(The hitchhiker‘s guide to) Imaging-Genetics

Andre Altmann[a], Marco Lorenzi[b,*]

[a] *CMIC, University College London, UK*
[b] *UCA, Inria Sophia Antipolis, France*

Abstract

We focus on the problem of imaging-genetics, where we generalise the classical analysis performed in genome-wide association (GWA) studies by modeling the association between medical imaging features (from ROIs to voxels/mesh-based measurements) and genetic variants (single nucleotide polymorphisms [SNPs]) in large cohorts. The main focus of this chapter will be on statistics, and on the use of appropriate statistical tools and inference methods for relating imaging phenotypes to complex and high-dimensional biological data.

Keywords: genetics, imaging, and more

1. Introduction

The human genome spans about 3 billion characters. The alphabet used to store the genetic information, which determines how the human body is assembled from a single cell and how the entire organism functions, comprises only four letters: `A, C, G` and `T`. The physical carrier of this information is the *deoxyribonucleic acid*, short: *DNA*. Each character of the genetic sequence is a nucleotide in the DNA chain; a nucleotide is composed of a sugar (deoxyribose), a phosphate group and one of the name-giving nitrogen-containing nucleobases: adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`). The nucleotides are joined together to form a chain via covalent bonds between the sugar of one nucleotide and the phosphate group of the next, thereby forming the backbone of the DNA molecule. Furthermore, if the sequences of two DNA molecules are compatible according to the base pairing rules (adenine can pair with thymine and cytosine with guanine), then the two separate DNA molecules can bind together via hydrogen bonds between the nucleobases and form the iconic double-helix. That is, the two chains have to be complementary: `AAAGGTCCAGGA` will form a bond with `TTTCCAGGTCCT`, but not with `CCCCCCAAAGGG` (see Figure 1).

The genetic information in humans is packaged into separate *chromosomes*: 22 autosomes (numbered simply based on their physical size from 1 through 22) and the sex chromosomes named `X` and `Y`. Each chromosome is a long uninterrupted DNA double helix. Commonly,

[*] Corresponding author

Email addresses: `a.altmann@ucl.ac.uk` (Andre Altmann), `marco.lorenzi@inria.fr` (Marco Lorenzi)

Figure 1: Base-pairing in two short DNA sequences. The left pair is complementary and form hydrogen-bonds (indicated by |) along the entire string and therefore forms a proper double strand. The pair on the right has many mismatching nucleotides (indicated by a space) and therefore cannot form a double-strand DNA molecule.

```
AAAGGTCCAGGA    AAAGGTCCAGGA
||||||||||||       ||
TTTCCAGGTCCT    CCCCCCAAAGGG
```

every cell in the body carries exactly two copies of each autosome, where each parent provided one set. The sex chromosomes, typically, come either in XX pairs (female) or in XY pairs (male).

Probably the most renowned part of the genome are the genes. Formally, a gene is a sequence of genetic code that encodes for a molecule, typically a protein or *ribonucleic acid (RNA)*, that carries out one or more functions in the organism. According to the GENCODE project, there are about 19,901 protein-coding genes and an additional 15,779 non-coding genes; estimates, however, vary based on the applied methodologies [1]. Before a protein can be synthesized from a gene, the genetic code is *transcribed* from the DNA into a working copy, named *messenger RNA* or *mRNA*, which serves as the template for the protein production. The RNA molecule uses ribose instead of deoxyribose as the sugar backbone. In eukaryotes, such as the human, many genes contain *exons* and *introns*: introns are removed from the mRNA through a mechanism termed *splicing* and only the sequence in the exons is retained and used to synthesize the protein. A protein is a chain of amino acid residues and the synthesis or *translation* from the mRNA follows the very precise genetic code: every group of three nucleotides (a *triplet*) encodes one of 20 amino acids or the code to terminate the synthesis. Splicing can lead to different proteins by altering the exon configuration of the mRNA that is produced from the DNA. This is phenomenon is known as *alternative splicing* and can occur, e.g., by skipping one or more exons or retaining introns. Alternative splicing is the rule rather than the exception and can occur in a condition- or tissue-specific manner. Thus, one gene can be the source of many different but related proteins.

Although genes are the most prominent part of the genome, they only make up for a minority of genetic sequence in the human genome: only about 2% of the genetic sequence in humans are at some point translated into proteins. The non-coding DNA, i.e., the remaining 98%, which comprises introns and non-coding genes among others, was referred to as *'junk DNA'* in the past. However, non-coding DNA has been found recently to contain important sequences that regulate the expression of genes, i.e., where and in which context a gene is to be read and how much protein should be produced [2].

1.1. Heritability

Differences in the appearance of individuals, that is their *phenotype* (such as hair color and height) are the result of the individual‘s genetic information (*genotype*) and environmental influences. Simply put, people appear different partly because of their difference in the genotype.

A means to quantify the degree by which variation of a trait in a population is due to the genetic variation between subjects, as opposed to environmental influence or chance, is the heritability statistic [3]. Formally, the phenotypic variance σ_P^2 is defined as:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + 2\text{Cov}(G, E),$$

where σ_G^2 and σ_E^2 are the genetic and environmental variance, respectively, and Cov(G,E) is the covariance between, genetics and environment (which can be set to 0 in controlled experiments). Thus, broad sense heritability (H^2) is simply defined as:

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

The most prominent method to estimate heritability is through twin studies that compare the phenotype similarity in monozygotic (i.e., identical or maternal) twins to the phenotype similarity in dizygotic (i.e., non-identical or fraternal) twins. Monozygotic (MZ) twins share nearly 100% of the genetic code, while dizygotic (DZ) twins, like non-twin siblings, share on average 50% of the genetic code. It is in general assumed that MZ and DZ twins share the same environment. Thus, one can compute H^2 using Falconer‘s formula [3]:

$$H^2 = 2(r_{\text{MZ}} - r_{\text{DZ}}),$$

where r_{MZ} and r_{DZ} are the Pearson‘s correlation coefficients of the trait between pairs of MZ twins and pairs of same-sex DZ twins, respectively. There are further methods to obtain heritability estimates outside twin studies, e.g., comparing trait similarities in close relatives including parents and non-twin siblings. Other methods are based on linear mixed models and can estimate heritability on more complex family trees (pedigrees). Furthermore, recent developments in statistical genetics enable us to estimate heritability from large datasets of unrelated subjects [4, 5].

1.2. Genetic variation

There are many changes that can shape an individual‘s genome. The most drastic ones are gains or losses of entire chromosomes as in trisomy 21 (*down syndrome*), in which three copies of chromosome 21 are present in cells, or monosomy X (*turner syndrome*), where only a single X chromosome is present in cells. In some cases parts of chromosomes are deleted, duplicated, inverted or moved to other chromosomes, these changes are referred to as either *copy number variations* or *structural variations*. For instance, a well characterized deletion of a small part on chromosome 22 (22q11.2 microdeletion) typically results in the loss of 30-40 genes in that region and causes *DiGeorge syndrome* and a 20-30-fold increased risk to develop Schizophrenia [6]. Other variations are on a much smaller scale and concern either the gain or loss of a few nucleotides (*insertion* or *deletion*, respectively), or simply the identity of the nucleotides at a given position in the reference genome, e.g., `A` is replaced by `G`. Depending on their location in the genome, these single nucleotide exchanges can have a serious effect on the organism. For instance, a nucleotide exchange or a deletion in the non-coding DNA is likely

	SNP_1	SNP_2	SNP_3	$\cdots$	SNP_{P-2}	SNP_{P-1}	SNP_P
1	1	0	0	$\cdots$	2	1	0
2	0	0	1	$\cdots$	0	1	0
3	0	0	0	$\cdots$	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
N-2	1	0	1	$\cdots$	1	0	0
N-1	1	0	1	$\cdots$	0	2	1
N	2	0	1	$\cdots$	0	1	0

Figure 2: Example of a SNP data matrix for N subjects organized in rows and P SNPs organized in columns. Each cell $c_{i,j}$ contains the count of the number of minor alleles in subject i for SNP j.

to be of little consequence to the organism, but replacement of a nucleotide in a gene‘s exon, which also leads to a different triplet in the mRNA and thus to an amino acid replacement in the final protein, may have severe consequences. While some amino acid replacements can be tolerated by the organism, e.g., if their biochemical properties are similar, and are considered benign; others have a deleterious effect and alter the protein‘s function. This is the case for instance with mutations in the *APP* gene causing the familial variant of Alzheimer‘s disease (a map of different deleterious and benign mutations in *APP* can be found online: `https://www.alzforum.org/mutations/app`). Luckily, exchanges with such drastic consequences are not shared by many people, i.e., their frequency in the population is low. Conversely, many of those nucleotide exchanges are part of the natural variation in our genome and are shared by many individuals, i.e., such variants show a high frequency (1% or more) in the population and are referred to as *Single Nucleotide Polymorphisms (SNPs)*. Interestingly, in most of the cases of such nucleotide exchanges, the reference nucleotide (or *major allele*), which occurs more than 50% in the population, is exchanged for only one of the other three nucleotides (referred to as *minor allele*), that is, we only see a replacement of `A` by `G` at a given position in the genome, but never a replacement of `A` by `C` or `A` by `T` at the same position (though, occasional exceptions do exist). Often the major allele is simply denoted by a capital 'A' while the minor allele is denoted by a lowercase 'a'. Discovered SNPs have been catalogued and can be referred to either by their position in the reference genome and nucleotide substitution (e.g., `19:44908684:T-C`) or by an identifier consisting of an integer number and the prefix `rs` (e.g., `rs429358`). Of note, the *minor allele frequency (MAF)* of some SNPs is close to 50%, therefore the identity of the minor and major allele in these SNPs may change between datasets due to sampling.

