix: Σ

Because of the LD between SNPs within a gene and since a SNP is used in many different pairs, the W_{ij} are correlated. By integrating over the multivariate normal distribution, Equation (3.7) explicitly accounts for Σ , the covariance structure of the W_{ij} . However, the estimation of Σ is not straightforward simply because the computation of W_{ij} in Equation (3.6) does not have a closed form. In order to compute Equation (3.7), we proposed in [JP3] an estimation of Σ based on the correlation (or LD) between variables.

Let $r_{i,i'} = \frac{p_{ii'} - p_i p_{i'}}{\sqrt{p_i(1-p_i)p_{i'}(1-p_{i'})}}$ be the widely used correlation measure. We proposed that:

$$\sigma_{(i,j),(i',j')} = \text{Cov}(W_{ij}, W_{i',j'}) \approx r_{i,i'} r_{j,j'} \quad (3.8)$$

3.4.3 Main results and application to Rheumatoid Arthritis

In [JP3], we evaluated the performance of our procedure AGGrEGATOr compared to 6 previously published methods: CCA for Canonical Correlation Analysis [Peng et al., 2010], KCCA for Kernel Canonical Correlation Analysis [Yuan et al., 2012, Larson et al., 2014], PCA for Principal Component Analysis based method [Li et al., 2009], CLD for Composite Linkage Disequilibrium [Rajapakse et al., 2012], PLSPM for Partial Least Square Path Modeling [Zhang et al., 2013] and GBIGM for Gene-Based Information Gain Method [Li et al., 2015].

The computation and comparative evaluations of the methods has been performed via the R package GeneGeneInter developed in [PP1].

Evaluation of type-I error rate

To investigate the control of the type-I error rate, we focused on three main disease models (No effect, One marginal effect and Multiplicative marginal effect) that express the relationship between the two blocks of variables and Y . As proposed in [Marchini et al., 2005], disease models were presented by a 3×3 table of odds where each cell characterizes the odds of the disease with respect to the genotype. Each model had two parameters: γ characterizes the baseline odds and θ quantifies the strength of the model.

As shown in Table 3.2, AGGrEGATOr is a valid statistical method for detecting block-block interaction and confirms that the estimation of the correlation matrix is accurate when the correlation between the two tested blocks is low. Furthermore, we proved in [JP3] that AGGrEGATOr is robust to various correlation patterns within the two blocks of interest.

Power studies

In [JP3], to evaluate the power of our minP procedure, we first considered disease models where the interaction between only one pair of variables was causal and then we investigated scenarios where 2, 5 and 10 pairs of variables were simulated as causal. In the case of one causal pair,

Models	α	θ	AGGrEGATOr	CCA	KCCA	CLD	PCA	PSLPM	GBIGM
No effect	0.05		0.061	0.046	0.043	0.043	0.071*	0.055	0.059
	0.01		0.010	0.009	0.009	0.010	0.017	0.009	0.047*
One marginal effect	0.05	1	0.053	0.051	0.067*	0.051	0.039	0.040	0.064
	0.01	1	0.011	0.010	0.008	0.009	0.005	0.008	0.063*
	0.05	4	0.058	0.055	0.070*	0.062	0.058	0.067*	0.052
	0.01	4	0.009	0.013	0.006	0.011	0.012	0.017	0.051*
Multiplicative marginal effects	0.05	1	0.047	0.043	0.049	0.062	0.054	0.029*	0.061
	0.01	1	0.012	0.015	0.012	0.01	0.011	0.006	0.059*
	0.05	4	0.041	0.040	0.084*	0.132*	0.067*	0.348*	0.104*
	0.01	4	0.009	0.014	0.028*	0.024*	0.014	0.302*	0.102*

Table 3.2: *Estimation of the false positive rate in several scenarios involving three disease models based the pair Locus 1-Locus 2: a model with no effect, a model with one marginal (recessive) effect and a model with two (multiplicative) marginal effects. α is the expected predefined type-I error rate and θ is the parameter of the disease. Results with an * indicate a significant deviation from the expected false positive rate.*

we focused on 8 disease models, all characterized by a 3×3 table of odds [Marchini et al., 2005] as described in the above section *Evaluation of the type-I error rate*. The 8 disease models, previously investigated in other studies [Li and Reich, 2000, Li et al., 2015] [JP7], were chosen to cover a wide spectra of epistatic models. We considered historical epistatic models (dominant-dominant, recessive-recessive and recessive-dominant models), and more sophisticated epistatic models (interaction multiplicative effect [Marchini et al., 2005], threshold [Neuman and Rice, 1992], XOR [Li and Reich, 2000], additive-additive [Li and Reich, 2000] and special interaction model [Li and Chen, 2008]). For each disease model, power was estimated from 1,000 simulations and for different $\theta \in [0, 5]$.

In [JP3], we remarked that under the 8 disease models with only one causal variable, AGGrEGATOr always outperformed the other methods. Our results demonstrated that, compared to the other methods, AGGrEGATOr has the capacity to accurately identify a wide range of interaction signals. Furthermore, AGGrEGATOr is the only method that is robust to the correlation pattern between and within blocks of interest.

When we considered a disease model with two causal pairs of variables, AGGrEGATOr outperformed the other methods in presence of interaction multiplicative and dominant-dominant effects. However, power for the AGGrEGATOr procedure was slightly lower than for CCA, KCCA, CLD and PCA when 10 causal pairs are involved in the disease model. Hence, our results prove, that CCA, KCCA, CLD and, to a lesser extent, PCA can aggregate multiple source of interaction more efficiently than AGGrEGATOr. Nevertheless, even for 10 causal pairs, our procedure AGGrEGATOr has reasonable power to detect gene-gene interaction. It can also be remarked that AGGrEGATOr has very similar power for other correlation structures while the other methods seemed to be sensitive to the correlation pattern.

Real data analysis

To assess the capacity of AGGrEGATOr to deal with real case-control phenotype, we first investigated the susceptibility of a set of pairs of genes (or blocks) to Rheumatoid Arthritis (RA). For doing so, we used the GSE39428 data set for which genotyping was performed using a custom-designed Illumina 384-SNP VeraCode microarray (Illumina) to determine possible associations of 17 genes to RA [Chang et al., 2013, Li et al., 2015]. The data contains 266 cases and 163 controls. We further used the WTCCC data set as a replication cohort [WTCCC, 2007]. In the WTCCC data set, 2,000 RA patients and 3,000 controls were genotyped in the British population using the Affymetrix GeneChip 500k Mapping Array Set.

In a second analysis, we aimed at replicating gene-gene interactions in susceptibility with complex diseases. For that purpose, we selected a list of publications that reported gene pairs statistically associated with three complex diseases: Rheumatoid Arthritis (RA), Crohn's Disease (CD) and Coronary Artery Disease (CAD). Prior to the analysis, SNPs within genes have been filtered with respect to Hardy-Weinberg Equilibrium, missing data and Minor Allele Frequency in the WTCCC data set [WTCCC, 2007]. After filtering, a total of 15 gene pairs, reported in 5 different publications, has been tested [Li et al., 2009, Jung et al., 2009, Peng et al., 2010, Liu et al., 2011, Musameh et al., 2015].

The application of the AGGrEGATOr procedure to the association between Rheumatoid Arthritis and 17 genes revealed a potential gene-gene interaction between PADI4 and CA1. Moreover, AGGrEGATOr was able to replicate 7 over 15 previously reported gene pairs associated with Rheumatoid Arthritis, or Crohn's Disease or Coronary Artery Disease. Again, additional investigation is needed to confirm the role played by these gene-gene interactions in the etiology of targeted diseases. The statistical replication of several gene pairs first confirmed the capacity for AGGrEGATOr to be a robust and valid method compared to competitive methods and also gives promising new insights in the etiology of Rheumatoid Arthritis, Crohn's Disease and Coronary Artery Disease.

3.4.4 Conclusion

In this contribution, we proposed a new statistical procedure to test for an association between a binary variable and the interaction between two categorical variables. Our method is based on a multinormal integration of individual tests so that the association is tested at the scale of a set of variables rather than at the scale of the variable. Our results proved that putting biological information, such as the block structure of the genome, in **hypothesis testing** may be efficient. Using such information in the statistical procedure is also crucial to correctly interpret the obtained results.

The method has been implemented in a Bioconductor R package that is available at <https://bioconductor.org/packages/devel/bioc/html/GeneGeneInter.html> and at <https://github.com/MathieuEmily/GeneGeneInter>.

3.5 Meta-prediction

Our research work presented in this contribution proposes a statistical framework to combine individual predictor into a meta-predictor. It is the result of a collaboration with Christian Delamarche (University of Rennes 1).

3.5.1 Context

In the previous sections, we focused on categorical variables, X , that are assumed to follow a multinomial distribution with three categories, in order to mimic SNPs data. Various methodologies developed in the previous sections relied on this modeling of data and can hardly be generalized to any kind of categorical variables. However for many omic data, variables have more than three categories. For example, in proteomics, a sequence is a set of amino acids where an amino acid can be modeled by a categorical variable with 20 categories.

One of the main goals of proteomics is the study of the structures and functions of proteins, since understanding protein function is one of the keys to understand life at the molecular level. Because amino acid sequence determines protein structure and protein structures dictate biochemical function, the study of protein sequences is crucial to give insights into functions of the pro-

teome. Furthermore, although three-dimensional structure is known to play important functional roles, many biological processes are assumed to be sequence-specific. Therefore, one of the big challenges in proteomics is to predict whether a given sequence of amino-acid is associated with the observation of a studied biological process. Studying the relationships between biological function and amino acid sequence is also important in the context of human disease because many conditions arise as a consequence of alterations of protein function due to modifications (mutations, insertion, deletion, etc.) of the native sequence.

For many applications, the analysis of an amino-acid sequence is performed by first calculating a score for each amino-acid. A score profile is then used to summarize the whole sequence. Based on this profile, a global score or predictive function is obtained to estimate the functional outcome of the sequence.

In that context, let introduce a sequence of length p , denoted by \mathbb{X} , as follows:

$$\mathbb{X} = [X_1, \dots, X_p]$$

where each X_i is a categorical variable with 20 categories in $[Alanine, \dots, Tyrosine]$ that corresponds to the set of amino acids that can be detected in mass spectrometry. For each X_i , a score $S(X_i)$ is given as an indicator of the propensity of X_i (or the subsequence in the vicinity of X_i) to give a targeted biological function, denoted by Y , to the whole protein. For any \mathbb{X} , a score can therefore be defined as a p -dimensional vector as follows:

$$S(\mathbb{X}) = [S(X_1), \dots, S(X_p)].$$

An example of such a profile is displayed in 3.8. A predictor of the targeted biological function Y is then characterized as a function f of \mathbb{X} , i.e in a general form:

$$\hat{Y} = f(S(\mathbb{X}))$$

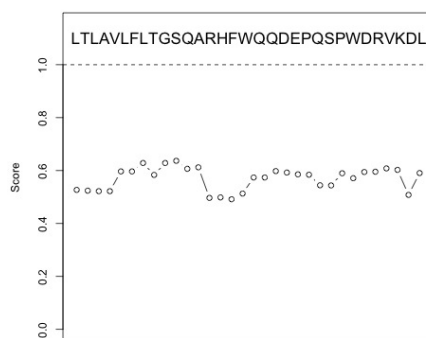


Figure 3.8: *Example of a sequence profile for sequence $\mathbb{X} = \text{LTLAVLFLTGSQARHFWQQDEPQSPWDRVKDL}$ of length 32. A score is assigned to each X_i (or amino-acid) of $\mathbb{X} = [X_1, \dots, X_{32}]$.*

A large number of features is involved in the functions of a protein, such as the its three-dimensional structure or its hydrophobicity. These features are commonly assumed to be correlated and/or to interact so that the functionality of a protein is an intricate phenomenon in which many features interplay. Since the amino-acid sequence is a key element in most of these features, it is very common that, for a given Y , numerous score functions, S , and predictor functions, f , are

proposed. However, existing predictors individually account for a very few number of features, thus reducing the overall predictive capacity of each method thus conferring them a lack of robustness. Therefore, there is a need for developing sequence-based predictors that can embrace the complexity of the targeted biological process.

3.5.2 Our approach: MetAmyl

In [JP6], we proposed statistical procedure, called MetAmyl: a METa-predictor for AMYloid proteins, to account for a wide range of features involved in a biological process. We developed a metapredictor that aims at selecting and combining a set of scores (used in individual predictors) into new score function. Although our method is dedicated to the detection of amyloid proteins, a biological process detailed in the next section, the procedure is sufficiently general to be applied in other contexts.

Let Y be a binary outcome variable and S_1, \dots, S_ℓ , a collection of ℓ score functions. We assumed that score functions are applied to a window $W_i = [X_{[i-k]}, \dots, X_{[i+k]}]$, with a fixed size k and centered in X_i . The size k of the window is usually chosen according to some biological knowledge. In [JP6], we proposed a new score function S_i^{meta} as a linear combination of the S_i :

$$S^{\text{meta}}(W_i) = \beta_0 + \sum_{j=1}^{\ell} \beta_j S_j(W_i) \quad (3.9)$$

The estimation of the linear combination is achieved through a logistic regression model and is decomposed into two main steps. In a first step, we automatically selected the most informative and complementary set of individual predictors using a stepwise procedure. Variable selection is commonly used in supervised classification to alleviate the effect of the curses of dimensionality and to enhance generalization by reducing overfitting [Venables and Ripley, 2002, Hastie et al., 2009]. In a second step, the weights assigned to each predictors, *i.e.* the regression coefficients in equation 3.9, are estimated by maximizing the likelihood of the logistic regression model. It is noteworthy that the second step performed simultaneously with the final step of the stepwise procedure. To be estimated, our methodology relies on an appropriate training dataset that is crucial to avoid overfitting.

We further proposed a thresholding method to predict the outcome Y with respect to W_i . For that purpose, we introduced τ so that the prediction function for a given W_i is:

$$f(W_i) = \mathbb{1}(S^{\text{meta}}(W_i) > \tau)$$

The threshold τ is estimated by maximizing the distance to the upper-left corner in the ROC curve, which corresponds to the cut-off that maximizes the quantity $(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2$. To prevent from overfitting, we used a leave-one-out cross-validation to estimate τ . We finally proposed a predictor of protein sequence \mathbb{X} as follows:

$$f(\mathbb{X}) = \begin{cases} 0 & \text{if } \forall W_i / f(W_i) = 0 \\ 1 & \text{if } \exists W_i / f(W_i) = 1 \end{cases}$$

3.5.3 Results and application to amyloidogenesis

A score function for predicting amyloid fibrils

In [JP6], we applied our methodology to the detection of amyloid fibrils that are protein aggregates insoluble and resistant to protease activity *in vivo* [Jiménez et al., 1999]. The formation and the accumulation of amyloid aggregates, as implicated in the cellular death process, are common

features of a variety of neurodegenerative diseases such as Parkinson's, Alzheimer's, and Huntington's diseases [Ross and Poirier, 2004, Chiti and Dobson, 2006]. Extensive researches have shown a large number of biological mechanisms involved in amyloidogenesis. Mutations, maturation, protein synthesis errors, inappropriate proteolysis and protein environment modification might lead to the formation of amyloid fibrils [Dobson, 2004]. Because of the complexity of amyloidogenesis, predicting the capacity for a given protein to form amyloid fibrils remains as of today a very challenging task.

Since it has been experimentally demonstrated that the length of six amino acids, corresponding to hexapeptides, is essential and sufficient for a segment to induce amyloid conversion of an entire protein domain [Ventura et al., 2004, Meng et al., 2012], the past few years have seen the development of a large number of methods dedicated to the prediction of amyloid hot spots in proteins. We therefore focused on the meta-prediction of 6-amino-acids length window, each W_i is composed of $k = 6$ variables. The large number of predictive methods reflects the complexity of the biological mechanisms involved in amyloidosis. It is very likely that the formation of amyloid fibrils is an intricate phenomenon in which many features interplay (secondary structures formation, disorder propensity, hydrophobicity, structural modeling energy, physico-chemical properties, amino-acid context). In [JP6], we combined well-known existing methods into a meta-predictor using the methodology described in the previous section and obtained the following score:

$$S_i^{\text{Meta}}(W_i) = \beta_0 + \beta_1 S_{\text{PAFIG}}(W_i) + \beta_2 S_{\text{SALSA}}(W_i) + \beta_3 S_{\text{Waltz}}(W_i) + \beta_4 S_{\text{FA1}}(W_i)$$

where

$$\beta_0 = -0.047727784, \beta_1 = 3.667188941, \beta_2 = 4.944766967, \beta_3 = 0.005114034, \beta_4 = -0.413373395$$

and PAFIG [Tian et al., 2009], SALSA [Zibae et al., 2007], Waltz [Maurer-Stroh et al., 2010] and FA1 [Garbuzynski et al., 2010] are the selected individual predictors.

