MULTI-STEP STATISTICAL METHODS FOR SIMULTANEOUS INFERENCE IN GENETICS
MULTI-STEP STATISTICAL METHODS FOR SIMULTANEOUS INFERENCE IN GENETICS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan de Vrije Universiteit Amsterdam, op gezag van de rector magnificus prof.dr. V. Subramaniam, in het openbaar te verdedigen ten overstaan van de promotiecommissie van de Faculteit der Exacte Wetenschappen op maandag 7 november 2016 om 11.45 uur in het auditorium van de universiteit, De Boelelaan 1105

door

Jakub Pečánka

geboren te Praag, Tsjehoslowakije
promotor: prof.dr. A.W. van der Vaart
copromotoren: dr. M.A. Jonker
dr. Z. Bochdanovits
The work on this thesis has been a very long project. Definitely much longer than I had anticipated. In fact, had I known that the production of this thesis was going to take eight years I am quite certain I would have thought much more carefully about starting the project in the first place. Despite that, or perhaps precisely because of that, looking back at my journey I feel a great sense of satisfaction. I have learned a lot in the past eight years and I feel that I can honestly say I have grown both as a researcher and, even more importantly, as a person. And here I would like to express my gratitude to those that have participated in my journey and provided the necessary professional and personal support and structure without which getting to this point would hardly have been possible.

Naturally, I would like to start by thanking my PhD supervisors: professor Aad van der Vaart and doctors Marianne Jonker and Zoltan Bochdanovits. To all three I owe a huge amount of gratitude for giving me the opportunity to start my PhD project in the first and for guiding my work throughout all these years. Aad’s calm personality and depth of knowledge, Marianne’s expertise, experience and general kindness as well as Zoltan’s ingenuity and boundless optimism have all been instrumental to my success. I also wish to thank professor Jelle Goeman with whom I collaborated on parts of this thesis and whose support and patience have greatly helped me to conclude the work on this thesis.

Besides the professional assistance I have enjoyed, of equal importance for my success have been the boundless support and love I received and continue to receive from my family, my beloved and eternally loving mom Jaroslava, my gentle and caring brother Marek and my helpful father Karel. A very special thank you also goes out to my former life partner Melania, whose love and kindness have guided me through many moments of darkness and whose optimistic and cheerful presence I had the pleasure to enjoy throughout most my journey. Hugely important to my success have also been the love and friendship I get from my friends Mehran Aflakparast, Jaroslav Pazdera, Bartek Knapik, Jurriaan Oostermann, Demeter Kiss, Kristýna Jungová, Tomáš Marada, Tomáš Soukup, Jan-Michal Mleziva, Jiří Horák, Melinda Paugschová-Niedrlová, Jan Horák, and many others.

I would also like to express my gratitude for the support of my colleagues at the De-
partment of Mathematics at VU University Mathisca de Gunst, Beata Ros, Gwenael Leday, Michael McAssey to name just a few. I also wish to thank all of the members of the Biostatistics group at the Department for Health Evidence at Radboud University Medical Center, and my colleagues at the Department of Medical Statistics and Bioinformatics at Leiden University Medical Center, where I sporadically worked on this thesis after my departure from VU. I also wish to thank all of the aforementioned institutions together with the VU University medical center for the institutional support I have received during the work on this project.

I wish to acknowledge the financial support by the Netherlands Scientific Organization. Large portions of the statistical analyses in this thesis were carried out using the Genetic Cluster Computer (www.geneticcluster.org) which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003 PI Posthuma) along with a supplement from the Dutch Brain Foundation and VU University.

Finally, a huge thank you to Marianne Jonker and Emke de Vries for translating the summary of this thesis into Dutch.

To these and others, with gratitude and love,

Jakub Pečánka

Amsterdam
September 2016
Contents

Acknowledgements v

1 Introduction 3
   1.1 Basic terminology of genetics 4
   1.2 Principles of association mapping 6
   1.3 Outline of the thesis 7
   1.4 Notational conventions 9

PART I  |  TWO-STAGE SEARCH FOR EPISTASIS 11

2 A brief introduction to testing for epistasis 13
   2.1 Genetic terminology 14
   2.2 Genetic interaction (epistasis) 15
   2.3 Linkage disequilibrium 18
   2.4 Multiple testing problem 22
   2.5 Examples of two-stage testing methods 24
   2.6 Our goal in Part I of the thesis 25

3 Two-stage testing for epistasis: Theory 27
   3.1 Interactions in logistic regression model 28
   3.2 Suitable pre-test statistics 32
   3.3 Post-test for pooled-sample pre-test 43
   3.4 Post-tests based on disjoint samples 44
   3.5 Post-tests based on adjusting for pre-test 46
   3.6 Error control in two-stage multiple testing 55

4 Two-stage testing for epistasis: Application 61
   4.1 Software tool: EpiDet 62
   4.2 Interaction penetrance models 62
   4.3 Analysis methods 64
   4.4 Tuning parameters 66
   4.5 Simulation study: Investigation of type I error control 73
   4.6 Simulation study: Investigation of power 78
   4.7 Real data analysis: Parkinson’s disease study 94
   4.8 Usage of EpiDet 97
Statistical genetics is an area at the intersection between genetics - the study of the genetic code inside cells of living organisms - and quantitative statistical analysis. While modern genetics started with Gregor Johann Mendel as far back as 1860s, it was not until the latter decades of the twentieth century that the discipline experienced a dramatic transformation from a mostly theoretical subject with limited room for empirical evidence to a heavily data-oriented discipline. As a result of novel genotyping and sequencing technologies and the progressive falling of costs of sequencing, the existence of large repositories of genetic data allows researchers to generate and explore new scientific hypotheses (Montana (2006), Pool et al. (2010)). Such shift dramatically increased the potential that the study of genetic code represents towards scientific understanding of life in general on the one hand, and towards advances in disciplines such as medicine, agriculture, and biotechnology on the other. For instance, as far as applications in medicine and health care are concerned, genetics promises to be the key to understanding and preventing and/or curing many diseases of both humans, animals and plants. Gene therapy could potentially treat certain disorders at their source by repairing the underlying genetic flaws, instead of just addressing the symptoms, and it can also be a way to create disease-resistant plants and animals. And it needs to be kept in mind that medical use is but one example of a broad range of possible applications of genetic research.

However, besides the promise of discovery and understanding, the current data-driven nature of genetic research also tests the limits of available models and computational analysis methods and fuels the need for further development of such models and methods. It is quite clear that for deeper understanding of the processes involved and for creation of usable tools strong statistical analysis methods are necessary. An example of such need is given by the study of (human) genetic makeup in its entirety and the efforts of linking sections of genetic code such as genes (see below) to external or internal observable characteristics of an individual called phenotypes (see below). Due to the size of such genome-wide data sets,
finding these links is often a truly challenging task in terms of both statistical and computational complexity. In this thesis we aim to contribute statistical methods to the already rich collection of tools available to tackle these challenges.

Below we introduce relevant basic genetic terminology and related concepts that create the context for and are necessary to understand the topics presented in this thesis. However, some of the underlying concepts are only touched upon and for full understanding a prior knowledge or further reading is generally advised. The thesis itself is divided into three parts. Here we start by defining terms such as genome, chromosome, gene, as well as basic principals of statistical analysis of genetic data, as these are relevant to both parts of the thesis. If necessary, at the beginning of each part these concepts are complemented by additional notions that are primarily relevant to only that part.

### 1.1 Basic terminology of genetics

In genetics, the term **genome** refers to the entire collection of hereditary information contained in cells of living organisms as coded by the DNA (deoxyribonucleic acid) for most organisms or RNA (ribonucleic acid) for many viruses (Ridley (2006)). Genetic information in the genome is encoded as a sequence of four nucleotides called guanine, adenine, thymine, and cytosine, which are coded using the letters G, A, T, and C, respectively. Most DNA molecules consist of two double-stranded helices, in which the nucleotides are bound together per pairs of guanine-cytosine and adenine-thymine. These pairs are usually referred to as nucleobases (or bases) and the exact knowledge of only one strand of the double-stranded helices therefore allows for complete reconstruction of the other strand of each helix. From a structural perspective, genetic information that makes up the genome of an organism is stored in distinct continuous portions of genetic sequence called **chromosomes**. A diagram of structural organization of DNA helix into chromosome is depicted in Figure 1.1a. The image shows a DNA helix wrapping around a set of proteins known as **histones**, which together make up a structure known as nucleosome that are organized into threads of **chromatin**, which finally condense into a chromosome. For diploid organisms such as humans there are two copies of each chromosome, one paternal and one maternal. A typical human has around 3 billion nucleobase pairs spread over 22 chromosomal pairs of autosomal (meaning non-sex) chromosomes plus 2 allosomal (meaning sex) chromosomes resulting in a total of 46 chromosomes. Figure 1.1b provides a truthful depiction of all 46 chromosomes of human male as seen under a microscope.

Genome can be divided into chunks of base sequences such as **genes**, regulatory sequences and others. Genes are those sections that have biological relevance such as containing information that codes for a protein. Although in a more informal way the word gene can also refer to any small section of genetic code irrespective of whether it actually codes for anything or not, we will stay away from this usage in this text. The position of a gene...
(or other significant sequence) on a chromosomal pair is called a locus (plural loci), where the position on a chromosomal pair refers to the two corresponding locations on each of the two copies of chromosome within a chromosomal pair. Alternative DNA sequences that occur for a given type of organism at the same physical locus are called alleles. A set of loci with their alleles at each locus is called genotype, but the word genotype can also refer to the entire genome of an individual depending on the context. Haplotype on the other hand refers to alleles within a collection of loci on a single chromosome from a chromosomal pair, i.e. genetic sequences of alleles that progeny inherits from one parent. If a locus has length 1 (measured in the number of nucleobases it consists of) and the nucleobase is not the same for all individuals of the considered species, it is referred to as single nucleotide polymorphism or SNP (pronounced as "snip") for short. For practical purposes the meaning of "not the same for all individuals" must be interpreted not in an absolute sense but rather in a relative sense with respect to a population of interest. Usually, it is required that at least a small fraction of individuals should have a different allele at the given position of the genome than the rest of the population and the cutoff is usually set at 1%. In the human genome there are around 10 million nucleobase pairs with both variants appearing with probability higher than 1%, or in other words there are around 10 million SNPs, which occur on average once in every 300 base pairs.

Figure 1.1: (a) Diagram of structural organization of DNA double-stranded helix into chromosome. Source: OpenStax College. The Nucleus and DNA Replication [Connexions Web site]. June 26, 2013. Available at http://cnx.org/content/m46073/1.4/ (b) Karyotype (a truthful depiction of the actual look under light microscope) of the 23 chromosomal pairs of a human male (using Giemsa staining). Source: Talking Glossary of Genetic Terms by the National Human Genome Research Institute (NHGRI) freely available at www.genome.gov/glossary.

In an organism the genetic information coded by genes is used to produce gene products such as proteins. The process through which the genetic information coded by a gene is utilized during production of gene products is called expression. Regulatory sequences are those parts of the genome which serve the purpose of increasing or decreasing the expression of
specific genes within an organism. While genotype is an internal hereditary characteristic of an individual, a phenotype or phenotypic value (Falconer and Mackay (1996)) of an individual is an observable characteristic of an individual, which is expressible in metric units, such as height, hair color, presence of a disease indicator, and blood type. For the purposes of analysis of such phenotypes within the context of a selected population it is useful to divide the phenotype into components that are attributable to different causes. In the usual setting two types of causes are considered: genetic and environmental. Genetic causes are understood to be the influence of the individual’s genome on the value of the phenotype, while environmental causes are all the non-genetic determinants of the phenotype. Therefore, these two causes by definition account for all of the variability or deviation in the given phenotype within the considered population. If we respectively denote by $P$, $G$ and $E$ the phenotypic value, the genotypic value and the environmental contribution to deviation we can symbolically write\footnote{We use the symbol “⊕” when denoting the combined effect of $G$ and $E$ instead of the often encountered “+” to avoid suggesting that the two causes necessarily combine in a linear fashion (i.e., additively). Combined effects can be decomposed into linear and non-linear components, where the non-linear combination of $G$ and $E$ is usually denoted as $G \times E$. In other words, we can write $G ⊕ E = G + E + G \times E$.}

\[ P = G ⊕ E. \]

### 1.2 Principles of association mapping

For the purposes of mapping complex diseases, one of two approaches is usually used. The first approach called linkage study uses family data (pedigree) to study links between genotypes and phenotypes. It attempts to discover genetic determinants, the so-called causal genes, from patterns of inheritance observed among the members of the pedigree. The second approach is called association study, which generally speaking can be classified as either family-based association study (FBAS) or population-based case-control study (PBCCS). FBAS use related individuals to assess the degree of association, which serves to avoid the potential confounding effects of population stratification. PBCCS, on the other hand, simply compare the genotypes of individuals affected by the disease (cases) with those of unaffected individuals (controls), where the individuals are randomly sampled from the population of interest based on their affection status in a prescribed proportion.

The possibility of accounting for population structure by FBAS, and the fact that in certain cases for rare diseases FBAS can be substantially more powerful than PBCCS, are two strong points on the side of FBAS (Laird and Lange (2006)). On the other hand, their usually more resource demanding nature, as measured in terms of money, time, as well as the amount of genotyping they require, made FBAS less popular in comparison with population-based designs.

Both linkage and association studies use markers in the analysis, which by definition are the observable characteristics of the genome. Typically, these markers are genetic loci that
were measured in the given data set, such as SNPs. The idea behind analyzing markers is to identify the causal loci directly if they are included among the markers, or at least to find markers that are associated with the phenotype in the hope that the true causal loci lie in close proximity of the associated markers. Such hope is fueled by the nature of dependence between loci as measured by linkage disequilibrium (see Section 2.3.1) and the fact that under random mating repeated recombination pushes loci towards independence. Since the rate of recombination tends to increase with physical distance between two loci, this provides justification for the physical proximity of associated marker and causal locus.

Small-scale association studies (SSAS) are traditionally deployed when the researchers suspect a specific gene or genetic region of association with the given phenotype. Such an approach focuses on a relatively small number of markers and by its nature does not allow to identify new genes or genetic regions in association with the phenotype. Relatively speaking, the multiple testing problem is not very severe in such studies.

Compared to SSAS, genome-wide association studies (GWAS) on the other hand are much more exploratory in their nature. The goal in GWAS is to scan large portions of the genome and identify novel genes or genetic regions that are associated with the studied phenotype. Conceptually, to perform a truly genome-wide search of a human genome, due to dependence among SNPs, it is not necessary to genotype all of the roughly 10 million SNPs. For instance, under the common disease/common variant hypothesis,\(^2\) it has been shown that a panel of well selected 500,000 to 1 million markers should identify common SNPs that are associated to common phenotypes (Bush and Moore (2012)). Such studies require cost-effective genotyping technologies that can accurately capture the alleles of up to 1 million SNPs for each studied individual in a cost-effective manner, which are becoming increasingly more available in the form of Next-Generation Sequencing methods (Grada and Weinbrecht (2013)). However, with association studies of such large size, an application of statistical methods unavoidably brings about a huge multiple testing problem. Moreover, in order to detect a sizable group of truly risk causing variants of mostly moderate size, large GWA studies require very large sample sizes. This is because of the obvious need to not only discover the risk causing variants, but also manage the number of false discoveries by accounting for multiple testing. Since available sample sizes are limited, accounting for multiple testing as efficiently as possible is critical.

### 1.3 Outline of the thesis

In this thesis we aim to address three general and important statistical problems highly relevant to modern genetics and numerous related fields. The common theme running through the thesis is the focus on efficient and reliable identification of rare signals (such as association

\(^2\)Common disease/Common variant hypothesis is the hypothesis that commonly occurring diseases in a population are caused in part by genetic variation that is common to that population (Bush and Moore (2012)).
of genes and phenotypes) in a large-scale setting via multi-step statistical analyses.

Part I
In Part I of the thesis, namely in Chapters 2 through 5, we present a novel statistical approach to simultaneous inference about the existence of links between a binary phenotype and several sections of genome that affect the phenotype as a group in a non-additive fashion, thus referred to as interaction (see Section 2.2). The methods of Part I are aimed at efficiently dealing with the multiple testing problem in a GWAS search for interactions by employing a two-stage testing scheme which proceeds by eliminating non-promising candidates in the first step.

The core results of Part I of the thesis have been submitted for publication in Pecanka et al. (2016).

Part II
In Part II of the thesis, which consists of Chapter 6, we present a novel variant of a cost-efficient two-stage experimental design and analysis procedure designed to be used for a large-scale simultaneous inference problems in a cost-restricted setting. Gains of statistical efficiency are achieved by improved allocation of budget towards the more promising candidates. The design builds up on existing methods published in the literature and improves on a number of these methods especially in application areas with non-homogeneous sparse effects.

The contents of Part II have been submitted for publication in Pecanka and Goeman (2016).

Part III
In Part III of the thesis, namely in Chapters 7 through 9, we present a novel usage of penalized regression for efficient identifying of small groups of genetic loci (markers) within a much larger collection of loci based on common association with a group of numerical mutually correlated phenotypes. The idea is to use the correlation among the phenotype variables to improve the selection process over methods that treat these phenotypes as independent. For instance, such problems arise in application in behavioral data analysis with data generated via a factor model, where related quantitative measurements are used to capture a medical condition of individuals, and the goal is the discovery of links between the underlying condition and genetic loci. Since the quantitative measurements all relate to a single underlying medical condition, they often exhibit correlation and should be treated simultaneously for efficiency purposes. In addition to exploiting the phenotype correlation, via an initial stage assessment of the marginal association of each covariate with the response prior to the actual fitting of the penalized regression model the method prioritizes promising covariates at the expense of the rest.

The contents of Part III are currently being prepared for two publications. The real data
analysis presented in Chapter 9 will be published in the upcoming article by de Menezes et al. (2016). It is also our plan to submit the contents of Chapter 8 for publication as a separate paper in the near future.

1.4 Notational conventions

Throughout this text, we use the following notation conventions. Random variables are denoted by capital letters such as $X, Y, Z$, etc., while their non-random (deterministic) values are denoted by small letters $x, y, z$, etc. Moreover, we differentiate between scalar and multidimensional objects, by using the regular weight font notation for scalars, such as $x, y, z$, $X, Y, Z$, $\mu, \theta, \psi$, etc., while multidimensional objects are generally denoted by bold font, such as random or deterministic vectors and matrices $x, y, z$, $X, Y, Z$, $\mu, \theta, \psi$, etc. For transposition of vectors and matrices we use the "prime" notation as in $a'$, $A'$, $a'$, $A'$, etc. The expectation, variance, covariance operators are denoted by $E$, $\text{var}$, $\text{cov}$, while their estimators are denoted using the "hat" notation $\hat{E}$, $\hat{\text{var}}$, $\hat{\text{cov}}$. Moreover, $I_{[\cdot]}$ denotes the set indicator function, $\|\cdot\|$ is the Euclidean norm of a vector or a matrix (in the latter case it is also called the Frobenius norm), $[\cdot]$ denotes the integer part function. $I$ and $I_n$ denote unit matrices, where the latter also carries explicit information about its dimension $n$, while $0$ denotes all-zero vectors or matrices (of appropriate dimension). Additionally, we use the standard mathematical notation such as $\mathbb{N}$, $\mathbb{R}$ for the sets of all positive integers and all real numbers, respectively, or $\text{diag}(x)$ for the diagonal matrix with vector $x$ as its diagonal.
PART I

TWO-STAGE SEARCH FOR EPISTASIS
A brief introduction to testing for epistasis

For many years now genetics has been haunted by the seeming inability of genome-wide studies to provide explanation of more than a minority of the heritability of most complex traits. Identification of variants linked to many diseases has generally been quite difficult and those variants that have been identified through these studies explain only a small part of the variability in the population. This is referred to as the problem of missing heritability of complex traits (Manolio et al. (2009), Eichler et al. (2010), Gibson (2010)). One of the main candidates for providing explanation of at least part of this missing heritability has been the concept of genetic interactions (Zuk et al. (2012)), also called epistasis, which in this text we use interchangeably. Epistasis was first described by Bateson (1909) and while there are many more or less conflicting definitions (Cordell (2002)), in general terms, epistasis occurs when the combined influence of multiple genetic loci is not just a simple linear addition of influences of individual loci as first described by Fisher (1918). More recent works on this topic include Wade et al. (2001), Cordell et al. (2001), Cordell (2009), Beckers et al. (2009), Mackay et al. (2009), Becker et al. (2011), Van Steen (2012), Hemani et al. (2013), Lehner (2013), Mackay (2014) among others. For a recent overview of the available methods we refer the reader to Niel et al. (2015).

While many rare genetic disorders, such as cystic fibrosis, are influenced by the effects of a single gene (Chial (2008)), common diseases are likely influenced by a collection of genes and their effects simultaneously. Recent research provides persuasive evidence of this and shows that epistasis detected via statistical and computational techniques may be relevant biologically (Ritchie (2011)). In general the number of interacting loci can be high, but in the simplest case epistasis involves two genetic loci. There have been many attempts at modeling and detecting such interaction but the inherent problem of doing that in a genome-wide analysis setting is the extremely strong interaction effect that the loci need to exhibit in order
2.1 Genetic terminology

Recall the case-control setup and assume we have a population $P$ with binary phenotype $V$ and a locus $L$ with three possible genotypes $G_0 = AA$, $G_1 = Aa$ and $G_2 = aa$. In order to for it to be detectable by the standard statistical tools. The main source of loss of power of these tools is the very large number of pairs that need to be investigated, which inevitably results in very low power of the standard tools through what is known as the multiple testing problem. An additional stumbling block of genome-wide searches for interaction is computational complexity of both the classical methods (e.g. likelihood ratio tests) and the more novel methods (e.g. some Bayesian approaches). While these methods are typically quite fast when applied to small problems, in a genome-wide setup they result in an unbearable amount of computational time needed to perform those searches. For these reasons modeling and discovery of three-way or even more complex interactions is largely utopian, however, many novel approaches aimed at reducing the severity of the above mentioned problems for the case of two-way interactions have been proposed over the years (Niel et al. (2015)). These include INTERSNP (Herold et al. (2009)), epiMODE (Tang et al. (2008), EpiGPU (Hemani et al. (2011)), EpiBLASTER (Kam-Thong et al. (2011)), Gini impurity index based decision tree algorithm (Li et al. (2011)), iLOCi (Piriyapongsa et al. (2012)), just to name a few.

In Part I of this thesis we focus on an alternative strategy of tackling the multiple testing and computational problems of genome-wide searches for epistasis. We formulate a two-stage testing procedure, which employs a simple and quick pre-screening of all relevant pairs of loci, henceforth referred to as the **pre-testing phase**. The second step of the procedure, which we refer to as the **post-testing phase**, relies on a more focused and complex testing method, which is applied only to those pairs of loci that had previously passed the pre-testing phase. This is done in a way that provides independence of the two steps, thus enabling us to lower the necessary multiple testing correction and achieve increased statistical power, while simultaneously retaining sufficient control of the type I error rate. In Chapter 3 we describe several variants of such tests and study their asymptotic distributions, while the supporting technical results are postponed until Chapter 5, which is of technical nature and effectively serves as an appendix to Part I. In Chapter 4 we apply the presented theoretical concepts to practical data analysis using both simulated and real data sets. We provide an extensive illustration of the behavior of several variants of the two-stage procedure, where we focus both on error control and power performance under several realistic parameter settings. We show that our two-stage procedures can be quite powerful in comparison to the standard statistical tools in this area and have the potential to enrich the landscape of analysis methods aimed at detecting epistasis.

The core results of Part I of the thesis have been submitted for publication in Pecanka et al. (2016).
characterize $\mathcal{P}$ in terms of the genotypes at locus $L$ we define the concepts of prevalence, penetrance, genotypic relative risk and odds ratio.

**Definition 2.1 (Prevalence, penetrance, genotypic relative risk, odds ratio)** Prevalence $P_V$ of the phenotypic variation $V$ is defined as $P_V = \mathbb{P}(V = 1)$, which is the probability that a random individual from the population $\mathcal{P}$ has phenotype $V = 1$. Penetrance $f_i$ of genotype $G_i$ with respect to phenotype $V$ is defined as $f_i = \mathbb{P}(V = 1 | G_i)$ for $i = 0, 1, 2$. Finally, genotypic relative risks (or risk ratios) of $G_1$ and $G_2$ relative to $G_0$ are then defined as $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$, respectively. Finally, odds ratio of exposure to genotype $G_i$ (or simply odds ratio of $G_i$) is defined as

$$OR_i = \frac{\mathbb{P}(V = 1 | G_i)/\mathbb{P}(V = 1 | \neg G_i)}{\mathbb{P}(V = 0 | G_i)/\mathbb{P}(V = 0 | \neg G_i)},$$

where $\neg$ is the logical negation with $\neg G_i$ meaning "not genotype $G_i".$

**Recessive, dominant, multiplicative and additive models**

A genetic model for locus $L$ is referred to as **recessive** (in a) if the penetrances $f_1$ and $f_0$ are equal, or in other words if $\mathbb{P}(V = 1 | G_0) = \mathbb{P}(V = 1 | G_1)$. In terms of genotypic relative risk it is equivalent to $\lambda_1 = 1$. A **dominant** model (in a) is a situation when $f_1 = f_2$, which is equivalent to $\mathbb{P}(V = 1 | G_1) = \mathbb{P}(V = 1 | G_2)$, and also to $\lambda_1 = \lambda_2$. A **multiplicative** (or log additive) model (in a) is such that $f_1^2 = f_2$, or equivalently $\mathbb{P}(V = 1 | G_1)^2 = \mathbb{P}(V = 1 | G_2)$. In terms of relative risks the multiplicative model requires that $\lambda_1^2 f_0 = \lambda_2$. A model is referred to as **additive** (in a) if $2f_1 = f_0 + f_2$, or equivalently $2\mathbb{P}(V = 1 | G_1) = \mathbb{P}(V = 1 | G_0) + \mathbb{P}(V = 1 | G_2)$, which is the same as requiring that $2\lambda_1 = 1 + \lambda_2$.

**2.2 Genetic interaction (epistasis)**

An important aspect of finding connection between genes and phenotypes is the identification of groups of genes that have a combined effect on the phenotype. In some cases this effect can be simply additive, which is the situation in which each gene contributes to a phenotype independently of other genes. However, in genetics it is suspected that the connection between genes and complex phenotypes (such as a complex disease) is non-linear and the influence of one gene on a phenotype depends on a number of other genes. If that is the case we talk of **genetic interaction** or epistasis.

For two loci $L_1$ and $L_2$ we denote their marginal genotypic values by $G_{L_1}$ and $G_{L_2}$, and the combined genotypic value by $G_{L_1L_2}$. The former two are the marginal contributions of the loci to the phenotype, while the latter is their joint contribution. If the effect of the marginal genotypic values $G_{L_1}$ and $G_{L_2}$ is additive we can write

$$G_{L_1L_2} = G_{L_1} + G_{L_2}.$$
2.2 Genetic interaction (epistasis)

If, however, it happens that for at least a single combination of possible genotypes at the two loci $L_1$ and $L_2$ it holds that their combined value is not the sum of the two marginal genotypic values we need to add a third term into the equality above to account for this extra deviation from additive combination. We obtain

$$G_{L_1,L_2} = G_{L_1} + G_{L_2} + I_{L_1L_2},$$

where the term $I_{L_1L_2}$ is called interaction deviation or epistatic deviation. This concept of interaction naturally generalizes to groups of more than just two loci and for the collection of $n$ loci $L_1, \ldots, L_n$ the combined genotypic value can be expressed as

$$G_{L_1,\ldots,L_n} = \sum_{i=1}^{n} G_{L_i} + I_{L_1\ldots L_n}.$$

For a further general discussion of the definition of genetic interactions see for example Wade et al. (2001) or Phillips (2008).

### 2.2.1 Statistical versus biological interactions

As we said, genetic interaction, or epistasis, occurs when the expression of one section of the genome depends on the genetic code at another location within the individual’s genome. These locations can be well defined collections such as genes, or arbitrary collections of one or more genetic loci. In an ideal situation an investigator would strive to discover the causal biological or mechanistical processes that underlie such dependence, henceforth referred to as biological epistasis (Van Steen (2012)). Such investigation is usually performed via an appropriate data generating experiment and the outcome is evaluated using statistical modeling techniques. If a statistical model yields evidence for interaction among genetic loci, such interaction should be called statistical epistasis. It is important to keep in mind that statistical epistasis and biological epistasis are not equivalent. In general, there is not a precise correspondence between models of biological epistasis and statistical epistasis (Cordell (2002)). It has been show that specific data pattern and statistical model can typically be derived using vastly different underlying biological models of disease development under epistasis (Thompson (1991), Cordell et al. (2001)). Furthermore, even if the biological and statistical models agree precisely, due to possibly strong dependence among genetic loci (such as linkage disequilibrium, see Section 2.3) there is no guarantee that the loci identified to be in statistical epistasis are actually causally linked with the phenotype. It is possible that loci strongly dependent with the causal loci involved in biological epistasis can explain the observed data equally well as the causal loci without being involved in biological epistasis at all. Since the actually causal and biologically interacting loci might not even be included in the experiment in the first place, it is important to replicate statistical findings by independent experiments and subsequent laboratory testing.
We wish to stress that the problem of identifying biological epistasis is not our aim in this thesis. Instead, we see the concept of epistasis as dependence in the statistical sense and we focus on the statistical modeling of interaction.

### 2.2.2 Logistic regression model of epistasis

Statistical tests of epistasis focus on precisely formulated hypotheses concerning well-defined parameters of statistical models. While there are many mathematical models of epistasis (Risch (1990), Neuman and Rice (1992), Risch et al. (1993), Park and Hastie (2008)), we focus in detail on two particular models of epistasis, namely the logistic regression model and the concept of linkage disequilibrium.

Suppose we have a population \( P \) and a collection of genetic loci \( L_1, \ldots, L_n \) and we are interested in investigating the interaction between the genotypes at loci \( L_1, \ldots, L_n \) with respect to a binary phenotypic variation (phenotype) \( \Delta \) with values \( 0 \) and \( 1 \) for every member of a selected population \( P \). For instance, such binary status variable \( \Delta \) can represent affection/non-affection by a disease or possession/non-possession of specific physical trait such as hair color. Based on \( \Delta \) we can divide the population \( P \) into two disjoint sub-populations denoted by \( P_0 \) and \( P_1 \), where \( P_0 \) represents the controls with \( \Delta = 0 \), while \( P_1 \) are the cases with \( \Delta = 1 \). For a randomly selected person \( P \) from population \( P \) the two events \( \{ P \in P_i \} \) and \( \{ P \text{ has phenotype } \Delta = i \} \) are therefore equivalent for both \( i = 0, 1 \). A study design that can be used to perform the investigation of interaction between loci \( L_1, \ldots, L_n \) is the case-control study, in which we collect a sample of both phenotype and genotype data from each of the two sub-populations \( P_0 \) and \( P_1 \) (Lewis and Knight (2012)). Despite the retrospective sampling scheme that is used to collect such data, where individuals are sampled based on the value of their phenotype, the resulting regression coefficients consistently estimate the associated log odds ratios when applied to case-control data (Piegorsch et al. (1994)).

As we said above the simplest case of genetic interaction involves only two loci \( L_1 \) and \( L_2 \), which is what we focus on in this thesis. To model the interaction between \( L_1 \) and \( L_2 \) with respect to a binary phenotype \( \Delta \) within the case-control design we represent the genotypes at the two loci numerically by variables \( X \) and \( Y \), respectively. Typically, \( X \) and \( Y \) count the number of minor alleles at each locus, where "minor" is defined by it being the rarer one in the population. We then define the logistic regression model of genetic interaction (LRM) by the relationship

\[
P(\Delta = 1 | X, Y) = \left(1 + \exp(-\beta_0 - \beta_1 X - \beta_2 Y - \beta_3 XY)\right)^{-1}.
\]  

(2.1)

 Needless to say, the LRM is a very popular statistical model that has been widely applied to the problem of modeling two-way interaction of genetic loci (for example Cordell (2002), Marchini et al. (2005), Park and Hastie (2008)). The parameter \( \beta_0 \) is called the penetrance, \( \beta_1, \beta_2 \) are referred to as main effects (also known as marginal effects) and \( \beta_3 \) represents the interaction effect. Since we are interested in detecting the presence of interaction, our main
focus is towards the value of $\beta_3$. A strong point of the model above is that additional covariates can be added to it, albeit at computational speed cost. For instance, population structure can be accounted for in the model by adding ancestry information covariates, which is important in genetic association studies that are susceptible to confounding due to population structure (Lewinger et al. (2013)).

In order to fit this model we must estimate the values of $\beta_0, \beta_1, \beta_2, \beta_3$. Typically, maximum likelihood estimators (MLEs) are used. The actual estimates are usually found by the method of \textit{iteratively reweighted least squares} (Holland and Welsch (1977), Marx (1996)). With this method the main effect estimates should accurately approximate the actual values of these parameters in the population, assuming the model is appropriate. However, the estimate of the penetrance $\beta_0$ in a case-control study is effectively determined by the sample size ratio of cases and controls and not the actual population penetrance of the phenotype $\Delta$ within $P$. For instance, if we assumed that $\beta_1 = \beta_2 = \beta_3 = 0$ then (3.1) yields the estimator $\hat{\beta}_0 = -\log(N/M)$, where $N$ and $M$ are the respective sample sizes of controls and cases. Therefore, it is not wise to try to extract much meaning from the value of $\hat{\beta}_0$.

Notice that in the formula above we chose $XY$ to be the interaction variable, which is referred to as the \textit{multiplicative interaction model}. While the choice of $XY$ is a popular one, there are certainly other options to combine the numerical representations of genotypes $X$ and $Y$ in order to model interactions. This is addressed in more detail in Section 3.1, but for the moment the current choice suffices. In any case, under the LRM we say that loci $L_1$ and $L_2$ do not interact if $\beta_3 = 0$. Whether $\beta_3 = 0$ is true or not can be tested using the usual statistics such as the \textit{likelihood ratio} or the \textit{score} statistics, which are asymptotically equivalent. Similarly to Zhang (2006), in the context of LRM within the case-control model we choose to focus on the score statistic, because of its relative computational simplicity that is highly desirable in a genome-wide analysis, where the test is applied for a large number number of loci pairs. In our two-stage testing procedure, the score statistic provides the basis for the post-test.

2.3 Linkage disequilibrium

The concept of \textit{linkage disequilibrium} (LD) between two or more genetic loci is of central importance to genetics. In general, LD can be defined in probabilistic terms as the \textit{non-random association} of alleles at two or more loci located on the same or on different chromosomes, where by non-random association we mean stochastic dependence between occurrences of genotypes at the two loci in the population. Contrariwise, two or more loci are said to be in \textit{linkage equilibrium} (LE) if there is no non-random association between them. In technical terms, for two haplotypes of loci $L_1$ and $L_2$ with alleles $(A/a)$ and $(B/b)$, respectively, denote as $p_{AB}$ the frequency of haplotype $AB$, and as $p_A, p_B$ the frequencies of individual alleles $A$ and $B$ in the population, respectively. LD is measured as the dependence of these haplotypes
by $D_{AB} = p_{AB} - p_A p_B$ and the loci $L_1$ and $L_2$ are said to be in LD if $D_{AB} \neq 0$.