In this chapter we will focus on the analysis of SNP data. The fact that SNPs are shared by many individuals renders them easy and cheap to measure: microarray technology allows the measurement of 100,000s of SNPs in one experiment with more recent iteration of the technology reaching up to 2 million SNPs in one experiment. Thus, producing (after processing the data) per subject an array of `0, 1, 2` entries quantifying how many copies of the minor allele are present in the individual, where `0, 1`, and `2` simply reflect the count of the minor (the non-reference) allele (Figure 2).

The availability of a method to cheaply measure 100,000s of genetic positions in the human genome enabled a type of analysis known as *Genome-Wide Association Study (GWAS).* Simply put, this framework univariately tests for every measured SNP whether the frequency of its non-reference allele in cases is higher (or lower) than the frequency in controls, thus indicating SNPs that increase (or decrease) the disease risk. With little change in statistical methodology, GWAS also allows to test the genetic influence of SNPs on quantitative measures, referred to as *traits*, such as height. GWAS conducted over the recent past uncovered many associations (`https://www.ebi.ac.uk/gwas/`) often identifying single SNPs that affect disease risk. Especially in the realm of brain disorders the question is often how these disease associated variants affect structure and function of the brain. A typical *imaging genetics* approach is to collect people with the disease increasing variant (*carriers*) and people without that variant (*non-carriers*) and compare their brain imaging data using standard tools (see Chapter **??**). For instance, the APOE-$\epsilon4$ variant increases the risk to develop Alzheimer‘s disease (AD), thus, one could assess the effect of APOE-$\epsilon4$ on gray matter by conducting a voxel-based morphometry (VBM) analysis and treating carriers as cases and non-carriers as controls (see Chapter **??**). Similarly, carriers of the 22q11.2 microdeletion can be compared to people without that genetic change.

The previous approach allowed to investigate effects of a single genetic variant on the brain. However, the limitation is that such suitable *candidate variants* must have been previously identified through other genetic (non-imaging) analyses. In brain disorders there is the strong assumption that changes in brain structure and function are causing the observed symptoms and altered cognitive performance. Thus, phenotypes derived from brain imaging data can act as *endophenotypes* in genetic studies of brain disorders, because they are closer to the actual biological substrate. E.g., in such studies a diagnosis of Alzheimer‘s disease would be the phenotype and hippocampal atrophy or cortical burden of the amyloid protein would be considered endophenotype. The underlying assumption is that the effect sizes of genetic variations on endophenotypes are greater than those on the disease risk or phenotype itself [7]. Therefore, one expects to increase the chance to identify genetic variants that would go unnoticed when studying just the case-control status alone. If there is just one (or a few) such endophenotypes of interest, then one can follow the standard pipeline for GWAS for quantitative traits (see Section 2) with the imaging derived phenotypes. For instance, one could analyze the genetic contribution to subcortical volumes [8]. After exploring imaging genetics approaches based on the standard GWAS framework, we will introduce machine learning based methods to link genetic variation to variation in brain imaging phenotypes at high resolution (see Section 3).

2. Genome-wide association studies

The field of genetics is rapidly expanding and recent advances in the technology enable us now to obtain the entire sequence of a human at a reasonable price (i.e., less than $1,000). In this chapter, however, we will focus on the type of genetic information that is utilized in GWAS and still constitutes the main genetic information used in imaging genetics.

2.1. Genotyping using microarray technology

Microarray technology is used to assess the identity of a genetic variant at a given position in the genome. The technology is based on the property that a single DNA strand will bind (*hybridize*) to its complementary DNA strand and form a double-stranded DNA molecule. The general principle of such SNP arrays is that they carry two copies of short DNA strands (e.g., 121 letters) to measure the identity of a SNP. Both such *DNA probes* contain the complementary sequence of the reference genome before and after the targeted SNP (e.g., 60 letters in each direction); one probe contains complement of the reference allele, while the other probe contains the complement of the alternative allele. The binding affinity of a subject‘s DNA sample will be higher to the probe containing the correct complementary SNP. Importantly, these DNA probes are immobilized on the array and their positions (e.g., x- and y-coordinates) are known. For this approach to work, the test subject‘s DNA is fragmented (i.e., chopped into small pieces) and a fluorescent dye is attached to each fragment. The amount of binding strength of a subject‘s genetic sample to both probes is then quantified through fluorescence microscopy. After analyzing the raw imaging data in number of subjects, one can assess whether a subject is *homozygous* for the reference allele (*AA*, i.e., both versions of the SNP show the reference allele), *heterozygous* (*Aa*, i.e., one copy is the reference the other the alternative allele), or *homozygous* for the alternative allele (*aa),* which is encoded for computations as `0, 1` or `2`, respectively, indicating the number of non-reference allele counts.

2.2. Processing SNP data

The first step in the process is to *call* the SNP from the fluorescence signal. In some subjects the signal may not be sufficiently strong to make a call, resulting in a missing entry for that SNP in these subjects. The end product of the calling step is essentially a data matrix of dimension $N \times P$ where N represents the number of subjects and P the number of measured SNPs (Figure 2). There are various popular data formats for efficiently storing SNP information along with programs to manipulate the data and conduct association studies, e.g., PLINK is a widely-used tool [9]. There are a number of quality control (QC) steps being carried out prior to the actual statistical analysis:

1. Removal of SNP that show a poor quality across the N subjects, i.e., SNPs that have more than 10% missing entries are typically removed.
2. Removal of subjects with too many missing calls, i.e., subjects missing more than 10% of SNP calls are removed. Many missing SNP calls are an indicator for poor sample quality.
3. Removal of subjects in whom the reported sex and the sex inferred from SNP data are mismatching.
4. Removal of SNPs with a low frequency in the study sample. Depending on the size of the study SNP with frequencies lower than 5% (or 1% in large samples) are removed before analysis.
5. Removal of SNPs deviating from the Hardy-Weinberg-Equilibrium (HWE) in control subjects. Here the allele counts for *AA, Aa*, and *aa* are tested for consistency with

population genetics theory. Violation of the HWE among healthy controls may indicate technical problems with measuring the SNP.

These are the main steps pertaining the SNP data quality. In practice, there exist further QC steps involving more advanced statistics capturing participants' relatedness, genetic ancestry and population structure [10], but they are outside of the scope of this chapter.

2.3. Imputation

Imputation is the process where missing (or unobserved) data is inferred from available data using statistical methods. When working with SNP data there are generally two types of missing data:

1. Missing SNP calls for a few subjects with low signal intensity preventing a high quality SNP call, resulting in missing entries within a column of the data matrix.
2. SNPs that were not measured due to lack of probes on the microarray platform, resulting in unobserved (missing) columns.

This latter type of missing data is a major challenge for large collaborative GWAS involving dozens of teams around the world. Teams generate the SNP data using different genotyping microarray platforms (often referred to as *SNP chips*) and chip versions, resulting in different SNPs being queried by different researchers. Moreover, often a team may have generated their data with different chip versions based on the availability at the time of data generation. Thus, simply taking the intersection of genotyped SNPs across all used platforms in a collaborative study often leaves too few SNPs for a meaningful analysis. One way to circumvent this problem is to impute the calls for SNPs that were not measured on these chips. This feat is possible because humans are a relatively young species, resulting in a great deal of correlation between genetic variants that are in close proximity on the genome; referred to as *linkage disequilibrium (LD)*. Intuitively, this has the same effect as spatial correlation between voxels in imaging data. Thus, knowing a few SNPs and having access to the LD information from a large database, missing SNPs in the vicinity of genotyped SNPs can be accurately imputed. The imputation results in posterior probabilities for each of the three possible genotype calls: *homozygous reference (AA)*, *heterozygous (Aa)* and *homozygous alternative (aa)*. See the review by Marchini and Howie for additional details [11]. In addition to the posterior probability a quality score for the imputed SNP across all subjects in the dataset is provided by most available tools. The imputation result can be converted into standard `0, 1 ,2` genotype calls based on a hard cutoff for the genotype posterior probability. For instance, if the cutoff was set to 0.9 and if none of the posterior probabilities for AA, Aa or aa reached 0.9, then the imputed SNP in that subject would be set to missing. Alternatively, the imputed genotypes can be converted into a *dosage*, i.e., sum of the minor allele counts (`0, 1, 2`) multiplied by their corresponding posterior probabilities (p_0, p_1, p_2), resulting in genotype values in $[0, 2]$:

$$\text{dosage} = 0 \times p_0 + 1 \times p_1 + 2 \times p_2.$$

The advantage of the latter approach is that no missing entries are produced within a column. Post imputation QC typically involves:

	A	a
Alzheimer‘s disease	358	256
Control	471	89

Table 1: Allele counts for the $\epsilon 4$ allele of the *APOE* gene (a) and the reference allele (A) in an Alzheimer‘s disease study compared to cognitively normal controls. There is a significant increase in APOE-$\epsilon 4$ in particpants with Alzheimer‘s disease: $P = 6.03 \times 10^{-22}$ using a χ^2-test with 1 degree of freedom.