Accuracy of the prediction

The evaluation of MetAmyl on three independent datasets (the training, the amyloyme and htt^{NT} datasets) revealed its accuracy to predict amyloidogenic segments in polypeptide chains and/or proteins. On the training dataset, MetAmyl has a significantly higher AUC, Accuracy and Matthews correlation coefficient than the other predictors (see Figure 3.9). Moreover, on the amyloyme dataset, MetAmyl has the best Q value, Matthews correlation coefficient and F1 score. The potential overfitting for MetAmyl on the training dataset has been controlled by the use of cross-validation which is enhanced by MetAmyl performance on the amyloyme subset and the htt^{NT} dataset. Although it is based on a large number of experimentally validated amyloid regions, the amyloyme subset suffers from a lack of validated non-amyloid regions, which can affect the results of performance calculations. For this reason, we used a third test set, named htt^{NT} and independent of the training dataset the amyloyme subset. The htt^{NT} dataset has been chosen as being unbiased with regard to the correct assignment of amyloid or non-amyloid.

Investigation of the human proteome

In [IC12], we investigated our score function in the prediction of amyloidogenic motifs in human proteins. Since the variability between human proteins is important, notably in their respective size, we further introduced a new measure of the amyloidogenesis of a protein, called *NHSA* for Normalized Hot Spot Area. *NHSA* is defined as the sum of the difference between each variable, X_i , and the threshold τ divided by the total length of the sequence, where only positive variables

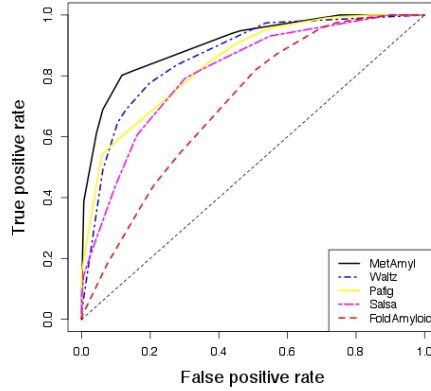


Figure 3.9: *Receiver operating characteristic (ROC curves) obtained for the 4 selected predictors, PAFIG, SALSA, FAI and Waltz, and leave-one-out cross validated MetAmyl on the training dataset.*

$(X_i > \tau)$ are considered. The general form of *NHSA* for a given sequence $\mathbb{X} = [X_1, \dots, X_p]$ with p variables is given by:

$$NHSA(\mathbb{X} = [X_1, \dots, X_p]) = \frac{1}{p} \sum_{i=1}^p \mathbb{1}_{x>\tau}(X_i)(X_i - \tau)$$

We computed *NHSA* for the 67,153 proteins of observed in humans. Our results, displayed in Figure 3.10, do not reveal islets or bias in the localization of genes related to the *NHSA* values.

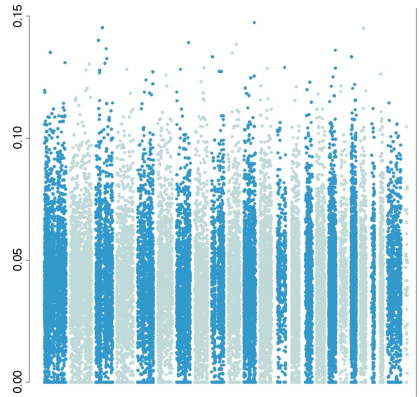


Figure 3.10: *Normalized Hot Spot Area of the 67,153 Human proteins.*

3.5.4 Concluding remarks

In [JP6], we proposed a statistical framework to integrate individual predictors into a meta-predictor. Our results demonstrated that, in the context of a complex trait such as amyloidogenesis, merging existing predictors allows accounting for a broad scale of features in a single predictor. In this contribution, rather than focusing on the **correlation** between features we aggregate methods that individually account for a very few number of features.

Our method is available online at the following url: <http://metamyl.genouest.org/> and its implementation is dedicated to the issue of detecting amyloidogenic profiles. For an input amino-acid sequence, an amyloidogenic profile is computed by the use of a sliding window with a fix number of 6 amino acids. The implementation has been designed to manipulate large scale datasets. We first computed the 64,000,000 hexapeptides scores corresponding to the combinatorial diversity of amino-acids and we stored them on the server. Thus, the building of the profile of an input sequence consists in uploading scores from the server instead of calculating it for each window, which accelerates the computation of our profiles.

Probabilistic modeling and statistical inference of spatial and time-to-event data

4.1 Introduction

In the following sections we propose an overview of our contributions in the probabilistic modeling of complex biological data and associated statistical inference techniques. Standard statistical inference and estimation theory rely on the modeling of observed data (categorized, ordered, discrete or continuous quantities) by random variables. Compared to our work presented in the previous chapters, we focus here in the statistical analysis of observed data that are not straightforwardly modeled by commonly used variables. In chapters 2 and 3, we indeed proposed statistical procedures to handle with data that are modeled by random variables with a relatively simple probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space (*i.e.* the set of all possible outcomes), \mathcal{F} is the set of events and \mathbb{P} the probability function. First, response variables Y were considered as Bernoulli variables, to mimic the case/control outcome for example, where the probability space is straightforward:

$$\{\Omega = \{0, 1\}; \mathcal{F} = \{\emptyset, 0, 1, 0 \cup 1\}; \mathbb{P}[Y = 1] = 1 - \mathbb{P}[Y = 0] = p\}.$$

Other variables, X , were categorical variables, such as SNP data with 3 categories, haplotype data or amino-acid data with 20 categories. The probability spaces for such data are also well known. Therefore, in the previous chapters, we more focused on the modeling of the multidimensionality of the data rather than on the probabilistic modeling.

However in many situations, biological data cannot be modeled by “simple” probabilistic random variables. For example, it is very common to study the geographic variation of individuals or features. More specifically, in the context of the spatial distribution of health outcomes, spatial epidemiology is concerned with the description and examination of spatial disease pattern in consideration of several risk factors such as demographic, environmental or genetic factors [Elliott and Wartenberg, 2004].

Another example is the study of the dependence of a response variable over time. For instance, environmental epidemiology is concerned with the study of environmental exposures that contribute to or protect against injuries, illnesses. The identification of public health and health care actions to manage the risks [Pekkanen and Pearce, 2001]. Data are typically available at regular time intervals (*e.g.*, daily pollution levels and daily mortality counts) and the aim is to explore short-term associations between them. In clinical trial, longitudinal experiments are designed to answer specific questions about biomedical or behavioral interventions, such as new treatments. During the trial, investigators recruit patients with the predetermined characteristics, administer the treatment(s) and collect data on the patients’ health for a defined time period.

Specific probabilistic objects have been developed to provide statistical frameworks for analyzing spatial data and time-dependent data. On the one hand, the theory of spatial point process is devoted to the probabilistic modeling of a set of points in general spaces, and especially points in geographical space, with a random process. On the other hand, time series models have been successfully developed to analyze time-dependent data using stochastic processes. In both cases, the probability space, especially Ω and \mathcal{F} can be complex. Therefore the definition of the probability function, \mathbb{P} , is crucial to provide formal assumptions of the biological processes under study. An appropriate definition of \mathbb{P} is therefore essential to provide efficient statistical frameworks related to hypothesis testing.

The complexity of the probability space requires specific development for answering traditional statistical questions such as goodness-of-fit, simulation-based inference, clustering, model selection, etc. In this chapter, we first tackle in section 4.2 the statistical issue of clustering spatial data in 2-dimension when covariates can be accounted for. We then proposed, in section 4.3 an estimation procedure related to a specific formulation of the probability function \mathbb{P} in the context of 2-dimensional spatial data. Finally, when dealing with temporal data, we introduced in section 4.4 a family of probability functions in the area of the evaluation of risk disease in public health.

4.2 Clustering in 2D spatial point process

Our research work presented in this section addresses the issue of detecting cluster(s) of points in a 2-dimensional space. It results from a collaboration with Avner Bar-Hen (University of Paris Descartes) and Nicolas Picard (CIRAD).

4.2.1 Spatial point process and context

In many fields, the observation of a configuration of points (e.g., trees in a forest, disease cases, or biological cells in a tissue) provides meaningful insights regarding the biological processes underlying the observed pattern. The overall study of such a configuration requires the analysis of the (1) location of points, (2) the distances between pairs of points but also (3) all possible k -uplets of points. In the pioneering example, designed by John Snow, of cholera cases in the London epidemic of 1854, the location of points was important since they were closed to a specific pump [Snow, 1855]. Furthermore, the inter distance between points was also crucial since points were abnormally closed together, thus characterizing a cluster of points. It is therefore necessary to define a configuration as a unique but complex object characterizing the whole set of points. From a statistical point-of-view, observing a configuration x should be interpreted as observing the realization of a random variable X .

Definition

For that purpose, spatial point processes have been introduced as a probabilistic environment to study pattern of points observed in 2-dimensional space (or in general d -dimensional space). Let $S \subseteq \mathbb{R}^2$ be a complete, separable metric space where the metric is the usual Euclidean distance. Let x be an observed configuration of points and $n(x)$ be the total number of points of x . For each bounded set $B \subset S$, x_B is the configuration of points of x restricted to B : $x_B = x \cap B$, so that $n(x_B)$ is the number of points of x in B .

The space of locally finite point configurations, N_{lf} , can be defined as follows:

$$N_{lf} = \{x \subseteq S : n_x(B) < \infty, \forall \text{ bounded } B \subseteq S\}$$

Let further consider that S is equipped with the Borel sigma algebra \mathcal{B} and let denote \mathcal{B}_0 the class of bounded Borel sets. The space N_{lf} can be equipped with the following sigma algebra:

$$\mathcal{N}_{lf} = \sigma\left(\{x \in N_{lf} : n(x_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0\right)$$

that is \mathcal{N}_{lf} is the smallest sigma algebra generated by the sets:

$$\{x \in N_{lf} : n(x_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0$$

where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Definition 4.2.1 A point process X on S is a measurable mapping defined on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ taking values in $(N_{lf}, \mathcal{N}_{lf})$. This mapping induces a distribution \mathbb{P}_X of X given by:

$$\mathbb{P}_X(F) = \mathbb{P}(\{\omega \in \Omega : \Phi(\omega) \in F\}), \forall F \in \mathcal{N}_{lf}.$$

In practice, the process lives in some subset W of \mathbb{R}^2 and patterns are only observed in a bounded area $S \subset W$, such as the experiment design plots in agronomy or the core of a tissue in histology.

In point process theory, the distribution of a point process X can be, equivalently, identified by three main characterizations: the finite dimensional distribution of X , the void probability and the generating functional of X . For example, the void probability of a bounded set $B \subseteq S$ is given by:

$$\nu(B) = \mathbb{P}(n_X(B) = 0), \quad \forall B \in \mathcal{B}_0.$$

Theorem 4.2.1 The distribution of a simple point process X , for which realizations does not contain coincident point, on S is uniquely determined by its void probabilities of bounded Borel sets $B \in \mathcal{B}_0$.

In other words, if two point processes share the same void probabilities, then they are equal in distribution.

The central role of the spatial Poisson process

Poisson point processes play a fundamental role in the theory of point processes. They possess the property of “no interaction” between points or “complete spatial randomness”. As such, they are practically useless as a model for a spatial point pattern as most spatial point patterns exhibit some degrees of interaction among the points. However, they serve as reference processes when summary statistics are studied and as a building block for more structured point process models.

Let $S \subseteq \mathbb{R}^2$ be a metric space. Poisson point processes are defined according to an intensity function $\lambda : S \rightarrow [0, \infty)$ that is locally integrable:

$$\int_B \lambda(x) dx < \infty \quad \forall \text{ bounded } B \subseteq S.$$

Intensity function is used to define the intensity measure, μ , of a Poisson point process as follows:

$$\mu(B) = \int_B \lambda(x) dx \quad \forall B \subseteq S$$

A formal definition of Poisson point process can be given as in the following definition [4.2.2](#).

Definition 4.2.2 A point process X on S is a Poisson point process with intensity function λ if the two following properties hold:

1. For any $B \subseteq S$, such as $\mu(B) < \infty$, $n_X(B)$ has a Poisson distribution with parameter $\mu(B)$ ($n_X(B) \sim \mathcal{P}(\mu(B))$).
2. For any $n \in \mathbb{N}$, $B \subseteq S$ such that $0 < \mu(B) < \infty$, knowing that $n_X(B) = n$, the point process X_B is made by n points i.i.d. with a density function given by $f(x) = \lambda(x)/\mu(B)$.

So, if S is bounded, this gives us a simple way to simulate a Poisson process on S . First draw $n_X(B)$ according the Poisson distribution with parameter $\lambda(B)$. Next draw $n_X(B)$ independent points uniformly on S .

Furthermore, the Poisson point process has nice tractable mathematical properties. First, it can be easily characterized using the void probabilities. According to the Poisson distribution, for a Poisson process with intensity function λ , we have:

$$\forall \text{ bounded } B \subseteq S, \nu(B) = \mathbb{P}(n_X(B) = 0) = \exp(-\mu(B)) \quad (4.1)$$

Furthermore, Proposition 4.2.1 gives other characterization of a Poisson point process.

Proposition 4.2.1 *Let X be a point process on S .*

1. X is a Poisson point process with intensity function λ if and only if $\forall B \subseteq S$, such as $\mu(B) = \int_B \lambda(x)dx$ and $\forall F \subseteq N_{lf}$:

$$\mathbb{P}(X_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B \mathbb{1}_{x_1, \dots, x_n \in F} \prod_{i=1}^n \lambda(x_i) dx_1, \dots, dx_n. \quad (4.2)$$

2. Let assume that X is a Poisson point process with intensity function λ . For any function $h : N_{lf} \rightarrow [0, \infty)$ and any $B \subseteq S$, such as $\mu(B) < \infty$:

$$\mathbb{E}[h(X_B)] = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B \mathbb{1}_{x_1, \dots, x_n \in F} h(x_1, \dots, x_n) \prod_{i=1}^n \lambda(x_i) dx_1, \dots, dx_n. \quad (4.3)$$

4.2.2 Context of clustering

The spatial distribution of tree species in forests, such as tropical forests for example, is known to be driven by ecological processes. The spatial pattern of a given population of a tree species can indeed be viewed as the result of interactions between the biology of the population and other ecological processes on the abiotic and biotic environments of the population. Describing the spatial distribution of a tree species, as a product of these processes, is an important tool for understanding its dynamic. Since tree locations are random in a natural forest, point process modeling is a classical approach to study these processes [Cressie, 1993, Diggle, 1983, Ripley, 1988, Stoyan et al., 1995, Stoyan and Stoyan, 1994, Daley and Vere-Jones, 1988].

In practice, exploration of a point process classically begins by descriptive statistics. These exploratory analyses rely on first testing complete spatial randomness (CSR) [Diggle, 1983]. Such a test allows the testing of the independence between points (or trees) and their marginal uniform distribution. If CSR is not rejected, the observed configuration of points is likely to be a realization of a Poisson point process.

However, when CSR is rejected (*i.e.* when the underlying point process is not a Poisson point process), one main purpose of point process modeling is the identification of spatial clusters that characterize spatial areas exhibiting a high concentration of events or points. Since detecting spatial pattern of events is essential in many fields (medicine, cosmology with spatial clustering of galaxies, social sciences and criminology, agronomy and more), a substantial literature has been

dedicated to the issue of spatial clustering [Murray et al., 2014]. The two most popular cluster detection approaches are the spatial scan statistics [Kulldorff and Nagarwalla, 1995, Kulldorff, 1997, Patil and Taillie, 2004, Tango and Takahashi, 2005, Duczmal and Assuncao, 2004, Demattei et al., 2007] and spatial autocorrelation [F Dormann et al., 2007, Ord and Getis, 1995, Stojanova et al., 2013]. On one hand, spatial scan statistics aim at scanning the studied area using windows of an imposed shape (circles, ellipses or squares): based on a likelihood ratio test, spatial clusters are defined by the windows that group together an abnormally high number of cases. On the other hand, spatial autocorrelation methods rely on the significance of local indicators calculated according to a weighted neighborhood matrix between observed points. Both classes of methods are based on a pre-existing spatial structure: the geometric shape of the scanning window for spatial scan statistics and the spatial weights set in the neighborhood matrix for autocorrelation indicators. The use of arbitrary spatial structures is a clear limitation regarding cluster detection since cluster structures are not restricted to regular shapes nor to known weighted neighborhood. Furthermore, the use of pre-defined spatial structures is likely to mask the underlying cause of clusters, such as the relationship between clusters and their ecological environment for instance. Therefore, existing methods are likely to fail at distinguishing true clusters from covariates dependence.

4.2.3 Our approach

In [JP5], we introduced a novel statistical procedure to detect spatial cluster based on a transformation of the 2-dimensional observed point process into a collection of 1-dimensional ordered trajectories. The two main advantages of our approach are the following. First, our transformation of the data does not depend on an imposed spatial structure. Next, transformed data are compared to a reference Poisson point process that can be either homogeneous or inhomogeneous. The use of an inhomogeneous Poisson process as a reference allows accounting for the effect of covariates.

We introduced a new measure of closeness between points that first relies on (1) a data transformation from \mathbb{R}^2 to \mathbb{R} . Cluster detection is then performed by using a (2) parallel between our measure of closeness and hierarchical clustering. We further introduced the (3) extension of our methodology to the non stationary case.