LD is a fundamental concept to gene mapping of complex disease genes in genome-wide association studies (GWAS) and also to discovering about evolution of populations. LD between two loci arises from intermixture of populations with different gene frequencies, but it can also arise by chance in small populations. Usually, LD is stronger for (but not exclusive to) loci that are linked, where by linkage it is meant a tendency of alleles to be inherited together by an offspring during meiosis. Under the effect of Mendelian law of segregation in a randomly mating population (Falconer and Mackay (1996)) without additional evolutionary forces (such as mutation) there exists a tendency for loci to approach such a random association or equilibrium. However, the rate at which it approaches such equilibrium is reduced by linkage between the loci, and hence linkage is said to generate a disequilibrium (King et al. (2006)). Although linkage is more likely to occur for loci that are close together on the original chromosome, since they are more likely to be inherited together by an offspring, it is not exclusive to them.

### 2.3.1 Linkage disequilibrium and epistasis

Of vital importance for us and our two-stage testing procedure to search for epistasis is that LD can be generated by an evolutionary process as a consequence of genetic interaction (Phillips (2008)). The link between LD at loci $L_1$ and $L_2$ and genetic interaction between $L_1$ and $L_2$ with respect to a given phenotype stems from the following genetic argument (Jin and Xiong (2008)):

1. If two loci $L_1$ and $L_2$ are unlinked, as would be expected for example for loci on different chromosomes, they should be in LE in the general population, but also among both cases and controls considered separately, unless the phenotype status is associated with the two loci in some way.
2. Simply put, interaction is a situation in which the effect of $L_1$ on the phenotype depends on $L_2$, and vice-versa.
3. In the diseased population (cases) this should lead to (high) correlation between genotypes at $L_1$ and $L_2$ with respect to the phenotype (disease indicator). Consequently, there should be (high) LD between the two loci among the cases.
4. Furthermore, if there is LD among the cases and the general population is in LE, then also the control population should be in LD.

Consequently, we can view LD among cases or LD among controls as an indication of interaction between the two loci. We can take the argument a bit further still:

5. If the general population is in LE, the direction of deviation from LE should be different between cases and controls. Moreover, if the prevalence of the case phenotype is different from 0.5, also the degree of deviation should be different between the two subpopulations.
6. Then, difference of LD between cases and controls can also be indicative of interaction between the two loci.

7. Moreover, if the proportion of the two sample sizes of cases and controls does not reflect the true phenotype prevalence, we can use LD within the pooled sample as indication of interaction as well.

The above argument provides several indirect ways to detect presence of interactions by looking at LD patterns in the case-control sample. On the other hand, it should be pointed out that LD between two loci is not necessarily a consequence of epistatic selection (see Wade et al. (2001)) and the above argument does not suggest that LD is the same thing as interaction. It merely says that the presence of LD within a phenotype based subpopulation might be a consequence of interaction. For a detailed discussion of LD a consequence of gene-gene interactions we shall refer to for example Wade et al. (2001) and Zhao et al. (2006).

For rare diseases we expect stronger LD within the cases, but it should also be induced among the controls, which means that we can use either population to detect interaction between two loci by looking at LD. Using the controls makes sense especially when the population prevalence of cases is high, while for low prevalence phenotypes it is turns out to be beneficial to rely on the cases instead. Alternatively, for rare diseases, the amount and/or the sign of correlation induced by interaction between two loci should be different between the two populations, thus allowing us to detect interactions by looking at the difference of LD measures between the two groups or within the pooled sample of cases and controls. In Chapter 4 we investigate these aspects more closely for several two-stage methods formulated in the next chapter, which use the case-only sample, or the control-only sample, or the pooled sample to detect interaction.

**Definition 2.2 (Interactions as difference of LD)** Denote by $D_{CA}^{cA}$ and $D_{CA}^{co}$ the LD measures in the subpopulation of cases and controls, respectively. We define the LD difference parameter as $D_{CA}^\delta = D_{CA}^{cA} - D_{CA}^{co}$. It is said that loci $L_1$ and $L_2$ do not interact under the LD model of interactions if $D_{CA}^\delta = 0$. We denote this as hypothesis $H_{\delta}^0$ and the complementary hypothesis as $H_{\delta}^1$.

A statistic that can be used to test the hypothesis $H_{\delta}^0$ is $T_{LD}^\delta = (\hat{D}_{CA}^\delta)^2 / \hat{\text{var}} \hat{D}_{CA}$. Where $\hat{D}_{CA}$ is computed by plugging estimated probabilities $\hat{p}_{AB}, \hat{p}_{A}, \hat{p}_{B}$ in the place of the unknown true probabilities into $D_{CA}^{\delta}$. For short we call $\hat{D}_{CA}$ the sample version of $D_{CA}^{\delta}$. As Zhao et al. (2006) showed the statistic is asymptotically $\chi^2_1$ distributed under $H_{\delta}^0$. In a typical population not all LD is caused by interaction of the two loci, there is often so called background LD. An advantage of the test above is that it is not sensitive to the presence of such population background LD, which refers to the LD not caused by interaction. Such background LD is filtered out when $D_{CA}^{\delta}$ is used, since the background LD should be approximately equal between the two subpopulations.

On the other hand, an undesirable aspect of $T_{LD}^\delta$ is that it requires sufficiently large samples from both subpopulations of cases and controls. This makes it quite costly, because in a
typical case-control study there tends to be (much) fewer cases than controls, since controls 
(e.g. non-diseased people) are typically easier to come by. Therefore, a procedure that 
detects LD using only the sample of controls would be highly beneficial from this point of view. 
Such statistic is

\[ T_{LD}^{co} = \frac{(\hat{D}_{AB}^{co})^2}{\hat{\text{var}} \hat{D}_{AB}^{co}}, \]

which only uses the data from the sample of controls. Due to the possible presence of back-
ground LD it might not be always advisable to solely rely on a procedure that looks at controls 
only when searching for interaction. However, if this procedure is used as an initial screen 
in conjunction with additional test in a two-stage procedure, it could yield substantial power 
 Improvements, especially in highly imbalanced case-control designs with a large excess of 
controls and high population prevalence of cases. On the other hand, for situations with low 
population prevalence of cases a case-based alternative of \( T_{LD}^{co} \) can be used.

### 2.3.2 Dependence measure based on genotypes instead of haplotypes

The need to estimate haplotype frequencies in order to calculate the above described LD 
measures presents a practical problem. Biological experiments for generation of haplotype 
data are expensive and time consuming (Jin and Xiong (2008)), while statistical estimation 
methods can be tricky and computationally expensive for certain haplotype frequencies. For 
example in the case of a two biallelic loci such as SNPs of a human who is heterozygous at 
both loci one cannot easily determine the haplotype from a genotype alone, as we illustrate 
in this section. Jin and Xiong (2008) also points out that errors incurred during estimation of 
 haplotypes from genotype data tend to lead to increased false detection of interaction, which 
is highly undesirable.

Suppose we have observed genetic information at loci \( L_1 \) and \( L_2 \) in a sample of indi-
viduals from our population and denote the alleles at locus \( L_1 \) by \( A \) and \( a \) and the alleles at 
locus \( L_2 \) by \( B \) and \( b \). In a case-control setting (without additional family data), in order to 
estimate haplotype frequencies using the observed genotypes, we can utilize the following 
genotype-haplotype conversion table

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Haplotype</th>
<th>Genotype</th>
<th>Haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA ( \times ) BB</td>
<td>( AB \star AB )</td>
<td>aa ( \times ) Bb</td>
<td>aB ( \star ab )</td>
</tr>
<tr>
<td>Aa ( \times ) BB</td>
<td>( AB \star aB )</td>
<td>AA ( \times ) bb</td>
<td>Ab ( \star Ab )</td>
</tr>
<tr>
<td>aa ( \times ) BB</td>
<td>( aB \star aB )</td>
<td>Aa ( \times ) bb</td>
<td>Ab ( \star ab )</td>
</tr>
<tr>
<td>AA ( \times ) Bb</td>
<td>( AB \star Ab )</td>
<td>aa ( \times ) bb</td>
<td>ab ( \star ab )</td>
</tr>
</tbody>
</table>

Above we omitted the genotype pair \( Aa \times Bb \), which corresponds to two haplotypes, namely 
\( AB \star ab \) and \( Ab \star aB \). In order to determine haplotype frequencies for such genotype additional 
estimation of the proportion of the two haplotypes in the population is required. An overview 
of available methods for this can be found in Browning and Browning (2011), where methods 
based on the EM-algorithm, hidden Markov chains, identity-by-descend, etc. are discussed.
Since the problem at hand is often complex, many of these methods run into difficulty in terms of estimation precision, but perhaps even more importantly computational complexity. For example, the popular EM-algorithm is computationally quite demanding since it is iterative.

### 2.3.3 Testing dependence of genotypes

To avoid the problem of estimating haplotype frequencies we can focus on the non-random association of genotypes instead of non-random association of haplotypes as was the case with LD. Analysis of non-random association of genotypes can be a suitable proxy for LD as argued by Weir (1979) or more recently by Wellek and Ziegler (2009). Therefore, in our two-stage testing procedure we use a two-locus genotype independence test during pre-testing which is followed by a LRM score test based test in the second step.

We present the details of such construct in the next chapter, but now we point out that it brings about at least two advantages. Firstly, shifting the focus from haplotype to genotype frequencies simplifies the estimation of parameters. This removes a level of complexity in terms of both estimation precision and computational difficulty. Secondly, the use of genotype frequencies allows for a convenient joint treatment of the pre-test and the post-test score statistic in terms of their (asymptotic) distributions and independence. The investigation of these properties is the focus of Chapter 3.

As suggested by the list of methods mentioned at the start of this chapter, there have been a number of attempts at devising powerful tools to search for epistasis through two-stage testing, some of which are quite similar to the theory presented in the next three chapters. For example, two-locus tests of independence have already been used as pre-tests in a two-stage setup by Herold et al. (2009) and Lewinger et al. (2013), where the approach of the latter is discussed in more detail in Section 3.2.1. It is interesting to compare the method by Herold et al. (2009) with our approach further developed in the next chapter. To some degree the two methods are somewhat similar, however, a major difference between the approach of Herold et al. (2009) and our methods lies with the multiple testing correction factors. Crucially, from a theoretical perspective the method Herold et al. (2009) requires much larger multiple testing correction factors by all tests involved in the analysis, which we perceive as the main shortcoming of their approach, as the reduction of multiple testing correction appears to be where all the potential power gains lie.

### 2.4 Multiple testing problem

The large number of genetic loci involved in a GWAS leads to a severe multiple testing problem that becomes even more pronounced in when searching for interaction of multiple loci, when the number of tests needed to be performed grows polynomially with the number of loci of interest. For instance, if we focus on two-way interactions between pairs of \( L \) loci,
Comparison of Sidak and Bonferroni corrections

Figure 2.1: Ratio of $\alpha_S = 1 - (1 - \alpha^p)^{1/K}$ and $\alpha_B = \alpha^p / K$ as function of the number of tests $K$ for $\alpha = 0.05$. The dashed line represents the limit ratio of $\lim_{K \to \infty} \alpha_S / \alpha_B = -\log(1 - \alpha) / \alpha \approx 1.026$.

the number of all tests $K$ is of the order of $L^2$. Even with only 15000 loci the number of pair-wise tests is already well over 100 million.

In order to retain control of the family-wise error rate (FWER), one usually performs a multiple testing correction (MTC). Classical methods to control FWER in multiple testing contexts are the methods of Bonferroni (Bonferroni (1936), Galambos and Simonelli (1996)), Šidák (Šidák (1967)), Holm (Holm (1979)), Hommel (Hommel (1983), Hommel (1986)), to name just a few. The Šidák method unfortunately requires independence of the parallel tests, which is not completely realistic in our case, while the other listed methods are fully applicable to the considered genetic setting allowing for arbitrary dependence structure among the tests. For the sake of simplicity we formulate the method using the Bonferroni method. We do so fully aware of the fact that the Bonferroni correction can be quite conservative if the number of tests is very large and there are many false hypotheses. Such conservativeness becomes even more severe if there are strong positive correlations among the $p$-values (Goeeman and Solari (2014)). However, the problem of many false null hypotheses should not be a severe complication for the purposes of detecting epistasis in a genome-wide search, where most tested pairs are expected to not be exhibiting interaction.

In a single-stage testing with $K$ tests with FWER desired to be controlled by $\alpha \in (0, 1)$, the Bonferroni method consists of performing each of the $K$ tests at level $\alpha_B = \alpha / K$. Its major advantage is that it provides strong control of FWER under arbitrary dependence of test statistics, where strong control of FWER means that for any set of null hypotheses the probability of having non-zero false positives is less than or equal to $\alpha$ (Holm (1979)). Figure 2.1 shows a comparison of the Bonferroni-corrected levels and the Šidák-corrected, which require independence. The comparison shows that the Bonferroni method is nearly optimal under independence. For the standard study-wide significance level of $\alpha = 5\%$ the ratio of the levels quickly approaches $-\log(1 - \alpha) / \alpha$ with increasing number of tests, which means that difference between the two levels is at most 2.6\%.
Alternatively, instead of controlling FWER, one can choose to control the *false discovery rate* (FDR) ([Benjamini and Hochberg (1995)](https://www.jstor.org/stable/2337909?origin=crossref), [Benjamini and Yekutieli (2001)](https://www.jstor.org/stable/1394159?origin=crossref)) and control the expected proportion of incorrectly rejected null hypotheses (so called "false discoveries"). Although compared with FWER the control of FDR leads to higher power, it is not straightforward to make claims about the truthfulness of any single discovery. For this reason we focus on FWER throughout Part I.

### 2.5 Examples of two-stage testing methods

The notion of two-stage testing is largely motivated by the expectation that in a multiple testing scenario a cleverly constructed two-stage test can be more powerful than a simple Bonferroni corrected single-stage test. As the idea itself is not new, there already exist quite a few examples of two-stage methods that illustrate the potential of such approach. In here we shall only focus on three examples of two-stage testing that are especially relevant to the methods we present in Part I of this thesis. The methods come from Zheng et al. (2006), Murcray et al. (2008) and Lewinger et al. (2013). First of these examples does not focus on genetic interactions, but it illustrates the need for independence between the steps of a two-stage testing procedure. On the other hand, the other two examples by Murcray et al. (2008) and Lewinger et al. (2013) focus on detecting interaction between gene and environment and between two genes. Since the method by Murcray et al. (2008) and especially that by Lewinger et al. (2013) use a very similar design to the two-stage method we present in the next chapter, their success shows the practical potential of the notion of two-stage testing.

The first example we present here comes from Zheng et al. (2006), who describes a self-replicating two-stage test for testing Hardy-Weinberg equilibrium (HWE) at a single locus. In their approach the pre-test is based on the Hardy-Weinberg disequilibrium trend test (HWDTT) ([Song and Elston (2006)](https://www.jstor.org/stable/2091356?origin=crossref)), while the post-test is the Cochran-Armitage trend test (CATT) ([Sasieni (1997)](https://www.jstor.org/stable/2337909?origin=crossref), [Freidlin et al. (2002)](https://www.jstor.org/stable/1394159?origin=crossref)). Zheng et al. (2006) prove that the pre-test and post-test statistics are independent, but only under the combined null hypothesis of HWE. This property motivates them to screen all $M$ candidate pairs by the HWDTT pre-test on a selected pre-test significance level $\alpha_1$ and subsequently perform MTC for only $\alpha_1 M$ tests in the CATT post-test. Essentially, they show that their two-stage procedure is more powerful than the corresponding single stage analysis providing additional support to the potential of the notion of two-stage testing in GWAS studies. However, the fact that the independence of the two-test requires that the combined null of both tests holds, which means that they must avoid taking extremely small values of $\alpha_1$. This arguably somewhat limits the usability of their approach in our opinion, since there is not too much space for lowering the MTC of the post-test in their method while retaining sufficient error control. In our two-stage method described in the next chapter, in order to remove such limitation, we specifically focus on achieving independence regardless of whether the pre-test null hypothesis holds or not.
The other two attempts at improving power performance through two-stage testing come from Murcray et al. (2008) and Lewinger et al. (2013). In the approach by Murcray et al. (2008) the authors focus on two-stage detection of gene-environment interactions in a multiple testing scenario of GWAS. The screening test (pre-test) of Murcray et al. (2008) is based on a case-only LRM, which models the environment variable on the dependent side, while genetic information plays the role of the independent variable. Using such model, they perform the likelihood ratio test of non-nulity of the independent variable coefficient using no multiple testing correction. The screening phase is subsequently followed by a case-control based likelihood ratio test within a LRM of the same form as (2.1) except with \( X \) and \( Y \) replaced by gene and environment variables \( G \) and \( E \). Unlike the screening phase, the second step employs the Bonferroni multiple testing correction for the number of tests in the second phase. While the argument by Murcray et al. (2008) for validity of such procedure in terms of type I error control based on asymptotic independence of the two tests is only approximate, they illustrate the power potential of such procedure using a simulation study, the results of which are quite promising.

On the other hand, the two-stage method by Lewinger et al. (2013) is specifically targeted at gene-gene interactions of two loci. The screening phase of their two-stage method is based on a pooled case-control sample based independence test, while the second phase is based on the LRM of (2.1) applied only to the pairs of loci that show signs of dependence as judged by the screening phase. In order to account for false rejections, Lewinger et al. (2013) use the same Bonferroni multiple testing correction for the number of tests in the second phase. They also illustrate the behavior of the method using simulation, which shows, under the considered modeling and parameter choices, that the power performance superiority of the two-stage method over the classical single-stage LRM based approach. The design by Lewinger et al. (2013) is similar to our method in that it uses the same two-stage setup which pre-tests for dependence of two loci. However, the use of the pooled sample of cases and controls is different from our approach since we focus on the case-only and control-only pre-tests. This seemingly subtle difference has important consequences with respect to independence of the two steps and with respect to power performance. As far as independence is concerned, using the pooled sample leads to asymptotically independent tests automatically, whereas in the control-only tests we must work for it. In Part I of the thesis we formulate theoretical results for both approaches.

### 2.6 Our goal in Part I of the thesis

At this point we provided practical examples of two-stage procedures, which serve as illustration of the potential of superior power performance over their corresponding single-stage counterparts. In the next three chapters we devise several related two-stage methods aimed at detecting interaction in GWAS setting. We formulate theoretical results concerning error
2.6 Our goal in Part I of the thesis

control and power of the procedures and the statistics on which they are based. This is done mostly in Chapter 3, which contains design of the methods and formulation of pivotal results. As the actual power performance of these methods is hugely dependent on properties of the data on the one hand, and modeling and tuning parameter choices made during analysis on the other, for practical usability it is important to investigate the behavior of the methods under many different scenarios. We present the results of such investigation in Chapter 4, where we focus both on power performance and error control, and also on the behavior of these methods for different choices of tuning parameter optimization. As far as the last of these is concerned, we also formulate results that can be used to make choices for these tuning parameters and thus allowing the user to fully exploit the power of the methods in question.
Two-stage testing for epistasis: Theory

In this chapter we present the theoretical results concerning several novel two-stage testing procedures applicable within the case-control design. Commonly, the first phase of these procedures screen all candidate pairs of loci for dependence of genotypes using a chisquare-type test. We call this phase the pre-test. It is followed by the second phase, which call the post-test, where tests of presence of interactions within the logistic regression model (LRM) are performed. Crucially, the post-tests are only applied for the pairs of loci that had their genotype independence rejected in the first stage.

In order to ease the flow of the text all of the supporting and technical results together with the proofs of the main theorems of this chapter are postponed until Chapter 5, which serves as an appendix to Part I. The division of this chapter is as follows. We start this chapter by presenting a formal definition of the LRM and the score test statistic on which our two-stage test for interactions is based (Section 3.1). This is followed by a formulation of several pre-tests in Section 3.2. The first option for a pre-test is based on the pooled sample of both cases and controls, while the rest of the presented tests are based on single population samples of only controls or only cases. We then proceed to describing how these pre-tests can be integrated with the post-test score statistic in a way that is statistically sound in terms of type I error control. First we formulate an asymptotic distribution and a pre-test independence result for the pooled-sample test in Section 3.3. Throughout this text we define asymptotic independence to be the situation in which two random vectors jointly converge in distribution to a limit whose distribution coincides with the product of its marginals (see Definition 3.2 in Appendix A). In the subsequent two sections we turn the attention to the control-only and case-only pre-tests. In Section 3.4 we describe a simple yet surprisingly powerful approach to achieving independence of the two-stages in the two-stage testing testing procedure. The approach of Section 3.4 is based on sample splitting and using one of the two resulting disjoint (and therefore independent) parts in the pre-testing phase, while utilizing the other disjoint part for the post-test. This general idea of sample splitting can be easily applied to any pair of
3.1 Interactions in logistic regression model

In general, we say that binary phenotype indicator variable $\Delta$ and genotype representation variables $X, Y$ follow the logistic regression model (LRM or LR model) with interactions if

$$P(\Delta = 1 \mid X, Y) = \left[1 + \exp(-\beta'z(X,Y))\right]^{-1} = \Psi(\beta'z(X,Y)),$$

(3.1)

for some constant $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ and a function $z(x,y) = (1, x, y, z(x,y))'$, where $z(x,y)$ is a non-constant and nonlinear function of $x$ and $y$ and where we denoted the logistic function $\Psi(x) = 1/(1 + e^{-x})$. By non-constant function $z(x,y)$ we understand a function that is not constant in at least one of the two variables and the non-linear requirement is explicitly stated because if $z$ is linear the interaction terms effectively drops out. This general notation with function $z(x,y)$ allows for simultaneous treatment of different forms of interaction. A specific interaction model is obtained by giving the function $z$ a specific form which is necessary for application. As we already mentioned in Section 2.2.2 the usual choice is $z(x,y) = xy$ or $z(x,y) = \min(1,xy)$. The null hypothesis of interest in the logistic regression model is the hypothesis that there are no interactions between loci $L_1$ and $L_2$. In model (3.1) this corresponds to the coefficient $\beta_3$ being equal to zero, thus we want to test the null hypothesis of no interactions $H_0^\Delta : \beta_3 = 0$ against the alternative $H_1^\Delta : \beta_3 \neq 0$. Consequently, under $H_0^\Delta$ it holds $\beta = \beta^0$, where $\beta^0 = (\beta_0, \beta_1, \beta_2, 0)'$ with $(\beta_0, \beta_1, \beta_2)' \in \mathbb{R}^3$. In order to test $H_0^\Delta$ the three nuisance parameters $\beta_0, \beta_1, \beta_2$ need to be estimated. This presents a theoretical challenge when dealing with distribution properties of such test statistic. In Theorem 3.1 we

1Note the difference in the notation between the 4-dimensional function $z(x,y)$ (bold font) and its last component $z(x,y)$ (regular font), which agrees with our convention of denoting multidimensional objects with bold font.
formulate a result that links the score statistic with estimated nuisance parameters and the corresponding score statistic in which the values of these parameters are known, which is convenient for proving the asymptotic results of this chapter.

The data set modelled by (3.1) consists of an independent sample of \( n \) triplets \((\Delta_i, X_i, Y_i)\), \( i = 1, \ldots, n \) and we refer to it as a case-control data set. We denote the number of individuals controls in the sample by \( N \) and the number of cases by \( M \), which means \( n = N + M \). Throughout this text we assume that each of the two subsamples consists of independent and identically distributed individuals. Moreover, we also assume that the individuals in the full sample of size \( n \) are independent, although naturally the full sample is not assumed to be identically distributed.

**Interpretation of parameters**

In Definition 2.1 we defined odds ratios of exposure with respect to genotypes \( G_0, G_1, G_2 \), which in the context of the logistic regression model are represented by variables \( X \) and \( Y \). The exposure odds ratio with respect the variables \( X \) or \( Y \) is the ratio of odds of being affected (\( \Delta = 1 \)) if, ceteris paribus, the value of \( X \) or \( Y \) is increased by 1. Under the LRM it conveniently turns out that the logarithms of odds ratios of the two main effects are equal to the coefficients \( \beta_1 \) and \( \beta_2 \), respectively. It is natural to view the interaction effect size parameter \( \beta_3 \) in the same way and define the log odds ratio associated with the interaction terms as \( \beta_3 \). Odds ratios therefore provide natural measures of risk associated with marginal increase of the number of alleles at a given locus and we use this terminology as the primary measure of effect sizes in Chapter 4.

**Case-control vs random sample design**

Unlike the classical random sample design (i.e. independent and identically distributed (iid) individuals), the case-control design consists of two random samples from each subpopulation with a fixed sample size ratio between them. However, from the perspective of inference about the parameters of the logistic regression model this difference between the two designs is of little importance. In fact, it can be shown (see for example Section 14.5 in van der Vaart (2015)) that coefficients of all non-trivial regressors in the logistic regression model are identifiable under both designs and (profile) likelihood functions for the two designs are proportional for these coefficients. It is only the penetrance parameter \( \beta_0 \) that is not estimable under the case-control design, which, however, is irrelevant for our purposes. Therefore, we can use this "near-equivalence" of the two designs throughout this chapter and Chapter 5 and conveniently prove the theoretical results using the iid design.

**Genetic explanatory variables**

In the context of genetics, the variable \( \Delta \) is a binary phenotype and the equation (3.1) links the probability of being a case (\( \Delta = 1 \)) to explanatory variables \( X \) and \( Y \), which represent genotypes at genetic loci \( L_1 \) and \( L_2 \). The variables \( X \) and \( Y \) are assumed to respectively take
3.1 Interactions in logistic regression model

finitely many values from sets $G_X$ and $G_Y$, which depend on the selected genotype model. For instance, if the two loci are biallelic with unordered genotypes $AA, Aa, aa$, the variables $X$ and $Y$ usually count the selected allele at each locus, usually the minor allele, i.e., the rarer one. Assuming $a$ is the minor allele, the counting is done according to the model assumed for the given loci and the values $X$ and $Y$ would be $0, 1, 2$ if additive model was assumed for both loci for the three genotypes. Under additive model we get $G_X = G_Y = \{0, 1, 2\}$. Alternatively, the values for $X$ and $Y$ could also be $0, 1, 1$ if a dominant model (in $a$) is assumed or $0, 0, 1$ if recessive model (in $a$) is chosen. In both of these cases it holds $G_X = G_Y = \{0, 1\}$. Naturally, a more general setting can be assumed and instead of these three models one could also use a general three-level factor variables to represent the genotypes.

In this chapter we formulate many of the results under the assumption that the loci $L_1$ and $L_2$ modelled by the logistic regression model (3.1) are biallelic with alleles $A/a$ at $L_1$ and alleles $B/b$ at $L_2$, which allows for less cumbersome formulations and a simpler notation. Thus, the three possible (unordered) genotypes for $L_1$ would be $AA, Aa, aa$, while for $L_2$ they would be $BB, Bb, bb$. Assuming an additive model for the two loci, the genotype variables $X_i, Y_i$ could for instance count the minor allele $a$ at each locus and take values $0, 1, 2$, where $X_i = k$ indicates there are $k = 0, 1, 2$ alleles $a$ at locus $L_1$ and $Y_i = l$ indicates there are $l = 0, 1, 2$ alleles $b$ at locus $L_2$.

Assumption of proportionate sample sizes

Another assumption that we make is of technical nature. In the following we focus on the investigation of asymptotic properties of several statistics based on the random samples of cases and controls, where by "asymptotic" properties we mean the behavior as $n$ goes to infinity. To simplify the treatment we shall assume that $N$ (and $M$) is very nearly constantly proportionate to $n$ for a fixed proportion $\tau \in (0, 1)$. Since $N$ is an integer, we effectively assume that $N$ does not differ from $\tau n$ by more than 1, which put in formula means that $\tau_n = N/n$ satisfies

$$\tau_n = \tau + O(n^{-1}).$$

This assumption makes the asymptotics with $N \to \infty$, $M \to \infty$ and $n \to \infty$ all equivalent.

3.1.1 Score statistic for interactions

We define the notation $\Psi_i = \Psi(\beta'z(X_i, Y_i))$ and $\Psi_i^0 = \Psi(\beta^0'z(X_i, Y_i))$ for the logistic functions of the $i$-th individual at the true parameter $\beta$ and the null hypothesis parameter $\beta^0$, respectively. With this notation, the likelihood function $L$ for one observation $\Delta, X, Y$ within

\footnote{As far as testing for interaction in the logistic regression model is concerned, the fact whether minor or major allele is counter does not actually make any difference with respect to $\beta_3$, as change from one to the other can only affect $\beta_0$, $\beta_1$ and/or $\beta_2$.}
the model (3.1) is equal to
\[L(X, Y, \Delta; \beta) = \left(\Psi(\beta'z(X, Y))\right)^\Delta \left(1 - \Psi(\beta'z(X, Y))\right)^{1-\Delta},\]
and the corresponding log-likelihood is \(\ell(X, Y, \Delta; \beta) = \log L(X, Y, \Delta; \beta)\). The joint likelihood function for a sample of \(n\) triplets \((\Delta_i, X_i, Y_i), i = 1, \ldots, n\), is equal to
\[L_n(X, Y, \Delta; \beta) = \prod_{i=1}^n L(X_i, Y_i, \Delta_i; \beta),\]
where \(\Delta = (\Delta_1, \ldots, \Delta_n)'\), \(X = (X_1, \ldots, X_n)'\) and \(Y = (Y_1, \ldots, Y_n)'\). The joint log-likelihood function \(\ell_n\) is
\[\ell_n(X, Y, \Delta; \beta) = \sum_{i=1}^n (\Delta_i - \Psi_i^0) z(X_i, Y_i).\]

By differentiation of \(\ell_n\) with respect to \(\beta\) we get \(\dot{\ell}_n(X, Y, \Delta; \beta) = \sum_{i=1}^n (\Delta_i - \Psi_i^0) z(X_i, Y_i)\), where we used the relationship between \(\Psi(x)\) and its derivative \(\Psi'(x) = e^{-x}/(1+e^{-x})^2\), which reads \(\Psi'(x) = \Psi(x)(1 - \Psi(x))\). To test \(H_0^c\), we standardize the score \(\dot{\ell}_n\) by \(n^{-1/2}\) and define the vector score statistic with known parameters as
\[S_n = n^{-1/2} \sum_{i=1}^n (\Delta_i - \Psi_i^0) z(X_i, Y_i).\]

In practice, the statistic above usually cannot be calculated because the true values \(\beta_0, \beta_1, \beta_2\) are unknown. That means that we usually need to replace \(\beta^0\) by its null hypothesis maximum likelihood (ML) estimator \(\widehat{\beta}^0 = (\widehat{\beta}_0^0, \widehat{\beta}_1^0, \widehat{\beta}_2^0)'\). Denoting \(\Psi_i^0 = \Psi(\widehat{\beta}^0_i z(X_i, Y_i))\), we get the vector score statistic with estimated parameters
\[\widehat{S}_n = n^{-1/2} \sum_{i=1}^n (\Delta_i - \widehat{\Psi}_i^0) z(X_i, Y_i).\]

The test of \(H_0^c\) is then performed using only the fourth coordinate of \(\widehat{S}_n\), which of course needs to be standardized using a suitable variance estimator.

**Variance of score statistic**

It is a classical result, which follows directly from the definition of the score statistic, that variance matrix of \(S_n\) is the Fisher information matrix \(I_{\beta} = \mathbb{E} \dot{\ell}(X, Y, \Delta; \beta) \dot{\ell}'(X, Y, \Delta; \beta)\). Using regularity\(^3\) of the logistic regression model, we can also write \(I_{\beta} = -\mathbb{E} \ddot{\ell}(X, Y, \Delta; \beta)\), where \(\ddot{\ell}(X, Y, \Delta; \beta) = (\partial^2/\partial \beta \partial \beta') \ell(X, Y, \Delta; \beta)\) denotes the Hessian matrix of \(\ell\). Calculating the Hessian matrix leads to
\[I_{\beta} = -\mathbb{E} \left[\Psi(\beta'z(X, Y))(1 - \Psi(\beta'z(X, Y))) z(X, Y) z'(X, Y)\right].\]

From the law of large numbers and the continuous mapping theorem it follows that under \(H_0^c\) we can consistently estimate \(I_{\beta}\) by \(\hat{I}_{\beta} = -n^{-1} \sum_{i=1}^n \ddot{\ell}(X_i, Y_i, \Delta_i; \beta^0)\). In terms of power using the null hypothesis estimator of the variance has only small effect, since \(I_{\beta}\) affects only the second or higher order terms in the corresponding power function (see Section 5.1).

\(^3\)Regularity conditions are discussed in Section 5.1 and specifically formulated in Theorem 5.1.
An added benefit of using the null hypothesis estimator is based on \( \hat{\beta}^0 \), which we must anyway estimate to calculate \( \hat{S}_n \). Estimating \( I_\hat{\beta} \) without assuming \( H_0^e \) requires obtaining the unrestricted ML estimator of \( \hat{\beta} \), which would add significantly to the already substantial computational burden.

### Asymptotic representation of score statistic

In (3.3) and (3.4) we defined the logistic regression model score statistics with true and estimated values of the parameters \( S_n \) and \( \hat{S}_n \). For practical purposes the latter of the two is of larger importance, since the true values of \( \beta_0, \beta_1, \beta_2 \) are usually unknown. On the other hand, \( S_n \) is simpler, which allows for more straightforward theoretical treatment. Conveniently, Theorem 3.1 below provides an asymptotic linear relationship between the two score statistics, which allows us to bypass the intricacy of direct theoretical treatment of asymptotic properties of the score statistic with estimated parameters \( \hat{S}_n \) and focus on \( S_n \) instead. The reader will also notice that the theorem below also yields an expression for the asymptotic variance of \( \hat{S}_n \), which we did not discuss above. Note that Theorem 3.1 is a special case of Theorem 5.5, which is formulated and proved in Section 5.1.

**Theorem 3.1 (Asymptotic representation of score statistic)** Assume that the logistic model of (3.1) holds and let \( H_0^e; \beta_3 = 0 \) hold, meaning that \( \beta = \beta^0 = (\beta_0, \beta_1, \beta_2, 0)' \). Let \( S_n \) and \( \hat{S}_n \) be defined by (3.3) and (3.4), respectively. Then, \( S_n \) has zero expectation and is asymptotically normal with variance matrix \( I_\beta \). Moreover, \( \hat{S}_n = A S_n + o_p(1) \) for \( A = I_\beta^{1/2} (1 - \Pi_{I/H}) I_\beta^{-1/2} \), where \( I \) is the identity matrix and \( \Pi_{I/H} \) is the projection onto \( I_\beta^{1/2} H_0 \) with \( H_0 = \{ \beta \in \mathbb{R}^3 \times \{0\} \} \) being the null hypothesis local parameter space. Consequently, the statistic \( \hat{S}_n \) is asymptotically zero-mean normal with asymptotic variance matrix \( A I_\beta A' \).

### 3.2 Suitable pre-test statistics

For a two-stage procedure to be sensible in terms of power, one necessary requirement is that the validity of the post-test alternative hypothesis should imply (with high chance at least) the validity of the pre-test alternative hypothesis (Dai et al. (2010)). In Section 2.3.2 we provided motivation for using the two-locus genotype test of independence to search for interactions of two loci. We argued that it is not unreasonable to expect dependence within two-locus genotypes if the two loci in question are involved in interaction (with respect to the considered phenotype). In this section we provide technical details of several genotype independence tests that can serve as pre-tests in our two-stage testing procedure.

We start with the definition of asymptotic independence, which requires the distribution of a sequence of random vectors, under suitable centering and scaling, to converge to a non-degenerate product measure.