1. Removal of SNPs with a low imputation quality, e.g., a quality score < 0.9 using recent tools.
2. Removal of SNPs with high missingness rate in case of hard-called SNPs.

Overall the imputation quality depends on the used reference set, where larger databases typically lead to better imputation results. A widely-used reference database is provided by the Haplotype Reference Consortium and comprises data on 32,488 subjects [12]. In fact, this reference dataset requires a lot of storage and it was more economical to be made available through cloud-based free for all imputation servers, e.g., the Michigan [13] (`http://imputationserver.sph.umich.edu/`) and the Sanger (`https://imputation.sanger.ac.uk/`) imputation server. A very positive side effect of this development is the increased reproducibility of research and that users are less likely to accidentally misuse the software.

2.4. Statistical Analysis

Once the SNP data have been cleaned, imputed and been cleaned again, one can proceed to the statistical analysis. The predominant approach to GWAS is a mass univariate testing strategy in which every SNP is tested individually for its association with the phenotype, regardless of the identity of the remaining SNPs. The initial method was to compare allele counts between cases and controls using a 1 degree of freedom (df) χ^2 test on a 2×2 contingency table of allele counts (Table 1).

Of note, each subject contributes two alleles (one inherited from each parent), thus the overall sample size for this test is $2N$. However, the χ^2-test approach does not allow for adjusting the association test for relevant covariates such as age, sex or genetic ancestry. Thus, the common method of choice are logistic and linear regression for case-control and quantitative studies, respectively. Here the output is modeled as Y and SNP is the minor allele count per subject, $C_1, \ldots, C_k$ are the covariates, and $\epsilon \sim N(0, 1)$ is the residual error:

$$Y = a_0 + b_0 \times \text{SNP} + b_1 C_1 + \ldots + b_k C_k + \epsilon. \tag{1}$$

For every SNP such a linear model is estimated using all N subjects, and the intercept (a_0) as well as the coefficients for the variables ($b_0, \ldots, b_k$) are estimated. Naturally, in genetic studies the main interest is the statistical significance of b_0, i.e., the influence of the SNP on the phenotype. There are different assumptions on how a SNP can act biologically: the *additive* model assumes that each copy of the minor allele has an independent effect; the *dominant* model assume that the presence of one allele has the same effect as both alleles, and the *recessive* model assumes that there is only an effect on the phenotype when both

	AA	Aa	aa
additive	0	1	2
dominant	0	1	1
recessive	0	0	1
over-/under-dominance	0	1	0

Table 2: Different encodings of the SNP variable in GWAS to realize different genetic models.

copies exhibit the minor allele. To realize these main models in the linear framework the encoding of the SNP is adjusted as shown in Table 2. There are also cases where the effect is only present in heterozygous subjects (Aa). Depending on the effect direction this is termed *overdominance* or *underdominance.* In brain science, examples for an overdominant effect are a variant in the nicotine receptor gene *CHRNA4* on cognitive control [14] as well as a variant in the longevity gene *KLOTHO* on enhanced cognition [15].

In addition to the classic linear models, there are more advanced methods that increase the computational speed on large datasets with many subjects and increase statistical power [16]. Genome-wide results are typically presented in the form of a *Manhattan plot*, where the $-\log_{10}(P)$ of b_0 for a SNP is given on the Y-axis and the SNPs‘ locations in the genome are arranged on the X-axis. Strong association between a lead SNP as well as its correlated neighboring SNPs with a phenotype lead to peaks in the plot mimicking the name-giving skyline of Manhattan. The mass univariate strategy on testing genetic variants leads to a massive multiple testing problem.

2.5. Multiple testing correction

In statistical hypothesis testing so-called *test statistics* are computed from the given observations, e.g., the t-score is the test statistic when conducting a t-test to compare the means of a measurement of two groups. Further, it is known how the test statistic is distributed under the *null hypothesis* or H_0. That is, it is known how the test statistic is distributed when there is no difference between two groups. In case of the t-test the t-distribution models the distribution of t-scores under the null hypothesis of no group differences. Thus, combining these two information, we can compute the likelihood of achieving the observed test statistic (or better) when there is no difference between two groups, i.e, when H_0 is correct. This probability is exactly what the p-value captures. A widely-accepted threshold (α-level) for calling a result *statistically significant* is $p = 0.05$. That is, there is only a 5% chance that the test statistic could be produced from random data where there is no difference between two groups.

In situations where more than a single hypothesis is tested, one has to reconsider the 5% cutoff. The problem is related to the situation of casting a die. The chance to get a 6 is $1/6 \approx 0.167$. However, when the die is cast 20 times, the chance of getting a single 6 in any of the rolls is $1.0 - (5/6)^{20} \approx 0.974$. So it is almost certain to hit at least one 6 in 20 rolls. The same holds true for a significant finding in statistical testing: if enough hypotheses are tested, then the likelihood increases that at least one of them will get a p-value of 0.05 or lower; even if the data are randomly generated. Therefore, the need to adjust the p-value for multiple testing arises. There are various methods available, but the most widely used ones are the Bonferroni method [17] for controlling family-wise error rate (FWER) and the method by Benjamini-Hochberg to control for false discovery rate (FDR) [18].

The Bonferroni methods aims to control the family-wise error rate, i.e, the probability to incorrectly reject the H_0 (calling something significant when it is not) once across all m alternative hypothesis tests $(H_1, \ldots, H_m)$ at a predefined α-level. This is simply achieved by testing each individual alternative hypothesis at α/m. That is, we divide the p-value threshold by the number of tests, or alternatively, we multiply each raw p-value with m to get a multiple testing adjusted p-value. This method assume that all the tested hypotheses are statistically independent.

The Benjamini-Hochberg method aims to control the false-discovery rate, i.e., the fraction of all hypothesis that were called 'significant' but where there was no true difference (and H_0 was correct). This is a less stringent correction than the FWER correction, where the α-level is adjusted such that the occurrence of a single false-discovery minimized. Practically, the p-values of the m tests are ordered from smallest to largest: $P_{(1)} < \ldots < P_{(m)}$. Next, the largest k such that $P_{(k)} < \frac{k}{m}\alpha$ is identified and all hypotheses corresponding to the p-values $P_{(1)}, \ldots, P_{(k)}$ are accepted. The Benjamini-Hochberg method (and its extensions) are also valid when hypotheses are dependent on (or correlated to) each other. FDR-correction is therefore often preferred when it is known that hypothesis are correlated, e.g., when testing voxels in the brain.

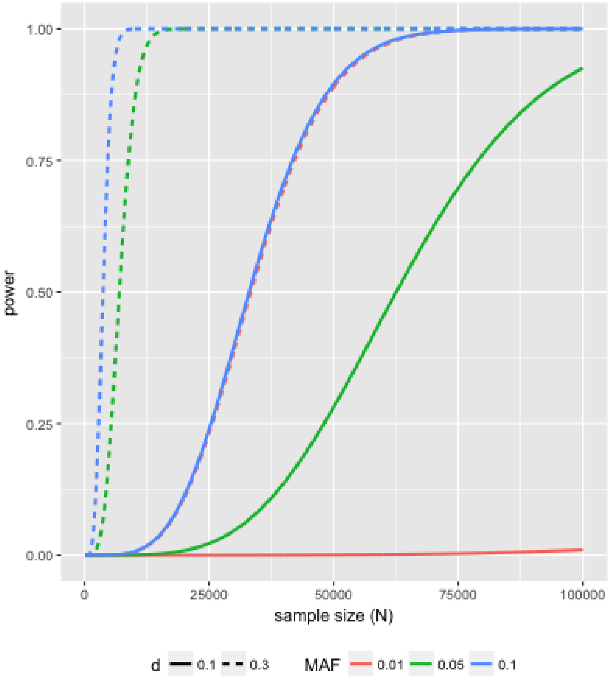
The widely-used approach of correcting p-values using the Bonferroni method for all the P tests (SNPs) is overly conservative in GWAS: SNPs are (spatially) correlated – this fact is exploited by the genotype imputation approach described above – reducing the number of effective independent tests below the number of tested SNPs. It has been estimated that in GWAS with common SNPs in the European population the effective number of independent tests, while accounting for the correlation between genetic variants (i.e., the LD structure), is about 1 million [19]. Thus, the Bonferroni adjusted p-value threshold for genome-wide significant was set to $P = 5 \times 10^{-8}$. However, with recent advances leading to the inclusion of rare SNPs and efforts to include non European subjects, the cut-off may need to be revised.