Data transformation to an ordered trajectory in \mathbb{R}

Data transformation is made by finding iteratively the next event, or point, in the trajectory as follows: (1) the first point in the trajectory $x_{(1)}$ is chosen arbitrarily in the set of the observed points; (2) $x_{(2)}$ is the nearest point, according to some distance d , from $x_{(1)}$; (3) assuming that the $i - 1$ first points of the trajectory have already been ordered, the i^{th} point $x_{(i)}$ is the nearest point from $x_{(i-1)}$ among the $n - i + 1$ points not yet selected; (4) the iterative process ends when all points are included in the trajectory. Such iterative process allows for ordering the points in the trajectory from $x_{(1)}$ to $x_{(n)}$ as displayed in the example shown in Figure 4.1.

Measure of closeness

The main idea of our measure of closeness is first to compare, with respect to a given trajectory, the observed distance between consecutive points to an expected distance according to a theoretical spatial point process. In order to avoid a strong dependency to the starting point of the chosen trajectory, we proposed in a second step to account simultaneously for several trajectories in the calculation of our measure of closeness.

Under the assumption that X is a homogeneous Poisson process with intensity λ , the cumulative probability distribution of the distance between the i^{th} and the $(i + 1)^{\text{th}}$ points can be easily obtained. Conditionally to the i points already included in the trajectory, the remaining $n - i$ points

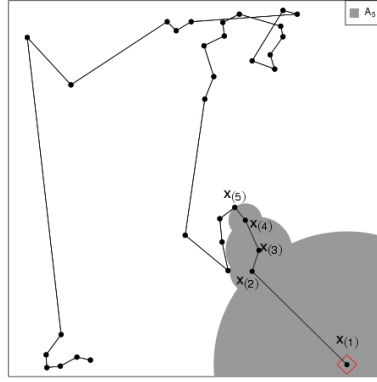


Figure 4.1: Example of a trajectory obtained with the data transformation algorithm. The grey part represents A_5 , the set of points where the nearest neighbor to $x_{(5)}$ cannot be located.

fall into the set $A \setminus A_i$ where:

$$A_1 = \emptyset \text{ and } A_i = \bigcup_{k=1}^{i-1} \mathcal{B}(x_{(k)}, d_k)$$

and $d_k = d(x_{(k)}, x_{(k+1)})$ is the observed distance between the k^{th} and the $(k+1)^{th}$ points in the trajectory. An illustrative example of A_i is given in Figure 4.1. Thus, since $x_{(i+1)}$ is the nearest neighbor to x_i among the $n - i$ remaining points, the probability that $x_{(i+1)}$ is closer than $d^* \in \mathbb{R}^+$ to x_i is given by the nearest neighbor distance distribution function for a homogeneous Poisson process. Considering D_i as the random variable characterizing the distance between x_i and its nearest neighbor in $A \setminus A_i$, we have:

$$\mathbb{P}(D_i \leq d^*) = 1 - \exp\{-\lambda|\mathcal{B}(x_{(i)}, d^*) \setminus A_i|\}.$$

where $\mathcal{B}(x_{(i)}, d^*)$ is the ball centered on $x_{(i)}$ with radius d^* . $|\mathcal{B}(x_{(i)}, d^*) \setminus A_i|$ is the area of the set of acceptable points at distance lower than d^* to x_i . Our measure of closeness between two nearest neighbors points in the trajectory, $x_{(i)}$ and $x_{(i+1)}$, is defined by p_i as follows:

$$p_i = \mathbb{P}(D_i \leq d_i | \lambda) = 1 - \exp\{-\lambda|\mathcal{B}(x_{(i)}, d_i) \setminus A_i|\}. \quad (4.4)$$

Thus, $|\mathcal{B}(x_{(i)}, d_i) \setminus A_i|$ is the area of the set of acceptable points closer to $x_{(i)}$ than $x_{(i+1)}$. Note that the p_i are an increasing function of the d_i . Thus, a low value for p_i indicates that $x_{(i)}$ and $x_{(i+1)}$ are close.

Two neighbor points on the trajectory, $x_{(i)}$ and $x_{(i+1)}$, can be considered within the same cluster if the measure of closeness p_i is less than a given threshold p^* . Unfortunately, the clusters are very sensitive to the choice of the first point. For different choices of initial points, the distances d_i and the derived p_i are different and therefore the clusters are unstable, especially for the points at the border of the clusters. In order to increase the robustness of our cluster detection, the aim of this second step is to combine the measures of closeness obtained from all possible trajectories into dissimilarities.

The maximum number of distinct possible trajectories equal to the number of points. Since two points can be neighbors for one trajectory but not for another trajectory, distance between two points cannot be directly compared among all trajectories. In order to compare the trajectories, each trajectory is converted to a matrix of dissimilarities between each pair of points. If two points are neighbors on the trajectory, the dissimilarity is equal to p_i . If the two points are not neighbors

we consider the maximum of the p_i on the path between the two points on the trajectory. For a trajectory t , we can define:

$$s(x_i, x_j|t) = \max\{p_k, \text{ for all } p_k \text{ from } x_{(b)} = x_i \text{ to } x_{(e)} = x_j \text{ on trajectory } t\} \quad (4.5)$$

A natural way to combine trajectories is to define the distance between two points as the minimum of the dissimilarities over all the possible trajectories t_l :

$$s(x_i, x_j) = \min_{t_l; l=1, \dots, n} \dots s(x_i, x_j|t_l) \quad (4.6)$$

Cluster identification

To identify cluster of points, in [JP5], we addressed the issue of determining an appropriate threshold value for s to aggregate points. Let first remark that choosing an appropriate threshold for p_i is equivalent to cut a dendrogram that represents a particular hierarchical clustering and, from Equation 4.6, that the hierarchical clustering induced by our dissimilarity is the single link in hierarchical cluster analysis. However, hierarchical clustering methods do not directly address the issue of determining the number of groups within the data and we used in [JP5] the Gap statistic [Tibshirani et al., 2001] for estimating the “best” clustering. The idea of the gap method is to compare the expected number of clusters under a uniform distribution with the empirical number of clusters computed from the original data. Let $W(r)$ be the number of clusters for a threshold r , the gap statistic is defined as:

$$G(r) = \mathbb{E}[\log(W(r))] - \log(W(r))$$

In case of repulsive distribution, $G(r)$ is positive while for aggregative distribution $G(r)$ is negative. Let r^* be the value that maximizes $G(r)$. The number of clusters is given by $W(r^*)$.

Extension to non stationary process

In the previous section, the underlying reference point process X is assumed to be a homogeneous Poisson process. However, in a practical framework, such hypothesis is rarely realistic. Thus, it is important to be able to separate the case of a clustered point process from a patch arising from the heterogeneity of the intensity of the point process (see [Waagepetersen, 2007] for example).

One can remark that the measure of closeness, introduced in Equation 4.4 for the homogeneous case, depends only on the intensity within $\mathcal{B}(x_{(i)}, d_i) \setminus A_i$. In order to account for covariates in the clustering detection, we proposed to compare the observed nearest-neighbor distance d_i to an inhomogeneous Poisson process with intensity $\lambda(x)$, where $x \in \mathbb{R}^2$. Thus, the definition of p_i can be extended to an inhomogeneous process by integrated the intensity over $\mathcal{B}(x_{(i)}, d_i) \setminus A_i$ as follows:

$$p_i = \mathbb{P}(D_i \leq d_i | \lambda(x)) = 1 - \exp\left(- \int_{\mathcal{B}(x_{(i)}, d_i) \setminus A_i} \lambda(x)\right). \quad (4.7)$$

The inhomogeneous intensity $\lambda(x)$ can be either a known function or estimated using covariates information of the area of interest. Various cases can be considered for estimated $\lambda(x)$ such as non-parametric, semi-parametric or parametric estimation.

4.2.4 Results and applications

Simulation-based evaluation of the Gap statistic

Our simulation study aims at investigating the clustering performances of our method to different methods. In this section we focus on two main objectives.

The first objective is to evaluate the performances of the Gap statistic to determine the optimal number of clusters in our framework and the second objective is to compare the clustering performances of our framework to methods based on scan statistics. We used Kulldorff's software SatScan v9.3.1 to compute the partitions based on scan statistics [Kulldorff and Information Management Services, 2009].

We tackle this objective by comparing the Gap statistic to 9 well-known methods designed to optimally cut a dendrogram according to various criteria. The 9 methods were chosen as the 4 best methods according to [Milligan and Cooper, 1985] (**ch**, [Calinski and Harabasz, 1974], **pseudot2**, [Duda and Hart, 1973], **cindex**, [Hubert and Levin, 1976], **gamma**, [Baker and Hubert, 1975]), and 5 other popular methods (**hartigan**: [Hartigan, 1975], **kl**: [Krzanowski and Lai, 1988], **silhouette** [Rousseeuw, 1987], **sindex**, [Halkidi et al., 2000], **sdbw**: [Halkidi and Vazirgiannis, 2001]). All methods have been used via the R programming software packages NbClust [Charrad et al., 2014] and cluster [Maechler et al., 2014]. The comparison of various clustering methods was based on the similarity between the true partition and partitions estimated by clustering methods. We measured the similarity between the true and observed partitions with the adjusted Rand index [Hubert and Arabie, 1985]. Note that the adjusted Rand index ranges from 0 to 1 where a value of 1 indicates a perfect match between the two partitions.

Putting results from the homogeneous and the inhomogeneous case together, the Gap statistic appears to be the most satisfying method for choosing the optimal number of clusters. More precisely, the Gap statistic is the only method showing a high mean adjusted Rand index and a reasonable 95% confidence interval.

Paracou

In [JP5], we illustrated our method on the analysis of the spatial pattern of a tree species (*Dicorynia guianensis*) surveyed at the Paracou experimental site in French Guiana [Gourlet-Fleury et al., 2004]. Figure 4.2 gives the classification tree derived from the methodology developed in the previous section. The use of the Gap statistic leads to 6 clusters: the two isolated points of the upper left corner, the isolated point of the right lower corner, and three clusters with a size larger than 6, called A, B and C in Figure 4.2.

However, differences in topography are known to led to different aggregation intensity for many species. (see for example [Ashton et al., 2000, Dalling et al., 2007, Traissac and Pascal, 2014]). *Dicorynia guianensis* fruits and seeds have many morphological features in common with wind-dispersed species. Furthermore, the slope of the plot is of particular interest since the seed spread is likely to be higher in steep area. In [JP5], we thus considered the slope as an exogenous continuous variable in the clustering of *Dicorynia guianensis*. We aimed at comparing the observed spatial distribution of trees with an inhomogeneous Poisson process with intensity $\lambda(x)$ (see Equation 4.7). We estimated the intensity function $\lambda(x)$ by fitting a spatial trend according to the slope in the overall plot. We fitted the conditional intensity as:

$$\lambda(u, x) = \exp\{\psi' B(u) + \varphi' C(u, x)\} \quad (4.8)$$

where $\theta = (\psi, \varphi)$ are the parameters to be estimated. Both ψ and φ are vectors of any dimension, corresponding to the dimensions of the vector-valued statistics $B(u)$ and $C(u, x)$ respectively. The term $B(u)$ depends only on the spatial location u , so it represents spatial trend, *i.e.* spatial covariate effects. The term $C(u, x)$ represents stochastic interactions, *i.e.* dependence between the points of the random point process. $C(u, x)$ is absent for Poisson process.

The fitted coefficients reveal a tendency for the intensity to increase as the slope decreases. Thus, trees are expected to be closer in flat regions than in steep regions.

Results of our clustering procedure in the inhomogeneous case, displayed in Figure 4.3, lead to 8 clusters: three isolated points, one cluster of 2 points and 4 clusters made by ≥ 6 points. Compared to the homogeneous case, where no exogenous variable was accounted for, 5 clusters

are identical and only cluster A was split in 3 parts, called A_1 , A_2 and A_3 . One can remark in the right panel of Figure 4.3 that the slope between clusters A_1 , A_2 and A_3 is small. As a consequence, the cumulative probability distribution of the distance between clusters A_1 , A_2 and A_3 is higher when accounting for the slope. Thus, points in clusters A_1 , A_2 and A_3 are seen as being more distant from each other. That is the reason why clusters A_1 , A_2 and A_3 are not connected in the heterogeneous case. From an ecological point of view, we might interpret our result as the fact that the range of dispersion of trees in flat area cannot be as high as in cluster A .

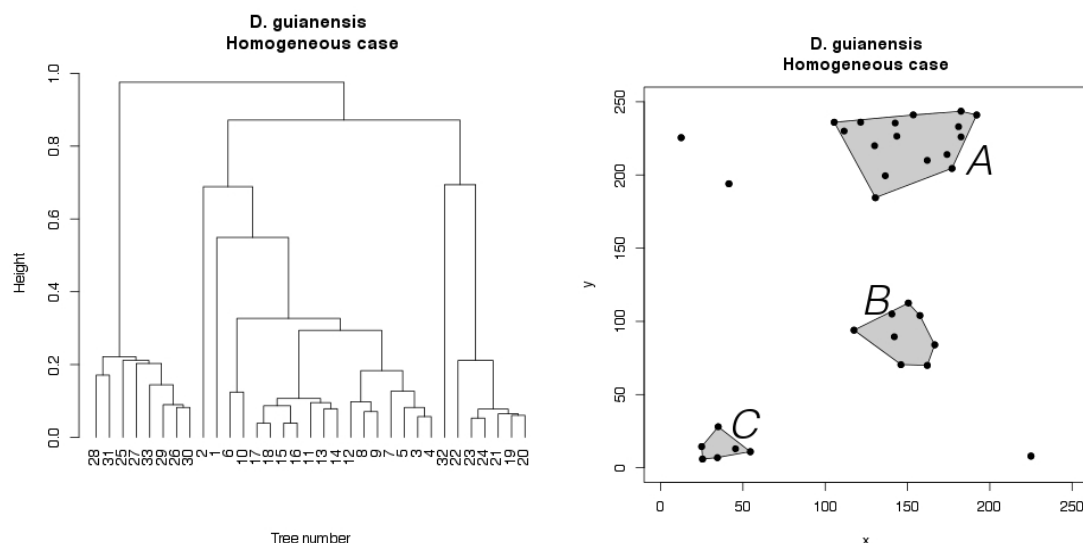


Figure 4.2: Left panel: Classification tree obtained from the spatial pattern of *Dicorynia guianensis*. Right panel: optimal clustering using gap statistics leads to 3 isolated points and 3 clusters in grey.

4.2.5 Conclusion

In this contribution, we proposed a statistical procedure to detect clusters of points in a 2-dimensional space. Our method is based on a **probabilistic characterization** of a configuration of points. Such a characterization allowed us to draw a parallel between **classification** and graph theory to propose an efficient procedure. Furthermore, the use of our formalization allows the adjustment for covariates that may prevent the detection of clusters by playing a cofounder role.

The method has been implemented in an R package that is available at <https://github.com/MathieuEmily/SpatialClustering>.

4.3 Statistical inference in 2D spatial marked point process

Our research work presented in this section proposes a stochastic model, accompanied with an inference procedure, for the self-organization of interacting points in a 2-dimensional space. It has been achieved within a collaboration with Radu Stoica (University of Lille 1).

4.3.1 Marked spatial Gibbs point process

Marked point process

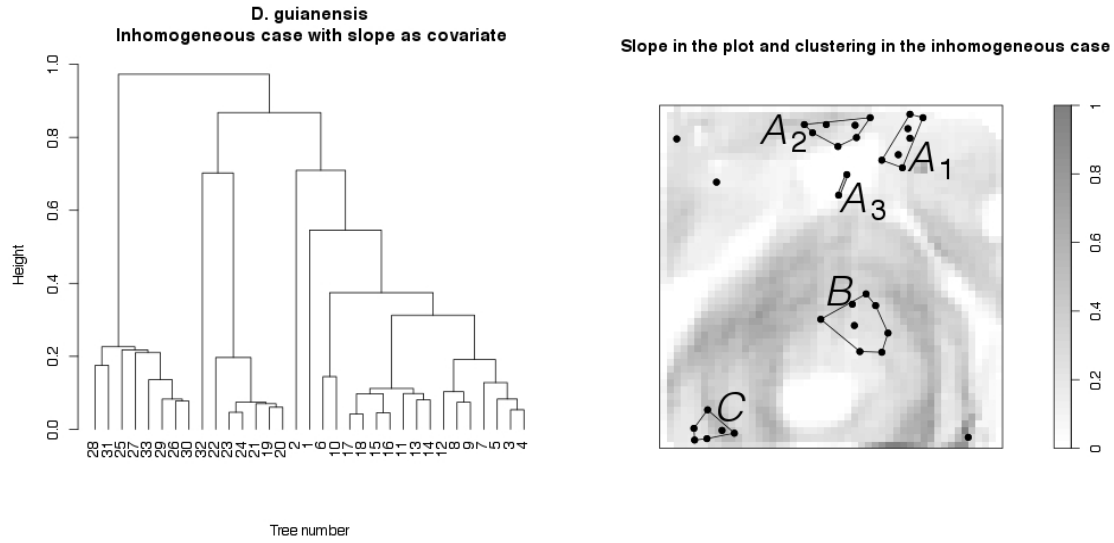


Figure 4.3: *Left panel: Classification tree obtained from the spatial pattern of *Dicorynia guianensis* with respect to the slope of the plot. Right panel: Optimal clustering using gap statistics leads to 3 isolated points and 5 clustered. The underlying grey contour represents the slope of the plot, normalized between 0 and 1. A slope closed to 0 means a flat region while a slope closed to 1 refers to the most steep area of the plot.*

In many situations, additional information exist on the points that form the observed configurations. These additional data are conventionally termed marks. The marks may be either quantitative variables, such as weight or height in the case of plants, or qualitative variables such as species, thus defining different types of points in the pattern.