**Definition 3.2 (Asymptotic independence)** For a fixed \( m \in \mathbb{N} \) and for \( n = 1, 2, \ldots \) let
3 Two-stage testing for epistasis: Theory

$X_1^{n}, \ldots, X_m^{n}$ be random vectors of dimensions $k_1, \ldots, k_m$, respectively. The random sequences

$(X_1^{1})_{n=1}^{\infty}, \ldots, (X_m^{n})_{n=1}^{\infty}$ are said to be (weakly) asymptotically independent (AI) as $n \to \infty$ if there exist sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ both of dimension $k = \sum_{i=1}^{m} k_i$, where the coordinates of $a_n$ are all positive, such that the sequence $(a_n \otimes (X_1^{1}, \ldots, X_m^{n})) - b_n)_{n=1}^{\infty}$, where $\otimes$ denotes coordinate-wise product, converges in distribution to a random vector with non-degenerate distribution function $F$ such that

$F(x_1, \ldots, x_m) = \prod_{i=1}^{m} F_i(x_i)$

for all $x_i \in \mathbb{R}^{k_i}, i = 1, \ldots, m$, and for some distribution functions $F_1, \ldots, F_m$.

3.2.1 Pooled-sample test of genotype independence

In the pre-test, we want to test the hypothesis of independence of loci $L_1$ and $L_2$. While this can be done in many ways, we choose to focus on chi-square tests on independence, for which there are several options, which differ by the exact form of the statistic used and/or the sample on which it is based. The first two pre-test statistics that we present are based on the full pooled sample of cases and controls, which are particularly convenient because of their inherent independence of the post-test full-sample statistic.

If we ignore the values of phenotype indicators $\Delta_i$, using the genotype variables $X_i, Y_i, i = 1, \ldots, n$ we can construct a contingency table with elements $n_{kl}$, which are the counts of observed individuals for whom $X_i = k$ and $Y_i = l$, where $k \in G_X$ and $l \in G_Y$. With the minor allele counting the genotypes sets are $G_1 = G_2 = \{0, 1, 2\}$, for which the genotype contingency table reads

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$n_{00}$</td>
<td>$n_{01}$</td>
<td>$n_{02}$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>1</td>
<td>$n_{10}$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{20}$</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>

Due to the nature of the case-control sampling, if we divide the pooled-sample contingency table above by $n$, we obtain the ML estimators of genotype probabilities within a hypothetical population $\overline{P}$, in which the prevalence of the phenotype $\Delta = 0$ is $\tau$ of (3.2). Therefore, we refer to $\overline{P}$ as the pooled population and throughout this text we define the genotype probabilities within the pooled population as

$\pi_{kl} = \mathbb{P}(X = k, Y = l), \quad \pi_k = \mathbb{P}(X = k), \quad \pi_l = \mathbb{P}(Y = l), \quad (3.5)$

and throughout this text we assume that for all $k, l = 0, 1, 2$ we have $\pi_{kl} > 0$.

Pre-test statistic based on the pooled sample: Pearson

Using the counts $n_{kl}$ defined above, the pooled-sample chi-square statistic is

$T_n^{po} = n \sum_{k,l} (\pi_{kl} - \overline{\pi}_k \overline{\pi}_l)^2 / (\overline{\pi}_k \overline{\pi}_l), \quad (3.6)$
where \( \pi_{k\ell} = n_{k\ell}/n \), \( \pi_k = n_k/n \) and \( \pi_j = n_j/n \) are the maximum likelihood (ML) estimators of probabilities \( \pi_{k\ell} \), \( \pi_k \), \( \pi_j \) defined in (3.5). Independence of two-locus genotypes within the pooled population \( P \) is equivalent to the null hypothesis \( H_0^{po} : \forall k, l : \pi_{k\ell} = \pi_k \pi_j \), which is tested against the alternative \( H_1^{po} : \exists k, l : \pi_{k\ell} \neq \pi_k \pi_j \). The rejection criterion is based on the asymptotic distribution of \( T_n^{po} \) under \( H_0^{po} \), which according to a classical result is the chi-square distribution with the degrees of freedom equal to \( (|G_X| - 1)(|G_Y| - 1) \). For the additive model there are four degrees of freedom.

As we show in Theorem 3.3 of Section 3.3, the pooled-sample statistic \( T_n^{po} \) and the score statistic \( S_n \) are asymptotically independent. This, as well as the computational simplicity of the pre-test based on \( T_n^{po} \), makes \( T_n^{po} \) extremely suitable to be used in the pre-test of the two-stage testing method with the post-test based on the score statistic \( S_n \). As we show in the simulation study of Chapter 4, the pooled-sample pre-test can indeed yield a very powerful two-stage testing method for interaction.

For the purposes of investigating the asymptotic properties of \( T_n^{po} \), it is convenient to substitute the ML estimators of probabilities \( \pi_k \) and \( \pi_j \) in the denominator of \( T_n^{po} \) by the true marginal probabilities \( \pi_k \) and \( \pi_j \) and instead focus on the squared Euclidean norm \( ||T_n^{po}||^2 \), where

\[
T_n^{po} = \left( \sqrt{n}(\pi_{k\ell} - \hat{\pi}_k \hat{\pi}_j)/\sqrt{\hat{\pi}_k \hat{\pi}_j} \right)_{k,l}.
\] (3.7)

It follows from Slutsky’s lemma (Lemma A.6) that the chi-square statistic \( T_n^{po} \) and the squared norm \( ||T_n^{po}||^2 \) have equal asymptotic distribution, since the maximum likelihood estimators of the single locus genotype probabilities are \( \sqrt{n} \)-consistent (see van der Vaart (1998), chapter 17). Therefore, asymptotic distribution results derived for \( ||T_n^{po}||^2 \) also hold for \( T_n^{po} \), allowing us to conveniently use \( T_n^{po} \) when formulating and proving the asymptotic independence of \( S_n \) of both \( T_n^{po} \) and \( T_n^{po} \) in Theorem 3.3.

From now on, we refer to \( T_n^{po} \), and other similar vectors whose Euclidian norm gives rise to one of our pre-test statistics, as (pre-test) generating vectors. Note that although \( T_n^{po} \) is indexed by two indices \( k \in G_X \) and \( l \in G_Y \), we view it as a vector and with components \( t_{i} \) indexed by a single index \( i \). While the particular ordering of the components of \( T_n^{po} \) into a vector is not important as long as it remains the same throughout the text. In this text we always consider the row-wise ordering, which for the additive model is given for instance by \( i = 3k + l \), meaning that \( t_i \) is indexed by \( i = 0, \ldots, 8 \).

A potential pitfall of using the pooled sample to measure dependence is that the evidence of dependence induced by interaction within each sample can be masked by the pooling of the samples. As we outline in Section 2.3.1, the direction of deviation from independence induced by interactions should be opposite within the population of controls relative to the population of cases. If the two samples of cases and controls are pooled, it can then happen that the resulting pooled sample shows very little or no dependence between the two loci. If, however, the two loci are unlinked, they are likely to be (almost) independent in the general population \( P \). Then, if the case-control ratio in our data set is such that the true prevalence
of the phenotype $\Delta = 0$ in $\mathcal{P}$ is different from $\tau$, the pooled sample is still likely to contain information about the possible dependence of the two loci that might be present within the individual subsamples. The objection to pooling is thus weakened, albeit not completely removed as some evidence of dependence between the loci is likely to be lost by pooling. As we show in Chapter 4, in a case-control data set with excess of controls and high prevalence of cases the effect of pooling is detrimental to the power performance when compared with a control-only test.

**Pre-test statistic based on the pooled sample: Trend test**

Lewinger et al. (2013) recently proposed to use in the pre-test an alternative pooled-sample single degree of freedom association statistic, which is based on a weighted correlation of numerically coded genotypes. They also mention the Pearson chisquare statistic $T_{p0}^n$, but they do not pursue it further for fear of low power and instead focus on a trend test statistic with a single degree of freedom. Using the notation $r = n^{-1/2} \sum_{k,l=0}^2 z(u_k, v_l) n_{kl}$, where $u_k$, $v_l$ are genotype based weights and $z(x, y)$ is the interaction function analogous to the one used in the LRM of (3.1), we define the test statistic

$$R_{p0}^n = \frac{(r-e_0)^2}{s_0^2},$$

where $e_0 = \bar{E}_0 r$ and $s_0^2 = \bar{\text{var}}_0 r$. These estimators of expectation of $r$ and the variance of $r - e_0$ are calculated under the null hypothesis of independence $H_0$. While the function $z$ expressed the mode of interaction of the two loci, the weights $u_k$ and $v_l$ depend on the selected single locus genotype model. For instance under the additive we would put $u_k = k$ and $v_l = l$. As far as the exact values of $e_0$ and $s_0^2$ are concerned, it is easy to show that $e_0 = n^{-1/2} \sum_{k,l=0}^2 z(u_k, v_l) n_{kl} n_{ij}$ is the ML estimator of of $E_0 r$. The expression for $s_0^2$ is slightly more complex, however, thus we postpone the expression and its calculation until Section 5.4.2. The statistic used by Lewinger et al. (2013) is obtained by putting $z(x, y) = xy$, however, it needs to be pointed out that the formula for standardizing variance estimator given by Lewinger et al. (2013) (on page 450) is not correct.

Lewinger et al. (2013) argue that the pooled-sample statistic $R_{p0}^n$ provides a convenient pre-test because it is asymptotically independent of the logistic regression score statistic (and other statistics that are asymptotically equivalent to it such as the likelihood ratio or Wald statistics). Their argument for asymptotic independence is based on presumed similarity of the statistic $R_{p0}^n$ and the logistic regression test of association statistic of Murcray et al. (2008), which is shown to be asymptotically independent of the likelihood ratio test within model 3.1. However, we feel the need to point out that the arguments for the asymptotic independence of $R_{p0}^n$ and the likelihood ratio statistic provided by Lewinger et al. (2013) and Murcray et al. (2008) are not precise enough. We address this in Theorem 3.3, which provides the asymptotic independence result for $\hat{S}_n$ and $R_{p0}^n$. At this point we only mention that the asymptotic independence of $\hat{S}_n$ and $R_{p0}^n$ is a consequence of the fact that the numerator term of $R_{p0}^n$ can be written as $r - e_0 = n^{1/2} \sum_{k,l=0}^2 z(u_k, v_l) (\hat{\pi}_{kl} - \hat{\pi}_k \hat{\pi}_l)$, which means that $R_{p0}^n$ is
a simple function of $T_{n}^{0\alpha}$, just like $T_{n}^{0\alpha}$ is.

3.2.2 Imbalanced case-control data sets

An alternative to the pooled-sample tests is to focus on the single population samples of either controls or cases. As it turns out, such approach can be particularly fruitful for case-control data sets with a substantial difference in the sample sizes of cases and controls. We refer to such data sets as *imbalanced*. In most situations actually, the number of controls is larger than the number of cases, while the reverse situation is less likely but also possible. As far as imbalance in favour of controls is concerned, even if a particular data set is not imbalanced, extra controls can usually be obtained from other available studies, while increasing the size of both samples is usually much more challenging. The possibility to gather extra controls relatively inexpensively is potentially quite beneficial. The fact that individuals from alternative studies, which perhaps focused on completely different phenotypes than our own study, can be used as extra controls by our method creates significant potential for power performance improvement over the corresponding single-stage test based on the same data. Naturally, when individuals from different studies are being pooled, attention needs to be paid to potential dissimilarity between the populations from which the data came, where population stratification can affect the results of an analysis.

**Power effects**

While the power benefits achieved by simultaneous increase of both the number cases and the number controls are rather obvious, the same does not immediately hold when only one of the two samples is enlarged, or at least not to the same extend as is typical for the single-stage test. It is intuitively clear, that the power functions of many single-stage test statistics, such as the score statistic in logistic regression, depend on the two respective sample sizes $N$ and $M$ through a factor that decreases as a function of $1/N + 1/M$. This translates into a upper limit for power gain that can be achieved by increasing only $N$, while keeping $M$ fixed, or vice-versa. In the logistic regression score test for interactions this means that adding extra controls into the study increases the score test’s power with hugely diminishing returns and the addition of only the controls translates less and less strongly into additional power.

We illustrate this behavior in Figure 3.1, where we plot the mean values of the score statistic (two left-most plots) and the corresponding asymptotic $p$-values (on the negative log10 scale, two right-most plots) against the total sample size $n$. The values were obtained using simulated genotype data for various interaction effects size under the assumption that $z(x, y) = xy$ in (3.1). The two pairs of plots show two different setups in terms of the case-control ratio. Under the first setup, the number of cases $M$ was fixed at 200 and the increasing

---

$^{4}$Alternatively, the power function can often be shown to depend on the control-case ratio $N/M$ in such a way that an increase of $N/M$ results in a vanishingly small increase of the power. For details see for example Foppa and Spiegelman (1997).
sample size was achieved by adding only the controls. Under the second setup, the sample sizes of cases \( M \) and controls \( N \) were equal, meaning that the sample consisted of equal number of cases and controls. In the left-most plot we observe how the increased sample size, which is only due to more controls, yields strongly diminishing returns in terms of the shift of the observed values of the score statistic away from zero. The effect translates into the \( p \)-values, which also show a levelling-off behavior as evidenced by the third plot from the left. On the other hand, if the sample is made up of equally many cases and controls, there is dramatic change in the observed behavior as illustrated by the second and fourth plots from the left. In the former we see a much more slowly vanishing increase with the overall sample size, while in the latter we see a non-vanishing linear increase of the corresponding \( p \)-values.

A consequence of such behavior is that in an imbalanced case-control data set, we should try to utilize the single-population excess individuals (typically controls) in a more productive way. One way to do this is to utilize pre-testing. By employing a two-stage testing setup, where the pre-test relies on the excess individuals, the effect of leveling-off power could potentially be limited or at least postponed until much larger imbalance relative to when it would occur for the single-stage test. With a single population pre-test, we effectively use the imbalance in our sample to our advantage much more effectively since the power of such pre-test does not need to suffer from diminishing returns of increasing sample size. Such possibility creates an intriguing motivation for employing single population based pre-tests, which is the focus of the following two subsections (Sections 3.2.3 and 3.2.4).

**Figure 3.1:** Illustration of dependence of the fourth coordinate of the score statistic \( \hat{S}_n \) from (3.4) on the sample size with different ratios between cases and controls. The two plots on the left show the observed score statistics for interaction, while the two plots on the right show the associated \( p \)-values (on the \(-\log_{10}\) scale). From left to right, the first and third plots show the results with the number of cases \( M \) fixed at 200 and \( N = n - 200 \), while the second and fourth plots show the results when \( M = N = \frac{1}{2} n \). The score statistics of model (3.1) with \( z(x,y) = xy \) were calculated from simulated case-control genotype data with phenotype status determined also according to (3.1) with \( z(x,y) = xy \). \( \beta_0 \) such that the population prevalence of cases was about 50\%, \( \beta_1 = \beta_2 = 0 \) and five different interaction effects \( \beta_3 = 0.2, 0.4, 0.6, 0.8, 1 \). Minor allele frequencies in the genotype data were \( f_1 = f_2 = 0.35 \) for both loci in each test.

### 3.2.3 Single population tests of genotype independence

If we only consider the \( N \) individuals that belong to the population of controls we get an analogous contingency table with elements \( N_{kl} \), where \( N_{kl} \) is the number of observed controls
for whom \( X_i = k \) and \( Y_i = l \) and \( \Delta_i = 0 \), where \( k \in G_X \) and \( l \in G_Y \). We can do the same with the sample of \( M \) cases using the counts \( M_{kl} \), where \( k \in G_X \) and \( l \in G_Y \). For the additive genotype model, with \( G_X = G_Y = \{0, 1, 2\} \), the resulting contingency tables are

<table>
<thead>
<tr>
<th>controls</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>( L_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( N_{00} )</td>
<td>( N_{01} )</td>
<td>( N_{02} )</td>
<td>( N_0 )</td>
</tr>
<tr>
<td>1</td>
<td>( N_{10} )</td>
<td>( N_{11} )</td>
<td>( N_{12} )</td>
<td>( N_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( N_{20} )</td>
<td>( N_{21} )</td>
<td>( N_{22} )</td>
<td>( N_2 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cases</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>( L_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( M_{00} )</td>
<td>( M_{01} )</td>
<td>( M_{02} )</td>
<td>( M_0 )</td>
</tr>
<tr>
<td>1</td>
<td>( M_{10} )</td>
<td>( M_{11} )</td>
<td>( M_{12} )</td>
<td>( M_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( M_{20} )</td>
<td>( M_{21} )</td>
<td>( M_{22} )</td>
<td>( M_2 )</td>
</tr>
</tbody>
</table>

Unlike the pooled-sample contingency table, both the control-only and case-only contingency tables above are directly linked with the corresponding population genotype probabilities. Within the control population these probabilities are

\[
p_{kl} = P(X = k, Y = l | \Delta = 0), \quad p_k = P(X = k | \Delta = 0), \quad q_l = P(Y = l | \Delta = 0),
\]

and the analogous probabilities within the case population by

\[
r_{kl} = P(X = k, Y = l | \Delta = 1), \quad r_k = P(X = k | \Delta = 1), \quad s_l = P(Y = l | \Delta = 1),
\]

Throughout the text all of these probabilities are assumed to be positive for all \( k, l \). Maximum likelihood estimators for the probabilities \( p_{kl} \) and \( r_{kl} \) can be obtained by dividing each element of the contingency tables by \( N \) and \( M \), respectively, which leads to \( \hat{p}_{kl} = N_{kl}/N \) and \( \hat{r}_{kl} = M_{kl}/M \). Similarly, for the marginal probabilities \( p_k, q_l, r_k, s_l \) the ML estimators are based on the marginal counts of the two tables above, which yields \( \hat{p}_k = N_k/N, \hat{q}_l = N_l/N, \hat{r}_k = N_k/M, \hat{s}_l = M_l/M \).

### Pre-test based on the control-only sample: Pearson statistic

With the genotype probabilities \( p_{kl}, p_k, q_l \) defined in (3.9), independence of two-locus genotypes at \( L_1 \) and \( L_2 \) within the control population is equivalent to \( p_{kl} = p_k q_l \) for all \( k, l \). Thus, the pair of hypotheses to test is \( H_0^\text{co}: p_{kl} = p_k q_l \) against \( H_1^\text{co}: p_{kl} \neq p_k q_l \). These hypotheses can be tested using the control-only chi-square statistic

\[
T_n^{\text{co}} = N \sum_{k,l} (\hat{p}_{kl} - \hat{p}_k \hat{q}_l)^2 / (\hat{p}_k \hat{q}_l),
\]

where \( \hat{p}_k = N_k/N, \hat{q}_l = N_l/N \) and \( \hat{p}_k \hat{q}_l = N_{kl}/N \) are the maximum likelihood (ML) estimators of \( p_{kl}, p_k, q_l \), respectively. Analogously to \( T_n^{\text{po}} \), the asymptotic distribution of \( T_n^{\text{co}} \) under \( H_0^\text{co} \) is the chi-square distribution with the same number of degrees of freedom as \( T_n^{\text{po}} \), that is \( (|G_X - 1|)|G_Y - 1| \). The vector whose squared Euclidean norm has the same asymptotic distribution as \( T_n^{\text{co}} \) is defined as

\[
T_n^{\text{co}} = \left( N^{1/2} (\hat{p}_{kl} - \hat{p}_k \hat{q}_l)/\sqrt{\hat{p}_k \hat{q}_l} \right)_{k,l}
\]

Analogously to \( T_n^{\text{po}} \), we again assume the elements of \( T_n^{\text{co}} \) are ordered row-wise, which for the additive model is achieved through indexing by \( i = 3k + l \).
Pre-test based on the control-only sample: Trend test statistic

Inspired by the single-degree-of-freedom pooled-sample statistic $R^p_n$, we additionally define an analogous control-only statistic

$$R^c_n = (r^c - e^c)^2/(s^c)^2,$$  \hspace{1cm} (3.13)

where $r^c = N^{-1/2} \sum_{k,l=0}^2 z(u_k, v_l)N_{kl}$ and $e^c$ and $s^c$ are the null hypothesis estimators of expectation and standard deviation of $r^c$. Analogously to $R^p_n$, under $H_0^c$, the asymptotic distribution of $R^c_n$ is chi-square with one degree of freedom. Analogously to $T^c_n$ of (3.12), we denote the generating vector of $R^c_n$ by

$$R^c_n = \left( M^{1/2} \left( \overline{r}_{kl} - \overline{r}_{kl} \overline{s}_i \right) / \sqrt{\overline{r}_{kl} \overline{s}_i} \right)_{k,l}. \hspace{1cm} (3.14)$$

To motivate this notation, we note that it is useful in Section 3.5, where we devise a modification of the score statistic that is independent of the pre-test statistic $R^c_n$.

Pre-test based on the case-only sample: Pearson and trend test statistics

A third option is to base the pre-test on the sample of cases only. With the framework of two-stage testing in mind, such independence test is only especially suitable for those case-control samples, in which there is an excess of the cases over the controls. Then, instead of $T^c_n$, we would use the case-only chi-square statistic

$$T^c_n = M \sum_{k,l}(\overline{r}_{kl} - \overline{r}_{kl} \overline{s}_i)^2 / (\overline{r}_{kl} \overline{s}_i),$$  \hspace{1cm} (3.15)

where $\overline{r}_{kl} = M_{kl}/M$, $\overline{r}_k = M_k/M$ and $\overline{s}_i = M_i/M$. The tested hypotheses would then be $H_0^c: r_{kl} = r_k s_i$ against $H_1^c: r_{kl} \neq r_k s_i$, where the probabilities are defined in (3.10). These hypotheses can be tested by comparing $T^c_n$ with the quantiles of the chi-square distribution with $(|G_X - 1)||G_Y - 1|$ degrees of freedom. The associated pre-test generating vector is

$$T^c_n = \left( M^{1/2} (\overline{r}_{kl} - \overline{r}_{kl}) / \sqrt{\overline{r}_{kl} \overline{s}_i} \right)_{k,l},$$  \hspace{1cm} (3.16)

while the analogue of $R^c_n$ is denoted by $R^c_n$. Using the case-only counts $M_{kl}$, we put $r^c = M^{-1/2} \sum_{k,l=0}^2 z(u_k, v_l)M_{kl}$, denote by $e^c$ and $s^c$ the null hypothesis estimators of expectation and standard deviation, and put

$$R^c_n = (r^c - e^c)^2/(s^c)^2,$$  \hspace{1cm} (3.17)

$$R^c_n = \left( M^{1/2} z(u_k, v_l)(\overline{r}_{kl} - \overline{r}_{kl}) \right)_{k,l}. \hspace{1cm} (3.18)$$

3.2.4 Partial-sample single population tests of genotype independence

As an alternative to the single population statistics $T^c_n$, $R^c_n$, $T^c_n$ and $R^c_n$, we can base the pre-test on only parts of the two samples. There are several benefits with such approach. First
of all, if only part of the sample of controls (or cases) is used, then the degree of dependence between the pre-test statistic and the full-sample post-test statistic such as \(\hat{S}_n\) is proportionately reduced. As it turns out, in some cases the dependence might be lowered enough for the Bonferroni correction by the actual number of tests in the post-test to be sufficient. We address this aspect in more detail in the applied setting of Chapter 4.

Another reason why a partial-sample pre-test statistic is useful is that it actually allows to remove the dependence between the two-stage completely. One way to do this is by not re-using the pre-test sample in the calculation of the post-test statistic at all, thus making the two steps stochastically independent as a consequence of independence of individuals in the case-control data set. This approach is further discussed in Section 3.4. Alternatively, using a partial-sample pre-test statistic allows for modification of the post-test to achieve independence of the pre-test statistic while at the same time making sure that the resulting post-test statistic is centered under the appropriate null hypothesis. Such approach is studied in Section 3.5. It is important to keep in mind that these approaches are potentially beneficial especially for strongly imbalanced case-control data sets, where as it turns out the use of partial-sample pre-test statistics is often not detrimental at all to the overall power performance of the resulting two-stage testing procedure. Actually, as we show in Chapter 4, quite the opposite often turns out to be true, especially if the issue of model misspecification within the score statistic is also considered.

**Partial sample control-only pre-test statistics**

Denote as \(\delta \in (0, 1)\) the proportion parameter that determines the fraction of the single population sample used in the pre-test. For the sample of controls, without loss of generality, we can use the first \(\lfloor \delta N \rfloor\) (integer part of \(\delta N\)) observed controls in the pre-test and denote \(N^\delta = \lfloor \delta N \rfloor\) the size of the pre-test sample. The ML estimators based on this partial sample are denoted \(\hat{p}_k^\delta = N_k^\delta/N^\delta\), \(\hat{p}_k = N_k/N^\delta\) and \(\hat{q}_l^\delta = N_l^\delta/N^\delta\), which means that a control-only partial sample pre-test statistic can be

\[
T_{n, \delta}^{co} = N^\delta \sum_{k,l} (\hat{p}_k^\delta - \hat{p}_k \hat{q}_l^\delta)^2 / (\hat{p}_k \hat{q}_l^\delta).
\]  

(3.19)

Quite obviously, using the partial sample does neither change the tested hypotheses, which are \(H_{0, \delta}^{co}\) and \(H_{1, \delta}^{co}\), nor the asymptotic distribution of \(T_{n, \delta}^{co}\), which is the chi-square distribution with \(|G_X - 1| |G_Y - 1|\) degrees of freedom, that is four under the additive model, the same as the asymptotic distribution of \(T_{n}^{co}\). Alternative pre-test statistic is the partial sample single degree of freedom chi-square statistic \(R_{n, \delta}^{co}\), which we define completely analogously to (3.13), except using the partial sample counts \(N_k^\delta\). Using obvious notation, we define

\[
R_{n, \delta}^{co} = (r_{0, \delta}^{co} - \hat{e}_{0, \delta}^{co})^2 / (\hat{e}_{0, \delta}^{co})^2.
\]  

(3.20)

In addition, we define the partial sample pre-test generating vectors

\[
T_{n, \delta}^{co} = \left( (N^\delta)^{1/2} (\hat{p}_k^\delta - \hat{p}_k \hat{q}_l^\delta) / \sqrt{\hat{q}_l^\delta} \right)_{k,l}.
\]  

(3.21)
For the purposes of developing theory it is useful to denote the size of the complementary partial samples by \( N^{\delta C} = N - \lfloor \delta N \rfloor \) and \( a \) and the associated generating vectors by

\[
T_n^{\delta,\delta C} = \left( (N^{\delta C})^{1/2} (\tilde{p}_{ik}^{\delta C} - \tilde{p}_k^{\delta C} \tilde{q}_l^{\delta C}) / \sqrt{p_k q_l} \right)_{k,l}. \tag{3.23}
\]

\[
R_n^{\delta,\delta C} = \left( (N^{\delta C})^{1/2} z(u_k,v_l)(\tilde{p}_{ik}^{\delta C} - \tilde{p}_k^{\delta C} \tilde{q}_l^{\delta C}) \right)_{k,l}. \tag{3.24}
\]

It is important that due to the independence of genotypes of individuals in the pre-test sample, the two vectors \( T_n^{\delta,\delta C} \) and \( R_n^{\delta,\delta C} \) are independent of \( T_n^{\delta,\delta C} \) and \( R_n^{\delta,\delta C} \). To avoid technical complications we additionally define \( T_n^{\delta,\delta C} = R_n^{\delta,\delta C} = 0 \) for \( \delta = 0 \) and \( T_n^{\delta,\delta C} = R_n^{\delta,\delta C} = 0 \) for \( \delta = 1 \), which then allows us to consider values \( \delta \in [0,1] \) and makes \( T_n^{\delta} \) and \( R_n^{\delta} \) special cases of \( T_n^{\delta,\delta C} \) and \( R_n^{\delta,\delta C} \), respectively.

**Partial sample case-only pre-test statistics**

In complete analogy to the control based statistics defined above, we can formulate their alternatives using the sample of cases. These are useful for imbalanced samples with excess of cases. While explicit definitions may seem rather cumbersome and unnecessary, we provide them here for easy reference within the theorems and lemmas formulated later in Part I.

With analogously denoted sample size and ML estimators within the partial sample of cases, two possible case-only partial sample pre-test statistics are

\[
T_n^{ca,\delta} = M_\delta \sum_{k,l} (\tilde{r}_{ik}^{\delta C} - \tilde{r}_k^{\delta C} \tilde{s}_l^{\delta C})^2 / (\tilde{r}_k^{\delta C} \tilde{s}_l^{\delta C}), \tag{3.25}
\]

\[
R_n^{ca,\delta} = (r_{ca,\delta} - \hat{e}_{ca,\delta})^2 / (\hat{e}_{ca,\delta}^2), \tag{3.26}
\]

where we trust the definitions of the terms in (3.25) and (3.20) are obvious at this point. The corresponding generating vectors are

\[
T_n^{ca,\delta} = \left( (M^{\delta})^{1/2} (\tilde{r}_{ik}^{\delta C} - \tilde{r}_k^{\delta C} \tilde{s}_l^{\delta C}) / \sqrt{r_k s_l} \right)_{k,l}, \tag{3.27}
\]

\[
R_n^{ca,\delta} = \left( (M^{\delta})^{1/2} z(u_k,v_l)(\tilde{r}_{ik}^{\delta C} - \tilde{r}_k^{\delta C} \tilde{s}_l^{\delta C}) \right)_{k,l}, \tag{3.28}
\]

while the generating vectors for complementary case-only subsample of size \( M^{\delta C} \) are

\[
T_n^{ca,\delta C} = \left( (M^{\delta C})^{1/2} (\tilde{r}_{ik}^{\delta C} - \tilde{r}_k^{\delta C} \tilde{s}_l^{\delta C}) / \sqrt{r_k s_l} \right)_{k,l}, \tag{3.29}
\]

\[
R_n^{ca,\delta C} = \left( (M^{\delta C})^{1/2} z(u_k,v_l)(\tilde{r}_{ik}^{\delta C} - \tilde{r}_k^{\delta C} \tilde{s}_l^{\delta C}) \right)_{k,l}. \tag{3.30}
\]

For \( \delta = 0 \) we put \( T_n^{ca,\delta} = 0 \) and \( R_n^{ca,\delta} = 0 \) and for \( \delta = 1 \) we put \( T_n^{ca,\delta C} = 0 \) and \( R_n^{ca,\delta C} = 0 \). The asymptotic distributions of \( T_n^{ca,\delta} \) and \( R_n^{ca,\delta} \) under two-locus genotype independence within the case population are chi-square with \( |(G_X - 1)(G_Y - 1)| \) and one degrees of freedom, respectively. Finally, we note that the vectors \( T_n^{ca,\delta}, R_n^{ca,\delta} \) are independent of \( T_n^{ca,\delta C}, R_n^{ca,\delta C} \).
**Choice of single population tests for testing interactions**

We should point out that tests based on the population of cases, such as $T_{n}^{ca}$ and $R_{n}^{ca}$, can be viable alternatives to the control-only tests, such as $T_{n}^{co}$ and $R_{n}^{co}$. Actually, from biological perspective, if the disease is rare, then one can estimate the interaction parameters by means of case data alone (Piegorsch et al. (1994)), suggesting that measuring genotype dependence among the cases should be productive. Piegorsch et al. (1994), Gauderman (2002), Lewinger et al. (2013) among others even argue that a case-only is more powerful than a traditional case-control test of interaction, although they do admit that single population tests, such as their case-only test, can have high type I error rates in the presence of population level (background) gene-gene associations. An alternative case-only approach based on principal component analysis has been proposed by Bhattacharjee et al. (2010), which was aimed at reducing bias due to population structure. However, their simulations show that the case-only method still suffers from the problem of inflated type I error rates. Fortunately, in a two-stage setup with pre-tests based on either only the controls or only the cases the possible increased false detection rates do not present a problem as the type I error control is preserved via the post-test, provided the two steps are sufficiently independent.

Comparing the single population pre-tests, it should be pointed out that the discussion presented at the end of the previous section about power gains that a score test for interaction can achieve by increasing the size of only one of the two samples while keeping the other sample size fixed applies here as well. This means that especially for imbalanced case-control data sets a two-stage procedure with a single-population pre-test should be ideal. For low prevalence phenotypes an ideal situation would be an excess of cases, although such excess of cases is rare for GWAS data sets. Moreover, unlike with the controls, the number of cases cannot simply be increased by merging data from various studies that focus on different phenotypes. Nonetheless, based on the results in Chapter 4 for low prevalence phenotypes even without an excess of cases a case-based pre-test two-stage methods are preferable. On the other hand, for high prevalence phenotypes it is often the controls that are in excess in many GWAS data sets and their use in pre-tests can be very fruitful as shown in Chapter 4L.

**Dealing with linked loci**

It is important to note that a case- or control-only test of dependence can be undesirably sensitive to background dependence of genotypes similarly to $T_{L}^{LID}$ of (2.2). If used as a pre-test, however, this behavior does not affect the overall type I error rate of the two-stage procedure. If there is a lot of background dependence among the considered pairs, this typically results in a higher number of rejections by the pre-tests, which results in a larger multiple testing correction in the post-test and a negative influence toward the power of the procedure. In such case one could apply the pre-tests only to the pairs of loci that are known to be unlinked. This would likely result in a relatively smaller number of undesirable rejections by the pre-tests, thus lowering the multiple testing burden in the post-tests. However, such removal of a priori known linked pairs from the analysis completely negates the possibility of finding interacting
pairs among them, which is an aspect worth considering. An alternative solution to the problem of linked loci seems to be not performing any pre-testing for those pairs of loci. However, paying special attention to linked loci makes matters less convenient. In order to avoid that one can also choose not to deal with this question and apply the two-stage procedure even to the pairs of linked loci. This is not unreasonable since the presence of background dependence does not influence the overall type I error rate of the procedure. Moreover, since the pre-test is likely to eliminate some of the pairs with background dependence from the post-test phase, the two-stage procedure is viable even if background dependence is present in the data. This is especially the case for the kinds of highly imbalanced sets we have in mind where we can afford to "spend" some individuals in the pre-test without losing much of the power in the post-test (as illustrated by Figure 3.1).

**Test of dependence structure difference of cases and controls**

Yet another alternative test can be devised using the combined sample of controls and cases. However, instead of discarding the phenotype information by pooling we calculate the generating vectors \( T_n^{co} \) and \( T_n^{ca} \) for the two samples and use them to measure the difference in dependence structure between the two populations. Motivation for using such difference is provided by the genetic argument presented in Section 2.3.1. However, in order to keep this text reasonably long in the rest of Part I we only focus on the single-sample and pooled-sample statistics defined above.

### 3.3 Post-test for pooled-sample pre-test

In this section we formulate a result concerning asymptotic independence of the score statistic \( \tilde{S}_n \) and the pre-test generating vector \( T_n^{po} \), and consequently of \( \tilde{S}_n \) and the norm \( T_n^{po} \), and also of \( \tilde{S}_n \) and \( R_n^{po} \). This is done in Theorem 3.3 and both the formulation and the proof of this theorem relies on the asymptotic representation of the score statistic provided by Theorem 3.1 of Section 3.1. We note that the simplicity of the result in Theorem 3.3 is an immediate consequence of the pooling employed by \( T_n^{po} \) and as such cannot be directly formulated for the single population full-sample statistics \( T_n^{co} \) and \( T_n^{ca} \). Nonetheless, independence of pre-test and post-test statistics can be achieved without pooling and we formulate the results in Sections 3.4 and 3.5. The proof of Theorem 3.3 is postponed until Section 5.2.