SNPs passing the genome-wide significance threshold are typically followed-up with additional research in order to reveal the underlying biology; this often involves costly wet-lab experiments. In order to limit the risk of false-positive associations, GWAS typically employ a two-step process where parts of the data are used as the *discovery* dataset (often referred to as stage `I`) and a *validation* dataset (stage `II`), where SNPs that passed the genome-wide threshold in stage `I` will be tested again at a more lenient threshold. This approach resembles the train-test split applied in statistical learning (Section 3).

2.6. Adjustments for quantitative traits

Initially GWAS were only conducted for dichotomous traits but minor changes in modeling facilitates the analysis of quantitative traits as well. However, additional precautions have to be taken when working with quantitative traits due to the sensitivity and model assumptions of the underlying linear models:

Figure 3: Power curves generated using a t-test with unequal sample sizes. The significance threshold was set to $\alpha = 5 \times 10^{-8}$, minor allele frequency (MAF) was varied from 0.01 to 0.1 (different colors), effect size (Cohen‘s d) was varied from 0.1 to 0.3 (different line types) and the sample size (N) was varied from 200 to $100,000$ subjects (X-axis). Alternatively, for quantitative traits the power can be computed based on test for the significance of Pearson‘s r, where in practice the correlation r is a function of the MAF and Cohen‘s d [21].



1. Phenotypic outliers can easily produce a false-positive association and outlier removal techniques such as windosorizing may be applied.
2. Phenotypes that are not approximately Gaussian distributed violate model assumptions, thus the raw phenotype may have to be transformed prior to analysis.

Furthermore, many datasets with brain imaging data comprise healthy subjects as well as diseased subjects. Thus, one recurring question is how to best handle the resulting bias and whether it is appropriate to adjust the model for disease status [20].

2.7. Statistical power

Before actually collecting data for a GWAS the question is often whether the sample size will be sufficiently large to discover variants given the many tests ($\sim$ 1 million). Like in many other disciplines this answer is addressed via a statistical power analysis. Statistical power quantifies the likelihood of rejecting the null hypothesis (H_0) when the alternative hypothesis (H_1) is correct (i.e., $1.0 -$ `type II` error probability). In the context of classification, statistical power would be equivalent to the a classifier‘s true positive rate. In GWAS the power depends on the SNP‘s minor allele frequency, its effect size and the size of the study sample.

As evident from Figure 3, larger samples in GWAS lead to an increased power to detect true associations at genome-wide significance ($\alpha = 5 \times 10^{-8}$) given that frequency and effect size of the variant remain unchanged. This fact led to an ever increasing size of genetic studies over the last decade with recent GWAS including more than 1 million subjects [22]. Further, power can be increased by studying a phenotype that is more directly influenced by the genetic variant, leading to the rationale of *endophenotypes*. For instance, a SNP with MAF of 0.1 (blue line) may have an effect of $d = 0.1$ (solid) on disease status achieving only a power of 0.25 with 25,000 subjects; the same SNP shows an effect of $d = 0.3$ (dashed line) on, say, hippocampal volume, thus increasing the power to 0.9 with only 5,700 subjects.

2.8. GWAS and imaging phenotypes

The concept of endophenotypes is clearly appealing from the power perspective. However, when there is more than one phenotype to study, then the statistical power will be reduced because the need to adjust for the number of (independent) tested phenotypes in addition to the number of genetic tests (1 million). For instance, in the case of ten brain imaging biomarkers this would amount to 10 million tests (= 10×1 million) leading to a p-value cutoff of 5×10^{-9}. Applying this adjusted α-level to the above power example in Figure 3 ($d = 0.3$, MAF=0.1,$N = 5,700$) would reduce the power to 0.83, which is still acceptable. However, in brain imaging it is not unlikely to study 1,000s of derived phenotypes, reducing the power even further (e.g., to 0.58 when corrected for 1 million SNPs and 1,000 imaging phenotypes). Thus, due to the increased multiple testing burden, the classic univariate GWAS approach works with large datasets and a few hundred imaging derived phenotypes, an approach followed by the ENIGMA consortium [8]. However, methodological advances are spurring ever finer-grained structural [23] and functional [24] brain parcellations, thereby increasing the number of available imaging phenotypes. Moreover, depending on the cohort there are many different imaging modalities that can be studied from gray matter volume, cortical thickness or measures based on diffusion weighted imaging, thus, easily amounting to a few 1,000 phenotypes with a rather coarse granularity [25]. In some cases imaging phenotypes are combined to generate novel biomarkers, e.g., Scelsi et al combined hippocampal volume and cortical amyloid burden into a novel multimodal imaging phenotype [26]. However, even despite finer grained parcellations, it is conceivable that genetic effects are not limited to boundaries of regions of interest, e.g., only a fraction of a candidate region may be affected by a given SNP, thereby weakening the association when the phenotype is defined over the entire region of interest. Taking the resolution spectrum to the extreme end leads to *voxel-wise genome-wide analyses (vGWAS)*. Though, the power analysis for this approach is not very encouraging: modern MRI whole brain acquisition techniques generate voxels of less than 1 mm^3, taken together with an approximate 0.8 dm^3 grey matter volume results in approximately 1 million relevant voxels to be tested in vGWAS. Fortunately, due to spatial correlation these tests will not constitute independent tests, but the estimate will still be around at least 100,000 independent tests pushing the threshold for genome-wide and voxel-wise significance to at least 5×10^{-13}. In their first vGWAS in Alzheimer‘s disease Stein et al analyzed data of 740 subjects and developed a novel method based on FDR to address the multiple comparisons problem [27]. While this first analysis was still limited in statistical power, the availability of large databases with paired imaging and genetics data such as the `UK BioBank` (`http://www.ukbiobank.ac.uk/`) or the collaborative efforts such as the `ENIGMA` consortium (`http://enigma.ini.usc.edu/`) [28] will ultimately lead to a statistical viability of vGWAS. Last but not least, the brute-force concept of vGWAS induces a massive computational burden: studies with a few hundred phenotypes can easily be parallelized on conventional hardware; running whole genome studies on millions of voxels requires advanced computational methods such as fast vGWAS [29].

GWAS-based approaches to analyze brain imaging phenotypes are by their nature mass-univariate on the genetic part, thereby they neither explore potential interactions between SNPs (referred to as *epistasis*) nor make explicit use of the spatial correlation within the imaging data. Multivariate approaches, which we will introduce in the next section, are a promising way to make efficient use of the available data.

3. GWAS. Multivariate approaches

In recent years many domains have seen an increased use of multivariate approaches including neuroscience [30] and GWAS [31]. Also recent methodological advances in the imaging genetics domain rely on multivariate approaches to capture meaningful genotype-phenotype interactions [32, 33]. The appeal of these methods lies in their ability to identify complex relationships between the genome and the brain by simultaneously modeling the joint effect of genetic variants on brain features. The promising potential of multivariate imaging-genetics approaches is to explicitly highlight the underlying biology of macroscopic processes, such as brain atrophy, by identifying sets of genetic variants that are jointly associated with the phenotype.

3.1. A sketch on classical multivariate approaches

Typical multivariate approaches for the joint analysis of brain imaging phenotype and large arrays of genetic variants are based on the identification of latent modes of maximal association between imaging and genetics features. The underlying principle of these procedures consists in looking for pairs of feature combinations - one “combination” or mode for each of the two distinct data types - that have maximal association. For example, while *Partial Least Squares* aims at maximising the covariance between these combinations (or projections on the modes’ directions), *Canonical Correlation Analysis* maximises their statistical correlation, and *Reduced-Rank Regression* minimises the error in predicting a target data-type from the optimal feature combination of a source one.

In what follows, genetic and imaging data for a given individual k are encoded by arrays $\mathbf{x}_k$ and $\mathbf{y}_k$, respectively, with dimensions $dim(\mathbf{x}_k) = D_g$ and $dim(\mathbf{y}_k) = D_i$. An imaging-genetics data matrix for N individuals is therefore represented by the pair of centered matrices $\mathbf{X}$ and $\mathbf{Y}$, with $dim(\mathbf{X}) = N \times D_g$, and $dim(\mathbf{Y}) = N \times D_i$.

The basic principle of classical multivariate analysis techniques relies on the identification of *linear transformations* of $\mathbf{X}$ and $\mathbf{Y}$ into a lower dimensional subspace where the projected data exhibits the desired statistical properties of similarity.