Combining a spatial point process with marks yields to a marked pointed process that can be defined as follows:

Definition 4.3.1 *Let X be a point process defined on S and M the mark space. A marked point process is defined as follows:*

$$\underline{X} = \{(x, \tau_x), x \in S \text{ and } \tau_x \in \mathcal{M}\}.$$

\underline{X} is a marked point process on $S \times M$.

In most applications, the mark space M is a subset of \mathbb{R}^m with $m \geq 1$. In the special case where $M = \{1, \dots, k\}$, \underline{X} is a multi-type process with k different types of points. The Poisson point process can be easily extended to the marked Poisson point process as defined in Definition 4.3.2.

Definition 4.3.2 *Let X be a Poisson point process on S with intensity function λ . Let assume that conditional to $X = x$, the observed marks τ_x are independent. The marked point process $\underline{X} = \{(x, \tau_x) : x \in X\}$ is a marked Poisson point process.*

Let p be the common density of the marks and $\lambda_M(x, \tau_x) = \lambda(x)p(\tau_x)$. \underline{X} is a Poisson point process on $S \times M$ with intensity function λ_M .

Marked Gibbs point process

When Complete Spatial Randomness (CSR) is rejected, observed point configurations are likely to come from an underlying point process that is not the Poisson point process. It is therefore essential to define processes that can fit data where the configuration does not look random.

A natural way of proposing new type of processes is to define distributions that are based on the Poisson process. Let consider \underline{X} a marked point process with a density f with respect to the marked Poisson process. Then we have:

$$\mathbb{P}(\underline{X} \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \int_M \cdots \int_B \int_M \mathbb{1}_{(x_1, \tau_{x_1}), \dots, (x_n, \tau_{x_n}) \in F} f(\{(x_1, \tau_{x_1}), \dots, (x_n, \tau_{x_n})\}) p(\tau_{x_1}), \dots, p(\tau_{x_n}) dx_1 d\tau_{x_1}, \dots, dx_n d\tau_{x_n}$$

In that context, Gibbs marked point processes are characterized by a class of density functions f . In general, a Gibbs point process has a density with respect to the Poisson point process of the following form:

$$\forall \underline{\varphi} \in S \times M, \quad f(\underline{\varphi}, \Theta) = \frac{\exp(-H(\underline{\varphi}, \Theta))}{Z(\Theta)} \quad (4.9)$$

where H is usually called an Hamiltonian in statistical physics and allows the modeling of particular patterns of points such as aggregative or repulsive patterns. Such a modeling is parameterized by the set of parameters Θ . Finally, for a given model, the constant $Z(\Theta)$ is a normalizing constant, also called partition function, used to ensure that f is a density:

$$Z(\Theta) = \int_{\underline{\varphi}} \exp(-H(\underline{\varphi}, \Theta))$$

Context of modeling

The development and the maintenance of multi-cellular organisms are driven by permanent rearrangements of cell shapes and positions. Such rearrangements are a key step for the reconstruction of functional organs [Armstrong, 1989]. In vitro experiments such as Holtfreter's experiments on the pronephros [Holtfreter, 1944] and the famous example of an adult living organism Hydra [Gierer et al., 1972] are illustrations of spectacular spontaneous cell sorting. Steinberg used the ability of cells to self-organize in coherent structures to conduct a series of pioneering experimental studies that characterized cell adhesion as a major actor of cell sorting [Steinberg, 1962b, Steinberg, 1962a, Steinberg, 1962c]. Following his experiments, Steinberg suggested that the interaction between two cells involves an adhesion surface energy which varies according to the cell type.

With the emergence of high-throughput tissue-based tools, such as Tissue microarrays (TMAs), the immunohistological analysis of large sample of tissues is allowed [Kononen et al., 1968]. Furthermore, TMAs are used to detect and characterize cell organization within a biological tissue since segmentation algorithms have been developed to automatically detect the nuclei of the cells of the tissue. Such algorithms allow the extraction of the bi-dimensional positions of the nuclei as well as additional features for each cell (such as fluorescence intensity, probe colocalization and other measures of protein expression).

The modeling of the self-organization of cells in a tissue has inspired the development of many mathematical models [Graner and Glazier, 1992, Mochizuki et al., 1996]. These models rely on computer simulations of physical processes and act by minimizing an energy functional, referred as Hamiltonian. Tuning the internal parameters of these models is usually achieved by direct comparison of the model output and the real data that they are supposed to mimic. An important challenge is to provide automatic estimation procedures for these parameters based on statistically consistent models and algorithms. Better understanding and estimating of the nature of cell-cell

interactions in tumorigenesis may play a key role for an early detection of cancer. In addition, the invasive nature of some tumors is directly linked to the modification of the strength of cell-cell interactions. Estimating this parameter could therefore be a step toward more accurate prognosis.

Although marked Gibbs point process provides a natural probabilistic framework to deal with such a data, the practical use of such a class of models remains challenging. First, biological processes have to be realistically modeled. Next, the inference of the set of parameters in a marked Gibbs point process is not straightforward. Due to the fact that the density in Equation 4.9 is known up to an intractable normalizing constant, the likelihood is not computable.

4.3.2 Our approach

In [NC7], we proposed an approach using marked Gibbs point process to mimic the organization of cells within a biological tissue. We further provided an estimation of the set of parameters based on simulation techniques.

Description of Gibbs model

We assumed that cells are localized in a bounded set $S \subset \mathbb{R}^2$. Furthermore, we focus on binary coding of cell types, so that, for each cell x_i , a mark m_{x_i} is attached to x_i where $m_{x_i} \in M = \{0, 1\}$. Our model aims at perturbing the marked Poisson point process by considering that points can interact. In [NC7], we therefore proposed the following marked Gibbs model:

$\forall \underline{x} \in S \times M$ where $\underline{x} = \{(x_i, m_i), \dots, (x_n, m_n)\}$:

$$\begin{aligned} f(\underline{x}, \Theta) &= \frac{h(\underline{x}, \Theta)}{Z(\Theta)} = \frac{\exp(-H(\underline{x}, \Theta))}{Z(\Theta)} \\ &= \frac{\exp\left(-\left(\theta_1 \sum_{i=1}^n p(x_i, m_i) + \theta_2 \theta_3 \sum_{i=1}^n r(x_i, m_i) + \sum_{i \sim j} q(m_i, m_j)\right)\right)}{Z(\Theta)} \\ &= \frac{\exp(-\langle t(\underline{x}, \Theta) \rangle)}{Z(\Theta)} \end{aligned} \quad (4.10)$$

where $\Theta = \{\theta_1, \theta_2, \theta_3\}$ is the set of parameters to be estimated and t is the vector of sufficient statistics. It is noteworthy that the proposed Hamiltonian (or functional energy) is decomposed into three main components corresponding to three summary statistics used to describe the spatial organization of cells (see Figure 4.4).

The first summary statistic, p , is a shape constraint of each cell. The second summary statistic, r was introduced to consider non-homogeneous cell types. Finally, the third summary statistic, q , allows the modeling of the interaction between neighboring cells. We therefore used the Voronoï tessellation to define the neighborhood so that $i \sim j$ means that cells x_i and x_j are neighbors in the Voronoï sense. Figure 4.4 also shows a modeling of the cells with the Voronoï tessellation.

MCMCML Estimation

The estimation of the set of parameters Θ is challenging since $Z(\Theta)$ is intractable. In [NC7], to overcome this issue, we proposed a Markov Chain Monte Carlo Maximum Likelihood (MCMCML) estimator. MCMCML is a class of estimation techniques that is based on approximation of some quantities using empirical simulations. Such a technique therefore requires the use of an efficient simulator of the process defined in Equation 4.10.

In our context, the use of a Voronoï neighborhood topology induces a local influence of the energy. It can be noticed that such a property is deduced from the fact the model proposed in this study belongs to the class of the nearest-neighbor markov point processes. Metropolis-Hastings

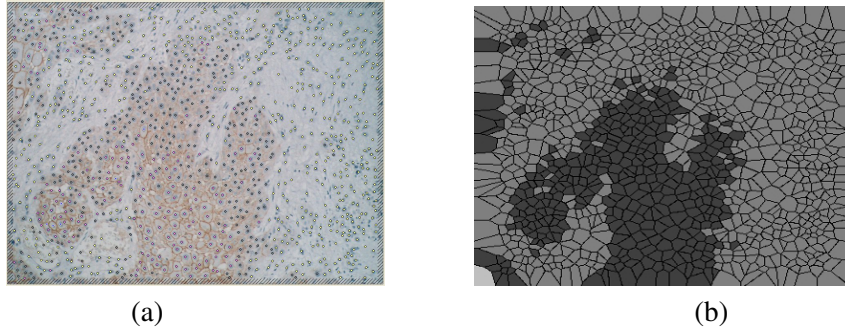


Figure 4.4: Example of an histological slide. (a) the real observed image (b) Voronoï tessellation computed from cell nuclei.

algorithm can therefore be used to correctly simulate such a process. We further used geometrical properties of the Voronoï graph to accelerate the rate of convergence as proved in [JP10].

MCMCML estimation can be derived as follows. According to Equation 4.10, the log likelihood of an observed marked configuration \underline{x} is:

$$\ell(\theta, \underline{x}) = \log(h(\underline{x}, \theta)) - \log(Z(\theta)), \quad \forall \theta \in \Theta$$

Let further consider a fixed set of parameter $\psi \in \Theta$. The log likelihood ratio of θ against ψ is usually more convenient since we have:

$$\ell(\theta, \underline{x}) - \ell(\psi, \underline{x}) = \log\left(\frac{h(\underline{x}, \theta)}{h(\underline{x}, \psi)}\right) - \log\left(\frac{Z(\theta)}{Z(\psi)}\right)$$

To maximize $\ell(\theta, \underline{x})$, it can be remarked that the first term $\log\left(\frac{h(\underline{x}, \theta)}{h(\underline{x}, \psi)}\right)$ is known in closed form. However, the second term $\log\left(\frac{Z(\theta)}{Z(\psi)}\right)$ is intractable. But, if $h(\underline{x}, \theta) = 0$ whenever $h(\underline{x}, \psi) = 0$, then it is known that:

$$\frac{Z(\theta)}{Z(\psi)} = \mathbb{E}_\psi \left[\frac{h(\underline{X}, \theta)}{h(\underline{X}, \psi)} \right]$$

The previous equation permits the calculation of the ratio of the normalizing constant by using an MCMC on the expectation. More precisely, let x_1, \dots, x_n be n simulated configurations obtained under our model in Equation 4.10 with $\Theta = \psi$, then we have:

$$\mathbb{E}_\psi \left[\frac{h(\underline{X}, \theta)}{h(\underline{X}, \psi)} \right] \approx \frac{1}{n} \sum_{i=1}^n \frac{h(x_i, \theta)}{h(x_i, \psi)}.$$

The log likelihood, $\ell(\theta, \underline{x})$ can therefore be approximated by $\ell_n(\theta, \underline{x})$ where:

$$\ell(\theta, \underline{x}) \approx \ell_n(\theta, \underline{x}) = \log\left(\frac{h(\underline{x}, \theta)}{h(\underline{x}, \psi)}\right) - \log\left(\frac{1}{n} \sum_{i=1}^n \frac{h(x_i, \theta)}{h(x_i, \psi)}\right) + \ell(\psi, \underline{x})$$

Maximizing $\ell_n(\theta, \underline{x})$ gives an approximation of the Maximum Likelihood Estimator (MLE), $\widehat{\theta}$. Such maximization was performed using a Newton-Raphson algorithm for which the gradient and Hessian expression of our model are simple:

$$\begin{aligned} \log(h(\underline{x}, \theta)) &= -\langle t(\underline{x}, \theta) \rangle \\ \nabla \log(h(\underline{x}, \theta)) &= -t(\underline{x}) \\ \nabla^2 \log(h(\underline{x}, \theta)) &= 0 \end{aligned}$$

4.3.3 Results

In [NC7], we performed a series of simulation in two situations. In a first situation, we simulated a large number of configurations with a known set of parameters θ . For each simulation, $\hat{\theta}$ was estimated with our MCMCML procedure by using another set of parameters, chosen to be equal to the targeted: $\psi = \theta$. In a second situation, we also simulated a large number of configurations according to θ , but we used several sets of value for ψ in our MCMCML simulation.

Our results demonstrated that the choice of ψ is crucial. When $\psi \approx \theta$, the whole set of parameters was correctly estimated. However, the further ψ was from θ , the worst the estimation was. In the latter case, there is no guaranty that the approximated likelihood has a maximum. Furthermore, the estimation of the gradient is very unstable and the efficiency of the estimation depends on the number of Monte Carlo simulations used to estimate $\mathbb{E}_\psi \left[\frac{h(\underline{X}, \theta)}{h(\underline{X}, \psi)} \right]$. Therefore, the computational cost of a good estimation is relatively high.

We further applied our model to histological slides from breast cancer tissues. Figure 4.4 displays an example of such data and estimation. The estimated coefficients obtained in Figure 4.4, can be interpreted as follows:

- the high value of β_1 indicates that the cell shapes are homogeneous;
- the low value of β_2 reveals an absence of multi-type effect: the size of clusters is similar from one cell type to another;
- the very high value β_2 allows the identification of a strong effect of aggregation between cell from the same type.

4.3.4 Conclusion

In this work we proposed a specific **probabilistic model** of the 2-dimensional organization of living cells in tissues. Our model is a mix between biological assumptions regarding the chemical and physical interaction between cells and point process theory. Such a formal framework allowed us to propose an appropriate statistical procedure to estimate interpretable parameters that drive the 2-dimensional configuration of cells. Based on techniques from **computational statistics**, our statistical procedure can be used in **hypothesis testing**.

4.4 Risk analysis in public health

Our research work presented in this section tackles the issue of preventing the risk of occurrence of hospital-acquired disease risk by using elements of stochastic process and non-parametric estimation. This work has been performed in collaboration with Olivier François (Institut National Polytechnique de Grenoble) and Pierre Casez (TIMC Laboratory). For this work, I received the Docteur Norbert MARX award delivered by the SFdS (“Société Française de Statistique” i.e. the French Statistical Society) .

4.4.1 Context

In a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography, many questions raise the issue of estimating or predicting the time needed to the occurrence of a particular event. The most commonly event that has been studied in the literature is death as exemplified in the pioneering studied of life table by Graunt [Graunt, 1662]. The event may also be the appearance of a tumor, the development of some disease (such as hospital-acquired diseases), recurrence of a disease, equipment breakdown, cessation of

breast feeding, and so forth. Furthermore, the event may be a good event, such as remission after some treatment, conception, cessation of smoking, and so forth.

Let T be the time until some specific event and consider that T is a nonnegative random variable. Four main functions are used to characterize the distribution of T , namely (1) the survival function, which is the probability of an individual surviving to time x , (2) the hazard rate function, sometimes termed risk function, which is the chance an individual of age x experiences the event in the next instant in time, (3) the probability density (or probability mass) function, which is the unconditional probability of the event's occurring at time x and (4) the mean residual life at time x , which is the mean time to the event of interest, given the event has not occurred at x .

In practice, the survival function, Ψ , is the basic quantity employed to describe time-to-event phenomena. It is defined as:

$$\Psi(t) = \mathbb{P}[T > t].$$

A substantial literature has been devoted to the analysis of survival curves, as for example the non-parametric estimator of Kaplan-Meier with a particular focus on the analysis and the modeling of censored data.

However, in some fields, there exists an interest in comparing survival curves to distinguish the distribution of time-to-event variables under various conditions [[Committee of Quality of Care in America, 2001](#)]. For example, in public health, several procedures or protocols can be compared in distinct hospitals to improve the safety and the quality of care. With reports describing the problem of medical errors and preventable complications, the safety and quality of health care has indeed become a major concern. Therefore, surveillance systems for hospital-acquired diseases (HAD) have been implemented. Based on Quality Indicators (QIs), developing statistical indices that can quantify HAD risk remains an important objective for improving health care systems [[Iezzoni, 2003](#)].

However, for a given procedure of care, such as hip replacement procedure, HAD risk is influenced by many factors. Among them, the length of stay (LOS) and the medical department where the procedure is performed are known to be the two most important factors [[Villemur, 1998](#)]. Accounting for these factors in the estimation of the survival function is not straightforward and requires the assumption that the individual probability of disease is a function of the length of stay in a department. For example, if we suppose a sequence of LOS (in days): {1, 4, 17, 3, 5} and further assume that disease events occur during the third and the fifth LOS. Then we have two realizations T , the time-to-event random variable, that are: $t_1 = 22 = 1 + 4 + 17$ and $t_2 = 8 = 3 + 5$. In our context, observed times-to-event are the sum of individual observed LOS.