**Theorem 3.3 (Normality and independence with pooling)** Let \( \tilde{S}_n \) be defined by (3.4), \( T_n^{po} \) by (3.7) and let \( H_n^c \) hold. Then, \( (\tilde{S}_n', (T_n^{po} - E T_n^{po})')' \), as \( n \to \infty \), converges in distribution to the zero-mean normal distribution with variance matrix

\[
W_{po} = \begin{pmatrix}
A \hat{I}_\beta A' & 0 \\
0 & V_T^{po}
\end{pmatrix},
\]

where \( \hat{I}_\beta \) is the Fisher information matrix, \( A \) is defined in Theorem 3.1 and \( V_T^{po} \) is the asympt-
Post-tests based on disjoint samples

It is a trivial observation that if the post-test and pre-test statistics are calculated from independent data, then they are also independent. Recall that genotypes of all of the individuals in the case-control data set are assumed to be independent. If a single case-control sample is split up into two disjoint portions and each of the parts is used in only one of the two steps, we refer to this as sample splitting or disjoint testing. The idea of sample splitting is very natural and can be used to obtain independence of any two tests that are based on a sample of inde-
Sample splitting means that for a case-control data set with $N$ controls and $M$ cases we define the pre-test sample size ratio parameter $\delta \in (0, 1)$. Using $\delta$ we then randomly select a subsample of $[\delta N]$ controls or $[\delta M]$ cases depending on which single-population pre-test we wish to use ($T_{n,\delta}^{\text{cn}}$ of (3.19), $R_{n,\delta}^{\text{cn}}$ of (3.20), or $T_{n,\delta}^{\text{ca}}$ of (3.25) or $R_{n,\delta}^{\text{ca}}$ of (3.26)). In the post-test we utilize the same type of score statistic as in (3.4), which is calculated using only the disjoint sample of $n_{\delta}$ remaining individuals. For the control-only pre-test we get $n_{\delta} = [\delta N] + M$ and for the case-only pre-test we get $n_{\delta} = N + [\delta M]$. To provide a formal definition of the post-test statistic, we define random variables $B_i, i = 1, \ldots, n$, to be indicators of the $i$-th individual’s belonging to the pre-test subsample. We then denote the null hypothesis maximum likelihood (ML) estimator calculated from the post-test subsample of $n_{\delta}$ individuals by $\hat{\beta}_{\delta}$. The disjoint score test statistic within the LRM is defined as

$$\hat{S}_{n,\delta}^\delta = n_{\delta}^{-1/2} \sum_{i=1}^{n} (1 - B_i)(\Lambda_i - \Psi(\hat{\beta}_{\delta}^\delta z(X_i, Y_i))) z(X_i, Y_i).$$ (3.31)

Similarly to $\hat{S}_n$, the statistic $\hat{S}_{n,\delta}^\delta$ is asymptotically normally distributed as $n_{\delta} \to \infty$ with zero mean under the null hypothesis $H_0^\delta$, while its asymptotic variance matrix is also analogous to that of $\hat{S}_n$, except modified to reflect the different case-control ratio with the $n_{\delta}$ sized subsample, which effectively changes the distribution of $\Delta$.

**Performance in a multiple testing setting**

In a multiple testing scenario with $K$ tests, sample splitting is an excellent and reliable way to achieve independence of the two stage, provided of course that the pre-test samples are the same over all tests so that the data between the two steps does not overlap at all. However, the question of good power performance of a sample splitting procedure is not at all obvious. As we discuss in Section 3.6, if the number of false hypotheses is small relative to $K$, the two-stage approach leads to decrease of the multiple testing burden in the post-test.

It is only fair to point out that in a balanced sample with (roughly) equal number of cases and controls the power performance gains might be only limited. As we show in Chapter 4, even if the selected logistic regression model (including the form of $z(x, y)$) is the true mechanism causing two-locus genotype dependence and even if the setting favour the use of control-based pre-test (i.e. high population prevalence of cases), the two-stage procedure is not more powerful test than the single-stage full-sample score test. For balanced samples this is not surprising, since according to the classical theory the full-sample score test is locally asymptotically optimal (see Theorem 5.6). However, the situation becomes vastly different if the case-control data set is highly imbalanced and/or the underlying interaction model is different from the one assumed by the LRM (i.e., the function $z(x, y)$ is misspecified). Under such scenarios the combination of the score test and a non-parametric pre-test in a multiple testing setting can yield a very powerful procedure while maintaining proper error control. This robustness is crucial for application when the exact shape of interaction between loci is
3.5 Post-tests based on adjusting for pre-test

In the previous section we described how independence of the two stages can be achieved via sample splitting. An alternative path to independence is to modify the full-sample score statistic \( \hat{S}_n \) defined in (3.4) in a way that accounts for the possible dependence with the pre-test generating vectors \( T^\text{co}_n \), \( R^\text{co}_n \), \( T^\text{ca}_n \), and \( R^\text{ca}_n \) respectively defined in (3.12), (3.16), (3.14) and (3.18). The idea we pursue here is to regress \( \hat{S}_n \) onto a pre-test generating vector \( T_n \), which is determined by which pre-test is used, and then base the post-test only on the residual vector of such regression. We refer to this modification-by-regression approach as *adjusting for pre-test* and call the resulting tests *adjusted tests*.

Regressing score statistic onto the pre-test generating vector

Let us treat all of the pre-test generating vectors simultaneously using generic pre-test generating vector \( T_n \). In order to regress \( \hat{S}_n \) onto \( T_n \), we want to find a sequence of matrices \( B^*_n \) such that \( \hat{S}_n - B^*_n T_n \) and \( T_n \) are (asymptotically) orthogonal. In geometric terms, to identify \( B^*_n \) we need to find an orthogonal projection of the components of the vector \( \hat{S}_n \) onto the space spanned by \( T_n - E T_n \). With such \( B^*_n \), and provided the expectation of the pre-test vector was known, the residual vector

\[
\gamma^*_n = \hat{S}_n - B^*_n (T_n - E T_n) \tag{3.32}
\]

yields a way to test \( H^c_0 \) using \( \gamma^*_n \) in a two-stage setup by taking only its fourth coordinate \( \gamma^*_n(4) \), which is the one corresponding to the interaction term \( \beta_3 \). The reason we focus on \( T_n - E T_n \) in (3.32) and not just \( T_n \) is that the presence of \( E T_n \) is essential to make sure that \( \gamma^*_n(4) \) is centered under \( H^c_0 \) without having to assume that the pre-test null hypothesis holds (which would make \( E T_n \) equal to zero). Making such assumption about the pre-test null hypothesis is highly undesirable in a two-stage setup since we test \( H^c_0 \) only if the pre-test null hypothesis was rejected, in which case it is no longer guaranteed that \( E T_n \) is zero.

In Theorem 3.4 we show that \( \hat{S}_n \) and the pre-test vector \( T_n - E T_n \) are jointly asymptotically normal under the null hypothesis of no interactions. This not only allows us to show asymptotic normality of \( \gamma^*_n \), but it also provides a way to show that \( \gamma^*_n \) is asymptotically independent of \( T_n - E T_n \), which in turn implies the same for \( \gamma^*_n(4) \) and \( T_n - E T_n \). The latter result is achieved in Theorem 3.5, the proof of which relies on the central limit theorem, Slutsky’s lemma and the continuous mapping theorem. Through the use of these tools the theorem also yields the asymptotic independence of \( \gamma^*_n \) and the pre-test statistics \( T^\text{co}_n \) of (3.11), \( R^\text{co}_n \) of (3.13), \( T^\text{ca}_n \) of (3.15), or \( R^\text{ca}_n \) of (3.17), depending on whether \( T_n = T^\text{co}_n \), or \( T_n = T^\text{ca}_n \), or \( T_n = R^\text{co}_n \), or \( T_n = R^\text{ca}_n \). It is important to point out that the theorem does not require the pre-test null hypothesis to be true, meaning that the assumption of two-locus genotype in-
dependence is not necessary. However, the calculation of $\gamma_n^*$ requires the knowledge of the true expectation $E T_n$, which presents a complication that is addressed in Section 3.5.1. In Sections 3.5.2 and 3.5.3 we present modifications of the statistic $\gamma_n^*$, which work around the problem of unknown $E T_n$.

**Joint asymptotic normality and independence of pre-test and post-test**

The following theorem postulates that under $H_0^I$ the joint asymptotic distribution of the vectors $\vec{S}_n$ and $T_n - E T_n$ is zero-mean normal for any of the four pre-test generating vectors. While the proof of this theorem is postponed until Section 5.3.1, we note that it is based on the asymptotic representation of $\vec{S}_n$ given by Theorem 3.1, which allows us to obtain the desired result for $\vec{S}_n$ by investigating $S_n$. Further notice that the asymptotic representation also provides a way to express the asymptotic covariance matrix of $\vec{S}_n$ and $T_n$ using the covariance matrix of $S_n$ and $T_n$.

**Theorem 3.4 (Normality I)** Let $\vec{S}_n$ be defined by (3.4) and let either $T_n = T_n^{co}$ of (3.12), or $T_n = T_n^{cai}$ of (3.16), or $T_n = R_n^{co}$ of (3.14), or $T_n = R_n^{cai}$ of (3.18). If $H_0^I$ holds, then the random vector $(\vec{S}_n, T_n - E T_n)'$, as $n \to \infty$, converges in distribution to the zero-mean normal distribution with variance

$$W_{ST} = \begin{pmatrix} A \bar{I}_B A' & A C_{ST} \\ C_{ST}' A' & V_T \end{pmatrix},$$

where $\bar{I}_B$ is the Fisher information matrix, $A$ is defined in Theorem 3.1, $V_T$ is the asymptotic variance matrix of $T_n$ and $C_{ST}$ is the asymptotic covariance matrix of $S_n$ and $T_n$.

After we have derived the joint asymptotic distribution of the vectors $\vec{S}_n$ and $T_n$ under the null hypothesis $H_0^I$, we show asymptotic independence of the centered pre-test vector $T_n - E T_n$ and the residual vector $\gamma_n^*$ defined above. This is accomplished in the following theorem with its proof formulated Section 5.3.1.

**Theorem 3.5 (Independence I)** Let $H_0^I$ hold and put either $T_n = T_n^{co}$ of (3.12), or $T_n = T_n^{cai}$ of (3.16), or $T_n = R_n^{co}$ of (3.14), or $T_n = R_n^{cai}$ of (3.18). Using the notation of Theorem 3.4, define projection matrices $B^* = AC_{ST} V_T$, where $V_T$ is a pseudoinverse of $V_T$. Let $\gamma_n^*$ be defined by (3.32) with $B^*$ in the place of $B_n^*$. Then, $\gamma_n^*$ is asymptotically normal with zero mean and variance matrix $V_{\gamma} = A(I_B - C_{ST} V_T C_{ST}) A'$ and $\gamma_n^*$ and $T_n$ are asymptotically independent (AI). Consequently, if $T_n = T_n^{co}$, then $\gamma_n^*$ and $T_n^{co}$ are AI, if $T_n = T_n^{cai}$, then $\gamma_n^*$ and $T_n^{cai}$ are AI, if $T_n = R_n^{co}$, then $\gamma_n^*$ and $R_n^{co}$ are AI, and if $T_n = R_n^{cai}$, then $\gamma_n^*$ and $R_n^{cai}$ are AI.

### 3.5.1 Problem of unknown expectation

In (3.32) we defined the statistic $\gamma_n^*$, which is adjusted for pre-test, and we showed that under $H_0^I$ it is asymptotically independent of $T_n - E T_n$ and asymptotically zero-mean normal. We
can use the asymptotic variance of $\gamma_n^*$ given by Theorem 3.5 to devise a consistent estimator of the variance of $\gamma_n^{(4)}$. Using such estimator $\sqrt{\text{var} \gamma_n^{(4)}}$ we can test $H_0$ by comparing the observed value of $\Gamma_n^* = \gamma_n^{(4)}/(\sqrt{\text{var} \gamma_n^{(4)}})^{1/2}$ with the appropriate standard normal quantiles. However, there is a problem with the test based on $\Gamma_n^*$, which comes from the fact that $\mathbb{E}T_n$ is generally unknown to us. More precisely, Lemma 5.8 shows that if $T_n = T_n^{(4)}$ the expectation of $T_n$ is $\mathbb{E}T_n = \sqrt{N}((p_{kl} - p_k q_l)/\sqrt{p_k q_l})_{k,l} + O(N^{-1/2})$. For the other generating vectors it yields analogous expressions. Unfortunately, in either case $\mathbb{E}T_n$ is an unknown value to us unless we assumed independence of genotypes within the pre-test population, in which case it would be zero. However, we want to avoid making that assumption since in the two-stage testing procedure we want to calculate $\gamma_n^*$ only for the pairs for which we rejected such independence. The expression for $\mathbb{E}T_n$ contains unknown probabilities, which we cannot unfortunately estimate using the same sample from which we calculated $T_n$. The problem is that replacing the unknown probabilities by their maximum likelihood estimators based on that sample would yield an estimator $\tilde{\mathbb{E}}T_n = T_n$, which effectively drops the regression term from $\gamma_n^*$.

### 3.5.2 Adjusted post-test based on centering by splitting pre-test sample

In (3.32) we defined $\gamma_n^*$ using the assumption of known expectation $\mathbb{E}T_n$. There are several options for replacing the unknown expectation $\mathbb{E}T_n$ in $\gamma_n^*$ with a known vector. First option is to split the pre-test sample into two disjoint parts and use only one part for the pre-test and use the other part to obtain an estimator for the expectation $\mathbb{E}T_n$, with which the pre-test generating vector inside the regression term in $\gamma_n^*$ is centered. For the control-only partial sample the pre-test is based on $T_{n,\delta}$ defined in (3.19) or $R_{n,\delta}$ defined in (3.20). The associated generating vector $T_{n,\delta}$ defined in (3.21) can easily be centered by $T_{n,\delta,\delta,\delta}^c$ defined in (3.23), while the same holds for $R_{n,\delta,\delta}^c$ and $R_{n,\delta,\delta}^{\delta,\delta}$ of (3.22) and (3.24). Analogously, for the case-only partial sample pre-test, the statistic is $T_{n,\delta}^c$ of (3.25) or $R_{n,\delta,\delta}^c$ of (3.26), the associated generating vectors are $T_{n,\delta}^c$ of (3.27) and $R_{n,\delta,\delta}^c$ of (3.28), while the centering is achieved through $T_{n,\delta,\delta,\delta}^c$ of (3.29) and $R_{n,\delta,\delta}^{\delta,\delta}$ of (3.30).

In either case, due to the assumption of equal distribution within each of the two samples, there is no question of equality of expectations of the two pre-test generating vectors, with appropriate scaling dependent on relative sizes of the two disjoint parts, have equal distributions. In terms of necessary assumptions, the option of centering by splitting the pre-test sample seems very viable, since only the assumption of independence and identical distribution (iid) within the sample of controls, or within the sample of cases, is used. In this section we formulate a post-test statistic that is based on splitting of the pre-test sample.

If only partial single population sample is used in the pre-test, be it controls or cases, we test two locus genotype independence at two loci $L_1$ and $L_2$ using the statistic $T_{\delta}$ equal to either $T_{n,\delta}^c$ or $T_{n,\delta}^{\delta,\delta}$, or the statistic $R_{n,\delta}^c$ equal to either $R_{n,\delta,\delta}^c$ or $R_{n,\delta,\delta}^{\delta,\delta}$. We then respectively use $T_{n,\delta,\delta,\delta}^c$, or $T_{n,\delta,\delta,\delta}^{\delta,\delta}$, or $R_{n,\delta,\delta}^{\delta,\delta}$, or $R_{n,\delta,\delta}^{\delta,\delta}$ instead of the expectation inside the regression.
term of the post-test statistic $\gamma_n^\delta$ of (3.32). In other word, using the generic notation $T_n^\delta$ and $T_n^{\delta C}$ for the pre-test generating and centering vectors, we define

$$X_n^\delta = T_n^\delta - \sqrt{\delta/(1-\delta)} T_n^{\delta C}. \quad (3.33)$$

and use it in the place of $T_n - \mathbb{E}T_n$ when regressing the score vector $\tilde{S}_n$ onto the pre-test generating vector to make those two uncorrelated. In Theorem 3.6, among other things, we show that the vector $X_n^\delta$ has asymptotically zero expectation and link its variance matrix to the variance matrix of $T_n$. Analogously to $\gamma_n^\delta$ of (3.32), we define the residual vector

$$\gamma_n^\delta = \tilde{S}_n - B^\delta X_n^\delta, \quad (3.34)$$

where $B^\delta$ is the analogue of $B^\star$, which makes $\gamma_n^\delta$ (asymptotically) orthogonal to $T_n$. As we show in Theorem 3.7, there is a direct and unsurprising relationship between $B^\delta$ and $B^\star$ of Theorem 3.5, namely $B^\delta = \delta^{1/2} B^\star$. As we show in Theorem 3.7, the score statistic $\tilde{S}_n$ and the vectors $T_n^\delta$ and $T_n^{\delta C}$ are jointly asymptotically normal with such asymptotic covariance matrices, that make $\gamma_n^\delta$ and $T_n^\delta$ asymptotically independent under $H_0^\delta$. Moreover, under $H_0^\delta$ the asymptotic expectation of $\gamma_n^\delta$ is zero, which means that we test $H_0^\delta$ using the fourth coordinate of $\gamma_n^\delta$ denoted by $\gamma_n^\delta(4)$. Since Theorem 3.7 also give the asymptotic expectation and variance of $\gamma_n^\delta(4)$, we can easily devise a consistent estimator of $\text{var} \gamma_n^\delta(4)$ denoted by $\hat{\text{var}} \gamma_n^\delta(4)$. The asymptotic normality of $\gamma_n^\delta(4)$ together with Slutsky’s lemma imply that

$$\Gamma_n^\delta = \gamma_n^\delta(4)/(\hat{\text{var}} \gamma_n^\delta(4))^{1/2} \quad (3.35)$$

is asymptotically standard normal. Therefore, $\Gamma_n^\delta$ exceeding the appropriate standard normal quantile gives evidence against $H_0^\delta$.

**Joint asymptotic normality and independence of pre-test and post-test**

The following Theorem 3.6 is a generalization of Theorem 3.4 and provides joint asymptotic normality of $\tilde{S}_n$, $T_n^\delta - \mathbb{E}T_n^\delta$ and $T_n^{\delta C} - \mathbb{E}T_n^{\delta C}$ under $H_0^\delta$. Theorem 3.6 also gives asymptotic normality of the joint vector of $\tilde{S}_n$ and $X_n^\delta$. The proof of the theorem is formulated in Section 5.3.2.

**Theorem 3.6 (Normality II)** Let $\tilde{S}_n$ be defined by (3.4), let $T_n$, $T_n^\delta$, $T_n^{\delta C}$ be respectively defined either by (3.12), (3.21), (3.23), or by (3.16), (3.27), (3.29), or by (3.14), (3.22), (3.24), or by (3.18), (3.28), (3.30). Then, under $H_0^\delta$ and for any $\delta \in (0,1)$ the random vector $(\tilde{S}_n^\prime, (T_n^\delta - \mathbb{E}T_n^\delta)^\prime, (T_n^{\delta C} - \mathbb{E}T_n^{\delta C})^\prime)$, as $n \to \infty$, converges to the zero-mean normal distribution with variance matrix

$$W^\delta = \begin{pmatrix}
      A\mathcal{I}_B A' & \sqrt{\delta} AC_{ST} & \sqrt{1-\delta} AC_{ST} \\
      \sqrt{\delta} C_{ST} A' & V_T & 0 \\
      \sqrt{1-\delta} C_{ST} A' & 0 & V_T
    \end{pmatrix},$$

where $\mathcal{I}_B$ is the Fisher information matrix, $A$ is defined in Theorem 3.1, $V_T$ is the asymptotic variance matrix of $T_n$ and $C_{ST}$ is the asymptotic covariance matrix of $S_n$ and $T_n$. 

Analogously to Theorem 3.5, Theorem 3.7 below shows that for suitable matrix $B^\delta$ the vectors $\gamma^\delta_n$ and $T_n$ are asymptotically independent under $H^c_0$. The proof is again postponed until Section 5.3.2.

**Theorem 3.7 (Independence II)** Let $H^c_0$ hold and let $T^\delta_n$, $T^\delta_n$, and $T^\delta_{nC}$ be respectively defined either by (3.19), (3.21), (3.23), or by (3.20), (3.22), (3.24), or by (3.25), (3.27), (3.29), or by (3.26), (3.28), (3.30). Using the notation of Theorem 3.6, put $B^\delta = \delta^{1/2} AC_{ST} V_T$ and use it to define $\gamma^\delta_n$ by (3.34). Then, for any $\delta \in (0,1]$ the random vectors $\gamma^\delta_n$ and $T^\delta_n$ are asymptotically independent, making $\gamma^\delta_n$ and $T^\delta_n$ asymptotically independent. Moreover, $\gamma^\delta_n$ is asymptotically normal with zero expectation and variance matrix $V^\delta_{\gamma} = A(\Sigma + \frac{1}{\sqrt{N}} C_{ST} V_T C_{ST}' A')$ with the notation of Theorem 3.4.

As a side note, we point out that for $\delta = 1$, in both cases covered by the theorem above, we get $T^\delta_n - E T^\delta_n = T_n - E T_n$, which makes Theorem 3.5 a special case of Theorem 3.7. However, we formulated the two cases separately for the sake of readability and because it makes the proofs better structured.

From a practical point of view, the important result of the previous theorem is that under the null hypothesis $H^c_0$ the vector $\gamma^\delta_n$ is asymptotically centered and jointly asymptotically normal with $T^\delta_n$, and asymptotically independent under $H^c_0$ of both $T^\delta_n$ and $T^\delta_n$. Consequently, it can be used to test $H^c_0$ in the post-test.

Instead of using a partial sample statistic $T^\delta_n$ and vectors $T^\delta_n$ and $T^\delta_{nC}$, we also could use a full-sample statistic such as $T_{nc}^\delta$, and base the regression on $T_{nc}^\delta$ and $T_{nc}^{\deltaC}$, provided certain additional assumptions are made. Such approach and the associated results are formulated next in Section 3.5.3. As the design and the results are largely parallel to those of the current Section 3.5.2, an anxious reader can skip Section 3.5.3.

### 3.5.3 Adjusted post-test based on centering by the other sample

Besides the use of partial samples, there is another option to center a pre-test generating vector $T_n$ inside the regression term of $\gamma^\star_n$. The idea is to use the pre-test generating vector based on the other single population sample to make the residual term inside the test statistic centered. However, unlike the splitting of pre-test sample option above, the current option requires substantially more stringent assumption of equal cell probabilities between the two single population samples. Therefore, in this section we assume that the cell probabilities $p_{kl}$ and $r_{kl}$ are equal for all $k,l$, while justifiability of such assumption is discussed at the end of this subsection.

For instance, if only the controls are used in the pre-test and the test is based on $T_{nc}^{\delta}$ of (3.12) or $R_{nc}^{\delta}$ of (3.14), we can use $T_{nca}^{\delta}$ of (3.16) or $R_{nca}^{\delta}$ of (3.18) to perform the centering. The expectations of these vectors can be obtained using Lemma 5.8, which yields

$$E T_{nc}^{\delta} = \sqrt{N}((p_{kl} - p_{kq})/\sqrt{p_{kq}})_{k,l} + O(N^{-1/2}),$$
3 Two-stage testing for epistasis: Theory

$$\mathbb{E}T^{ca}_n = \sqrt{M}((r_{kl} - r_{k,l}s_1)/\sqrt{r_{k,l}s_1})_{k,l} + O(M^{-1/2}),$$

$$\mathbb{E}R^{co}_n = \sqrt{N}(z(u_k,v_l)(p_{kl} - p_kq_l))_{k,l} + O(N^{-1/2}),$$

$$\mathbb{E}R^{ca}_n = \sqrt{M}(z(u_k,v_l)(r_{kl} - r_{k,l}s_1))_{k,l} + O(M^{-1/2}).$$

Therefore, if $\mathbb{E}T^{ca}_n$ is rescaled by $(N/M)^{1/2}$, and $\mathbb{E}T^{co}_n$ by $(M/N)^{1/2}$, the resulting expectations are asymptotically equal\(^6\), provided we assume that the genotype probabilities $p_{kl}$ and $r_{kl}$ are equal for all $k,l$. The same holds also for $\mathbb{E}R^{ca}_n$ and $\mathbb{E}R^{co}_n$.

In light of this, we put $\rho_n = (N/M)^{1/2}$ and define

$$Y_n^{co} = T_n^{co} - \rho_n T^{ca}_n,$$

$$Z_n^{co} = R_n^{co} - \rho_n R^{ca}_n,$$

and replace $T_n = \mathbb{E}T_n$ by either $Y_n = Y_n^{co}$, or $Y_n = Y_n^{ca}$, or $Y_n = Z_n^{co}$, or $Y_n = Z_n^{ca}$ inside $\gamma_n^*$, depending on whether $T_n^{co}$, $T_n^{ca}$, $R_n^{co}$, or $R_n^{ca}$ was used in the pre-test. We then define

$$\gamma_n^Y = \hat{S}_n - B_n Y_n,$$  \quad (3.38)

where $B_n$ are appropriate projection matrices which provide asymptotic independence of $\gamma_n^Y$ and the pre-test generating vector $Y_n$.

In Theorem 3.8 we show the joint asymptotic normality of $\hat{S}_n$ and $Y_n$ for any of the above choices for $Y_n$, which among other implies asymptotic normality of $\gamma_n^Y$. Then, in Theorem 3.9 we show that if we put $T_n = T_n^{ca}$ in Theorem 3.5, due to the independence of cases and controls, the resulting matrix $B^*$ is what $\gamma_n^Y$ needs in the place of $B_n$ to make it asymptotically independent of $T_n^{co}$. Unsurprisingly, analogous projection matrices achieve the same feat for $T_n^{ca}$, $R_n^{co}$ and $R_n^{ca}$, which results in asymptotic independence of $\gamma_n^Y$ with $T_n^{co}$, $T_n^{ca}$, $R_n^{co}$, or $R_n^{ca}$, depending on the choice of $Y_n$.

All of this means that we can base a test of $H_n^Y$ on the fourth coordinate of $\gamma_n^Y$, which we denote as $\gamma_n^{Y(4)}$. Under admittedly more restricting assumptions of both $H_n^Y$ and $p_{kl} = r_{kl}$ for all $k,l$ the statistics $\gamma_n^{Y(4)}$ is asymptotically zero-mean normal. With its asymptotic variance given by Theorem 3.9, standardizing $\gamma_n^{Y(4)}$ by a consistent estimator of variance and comparing the resulting value with the appropriate standard normal quantiles then yields a valid statistical test of $H_n^Y$.

Joint asymptotic normality and independence of pre-test and post-test

The following theorem yields analogous properties of the joint vector of $\hat{S}_n$, $T_n^{co}$ and $T_n^{ca}$ that were provided by Theorem 3.6. Its proof is postponed until Section 5.3.3. We also point out that in Section 5.4 we present several lemmas, which give expressions for the asymptotic covariance matrices in the following theorem.

\(^6\)Actually, after rescaling the two expectations would be exactly equal, since even the $O$ terms would be equal under $p_{kl} = r_{kl}$. As is evident from the proof of Lemma 5.8.
Equal expectations among cases and controls

In order to use $T_n^{CA}$ for centering $T_n^{CO}$ within $Y_n^{Y}$, it is important that $T_n^{CO}$ and $\rho_n T_n^{CA}$ have equal expectation. As we said, the two expectations are equal after sample size rescaling if
the two-locus genotype probabilities $p_{kl}$ and $r_{kl}$ are equal for all $k,l = 0,1,2$. It would be ideal if we could show that in the LRM $\beta_3 = 0$ implies $p_{kl} = r_{kl}$. While this is not true in general, it is true if the main effects are also zero (i.e. $\hat{\beta}_1 = \hat{\beta}_2 = 0$). For the controls the two-locus and single-locus genotype probabilities $p_{kl}$, $p_k$, $q_l$ are defined by (3.9), while the probabilities $r_{kl}$, $r_k$, $s_l$ for the cases are defined in (3.10). Since the marginal probabilities are sums of the cell probabilities, we just need to show the equality $p_{kl} = r_{kl}$ for all $k,l$. The equality $p_{kl} = r_{kl}$ is equivalent to

$$\frac{\mathbb{P} (\Delta = 1)}{\mathbb{P} (\Delta = 0)} = \frac{\mathbb{P} (\Delta = 1 | X = k, Y = l)}{\mathbb{P} (\Delta = 0 | X = k, Y = l)}.$$  

Under the logistic model (3.1) and $H^*_0$ the right hand side of this equality turns into

$$\frac{\mathbb{P} (\Delta = 1 | X = k, Y = l)}{\mathbb{P} (\Delta = 0 | X = k, Y = l)} = e^{\beta_0 + \beta_1 k + \beta_2 l},$$  

implying that $p_{kl} = r_{kl}$ for all $k,l = 0,1,2$ if and only if $\mathbb{P} (\Delta = 1) / \mathbb{P} (\Delta = 0) = e^{\beta_0 + \beta_1 k + \beta_2 l}$, for all $k,l$. This is true only if $\beta_1 = \beta_2 = 0$, which means that in order for the cell and marginal probabilities to be the same within the cases and controls under the LRM, there must be no main effects of the two loci in question, meaning that the phenotype indicator $\Delta$ is independent of the genetic information variables $X,Y$. It seems potentially useful to make this an operating assumption, since a two-stage procedure which uses all of the controls in the pre-test and then centers the adjusted post-test statistic using the sample of cases may have a very good statistical power. How good the power actually is under realistic applied setting is investigated in Chapter 4.

### 3.5.4 Post-tests adjusted for multiple parallel pre-tests

In the previous two sections we addressed the problem of how to adjust the post-test score statistic to make it (asymptotically) independent of the corresponding pre-test statistic. In this section we formulate a method that builds on the regression idea and results in a post-test statistic that is independent of multiple pre-test statistics for different pairs of loci simultaneously. For simplicity, we focus on the controls-only pre-test based on $T_n^d$ of (3.21).

Suppose we have a number of different loci located over certain number of chromosomes and we want to test each pair of these loci for genetic interactions with respect to the binary phenotype $\Delta$. Suppose that the case-control data set defined in Section 3.1 contains data for all of the considered loci of the $n$ independent individuals. In the absence of interactions, in genetic applications it is often assumed that physically unlinked loci (i.e., loci on different chromosomes) are mutually independent unless associated with a phenotype. In our setting of a two-stage testing procedure this means that for pairs of loci $(L_a,L_b)$ and $(L_c,L_d)$, where loci $L_a$ and $L_b$ are on different chromosome(s) than $L_c$ and $L_d$, the score statistic $S_n$ that corresponds to the pair $(L_a,L_b)$ is stochastically independent of the pre-test generating vector $T_n$.
that was computed using the genetic information at \((L_c, L_d)\). However, if the loci pairs \((L_a, L_b)\) and \((L_c, L_d)\) (partially) share chromosomes, there may still be correlation of genetic information they contain even if the assumption of independent chromosomes is correct. Therefore, when performing multiple tests using the two-stage procedure we need to take into account the possible dependence of the score test on loci \(L_a, L_b\) and chisquare test on loci \(L_c, L_d\).

### Independence of multiple pre-tests

Using the regression and splitting of controls approach of Section 3.5.2, we can solve the problem of dependence between loci on the same chromosome by regressing the score statistic for a given loci pair not only on the corresponding pre-test generating vector but also on pre-test generating vectors for other pairs of loci. This provides a way to make the score statistic asymptotically independent of multiple chisquare pre-tests. However, regressing onto all \(K\) pairs may result in excessive computational burden, in which the regression can be limited to a smaller number of \(R\) loci pairs. For instance, an option is to include only loci pairs with at least one locus located on the same chromosome as one of the loci on which the score statistic is based. Alternatively, we can utilize known LD patterns provided for instance by the HapMap project to gain additional information about locus dependence patterns. With such information we could include into regression only those loci that are known to be strongly dependent with one or both of the loci of the pair on which the score statistic is based.

Denote the full-sample score statistics for all \(K\) loci pairs by \(\hat{S}_n^{(1)}, \ldots, \hat{S}_n^{(K)}\). Assume that for a selected \(i \in \{1, \ldots, K\}\) there are \(R_i\) loci pairs that \(\hat{S}_n^{(i)}\) is to be regressed on. For simplicity, assume that these loci pairs are indexed by \(r=1, \ldots, R_i\). Fix \(\delta \in (0, 1)\) and put \(N_\delta = [\delta N]\) and \(N_\delta^{\mathcal{C}} = N - N_\delta\). Splitting the sample of controls into subsamples of size \(N_\delta\) and \(N_\delta^{\mathcal{C}}\) and using the split for each of the \(R\) loci pairs, we denote the pre-test generating vectors by \(T_n^{\delta}, \ldots, T_n^{\delta}_{R_i}\) and \(T_n^{\delta_{\mathcal{C}}}, \ldots, T_n^{\delta_{\mathcal{C}}}_{R_i}\). These vectors are the analogues of \(T_n^{1}\) and \(T_n^{\mathcal{C}}\) defined in (3.21) and (3.23). We further denote the corresponding analogues of \(X_n^{\delta}\) from (3.33) by \(X_n^{\delta_{1}}, \ldots, X_n^{\delta_{R_i}}\), meaning that for \(r=1, \ldots, R_i\) we put

\[
X_n^{\delta_{r}} = T_{nr}^{\delta} - \left(\frac{\delta}{1-\delta}\right)^{1/2} T_{nr}^{\delta_{\mathcal{C}}}. \tag{3.39}
\]

The fact that \(T_n^{\delta_{1}}, \ldots, T_n^{\delta_{R_i}}\) are computed using the same individuals (which are therefore different from those that \(T_n^{\mathcal{C}}, \ldots, T_n^{\mathcal{C}_{R_i}}\) are computed on) makes \(T_{nr}^{\delta}\) and \(T_{ns}^{\delta_{\mathcal{C}}}\) independent for all \(r,s = 1, \ldots, R\). In order to make \(\hat{S}_n^{(i)}\) asymptotically independent of all \(T_n^{\delta_{1}}, \ldots, T_n^{\delta_{R_i}}\), we regress it onto \(X_n^{\delta_{1}}, \ldots, X_n^{\delta_{R_i}}\). Similarly to the independence Theorems 3.5, 3.7 and 3.9, we look for suitable matrices \(\mathbf{B}^{\delta}_{1i}, \ldots, \mathbf{B}^{\delta}_{R_i}\) such that under \(H_{0,j}^{\mathcal{C}}\)

\[
\mathbb{E}\left(\hat{S}_n^{(i)} - \sum_{k=1}^{R} \mathbf{B}^{\delta}_{ik} X_{nk}^{\delta}\right) T_{nj}^{\delta_{\mathcal{C}}} = o(1), \quad \text{for all} \quad j=1, \ldots, R. \tag{3.40}
\]

Using \(\mathbf{B}^{\delta}_{1i}, \ldots, \mathbf{B}^{\delta}_{R_i}\), we define the multiple-pre-test-adjusted score statistic

\[
\gamma_n^{(i)} = \hat{S}_n^{(i)} - \sum_{k=1}^{R} \mathbf{B}^{\delta}_{ik} X_{nk}^{\delta}. \tag{3.41}
\]
When solving (3.40) we can utilize the formulas for \( \text{cov}(\hat{S}_n, T_n^k) \) and \( \text{cov}(T_n^k, T_n^r) \) formulated in Section 5.3.4. Denote the fourth coordinate of \( \gamma_{n}^{(i)\delta} \) by \( \gamma_{n}^{(i)\delta} \). Since we can estimate the variance of \( \gamma_{n}^{(i)\delta} \) by combining Propositions 5.14 and 5.15, we can test \( H_{0,i}^{c} \), using the asymptotically normal statistic \( \Gamma_{n}^{(i)\delta} = \gamma_{n}^{(i)\delta}/(\text{var} \gamma_{n}^{(i)\delta})^{1/2} \), where the normality under \( H_{0,i}^{c} \) for all \( i = 1, \ldots, R \) follows by Proposition 5.13 in Section 5.3.4.