3.1.1. Canonical Correlation Analysis

In canonical correlation analysis (CCA), this problem is formulated by looking for linear transformations parameterized by the vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ such that $\mathbf{X}\mathbf{w}_x$ and $\mathbf{Y}\mathbf{w}_y$ are maximally correlated. In mathematical terms:

$$\mathbf{w}_x, \mathbf{w}_y = \underset{\mathbf{w}_x, \mathbf{w}_y}{\operatorname{argmax}} \rho(\mathbf{X}\mathbf{w}_x, \mathbf{Y}\mathbf{w}_y), \tag{2}$$

where $\rho(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}/(\sqrt{\mathbf{a}^T\mathbf{a}}\sqrt{\mathbf{b}^T\mathbf{b}})$. Developing (2) we obtain

$$\rho(\mathbf{X}\mathbf{w}_x, \mathbf{Y}\mathbf{w}_y) = \frac{\mathbf{w}_x^T \mathbf{S}_{\mathbf{XY}} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{S}_{\mathbf{XX}} \mathbf{w}_x}\sqrt{\mathbf{w}_y^T \mathbf{S}_{\mathbf{YY}} \mathbf{w}_y}}, \tag{3}$$

where $\mathbf{S}_{\mathbf{XY}}$ is the cross-covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$, while $\mathbf{S}_{\mathbf{XX}}$ and $\mathbf{S}_{\mathbf{YY}}$ are the sample covariances of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

Since the optimization of CCA (2) is independent from the norm of the projection vectors, without loss of generality we can restrict the problem by introducing the constraints $||\mathbf{X}\mathbf{w}_x||^2 = 1$ and $||\mathbf{Y}\mathbf{w}_y||^2 = 1$. This implies that we look for projections into a low-dimensional subspace of unitary variance. The maximisation of (2) can be tackled through the optimization of the Lagrangian

$$\mathcal{L}(\mathbf{w}_x, \mathbf{w}_y, \lambda_x, \lambda_y) = \mathbf{w}_x^T \mathbf{S}_{\mathbf{XY}} \mathbf{w}_x - \lambda_x(\mathbf{w}_x^T \mathbf{S}_{\mathbf{XX}} \mathbf{w}_x - 1) - \lambda_y(\mathbf{w}_y^T \mathbf{S}_{\mathbf{YY}} \mathbf{w}_y - 1), \tag{4}$$

and it can be easily shown that the CCA solution can be obtained through the following generalized eigenvalue problem [34]:

$$\begin{bmatrix} \mathbf{0} & \mathbf{S}_{\mathbf{XY}} \\ \mathbf{S}_{\mathbf{YX}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{S}_{\mathbf{XX}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{\mathbf{YY}} \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix}, \tag{5}$$

or equivalently,

$$\begin{bmatrix} \mathbf{0} & \mathbf{S}_{\mathbf{XX}}^{-1/2} \mathbf{S}_{\mathbf{XY}} \mathbf{S}_{\mathbf{YY}}^{-1/2} \\ \mathbf{S}_{\mathbf{YY}}^{-1/2} \mathbf{S}_{\mathbf{YX}} \mathbf{S}_{\mathbf{XX}}^{-1/2} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_x \\ \mathbf{v}_y \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{v}_x \\ \mathbf{v}_y \end{bmatrix}, \tag{6}$$

where $\mathbf{v}_x = \mathbf{S}_{\mathbf{XX}}^{1/2T} \mathbf{w}_x$, and $\mathbf{v}_y = \mathbf{S}_{\mathbf{YY}}^{1/2T} \mathbf{w}_y$. This last formula shows that the CCA functional can be rewritten as

$$\rho(\mathbf{X}\mathbf{v}_x, \mathbf{Y}\mathbf{v}_y) = \frac{\mathbf{v}_x^T \mathbf{S}_{\mathbf{XX}}^{-1/2} \mathbf{S}_{\mathbf{XY}} \mathbf{S}_{\mathbf{YY}}^{-1/2} \mathbf{v}_y}{\sqrt{\mathbf{v}_x^T \mathbf{v}_x}\sqrt{\mathbf{v}_y^T \mathbf{v}_y}}. \tag{7}$$

Through standard algebraic derivations formula (7) provides an alternative formulation for the projection vectors. Introducing the vector $\mathbf{u}^T = \mathbf{v}_x^T \mathbf{S}_{\mathbf{XX}}^{-1/2} \mathbf{S}_{\mathbf{XY}} \mathbf{S}_{\mathbf{YY}}^{-1/2}$, we observe that the product $\mathbf{u}^T \mathbf{v}_y$ at the numerator of formula (7) is bounded by $\sqrt{(\mathbf{u}^T\mathbf{u})(\mathbf{v}_y^T \mathbf{v}_y)}$, with the equality holding if and only if $\mathbf{v}_y$ if parallel to $\mathbf{u}$ (Cauchy-Schwarz inequality). In this case, formula (7) can be rewritten as the new functional

$$\rho'(\mathbf{X}\mathbf{v}_x, \mathbf{X}\mathbf{v}_x) = \frac{\mathbf{u}^T\mathbf{u}}{\sqrt{\mathbf{v}_x^T \mathbf{v}_x}} = \frac{\mathbf{v}_x^T \mathbf{S}_{\mathbf{XX}}^{-1/2} \mathbf{S}_{\mathbf{XY}} \mathbf{S}_{\mathbf{YY}}^{-1} \mathbf{S}_{\mathbf{YX}} \mathbf{S}_{\mathbf{XX}}^{-1/2} \mathbf{v}_x}{\sqrt{\mathbf{v}_x^T \mathbf{v}_x}}. \tag{8}$$

By following an analogous derivation as in (4) we conclude that the solution of this functional is the eigenvector of the matrix $\Sigma = \mathbf{S}_{\mathbf{XX}}^{-1/2}\mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{YY}}^{-1}\mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1/2}$. Thus, the CCA solution can be alternatively obtained by computing $\mathbf{w}_x = \mathbf{S}_{\mathbf{XX}}^{-1/2}\mathbf{v}_x$, where $\mathbf{v}_x$ is eigenvector of Σ, and by defining $\mathbf{w}_y = \mathbf{S}_{\mathbf{YY}}^{-1/2}\mathbf{u}$.

Formula (5) highlights the numerical drawbacks of CCA, as its computation depends on the sample covariances $\mathbf{S}_{\mathbf{XX}}$ and $\mathbf{S}_{\mathbf{YY}}$. In particular, being these matrices estimated from sample data, the associated eigenvalues may quickly become small and unavoidably compromise the stability of the estimation. For this reason it is common practice to reformulate the CCA problem with a *regularized* version aimed at improving stability. This is performed by stabilizing the right hand side of (4) by adding a constant diagonal term depending on a regularization parameter δ:

$$\begin{bmatrix} \mathbf{0} & \mathbf{S}_{\mathbf{XY}} \\ \mathbf{S}_{\mathbf{YX}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{S}_{\mathbf{XX}} + \delta\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{\mathbf{YY}} + \delta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix}, \tag{9}$$

3.1.2. Partial Least Squares

Analogously to CCA, PLS is based on the identification of linear projections maximising the *covariance* between the projected data:

$$\mathbf{w}_x, \mathbf{w}_y = \underset{\mathbf{w}_x, \mathbf{w}_y}{\operatorname{argmax}}\, cov(\mathbf{X}\mathbf{w}_x, \mathbf{Y}\mathbf{w}_y), \tag{10}$$

where

$$cov(\mathbf{X}\mathbf{w}_x, \mathbf{Y}\mathbf{w}_y) = \frac{\mathbf{w}_x^T \mathbf{S}_{\mathbf{XY}} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{w}_x}\sqrt{\mathbf{w}_y^T \mathbf{w}_y}}. \tag{11}$$

As before, the PLS problem can be optimized through the solution of an ordinary eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{S}_{\mathbf{XY}} \\ \mathbf{S}_{\mathbf{YX}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix}, \tag{12}$$

PLS has interesting analogies with CCA. In particular, the PLS solution can be seen as the maximally regularized version of CCA obtained when the regularizing parameter δ of Formula (9) tends towards infinity. The solution corresponding to the eigenmodes of the problem (12) is known as PLS-SVD, and has been popularized in the field of neuroimaging in the seminal works [35, 36], for the study of positron emission tomography (PET) and functional magnetic resonance images (fMRI) through the analysis of the associated eigenmodes of intensity variation.

3.1.3. Iterative Numerical Schemes for PLS and CCA: NIPALS

In practice, to avoid the numerical instabilities related to the decomposition of potentially large sample covariance matrices, both PLS and CCA can be computed by leveraging on stable numerical schemes. In particular, the *non-linear iterative partial least squares* (NIPALS) is a classical algorithm proposed by H. Wold [37] for the iterative computation of the PLS (and CCA) solution. Within this method, the principal eigenmodes $\mathbf{w}_x^{(0)}$ and $\mathbf{w}_y^{(0)}$ are initially computed from the data matrices $\mathbf{X}^{(0)} = \mathbf{X}$ and $\mathbf{Y}^{(0)} = \mathbf{Y}$ with the iterative scheme detailed in Algorithm 1 [38].