More formally, T is the sum of random variables $T = \sum_i X_i$ where the event occurs in X_i (the i^{th} LOS) with a given probability. Therefore T can be written as a random sum of random variables as follows:

$$T = \sum_{i=1}^I X_i \quad (4.11)$$

where:

$$I_0 = 0 \text{ and } I = \min_i \{Z_i = 1\}$$

with:

$$Z_i = \begin{cases} 0 & \text{if no event occur during } X_i \\ 1 & \text{otherwise} \end{cases}$$

so that Z_i is bernoulli variable with the conditional probability of an event as parameter:

$$Z_i \sim \mathcal{B}(p(x_i)) \text{ where } p(x_i) = \mathbb{P}[\text{An event occur} | X_i = x_i]$$

Given a specific medical department, to account for the LOS in the estimation of the time-to-event distribution, it is therefore appropriate to use the definition of T as proposed in Equation 4.11.

4.4.2 Approach

In [JP9], we proposed a model to estimate the survival function, Ψ , of time-to-event variable T , as defined in Equation 4.11. In our model we assumed that the sequence of LOS X_1, X_2, \dots , are i.i.d. random variables with unknown probability density $f(x)$ and cumulative distribution function $F(x)$, for $x > 0$.

Renewal theory

Applying renewal arguments to the risk function $\Psi(t)$ leads us to the following defective renewal equation:

$$\Psi(t) = 1 - F(t) + \int_0^t \Psi(t-x)q(x)f(x)dx \quad t > 0 \quad (4.12)$$

where $q(x) = \mathbb{P}[\text{No event occur}|X=x] (= 1 - p(x))$

Cramer-Lundberg approximation

In [JP9], because Equation 4.12 could not be solved explicitly, we sought an exponential approximation for $\Psi(t)$, analogous to the so-called Cramer-Lundberg approximation of actuarial theory. This approximation assumes a positive solution of the Lundberg equation, called the adjustment coefficient. Here, the adjustment coefficient translates into the smallest positive solution R of the following Lundberg-like equation:

$$\int_0^\infty e^{Rx}q(x)f(x)dx = 1 \quad (4.13)$$

Assuming the existence of R , we can define $B(t) = \Psi(t)e^{Rt}$, $t > 0$. From the defective renewal Equation 4.12 we have:

$$B(t) = b(t) + \int_0^t B(t-x)dG(x), \quad t > 0,$$

where $b(t) = e^{Rt}(1 - F(t))$, and $dG(x) = e^{Rx}q(x)f(x)dx$. Now G is a probability distribution, and $B(t)$ is the solution of a standard renewal equation. The renewal theorem can be applied to study both the asymptotic behavior of $B(t)$, and the risk probability $\Psi(t)$, provided that $e^{Rt}(1 - F(t))$ is integrable and that $\int_0^\infty xdG(x) < \infty$. In fact, we obtain that:

$$\Psi(t) \sim C_R e^{-Rt}, \quad \text{as } t \rightarrow \infty$$

with:

$$C_R = \frac{\int_0^\infty e^{Rx}(1 - F(x))dx}{\int_0^\infty x e^{Rx}q(x)f(x)dx}.$$

Cramer-Lundberg approximation

Since the formulas obtained for R and C_R cannot be used for the inference of the disease risk directly, we introduced, in [JP9] an alternative approach based on the Bayes formula. To this aim, we looked at the distribution of the X conditional on no disease. In other words, we reject X_i , for which a disease event is observed, $Z_i = 1$, and we keep only the durations for which $Z_i = 0$. Denoting a duration resulting from this rejection procedure by Y_i , we have:

$$\mathbb{P}[Y_i \leq s] = \mathbb{P}[X_i \leq s | Z_i = 0], \quad s \geq 0,$$

and we write the common probability density function of all the Y_i 's by $f_Y(\cdot)$. Using the definition of $q(x)$ and applying the Bayes formula, we obtain that:

$$q(x) = f_Y(x) \frac{\mathbb{P}[\text{No disease}]}{f(x)}, \quad x > 0$$

Replacing $q(x)$ by the above expression in Equation 4.13 leads us to the following equation:

$$\int_0^\infty e^{Rx} f(x) f_Y(x) \frac{\mathbb{P}[\text{No disease}]}{f(x)} dx = 1$$

Finally, we obtain that:

$$\int_0^\infty e^{Rx} f_Y(x) dx = \frac{1}{\mathbb{P}[\text{No disease}]}$$

which can be rewritten as:

$$\mathbb{E}[e^{RY}] = \frac{1}{\mathbb{P}[\text{No disease}]}.$$

This equation allows us to build an estimator \widehat{R} for the adjustment coefficient R by replacing the expected values by their corresponding empirical averages.

Regarding the estimation of C_R , the same type of arguments also leads to a natural estimator. First, we can use integration by parts to obtain the following expression:

$$\int_0^\infty e^{Rx} (1 - F(x)) dx = \frac{\mathbb{E}[e^{RX}] - 1}{R},$$

where X has probability density $f(x)$. Using the Bayes formula, the denominator in CR can be simplified, and we obtain:

$$\int_0^\infty x e^{Rx} q(x) f(x) dx = \mathbb{P}[\text{Nodisease}] \times \mathbb{E}[Y e^{RY}].$$

An estimate, \widehat{C}_R , can be built by replacing the expected values by their corresponding empirical averages in the above formulas, using the value of \widehat{R} instead of the adjustment coefficient R . To finish, the disease risk can be estimated as $1 - \widehat{C}_R \exp(\widehat{R}t)$.

4.4.3 Results

To evaluate the accuracy of our proposed estimator, we first performed in [JP9] a simulation study by simulating X_i 's according to the exponential distribution, and we used 10 distinct models for the conditional individual risk, $p(x)$. Each model corresponded to the cumulative distribution function of a classical probability distribution, exponential distribution, mixture of exponential distributions, gamma distributions, Weibull distributions and uniform distribution over $[0, 1]$. The distributions were chosen as being representative of a spectrum of distribution of practical interest, and they also corresponded to models for which diseases generally occur at non constant rates over the individual stay. Using Monte-Carlo simulations, we proved that \widehat{R} and \widehat{C}_R are unbiased (except for \widehat{C}_R under the Weibull distribution), with an acceptable standard error.

Application to pulmonary embolism after hip replacement procedure

As a typical disease illustrating our approach, we considered pulmonary embolism (PE) after hip replacement procedure (HRP) in hospitals of the Rhône-Alpes area, France. PE is a common but preventable complication following HRP, and it is a cause of morbidity and mortality

[Tapson, 2008]. For that complication, LOS and confinement to bed are acknowledged to be among the principal determinants [Villemur, 1998].

For the purpose of this study, we selected patients older than 18 years who were admitted to a sample of 20 hospitals in the Rhône-Alpes area from January to November 2006 for a first HRP (3,569 patients). For each of the 20 hospitals, we estimated an (unconditional) individual risk, \bar{p} , and we formed the set of durations without declared disease events, y_i 's. Since the adjustment coefficient was estimated by solving the equation $(\sum_{i=1}^n e^{\bar{R}y_i}/n)/(1 - \bar{p})$, we can expect an approximate linear relationship between the logarithm of R and the logarithm of the average duration when the coefficient is small $R \approx 0$. In Figure 4.5, the 20 estimated adjustment coefficients were plotted against the inverse average X , $1/\bar{y}$, and we then normalized the plot using the regression of the risk estimate on the inverse X . From Figure 4.5, two hospitals displayed outlier adjustment coefficients (Hospital 3 and Hospital 13) higher than those obtained for the other hospitals.

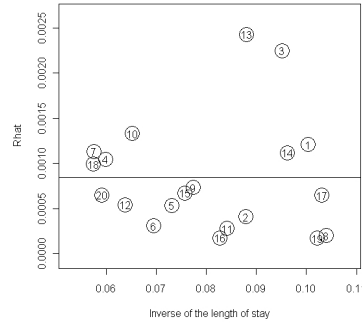


Figure 4.5: *Estimated risk coefficients for the 20 hospital data on PE after HRP recorded in the FPPS database. The coefficients were adjusted after regression on the inverse average LOS in each hospital.*

4.4.4 Concluding remarks

In this contribution, we addressed the issue of estimating the survival function of a variable that is interpretable as the result of a stochastic process. Based on elements from actuarial theory, we provided an original **probabilistic characterization** of the process that allowed us to give a parameterization of the survival function. We described **non-parametric** estimators of the set of parameters and used these estimations to compare 20 hospitals.

Conclusion and perspectives

The previous chapters presented the actual state of my research activities in biostatistics and addressed four main statistical challenges: designing powerful experiments, formalizing statistical hypothesis, modeling heterogeneous data type and accounting for the structure of the data. These challenges have been tackled through three axes of contributions: the analysis of categorical data, the modeling of structured data and the probabilistic modeling of data in biosciences. In my contributions, I designed statistical procedures motivated by issues encountered in various fields of biosciences and applied them in genomics, proteomics, cancerology, ecology and public health. Most of my contributions are accompanied by computational tools implementing our methods.

My contributions have however raised additional questions and open new research directions that we aim at addressing in the coming years. In the following paragraphs, we provide details of these future works in the light of the current research challenges in biostatistics.

Categorical variables analysis

► Improving experimental design in three-way association testing

In [JP4, PP3], we proposed a general framework to compare the power of several tests of association between two categorical variables. However, in genomics and many other fields, it is very common to investigate the relationship between more than two categorical variables. For instance, a large number of statistical methods have been developed to the detection of interaction in three-way contingency tables, such as the search for epistasis in genetic association studies [Cordell, 2009]. Since power for detecting epistasis is known to be limited in large scale genetic association studies, it is necessary to understand the roles played by experimental design parameters in power functions. We therefore aim at proposing a modeling of power functions in order to suggest new perspectives in the detection of three-way interaction testing.

► Assessing the impact of imputation in genome-wide association testing

The explosion of the amount of data has led to the development of many statistical methods dedicated to the imputation of missing values. In genome-wide association studies, imputation refers to the inference of unobserved genotypes and allows testing initially-untyped genetic variants for association with a trait of interest [Marchini and Howie, 2010]. Genotype imputation hence helps tremendously in narrowing-down the location of probably causal variants. There are several software packages available to impute genotypes (Beagle [Browning and Browning, 2009], MaCH [Li et al., 2010] and IMPUTE2 [Howie et al., 2012]). These methods are based on reference panels such as the 1000 Genomes Project. However,

imputed values are subjected to uncertainty due the inference procedure and such uncertainty has to be accounted for in association testing. We aim at addressing this issue by incorporating the uncertainty in the computation of association p-values.

Modeling of the dependence in high-dimensional data

► Using generalized linear factor models in high-dimensional data

To cover the entire genome, the number of SNPs in GWAS has to be very large inducing a tremendous number of statistical tests. Thus a big challenge in the interpretation of GWAS is the evaluation of the statistical significance level, for which multiple testing adjustments are commonly performed to either control the family-wise error-rate (FWER) or the false-discovery-rate (FDR). However SNPs data are known to be dependent, which leads to some correlation between statistical tests and seriously affects the consistency of SNP ranking [Friguet et al., 2009, Fan et al., 2012]. Accounting for the dependence in the joint distribution of categorical variables is not straightforward. To overcome that issue, we proposed in [IC9] a generalized factor model that aims at identifying a linear kernel of dependence in a family of log-linear models.

We introduced Y as the phenotype of interest considered as binary. The i^{th} SNP was denoted by X_i , also considered as binary in its bi-allelic form. The following multivariate logistic framework was assumed for the SNP profile:

$$\text{logit}(P(X_i = 1|Y = y, Z = z)) = \mu_i + \alpha_{iy} + b_i'Z, \quad (5.1)$$

where Z is a q -vector of unobservable random variables assumed to be independently normally distributed with mean 0 and standard deviation 1. The estimation of the set of parameters in a factor model is usually performed with a Expectation-Maximization (EM) algorithm.

However, in the model proposed in Equation 5.1 the expectation of the deviance is not tractable. Our goal is to propose alternative strategy, such as coordinate descent algorithm, for estimating such class of model. Furthermore, it can be remarked that the model in Equation 5.1 is very closed to “item response theory (IRT)” models. In that context, we further aim at addressing the stability of the model.

► Using multilevel modeling in multiple testing

Correlation among test statistics is known to affect the control of the proportion of false discoveries in high dimensional data. Multiple hypotheses testing is therefore a critical issue under dependence. However, in many fields, data are self-organized through a hierarchical scheme, going from microscopic to macroscopic levels of structures. In genomics, for example, genomes are known to have a block structure of correlation induced by the patterns of linkage disequilibrium. The dependence among tests in genome-wide studies can therefore be decomposed into a within and a between blocks correlations. To tackle the issue of multiple testing in GWAS, we therefore aim at proposing a model where within and between correlations are modeled independently.

First, we can assume that blocks are independent so that traditional techniques for correction for multiple testing under independence (such as Benjamini-Hochberg [Benjamini and Hochberg, 1995] or Simes [Simes, 1986]) can be applied to account for the between correlation structure. However, the application of such traditional methods requires the calculation of a combined p-value characterizing the association of a single block. Since the size of a block is reasonable, several methods can be applied to combine statistical tests within a block (minP

[Conneely and Boehnke, 2007], VEGAS [Liu et al., 2010], whitening [Kessy et al., 2015]). However, these methods have been designed for normally distributed statistics. Our aim is extend these procedure to χ^2 distributed statistic that are largely encountered in association testing.

► **Modeling the spatial architecture of the genome with Gibbs point processes**

The architecture of the genome is known to be highly complex thus reflecting the multiple mechanisms involved in maintaining the correct functionality of the genome. These mechanisms include recombination, linkage disequilibrium, physical constraints in the 3D structure. Our aim is to propose a spatial modeling of the architecture of the genome that accounts for most of the heterogeneous source of constraints. Gibbs point processes is adapted to such modeling since constraints can be formalized in the Hamiltonian of the model.

By improving the modeling of the spatial architecture, our goal is first to propose adapted decomposition of the correlation structure among association tests. Models of the spatial pattern of the genome can indeed be used as *a priori* class of covariance structure that can help addressing the multiple testing issue. Extension to spatio-temporal Gibbs point process is also considered to improve our knowledge of the evolution of the structure of the genome. Such knowledge is important to understand and control the impact of structure modifications of the genome on its functionality.

Probabilistic modeling of sequencing data

► **Using zero-inflated models in next-generation sequencing data**

High-throughput technology allows the quantification of the expression of a large number of biological features. However, the absence, or the very low level of expression, of the targeted feature leads to the observation of zeros. In many situations, measuring a zero is particularly informative and, therefore, it is important to propose probabilistic modeling of such measurements that specifically accounts for the observation of zeros.

In that context, the past few years have seen the emergence of zero-inflated models in bioinformatics. Our aim is to propose a specific modeling of data obtained in whole transcriptome sequencing technologies (RNA-Seq) by using zero-inflated negative binomial distributions. Applied to cancer diagnosis, our goal is to compare the expression signature of microRNAs (mRNAs) with long non-coding RNAs (lncRNAs). The main goal of our probabilistic modeling is to capture more subtle changes that could not be seen with current models. Based on this model, another perspective consists in proposing original statistical classification and regression procedures to (a) refine the classification of tumor samples into subtypes based on the comparative expression profiling of matched paired (tumor/control) samples, (b) identify the most statistically significant expression profiles as potential cancer biomarkers and (c) estimate regulatory network of lncRNAs by mRNAs. We aim at proposing models that are not restricted to RNA-seq data but that can also been used in other context such as single-cell gene expression analysis [Pierson and Yau, 2015].

Towards the use of biostatistics for personalized medicine

In the longer term, I would like to focus my research activities towards the new era of **personalized medicine**. A need for personalized medicine stems from several major factors, including failure of the current research and development practices, based on population-based studies, to develop effective therapies for an entire population of patients [Hamburg and Collins, 2010, Offit, 2011].

However, personalized medicine is one the biggest current challenges in life science and raises a large number of biotechnological, computational and also statistical issues. The latter include issues in experimental design, high-dimensional modeling and data integration.

First, when approaching the ultimate level of personalized medicine, each group of individuals consists of a single patient and a dense serie of measures is a prerequisite for reliable inference and predictions. Such longitudinal experiments are based on experimental designs involving the baseline and several follow-up time points, for instance, before and after a particular disease status, intervention or development of resistance to a particular drug treatment. The challenge here is to determine the key factors influencing statistical power in order to propose optimal experimental design.

Next, since the biological functions of organisms depend on complex and highly interactive systems of biomolecules, including DNA, RNA, proteins, metabolites, and lipids, data used in personalized medicine are characterized by new high-throughput multi-omics data from genomics, metagenomics, transcriptomics, proteomics, metabolomics, and lipidomics experiments. However high-dimensionality is one of the main challenges that biostatisticians face when deciphering omics data. All estimate instability, model overfitting, local convergence, and large standard errors compromise the prediction advantage provided by multiple measures. Controlling the high rates of false-positives requires researchers to adjust for multiple testing to control for type 1 error rate. Another solution to overcome multiple testing issues is to reduce dimensionality via sparse methods that provide sparse linear combinations from a subset of relevant variables (*i.e.* sparse canonical correlation analysis, sparse principal components analysis, sparse regression). However, stochastic processes to select “best” subsets of variables inferred from a given sample population may not contain the best information on another independent study, and certainly not at an individual level (*i.e.* selection-bias) [Bühlmann and van de Geer, 2011]. Reducing dimensionality remains still very challenging but is a key step in reducing the loss of information.