### 3.6 Error control in two-stage multiple testing

In this section we discuss in detail the multiple testing correction aspect of two-stage testing.

In this chapter we formulated several strategies how to come up with a post-test statistic that is independent of the pre-test. We showed that using \( T_{n}^{0} \) and \( S_{n} \) leads to asymptotically independent pre-test and post-test, which allows for lowering of the multiple testing correction needed in the post-test to retain FWER control. In this section we address the theoretical aspects of this in more detail.

#### 3.6.1 Two-stage multiple testing setup

Suppose we have a case-control data set with a total of \( K \) pairs of loci and to this data set we want to apply a two-stage testing procedure, which pre-tests independence of genotypes for all \( K \) pairs and subsequently post-tests for interactions only to the pairs with rejected independence by the pre-tests. It is natural to require that such procedure controls the overall type I error rate by a given significance level \( \alpha \in (0,1) \). Denoting the set of all tests by \( \mathcal{K} = \{1, \ldots, K\} \), the pre-test hypotheses are

\[
H_{0,k}^{\text{pre}}: p_{rs}^{k} = p_{r}^{k}q_{s}^{k}, \quad \text{for all} \quad r \in G_{X}^{k}, s \in G_{Y}^{k}, \quad k \in \mathcal{K},
\]

\[
H_{1,k}^{\text{pre}}: p_{rs}^{k} \neq p_{r}^{k}q_{s}^{k}, \quad \text{for some} \quad r \in G_{X}^{k}, s \in G_{Y}^{k}, \quad k \in \mathcal{K},
\]

where \( G_{X}^{k}, G_{Y}^{k} \) are the possible genotype sets for the \( k \)-th locus and \( p_{rs}^{k}, p_{r}^{k}, q_{s}^{k} \) are the two-locus and single-locus genotype probabilities for the \( k \)-th pair loci within the pre-test population. In the additive genotype model we have \( G_{X}^{k} = G_{Y}^{k} = \{0,1,2\} \) for all \( k = 1, \ldots, K \).

In order to test \( H_{0,k}^{\text{pre}} \) for a given \( k \), we compute the pre-test statistics \( T_{n}^{k} \) and reject \( H_{0,k}^{\text{pre}} \) if \( T_{n}^{k} \) is significant on level \( \alpha_{1} \in (0,1) \). Denote by \( \mathcal{K}_{1} \) the subset of \( \mathcal{K} \) that contains those indices for which \( H_{0,k}^{\text{pre}} \) is rejected and denote \( K_{1} \) the size of \( \mathcal{K}_{1} \), i.e., \( K_{1} = |\mathcal{K}_{1}| \). The corresponding post-test pairs of hypotheses are \( H_{0,k}^{c} : \beta_{3}^{k} = 0 \) and \( H_{1,k}^{c} : \beta_{3}^{k} \neq 0 \) for \( k \in \mathcal{K}_{1} \), where \( \beta_{3}^{k} \) is the corresponding interaction term parameter in the logistic regression model of the \( k \)-th pair of loci. For each \( k \in \mathcal{K}_{1} \) we then compute the post-test statistic \( \Gamma_{n}^{k} \) and we perform a statistical test of the hypothesis \( H_{0,k}^{c} \) on a suitably chosen level \( \alpha_{2} \in (0,1) \). In the rest of this section we work under the assumption that the statistics \( T_{n}^{k} \) and \( \Gamma_{n}^{k} \) are pair-wise independent while keeping in mind that pair-wise independence does not imply cross-dependence of \( T_{n}^{k_{1}} \) and \( \Gamma_{n}^{k_{2}} \).
3.6 Error control in two-stage multiple testing

for \( k_1 \neq k_2 \), which in the GWAS setting can easily occur.

### 3.6.2 FWER in two-stage testing

As would be the case with the chisquare pre-tests, we focus on tests where large values of \( T^k_{n} \) indicate violation of \( H^\text{pre}_{0,k} \), meaning that we reject \( H^\text{pre}_{0,k} \) if \( T^k_{n} \geq \tau_{\alpha_1} \), where \( \alpha_1 \in (0, 1) \) is the pre-test level and \( \tau_{\alpha_1} \) is the corresponding critical value of the distribution of \( T^k_{n} \). Let us also make such assumption about the significance regions of the post-tests, meaning that we reject \( H^\text{e}_{0,k} \) if \( \Gamma^k_{n} \geq \xi_{\alpha_2} \), where \( \xi_{\alpha_2} \) is the corresponding critical value of the distribution of \( \Gamma^k_{n} \). Denote by \( K_0 \subset K \) the set of those pairs of loci for which the null hypotheses \( H^\text{e}_{0,k} \) are true. Since type I error only occurs if we reject in both steps of the procedure, the family-wise error rate (FWER) of the two-stage multiple testing procedure is defined as

\[
\lambda_n = P(\exists k \in K_0 : \Gamma^k_{n} \geq \xi_{\alpha_2}, T^k_{n} \geq \tau_{\alpha_1}).
\]

For the sake of simplicity of notation we use the pre-test levels of significance are the same for all pre-tests and that the critical values are the same for all pre-tests, which means that the pre-test statistics all have the same distribution. And the same is assumed for all of the post-tests. More general formulation is easily achievable.

In a multiple testing scenario, we must also address the question of what exactly should the post-test level \( \alpha_2 \) be so that FWER is controlled by a chosen value \( \alpha \). In the following we focus only on Bonferroni-type corrections.

**Bonferroni correction by the total number of tests**

It is well known, that if we use the post-test level \( \alpha_2 = \alpha/K \), where \( K = |K| \), then the Boole inequality yields for any set of true hypotheses \( K_0 \subset K \) that

\[
\lambda_n \leq P(\exists k \in K_0 : \Gamma^k_{n} \geq \xi_{\alpha_2/K}) \leq \sum_{k \in K_0} P(\Gamma^k_{n} \geq \xi_{\alpha_2/K}) = |K_0| \alpha/K \leq \alpha,
\]

This means that the Bonferroni correction for all tests provides strong control of the FWER. However, it is also well known that taking \( \alpha_2 = \alpha/K \) often results in a significant and unnecessary loss of power, which makes such procedure suboptimal. Not to mention the fact that if we correct by \( K \), there is no point in performing any pre-tests. Since we seek to achieve better power performance through two-stage testing, we would like to choose \( \alpha_2 \) differently.

**Bonferroni correction by expected number of pre-test rejections**

When correcting by the fixed number of tests \( K \), we utilized the implied independence of \( \Gamma^k_{n} \) and \( K \) for all \( k \in K \). If we put \( \alpha_2 = \alpha/K_2 \) where \( K_2 \) is a priori fixed, we can employ this reasoning and show strong control of the FWER. An obvious suitable candidate for the choice of \( K_2 \) appears to be \( EK_1 \), where \( K_1 = |K_1| \), which is the expected number of rejected tests

\[7\] Alternatively we could also use Bonferroni inequality here. For both see Lemma A.11.
by the pre-tests. For shorter notation we denote $A_k = \{T^k_n \geq \tau_{\alpha_1}\}$. If we put $\alpha_2 = \alpha / \mathbb{E}K_1$, provided $\Gamma^k_n$ is (made) independent of $T^k_n$, we can write

$$
\lambda_n = \mathbb{P}(\exists k \in K_0 : \Gamma^k_n \geq \xi_{\alpha/\mathbb{E}K_1}, A_k) \leq \sum_{k \in K_0} \mathbb{P}(\Gamma^k_n \geq \xi_{\alpha/\mathbb{E}K_1}, A_k) = \sum_{k \in K_0} \mathbb{P}(\Gamma^k_n \geq \xi_{\alpha/\mathbb{E}K_1}) \mathbb{P}(A_k) 
\leq \sum_{k \in K_0} \mathbb{P}(A_k) \alpha / \mathbb{E}K_1 \leq \sum_{k \in K} \mathbb{P}(A_k) \alpha / \mathbb{E}K_1 = \alpha,
$$

where in $\triangleq$ we used the independence$^8$ of $\Gamma^k_n$ and $A_k$ implied by the independence of $\Gamma^k_n$ and $T^k_n$, and the last equality follows from

$$
\mathbb{E}K_1 = \mathbb{E} \sum_{k \in K} I(T^k_n \geq \tau_{\alpha_1}) = \mathbb{E} \sum_{k \in K} I(A_k) = \sum_{k \in K} \mathbb{P}(A_k).
$$

The inequality $\lambda_n \leq \alpha$ found in (3.42) makes correcting by $\mathbb{E}K_1$ very appealing, however, the complication that arises when taking $\alpha_2 = \alpha / \mathbb{E}K_1$ is the difficulty to accurately estimate $\mathbb{E}K_1$. A solution could for example be a use of an independent "representative" source of information (such as the HapMap project) and estimate $\mathbb{E}K_1$ through that. Another solution could be to use $K_1$ as an approximation for $\mathbb{E}K_1$, which is discussed next.

**Rare interaction setup**

In a situation when we expect that only a small fraction of pairs truly interact, the approximation of $\mathbb{E}K_1 = \mathbb{E}[K_1]$ by $\alpha_1 K$ appears reasonable and attractive. We refer to such situation as the rare interaction setup. It is worth pointing out that for extremely small pre-test levels even in the rare interaction setup this number might be too small. If there is a reason to worry that $\alpha_1 K$ is not sufficient, an option is to approximate $\mathbb{E}K_1$ by a multiple of $\alpha_1 K$, and take $mK$ with $m > \alpha_1$ instead of $\alpha_1 K$. The problem that remains of course is deciding how to choose the value $m$ appropriately. For instance, if we assume that the fraction of interacting pairs denoted by $\nu$ is below some value $\epsilon \in (0, 1)$, using $m = \alpha_1 + \epsilon$ and correct by $(\alpha_1 + \epsilon)K$, we gain a kind of "semi-strong" control of FWER (the prefix semi is to indicate that the control is strong only under the assumption $\nu \leq \epsilon$).$^9$

**Bonferroni correction by the actual number of pre-test rejections**

Next we consider the possibility of correcting the levels of the post-test by the actual number of post-tests. In other words, we put $\alpha_2 = \alpha / K_1$ (assuming $K_1 > 0$). Since $K_1 = |K_1|$ is a random variable, such post-test level $\alpha_2$ is also a random variable. In order to show that Bonferroni-type correction by the actual number of pre-test rejections provided FWER control, we need to further assume the independence of $\Gamma^k_n$ and $K_1$ (and consequently $K_1$)

---

$^8$As we showed earlier in this chapter if two statistics are not independent they can be made asymptotically independent using regression. If only asymptotic independence holds for $\Gamma^k_n$ and $T^k_n$ then the inequality in (3.42) should hold approximately.

$^9$The value $(\alpha_1 + \epsilon)K$ can be replaced by an even smaller value $\epsilon K + (1 - \epsilon)\alpha_1 K = (\alpha_1 + \epsilon)K - \alpha_1 \epsilon K$ with the same effect.
for all $k \in \mathcal{K}$. For instance, such independence is valid for the disjoint testing approach of Section 3.4, provided the sample split between the two steps remains fixed over all of the tests. Denoting $\mathcal{P}^K$ the set of all subsets of $\mathcal{K}$, under this stronger independence assumption, we can write

$$
\lambda_n = \Pr(\exists k \in \mathcal{K}_0 : \Gamma_n^k \geq \xi_{\alpha/\mathcal{K}_1}, A_k)
= \sum_{\mathcal{K}_1 \in \mathcal{P}^\mathcal{K}} \Pr(\mathcal{K}_1) \Pr(\exists k \in \mathcal{K}_0 \cap \mathcal{K}_1 : \Gamma_n^k \geq \xi_{\alpha/\mathcal{K}_1}, A_k | \mathcal{K}_1)
\leq \sum_{\mathcal{K}_1 \in \mathcal{P}^\mathcal{K}} \Pr(\mathcal{K}_1) \sum_{k \in \mathcal{K}_0 \cap \mathcal{K}_1} \Pr(\Gamma_n^k \geq \xi_{\alpha/\mathcal{K}_1} | \mathcal{K}_1)
\leq \alpha \sum_{\mathcal{K}_1 \in \mathcal{P}^\mathcal{K}} \Pr(\mathcal{K}_1) \leq \alpha.
$$

Note that the independence assumption of $\Gamma_n^k$ and $\mathcal{K}_1$ was necessary for the equality $**$ of the preceding display to hold. Full independence of $\Gamma_n^k$ and $\mathcal{K}_1$ for all $k$ can be achieved for example by sample splitting described in Section 3.4, while asymptotic independence occurs as a result of adjusting described in Section 3.5.

The problem of FWER control of the two-stage testing procedure described in this section was studied by Dai et al. (2010), where the focus is on asymptotically normal uncorrelated Wald-type statistics based on asymptotically linear estimators. They assume joint asymptotic normality of all pre-test and post-test statistics and show that under the joint pre-test and post-test null hypothesis the two-stage testing procedure controls the FWER in the strong sense, though Bonferroni correction is only applied to the second-stage testing. Moreover, they show that in high-dimensional hypothesis testing (with $\mathcal{K}_1/\mathcal{K} \rightarrow_p \alpha_1$) it is sufficient for strong FWER control that their pre-test and post-test statistics need only be pair-wise asymptotic normal and uncorrelated (under the joint null hypothesis). Concerning our setup such results suggest that asymptotic independence can be enough to achieve strong control of FWER. By Theorem 3.3 the pre-test-post-test pairs $T_n^{po}$ and $\mathcal{S}_n$ or $\mathcal{R}_n^{po}$ and $\mathcal{S}_n$ are asymptotically pair-wise independent. Moreover, using the regression approach described in Section 3.5, we can achieve asymptotic independence of the post-test and any of the single population pre-tests. However, in the GWAS setup it is possible to observe dependence of two-locus genotypes between many of the loci pairs, which we might want to account for within our test statistic. Relying of the regression approach, we can achieve strong control of FWER $\lambda_n$ by $\alpha$, at least asymptotically, by regressing each score statistic onto many or even all pre-test vectors. In Section 3.5 we presented the details of such regression. However, it is necessary to point out that regressing the score statistic onto many pre-test vectors can add substantially to the computational burden, which is undesirable in a GWAS setting. Moreover, such regression might cause difficulties with parameter estimation, making the resulting statistic unreliable. Therefore, if inter-pair dependence is of concern it may be a reasonable compromise to regress each score statistic onto only a relatively small subset of pre-tests with which
it is correlated the most. The selection of pairs to include in such regression could be based for instance on known patterns of LD, using the HapMap project, or similar sources.

In the next chapter we focus on putting the methods formulated in the current chapter to use on practical data sets. To illustrate the behavior we perform a large number of simulations. For reasons of tractability, we focus on the adjusted methods that regress on the corresponding pre-test only. This might be viewed as a slight omission in simulation, but it becomes perhaps even more relevant in real data analysis. However, since the main application of our methods is towards exploratory analysis, it seems to us that explicitly accounting only for the dependence between corresponding pre-tests and post-tests is a reasonable compromise between our efforts to avoid false rejections and the desire to have a powerful and reasonably fast interaction detection tool.
Building on the theory of Chapter 3, in this chapter we present the results of our investigation into the practical behavior of the two-stage testing procedures which utilize either the control-only or the case-only pre-tests. During the investigation we analyzed both simulated and real data sets. In the simulation we focused on both balanced and imbalanced data sets which were simulated under a number of different scenarios including different simulation and/or analysis interaction models, various sample size as well as different minor allele frequencies (MAFs). Moreover, we considered two distinct settings in terms of the population prevalence of cases. First, we focused on a setting with high population prevalence (around 50%), where it is suitable to use the two-stage methods with control-only pre-tests. In addition, we consider a setting with low population prevalence of cases (around 5%), where the case-only pre-test based two-stage methods turn out to be more appropriate. The case-only pre-test based two-stage methods were also used in the real data analysis of four Parkinson’s disease cohorts (courtesy of IPDGC\textsuperscript{1}).

The chapter is structured as follows. We first introduce the purpose built software which implements all of the discussed method. Then we specify several models of interaction within the framework of the logistic regression model, which were used in the simulation study, and provide a list of specific testing methods that are considered for the comparison. Given that the two-stage methods require the user to specify one or more tuning parameters, the problem of how to choose suitable values for these parameters is addressed. The attention is then turned to an investigation of type I error control by the considered methods, where we numerically verify the theoretical results of the previous chapter as well as provide an example of badly designed two-stage methods. After that the issue of power performance is

\textsuperscript{1}A full list of the IPDGC members and their affiliations is provided at the end of this chapter.
addressed, where we consider various combinations of data generating and data analysis interaction models. The question of robustness of the methods with respect to a misspecification of the interaction model is crucial for application in a practical genome-wide setting, where the exact knowledge of the data generating processes is essentially impossible. In the power performance comparison we show that in a realistic setting the standard single-stage method based on the logistic regression model is easily outperformed by the considered two-stage methods especially under model misspecification.

4.1 Software tool: EpiDET

The results presented in this chapter were obtained using the analysis tool called EpiDET, which we developed for this purpose. EpiDET was coded in Fortran and represents a major part of our work on developing a method for detecting epistasis. The fortran code contains over 50,000 lines of our own code, which build upon the module Logistic_Regression by Alan Miller (Miller (2003)) and the Fortran90 EISPACK library (Burkardt (2009)). We believe that the value and utility of the theoretical results presented in Part I of this thesis are greatly increased by the existence of a software package that implements them and provides an end-user with an easy way of performing the analysis of their genetic data. The tool provides an efficient multi-threaded implementation of all of the single-stage and two-stage methods described in Chapter 3. It is a command line software similar to the popular genetic data analysis package PLINK (Purcell et al. (2007)). Similarly to PLINK, EpiDET allows the user to specify a large number of options via command line flags. EpiDET takes on input the standard PLINK format data (both plain and binary) and it provides several options for data processing and conversion. The software can be readily deployed to perform a GWAS search for epistasis and its source code of this software is publicly available upon request. A brief documentation of EpiDET’s functions and usage is presented at the end of this chapter.

4.2 Interaction penetrance models

The use of the logistic regression model for modeling genetic interaction requires that we make choices about how to code the genetic information at hand. These choices concern the genotype variables $X$ and $Y$ and the interaction variable $z(X, Y)$ inside (3.1). As far as the genotype variables $X, Y$ are concerned, we already discussed the additive, recessive, dominant single locus genotype models and we also hinted at a general factor model in Section 3.1. Recall that we also made the assumption that $X$ and $Y$ are based on the additive model and count the minor alleles at each locus, thus take values in $G_X = G_Y = \{0, 1, 2\}$. In this chapter we choose the additive model.

A second choice we need to make when applying the logistic regression model (3.1) is
how to code the interaction term $z(x, y)$. As we said in Section 3.1, $z$ must be a non-constant and nonlinear function of $y$ and $x$. A particular choice of $z$ is referred to as interaction penetrance model, or shortly penetrance model (Marchini et al. (2005), Piriyapongsa et al. (2012)). While there are many reasonable choices for the values of $z$, picking one in particular can be a challenge, especially in exploratory analysis in which we need to analyze many loci pairs simultaneously without detailed knowledge of the potential form of interaction within each locus. In the following, we therefore focus on investigating the behaviour of our analysis methods under several different penetrance models. Namely, we focus on the six penetrance models given in Table 4.1. We refer to those models as $A$, $C$, $I$, $J$, $O$, $P$, where the naming is chosen arbitrarily. Model $A$ is arguably among the most popular models of interaction and under this model the interaction effect is multiplicative in the minor allele counts represented by variables $X$ and $Y$. Another popular choice is model $I$, which can be referred to as a threshold model. Under this model the interaction effect is only present if there is at least 1 minor allele at each locus but the size of effect does not change as more minor alleles are added to any of the two loci. Model $C$ on the other hand represents a situation in which the size of the interaction effect is not a monotone function in the minor allele counts $X$ and $Y$, resulting in a very different kind of interaction model. Under model $I$ the interaction term of the two loci appears only if the variables $X$ and $Y$ have equal values, while the size of the effect is a non-linear function of the allele count represented by these variables. Models $O$ represents a relatively mild deviation from model $A$ and $I$. Finally, perhaps despite a somewhat similar appearance of the penetrance table to the one for model $I$, model $P$ represents a strong deviation from models $A$ and $I$, as evidenced by the results presented later in this chapter. It seems that the difficulty with model $P$ is that the interaction effect occurs only between pairs of genotypes with at least three out of four alleles being the minor one. Especially if the MAF is small, such combinations rarely occur. Consequently, in a large testing scenario, the chance of detection of the interaction under models such as $P$ is negligible unless very large sample sizes are available and/or very large interaction effect sizes are present. In a genome-wide exploratory search, especially the latter is quite unlikely.

In this chapter we investigate the performance of several testing methods under two kinds of scenarios in terms of model selection. On the one hand, we focus on a setup in which the analysis method assumes the correct interaction model that was used to generate the data. On the other hand, since in a practical setup the exact shape of interaction is often unknown, it is perhaps even more important to investigate the behavior of any testing method also under model misspecification, where the analysis and simulation models differ. The difference between analysis and simulation models can have a strong influence on the score test, as it is a parametric test. Similarly, the single degree of freedom pre-tests are also parametric. On the other hand, the Pearson chi-square pre-tests are non-parametric since they make no direct assumptions about the form of interaction between the two loci. As we show below, under the model misspecification scenarios the two-stage testing methods with non-parametric pre-tests yield particularly strong power performance relative to the single degree of freedom pre-test
4.3 Analysis methods

During the analysis we fixed the desired overall level of significance $\alpha$ at $\alpha = 0.05$. In all of the two-stage procedures each of the $K$ pairs of loci is first tested for independence of the single-locus genotypes. Those pairs that exhibit dependence are passed on to the post-test, in which a variant of a score test is performed at an appropriately multiple testing corrected level of significance. The following analysis methods were investigated in the simulation study.

**CS: Classical full-sample score test (single-stage)**

As a benchmark in our study we use the single-stage score test with estimated parameters based on the full case-control sample ($n$ individuals). The actual test is performed using the fourth coordinate of the statistic $\hat{S}^n$ defined in (3.4), which is asymptotically normal with zero mean under the null hypothesis $H^0$ with asymptotic variance matrix given by Theorem 3.1. With $K$ pairs of loci we perform the Bonferroni correction for all $K$ tests and perform each single-stage test on level $\alpha_2 = \alpha/K$. We refer to this procedure as CS.

**Ppo$_1$-CS, Ppo$_2$-CS: Full-sample score test with pooled-sample pre-tests (two-stage)**

First two of the considered two-stage testing methods are based on the pooled-sample pre-tests that are based on the statistics $T^p$ or $R^p$, which are respectively defined in (3.6) and (3.8). Each of the pre-tests are combined with a post-test that is based on the full-sample score test statistic $\hat{S}^n$. We respectively refer to these combinations as Ppo$_2$-CS and Ppo$_2$-CS.

In Theorem 3.3 we showed that under the null hypothesis of no interaction $H^0$ both $T^p$ and $R^p$ are asymptotically independent of the full-sample score test statistic $\hat{S}^n$. Crucially, this means that for large enough sample size, if combined with $T^p$ or $R^p$ the post-test statistic $\hat{S}^n$ only needs to be multiple testing corrected for the number of tests in the post-test instead of the total number of tests $K$. In either case the $K$ pre-tests are performed on a

<table>
<thead>
<tr>
<th>$A$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$y = 2$</th>
<th>$C$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$y = 2$</th>
<th>$I$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$y = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$x = 0$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$x = 0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>$x = 1$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>$x = 1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$x = 2$</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>$x = 2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$x = 2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$J$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$y = 2$</th>
<th>$O$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$y = 2$</th>
<th>$P$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$y = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$x = 0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$x = 0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>$x = 1$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>$x = 1$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$x = 2$</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>$x = 2$</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>$x = 2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.1: Six combinations of interaction penetrance function $z(x,y)$ within the logistic regression model of (3.1). Each table represents one of the interaction models $A$, $C$, $I$, $J$, $O$, $P$ and shows values of $z(x,y)$ for the specified values of $x$ (invariant within rows) and $y$ (invariant within columns).
suitably chosen level $\alpha_1$, while the post-tests are performed on the level $\alpha_2 = \alpha/K_1$, where $K_1$ is the number of observed rejections by the given pre-test. If the selected level of significance in the pre-test is $\alpha_1$, assuming that the number of pairs with dependent genotypes is small relative to $K$, the actual multiple testing correction of $\hat{S}_n$ in the two-stage setup should be approximately $\alpha_1 K$. Therefore, if testing genotype independence proves to be a sufficiently powerful way to detect interactions under a given interaction model, these methods should provide strong competitors to CS.

**Pco$_1$-DS, Pco$_1$-DS, Pca$_4$-DS, Pca$_1$-DS: Disjoint score test methods**

An alternative to the pooling of samples is the sample splitting approach described in Section 3.4, where the available case-control data are split and disjoint samples are used in each step. Using the pre-test sample size ratio parameter $\delta$ we split either the control-only or the case-only subsample into two disjoint parts. Depending on which subsample the user chooses to utilize, a test is selected from among $T_n^{ca,\delta}$, $R_n^{co,\delta}$, $T_n^{ca,\delta}$, $R_n^{co,\delta}$ defined in (3.19), (3.20), (3.25), (3.26), respectively, where $T_n^{co,\delta}$ and $R_n^{co,\delta}$ would be calculated using the selected $[\delta N]$ controls, while $T_n^{ca,\delta}$ and $R_n^{ca,\delta}$ would be based on the selected $[\delta M]$ cases. The remaining part of the data of size $n_0 = n - [\delta N]$ is used to calculate $\hat{S}_n^0$ defined in (3.31). In the pre-test, the $K$ pre-tests are performed on a suitably chosen level $\alpha_1$. The post-tests are performed using $\hat{S}_n^0$ on the level $\alpha_2 = \alpha/K_1$, where $K_1$ is the number of observed rejections by the pre-tests. The critical values for the post-tests are derived using the asymptotic normal distribution of $\hat{S}_n^0$. As showed in Section 3.6.2 this correction yields a strong error control due to the independence of pre-test and post-test statistics, provided the pre-test subsamples are the same over all $K$ pre-tests. We refer to the two-stage procedures with pre-test based on $T_n^{co,\delta}$, $R_n^{co,\delta}$, $T_n^{ca,\delta}$, $R_n^{ca,\delta}$ as Pco$_4$-DS, Pco$_1$-DS, Pca$_4$-DS, Pca$_1$-DS, respectively.

**Pco$_4$-AS, Pco$_1$-AS, Pca$_4$-AS, Pca$_1$-AS: Adjusted score test methods**

In a completely analogous way to Pco$_4$-DS, Pco$_1$-DS, Pca$_4$-DS, Pca$_1$-DS above we define methods Pco$_4$-DS, Pco$_1$-DS, Pca$_4$-DS, Pca$_1$-DS using the adjusted score statistic $\Gamma_n^\delta$ defined in (3.35) in the place of $\hat{S}_n^0$. The asymptotic independence of $\Gamma_n^\delta$ with all four pre-test statistics follows from Theorem 3.7, which also yields asymptotic standard normality of $\Gamma_n^\delta$ under the null hypothesis $H_n^0$.

**Pco$_4$-CS, Pco$_1$-CS, Pca$_4$-CS, Pca$_1$-CS: Invalid two-stage procedures**

Finally, also analogously to the disjoint and adjusted score methods we define Pco$_4$-CS, Pco$_1$-CS, Pca$_4$-CS, Pca$_1$-CS where instead of $\hat{S}_n^0$ or $\Gamma_n^\delta$ we use the full-sample score test statistic $\hat{S}_n$. We must stress that due to the dependence between $\hat{S}_n$ and all of the pre-test statistics these procedures are invalid in the sense that they do not possess sufficient type I error control. We include these methods in the simulation study primarily to illustrate the degree of type I error control violation. The results in Section 4.5 should serve as a severe warning against the usage of such improperly designed methods.
4.4 Tuning parameters

An important aspect of the two-stage procedures is that they require the user to make choices about one or more tuning parameters. These tuning parameters are the pre-test level of significance $\alpha_1$, which is required by all of the two-stage procedures discussed in this text, and the pre-test sample size ratio $\delta$, which is required by all sample-splitting two-stage methods.

In this section we formulate a theory-based approach to choosing suitable values of these two parameters once sample sizes, expected effect sizes and population genotype distribution is specified. This theoretical approach is implemented in EpiDET, which increases the user-friendliness of the software.

Theoretical setup

For convenience in this section we focus only on the method $Pco\_DS$ but analogous results could be obtained for the other two-stage methods as well. Suppose that for a given pair of loci $L_1$ and $L_2$ we have a case-control data set $\{(\Delta_i, X_i, Y_i), i = 1, \ldots, n\}$ of $n$ independent individuals that fits the logistic regression model (3.1). Similarly to before, we assume that of these $n$ individuals there are $N$ controls and $M$ cases and that the genotype variables $X_i$ and $Y_i$ represent the genotypes under the additive model, in which these variables count the minor alleles at each locus, meaning $G_X = G_Y = \{0, 1, 2\}$. Denote simply as $T_n$ the pre-test statistic $T_n^{co, E\delta}$ computed using the first $n_1 = \lfloor \delta N \rfloor$ controls. Under genotype independence $T_n$ follows the chisquare-4 distribution.\footnote{Using the first $n_1$ controls during pretest, as opposed to a randomly selected subsample of size $n_1$, serves only the purpose of simpler notation, since the controls form a random sample.} The post-test utilizes the score statistic $\hat{S}_n^{\delta}$ defined in (3.31), which is based on the disjoint sample of $N - n_1$ controls and $M$ cases. With $K$ loci pairs, denote by $K_1$ the number of pairs that are rejected in the pre-test. For simplicity assume that we perform each post-test on the Bonferroni corrected level $\alpha/EK_1$, where we correct by the expected number of rejections in the pre-test. In Section 3.6.2, namely (3.42), we showed that such correction is sufficient for the FWER to be controlled by $\alpha$. Under the rare interaction setup (Section 3.6.2) $E K_1$ is a reasonable approximation for $K_1$ for the purposes of deriving sensible values for the tuning parameters. In this setup we reject the null hypothesis of no interaction for a given pair of loci if both $T_n \geq G^{-1}(1 - \alpha_1, 4, 0)$ and $\hat{S}_n \geq \Phi^{-1}(1 - \alpha/EK_1)$, where $\Phi^{-1}$ is the standard normal distribution quantile function and $G^{-1}(\cdot, p, \eta)$ is the quantile function of the chisquare distribution with $p$ degrees of freedom and non-centrality parameter $\eta$, respectively.

4.4.1 Power optimization

For the sake of notational simplicity, in this section we only focus on testing $\beta_3 = 0$ against the one-sided alternative $\beta_3 > 0$. The power investigation is then performed for the local
alternative of type $\hat{\beta}_3 = h/\sqrt{n_2}$ for some $h > 0$, where $n_2 = n - n_1$ is the sample size on which the post-test statistic is based. To simplify things still we also assume that in the general population for each $i = 1, \ldots, n$ the genotype variables $X_i$ and $Y_i$ are independent unless an interaction between $X_i$ and $Y_i$ present. We refer to this as the assumption of no background dependence. Under this assumption any dependence between two loci $L_1$ and $L_2$ within the control population is equivalent to the presence of interaction between the two loci. Let us further assume that the procedure is applied in the so-called rare interaction setup (see Section 3.6.2), which implies that $\mathbb{E}K_1$ can be well approximated by $\alpha_1 K$, as in such setup vast majority of rejections in the pre-test are false positives. This appears reasonable if the fraction of interacting pairs $\nu$ is much smaller than the chosen pretest level $\alpha_1$, in which case correcting by $\alpha_1 K$ should suffice.

Working under the local alternative and with the assumption of no background dependence, for each of the two steps we utilize their local limiting power functions. According to Cohen (1988), the distribution of pre-test statistic $T_n$ is increasingly well-approximated by the non-central chi-square distribution. More precisely, its limiting power function is

$$\Pi^{(1)}(\alpha_1, \delta) = 1 - G(G^{-1}(1 - \alpha_1, 4, 0), \eta_\delta),$$

where $G$ is the non-central chi-square distribution function with non-centrality parameter $\eta_\delta$. Note the explicitly denoted dependence on $\delta$. Under the current assumptions, the non-centrality parameter stabilizes in the limit and thus denoting its dependence on the sample size is not necessary. Moreover, based on the discussion in Section 5.1.4, namely Theorem 5.6, the limiting power function of the score post-test is

$$\Pi^{(2)}(\alpha_1, \delta) = 1 - \Phi(\Phi^{-1}(1 - \alpha/(\alpha_1 K)) - Bh),$$

where $\Phi$ denotes the normal distribution function. Note again the explicitly denoted dependence of the power function on $\delta$. The parameter $B$ in the power function is called the slope and its value can be calculated using Theorem 5.6 in Section 5.1.4. As showed in Lemma 5.7 therein, the slope of the statistic $\hat{S}_n$ for the current setup is equal to $B = (e_4', \tilde{\nu}^{-1} \tilde{e}_4)^{-1/2}$, where $e_4 = (0, 0, 0, 1)'$.