After the estimation of the eigenmode of iteration i, the high-order components $\mathbf{w}_x^{(i)}$ and $\mathbf{w}_y^{(i)}$ are subsequently computed by applying NIPALS on the deflated data matrices $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$ obtained by regressing out the current projections in the latent space $\boldsymbol{t}^{(i)} = \mathbf{X}^{(i)}\mathbf{w}_x^{(i)}$, $\boldsymbol{u}^{(i)} = \mathbf{Y}^{(i)}\mathbf{w}_y^{(i)}$:

$$\mathbf{X}^{(i+1)} = \mathbf{X}^{(i)} - \boldsymbol{t}^{(i)}\frac{\boldsymbol{t}^{(i)T}\mathbf{X}^{(i)}}{\boldsymbol{t}^{(i)T}\boldsymbol{t}^{(i)}},$$
$$\mathbf{Y}^{(i+1)} = \mathbf{Y}^{(i)} - \boldsymbol{u}^{(i)}\frac{\boldsymbol{u}^{(i)T}\mathbf{Y}^{(i)}}{\boldsymbol{u}^{(i)T}\boldsymbol{u}^{(i)}}.$$

Algorithm 1 NIPALS iterative eigenmode computation for component i [38].

```
Initialize $\mathbf{w}_y^{(i)}$. While not converged do:
    1. $\boldsymbol{u}^{(i)} = \mathbf{Y}^{(i)}\mathbf{w}_y^{(i)}$
    2. Estimate the projection for $\mathbf{X}^{(i)}$ from the latent space of $\mathbf{Y}^{(i)}$
         PLS. $\mathbf{w}_x^{(i)} = \mathbf{X}^{(i)T}\boldsymbol{u}^{(i)}/\boldsymbol{u}^{(i)T}\boldsymbol{u}^{(i)}$.
         CCA. $\mathbf{w}_x^{(i)} = \mathbf{X}^{(i)*}\boldsymbol{u}^{(i)}$,
       where $\mathbf{X}^{(i)*}$ is the Moore-Penrose inverse of $\mathbf{X}^{(i)}$.
    3. Normalize $\mathbf{w}_x^{(i)}$.
    4. $\boldsymbol{t}^{(i)} = \mathbf{X}^{(i)}\mathbf{w}_x^{(i)}$
    5. Estimate the projection for $\mathbf{Y}^{(i)}$ from the latent space of $\mathbf{X}^{(i)}$
         PLS. $\mathbf{w}_y^{(i)} = \mathbf{Y}^{(i)T}\boldsymbol{t}^{(i)}/\boldsymbol{t}^{(i)T}\boldsymbol{t}^{(i)}$,
         CCA. $\mathbf{w}_y^{(i)} = \mathbf{Y}^{(i)*}\boldsymbol{t}^{(i)}$,
       where $\mathbf{Y}^{(i)*}$ is the Moore-Penrose inverse of $\mathbf{Y}^{(i)}$.
    6. Normalize $\mathbf{w}_y^{(i)}$.
```

The implementation illustrated in Algorithm 1 can be found in classical statistical and machine learning packages, such as Scikit-learn[1] [39].

3.1.4. Reduced Rank Regression

Reduced Rank Regression (RRR) differs from CCA and PLS since the relationship between the data matrices $\mathbf{X}$ and $\mathbf{Y}$ is estimated by optimizing a standard regression problem with Gaussian noise ϵ:

$$\mathbf{Y} = \mathbf{XC} + \epsilon. \tag{13}$$

[1] `http://scikit-learn.org/stable/`.

The model parameters $\mathbf{C}$ are a matrix of dimensions $D_i \times D_g$ with rank $R \leq min(D_i, D_g)$ [40]. The decomposition $\mathbf{C} = \mathbf{AB}$, with $dim(A) = D_i \times R$ and $dim(B) = R \times D_g$ provides an way to interpret RRR as the composition of respectively a linear projection of $\mathbf{X}$ into a latent space of dimension R, and a linear reconstruction of $\mathbf{Y}$ from the latent representation. Thanks to this representation, $\mathbf{A}$ and $\mathbf{B}$ can be interpreted to jointly provide a quantification of the relationship between imaging and genetic features [41].

The optimization of RRR is equivalent to the minimization of the loss:

$$f(\mathbf{A}, \mathbf{B}) = tr\{(\mathbf{Y} - \mathbf{XAB})\Gamma(\mathbf{Y} - \mathbf{XAB})^T\}, \tag{14}$$

for any given positive definite matrix Γ [40]. Within this formulation, the RRR solution is

$$\mathbf{A} = \Gamma^{-1/2}\mathbf{U}, \qquad \mathbf{B} = \mathbf{U}^T\Gamma^{1/2}\mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1},$$

where $\mathbf{U}$ are the eigenvectors associated to the matrix $\mathbf{R} = \Gamma^{1/2}\mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{XY}}\Gamma^{1/2}$. The role of Γ is to account for the covariance between the features of $\mathbf{Y}$, and can be set to identity or any other suitable form. For example, by setting $\Gamma = \mathbf{S}_{\mathbf{YY}}$, we obtain the CCA solution derived from formula (8).

3.1.5. Parallel Independent Component Analysis

Although less standard in the multivariate analysis literature, *Parallel ICA* (pICA) is an established method in medical imaging and imaging-genetics [42], building on the classical framework of *independent component analysis* (ICA) [43].

ICA-based methods are conceptually different from the multivariate approaches previously described in this chapter (CCA, PLS, RRR). Indeed, the latter are essentially based on the identification of a low-dimensional data representation in which the statistical association expressed by the covariance, or the correlation, is maximized. In practice, this amounts at assuming a Gaussian distribution of the data, and to the identification of an opportune linear change of coordinates leading to maximum data likelihood.

On the contrary, when applying ICA to a single modality $\mathbf{X}$, we assume that the observed data is a linear mixture of *non-Gaussian distributed and independent* latent sources (or there is at most a single Gaussian component). The associate generative model is:

$$\mathbf{X} = \mathbf{SW}_x, \tag{15}$$

where $\mathbf{S}$ and $\mathbf{W}$ are respectively sources and mixing matrix. Several ICA approaches for the joint optimization of $\mathbf{S}$ and $\mathbf{W}$ have been proposed in the literature [44]. For example, when the measure of non-Gaussianity is given by the negentropy [45], the FastICA algorithm is a very efficient ICA approach based on the Newton's gradient-based optimization [46].

Given imaging-genetics data matrices, pICA joinlty optimizes the ICA problem separately on the two modalities [47]:

$$\mathbf{X} = \mathbf{S}_x\mathbf{W}_x,$$
$$\mathbf{Y} = \mathbf{S}_y\mathbf{W}_y,$$

while also maximising the correlation between the columns of the respective mixing matrices $Corr(\mathbf{W}_x[:, i], \mathbf{W}_y[:, j])$. This approach thus aims at identifying the mixing parameters to reconstruct the data while at the same time highlighting their relationship.

3.2. Regularization in Multivariate Models

When dealing with high-dimensional data, such as the one available imaging-genetics, multivariate approaches are highly prone to the problem of over-fitting and to the identification of spurious associations. This problem ultimately leads to low stability and poor interpretability of the analysis results.

A common strategy for mitigating this problem is through model regularization, achieved by constraining the solution to belong to a suitable parameter space. These constraints are usually specified by introducing a penalization term on the parameters' norm. In the following section we illustrate the main approaches to regularization along with their principal applications in imaging-genetics.

3.2.1. Sparsity an Smoothness

Sparsity is a classical regularization choice, known as *least absolute shrinkage and selection operator* (LASSO), looking for a limited subset of non-zero parameters' coefficients. In its simplest formulation, given an objective function $f(\mathbf{w})$, sparsity on the solution $\mathbf{w}$ is imposed by solving the associated functional:

$$\underset{\mathbf{w}}{\operatorname{argmin}} f(\mathbf{w}) + \lambda ||\mathbf{w}||_1, \tag{16}$$

which corresponds to the Lagrangian of the optimization of $f(\mathbf{w})$ constrained to $|\mathbf{w}| \leq t$, for an opportune t. The left panel of figure 4 illustrates the sparsity effect of the LASSO penalty on the identification of the minimum for a given function f (red contours). Increasing the sparsity constraint reduces the size of the regularization contour (black line), enforcing the solution parameters to zero (in this case the parameter w_1).

Another classical form of regularization, known as *Ridge* or *Tikhonov* regularization, consists in introducing a penalization of the ℓ_2 norm of the parameters:

$$\underset{\mathbf{w}}{\operatorname{argmin}} f(\mathbf{w}) + \lambda ||\mathbf{w}||_2^2. \tag{17}$$

While this regularization form does not promote sparsity (Figure 4, right), it is used to enhance regularity of the solution and promote the well-posedness of the optimization problem.

In applied studies, approaches to regularized CCA [48, 49], PLS [50, 51], and RRR [41] have been proposed by introducing an additional parameter penalization term of the ℓ_1 norm (LASSO, [52]), or jointly to ℓ_1 and ℓ_2 (so-called Elastic Net regularization [53]).