Lastly, efficient integration of complementary information sources from multiple levels, including tissue characteristics from cellular imaging, the genome, transcriptome, proteome, metabolome and interactome, can greatly facilitate the discovery of true causes and states of disease in specific subgroups of patients sharing a common genetic background. However, the depiction of biological systems through the integration of omics data requires appropriate mathematical and statistical methodologies to infer and describe causal links between different subcomponents [Brown et al., 2014]. The integration of omics data is both a challenge and an opportunity in biostatistics since costs of omics profiles is decreasing. Aside from the computational complexity of analyzing thousands of measurements, the extraction of correlations as true and meaningful biological interactions is not trivial. Biological systems include non-linear interactions and joint effects of multiple factors that make it difficult to distinguish signals from random errors. Data integration of heterogeneous data types is therefore one of the biggest challenge in biostatistics and there is a need to develop statistical methods to improve data utilization and scientific discovery [Gomez-Cabrero et al., 2014].

As a biostatistician, facing the above mentioned statistical challenges is a way to fulfill the promises provided by the emergence of personalized medicine.



On the use of mixed models in life science

In this appendix, I introduce a series of works that have been achieved with some of my collaborators in various fields of life science at the frontiers with statistics. All these contributions rely on the analysis of data obtained with human-based experiments characterized by a relatively low number of subjects and a substantial within and between subjects variability. In these works, my contributions fall into the modeling of observed data in order to address three main statistical issues: predictive inference of an outcome variable, comparison of the accuracy of several concurrent methods and the modeling of the residual variability.

In the remainder of this appendix, I first formalize the common statistical context of these analyses that fall into the area of linear mixed models. I then briefly summarize the contributions of such a modeling in physiopathology, virtual reality and peptidomics. Finally, I provide some statistical perspectives arisen from the practical interpretation of these works.

A.1 Introduction and motivation

Life science data are characterized by the measurement of variables in human and animal subjects. Data are therefore subjected to multiple correlated measurements and the use of multilevel models is required to the statistical analysis. Multilevel models are indeed increasingly employed across a variety of disciplines to analyze nested or hierarchically-structured data. There are many types of multilevel models, which differ in terms of the number of levels, type of design (e.g., cross-sectional, longitudinal with repeated measures, cross-classified), scale of the outcome variable (e.g., continuous, categorical), and number of outcomes (e.g., univariate, multivariate). These models have been used to address a variety of research questions involving model parameters that include fixed effects, random level-1 coefficients, and variance-covariance components.

Typically, when repeated measures are observed from the same object, data are considered as clustered in groups where a group corresponds to an individual. In that context, linear mixed models are commonly used to account for the cluster structure of the data. The (general) linear mixed model has therefore become a standard tool for modeling correlated continuous data from longitudinal and clustered sampling.

To deal with the common characteristics (low number of individuals and substantial within and between individual variability) of the data considered in this appendix, we therefore focus our modeling effort on the definition of appropriate linear mixed models. Before summarizing our contributions, I will briefly introduce the main statistical concepts of linear mixed model where a single level of grouping is considered.

Linear mixed model for single level of grouping

Let consider \mathbf{y} a set of n observations of a continuous response variable that are clustered into I groups, where group i is made by n_i observations so that $n = \sum_{i=1}^I n_i$. The n observed responses can therefore be grouped as follows:

$$\mathbf{y} = [\underbrace{y_{1,1}, \dots, y_{1,n_1}}_{\mathbf{y}_1}, \underbrace{y_{2,1}, \dots, y_{2,n_2}}_{\mathbf{y}_2}, \dots, \underbrace{y_{I,1}, \dots, y_{I,n_I}}_{\mathbf{y}_I}]'$$

so that \mathbf{y}_i is a n_i -dimensional vector of observations for the i^{th} group.

For each response k in the i^{th} group ($k = 1, \dots, n_i$), we considere that a p -dimensional vector $\mathbf{x}_{i,k} = [x_{i,k,1}, \dots, x_{i,k,p}]$ and a q -dimensional vector $\mathbf{z}_{i,k} = [z_{i,k,1}, \dots, z_{i,k,q}]$ are attached as explanatory variables. For all $i = 1, \dots, I$, we defined \mathbf{X}_i (of size $n_i \times p$) and \mathbf{Z}_i (of size $n_i \times q$) as follows:

$$\mathbf{X}_i = \begin{pmatrix} \overbrace{x_{i,1,1} \dots x_{i,1,p}}^p \\ \vdots & \ddots & \vdots \\ x_{i,n_i,1} \dots x_{i,n_i,p} \end{pmatrix} \quad \text{and} \quad \mathbf{Z}_i = \begin{pmatrix} \overbrace{z_{i,1,1} \dots z_{i,1,q}}^q \\ \vdots & \ddots & \vdots \\ z_{i,n_i,1} \dots z_{i,n_i,q} \end{pmatrix}$$

The linear mixed-effects model expresses \mathbf{y}_i ($\forall i = 1, \dots, I$) as follows:

$$\mathbf{y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i u_i + \varepsilon_i \quad (\text{A.1})$$

where \mathbf{X}_i is the design matrix for the fixed effect (or fixed-effects regressor matrix) and \mathbf{Z}_i the design matrix for the random effects (or random-effects regressor matrix), $\beta \in \mathbb{R}^p$ is a p -dimensional vector of fixed effects and $u_i \in \mathbb{R}^q$ a q -dimensional vector of random effects, for which it is assumed that:

$$u_i \sim \mathcal{N}(0, \Sigma) \quad (\text{A.2})$$

Finally, $\varepsilon_i \in \mathbb{R}^{n_i}$ is a n_i -dimensional vector, also called within-group error vector with:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n_i})$$

where \mathbf{I}_{n_i} is the $n_i \times n_i$ identity matrix.

The random group effects u_i and the within-group errors ε_i are assumed to be independent for different groups and to be independent of each other for the same group.

This decomposition of the variability with both a within and between error terms is particularly adapted to the type of data under consideration in our following contributions. First, accounting for random effects in the model allows for a better selection of the fixed effects to improve the prediction [Pinheiro and Bates, 2009]. Next, considering random effects allow a more reliable evaluation of fixed effects [Pinheiro and Bates, 2009]. Finally, the comparison of random effects models give insights in experimental design [Davis, 2002].

A.2 Application to physiopathology, virtual reality and peptidomics

This section is devoted to the summary of our contributions regarding the use of linear mixed models in three emerging domains of life science: physiopathology, virtual reality and peptidomics.

A.2.1 Designing equations for predicting the metabolic rate during outdoor walking

In [PP2], we tackled the issue of assessing energy expenditure (EE) related to walking, at both individual and group levels. Walking has indeed been considered as of substantial importance to public health and as a leading therapeutic modality and the amount EE achieved while walking is related to the risk of mortality from chronic diseases and the risk of cardiovascular events. The main goal of our study is to determine the accuracy of using Global Positioning System (GPS) in combination with other sensors for estimating EE during outdoor walking testing several conditions of speed and grade. Thus, the purpose of the present study was two-fold: 1) to compare GPS, accelerometry, and heart rate methods for estimating EE during level, uphill and downhill outdoor walking; and 2) to determine to which extent combining these methods, increases, if any, the accuracy of EE estimation compared with using a single method.

This twofold objective was addressed throughout the development of new equations obtained from the estimation of linear mixed models. We introduced a general framework where fixed and random effects are selected with variable selection procedures. Based on cross-validated root-mean-squared-error (RMSE), predictive performances of each model are compared and proved that a GPS device is a valuable method to be used for estimating EE during outdoor walking. However, although our results confirm the importance of accounting for random effects in the predictive model, the amount of variability retrieved by our model is very high compared to the residual variability. Therefore, the inclusion of other random factors in $\mathbf{Z}_i u_i$ in Equation A.1 may improved the predictive performance of the models.

A.2.2 Measuring the human perception in virtual worlds

Since almost three decades, Virtual Reality (VR) has become a huge field of exploration for researchers in computer science. The virtual could assist the surgeon, help the prototyping of industrial objects, simulate natural phenomena or entertain users through games or films. Virtual Reality technologies simulate digital environments with which users can interact and, as a result, perceive through different modalities the effects of their actions in real time. The main idea is that the user's motions need to be perceived and to have an immediate impact on the virtual world by modifying the objects in real-time. In addition, the targeted immersion of the user is not only visual: auditory or haptic feedback need to be taken into account, merging all the sensorial modalities of the user into a multimodal answer.

The user's perception of the virtual world generally represents the key indicator of the degree of interactivity that the virtual environment can generate. The more believable the interaction and its feedback, the more it makes the user unconsciously shift his reality from the real to the virtual environment, developing a true sense of presence, as defined by the "sense of being there" by Slater in 1995, an illusion of being located inside the virtual environment depicted by the VR system. Thus, interaction significantly contributes in making Virtual Reality such a powerful and immersive tool.

Concerning interaction, each of us uses all our body possibilities to interact with our real environment. If VR was limited to specific hardware using mainly hand or head motions one decade ago, there have been these last years increasing novel hardware setups tracking and measuring full-body motions and feedback. Following these improved hardware devices, 3D interaction techniques have to adapt their properties to propose novel interaction metaphors exploiting similar body inputs. The entire user body can now be used to interact with a virtual environment.

In this context, I participated to the evaluation of novel 3D interaction techniques exploiting novel body parts such as the feet [IC18, BC1], the full-body [IC16], the two hands simultaneously [IC10], or at a smaller scale the fingers [IC3]. Deformed body parts such as the arms were also

studied and compared to real-world experiments [PP4]. The evaluations are essential to assess the performance of the interaction techniques. They are composed of two main parts: quantitative measurements such as the speed or the accuracy of the participant to achieve a given task in the virtual environment, and subjective questionnaires to gather the participants' opinion.

In our contributions, we used linear mixed models as well as generalized linear mixed models to select the main factors influencing the quantitative measurements as well as participants' opinion. In [IC3, PP4], we also proposed specific formulations of the design matrix for the fixed effect (\mathbf{X}_i in Equation A.1), to evaluate learning effects in the studied tasks.

A.2.3 Modeling of the kinetics of milk digestion

In [JP1], we addressed the question of the impact of Holder milk pasteurization (a type of pasteurization that ensures sanitary quality of donor's human milk but also denatures beneficial proteins) in the kinetics of peptide release during the gastrointestinal digestion of term human milk. Digestion was measured through a measure of abundance for a collection of peptide of $n = 1054$ peptides. For each peptide, we proposed a linear mixed model to estimate the fixed coefficients that measure the interaction between two factors: Pasteurization and Digestion time. Ascending hierarchical clustering was then conducted on these interaction coefficients and allowed the identification of height clusters of peptides. The characterization of the clusters was further undertaken using the biochemical characteristics and the bioactivity prediction of each peptide. Our results demonstrated that human milk pasteurization impacted selectively the release kinetics of more than half of the peptides during term newborn *in vitro* dynamic digestion. It also increased the number and abundance of peptides present before gastrointestinal digestion which may have further nutritional consequences.

More recently, we extended our work to *in vivo* data. For that purpose, we used simulations and real data to investigate the impact of the modeling of the variability in linear mixed models. We therefore compared the following situations: no random effects, simple random effects and repeated measurements modeling of random effects with various structures (such as general, compound symmetry, Gaussian, etc.) by using corresponding correlation structure in the matrix Σ in Equation A.2. As expected according to parsimonious arguments [Davis, 2002], our results show that the correct modeling of correlation structure of repeated measurements require a relatively large number of individuals. Such result give new insights in the design of experiment and put the experimental interpretation of the results into novel perspectives.

A.3 Perspectives

A common characteristic of our contributions in physiopathology, virtual reality and peptidomics is the small size of the sample. Since experiments were based on human evaluation satisfying several constraints, recruiting participants was costly thus reducing the sample size. However, with the expansion of the use of multilevel models, questions have emerged concerning how well these models work under design conditions such as sample size at each level of the analysis [Bell et al., 2010]. This issue is central in most quantitative studies but is more complex in multilevel models because of the multiple levels of analysis. Currently there are few sample size guidelines referenced in the literature.

Another interesting perspective of our research is the evaluation of the impact of the parametric modeling of the fixed and the random effects in our proposed models. For example, in several situations, the perceptive evaluation of a virtual task hardly follows a normal distribution. In that context other class of mixed models should be considered such as non-parametric mixed models [Karcher and Wang, 2001].

Addressing these two challenges is an opportunity to push the analysis of data one step further in various fields of application such as physiopathology, virtual reality and peptidomics.



Scientific production

Papers

Preprint

- [PP1] **M. Emily**, N. Sounac, F. Kroell and M. Houée-Bigot, *Gene-based methods to detect Gene-Gene Interaction in R: the GeneGeneInteR package*, Submitted.
- [PP2] P.-Y. de Müllenheim, S. Chaudru, **M. Emily**, M. Gernigon, G. Mahé, S. Bickert, J. Prioux, B. Noury-Desvaux and A. Le Faucheur, *GPS, accelerometry, and heart rate to estimate outdoor walking energy expenditure*, Submitted.
- [PP3] **M. Emily**, *Power comparison of Cochran-Armitage Test of Trend against allelic and genotypic tests in case-control genetic association studies*, Submitted.
- [PP4] A. Girard, **M. Emily**, A. Lécuyer and M. Marchal *Virtual arms*, Submitted.

Journal papers

- [JP1] A. Deglaire, S. de Oliveira, J. Jardin, V. Briard-Bion, **M. Emily**, O. Ménard, C. Bourlieu and D. Dupont (2016) *Impact of human milk pasteurization on the kinetics of peptide release during in vitro dynamic term newborn digestion*, Accepted in Electrophoresis.
- [JP2] **M. Emily**, C. Hitte and A. Mom (2016) *SMILE: a novel Dissimilarity-based Procedure for Detecting Sparse-Specific Profiles in Sparse Contingency Tables*, Computational Statistics and Data Analysis, Vol. 99, pages 171-188.
- [JP3] **M. Emily** (2016) *AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies*, Statistical Application in Genetics and Molecular Biology, Vol. 15(2), pages 151-171.
- [JP4] **M. Emily** and C. Friguet. (2015) *Power evaluation of asymptotic tests for comparing two binomial proportions to detect direct and indirect association in large-scale studies*, Accepted in Statistical Methods in Medical Research.
- [JP5] A. Bar-Hen, **M. Emily** and N. Picard. (2015) *Spatial Cluster Detection Using Nearest Neighbour Distance*, Spatial Statistics, Vol. 14, pages 400-411.
- [JP6] **M. Emily**, A. Talvas and C. Delamarche. (2013) *MetAmyl: a METa-predictor for AMYLoiD proteins*, PLoS One; 8(11): e79722.

- [JP7] **M. Emily** (2012) *IndOR: A new statistical procedure to test for SNP-SNP epistasis in Genome-Wide Association Studies*, Statistics In Medicine, Vol. 31, No. 21, pages 2359-2373.
- [JP8] **M. Emily**, T. Mailund, J. Hein, L. Schauser and M. H. Schierup. (2009) *Using Biological Networks to Search for Interacting Loci*, European Journal of Human Genetics, Vol. 17, pages 1231-1240.
- [JP9] **M. Emily**, P. Casez and O. François. (2009) *Risk assessment for hospital-acquired infections and occupational diseases: a risk-theory approach*, Risk Analysis, Vol. 29, No. 4, pages 565-575.
- [JP10] **M. Emily** and O. François. (2007) *A statistical approach to estimating the strength of cell-cell interactions under the differential adhesion hypothesis*, Theoretical Biology and Medical Modelling, Vol. 4:37.
- [JP11] **M. Emily** and O. François. (2006) *Conditional coalescent trees with two mutation rates and their application to genomic instability*, Genetics, Vol. 172, pages 1809-1820.
- [JP12] **M. Emily**, D. Morel, R. Marcelpoil and O. François. (2005) *Spatial correlation of gene expression measures in Tissue Microarray core analysis*, Journal of theoretical Medicine, Vol. 6, No. 1, pages 33-39.

Book chapters

- [BC1] M. Marchal, A. Lécuyer, G. Cirio, L. Bonnet, **M. Emily** (2012) *Pseudo-haptic Walking*. Chapter in: Walking with the Senses, eds. Y. Visell and F. Fontana. Logos Verlag.

Popular science

- [PS1] M. Emily (2007) *Détecter plus tôt le cancer?* Le mensuel de l'université.

Thesis

- [TS1] **M. Emily** (2006) *Modèles statistiques du développement de tumeurs cancéreuses*, PhD thesis, Institut National Polytechnique de Grenoble.
- [TS2] **M. Emily** (2003) *Classification automatique de coupes histologiques : application au cancer du sein*, Master's thesis, Université Joseph Fourier..