In order to determine the optimal pre-test level $\alpha_4^*$ and optimal ratio $\delta^*$ for this two-stage setup, we need to maximize the overall power function of the two-stage testing procedure, which due to the assumed independence of the two steps is equal to

$$\Pi_\mu(\alpha_1, \delta) = \mathbb{P}[\mathbb{P}(T_n \geq G^{-1}(1 - \alpha_1, 4, 0)) \mathbb{P}(\hat{S}_n \geq \Phi^{-1}(1 - \alpha/(\alpha_1 K)))]$$

For the purposes of finding $\alpha_4^*$ and $\delta^*$, we can asymptotically (for $N, M \to \infty$) approximate $\Pi_\mu(\alpha_1, \delta)$ by the local limiting power function

$$\Pi(\alpha_1, \delta) = \Pi^{(1)}(\alpha_1, \delta) \Pi^{(2)}(\alpha_1, \delta).$$

(4.1)

Under our simplified setup, the unknown quantities inside $\Pi$ such as the slope $B$ and the non-centrality parameter $\nu$ can be determined from the values of the parameters in the logistic regression model and the distribution of genotypes in the population in question. As far as $B = (e_4', \tilde{\nu}^{-1} \tilde{e}_4)^{-1/2}$ is concerned, these quantities are needed to calculate $\tilde{\nu}^\rho$. The case of $\eta_\delta$ is addressed below.
Non-centrality parameter $\eta_\delta$

Having specified the slope parameter $B$ under the logistic regression model, we can calculate $\Pi^{(2)}$. However, in order to be able to maximize $\Pi$ we also need to give an expression for the non-centrality parameter $\eta_\delta$. According to Cohen (1988) and Lemma 5.8, with or without the assumption of independent genotypes, for large $n_1$ the statistic $T_n$ has approximately the non-central chi-square-four distribution with non-centrality parameter $\eta_\delta = ||\mu||^2$, where $\mu = (\mu_{kl})_{k,l}$ has elements $\mu_{kl} = \sqrt{n_1}(p_{kl} - p_k q_l)/\sqrt{p_k q_l}$, which depend on $\delta$ through $n_1$. Since the value of parameter $\eta_\delta$ determines the power of the pre-test statistic $T_n$, we rewrite the terms $\mu_{kl}$ in terms of the parameter $\beta$ of the logistic regression model. We note that under the logistic regression model the two-locus genotype probabilities $p_{kl}$ defined in (3.9) can be written as $p_{kl} = \tau^{-1} \pi_{kl} (1 - \Psi_{kl})$ with $\tau = \mathbb{P}(\Delta = 0)$ defined in (3.2), $\pi_{kl}$ defined in (3.5), and $\Psi_{kl} = \Psi(\beta_0 + \beta_1 k + \beta_2 l + \beta_3 k l)$. This follows from

$$p_{kl} = \mathbb{P}(X = k, Y = l | \Delta = 0) = \frac{\mathbb{P}(\Delta = 0 | X = k, Y = l) \mathbb{P}(X = k, Y = l)}{\mathbb{P}(\Delta = 0)}.$$

Consequently, the single-locus genotype probabilities are $p_k = \tau^{-1} \sum_l \pi_{kl} (1 - \Psi_{kl})$ and $q_l = \tau^{-1} \sum_i \pi_{il} (1 - \Psi_{il})$, which yields $p_k q_l = \tau^{-2} \sum_{i,j} \pi_{ij} \pi_{kl} (1 - \Psi_{ij}) (1 - \Psi_{kl})$. Moreover, since the probabilities $p_{kl}$ sum up to one, we can express $\tau$ as $\tau = (\sum_{k,l} \pi_{kl} (1 - \Psi_{kl}))^{-1}$. Plugging these into $\mu_{kl}$ finally yields the non-centrality parameter

$$\eta_\delta = n_1 \sum_{k,l} \frac{\pi_{kl} (1 - \Psi_{kl}) - (\sum_{i,j} \pi_{ij} (1 - \Psi_{ij})) \sum_{i,j} \pi_{ij} \pi_{kl} (1 - \Psi_{ij}) (1 - \Psi_{kl})}{\sum_{i,j} \pi_{ij} \pi_{kl} (1 - \Psi_{ij}) (1 - \Psi_{kl})^2}.$$

Maximizing limiting power function

A useful conclusion that we can make based on the above display is that the non-centrality parameter $\eta_\delta$ can be expressed as a function of the probabilities $\pi_{kl}$ and parameters of the logistic regression model $\beta$. Moreover, as we showed, the slope parameter $B$ is also fully specified once these quantities are given. It seems reasonable that for the purposes of finding $\alpha^*_1$ and $\delta^*$ these parameters can be set to certain values based on reasonable expectations about the population at hand. The population genotype probabilities $\pi_{kl}$ and the population prevalence parameter $\beta_0$ either accompanies the case-control sample or can often be reasonably estimated using either external knowledge. The main effects $\beta_1, \beta_2$ can either be put equal to zero, as that should be close to the true values for most loci, or alternatively if there are known main effects for given loci with respect to the given phenotypes, other values can be more appropriate. Finally, as far as the interaction effect $\beta_3$ is concerned, for the purposes of maximizing $\Pi(\alpha_1, \delta)$ a value can be selected from a range that can be expected to be reasonably discoverable within the given data set, which depends among other on the chosen interaction model via the function $z$ in (3.1). An advantage of this approach is that it allows to tune the two-stage procedure’s sensitivity to specific values of $\beta$ with a specific interaction model that the user considers applicable and uses in the analysis. With given values of $p$,
\( \eta \) and \( B \), maximizing of the power function \( \Pi \) can be done either numerically or analytically. Unfortunately, it seems that the analytical approach for this setup is intractable, which suggests to maximize \( \Pi \) with respect to both \( \alpha_1 \) and \( \delta \) numerically.

### 4.4.2 Numerical maximization of power

For the one-sided alternative setting, we implemented the theory to calculate the power function of the two-stage disjoint score test \( P_{co_4-DS} \) as function of the pre-test level of significance \( \alpha_1 \) and pre-test sample size ratio \( \delta \). The theoretical power functions were then evaluated (not estimated) for large number of combinations of \( \alpha_1 \) and \( \delta \). The purpose of this section is to illustrate the dependence of the two-stage procedure’s power on these two tuning parameters under several different scenarios. However, this procedure can be readily applied to optimize the power function with respect to both \( \alpha_1 \) and \( \delta \) for a given data set, provided the necessary choices about LR model parameters and distribution of genotypes are made.

**Considered settings**

The results of numerically evaluated theoretical power with optimal values of tuning parameters are presented in Figures 4.1 and 4.2. The two figures show the power functions for the six interaction models \( A, I, C, J, O \) and \( P \). The power functions in these plots correspond to a single test assumed to be performed within a collection of 100 million tests, meaning \( K = 10^8 \), which is intended to represent a moderately sized GWAS data set. Different plots in each figure represent various combinations of input settings such as interaction models, interaction odds ratios (\( OR_3 = e^{\beta_3} \)) and different sample sizes. In every plot the number of cases is the same and equal to 2000 and the minor allele frequencies (MAFs) are fixed at 0.35 for both loci within each pair. On the other hand, the values of \( OR_3 \) range between 1.1 and 1.6 (left to right), while the control counts are 3000, 7000 or 11000 (top to bottom for each interaction model). For simplicity, no main effects are present (\( \beta_1 = \beta_2 = 0 \)) and the population prevalence of cases was around 50%.

The individual plots present the following situation. For each interaction model and each value of odds ratio a total of 95 different values of \( \delta \) were considered ranging between 0.01 and 0.95 in step of 0.01. Note that we intentionally did not consider \( \delta \) equal to zero, since then \( P_{co_4-DS} \) effectively turns into \( CS \), which we wanted to avoid in these plots. For each \( \delta \) the power function of the \( P_{co_4-DS} \) two-stage procedure was determined for a range of pre-test levels of significance \( \alpha_1 \) and the value \( \delta^* \) which yielded the highest maximum power was selected and the power function for that \( \delta \) was plotted. As we said, the procedure is assumed to be calculated within the rare interaction setup, which means that it is sufficient to perform multiple testing correction of the disjoint post-test score statistic by \( EK_1 = \alpha_1 K = \alpha_1 10^8 \). The maximizing values \( \delta^* \) and \( \alpha_1^* \) within each plot are presented in the upper-left and upper-right corners, respectively. Each plot also contains a red horizontal line, which denotes the power of the single-stage \( CS \) test. The two power functions are then compared and areas in
4.4 Tuning parameters

Theoretical power of DS for model A as a function of pretest levels for $N_{CA} = 2000$ and $K = 100M$

![Graph of DS power for model A]

Theoretical power of DS for model I as a function of pretest levels for $N_{CA} = 2000$ and $K = 100M$

![Graph of DS power for model I]

Theoretical power of DS for model J as a function of pretest levels for $N_{CA} = 2000$ and $K = 100M$

![Graph of DS power for model J]

Figure 4.1: Theoretical powers as functions of pre-test levels $\alpha_1$ (x axes) for various odds ratio (OR) of $\beta_3$ for analysis models A, I, C. No main effects ($\beta_1 = \beta_2 = 0$), 2000 cases and 3000, 7000, 11000 controls (top to bottom for each model) with MAFs fixed at 0.35 for both loci in each test, which are assumed to be one of 100 million.
Theoretical power of DS for model J as a function of pretest levels for \( N_{pa} = 2000 \) and \( K = 100M \)

- DS power
- CS power
- Max DS power
- DS power > 99% of max
- DS power > 90% of max
- DS power < CS power

Theoretical power of DS for model O as a function of pretest levels for \( N_{pa} = 2000 \) and \( K = 100M \)

- DS power
- CS power
- Max DS power
- DS power > 99% of max
- DS power > 90% of max
- DS power < CS power

Theoretical power of DS for model P as a function of pretest levels for \( N_{pa} = 2000 \) and \( K = 100M \)

- DS power
- CS power
- Max DS power
- DS power > 99% of max
- DS power > 90% of max
- DS power < CS power

Figure 4.2: Theoretical powers as functions of pre-test levels \( \alpha_1 \) (x axes) for various odds ratio (OR) of \( \beta_3 \) for analysis models J, O, P. For details see Figure 4.1.
which CS outperforms $Pco_4$-DS are denoted by maroon colored dashed lines. Additionally, ranges of pre-test levels for which $Pco_4$-DS attains near maximum power, that is power within 1% and 10% of the peak power for given $\delta^*$, are marked as green filled area and dashed green areas, respectively. The plots illustrate the differences between the models in terms of optimal pre-test level $\alpha_1^*$ and optimal pre-test sample size ratio $\delta^*$ as they relate to different combinations of parameters such as interaction effect size, sample sizes, case-control sample ratios, MAFs, etc. These plots, or more precisely the theory behind these plots, can be used to determine the input parameter values in a practical data analysis. As these plots are based on theoretical arguments and rely on quantities that are generally (at least approximately) known to the user, analogous plots can be produced for any other interaction model and any combination of input parameters.

Results of optimization

Looking at the plots in Figures 4.1 and 4.2 reveals several unsurprising patterns of optimality for $\alpha_1$ and $\delta$. First of all, the plots suggest that for the smallest number number of controls (3000), which is the setting closest to a balanced case-control sample, the power of two-stage procedure is maximized when $\delta$ is the smallest and $\alpha_1$ is the biggest. Since the smaller $\delta$ is and the bigger $\alpha_1$ is the less weight is put on the pre-test, meaning that in that case $Pco_4$-DS virtually turns into CS. Unsurprisingly then, for the smallest considered number of controls $Pco_4$-DS is outperformed by CS as indicated by the fact that in that case the red horizontal line is always above the black colored power function of $Pco_4$-DS. On the other hand, as the number of controls increases and the sample becomes more imbalanced, the situation turns in favour of the two-stage procedure. For most of the considered interaction models the upper hand in terms of power is gained by $Pco_4$-DS over CS already for 7000 controls, while the dominance becomes even more persuasive with 11000 controls for all models. The dominance is often even stronger for the smaller values of odds ratios, which is very appealing since those are the ones most difficult to discover. As far as suitable values $\delta$ are concerned, the plots in Figures 4.1 and 4.2 also tell us that the more imbalanced in favour of the controls the sample is, the larger the optimal value $\delta^*$ tends to be. Existence of such effect should not come as a surprise, since the power function of the score statistic in a case-control setup levels off if only one of the sample sizes keeps rising. This leveling off makes it increasingly more fruitful to use those extra controls in the pre-test, as the power of the pre-test statistic does not suffer from such limitation. The consequent increase of power of the pre-test translates into smaller optimal pre-test levels, which, under the rare interaction setup, allows for smaller multiple testing correction factor in the post-test, thus improving the overall power of the two-stage procedure.

Next we discuss the areas of optimality of pre-test levels. For most of the considered settings $Pco_4$-DS outperforms CS and the optimal values of $\alpha_1$ for which this happens seem to fall roughly between $10^{-4}$ and $10^{-8}$, which is the smallest considered pre-test level. However, general claims about points of optimality are quite difficult, since the selected model
and input parameters come into play in complex manner. Additionally, Figures 4.1 and 4.2 show that both $\alpha_1^*$ and $\delta^*$ depend also on the interaction effect odds ratio. Fortunately, the figures also suggest that the power-maximizing combinations do not vary too rapidly for the considered range of odds ratios, and neither does the corresponding power, especially given the logarithmic scale of the axes. This is very good news for practical use of the two-stage method as it suggests that the power performance is not extremely sensitive to the chosen pre-test level and reasonable performance improvement over CS can be expected for a range of different pre-test levels. We must also stress the importance of this behavior for the current method of determining input values of $\alpha_1$ and $\delta$ by maximizing the power function $\Pi$ in (4.1), where a target value of OR$_3$ needs to be selected before such maximization can take place. The observed relatively low sensitivity of the power-maximizing combinations of $\alpha_1^*$ and $\delta^*$ to the choice of OR$_3$, within reason of course, means that relatively broad ranges of values of OR$_3$ result in largely similar values of $\alpha_1^*$ and $\delta^*$, thus lowering the influence of this particular step on the outcome. This allowed us to set more or less reasonable default values in EpiDET for OR$_3$ for each implemented analysis model. Additionally, making a choice for the default analysis model in EpiDET, which we set to be model A, allowed us to remove the need for input from the user and allowing the software to automatically tailor tuning parameters specifically for each locus pair based on the observed allele frequencies. Naturally, in an applied setting with a more detailed knowledge of the data at hand the implementation of the method allows the user to improve on the default choices in order to achieve optimal performance. Nonetheless, model A seems like a reasonable choice if the high interaction effect is expected to lie with the minor alleles at both loci.

4.5 Simulation study: Investigation of type I error control

Next we focus on investigating error rate control by the considered methods. We illustrate the null hypothesis behavior (no interactions) of the valid two-stage methods defined in Section 4.3 for various values of input parameters $\alpha_1$ and $\delta$ (where applicable). The aim is to show that the probability of false rejection is properly controlled at a chosen level $\alpha \in (0, 1)$ by the valid two-stage methods and thus numerically verify the asymptotic theoretical results of Chapter 3. In contrast to that, we show that for the invalid two-stage methods lack such control due to the dependence between the stages.

4.5.1 Simulation setup

The simulation setup used to investigate the error rate behavior of our methods is the following. At the start of each simulation we fixed the minor allele frequencies (MAFs) to be either $f_1 = f_2 = 0.35$ or $f_1 = f_2 = 0.1$. Then we repeatedly simulated genotypes for about 12500 independent loci of an individual using the given value of MAF for each locus. This pro-
4.5 Simulation study: Investigation of type I error control

Error rate comparison: analysis model A, no main effects ($\beta_1 = \beta_2 = 0$), $f_1 = f_2 = 0.35$

Figure 4.3: Type I error rates for methods $PcO_1$-DS, $PcO_1$-AS, $PcO_4$-AS, $PpO_1$-CS, $PpO_4$-CS, $Pco_1$-CS, $Pco_4$-CS and $Pco_1$-CS with MAFs near 0.35 for both loci in every test. Error rates are plotted as functions of pre-test level of significance $\alpha_1$ under analysis model A with 5 different values of pre-test sample size ratio $\delta$ between 0.1 and 0.9 (left to right) and 1000, 3000, 5000 controls (top to bottom) with 500 cases, no main effects ($\beta_1 = \beta_2 = 0$). Each of the 30 plots is based on about 80 million replications.

Error rate comparison: analysis model A for $f_1 = f_2 = 0.1$ with no main effects ($\beta_1 = \beta_2 = 0$)

Figure 4.4: Type I error rates for methods $PcO_1$-DS, $PcO_1$-AS, $PcO_4$-AS, $PpO_1$-CS, $PpO_4$-CS, $Pco_1$-CS, $Pco_4$-CS and $Pco_1$-CS with MAFs near 0.1 for both loci in every test. For details see Figure 4.3.
Two-stage testing for epistasis: Application

MAF setting and each sample size combination the 12500 loci were paired up, which resulted to 500, while the number of controls ranged over the values 1000, 4000 and 10000. For each interaction effect means that the phenotypes were effectively drawn from a Bernoulli distribution in the scope of our simulation, since the nullity of main effects has little effect on the analysis independently of the genotypes. This, however, does not introduce a significant limitation to the scope of our simulation, since the nullity of main effects has little effect on the analysis of the interaction effect. The number of cases in the case-control data sets was always equal to 500, while the number of controls ranged over the values 1000, 4000 and 10000. For each MAF setting and each sample size combination the 12500 loci were paired up, which resulted

Figure 4.5: Type I error rates for methods Pca4-DS, Pca1-DS, Pca4-AS, Pca1-AS, Ppo4-CS, Ppo1-CS, Pca4-CS and Pca1-CS with MAFs near 0.35 for both loci in every test. Error rates are plotted as functions of pre-test level of significance $\alpha_1$ under analysis model A with 5 different values of pre-test sample size ratio $\delta$ between 0.1 and 0.9 (left to right) and 500, 3000, 7000 controls (top to bottom) with 1000 cases, no main effects ($\beta_1 = \beta_2 = 0$). The y-axes in each plot have the same range from 0 to 0.3.

cess was independently repeated until the desired number of cases and controls was reached. Under the null hypothesis, the phenotype of each individual was determined using the logistic regression model with $\beta_0 = 0.01$ (i.e. population prevalence of cases was about 50%), $\beta_1 = \beta_2 = 0$ and most importantly $\beta_3 = 0$. The lack of main effects in addition to no interaction effect means that the phenotypes were effectively drawn from a Bernoulli distribution independently of the genotypes. This, however, does not introduce a significant limitation to the scope of our simulation, since the nullity of main effects has little effect on the analysis of the interaction effect. The number of cases in the case-control data sets was always equal to 500, while the number of controls ranged over the values 1000, 4000 and 10000. For each MAF setting and each sample size combination the 12500 loci were paired up, which resulted
in approximately 80 million pairs of loci for each combination of sample size and MAF. For each pair of loci we calculated the pre-test and post-test \( p \)-values of the eight two-stage methods described in Section 4.3 and the single-stage \( p \)-value by CS. For the two-stage methods that require the choice of \( \delta \), which were the “valid” methods \( Pco_4 \)-DS, \( Pco_1 \)-DS, \( Pco_4 \)-AS, \( Pco_1 \)-AS and the “invalid” methods \( Pco_4 \)-CS and \( Pco_1 \)-CS, the \( p \)-values were calculated using a random split of the data according to five different values of \( \delta \), which ranged over 0.1, 0.3, 0.5, 0.7, 0.9. As a result, we obtained a total of approximately 30 billion \( p \)-values over the 30 combinations of MAFs, sample size and \( \delta \).

In the two-stage tests we combined the pairs of corresponding \( p \)-values using pre-test levels \( \alpha_1 \) ranging in small steps between 0.9 and \( 10^{-4} \) or even \( 10^{-6} \). For each each sample size, \( \delta \) and \( \alpha_1 \) the post-test \( p \)-values were corrected for multiplicity by the actual number of tests performed in the post-test out of the 80 million. The \( p \)-values for the single-stage CS method were corrected for 80 million tests. Such setup was used because it is suitable for investigating independence of pre-test and post-test in the sense that it allows to see whether the overall false rejection rate is independent of the selected pre-test level. In other words, for a given two-stage method to reliably control type I error, we want the associated false rejection rate to be near \( \alpha = 0.05 \) irrespective of the value of the pre-test level \( \alpha_1 \), and ideally also for all parameter settings. Based on the theoretical results concerning independence formulated in Chapter 3, we should be able to observe this kind of behavior for \( Pco_4 \)-DS, \( Pco_1 \)-DS, \( Pco_4 \)-AS, \( Pco_1 \)-AS, \( Ppo_4 \)-CS and \( Ppo_1 \)-CS. As the theoretical results for the latter four methods are of asymptotic nature, it is important to show that our expectations apply in a realistic setting ideally with small and moderate sample sizes under reasonable parameter setting. On the other hand, the expectation is that the false rejection rates of methods \( Pco_4 \)-CS and \( Pco_1 \)-CS are not likely to be controlled by \( \alpha \). As we explained before, this is a consequence of dependence of the control-only pre-test statistics \( T_{n,\delta}^{co} \) and \( R_{n,\delta}^{co} \) and the full-sample score statistic \( \hat{S}_n \), resulting from the (partial) overlap of the control samples on which these statistics are based. The problem should be especially severe if \( \delta \) is large (close to 1), when the dependence between the pre-test and post-test is strongest. Also, we expect the false rejection rate to be most inflated if the case-control sample is almost balanced, and less so if the sample is strongly imbalanced. As the thread of such undesirable behavior of \( Pco_4 \)-CS and \( Pco_1 \)-CS is the main motivation for developing the “valid” two-stage methods, it is important to provide illustration of such behavior.

4.5.2 Results of type I error control simulation

The results of the simulations under the null hypothesis of no interactions are presented in Figures 4.3 and 4.4 for the methods with control-based pre-tests and in Figure 4.5 for the methods with case-based pre-tests.\(^4\) The plots show the dependence of the post-test type I error...
error rate on the pre-test level $\alpha_1$. Effectively, the difference between Figure 4.3 and Figure 4.4 are the values of MAFs used during the simulation of genotypes at each locus, which were 0.35 in Figure 4.3 and 0.1 in Figure 4.4. For the case-based methods we present the type I error results only for a single value of MAF, namely 0.35.

**Behavior of the "valid" methods**

The primary question we want to answer using Figures 4.3 – 4.5 is whether the post-test false rejection probabilities for any of the "valid" methods depend on the the pre-test level of significance $\alpha_1$. Satisfyingly, a general conclusion from the three figures is that the "valid" methods, be they based on either the control-only or the case-only pre-tests, do maintain the post-test type I errors quite well regardless of the value of $\alpha_1$. Although, a closer look at Figures 4.3 and 4.4 might suggest that perhaps for some of the most extreme pre-test levels there is an increased variability of the false rejection rate of the "valid" two-stage methods especially for the smaller MAFs and $P_{co1}$-AS and somewhat surprisingly also for $P_{co1}$-DS. Given that we only observe this behavior in the top row plots in Figure 4.4, it seems likely that this behavior is caused by the relatively low sample sizes underlying those plots, which increase the difficulty of estimation of variance and covariance matrices necessary to calculate both the disjoint and especially the adjusted methods. As soon as the available sample sizes increase, the problem seems to vanish. Moreover, the results must also be considered in light of the number of considered scenarios as well as the range of pre-test levels in question. Since the considered pre-test levels were as small as $10^{-6}$, some fluctuations of the type I error rates are expected, since the numbers of post-tests on which the plotted lines are based are not very large for the most extreme pre-test levels. In fact, under the rare interaction setup together with no background dependence the expected number of tests in the post-test can be approximated by $\alpha_1 K$, where $K$ is the total number of pre-tests equal to 80 million in our simulation. For the most extreme levels this means that the expected number of tests in the post-test phase should be somewhere near $\alpha_1 K = 10^{-6} \cdot 80 \cdot 10^6 = 80$, which indeed is not a very large number. Moreover, based on the simulation, especially in the case of smaller MAFs, this number often even overestimates the number of tests actually passing the pre-test, as with small MAFs it is more likely to have zero cell counts or even marginal counts in the pre-test genotype counts, which can lead to the actual distribution of the pre-test statistics to be less heavy tailed than the used asymptotic approximation by the chi-square distributions. As far as the disjoint sample based methods $P_{co4}$-DS, $P_{co1}$-DS, $P_{ca4}$-DS and $P_{ca1}$-DS are concerned, the independence is guaranteed by the independence of individuals in the full case-control sample. Since the data was simulated in an independent manner, the methods disjoint score methods can be considered a point of reference for the behavior of the other methods, as there is no reason to doubt their ability to control the type I error (especially with

---

based and those for the case-based methods is that they were produced several years apart and each for a different outlet (i.e. journals). The same explanation applies also to the differences between the power plots in Figures 4.6 – 4.7 and Figures 4.10 – 4.11. We hope this satisfies the reader’s possible curiosity.
large sample sizes). Indeed, the validity of such expectation is well supported by the green lines in Figures 4.3 – 4.5, perhaps with the exception of the top row plots in Figure 4.3. In the case of control-only based pre-tests in Figures 4.3 – 4.4, virtually all combinations of control sample sizes and pre-test ratios δ lead to essentially straight green horizontal lines, thus indicating that the rejection rates of both $Pco_4$-DS and $Pco_1$-DS are indeed independent of the pre-test level $\alpha_1$. Although the plots in Figure 4.5 for the case-only based pre-tests show an increased variability of the observed type I error rates, the overall lack of any trend is reassuring. Overall it seems fair to say that the two figures confirm that these methods behave as expected based on the (asymptotic) independence of the two-stages in each of these procedures, especially if the sample sizes are large enough. Naturally, the usual caution is advised when dealing with very small sample sizes and/or very small MAFs.

Behavior of the "invalid" methods

A secondary question we can address based on Figures 4.3 – 4.5 is the issue of inflated rejection rate for the "invalid" methods $Pco_4$-CS, $Pco_1$-CS, $Pca_4$-CS and $Pca_1$-CS, which are based on the combination of non-independent statistics and thus are expected to fail at controlling the overall type I error. Indeed, all three of the figures provide strong evidence of such undesirable behavior. Clearly, the severity of the problem increases with $\delta$, which is expected since larger $\delta$ leads to stronger dependence between the statistics of the two stages.

4.6 Simulation study: Investigation of power

With a large genetic study in mind, in our investigation of the power performance of the various methods we focus on the rare interaction setup described above. Under such setup we assume a large initial number of tests (loci pairs) $K$, which all need to be tested for presence of interactions, but only a small fraction (in our case only one) of these pairs is actually interacting with respect to the given phenotype. Therefore, in the following we consider statistical power of a test not in the classical sense but within a multiple testing setup where power is the probability of detection after appropriate multiple testing correction is performed. For multiple methods with identical multiple testing correction there is a one-to-one correspondence between classical power and MTC-power. However, if one wishes to compare performance of several testing methods each with a different multiple testing correction, MTC-power is a much more relevant measure of performance. Since we only focus on MTC-power here, we simply refer to it as power.

4.6.1 Simulation setup

In this part of the simulation study investigated the same methods as in the error rate investigation (see Section 4.3), although this time we also included the full-sample score method
Two-stage testing for epistasis: Application

CS into the mix. In the rare interaction setup roughly $\alpha_1 K$ pairs of loci are expected to pass the pre-testing phase resulting in a multiple testing factor near $\alpha_1 K$ in the post-test, which is the value we correct for in all of the post-tests in the power analysis of this section. In the results below, in addition to the valid two-stage methods we also included the "invalid" methods $P_{co_4}$-CS and $P_{co_1}$-CS. However, it is essential to keep in mind that the "invalid" methods do not control type I error and the "valid" two-stage methods cannot be reasonably expected to have the same power as $P_{co_4}$-CS and $P_{co_1}$-CS especially for nearly balanced samples with large $\delta$ and small $\alpha_1$. Nonetheless it is interesting to see how close the "valid" two-stage methods come to matching the performance of the "invalid" ones.

Models of interaction

In Section 4.2 we defined several interaction models (see Table 4.1). While in rare practical situations the choice of an interaction model can be clear, in most exploratory analyses the choice is often made somewhat arbitrary. It is therefore relevant to investigate the behavior of our methods under various analysis models and various data generating models, judge each method based on its overall performance under these scenario, and identify those that perform well from such perspective. While the number of possible models is infinite, we must of course limit the investigation only to a few, namely models $A$, $I$, $C$, $J$, $O$ and $P$ defined in Table 4.1. Using each model we simulated numerous data sets, which were subsequently analyzed using the two considered methods with the score tests employing models $A$ and $I$, which are the two models most popular in practice. This pairing of simulation and analysis model leads to situations with various degrees of model misspecification ranging from the setup where the simulation and analysis models are identical (pairings $A \times A$ and $I \times I$ meaning "analysis model $\times$ simulation model"), or slightly misspecified (for instance $A \times O$ or $I \times J$), or grossly misspecified (such as $A \times P$, $A \times C$, $I \times P$, $I \times C$).

Parameter settings

In our study we considered the following parameter settings. We assumed that the total number of pairs of loci $K$ is equal to 100 million, which is the number of tests that corresponds to about 14000 genetic loci all paired up to investigate interactions. While 14000 loci is less than what would be the case in a true genome-wide search, it is of the same order of magnitude as a gene-only GWAS in humans, which makes it a relevant setting for power performance comparison of the methods. In the simulation we fixed the number of cases at 2000 and the number of controls was set to either 3000, 7000 or 11000. In terms of MAFs, we considered two combinations, namely $f_1 = f_2 = 0.35$ and $f_1 = f_2 = 0.1$. The former of the two choices yields a comparison under non-extreme conditions in terms of allele frequencies, while the latter reflects a situation with rather small MAFs, for which the

---

5 A gene-only GWAS data set is such where every gene of the given species is represented by a single locus. Since humans have approximately 20000–25000 protein coding genes, their analysis for interaction would require a number of tests in the order of hundreds of million.
power performance of all methods is generally weaker and estimating some of the necessary variance and covariance matrices is more difficult. The main effects $\beta_1$ and $\beta_2$ at were set to 0. Although the lack of main effects might seem like a special case scenario, in our experience such choice is of minor consequence. In the simulation we observed that the results both with and without main effects were almost indistinguishable, which negates the need to present both of them simultaneously. In the power simulation we considered two different population prevalence scenarios. First, we focused on a setting with high population prevalence by setting $\beta_0 = 0.01$ (prevalence of around 50%) where we used the two-stage methods with control-only pre-tests. Second, we considered a setting with low population prevalence of cases by setting $\beta_0 = -3.3$ (prevalence of around 5%), where we used the case-only pre-test based two-stage methods. In the logistic regression model the degree of deviation from the null hypothesis is expressed through the interaction effect parameter $\beta_3$, where larger absolute value of $\beta_3$ means stronger interaction effect.\(^6\) In the plots on the x-axis we use the associated interaction odds ratio $\text{OR}_3 = \exp(\beta_3)$ to express effect size and we let it range from 1 (no interaction) upwards to $\text{OR}_3 = 3$ in steps of 0.05. For each effect size and a given combination of input parameters we simulated 4000 independent pairs of loci,\(^7\) which were subsequently analyzed by the considered methods.

**Tuning parameters**

Each of the considered two-stage methods requires choice of suitable values of the tuning parameters $\alpha_1$ and $\delta$, which must be set before an analysis can be performed. In Section 4.4 we presented an approach that addressed the problem of how these values can be chosen using a theoretical framework. Since we are currently primarily interested in the best possible power performance of each method, we did not use that approach here. Instead, the analyses were performed using a large number of combinations of values for $\alpha_1$ and $\delta$ and the power maximizing combinations were selected to illustrate the potential of each two-stage testing method. We considered 68 different values for $\alpha_1$ ranging between 0.5 and $10^{-8}$, where the maximum value of 0.5 was chosen instead of 1 only to keep the two-stage methods and CS from becoming indistinguishable as it might have for certain parameter combinations. As a consequence of the rare interaction setting, the corresponding post-test multiple testing correction factors $\alpha_1 K$ ranged between 50 million for the largest pre-test level 0.5 and 1 for the smallest pre-test level $10^{-8}$. Similarly, the power functions for the partial sample methods were maximized over 19 different values of $\delta$ ranging between 0.05 and 0.95 in steps of 0.05.

Naturally, in practice it is difficult to select the optimal values of unknown tuning parameters, although theoretical arguments such as those presented in Section 4.4 can allow the user to come close to optimality. Furthermore, the theoretical arguments allowed us to demон-

\(^6\)It should be noted that the relevant values of $\beta_3$ and $\text{OR}_3$ are strongly linked to the parametrization of the interaction model such as those in Table 4.1, which makes their interpretation on an absolute scale rather pointless. In any case, however, we are mostly interested in a relative comparison of the methods, which is the relevant point of view here.

\(^7\)The number 4000 is effectively the number of replications for each effect size and it is unrelated to the assumed 100 million tests for each pair.
strate that the actual power performance of the two-stage methods are not extremely sensitive to the chosen values of the tuning parameters as discussed in Section 4.4.2. Finally, as one can see below the power performance advantage on the side of the two-stage procedures over the single-stage alternative under many scenarios can be very large. It then stands to reason that as long as the values of the tuning parameters are in a reasonable range, the two-stage procedures should outperform the single-stage alternative.

4.6.2 Results under high prevalence setting

In this section we discuss in depth the results of our power study under the high prevalence setting. Figures 4.6 – 4.7 show the performance of the considered methods under analysis models A or I and the two considered MAF combinations. We also provide an illustration of optimal tuning parameters in Figures 4.8 – 4.9 for both considered MAF combinations. For a comparison of the methods under the low prevalence setting see Section 4.6.3.

Based on Figures 4.6 – 4.7 we can compare the performance of the "valid" two-stage methods against the standard provided by the single-stage full-sample score test CS and also against the "invalid" methods $Pco\_4$-CS and $Pco\_1$-CS. In each figure, the top 18 plots relate to analysis model A, while the bottom 18 plots relate to analysis model I. For each analysis model, we consider the 6 simulations models from Table 4.1, which differ over the columns of the figures. For each combination of analysis and simulation models we consider three different combinations of sample sizes with constant 2000 cases and 3000, 7000 and 11000 controls arranged from top to bottom, respectively. The MTC-power functions of different methods are indicated by color, which are indicated in the legends of each collection of 18 plots. At the top of each plot there are horizontal lines that indicate optimal method or methods for given value of $OR_3$ using the same colors as the power functions. Among these indicators of optimality we do not include the "invalid" two-stage methods $Pco\_4$-CS and $Pco\_1$-CS for the reasons of unreliable type I error rate as discussed at the top of this section. Also, since the results are based on simulation, there is a tolerance for optimality of methods, which is set at 1%. This means that if a ratio of two power functions for a given $OR_3$ is above 99%, the two methods corresponding to the two power functions are considered equally good for that $OR_3$. Additionally, if all of the "valid" methods do not differ by at least 0.5% for given $OR_3$, no optimality lines are plotted for that value of $OR_3$. For each analysis model the columns of the figures are ordered according to the performance by CS relative to the best method under each of the simulation models. The ordering is determined separately for each analysis model based on Figure 4.6, and kept the same in the remaining figures for ease of comparison. The ordering of simulation models according to relative performance by CS can be viewed as indicative of the degree of misspecification the analysis model represents under the given simulation model.

Figures 4.6 and 4.7 show how the power performance for all methods improves as sample size (due to more controls only) increases. This is completely unsurprising as power usually
4.6 Simulation study: Investigation of power

MTC – power functions: analysis model A with \( f_1 = f_2 = 0.35 \) and no main effects (\( \beta_1 = \beta_2 = 0 \)), \( N_{ca} = 2K \), MTC = 100M

Figure 4.6: Maximum empirical MTC-powers for various methods as functions of interaction odds ratio OR\( \beta \) for analysis models A and (top 18 plots) and I (bottom 18 plots) for 6 different data simulation models (left to right). There 2000 cases and 3000, 7000, 11000 controls (top to bottom) with both MAFs fixed at 0.35 for both loci in each test and no main effects are present (\( \beta_1 = \beta_2 = 0 \)). Pre-test levels of significance \( \alpha_1 \) and pre-test sample size ratios \( \delta \) (where applicable) are selected to give the best possible power performance. Each test is assumed to be one of 100 million and is Bonferroni corrected by K for CS or by \( \alpha_1K \) for the eight two-stage methods. Horizontal lines at the top of each plot indicate which tests have the best power performance for the given OR\( \beta \) within each setting with a relative tolerance of 1%.
increases as sample size grows. A more nuanced aspect of this behavior, however, are the relative paces of power improvement of each method when stacked up against the rest. What we notice is that the stronger the imbalance between case and control sample size is, the better all six "valid" two-stage methods perform relative to the single-stage CS. A very interesting conclusion can be drawn from the plots in Figures 4.6 concerning the difference between the two approaches to modification of the score statistic described in Section 3.4 \((Pco_4-DS)\) and \((Pco_1-DS)\) and Section 3.5 \((Pco_4-AS)\) and \((Pco_1-AS)\). It seems that for large MAFs the power functions of \(Pco_4-DS\) and \(Pco_4-AS\), as well as of \(Pco_1-DS\) and \(Pco_1-AS\), are virtually overlapping. This suggests that for moderate or large MAFs the two approaches result in largely equivalent tests. On the other hand, for small MAFs in Figure 4.7, more visible differences between the two approaches can be identified. In addition to this, it also appears that the disjoint tests \(Pco_1-DS\) and \(Pco_4-DS\) yield highest power under analysis model \(A\), while the adjusted tests \(Pco_4-AS\) and \(Pco_1-AS\) catch up to them or perhaps take over their disjoint counterparts under analysis model \(I\). We also notice a shift of power superiority from the single degree test \(Pco_1-DS\) and \(Pco_1-AS\) to the Pearson test based methods \(Pco_4-DS\) and \(Pco_4-AS\) as the analysis model becomes increasingly misspecified.