In its simplest formulation, sparsity in CCA and PLS can be implemented by soft-thresholding of the parameters' weights, based on a pre-defined ratio of desired resulting non-zero coefficients [48, 51]. Regularity can be also introduced in NIPALS [50], by estimating regularized projections $\overline{\mathbf{w}_x}$ and $\overline{\mathbf{w}_y}$ through Elastic Net (Algorithm 2). It is interesting

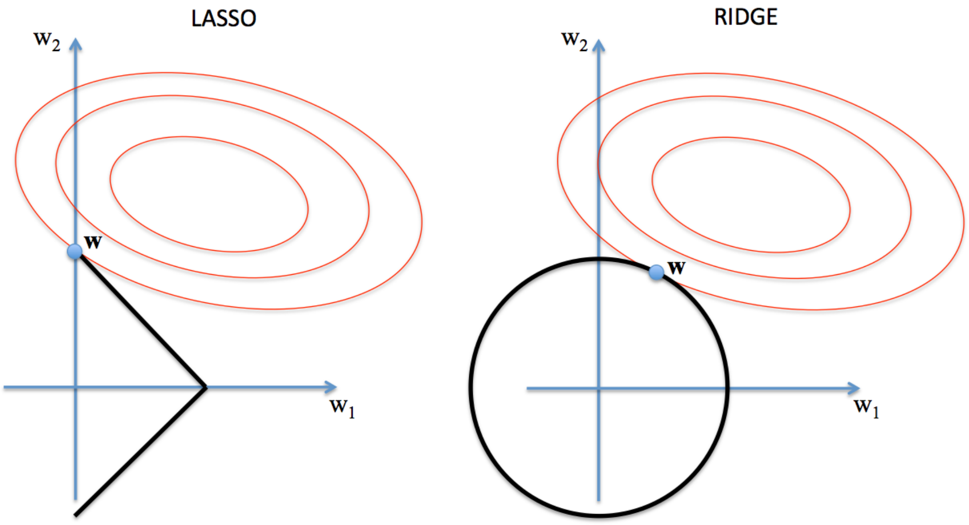


Figure 4: An illustration of regularized regression with LASSO and Ridge penalization. Red: contours for the solution of the function $f(\mathbf{w})$. Black: Regularization contours. In LASSO, the geometric constraint associated to the ℓ_1 penalization promotes sparse solutions (e.g. $w_1 = 0$).

to notice that the method reduces to univariate LASSO soft-thresholding when the ℓ_2 parameters $\lambda_{y2}, \lambda_{x2}$ are set to infinity [53].

Algorithm 2 Regularization of projections parameters $\mathbf{w}_x$ and $\mathbf{w}_y$ in NIPALS.

```
Given current estimates of $\mathbf{w}_x$ and $\mathbf{w}_y$.
While not converged do:
    1. compute $\mathbf{t} = \mathbf{X}\mathbf{w}_x$,
    2. compute $\mathbf{u} = \mathbf{Y}\mathbf{w}_y$,
    3. compute $\overline{\mathbf{w}_x}$ by solving the Elastic-Net regression:
        $\overline{\mathbf{w}_x} = \arg\min_{\mathbf{v}} (\mathbf{t} - \mathbf{X}\mathbf{v})^2 + \lambda_{x2}\|\mathbf{v}\|_2^2 + \lambda_{x1}\|\mathbf{v}\|_1,$
    4. compute $\overline{\mathbf{w}_y}$ by solving the Elastic-Net regression:
        $\overline{\mathbf{w}_y} = \arg\min_{\mathbf{v}} (\mathbf{u} - \mathbf{Y}\mathbf{v})^2 + \lambda_{y2}\|\mathbf{v}\|_2^2 + \lambda_{y1}\|\mathbf{v}\|_1,$
    3. Normalize $\overline{\mathbf{w}_x}$ and $\overline{\mathbf{w}_x}$,
    4. Set $\mathbf{w}_x = \overline{\mathbf{w}_x}$, $\mathbf{w}_y = \overline{\mathbf{w}_y}$.
```

The experimental investigation proposed in [51] provides a useful comparison of sparse vs non-sparse implementations of multivariate approaches in imaging-genetics. In particular, although sparsity was shown to generally help in identifying more stable solutions, the choice of regularization parameters, cross-validation and data pre-processing strategies (especially data filtering and dimensionality reduction) seems to play a critical role for the reliability of the solution.

Moreover, an important issue is still represented by the intrinsic signal correlation characterizing both imaging and genetic data. While the imaging data is characterized by non-trivial spatial correlations, affecting for example the signal of neighbouring voxels, SNPs are spatially correlated due to linkage disequilibrium. In this case, the solutions provided by LASSO and Ridge are in opposition and thus not straightforwardly interpretable. Indeed, LASSO tends to isolate a single variable from a set of correlated ones, while Ridge identifies the whole set of correlated predictors. However, in the practical context, this behaviour is not guaranteed and may be strongly related to data variability and dimensions [51].

For this reason, more complex regularization strategies accounting for known relationships across imaging and genetic features have been proposed in the last years. The next section provides an overview of this more advanced modeling topic.

3.2.2. Group-wise Penalization

Imaging-genetics data is characterized by non-trivial correlations representing precise biological and anatomical mechanisms.

On one hand SNPs are known to act “in concert” through *biological pathways* representing specific biological processes (metabolic, cellular, genetic). These processes can be represented via relation networks, as provided in the KEGG pathway database[2]. Similarly, we can rely on hand-crafted ontologies describing gene functions and relationships (Gene Ontology Consortium[3]). On the other hand, medical imaging data is characterized by both local correlations across neighbouring voxels, as well as non-local ones. For example, due to the organization of the brain in anatomical and functional networks, the signal measured in brain imaging data exhibits important network properties representing specific physiological mechanisms.

In this context, the stability and interpretability of the multivariate analysis results can be promoted by introducing an additional regularization constraint on the model parameters. This constraint aims at enforcing the model to provide solutions compatible with our prior information on the relationship between features.

This problem can be addressed by introducing a group-wise penalization [54]. Within this approach, variables known to act together are jointly regularized with respect to the same penalization parameter. In its basic formulation, we assume the D features of a predictor $\mathbf{x} = (x_1, \ldots, x_D)$ are grouped in L sets $\mathcal{L}_l = (x_{i_1}, \ldots, x_{i_l})$. The group-wise penalized regression of the independent variable $\mathbf{y}$ with respect to the predictors $\mathbf{X}$ is thus expressed as:

$$f(\mathbf{a}) = ||\mathbf{y} - \mathbf{X}\mathbf{a}||_2^2 + \lambda \sum_{l=1}^{L} w_l ||\mathbf{a}_l||_2,$$

where the vectors $\mathbf{a}_l = (a_{i_1}, \ldots, a_{i_l})$ group the elements of $\mathbf{a}$ associated to each set of features $\mathcal{L}_l$. Computational issues may arise when the sets $\mathcal{L}_l$ are not disjoint, and specific optimization strategies need to be defined to address the problems of identifiability and computational efficiency [55].

[2] https://www.genome.jp/kegg/pathway.html

[3] http://geneontology.org/

Building on this methodological framework, an extension of RRR to account for group-wise relationships was introduced in [56]. In particular, by considering a rank-one RRR model, formula (14) can be extended to

$$f(\mathbf{a}, \mathbf{b}) = tr\{(\mathbf{Y} - \mathbf{Xab})(\mathbf{Y} - \mathbf{Xab})^T\} + \lambda \sum_{l=1}^{L} w_l ||\mathbf{a}_l||_2, \tag{18}$$

where the coefficients $\mathbf{a}_l$ account for known relationships on the genetic data. Similarly to the RRR case, other extensions to group-wise penalization in multivariate imaging-genetics models have been proposed, for example applied to CCA [57], or to sparse regression [58].

3.3. Stability and validation of multivariate models

The promise of multivariate models to unveil hidden relationship in high-dimensional data is often hindered by the problem of overfitting and lack of stability of the results. In particular, overfitting is a critical issue concerning the drop in the generalization ability of the model when tested on independent data. In the previous sections we have seen that regularization can be used to identify more meaningful and stable solutions. However, the introduction of regularization comes with the problem of tuning the associated regularity parameter (e.g., the LASSO parameter λ of Equation (16)). Since the model solution often greatly varies depending on the imposed degree of regularization, this problem ultimately affects the stability of the results. For this reason it is common practice to tune the parameters by grid search on a subset of hold-out data, to identify the parameter set leading to the optimal cross-validation results [51, 56, 41]. Unfortunately this practice may also lead to overfit and selection-bias [59], especially when the sample size is low compared to the data dimensionality. Again, in this case the marginal improvement related to the tuning of the parameters does not lead to an improvement in the generalization of the model in unobserved data.

All in all, the validity of multivariate models should be ultimately established via testing on datasets not used to estimate the model parameters. This is a critical aspect related to the availability of independent data compatible with the one used for the model fitting. Current effort in standardization and opening of high-quality datasets are becoming fundamental for the reliable use of multivariate methods in imaging-genetics and, more generally, in biomedical applications.

References

[1] C. Willyard, New human gene tally reignites debate, Nature 558 (7710) (2018) 354.

[2] E. P. Consortium, et al., An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57.

[3] D. S. Falconer, Introduction to quantitative genetics, Oliver And Boyd; Edinburgh; London, 1960.

[4] J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis, The American Journal of Human Genetics 88 (1) (2011) 76–82.

[5] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, S. W. G. of the Psychiatric Genomics Consortium, et al., LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, Nature genetics 47 (3) (2015) 291.