Conferences

Contributions (international)

- [IC1] A. Le Faucheur, , P.-Y. de Müllenheim, **M. Emily**, S. Chaudru, J. Prioux, G. Mahé and B. Noury-Desvaux (2016) *Comparison of GPS, accelerometry and heart rate for estimating metabolic rate during level and uphill outdoor walking*. Proceedings of 21th annual congress of the European College of Sport Science, Vienna, Austria.
- [IC2] P.-Y. de Müllenheim, S. Chaudru, **M. Emily**, M. Franconnet, R. Moreau, J. Prioux, G. Mahé and A. Le Faucheur (2016) *The relationship between walking capacity and previous stop duration in patients with peripheral artery disease*. Proceedings of 21th annual congress of the European College of Sport Science, Vienna, Austria.

- [IC3] Y. Gaffary, M. Marchal, A. Girard, M. Pellan, A. Asselin, B. Peigne, **M. Emily** and A. Lécuyer (2016) *Studying one and two-finger perception for the discrimination of tactile directional cues*, Proceedings of Eurohaptics, London, United Kingdom.
- [IC4] **M. Emily**, N. Sounac, F. Kroell and M. Houée-Bigot, (2016) *Statistical methods for gene-based gene-gene interaction detection in R*. Journées ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2016), Lyon, France.
- [IC5] **M. Emily**, N. Sounac and F. Kroell (2016) *Analyzing gene-based gene-gene interactions with R*. European Mathematical Genetics Meeting EMGM, Newcastle Upon Tyne, United Kingdom.
- [IC6] **M. Emily** (2015) *Impact of tagging on the statistical power of association tests in Genome-Wide Association Studies*. European Mathematical Genetics Meeting EMGM, Brest, France.
- [IC7] **M. Emily**. and C. Friguet (2013) *Biological marker selection for detecting gene-gene interaction in genome-wide association studies*. 6th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2013), London, United Kingdom..
- [IC8] **M. Emily** (2013) *A case-only measure of SNP-SNP interaction in GWAS*. European Mathematical Genetics Meeting EMGM, Leiden, The Netherlands.
- [IC9] **M. Emily** and D. Causeur (2013) *Generalized linear factor modeling for dependence between SNPs in GWAS*. Statistical for (Post) Genomic Data (SMPGD'13), Amsterdam, The Netherlands.
- [IC10] A. Talvas, M. Marchal, C. Nicolas, G. Cirio, **M. Emily**, A. Lécuyer (2012) *Novel Interactive Techniques for Bimanual Manipulation of 3D Objects with Two 3DoF Haptic Interfaces*. Proceedings of Eurohaptics , Tampere, Finland.
- [IC11] H. Jean-Baptiste-Adolphe, **M. Emily**, A. Vaysse, C. André, C. Hitte (2012) *Haplotype-based method for detecting regions under selection in domestic dog*. Journées ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2012), Rennes, France.
- [IC12] **M. Emily**, C. Schirmer, A. Jan, A. Talvas, C. Garnier and C. Delamarche (2012) *Prediction of amyloidogenic motifs in Human proteins*. Journées ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2012), Rennes, France.
- [IC13] A. Talvas, C. Delamarche and **M. Emily** (2011) *Meta-prediction of amyloidogenic fragments using logistic regression*. The 6th IAPR International Conference on Pattern Recognition in Bioinformatics, Delft, The Netherlands.
- [IC14] **M. Emily** (2011) *SNP-SNP interaction in Association Studies*. 2nd international BIO-SI Workshop on biostatistics, Rennes, France.
- [IC15] A. Talvas, C. Delamarche, **M. Emily** (2011) *Meta-prediction of amyloidogenic fragments using logistic regression*. Journées ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2011), Paris, France.
- [IC16] L. Terziman, M. Marchal, **M. Emily**, F. Multon, B. Arnaldi, A. Lécuyer (2010) *Shake-Your-Head: Revisiting Walking-In-Place for Desktop Virtual Reality*. Proceedings of 17th ACM Symposium on Virtual Reality Software and Technology (VRST 2010), Hong Kong.

- [IC17] A. Le Behec, A. Talvas, E. Rio, **M. Emily**, C. Garnier and C. Delamarche (2010) *A large scale comparison of predicted amyloigogenic regions using several published methods*. Conference Jacques Monod, Roscoff, France.
- [IC18] M. Marchal, A. Lécuyer, G. Cirio, L. Bonnet and **M. Emily** (2010) *Walking Up and Down in Immersive Virtual Worlds: Novel Interactive Techniques Based on Visual Feedback*. Proceedings of IEEE Symposium on 3D User Interface (3DUI'10), Waltham, USA.
- [IC19] **M. Emily**, L. Schausser, T. Mailund and M. Schierup (2008) *Using Biological Networks to Search for Interacting SNPs in Genome-wide Association Studies*, The International Congress of Genetics, Berlin, Germany.
- [IC20] **M. Emily**, L. Schausser, T. Mailund and M. Schierup (2007) *Biological networks and epistasis in genome-wide association studies*, The Genomics of Common Diseases, Hinxton - Wellcome Trust Genome Campus, United Kingdom.
- [IC21] **M. Emily** and O. François (2005) *Number of segregating sites in a sample of genes under the genetic instability hypothesis*, XIth International Symposium on Applied Stochastic Models and Data Analysis, ASMDA 2005, Brest, France.
- [IC22] **M. Emily** and O. François (2005) *Estimating the raised mutation rate from a sample of genes with mutators*, 33rd European Mathematical Genetics Meeting, EMGM 2005, Le Kremlin-Bicetre Paris, France.
- [IC23] **M. Emily** and O. François (2005) *Estimating the raised mutation rate in a sample of gene with mutators*, Journées ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2005), Lyon, France.
- [IC24] **M. Emily**, D. Morel, R. Marcelpoil, O. François (2004) *Spatial correlation of gene expression measures in Tissue Microarray core analysis*, Journées ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2004), Montréal, Canada..

Contributions (french)

- [NC1] **M. Emily** and A. Mom (2015) *Détection de profils conditionnels dans des matrices creuses pour la sélection génomique*. 47ème Journées de Statistique, Lille, France.
- [NC2] **M. Emily** and C. Friguet (2014) *Etude de puissance pour la détection d'association directe et indirecte à partir de table de contingence 2x2*, 46ème Journées de Statistique, Rennes, France.
- [NC3] C. Friguet and **M. Emily** (2013) *Sélection de marqueurs biologiques pour la détection d'interaction de gènes*, 45ème Journées de Statistique, Toulouse, France.
- [NC4] **M. Emily** and A. Bar-Hen (2012) *Spatial clustering using nearest-neighbour distance*, 44ème Journées de Statistique, Bruxelles, Belgium.
- [NC5] L. Terziman, M. Marchal, **M. Emily**, F. Multon, B. Arnaldi, A. Lécuyer (2010) *A Novel Walking-In-Place Technique for Navigating in Virtual Worlds Using Head Motions*, 5ème journées de l'AFRV (Association Française de Réalité Virtuelle, Augmentée, Mixte et d'Interaction 3d), Orsay, France.
- [NC6] **M. Emily** (2010) *Détection d'interaction de gènes à l'échelle du génome*, 42ème Journées de Statistique, Marseille, France.

- [NC7] **M. Emily** and R. S. Stoica (2009) *Estimation Monte Carlo dans les processus ponctuels marqués en biologie tissulaire*, 41ème Journées de Statistique, Bordeaux, France.
- [NC8] **M. Emily** and O. François (2006) *Estimateur de pseudo-vraisemblance pour un processus ponctuel de Markov : application à l'histologie*, 38ème Journée de Statistique, Clamart France.

Invited talks (french)

- [IT1] **M. Emily** (2009) *Théorie du risque en santé publique: Application aux infections en milieu hospitalier*, 41ème Journées de Statistique, Bordeaux, France.
- [IT2] **M. Emily** (2008) *Protein interaction networks and genome-wide association studies*, Journées MAS de la SMAI - Modélisation et Statistiques des Réseaux Rennes, France.

Bibliography

- [Adams, 2008] Adams, J. (2008). The proteome: Discovering the structure and function of proteins. *Nature Education*, 1(3):6.
- [Aggarwal and Zhai, 2012] Aggarwal, C. and Zhai, C. (2012). A survey of text clustering algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 77–128. Springer US.
- [Agresti, 2013] Agresti, A. (2013). *Categorical Data Analysis*. Wiley, New York, third edition.
- [Agresti and Yang, 1987] Agresti, A. and Yang, M. C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis*, 5:9–21.
- [Armitage, 1955] Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.
- [Armstrong, 1989] Armstrong, P. B. (1989). Cell sorting out: the self assembly of tissues in vitro. *Critical Reviews in Biochemistry and Molecular Biology*, 24:119–149.
- [Ashton et al., 2000] Ashton, R., Baker, P. S., Bunyavejchewin, P., Gunatilleke, S., Gunatilleke, S., Hubbell, N., Foster, S. P., Itoh, R. B., LaFrankie, A., Lee, J. V., Losos, H. S., Manokaran, E., Sukumar, N., and T.Yamakura, T. (2000). Spatial patterns in the distribution of tropical tree species. *Science*, 288:1414–1418.
- [Baker and Hubert, 1975] Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Psychological Bulletin*, 70:31–38.
- [Balding, 2006] Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–791.
- [Bateson, 1909] Bateson, W. (1909). *Mendel’s Principles of Heredity*. Cambridge University Press, Cambridge, UK, first edition.
- [Bell et al., 2010] Bell, B., Morgan, G., Kromrey, J., and Ferron, J. (2010). The impact of small cluster size on multilevel models: a monte carlo examination of two-level models with binary and continuous predictors. *JSM Proceedings, Survey Research Methods Section*, pages 4057–4067.

- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.
- [Bonetta, 2010] Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. *Nature*, 468:851–854.
- [Boulesteix, 2010] Boulesteix, A.-L. (2010). Over-optimism in bioinformatics research. *Bioinformatics*, 26(3):437–439.
- [Brown et al., 2014] Brown, N., MacDonald, D., Samanta, M., Friedman, H., and Coyne, J. (2014). A critical reanalysis of the relationship between genomics and well-being. *Proceedings of National Academy of Science*, 111(35):109–114.
- [Browning and Browning, 2009] Browning, B. and Browning, S. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84:210–223.
- [Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.
- [Calinski and Harabasz, 1974] Calinski, R. B. and Harabasz, J. (1974). An examination of procedures for determining the number of clusters in a data set. *Communications in Statistics*, 3:1–27.
- [Calinski and Corsten, 1985] Calinski, T. and Corsten, L. C. A. (1985). Clustering means in anova by simultaneous testing. *Biometrics*, 41(1):39 – 48.
- [Carlson et al., 2004] Carlson, C., Eberle, M., Rieder, M., Yi, Q., Kruglyak, L., and Nickerson, D. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, 74:106–120.
- [Casella and Berger, 1990] Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press, Belmont CA.
- [Chang et al., 2013] Chang, X., Xu, B., Wang, L., Wang, Y., Wang, Y., and Yan, S. (2013). Investigating a pathogenic role for *txndc5* in tumors. *International Journal of Oncology*, 43(43):1871–1884.
- [Charrad et al., 2014] Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61.
- [Chen and Chatterjee, 2007] Chen, J. and Chatterjee, N. (2007). Exploiting hardy-weinberg equilibrium for efficient screening of single snp associations from case-control studies. *Human Heredity*, 63:196–204.
- [Chiti and Dobson, 2006] Chiti, F. and Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, 75(1):333–366.
- [Committee of Quality of Care in America, 2001] Committee of Quality of Care in America, I. o. M. (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington DC: National Academy Press.

- [Conneely and Boehnke, 2007] Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168.
- [Consortium, 2003] Consortium, I. H. (2003). The international hapmap project. *Nature*, 426:789–796.
- [Cordell, 2009] Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Review Genetics*, 10(2):392–404.
- [Cressie, 1993] Cressie, N. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics, New York, second edition.
- [Cressie and Read, 1984] Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, 46:440 – 64.
- [Cressie and Read, 1989] Cressie, N. and Read, T. (1989). Pearson’s χ^2 and the loglikelihood ratio statistic G^2 : a comparative review. *International statistical review*, 57(1):19 – 43.
- [Daley and Vere-Jones, 1988] Daley, D. and Vere-Jones, D. (1988). *An introduction to the theory of point processes*. Springer-Verlag, New York, first edition.
- [Dalling et al., 2007] Dalling, J. R., Harms, J. W., Yavitt, K. E., Stallard, J. B., Mirabello, R. F., Hubbel, M., Valencia, S. P., Navarrete, R., Vellejo, H., and Foster, R. B. (2007). Soil nutrients influence spatial distributions of tropical tree species. *Proceedings of the National Academy of Sciences of the United States of America*, 104:864–869.
- [Davis, 2002] Davis, D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer Verlag, New York, 1 edition.
- [de Bakker et al., 2005] de Bakker, P., Yelensky, R., Pe’er, I., Gabriel, S., Daly, M., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, 37:1217–1223.
- [Demattei et al., 2007] Demattei, C., Molinari, N., and Daurès, J. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data. *Comput. Stat. Data An.*, 51:3931–3945.
- [Diggle, 1983] Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London, first edition.
- [Dobson, 2004] Dobson, C. M. (2004). Experimental investigation of protein folding and misfolding. *Methods*, 34:4–14.
- [Donoho, 2000] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS CONFERENCE ON MATH CHALLENGES OF THE 21ST CENTURY*.
- [Drost et al., 1989] Drost, F., Kallenberg, W., Moore, D., and Oosterhoff, J. (1989). Power approximations to multinomial tests of fit. *Journal of the American Statistical Association*, 84(405):130 – 141.
- [Duczmal and Assuncao, 2004] Duczmal, L. and Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Comput. Stat. Data An.*, 45:269–286.

- [Duda and Hart, 1973] Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York, first edition.
- [Elliott and Wartenberg, 2004] Elliott, P. and Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112(9):998–1006.
- [F Dormann et al., 2007] F Dormann, C., M McPherson, J., B Araújo, M., Bivand, R., Bolliger, J., Carl, G., G Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.
- [Fagerland et al., 2015] Fagerland, M. W., Lydersen, S., and Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, 24(2):224–254.
- [Fan et al., 2012] Fan, J., Han, X., and Gu, W. (2012). Control of the false discovery rate under arbitrary covariance dependence. *Journal of American Statistical Association*, 107:1019–1045.
- [Fisher, 1922] Fisher, R. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85:87–94.
- [Fisher, 1925] Fisher, R. (1925). *Statistical Methods for research workers*. Edinburgh: Oliver and Boyd.
- [Fisher, 1935] Fisher, R. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd, first edition.
- [Fleming et al., 2006] Fleming, K., Kelley, L. A., Islam, S. A., MacCallum, R. M., Muller, A., Pazos, F., and Sternberg, M. J. (2006). The proteome: structure, function and evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1467):441–451.
- [Franke et al., 2010] Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. A., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Büning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., Vos, M. D., D’Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Gossu, A. V., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J.-P., Karban, A., Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panés, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhardt, A. H., Stokkers, P. C. F., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D’Amato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C., Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annesse, V., Hakonarson, H., Daly, M. J., and Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn’s disease susceptibility loci. *Nature Genetics*, 42:1118–1125.
- [Freedman, 2000] Freedman, D. A. (2000). *Notes on the odds-ratio and the δ -method*, <http://www.stat.berkeley.edu/~census/oddsrat.pdf>.

- [Friguet et al., 2009] Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of American Statistical Association*, 104:1406–1415.
- [Gabriel et al., 2002] Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.
- [Garbuzynskiy et al., 2010] Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3):326–332.
- [Gierer et al., 1972] Gierer, A., Berking, S., Bode, H., David, C., Flick, K., Hansmann, G., Schaller, H., and Trenkner, E. (1972). Regeneration of hydra from reaggregated cells. *Nat New Biol*, 239:98–101.
- [Giraud, 2014] Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 1st edition.
- [Gomez-Cabrero et al., 2014] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merckenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2):1–10.
- [Gourlet-Fleury et al., 2004] Gourlet-Fleury, S., Ferry, B., Molino, J. F., and Petronelli, P. (2004). Paracou experimental plots: key features. In S. Gourlet-Fleury, J. M. G. . O. L., editor, *Ecology and management of a neotropical rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana*, pages 17–34.
- [Gower and Ross, 1969] Gower, J. and Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(17):54–64.
- [Graner and Glazier, 1992] Graner, F. and Glazier, J. (1992). Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical Review Letters*, 69:2013–2016.
- [Graunt, 1662] Graunt, J. (1662). *Natural and Political Observations Made upon the Bills of Mortality*. London : Printed by John Martyn, first edition.
- [Greenacre, 1988] Greenacre, M. (1988). Clustering the rows and columns of a contingency table. *Journal of Classification*, 5(1):39–51.
- [Halkidi and Vazirgiannis, 2001] Halkidi, M. and Vazirgiannis, M. (2001). Clustering validity assessment: finding the optimal partitioning of a data set. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 187–194.
- [Halkidi et al., 2000] Halkidi, M., Vazirgiannis, M., and Batistakis, Y. (2000). Quality scheme assessment in the clustering process. In Zighed, D., Komorowski, J., and Żytkow, J., editors, *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *Lecture Notes in Computer Science*, pages 265–276. Springer Berlin Heidelberg.
- [Halldorsson et al., 2004] Halldorsson, B., Bafna, V., Lippert, R., Schwartz, R., De La Vega, F., Clark, A. G., and Istrail, S. (2004). Optimal haplotype block-free selection of tagging snps for genome-wide association studies. *Genome Research*, 14:1633–1640.