**Comparison under well-specified model (high prevalence)**

Next we focus on the left-most columns in Figures 4.6 and 4.7, which represent a situation where the analysis models \(A\) and \(I\) are well-specified. In Figure 4.6 we notice that if the sample size of case and controls are close, then the power performance of CS is generally comparable to, and under analysis model \(I\) even better than, the "valid" two-stage methods. Such behavior is not surprising, since under well-specified model with near-balanced samples the score statistic, on which CS is based, is expected to be optimal or near-optimal based on classical statistical theory, as it is in certain sense asymptotically optimal (see Theorem 5.6). However, even for the well-specified analysis model setup, as the number of controls increases, the classical single-stage method starts to lack behind the two-stage methods. This behavior of CS can be explained by recalling that in the score test adding extra controls without also adding cases had diminishing returns in terms of power, while the pre-tests keep benefiting from extra individuals even if they come from only one of the subpopulations.

Under the small minor allele setup of Figure 4.7, where the power functions are much more differentiated, the superiority of (some of) the two-stage methods is still present, with all the two-stage methods outperforming CS for the most imbalanced samples. As can be seen in the left-most column of Figure 4.7, especially the disjoint two-stage method \(Pco_1-DS\) has the upper hand over all of the other methods, including even CS for the most balanced sample sizes. As the sample size imbalance increases, the inferiority of CS increases, while the other two-stage methods show relative improvement.

Next we compare the "valid" method and the "invalid" two-stage methods \(Pco_4-CS\) and \(Pco_4-CS\) under well-specified analysis models and large MAFs. In Figure 4.6 we identify a clear trend of diminishing difference of performance between the "invalid" and "valid"
methods as the imbalance of samples grows. The power functions of the "valid" two-stage methods are more or less similar and they converge to the also similar power functions of the "invalid" methods. With small MAFs in Figure 4.7, the convergence of the power functions is no longer as strongly visible as before, but it can still be identified. In that figure we also notice a more pronounced difference between $P_{co1-}CS$ and $P_{co4-CS}$. For large control counts, the "invalid" method $P_{co4-CS}$ is in most cases caught up by the "valid" methods $P_{co1-DS}$ and $P_{co1-AS}$, and also by the remaining two-stage methods under the more misspecified simulation models.

Comparison under misspecified model (high prevalence)

We turn to the remaining five columns for each of the two analysis models $A$ and $I$ in Figures 4.6 and 4.7, which show the results under misspecified analysis models, where the columns are ordered based on the degree of relative difference between the best "valid" two-stage method and $CS$. Moving through the columns of Figures 4.6 and 4.7 from left to right, we notice that despite the model misspecification the two-stage methods hold their ground and in most cases retain strong power performance. As we would expect, in both Figure 4.6 and Figure 4.7 the inferiority of $CS$ is best visible in the third row (analysis model $A$) and sixth row (analysis model $I$), where the case and control samples are most imbalanced.

The fact that the two-stage procedures behave better under model misspecification can be attributed to the nonparametric nature of the employed pre-tests, which are relatively much less sensitive to which simulation model was used. If the score based post-test statistics were considered alone in a single-stage procedure, they would be just as sensitive or even more sensitive to model misspecification as the unadjusted full-sample score statistic within $CS$. However, when these post-test statistics are combined with the non-parametric pre-tests, the overall power performance suffers much less compared to $CS$. This is because the non-parametric pre-tests are able to significantly reduce the multiple testing correction in the post-tests, thus making it easier for the post-tests to discover the interaction effect even with the analysis model misspecified. This is especially true for the Pearson chi-square test of independence employed by $P_{co4-DS}$ and $P_{co4-AS}$ for $f_1 = f_2 = 0.35$ under both analysis models, where these two methods are clearly the best of the lot. As the chi-square-1 pre-test employed by $P_{co1-DS}$ and $P_{co1-AS}$ is comparably more sensitive to analysis model choice (see function $z(u,v)$ in (3.22)), under model misspecification with $f_1 = f_2 = 0.35$ the two methods $P_{co1-DS}$ and $P_{co1-AS}$ generally perform better than $CS$ on the one hand, but worse than $P_{co4-DS}$ and $P_{co4-AS}$ and $P_{po4-CS}$ on the other, while $P_{po1-CS}$ appears to be possesses the least favourable power performance of the considered two-stage methods.

On the other hand, with $f_1 = f_2 = 0.1$ the situation is somewhat altered. Under mild model misspecification it turns out that the single degree of freedom pre-tests yield overall best performing methods $P_{co1-DS}$ and $P_{co1-AS}$, especially if the sample sizes are strongly imbalanced. It can also be seen that $CS$ is clearly outperformed by the two-stage methods provided the case/control imbalance is large enough. The fact that $P_{co1-DS}$, $P_{co1-AS}$ and
Figure 4.7: Maximum empirical powers for various methods as functions of ORs for analysis models A and (top 18 plots) and I (bottom 18 plots) for 6 different data simulation models (left to right) with both MAFs fixed at 0.1 for both loci in each test. For parameter settings see caption of Figure 4.6.
Simulation study: Investigation of power

Ppo_{1}-CS appear to have the upper hand over Pco_{4}-DS, Pco_{4}-AS and Ppo_{4}-CS is likely a consequence of the nature of the Pearson chisquare test employed by Pco_{4}-DS, Pco_{4}-AS and Ppo_{4}-CS. With small MAFs, it is more likely to have zero counts within the contingency table or even zero observed marginal frequencies, in which case the Pearson test does not perform well, as it essentially "wastes" the available degrees of freedom and the effective level of significance is actually smaller than it would be with large MAFs. This behavior is somewhat negated in the right-most columns of the figures, where the strong nonparametric nature of the Pearson pre-tests in Pco_{4}-DS, Pco_{4}-AS and Ppo_{4}-CS becomes the driving force of the power performance, which results in Pco_{4}-DS, Pco_{4}-AS and Ppo_{4}-CS catching up to, or even overtaking, Pco_{1}-DS, Pco_{1}-AS and Ppo_{1}-CS in terms of power.

Figures 4.6 and 4.7 also illustrate the problems with model \( P \), which we discussed in Section 4.2. We notice that only under analysis model \( A \) with the larger MAFs the detection of interaction occurs for reasonably small values of OR_{3}. In the other three cases, that is with small MAFs under model \( A \) and with both MAF combination under analysis model \( I \), the performance of any of the methods is quite abysmal. Despite the large value of OR_{3} required for detection in those cases, even under model \( P \) there is a noticeably large relative difference in performance between the two-stage methods and the single step CS. This provides further evidence for the superiority of the two-stage methods.

Optimal tuning parameters \( \alpha^*_1 \) and \( \delta^* \) under high prevalence

Besides the power performance comparison, we focused on ways of determining suitable values of the tuning parameters. In Figures 4.8 and 4.9 we plotted the optimal values for \( \alpha_1 \) and \( \delta \) that correspond to the power plots in Figures 4.6 and 4.7. We point out that the plotted lines in the two figures are smoothed versions of the actually observed values.\(^8\) We smoothed out the observed functions in order to emphasize the observed trends. However, we must stress that we paid close attention to preserving the relevant patterns in the raw results.

Figure 4.8 shows the power maximizing pre-test levels \( \alpha^*_1 \) for all eight two-stage methods including the "invalid" methods Pco_{4}-CS and Pco_{1}-CS. First we focus on the optimal pre-test levels in the case of large MAFs. Going through the columns of plots for MAFs equal to 0.35, we immediately notice how the optimal value \( \alpha^*_1 \) tends to become smaller as the analysis model becomes more misspecified. In most plots, we observe a U-shaped dependence of \( \alpha^*_1 \) on OR_{3}, where the depth of the valley increases with severity of model misspecification. Such behavior can again be attributed to the corresponding increasingly poor performance of the score test and the resulting shift of balance towards the pre-test. In such cases, the pre-tests must strongly decrease the multiple testing correction factor in the post-test, as it progressively becomes the only way for the score tests to identify the interaction under increasingly misspecified analysis model. Consequently, with MAFs equal to 0.35, for both analysis models the optimal values of \( \alpha^*_1 \) shift from relatively moderate towards rather ex-

\(^8\)The smoothing was done using the R function \texttt{smooth.spline()} with parameter \texttt{spar=0.75}.
Figure 4.8: Observed optimal (power-maximizing) pre-test levels of significance $\alpha_0^*$ as functions of OR, for analysis models A and I for 6 different data simulation models (left to right) with no main effects and both MAFs fixed at 0.35 (top 36 plots) and 0.1 (bottom 36 plots) in each test. For parameter settings see caption of Figure 4.6.
treme. For instance, the optimal $\alpha^*_1$ for $Ppo_4$-CS in the left-most columns is between 0.5 and roughly $10^{-3}$ depending on control counts, whereas in the right-most column the optimal $\alpha^*_1$ ranges between 0.5 and $10^{-8}$, the minimum considered pre-test level.

For the other two-stage methods such decreasing trend in $\alpha^*_1$ accelerated, which means that the burden of performance shifted faster towards the pre-testing phase for these methods. This is quite intuitive for two reasons. Firstly, the full-sample post-tests used by $Ppo_4$-CS and $Ppo_1$-CS are more powerful than those of the other two-stage methods (except for the "invalid" methods). This is because the adjusted post-tests in $Pco_4$-AS and $Pco_1$-AS, despite using the full sample, appear to be weakened by the regression, while the post-tests in $Pco_4$-DS and $Pco_1$-DS use only partial sample. The second reason is that the pooled pre-tests in $Ppo_4$-CS and $Ppo_1$-CS tends to be inferior to the single sample pre-tests employed by the rest of the two-stage methods, apparently due to some of the dependence of two-locus genotypes induced by the presence of interaction being masked away in the pooled pre-test. We additionally notice that in most plots in Figure 4.8 that correspond to MAFs of 0.35, there is only a small difference between optimal pre-test levels of the corresponding adjusted and disjoint test. This provides further evidence that the two approaches to accounting for pre-testing are largely equivalent.

Turning to the results for the case of small MAFs, which are captured in the bottom 36 plots in Figure 4.8, we notice a more volatile dependence of $\alpha^*_1$ on OR. While in most of these plots we can still notice a kind of U-shaped dependence, it is generally less pronounced in most cases, and in some cases it is no longer present. Similarly to power plots in Figure 4.7, we notice a more apparent difference between the disjoint and adjusted methods, as in some cases the green and blue lines no longer coincide with each other so well. Nonetheless the optimal values $\alpha^*_1$ for the two approaches are still similar, especially for the case of the Pearson pre-test based methods $Pco_4$-DS and $Pco_4$-AS.

Finally, we focus on the optimal pre-test sample size ratios $\delta$. The observed optimal values $\delta^*$ for analysis model $A$ and $I$ and both MAF combination can be found Figure 4.9. A closer look at the plots reveals a complementary pattern to that which we noticed in Figure 4.8. It turns out that whenever a relatively strict pre-test level is optimal, it tends to be accompanied by a large value of $\delta^*$. This is particularly visible in the plots corresponding to MAFs equal to 0.35, as evidenced by the top 36 plots in Figure 4.9, where the observed patterns for $\delta^*$ are almost perfect upside-down mirroring of the patterns for $\alpha^*_1$ in the top 36 plots in Figure 4.8. Such inverse relationship between $\alpha^*_1$ and $\delta^*$ is intuitively quite clear, since if the burden of power lies with the pre-test, then it should perform well in the sense of reducing the multiple testing correction in the post-test only with relatively strict pre-test levels. In other words, if a strict pre-test level is used, it should be accompanied by a large sample size, thus large values of $\delta$ are optimal. Complementarily, if a relatively large value of $\alpha_1$ is optimal in terms of power, then the potential for reduced multiple testing correction achieved by the pre-test is small, as it is expected that $\alpha_1 K$ or more tests pass the pre-test. In that case it is probably not desirable to use a large portion of the available sample in the pre-test, as both
Figure 4.9: Observed optimal (power-maximizing) pre-test sample size ratios $\delta^*$ as functions of OR for analysis models A and I for 6 different data simulation models (left to right) with no main effects and both MAFs fixed 0.35 (top 36 plots) and 0.1 (bottom 36 plots) in each test. For parameter settings see caption of Figure 4.6.
the adjusted and the disjoint post-tests in $Pc0_4$-AS, $Pc0_1$-AS, $Pc0_4$-DS and $Pc0_1$-DS must pay for it in terms of power (compared with CS).

4.6.3 Results under low prevalence setting

Figures 4.10 – 4.11 show the results. The total of 63 power plots provides a substantial insight into how the various methods stack up against each other and against the classical single-stage test for different case-control ratios and MAF combinations under various degrees of misspecified interaction model during the analysis ranging from well-specified to strongly misspecified. Without going into as much detail as in the high prevalence setting, it is crucial to point out the even more pronounced role of model misspecification as a differentiating factor between the methods. Unlike in the high prevalence setting, with well-specified interaction model the disjoint and adjusted score based methods are outperformed by the pooled-sample two-stage methods $Ppo_4$-CS and $Ppo_1$-CS especially when the data imbalance is low. Moreover, the low prevalence setting differentiates also between the two types of pre-tests, namely the Pearson test of independence (employed by $T_{\text{co},a}^{\alpha,\delta}$ and $T_{\text{ca},a}^{\alpha,\delta}$) and the independence trend test (employed by $R_{\text{co},a}^{\alpha,\delta}$ and $R_{\text{ca},a}^{\alpha,\delta}$). This is to say that in the low prevalence setting model misspecification alters the relative performance of the methods and that the parametric single-stage CS test and the parametric two-stage methods which require specifying $z(x,y)$ in both stages (i.e. $Pca_1$-AS, $Pca_1$-DS, $Ppo_1$-CS) suffer more pronounced decrease of power due to model misspecification compared to the other methods (i.e. $Pca_4$-AS, $Pca_4$-DS, $Ppo_4$-CS).

It is also extremely noteworthy that under model misspecification (with fixed population prevalence and fixed allele frequencies) not all methods benefit from an overall larger sample size. This is particularly evident for the pooled-sample methods $Ppo_4$-CS and $Ppo_1$-CS, where under all scenarios with misspecified model the two methods in fact lose power when the number of controls is increased. Although this might seem somewhat counterintuitive, it is a direct consequence of the pooled-sample tests losing power as the fraction of cases in the sample approaches the prevalence of cases in the population. In the three considered case-control ratios in the simulation, by keeping the number of cases fixed and increasing only the number of controls the sample prevalence of cases approaches the population prevalence (around 5%), which results in a vast loss of overall power of the two-stage procedures based on the pooled-sample test of independence. This effect is most pronounced under the high MAF scenario in Figure 4.10, but it is present in Figure 4.11 as well. On the other hand, the partially non-parametric $Pca_4$-DS and $Pca_4$-AS very desirably benefit from the increased sample size. Moreover, not only are $Pca_4$-DS and $Pca_4$-AS positively affected by the increased sample size, they are the top performers overall in many of the model misspecification scenarios and their advantage over the competition increases with increased imbalance towards the controls (e.g. simulation models $C$, $J$, $P$). This is the case for both $Pca_4$-DS and $Pca_4$-AS with high MAF scenario of Figure 4.10, while for the low MAF scenario in Figure 4.11 it is $Pca_4$-DS that takes the upper hand over $Pca_4$-AS. Given that both $Pca_4$-DS and
Two-stage testing for epistasis: Application

$Pca_{4}$-AS yield reasonable performance compared to the other methods also under the correctly specified interaction model, it seems safe to say that the two methods, and especially $Pca_{4}$-DS, are very strong performers that overall can be seen as superior to the other considered methods. As for the comparison among $Pca_{4}$-DS and $Pca_{1}$-DS versus $Pca_{4}$-AS and $Pca_{1}$-AS it seems fair to say that the disjoint tests ($Pca_{4}$-DS and $Pca_{1}$-DS) generally outperform the adjusted tests ($Pca_{4}$-AS and $Pca_{1}$-AS). In terms of power, under many considered scenarios the two methods are equivalent, however, under several scenarios the disjoint test was clearly superior. We also observed that the adjusted score test is more susceptible to numerical instability problems especially for very small allele frequencies. This is likely due to the more involved nature of the statistic, which requires an estimation of higher number of variance and covariance matrices.

4.6.4 Simulation study: Conclusions

As we showed in Section 4.5, all of the these methods showed sufficient type I error control. Our investigation also showed that it can be extremely risky in terms of false rejection rate to combine dependent tests into a two-stage testing procedure. Since proper type I error control without good power is not very useful, we put the two-stage methods to further scrutiny in Section 4.6, where we stacked them up against the full-sample score test and the two "invalid" two-stage methods. The results of that section show that under many simulation scenarios almost all of the methods have the potential to provide excellent power performance if the minor allele frequencies of the loci are moderate or large. This holds both for the two-stage methods with control-only based pre-tests in the high prevalence setting and for the two-stage methods with case-only based pre-tests in the low prevalence setting. Additionally, even under small minor allele frequencies many of the two-stage methods perform well. As is typically the case with two stage methods, their performance is influenced by the user’s ability to specify suitable values for their tuning parameters. In Sections 4.4 and 4.6 we addressed this problem and formulated a theory-based method to select these values specifically for the data at hand in a way that promises good performance. We also implemented automatic tuning parameter selection within EpiDET, which makes performing data analysis via the two-stage testing methods quite straightforward and effortless.

Our investigation also included a comparison of the performance of methods that differ in the way that they achieve independence between the pre-test and post-test statistics. While the disjoint testing approach is very simple, it is surprisingly robust and powerful. We also observed that the second simplest methods, namely the pooled-sample pre-test based methods, yielded good performance. This was more the case in the low prevalence setting and balanced data sets, while for the imbalanced data and/or high prevalence setting the methods performed noticeably poorer. Finally, the conceptually most involved methods which are based on the adjusted score statistic performed in most cases identically to the disjoint approach. Moreover, the disjoint methods have several advantages on their side compared
92 4.6 Simulation study: Investigation of power

Figure 4.10: Empirical MTC-powers at optimal input parameters \(\alpha_1\) and \(\delta\) for various methods as functions of interaction odds ratio \(\text{OR}_3 = \log \beta_3\) for simulation models A,C,I,O,P with analysis model A, 2000 cases and 3000, 7000, 11000 control (denoted as \(m_u\)) with MAF close to 0.35 for both loci in each test, phenotype prevalence of 5\% and \(\beta_1 = \beta_2 = 0\). (Note the extreme closeness of the power functions for \(\text{Pca}_4\)-DS (bright green) and \(\text{Pca}_4\)-AS (purple) and for \(\text{Pca}_1\)-DS (dark green) and \(\text{Pca}_1\)-AS (dark blue) in several plots, where the power functions of \(\text{Pca}_4\)-AS and \(\text{Pca}_1\)-AS essentially disappear behind the power functions of \(\text{Pca}_4\)-DS and \(\text{Pca}_1\)-DS, respectively.)
Figure 4.11: Empirical MTC-powers at optimal input parameters $\alpha_1$ and $\delta$ for various methods as functions of interaction odds ratio $OR_3 = \log \beta_3$ for simulation models A,C,I,J,O,P with analysis model A, 2000 cases and 3000, 7000, 11000 control (denoted as $m_0$) with MAF close to 0.1 for both loci in each test, phenotype prevalence of 5% and $\beta_1 = \beta_2 = 0$. 

MTC – power functions ($t_1 - t_2 = 0.1$, $\beta_1 = \beta_2 = 0$, analysis model A, assumed $N = 10^8$ tests, S2 correction $\alpha_1(N)$)
4.7 Real data analysis: Parkinson’s disease study

We applied the two-stage methods to the analysis of real data relating to Parkinson’s disease (PD).

Data

For this analysis four independent PD cohorts from the International Parkinson’s Disease Genomics Consortium (IPDGC) have been used (Table 4.2). These cohorts and the quality control procedures applied to the data have been described in detail elsewhere ([104], [69], [70]). From the available genotype data we extracted those SNPs that could be assigned to protein coding genes. This resulted in 11382 (USA-NIA), 13484 (NL), 13116 (GE) and 10602 (NINDS-CIDR) gene based SNPs to be tested for gene-gene interactions on PD status. The main numerical characteristics of the four data sets are provided in Table 4.2.

Method of analysis

For the analysis of the four Parkinson’s disease data we represented the genotypes numerically by counting the rarer allele at each SNP. Out of the desire to compare the studied methods in terms of real data performance, we deployed six two-stage methods, namely $Pca_4$-
Two-stage testing for epistasis: Application

$DS$, $Pca_1$-$DS$, $Pca_4$-$AS$, $Pca_1$-$AS$, $Ppo_4$-$CS$, $Ppo_1$-$CS$, where $Pca_4$-$DS$, $Pca_1$-$DS$, $Pca_4$-$AS$ and $Pca_1$-$AS$ are the case-only based pre-test analogues of $Pco_4$-$DS$, $Pco_1$-$DS$, $Pco_4$-$AS$ and $Pco_1$-$AS$, respectively. In order to limit the influence of possible population differences among the four data sets, instead of merging the data sets and analyzing the combined data, we performed the S1 tests for each SNP-pair separately in each data set where the given pair was present using 70% of the available case in each data set (i.e. $\delta = 0.7$). This resulted in up to four S1 $p$-values for each SNP-pair, which we subsequently combined using a Fisher-type weighted $p$-value combination method (Fisher (1932), Box (1954)). For a general set of $k$ $p$-values $p_1, \ldots, p_k$ the combination method utilizes the statistic $F = -2 \sum_{i=1}^{k} w_i \log p_i$, where the weights $w_i$ are based directly on the sample sizes underlying each $p$-value. The null hypothesis distribution of $F$ is given by Theorem 2.4 of Box (1954). This yielded a single combined S1 $p$-value for each SNP-pair, which we compared with a pre-set level for each method. Based on the observed optimality of S1 levels in the simulation we selected $\alpha_1 = 10^{-6}$ for $Pca_4$-$DS$, $Pca_1$-$DS$, $Pca_4$-$AS$ and $Pca_1$-$AS$, while for $Ppo_4$-$CS$, $Ppo_1$-$CS$ we used $\alpha = 10^{-4}$. For each of the methods we performed the S2 tests for those SNP-pairs that passed the corresponding S1 tests. The S2 tests were again performed separately for each data set (where present) and the resulting $p$-values for each method were again combined using the weighted $p$-value combination method with the weights based on the harmonic means of the case and control counts used in each S2 test, since the harmonic mean gives the effective rate of convergence of the score statistic in the LRM (Foppa and Spiegelman (1997)). Given that the resulting combined $p$-values in S1 were independent of the combined $p$-values in S2, for proper type I error control it is sufficient to correct the combined S2 $p$-values only by the number of tests in S2. In addition to the two-stage methods, we also calculated the single-stage score test $p$-values in each cohort, which we combined using the same $p$-value combination method. Given that there were 92,250,923 tests in total, the resulting combined $p$-values were compared with the Bonferroni corrected level $5.4 \cdot 10^{-10}$.

Results

Based on the multiple testing corrected combined S2 $p$-values we identified several SNP pairs to be genome-wide significant, while the single-stage tests yielded none. Aiming at maximizing the unambiguity of the results, we took advantage of the availability of multiple cohorts and report only the SNP pairs for which the significant combined S2 test was nominally supported by more than one cohort. In light of the availability of data for multiple cohorts it would have been possible to attempt replication directly by reserving a single data set. However, given that majority of the replication data set would not have been used at all, attempting replication would have come at a relatively high cost in terms of power. Therefore, we opted for a single analysis but present only the multi-cohort supported findings. While technically this does not amount to a replication of our findings, we suggest that it increases their credibility substantially. Note that a significant S1 test was not required to be supported by multiple cohorts, which is consistent with the rationale for the pre-tests as providing pri-
4.8 Real data analysis: Parkinson’s disease study

<table>
<thead>
<tr>
<th>GENE1</th>
<th>Chr1:RS1</th>
<th>GENE2</th>
<th>Chr2:RS2</th>
<th>$T_n^b$</th>
<th>$A_n$</th>
<th>$D_n$</th>
<th>$C_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBE2J1</td>
<td>6:rs17798549</td>
<td>GPR107</td>
<td>9:rs4837460</td>
<td>8.27e-07</td>
<td>1.34e-02</td>
<td>5.97e-06</td>
<td>8.65e-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$T_n^b$ (NL)</td>
<td>$A_n$ (NL)</td>
<td>$D_n$ (NL)</td>
<td>$C_n$ (NL)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.32e-04</td>
<td>8.59e-02</td>
<td>2.60e-03</td>
<td>4.48e-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$T_n^b$ (GE)</td>
<td>$A_n$ (GE)</td>
<td>$D_n$ (GE)</td>
<td>$C_n$ (GE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.60e-05</td>
<td>1.89e-02</td>
<td>4.56e-05</td>
<td>2.44e-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$T_1^e$ (NL)</td>
<td>$t_2$ (NL)</td>
<td>$t_1$ (GE)</td>
<td>$t_2$ (GE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.043</td>
<td>0.011</td>
<td>0.035</td>
<td>0.023</td>
</tr>
<tr>
<td>DUSP12</td>
<td>1:rs1063179</td>
<td>DOCK4</td>
<td>7:rs1270579</td>
<td>8.30e-07</td>
<td>3.59e-06</td>
<td>1.30e-01</td>
<td>8.88e-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$T_n^b$ (NL)</td>
<td>$A_n$ (NL)</td>
<td>$D_n$ (NL)</td>
<td>$C_n$ (NL)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.11e-08</td>
<td>3.45e-02</td>
<td>4.54e-02</td>
<td>9.17e-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$T_n^b$ (GE)</td>
<td>$A_n$ (GE)</td>
<td>$D_n$ (GE)</td>
<td>$C_n$ (GE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.75e-01</td>
<td>2.78e-06</td>
<td>9.82e-01</td>
<td>5.79e-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$t_1$ (NL)</td>
<td>$t_2$ (NL)</td>
<td>$t_1$ (GE)</td>
<td>$t_2$ (GE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
<td>0.062</td>
<td>0.05</td>
<td>0.064</td>
</tr>
</tbody>
</table>

Table 4.3: Results of the analysis of PD data sets. The table shows both combined and single-cohort $p$-values and MAFs for two pairs of genes that were identified for interaction using $Pcα1\text{-DS}$ (first pair) and $Pcα1\text{-AS}$ (second pair) with $δ = 0.7$ and $α_1 = 10^{-6}$. The implicating $p$-values for each pair are shown in bold.

marily circumstantial evidence that guides the verification stage analyses. Any reasonable a priori suggestion of a genetic interaction would then be formally tested in the S2 stage, hence nominal support from multiple sources was only required for this stage.

Our analysis identified two SNP pairs with multi-cohort evidence of epistasis. For both pairs the multi-cohort evidence for interaction comes from the Dutch (NL) and German (GE) cohorts, while the other two cohorts did not contain data for either one or both of the incriminated SNPs in each pair. Given that all SNPs in the analysis were gene based, our results yield evidence for genome-wide significant gene-gene interaction for the risk to develop PD (Table 4.3). The two identified gene pairs are DUSP12 in combination with DOCK4 and UBE2J1 in combination with GPR107. Interestingly, at least three out of these four genes have biological functions that are closely related to the pathogenesis of PD. DUSP12, also known as hYVH1, is a dual-specificity phosphatase that was shown to physically interact with Hsp70 in order to prevent heat-shock induced cell death (Sharda et al. (2009)). Hsp70 in turn has been shown to affect PD pathogenesis by affecting aberrant alpha-synuclein aggregation (Gao et al. (2015), Zhang and Cheng (2014)). DOCK4 is known to regulate neurite differentiation by activation of Rac1 (Xiao et al. (2013)). Rac1 in turn was shown to rescue neurite retraction caused by G2019S LRRK2, a well-known pathogenic mutation causing familiar PD (Chan et al. (2011)). UBE2J1, also known as UBC6, is a member of the Parkin-Ubiquitin Proteasomal System pathway (see NCBI - BioSystems) and directly interacts with Parkin (Mengesdorf et al. (2002)), another well-known PD gene. Finally, GPR107 is a G-protein coupled receptor. While a different G-protein coupled receptor gene, namely GPR37, is known to be a risk gene with respect to PD, a direct connection between GPR107 and PD is unclear at this point.
4 Two-stage testing for epistasis: Application

4.8 Usage of EpiDET

Our purpose built software tool EpiDET is in usage very similar to PLINK (Purcell et al. (2007)). A number of command line flags can be used to modify its behavior. Here we present a sample of the most relevant flags in order to give the reader an impression of usability of the tool. We provide a list of flags and very brief description of its functionality. We divide the flags into several categories and present each category in a separate table. If a flag is followed by a capital letter (e.g. X) it means that it requires a value to be specified.

<table>
<thead>
<tr>
<th>Flags</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--cmdfile X</td>
<td>Specifies input file with a set of commands (any below)</td>
</tr>
<tr>
<td>--bfile X, --file X</td>
<td>Specifies input files X (in PLINK binary format)</td>
</tr>
<tr>
<td>--bed X, --bim Y, --fam X</td>
<td>Specifies input files X in PLINK plain text format)</td>
</tr>
<tr>
<td>--ped X, --map Y</td>
<td>Specifies input files X in PLINK plain text format)</td>
</tr>
<tr>
<td>--ped-list X, --map-list Y</td>
<td>Multiple plain text input files can be specified via list files</td>
</tr>
<tr>
<td>--ps X</td>
<td>Specifies the pre-test subsample output file</td>
</tr>
<tr>
<td>--inc X</td>
<td>Specifies the pre-test subsample inclusion input file</td>
</tr>
<tr>
<td>--submap-include X</td>
<td>Specifies a file with a list of markers to be included in analysis</td>
</tr>
<tr>
<td>--submap-exclude X</td>
<td>Specifies a file with a list of markers to be excluded from analysis</td>
</tr>
<tr>
<td>--simulate-input</td>
<td>Instructs to simulate input data (further modifiable via other flags)</td>
</tr>
<tr>
<td>--simulation-model</td>
<td>Specifies the interaction penetrance model for simulation</td>
</tr>
<tr>
<td>--markers</td>
<td>Specifies the simulated number of markers</td>
</tr>
<tr>
<td>--prev,--OR1,--OR2</td>
<td>Specifies the simulation prevalence and main effects (odds ratios)</td>
</tr>
<tr>
<td>--HWE,--no-HWE</td>
<td>Specifies whether simulated markers are in HWE</td>
</tr>
<tr>
<td>--maf,--maf1,--maf2</td>
<td>Specifies simulation minor allele frequencies</td>
</tr>
<tr>
<td>--OR</td>
<td>Specifies the simulation interaction effect (odds ratio)</td>
</tr>
<tr>
<td>--OR-min X,--OR-max Y</td>
<td>Simulate multiple samples with OR ranging from X to Y ... ... in steps of size Z</td>
</tr>
<tr>
<td>--OR-step Z</td>
<td>Amount of LD for marker pairs in simulation</td>
</tr>
<tr>
<td>--LD X</td>
<td>Specifies how many marker pairs are in LD</td>
</tr>
<tr>
<td>nLDpairs X</td>
<td>Specifies the input/simulation sample sizes</td>
</tr>
<tr>
<td>--save-binary,--save-plain</td>
<td>Saves input data into a file</td>
</tr>
<tr>
<td>--save-snp-major</td>
<td>Saves input data in a binary or plain text format</td>
</tr>
<tr>
<td>--save- indiv-major</td>
<td>Data saved in SNP-major format (binary only)</td>
</tr>
<tr>
<td>--save-maf</td>
<td>MAFs are calculated and saved into a separate file</td>
</tr>
<tr>
<td>--out X</td>
<td>Specifies the name of the output file</td>
</tr>
<tr>
<td>--out-zip</td>
<td>Results file(s) is compressed</td>
</tr>
<tr>
<td>--out-interaction</td>
<td>Interaction effect estimates are outputted in results file</td>
</tr>
<tr>
<td>--out-maf</td>
<td>Whole sample MAFs are outputted in results file</td>
</tr>
<tr>
<td>--out-snp-coca</td>
<td>Subsample MAFs are outputted in results file</td>
</tr>
<tr>
<td>--out-minimalistic</td>
<td>Only p-values and location info are outputted in results file</td>
</tr>
<tr>
<td>--out-errors</td>
<td>Erroneous results are outputted in results file</td>
</tr>
<tr>
<td>--hide-errors</td>
<td>Erroneous results are not outputted in results file</td>
</tr>
<tr>
<td>--compress-pval</td>
<td>p-values are reported in a shorthand format</td>
</tr>
<tr>
<td>--comment X</td>
<td>Sets the input and output comment character</td>
</tr>
<tr>
<td>--status-case X</td>
<td>Set the input case status character (length 1)</td>
</tr>
<tr>
<td>--status-control X</td>
<td>Set the input control status character (length 1)</td>
</tr>
<tr>
<td>--missing-geno X</td>
<td>Sets the input missing genotype character (length 1)</td>
</tr>
<tr>
<td>--missing-status X</td>
<td>Sets the input missing case/control status character (length 1)</td>
</tr>
</tbody>
</table>
EPIDET

--missing-sex X
Sets the input missing sex character (length 1)

--nsamples X
X equal sized samples expected on input

--ndig-stat X
Number of significant digits for statistics on output

--ndig-pval X
Number of significant digits for p-values on output

--ndig-mafs X
Number of significant digits for MAFs on output

--outreport1
Output report bound for stage 1 p-values

--outreport2
Output report bound for stage 2 p-values (raw)

--outreport3
Output report bound for stage 2 p-values (corrected)

**ANALYSIS**

--do-test
Performs interaction analyses

--no-test
No interaction analyses are performed (data processing only)

--no-pretest
Disables stage 1 tests for two-stage methods

--do-pretest
Enables stage 1 tests for two-stage methods

--pretest X
Selects the test for stage 1 analyses

--analysis-model X
Sets the interaction penetrance model for analysis

--random-subsample
Stage 1 subsample is selected randomly for each test

--fixed-subsample
Stage 1 subsample are fixed the same for all tests

--reject-all
Stage 2 tests are performed for all marker pairs

--only-pretest
Only stage 1 tests are performed

--level1
Sets the rejection threshold for stage 1 tests

--auto-level
Stage 1 rejection threshold is determined automatically

--test-same-chr
Markers located on the same chromosome are also analyzed

--skip-same-chr
Markers located on the same chromosome are not analyzed

--report-AS
Adjusted score (AS) two-stage tests are performed

--no-centering
AS: regression without centering

--center-by-self
AS: regression centering by the same subpopulation

--center-by-other
AS: regression centering by the other subpopulation

--center-by-all
AS: regression centering by the all individuals

--report-DS
Disjoint score (DS) two-stage tests are performed

--report-PO
Pooled-sample pre-test two-stage tests are performed

--report-CS
Single-stage score tests are performed

--all-pretests
All implemented two-stage analyses are performed

--delta X
Sets the stage 1 sample portion (for all tests)

--delta-AS X
Sets the stage 1 sample portion for AS

--delta-DS X
Sets the stage 1 sample portion for DS

--delta-min X
Range the stage 1 sample portion from X

--delta-max Y
... to Y ...