[6] E. Perez, K. E. Sullivan, Chromosome 22q11. 2 deletion syndrome (DiGeorge and velocardiofacial syndromes), Current opinion in pediatrics 14 (6) (2002) 678–683.

[7] J. Flint, M. R. Munafò, The endophenotype concept in psychiatric genetics, Psychological medicine 37 (2) (2007) 163–180.

[8] D. P. Hibar, J. L. Stein, M. E. Renteria, A. Arias-Vasquez, S. Desrivières, N. Jahanshad, R. Toro, K. Wittfeld, L. Abramovic, M. Andersson, et al., Common genetic variants influence human subcortical brain structures, Nature 520 (7546) (2015) 224.

[9] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, The American Journal of Human Genetics 81 (3) (2007) 559–575.

[10] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al., Genes mirror geography within Europe, Nature 456 (7218) (2008) 98.

[11] J. Marchini, B. Howie, Genotype imputation for genome-wide association studies, Nature Reviews Genetics 11 (7) (2010) 499.

[12] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, et al., A reference panel of 64,976 haplotypes for genotype imputation, Nature genetics 48 (10) (2016) 1279.

[13] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, et al., Next-generation genotype imputation service and methods, Nature genetics 48 (10) (2016) 1284.

[14] S. Sadaghiani, B. Ng, A. Altmann, J.-B. Poline, T. Banaschewski, A. L. Bokde, U. Bromberg, C. Büchel, E. B. Quinlan, P. Conrod, et al., Overdominant effect of a CHRNA4 polymorphism on cingulo-opercular network activity and cognitive control, Journal of Neuroscience 37 (40) (2017) 9657–9666.

[15] D. B. Dubal, J. S. Yokoyama, L. Zhu, L. Broestl, K. Worden, D. Wang, V. E. Sturm, D. Kim, E. Klein, G.-Q. Yu, et al., Life extension factor klotho enhances cognition, Cell reports 7 (4) (2014) 1065–1076.

[16] P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, A. L. Price, Mixed-model association for biobank-scale datasets, Nature genetics (2018) 1.

[17] C. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8 (1936) 3–62.

[18] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the royal statistical society. Series B (Methodological) (1995) 289–300.

[19] C. J. Hoggart, T. G. Clark, M. De Iorio, J. C. Whittaker, D. J. Balding, Genome-wide significance for dense SNP and resequencing data, Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 32 (2) (2008) 179–185.

[20] J. Kim, W. Pan, A cautionary note on using secondary phenotypes in neuroimaging genetic studies, NeuroImage 121 (2015) 136–145.

[21] B. Aaron, J. D. Kromrey, J. Ferron, Equating" r"-based and" d"-based effect size indices: problems with a commonly recommended formula, ERIC Clearinghouse, 1998.

[22] J. J. Lee, R. Wedow, A. Okbay, E. Kong, O. Maghzian, M. Zacher, T. A. Nguyen-Viet, P. Bowers, J. Sidorenko, R. K. Linnér, et al., Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals, Nature genetics (2018) 1.

[23] J. E. Iglesias, J. C. Augustinack, K. Nguyen, C. M. Player, A. Player, M. Wright, N. Roy, M. P. Frosch, A. C. McKee, L. L. Wald, et al., A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI, Neuroimage 115 (2015) 117–137.

[24] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, et al., A multi-modal parcellation of human cerebral cortex, Nature 536 (7615) (2016) 171–178.

[25] L. Elliott, K. Sharp, F. Alfaro-Almagro, S. Shi, K. Miller, G. Douaud, J. Marchini, S. Smith, Genome-wide association studies of brain structure and function in the UK Biobank, bioRxiv.

[26] M. A. Scelsi, R. R. Khan, M. Lorenzi, L. Christopher, M. D. Greicius, J. M. Schott, S. Ourselin, A. Altmann, Genetic study of multimodal imaging AlzheimerâĂŹs disease progression score implicates novel loci, Brain.

[27] J. L. Stein, X. Hua, S. Lee, A. J. Ho, A. D. Leow, A. W. Toga, A. J. Saykin, L. Shen, T. Foroud, N. Pankratz, et al., Voxelwise genome-wide association study (vGWAS), neuroimage 53 (3) (2010) 1160–1174.

[28] P. M. Thompson, J. L. Stein, S. E. Medland, D. P. Hibar, A. A. Vasquez, M. E. Renteria, R. Toro, N. Jahanshad, G. Schumann, B. Franke, et al., The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data, Brain imaging and behavior 8 (2) (2014) 153–182.

[29] M. Huang, T. Nichols, C. Huang, Y. Yu, Z. Lu, R. C. Knickmeyer, Q. Feng, H. Zhu, FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data, Neuroimage 118 (2015) 613–627.

[30] J. Schrouff, M. J. Rosa, J. M. Rondina, A. F. Marquand, C. Chu, J. Ashburner, C. Phillips, J. Richiardi, J. Mourão-Miranda, PRoNTo: pattern recognition for neuroimaging toolbox, Neuroinformatics 11 (3) (2013) 319–337.

[31] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, Y. V. Sun, Machine learning in genome-wide association studies, Genetic epidemiology 33 (S1) (2009) S51–S57.

[32] J. Liu, V. D. Calhoun, A review of multivariate analyses in imaging genetics, Frontiers in neuroinformatics 8 (2014) 29.

[33] M. Lorenzi, A. Altmann, B. Gutman, S. Wray, C. Arber, D. P. Hibar, N. Jahanshad, J. M. Schott, D. C. Alexander, P. M. Thompson, S. Ourselin, Susceptibility of brain atrophy to trib3 in alzheimer's disease, evidence from functional prioritization in imaging genetics, Proceedings of the National Academy of Sciences 115 (12) (2018) 3162–3167. doi:10.1073/pnas.1706100115.

[34] T. De Bie, N. Cristianini, R. Rosipal, Eigenproblems in pattern recognition, in: Handbook of Geometric Computing, Springer, 2005, pp. 129–167.

[35] A. McIntosh, F. Bookstein, J. V. Haxby, C. Grady, Spatial pattern analysis of functional brain images using partial least squares, Neuroimage 3 (3) (1996) 143–157.

[36] K. J. Worsley, An overview and some new developments in the statistical analysis of PET and fMRI data, Human Brain Mapping 5 (4) (1997) 254–258.

[37] H. Wold, Path models with latent variables: The NIPALS approach, in: Quantitative sociology, Elsevier, 1975, pp. 307–357.

[38] M. Tenenhaus, L'approche pls, Revue de statistique appliquée 47 (2) (1999) 5–40.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[40] R. Velu, G. C. Reinsel, Multivariate reduced-rank regression: theory and applications, Vol. 136, Springer Science & Business Media, 2013.

[41] M. Vounou, T. E. Nichols, G. Montana, Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, Neuroimage 53 (3) (2010) 1147–1159.

[42] G. D. Pearlson, V. D. Calhoun, J. Liu, An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders, Frontiers in genetics 6 (2015) 276.

[43] P. Comon, Independent component analysis, a new concept?, Signal processing 36 (3) (1994) 287–314.

[44] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural networks 13 (4-5) (2000) 411–430.

[45] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[46] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE transactions on Neural Networks 10 (3) (1999) 626–634.

[47] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, V. Calhoun, Combining fMRI

and SNP data to investigate connections between brain function and genetics using parallel ICA, Human brain mapping 30 (1) (2009) 241–255.

[48] E. Parkhomenko, D. Tritchler, J. Beyene, Genome-wide sparse canonical correlation of gene expression with genotypes, in: BMC proceedings, Vol. 1, BioMed Central, 2007, p. S119.

[49] D. M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.

[50] S. Waaijenborg, A. H. Zwinderman, Penalized canonical correlation analysis to quantify the association between gene expression and dna markers, in: BMC proceedings, Vol. 1, BioMed Central, 2007, p. S122.

[51] É. Le Floch, V. Guillemot, V. Frouin, P. Pinel, C. Lalanne, L. Trinchera, A. Tenenhaus, A. Moreno, M. Zilbovicius, T. Bourgeron, et al., Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares, Neuroimage 63 (1) (2012) 11–24.

[52] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[53] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[54] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1) (2006) 49–67.

[55] M. Silver, G. Montana, Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps, Statistical applications in genetics and molecular biology 11 (1) (2012) 1–43.

[56] M. Silver, E. Janousova, X. Hua, P. M. Thompson, G. Montana, Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression, NeuroImage 63 (3) (2012) 1681–1694.

[57] D. Lin, J. Zhang, J. Li, V. D. Calhoun, H.-W. Deng, Y.-P. Wang, Group sparse canonical correlation analysis for genomic data integration, BMC bioinformatics 14 (1) (2013) 245.

[58] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant snps, Bioinformatics 28 (18) (2012) i619–i625.

[59] A. F. Mendelson, M. A. Zuluaga, M. Lorenzi, B. F. Hutton, S. Ourselin, Selection bias in the reported performances of ad classification pipelines, NeuroImage: Clinical 14 (2017) 400–416.