- [Hallgrimsdottir and Yuster, 2008] Hallgrimsdottir, I. B. and Yuster, D. S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, 9(17).
- [Hamburg and Collins, 2010] Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304.
- [Hartigan, 1975] Hartigan, J. A. (1975). *Clustering algorithms*. Wiley, New York, first edition.
- [Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.
- [Hindorff et al., 2009] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and A., T. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceeding of the National Academy of Sciences*, 106(23):9362–9367.
- [Hirji et al., 1991] Hirji, K., Tau, J., and Elashoff, R. (1991). A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine*, 10:1137–1153.
- [Hirotzu, 2009] Hirotzu, C. (2009). Clustering rows and/or columns of a two-way contingency table and a related distribution theory. *Computational Statistics & Data Analysis*, 53(12):4508 – 4515.
- [Holfreter, 1944] Holfreter, J. (1944). Experimental studies on the development of the pronephros. *Rev Can Biol*, 3:220–250.
- [Howie et al., 2012] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44:955–959.
- [Huang et al., 2011] Huang, H., Chanda, P., Alonso, A., Bader, J. S., and Arking, D. E. (2011). Gene-based tests of association. *PLoS Genetics*, 7(7):e1002177.
- [Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of the Classification*, 2:193–218.
- [Hubert and Levin, 1976] Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1080.
- [Iezzoni, 2003] Iezzoni, L. (2003). *Risk Adjustment for Measuring Health Care Outcomes*. Chicago: Academy Health and Health Administration Press, 3rd edition.
- [Jardine and Sibson, 1971] Jardine, N. and Sibson, R. (1971). *Mathematical taxonomy*. J. Wiley and Sons, London.
- [Jelizarow et al., 2010] Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and Boulesteix, A.-L. (2010). Over-optimism in bioinformatics: an illustration. *Bioinformatics*, 26(16):1990–1998.
- [Jiménez et al., 1999] Jiménez, J. L., Guijarro, J., Orlova, E., Zuro, J., Dobson, C. M., Sunde, M., and Saibil, H. R. (1999). Cryo-electron microscopy structure of an sh3 amyloid fibril and model of the molecular packing. *The EMBO Journal*, 18:815–821.
- [Jorgenson and Witte, 2006] Jorgenson, E. and Witte, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nature Review Genetics*, 7(11):885–891.

- [Jung et al., 2009] Jung, J., J.J., S., and D., K. (2009). Allelic based gene-gene interactions in rheumatoid arthritis. *BMC Proc*, S7:S76.
- [Karcher and Wang, 2001] Karcher, P. and Wang, Y. (2001). Generalized nonparametric mixed effects models. *Journal of Computational and Graphical Statistics*, 10:641–655.
- [Kessy et al., 2015] Kessy, A., Lewin, A., and Strimmer, K. (2015). Optimal whitening and decorrelation. *arXiv:1512.00809*.
- [Kettenring et al., 2003] Kettenring, J., Lindsay, B., and Siegmund, D. (2003). *Statistics: challenges and opportunities for the twenty-first century*.
- [Kononen et al., 1968] Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M., Sauter, G., and Kallioniemi, O. (1968). Tissue microarray for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4:844–847.
- [Kruskal, 1956] Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In *Proceedings of the American Mathematical Society*, 7, pages 48–50.
- [Krzanowski and Lai, 1988] Krzanowski, W. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44:23–34.
- [Kulldorff, 1997] Kulldorff, M. (1997). A spatial scan statistic. *Commun. Stat. Theory*, 26:1481–1496.
- [Kulldorff and Information Management Services, 2009] Kulldorff, M. and Information Management Services, I. (2009). SatScantm v9.1: Software for the spatial and space-time scan statistics. <http://www.satscan.org/>.
- [Kulldorff and Nagarwalla, 1995] Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14:799–810.
- [Landauer et al., 1998] Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.
- [Larson et al., 2014] Larson, N. B., Jenkins, G. D., Larson, M. C., Vierkant, R. A., Sellers, T. A., Phelan, C. M., Schildkraut, J. M., Sutphen, R., Pharoah, P. P. D., Gayther, S. A., Wentzensen, N., Goode, E. L., and Fridley, B. L. (2014). Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *European Journal of Human Genetics*, 22(1):126–131.
- [Larson and Schaid, 2013] Larson, N. B. and Schaid, D. J. (2013). A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology*, 37(7):695–703.
- [Lee and Seung, 2001] Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press.
- [Lequarré et al., 2011] Lequarré, A.-S., Andersson, L., André, C., Fredholm, M., Hitte, C., and Leeb, T. *et. al.* (2011). Lupa: A european initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *The Veterinary Journal*, 189(2):155 – 159. Special Issue: Canine Genetics.

- [Li and Li, 2008] Li, C. and Li, M. (2008). Gwasimulator: a rapid whole-genome simulation program. *Bioinformatics*, 24:140–142.
- [Li and Chen, 2008] Li, J. and Chen, Y. (2008). Generating samples for association studies based on hapmap data. *BMC Bioinformatics*, 9(1):44.
- [Li et al., 2015] Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., Zhang, R., Jiang, Y., Lv, H., and Wang, L. (2015). A gene-based information gain method for detecting gene-gene interactions in case-control studies. *European Journal of Human Genetics*, Online:Online.
- [Li and Ji, 2005] Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–227.
- [Li et al., 2009] Li, J., Tang, R., Biernacka, J., and de Andrade, M. (2009). Identification of gene-gene interaction using principal components. *BMC Proceedings*, 3(Suppl 7):S78.
- [Li and Reich, 2000] Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50(6):334–349.
- [Li et al., 2010] Li, Y., Willer, C., Ding, J., Scheet, P., and Abecasis, G. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34:816–834.
- [Lin and Yang, 2009] Lin, C. and Yang, M. (2009). Improved p-value tests for comparing two independent binomial proportions. *Communications in statistics - Simulation and Computation*, 38:78–91.
- [Lind, 1753] Lind, J. (1753). *A treatise on the scurvy : in three parts*. London : Printed for A. Millar in the Strand, second edition.
- [Little, 2005] Little, P. F. (2005). Structure and function of the human genome. *Genome Research*, 15(12):1759–1766.
- [Liu et al., 2010] Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139 – 145.
- [Liu et al., 2011] Liu, Y., Xu, H., Chen, S., Chen, X., Zhang, Z., Zhu, Z., Qin, X., Hu, L., Zhu, J., Zhao, G.-P., and Kong, X. (2011). Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genetics*, 7(3):e1001338.
- [Lui, 2004] Lui, K. J. (2004). *Statistical Estimation of Epidemiological Risk*. Wiley, New York.
- [Lydersen et al., 2009] Lydersen, S., Fagerland, M., and Laake, P. (2009). Recommended tests for association in 2x2 tables. *Statistics in Medicine*, 28(7):1159–1175.
- [Maechler et al., 2014] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2014). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.15.3 — For new features, see the 'Changelog' file (in the package source).
- [Maher, 2008] Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456:18–21.

- [Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.
- [Marchini et al., 2005] Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417.
- [Marchini and Howie, 2010] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11:499–511.
- [Maurer-Stroh et al., 2010] Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., Lopez de la Paz, M., Martins, I. C., Reumers, J., Morris, K. L., Copland, A., Serpell, L., Serrano, L., Schymkowitz, J. W. H., and Rousseau, F. (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3):237–242.
- [Meng et al., 2012] Meng, S.-R., Zhu, Y.-Z., Guo, T., Liu, X.-L., Chen, J., and Liang, Y. (2012). Fibril-forming motifs are essential and sufficient for the fibrillization of human tau. *PLoS ONE*, 7(6):e38903.
- [Milligan and Cooper, 1985] Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.
- [Mochizuki et al., 1996] Mochizuki, A., Iwasa, Y., and Y., T. (1996). A stochastic model for cell sorting and measuring cell-cell-adhesion. *Journal of Theoretical Biology*, 179:129–146.
- [Moore, 2003] Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56:73–82.
- [Moskvina and O'Donovan, 2007] Moskvina, V. and O'Donovan, M. C. (2007). Detailed analysis of the relative power of direct and indirect association studies and the implications for their interpretation. *Human Heredity*, 64:63–73.
- [Murray et al., 2014] Murray, A., Grubestic, T., and Wei., R. (2014). Spatially significant cluster detection. *Spatial statistics*, 10:103–116.
- [Musameh et al., 2015] Musameh, M. D., Wang, W. Y. S., Nelson, C. P., LluÀns-Ganella, C., Debiec, R., Subirana, I., Elosua, R., Balmforth, A. J., Ball, S. G., Hall, A. S., Kathiresan, S., Thompson, J. R., Lucas, G., Samani, N. J., and Tomaszewski, M. (2015). Analysis of gene-gene interactions among common variants in candidate cardiovascular genes in coronary artery disease. *PLoS ONE*, 10(2):e0117684.
- [Neale and Sham, 2004] Neale, B. M. and Sham, P. C. (2004). The future of association studies: Gene-based analysis and replication. *The American Journal of Human Genetics*, 75(3):353–362.
- [Neuman and Rice, 1992] Neuman, R. J. and Rice, J. P. (1992). Two-locus models of diseases. *Genetic Epidemiology*, 9:347–365.
- [Offit, 2011] Offit, K. (2011). Personalized medicine: new genomics, old lessons. *Human Genetics*, 130(1):3–14.

- [Oksanen et al., 2015] Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2015). *vegan: Community Ecology Package*. R package version 2.2-1.
- [Ord and Getis, 1995] Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4):286–306.
- [Patil and Taillie, 2004] Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ. Ecol. Stat.*, 11:183–197.
- [Pease et al., 1994] Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., and Fodor, S. P. (1994). Light-generated oligonucleotide arrays for rapid dna sequence analysis. *PNAS*, 91(11):5022–5026.
- [Pekkanen and Pearce, 2001] Pekkanen, J. and Pearce, N. (2001). Environmental epidemiology: challenges and opportunities. *Environmental Health Perspectives*, 109(1):1–5.
- [Peng et al., 2010] Peng, Q., Zhao, J., and Xue, F. (2010). A gene-based method for detecting gene-gene co-association in a case-control association study. *European Journal of Human Genetics*, 18(5):582–587.
- [Phillips, 2008] Phillips, P. (2008). Epistasis, the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Review Genetics*, 9:855–867.
- [Pierson and Yau, 2015] Pierson, E. and Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16.
- [Pinheiro and Bates, 2009] Pinheiro, J. and Bates, D. M. (2009). *Mixed-effects models in S and S-PLUS*. Springer Verlag, New York, 1 edition.
- [Pritchard and Przeworski, 2001] Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *The American Journal of Human Genetics*, 69:1 – 14.
- [Quinn and Keough, 2002] Quinn, G. and Keough, M. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, first edition.
- [Rajapakse et al., 2012] Rajapakse, I., Perlman, M. D., Martin, P. J., Hansen, J. A., and Kooperberg, C. (2012). Multivariate detection of gene-gene interactions. *Genetic Epidemiology*, 36(6):622–630.
- [Ripley, 1988] Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge University Press, Cambridge, first edition.
- [Ross and Poirier, 2004] Ross, C. A. and Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine*, 10:S10–S17.
- [Rousseeuw, 1987] Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- [Ruxton and Neuhauser, 2010] Ruxton, G. and Neuhauser, M. (2010). Good practice in testing for an association in contingency tables. *Behavioral Ecology and Sociobiology*, 64(9):1505–1513.
- [Sammut and Webb, 2011] Sammut, C. and Webb, G. I. (2011). *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st edition.

- [Simes, 1986] Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754.
- [Singhal, 2001] Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.
- [Snow, 1855] Snow, J. (1855). *On the mode of communication of Cholera*. London : John Churchill, first edition.
- [Sokal and Rohlf, 1995] Sokal, R. and Rohlf, F. (1995). *Biometry*. Freeman, New York, first edition.
- [Speed, 1985] Speed, T. (1985). Teaching statistics at the university level: how computers can help us find realistic models for real data and reasonably assess their reliability. In Rade, L. and Speed, T., editors, *Teaching of statistics in the computer age*, pages 184–195, Lund, Sweden.
- [Steinberg, 1962a] Steinberg, M. (1962a). Mechanism of tissue reconstruction by dissociated cells, ii. time-course of events. *Science*, 137:762–763.
- [Steinberg, 1962b] Steinberg, M. (1962b). On the mechanism of tissue reconstruction by dissociated cells, i. population kinetics, differential adhesiveness, and the absence of directed migration. *Proceedings of the National Academy of Science*, 48:1577–1582.
- [Steinberg, 1962c] Steinberg, M. (1962c). On the mechanism of tissue reconstruction by dissociated cells, iii. free energy relations and the reorganization of fused, heteronomic tissue fragments. *Proceedings of the National Academy of Science*, 48:1577–1582.
- [Stojanova et al., 2013] Stojanova, D., Ceci, M., Appice, A., Malerba, D., and Džeroski, S. (2013). Dealing with spatial autocorrelation when learning predictive clustering trees. *Ecological Informatics*, 13:22–39.
- [Stoyan et al., 1995] Stoyan, D., Kendall, W. S., and Mecke, J. (1995). *Stochastic geometry and its applications*. John Wiley & Sons, Chichester, second edition.
- [Stoyan and Stoyan, 1994] Stoyan, D. and Stoyan, H. (1994). *Fractals, random shapes and point fields*. John Wiley & Sons, Chichester, first edition.
- [Stram, 2004] Stram, D. (2004). Tag-snp selection for association studies. *Genetic Epidemiology*, 27:365–374.
- [Tango and Takahashi, 2005] Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *Int. J. Health Geogr.*, 4.
- [Tapson, 2008] Tapson, V. (2008). Acute pulmonary embolism. *New England Journal of Medicine*, 358:1037–1052.
- [Thomas, 2004] Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, New York, first edition.
- [Tian et al., 2009] Tian, J., Wu, N., Guo, J., and Fan, Y. (2009). Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics*, 10:1–8.
- [Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *J. Roy. Stat. Soc. B*, 63:411–423.

Résumé

La biostatistique est confrontée depuis quelques années à la modélisation et à l'analyse de données de plus en plus complexes. Cette complexité croissante est le fruit d'une évolution des données marquée par deux révolutions majeures : l'explosion des capacités de calculs informatiques et l'émergence des technologies d'acquisition haut-débit. Bien que les évolutions technologiques aient permis l'acquisition massive de données, l'analyse statistique laisse aujourd'hui de nombreuses questions ouvertes. Ces dernières années, la typologie des données biologiques a en effet profondément changé, faisant ainsi émerger de nouveaux défis statistiques. Dans ce contexte, l'objectif principal de mes activités de recherche consiste à proposer, en réponse à une question biologique d'intérêt, des procédures statistiques s'appuyant sur quatre grands défis principaux : la conception et la planification d'expériences optimisant la puissance statistique, la modélisation des types de variables mesurées, la formalisation d'hypothèses de tests pertinentes du point de vue biologique et la prise en compte de la structure des données. Mon approche peut être décrite au travers de trois axes principaux de recherche : (1) l'analyse et la modélisation de données catégorielles, (2) la modélisation statistique de données fortement structurées et (3) la modélisation probabiliste et l'inférence statistique pour le traitement de données spatialisées ou temporelles. Mes contributions ont notamment porté sur les tests d'association avec l'estimation de fonctions de puissances, la détection d'interaction, la sélection de variables et la correction pour les tests multiples. J'ai également traité des problèmes liés à la classification, à l'agrégation de tests et à l'estimation de fonction survie. Les approches développées ont été systématiquement évaluées à l'aide de données réelles provenant de nombreux domaines des sciences du vivant comme la génomique, la protéomique, la cancérologie, l'écologie et la santé. Les résultats obtenus ouvrent de nombreuses perspectives en biostatistique, notamment dans l'intégration de données hétérogènes et la médecine personnalisée.

Abstract

The recent years have seen biostatistics facing issues regarding the modeling and the analysis of more and more complex data. The evolution of biological data has been paved by two main technological revolutions: the explosion of computer capacities and the advent of high-throughput technologies. Although the fast evolution of biotechnologies has allowed the collection of massive amount of data, it has raised a large number of open questions. In the last few years, the deep modification of biological data type has contributed to the emergence of novel statistical challenges. In this context, the main goal of my research is to provide, in response to a biological question of interest, statistical procedures based on four main challenges: the design and the experimental planning to optimize statistical power, the statistical modeling of the types of measured variables, the formalization of relevant biological assumptions and the modeling of the structure of the data. My approach can be described through three main research axes: (1) the analysis and the modeling of categorical data, (2) the statistical modeling of highly structured data and (3) the probabilistic modeling and the statistical inference of spatial and temporal data. My contributions especially tackle the issue of association testing through the estimation of power functions, the detection of interaction, the variable selection and the correction for multiple testing. I have also focused my research activities on classification issues, the aggregation of statistical tests and the estimation of survival function. A systematic evaluation of the proposed methods has been performed through the analysis of real data from many fields of biosciences, such as genomics, proteomics, cancer, ecology and health. The results open novel perspectives in biostatistics, including the integration of heterogeneous data and personalized medicine.