--delta-step Z
... in steps of size Z

--correct-cell-count
Stage 1 contingency table cell count correction

--grouped-variance
Score test variance matrix estimated using all individuals

--non-grouped-variance
Score test variance matrix estimated using only pre-test individuals

--grouped-covariance
Score test covariance matrix estimated using all individuals

--non-grouped-covariance
Score test covariance matrix estimated using only pre-test individuals

--exclude-males
Males are excluded during analysis

--exclude-females
Females are excluded during analysis

--min-var
Lower bound for the score test variance estimate

--min-samplesize X
Minimum sample for each test (otherwise error)

--min-controls X
Minimum number of controls for each test (otherwise error)

--min-cases X
Minimum number of cases for each test (otherwise error)

--min-maf X
Minimum MAF for each test (otherwise error)

--min-margin-count X
Minimum marginal count for stage 1 tests

--min-cell-count X
Minimum cell count for stage 1 tests

--cell-correction X
Cell count correction if too low

--min-cell-correct-all
Determines how cell count correction is applied
4 Two-stage testing for epistasis: Application

MISCELLANEOUS SETTINGS

--random-seed X
Sets the random seed

--nthreads X
Sets the number of parallel threads for the analysis

--max-runtime X
Limits the total runtime

--max-ntests X
Limits the maximum number of tests

--max-out-size X
Sets the output file size limit

--temp-cycle X
Determines how often intermediary results are saved to file

As an example of the usage of EpiDet in Figures 4.12–4.13 we provide two screenshots of the program output obtained during the analysis of a Dutch PD cohort, which was performed via the call

```
./EpiDetector --bfile PD-Dutch --out PD-Dutch --method DS1 --model A --S1-sample cases --test-same-chr --level1 0.000001 --delta 0.7 --out-minimal --out-maf --nthreads 16
```

Through this call we selected the method Pca1-DS and model A, we enabled the analysis of SNP pairs with both SNPs located on the same chromosome and used 70% cases during S1 with a p-value threshold $\alpha_1 = 10^{-6}$. The analysis of the 90 million SNP pairs was parallelized over 16 threads and took about 2 hours. From the results summary it is clear that the Dutch PD cohort does not yield any genome-wide significant results. However, that changes when the results for the Dutch cohort are combined with the results for the other three cohorts as discussed in Section 4.7.

Acknowledgements

We wish to thank all of the subjects who donated time and biological samples to be a part of the Parkinson’s disease study. A full list of the IPDGC members and their affiliations follows.

IPDGC members and affiliations

Mike A Nalls (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA), Vincent Plagnol (UCL Genetics Institute, London, UK), Dena G Hernandez (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Manu Sharma (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Una-Marie Sheerin (Department of Molecular Neuroscience, UCL Institute of Neurology), Mohamad Saad (INSERM U563, CPTP, Toulouse, France; and Paul Sabatier University, Toulouse, France), Javier Simón-Sánchez (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Claudia Schulte (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Suzanne Lesage (INSERM, UMR_S975 [formerly UMR_S679], Paris, France; Université Pierre et Marie Curie-Paris, Centre de Recherche de l’Institut du Cerveau et de la Moelle épinière, Paris, France; and CNRS, Paris, France), Sigurlaug Sveinbjörnsdóttir (Department of Neurology, Landspítl University Hospital, Reykjavík, Iceland; Department of Neurology, MEHT Broomfield Hospital, Chelmsford, Essex, UK; and Queen Mary College, University of London, London, UK), Sampath
Figure 4.12: A sample screenshot of EPIDET obtained during the analysis of a the Dutch PD cohort (A).
Figure 4.13: A sample screenshot of EPIDET obtained during the analysis of the Dutch PD cohort (B).
4.8 Usage of EPIDET

Arepalli (Laboratory of Neurogenetics, National Institute on Aging), Roger Barker (Department of Neurology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK), Yoav Ben-Shlomo (School of Social and Community Medicine, University of Bristol), Henk W Berendse (Department of Neurology and Alzheimer Center, VU University Medical Center), Daniela Berg (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and DZNE, German Center for Neurodegenerative diseases), Kailash Bhatia (Department of Motor Neuroscience, UCL Institute of Neurology), Rob M A de Bie (Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands), Alessandro Biffi (Center for Human Genetic Research and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA), Bas Bloem (Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Zoltan Bochdanovits (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Michael Bonin (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen, Tübingen, Germany), Jose M Bras (Department of Molecular Neuroscience, UCL Institute of Neurology), Kathrin Brockmann (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and DZNE, German Center for Neurodegenerative diseases), Janet Brooks (Laboratory of Neurogenetics, National Institute on Aging), David J Burn (Newcastle University Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne, UK), Elisa Majounie (Laboratory of Neurogenetics, National Institute on Aging), Gavin Charlesworth (Department of Molecular Neuroscience, UCL Institute of Neurology), Codrin Lungu (National Institutes of Health Parkinson Clinic, NINDS, National Institutes of Health), Honglei Chen (Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, NC, USA), Patrick F Chinnery (Neurology M4104, The Medical School, Framlington Place, Newcastle upon Tyne, UK), Sean Chong (Laboratory of Neurogenetics, National Institute on Aging), Carl E Clarke (School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK; and Department of Neurology, City Hospital, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK), Mark R Cookson (Laboratory of Neurogenetics, National Institute on Aging), J Mark Cooper (Department of Clinical Neurosciences, UCL Institute of Neurology), Jean Christophe Corvol (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and INSERM CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France), Carl Counsell (University of Aberdeen, Division of Applied Health Sciences, Population Health Section, Aberdeen, UK), Philippe Damier (CHU Nantes, CIC0004, Service de Neurologie, Nantes, France), Jean-François Dartigues (INSERM U897, Université Victor Segalen, Bordeaux, France), Panos Deloukas (Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK), Günther Deuschl (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein, Campus Kiel, Christian-AlbrechtsUniversität Kiel, Kiel, Germany), David T Dexter (Parkinson's Disease Research Group, Faculty of Medicine, Imperial College London, London, UK), Karin D van Dijk (Department of Neurology and Alzheimer Center, VU University Medical Center), Alissa Dillman (Laboratory of Neurogenetics, National Institute on Aging), Frank Durif (Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France), Alexandra Dür (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and AP-HP, Pitié-Salpêtrière Hospital), Sarah Edkins (Wellcome Trust Sanger Institute), Jonathan R Evans (Cambridge Centre for Brain Repair, Cambridge, UK), Thomas Foltynie (UCL Institute of Neurology), Jing Dong (Epidemiology Branch, National Institute of Environmental Health Sciences), Michelle Gardner (Department of Molecular Neuroscience, UCL Institute of Neurology), J Raphael Gibbs (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology), Alison Goate (Department of Psychiatry, Department of Neurology,
Washington University School of Medicine, MI, USA), Emma Gray (Wellcome Trust Sanger Institute), Rita Guerrero (Department of Molecular Neuroscience, UCL Institute of Neurology), Clare Harris (University of Aberdeen), Jacobus J van Hilten (Department of Neurology, Leiden University Medical Center, Leiden, Netherlands), Albert Hofman (Department of Epidemiology, Erasmus University Medical Center, Rotterdam, Netherlands), Albert Hollenbeck (AARP, Washington DC, USA), Janice Holton (Queen Square Brain Bank for Neurological Disorders, UCL Institute of Neurology), Michele Hu (Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Xuemei Huang (Departments of Neurology, Radiology, Neurosurgery, Pharmacology, Kinesiology, and Bioengineering, Pennsylvania State University Milton S Hershey Medical Center, Hershey, PA, USA), Isabel Wurster (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and German Center for Neurodegenerative diseases), Gavin Hudson (Neurology M4104, The Medical School, Newcastle upon Tyne, UK), Sarah E Hunt (Wellcome Trust Sanger Institute), Johanna Huttonlocher (deCODE genetics), Thomas Illig (Institute of Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Pámi V Jónsson (Department of Geriatrics, Landspítali University Hospital, Reykjavík, Iceland), Jean-Charles Lambert (INSERM U744, Lille, France; and Institut Pasteur de Lille, Université de Lille Nord, Lille, France), Cordelia Langford (Cambridge Centre for Brain Repair), Andrew Lees (Queen Square Brain Bank for Neurological Disorders), Peter Lichtner (Institute of Human Genetics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Patricia Limousin (Institute of Neurology, Söbeck Department, Unit of Functional Neurosurgery, London, UK), Grisel Lopez (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI, National Institutes of Health), Delia Lorenz (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein), Codrin Lungu (National Institutes of Health Parkinson Clinic, NINDS, National Institutes of Health), Alisdair McNeill (Department of Clinical Neurosciences, UCL Institute of Neurology), Catriona Moorby (School of Clinical and Experimental Medicine, University of Birmingham), Matthew Moore (Laboratory of Neurogenetics, National Institute on Aging), Huw R Morris (National Hospital for Neurology and Neurosurgery, University College London, London, UK), Karen E Morrison (School of Clinical and Experimental Medicine, University of Birmingham; and Neurosciences Department, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK), Valentina EscottPrice (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK), Ese Mudanohwo (Neurogenetics Unit, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery), Sean S O’ Sullivan (Queen Square Brain Bank for Neurological Disorders), Justin Pearson (MRC Centre for Neuropsychiatric Genetics and Genomics), Joel S Perlmutter (Department of Neurology, Radiology, and Neurobiology at Washington University, St Louis), Hjörvar Pétursson (deCODE genetics; and Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Pierre Pollak (Service de Neurologie, CHU de Grenoble, Grenoble, France), Bart Post (Department of Neurology, Radboud University Nijmegen Medical Centre), Simon Potter (Wellcome Trust Sanger Institute), Bernard Ravina (Translational Neurology, Biogen Idec, MA, USA), Tamas Revesz (Queen Square Brain Bank for Neurological Disorders), Olaf Riess (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Fernando Rivadeneira (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Patrizia Rizzu (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Mina Ryten (Department of Molecular Neuroscience, UCL Institute of Neurology), Stephen Sawcer (University of Cambridge, Department of Clinical Neurosciences,
Addenbrooke’s hospital, Cambridge, UK), Anthony Schapira (Department of Clinical Neurosciences, UCL Institute of Neurology), Hans Scheffer (Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Karen Shaw (Queen Square Brain Bank for Neurological Disorders), Ira Shoulson (Department of Neurology, University of Rochester, Rochester, NY, USA), Joshua Shulman (Baylor College of Medicine, Houston, Texas), Ellen Sidransky (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI), Colin Smith (Department of Pathology, University of Edinburgh, Edinburgh, UK), Chris C A Spencer (Wellcome Trust Centre for Human Genetics, Oxford, UK), Hreinn Stefánsson (deCODE genetics), Francesco Bettella (deCODE genetics), Joanna D Stockton (School of Clinical and Experimental Medicine), Amy Strange (Wellcome Trust Centre for Human Genetics), Kevin Talbot (University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Carlie M Tanner (Clinical Research Department, The Parkinson’s Institute and Clinical Center, Sunnyvale, CA, USA), Avazeh Tashakkori-Ghanbaria (Wellcome Trust Sanger Institute), François Tison (Service de Neurologie, Hôpital Haut-Lévêque, Pessac, France), Daniah Trabzuni (Department of Molecular Neuroscience, UCL Institute of Neurology), Bryan J Traynor (Laboratory of Neurogenetics, National Institute on Aging), André G Uitterlinden (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Daan Velseboer (Department of Neurology, Academic Medical Center), Marie Vidailhet (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225), Robert Walker (Department of Pathology, University of Edinburgh), Bart van de Warrenburg (Department of Neurology, Radboud University Nijmegen Medical Centre), Mirdhu Wickremaratchi (Department of Neurology, Cardiff University, Cardiff, UK), Nigel Williams (MRC Centre for Neuropsychiatric Genetics and Genomics), Caroline H Williams-Gray (Department of Neurology, Addenbrooke’s Hospital), Sophie Winder-Rhodes (Department of Psychiatry and Medical Research Council and Wellcome Trust Behavioural and Clinical Neurosciences Institute, University of Cambridge), Kári Stefánsson (deCODE genetics), Maria Martinez (INSERM UMR 1043; and Paul Sabatier University), Nicholas W Wood (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology), John Hardy (Department of Molecular Neuroscience, UCL Institute of Neurology), Peter Heutink (DZNE, German Center for Neurodegenerative Diseases and Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany), Alexis Brice (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225, AP-HP, Pitié-Salpêtrière Hospital), Thomas Gasser (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, and DZNE, German Center for Neurodegenerative Diseases), Andrew B Singleton.
Two-stage testing for epistasis: Proofs

This chapter is of technical nature and perhaps can be seen as a sort of appendix to Part I. It contains supporting theory and proofs of the results formulated in Chapter 3 and used in Chapter 4. The structure of this chapter is as follows. In Section 5.1 we formulate theory necessary to prove Theorem 3.1, which is achieved by deriving its generalization in the form of Theorem 5.5. The section uses the theoretical machinery presented in van der Vaart (1998), which should be viewed as a general point of reference for this section especially for the more involved concepts that could not be treated here in sufficient detail. At the end of Section 5.1 we also briefly focus on the asymptotic power of the score statistic in the logistic regression model. In Section 5.2 we formulate the proof of the normality and independence Theorem 3.3 by showing asymptotic normality with block-diagonal covariance matrix. The proof of Theorem 3.3 is based on writing the joint vector of interest as a continuous function of a sum of an independent and identically distributed terms and invoking such classical results as the multivariate central limit theorem (Theorem A.10), the delta method (Theorem A.9) and Slutsky’s lemma (Lemma A.6). We point this out because the same machinery is repeated in slight variations throughout this chapter. In Section 5.3, after the proof of Theorem 3.3 is formulated, we provide arguments that prove the three normality Theorems 3.4, 3.6 and 3.8. This is accomplished by arguments that prove the three independence Theorems 3.5, 3.7 and 3.9, which rely on uncorrelatedness and joint asymptotic normality provided by the three normality theorems. Along the way we also formulate and prove several technical lemmas required by the proofs of normality and independence theorems. Since all of the normality and independence results are largely analogous to Theorem 3.3 and rely on the same machinery, the proofs of the latter theorems are formulated in less detail than the proof of Theorem 3.3. The chapter is concluded by a calculation of several variance and covariance matrices that appear in the formulations of the normality and independence theorems and that were required by the implementation of the methods in EPIDET.

Note that throughout this chapter we assume that the data form a single random sample.
from our population of interest, which simplifies the theoretical treatment of the test statistics. As already explained at the beginning of Chapter 3.1 (see Section 14.5 in van der Vaart (2015) for a technical argument), the obtained results can be readily used also within the logistic regression model under the case-control design (with the exception of inference about $\beta_0$ in the general population).

5.1 Score statistic in general parametric model

In this section we focus on the score statistic in a general parametric model. In general, if the model has log-likelihood function $\ell(\cdot; \theta)$, we define the score function as the gradient of $\ell(\cdot; \theta)$, which is the vector of partial derivatives of $\ell(\cdot; \theta)$ with respect to components of the parameter $\theta$. We address this in more detail below, but the idea is to go beyond the score statistic in the logistic regression model and derive a general asymptotic result concerning the score statistic. The goal is met in Theorem 5.5, which provides a linear asymptotic relationship between the score statistic with true values of parameters and its estimated parameter counterpart. It is a direct generalization of Theorem 3.1 of Section 3.5, and therefore proves it.

5.1.1 Score statistic

Assume we have a parametric statistical model $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is an open subset of $\mathbb{R}^k$ and $P_\theta$ is a probability distribution for every $\theta \in \Theta$. A random sample (iid) of variables $X_1, \ldots, X_n$ can be regarded as a single observation of a random vector $X = (X_1, \ldots, X_n)'$ and the model for this vector is $\{P^n_\theta : \theta \in \Theta\}$, where $P^n_\theta$ is a product of $n$ copies of probability distribution $P_\theta$. The collection of probability distributions $M^n_\theta = \{P^n_\theta : \theta \in \Theta\}$ is referred to as a statistical experiment. Assume that the model $P_\theta$ has density $p_\theta(x)$ with respect to a $\sigma$-finite measure $\mu$ and that the likelihood function $p_\theta$ is sufficiently regular in the sense of the assumptions specified in Theorems 5.1 and 5.5. The log-likelihood of the random vector $X$ is equal to $\ell_n(X; \theta) = \sum_{i=1}^n \ell(X_i; \theta)$, where we put $\ell(x; \theta) = \log p_\theta(x)$. Moreover, the log-likelihood ratio of $\theta$ against $\theta_0$ is $\Lambda(\theta) = \log \prod_{i=1}^n (p_\theta/p_{\theta_0})(X_i)$, the ML estimator of $\theta$ in the experiment $M^n_\theta$ is $\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(X; \theta)$, and the score statistic equals

$$S_{n,\theta} = n^{-1/2} \dot{\ell}_n(X; \theta),$$

where $\dot{\ell}_n = (\partial/\partial \theta) \ell_n$ is the score function. The well known Fisher information matrix $I_\theta$ of the model $\{P_\theta : \theta \in \Theta\}$ is defined as $I_\theta = \mathbb{E}_\theta (\dot{\ell}(X; \theta)(\dot{\ell}(X; \theta))'$ and it is easy to see from

1By notation $\ell(\cdot; \theta)$ we understand a function of the parameter $\theta$ which also depends on some other quantity, for instance a random variable.

2As the reader will notice, Section 5.1 is the only exception to our convention of denoting non-scalar objects by bold font. We believe that in this case the strict use of the bold font would unnecessarily complicate the notation.
the definition of $\mathcal{I}_\theta$ and from $3 \mathbb{E}_\theta S_{n,\theta} = 0$ that $\mathcal{I}_\theta$ is the variance matrix of $S_{n,\theta}$. Moreover, according to the central limit theorem $S_{n,\theta}$ is asymptotically normal with zero mean and variance matrix $\mathcal{I}_\theta$, as $n \to \infty$. Additionally, in order to guarantee the existence of the maximum likelihood estimator of $\theta$ and its consistency, we assume throughout the text that the parameter $\theta$ is identifiable, meaning that $\theta_1 \neq \theta_2$ implies $\ell(\theta_1) \neq \ell(\theta_2)$ for all $\theta_1, \theta_2 \in \Theta$.

### 5.1.2 Taylor expansion of score statistic

For the purposes of Taylor expansion it is convenient to reparametrize the experiment $\mathcal{M}_\theta^n$ in the following equivalent way. Assume a fixed value $\delta \in \Theta$ and for each $\theta \in \Theta$ define a local parameter $h = \sqrt{n}(\theta - \delta)$. The local parameter space is then defined as $\mathcal{H}_n = \sqrt{n}(\Theta - \delta)$. For each $n$ the set $\mathcal{H}_n \subset \mathbb{R}^k$ is a collection of all parameters $h = \sqrt{n}(\theta - \delta)$, where $\theta \in \Theta$. The statistical experiments $\mathcal{M}_\theta^n = \{P^n_h : \theta \in \Theta\}$ and $\mathcal{M}_\theta^n = \{P^n_h + h/\sqrt{n} : h \in \mathcal{H}_n\}$ are equivalent, since with $\delta$ (and $n$) fixed, the mapping $\theta = \delta + h/\sqrt{n}$ is one-to-one, meaning that each $\theta \in \Theta$ in the experiment $\mathcal{M}_\delta^n$ corresponds to a single value of $h \in \mathcal{H}_n$ in $\mathcal{M}_\theta^n$ and vice-versa. Analogously to the ML estimator of $\theta$ within $\mathcal{M}_\theta^n$, we define the local ML estimator of $h$ within the experiment $\mathcal{M}_\delta^n$ as $\tilde{h}_n = \arg\max_{h \in H_n} \ell_n(X; \delta + h/\sqrt{n})$.

If $p_\delta$ is the likelihood function within the experiment $\mathcal{M}_\delta^n$, then $p_{\delta + h/\sqrt{n}}$ is the likelihood function within $\mathcal{M}_{\delta,h}^n$. The log-likelihood ratio within $\mathcal{M}_{\delta,h}^n$ of $h$ against $h = 0$ is

$$\Lambda_n^\delta(h) = \log \prod_{i=1}^n \frac{p_{\delta + h/\sqrt{n}}}{p_\delta}(X_i).$$

(5.2)

for which $\tilde{h}_n = \arg\max_{h \in H_n} \Lambda_n^\delta(h)$ . Since the terms $p_{\delta}(X_i)$ do not depend on $h$, differentiating $\Lambda_n^\delta(h)$ with respect to $h$ and denoting the gradient vector by $\nabla \Lambda_n^\delta(h)$, we get

$$S_{n,\delta} = \nabla \Lambda_n^\delta(h)|_{h=0}.$$  

(5.3)

Furthermore, under twice differentiability of $\ell(x; \theta)$ with respect to $\theta$, for every $x$, the Hessian matrix of $\Lambda_n^\delta(h)$ is $\nabla^2 \Lambda_n^\delta(h) = n^{-1} \sum_i \ell'(X_i; \delta + h/\sqrt{n})$. By the law of large numbers, as $n \to \infty$, the sample mean $n^{-1} \sum_{i=1}^n \ell'(X_i; \theta)$ converges in probability to the expectation $\mathbb{E}_\theta \ell'(\theta) = -\mathcal{I}_\theta$, or in other words $\nabla^2 \Lambda_n^\delta(h)|_{h=0} = -\mathcal{I}_\theta + o_p(1)$. This together with the fact that $\Lambda_n^\delta(0) = 0$ can be used to get a Taylor expansion of $\Lambda_n^\delta(h)$ with respect to $h$ around 0 and obtain the relationship

$$\Lambda_n^\delta(h_n) = h' S_{n,\delta} - \frac{1}{2} h' \mathcal{I}_\theta h + o_p(1),$$

(5.4)

for every sequence $h_n$ converging to $h$. The convergence of the remainder term in (5.4) relies on regularity of the statistical model in question. A sufficient regularity condition for this is differentiability in the quadratic mean of $\sqrt{p_\theta}$ established in Theorem 5.1. The theorem and its proof can be found in van der Vaart (1998) (Theorem 7.2).

---

3 $\mathbb{E}_0 S_{n,0}$ denotes the expectation of $S_{n,0}$ under density $p_0$. 
Theorem 5.1 (Taylor expansion of log-likelihood) Let $\Theta \in \mathbb{R}^k$ be an open set and let $\{p_\theta : \theta \in \Theta\}$ be a statistical model differentiable in the quadratic mean at $\theta$, meaning that the density function $p_\theta(x)$ of $P_\theta$ with respect to $\sigma$-finite measure $\mu$ and some measurable functions $\ell(x;\theta) : \mathbb{R}^k \to \mathbb{R}^k$ defined on $\Theta$ for every $x$ satisfy

\[
\int \left( \sqrt{p_{\theta+t}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} t \ell(x;\theta) \sqrt{p_\theta} \right)^2 d\mu(x) = o(||t||^2), \quad t \to 0. \tag{5.5}
\]

Then, $E_\theta \ell_\theta(X) = 0$ and the Fisher information matrix $I_{\theta} = E_\theta (\ell_\theta(X;\theta)(\ell_\theta(X;\theta))')$ exists. Furthermore, let $\Lambda_n^\theta(h)$ be defined by (5.2) and fix $h \in \mathbb{R}^k$. Then, for every sequence $h_n$ such that $h_n \to h$ as $n \to \infty$, it holds $\Lambda_n^\theta(h_n) = h' S_{n,\theta} - \frac{1}{2} h' I_{\theta} h + o_P(1)$.

Theorem 5.1 provides justification for the Taylor expansion of the log-likelihood $\Lambda_n^\theta(h)$ around $h = 0$ in (5.4). The theorem applies to models that are differentiable in the quadratic mean, a condition defined within the theorem. It is worth pointing out that differentiability in the quadratic mean does not require $\ell(\theta)$ to be differentiable in the usual sense with respect to $\theta$ for any single $x$. Instead, it merely requires the existence of measurable function $\ell(\theta)$, which acts like the derivative of $\ell(\theta)$ in the linear approximation sense of (5.5). Whether a given likelihood function $p_\theta$ satisfies (5.5) can often be verified using Lemma 7.6 in van der Vaart (1998) by looking at continuous differentiability of $\sqrt{p_\theta}$ for every $x$ and continuity (in $\theta$) of $I_\theta$.

Regarding the logistic regression model of (3.1), its score function satisfies the assumptions of Theorem 5.1, because the likelihood function in (3.1) is clearly smooth (infinitely differentiable) with respect to $\beta$. As it turns out, however, the assumptions of Theorem 5.1 are not sufficient for what is required to prove the results in the rest of Section 5.1. In order to prove Theorem 5.5 we need to (asymptotically) approximate $\nabla \Lambda_n^\theta(h)$. In other words we need sufficiently strong assumptions concerning $\ell(\theta)$ that allow us to write

\[
\nabla \Lambda_n^\theta(h) = S_{n,\delta} - I_{\delta} h + o_p(1), \tag{5.6}
\]

which is the focus of Section 5.1.3.

5.1.3 Asymptotic representation of the score statistic

In the experiment $\mathcal{M}_\theta^\nu = \{P_\theta^\nu : \theta \in \Theta\}$ we denote by $\theta^\nu$ the (unknown) true value of the parameter $\theta$ and suppose we want to test a hypothesis $H_0: \theta = \Theta_0$ for some $\Theta_0 \subset \Theta$. Under $H_0$ we can estimate $\theta$ by its null hypothesis ML estimator $\hat{\theta}_n^0 = \arg\max_{\theta \in \Theta_0} \ell_n(X;\theta)$ and define the null hypothesis score statistic with estimated parameters as

\[
\hat{\ell}_n^0 = n^{-1/2} \ell_n(X;\hat{\theta}_n^0), \tag{5.7}
\]

\[\text{In the logistic regression model of (3.1) we are interested in the hypothesis that exactly one component of the parameter is equal to zero, where we hypothesize that the parameter vector } \beta = (\beta_0, \beta_1, \beta_2, \beta_3)^t \text{ is such that } \beta_3 = 0.\]
which is obtained from (5.1) by replacing $\theta$ with $\hat{\theta}_n^0$. In the equivalent experiment $\mathcal{M}_{\theta, \delta}^n$ parametrized by $h = \sqrt{n}(\theta - \delta)$, we define the null hypothesis local parameter space as $\mathcal{H}_{n,0} = \sqrt{n}(\Theta_0 - \delta)$ and the null hypothesis ML estimator of $h$ as $\hat{h}_n^0 = \arg\max_{h \in \mathcal{H}_{n,0}} \Lambda_n^0(h)$. As can be easily seen, the two ML estimators $\hat{\theta}_n$ and $\hat{h}_n^0$ are linked by $\hat{\theta}_n = \delta + \hat{h}_n^0/\sqrt{n}$. Moreover, it holds $\hat{S}_n^0 = \nabla \Lambda_n^0(h)|_{h=\hat{h}_n^0}$, as follows from (5.2) and

$$\nabla \Lambda_n^0(h)|_{h=\hat{h}_n^0} = \left( \frac{\partial}{\partial h} \log \prod_{i=1}^{n} \frac{p_{\theta_i}(X_i)}{p_{\hat{\theta}_n}(X_i)} \right)|_{h=\hat{h}_n^0} = n^{-1/2} \sum_{i=1}^{n} \ell(X_i; \hat{\theta}_n) = \hat{S}_n^0. $$

Then, with sufficient regularity for (5.6) to hold, it follows

$$\hat{S}_n^0 = S_{n, \delta} - \mathcal{I}_\delta \hat{h}_n^0 + o_p(1).$$

(5.8)

This gives us a useful link between $\hat{S}_n^0$ and $S_{n, \delta}$. In fact, we can go further than this and establish an even more useful asymptotic relationship between the two vectors, which is done in this section via Theorem 5.5. In order to formulate the result it is necessary to define the concept of convergence of sets.

**Definition 5.2** We say that a sequence of sets $\mathcal{H}_n$ converges to a set $\mathcal{H}$ if the following two conditions hold:

i) For every $h \in \mathcal{H}$ there exists a sequence $h_n \in \mathcal{H}_n$ such that $h_n \rightarrow h$ as $n \rightarrow \infty$.

ii) For every sequence $h_n \in \mathcal{H}_n$ such that $h_n \rightarrow h$ for some $h$ it follows that $h \in \mathcal{H}$.

The following lemma gives two useful inequalities for a sequence of sets. The proof of the lemma can be found in van der Vaart (1998) (Lemma 7.13).

**Lemma 5.3** Let $X_n$ be a sequence of random vectors that converges in distribution to $X$ as $n \rightarrow \infty$. Let the sequence of sets $\mathcal{H}_n \subset \mathbb{R}^k$ converge to a nonempty set $\mathcal{H}$. Then

i) $||X_n - \mathcal{H}_n \cap F|| \geq ||X_n - \mathcal{H} \cap F|| + o_p(1)$ for every closed set $F$.

ii) $||X_n - \mathcal{H}_n \cap G|| \leq ||X_n - \mathcal{H} \cap G|| + o_p(1)$ for every open set $G$.

With the concept of set convergence defined, we can formulate a theorem which provides asymptotic approximations for the ML estimator $\hat{h}_n^0$ based on maximization of the non-vanishing part of the Taylor expansion in (5.4). The assumptions of the theorem include stronger regularity conditions compared to Theorem 5.1, which are necessary for sufficient smoothness of the second order remainder terms in the Taylor expansion.

**Theorem 5.4** Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model differentiable in quadratic mean at $\delta$ with likelihood function $p_\theta$ and non-singular $\mathcal{I}_\delta$. Suppose that there exists a measurable function $f(x)$ such that under $X \sim P_\delta$ the expectation of $f^2(X)$ is finite and

$$|\ell(x; \theta_1) - \ell(x; \theta_2)| \leq f(x)||\theta_1 - \theta_2||,$$
for all $\theta_1, \theta_2 \in \mathbb{R}^k$ such that $\|\theta_1 - \theta\| < K$ and $\|\theta_2 - \theta\| < K$ for some $K \in (0, \infty)$. Let $\hat{\theta}_n^0$ be a $\sqrt{n}$-consistent estimator of $\theta$ and let $\mathcal{H}_{n,0}$ converge to a nonempty convex set $\mathcal{H}_0$ as $n \to \infty$. Denote $\tilde{h}_n^0 = \arg \min_{h \in \mathcal{H}_{n,0}} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2$ and $\hat{h}_n^0 = \arg \min_{h \in \mathcal{H}_0} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2$.

Then
\[
\hat{h}_n^0 = \tilde{h}_n^0 + o_P(1) = \tilde{h}_n^0 + o_P(1). \tag{5.9}
\]

**Proof.** By definition, the ML estimator $\hat{h}_n^0$ maximizes $\Lambda_n^0(h)$, for which the Taylor expansion (5.4) holds. We denote the $o_P(1)$ term in (5.4) as $R_n(h)$ and rewrite $\tilde{h}_n^0$ as
\[
\tilde{h}_n^0 = \arg \max_{h \in \mathcal{H}_{n,0}} (h^T S_{n,\theta} - \frac{1}{2} h^T I_{\theta} h - \frac{1}{2} S_{n,\theta} I_{\theta}^{-1} S_{n,\theta} + R_n(h))
= \arg \min_{h \in \mathcal{H}_{n,0}} (\| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2 + R_n(h)). \tag{5.10}
\]

The first equality in (5.10) is trivial since $\frac{1}{2} S_{n,\theta} I_{\theta}^{-1} S_{n,\theta}$ does not depend on the maximization argument $h$, which means that subtracting it inside the arg max does not influence the argument of the maximization. The second equality in (5.10) is only a straightforward rewrite.

As shown in the proofs of Theorems 7.12 and 16.7 in van der Vaart (1998), under differentiability in quadratic mean, the remainder term $R_n(h)$ converges to zero in probability uniformly over balls of radius $M_n$ for $M_n \to \infty$ sufficiently slowly. Fix such $M_n$. By the $\sqrt{n}$-consistency of $\hat{\theta}_n^0$, the ML estimators $\hat{h}_n^0$ are bounded in probability. Hence, with probability tending to 1, $\hat{h}_n^0$ belong to any sequence of balls $M_n$ with radius going $\infty$. Consequently, $\hat{h}_n^0$ can be assumed to belong to $\mathcal{H}_{n,0} \cap \text{ball}(0, M_n)$. This gives us uniform convergence of $R_n(h)$ to zero in probability on $\mathcal{H}_{n,0}$, which in turn yields
\[
\arg \min_{h \in \mathcal{H}_{n,0}} (\| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2 + R_n(h)) = (\arg \min_{h \in \mathcal{H}_{n,0}} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2) + o_P(1),
\]
thus proving the first equality in (5.9). Since the intersections of sets $\mathcal{H}_{n,0}$ with the balls of radius $M_n \to \infty$ still converge to $\mathcal{H}_0$, Lemma 5.3 yields
\[
\arg \min_{h \in \mathcal{H}_{n,0}} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2 = (\arg \min_{h \in \mathcal{H}_0} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2) + o_P(1),
\]
which concludes the proof of the second equality in (5.9).\[\Box\]

Under the assumptions of Theorem 5.4, we can combine (5.8) with (5.9), which yields
\[
\tilde{S}_n^0 = S_{n,\theta} - \arg \min_{h \in \mathcal{H}_0} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2 + o_P(1).
\]

Further rewriting the right-hand side yields
\[
\tilde{S}_n^0 = I_{\theta}^{1/2} (I_{\theta}^{-1/2} S_{n,\theta} - I_{\theta}^{1/2} \arg \min_{h \in \mathcal{H}_0} \| I_{\theta}^{1/2} h - I_{\theta}^{-1/2} S_{n,\theta} \|^2) + o_P(1)
= I_{\theta}^{1/2} (I_{\theta}^{-1/2} S_{n,\theta} - \arg \min_{h \in \mathcal{H}_0} \| h - I_{\theta}^{-1/2} S_{n,\theta} \|^2) + o_P(1). \tag{5.11}
\]
The term $g_{\text{min}}^n = \arg \min_{g \in \mathcal{I}_H^{1/2}H_0} \|g - \mathcal{I}_H^{1/2}S_{n,0}\|^2$ can be seen as a solution to the minimization problem of finding an element in the space $\mathcal{I}_H^{1/2}H_0$ that is closest to $\mathcal{I}_H^{1/2}S_{n,0}$, or in other words, finding an orthogonal projection of $\mathcal{I}_H^{1/2}S_{n,0}$ onto $\mathcal{I}_H^{1/2}H_0$. Denoting $\Pi_{\mathcal{I}_H}$ for the projection operator onto $\mathcal{I}_H^{1/2}H_0$, we get $g_{\text{min}}^n = \Pi_{\mathcal{I}_H}(\mathcal{I}_H^{1/2}S_{n,0})$. Finally, plugging this into (5.11) yields

$$\hat{S}_{n}^0 = \mathcal{I}_H^{1/2}(I - \Pi_{\mathcal{I}_H})I_H^{1/2}S_{n,0} + o_P(1),$$

where $I$ is the identity matrix (of the same dimension as $\hat{\theta}$). This is the asymptotic linear relationship between the two score statistics $\hat{S}_{n}^0$ and $S_{n,0}$ under $H_0$, that was the goal of this section. We summarize the results derived in this section by formulating Theorem 5.5, which in turn implies Theorem 3.1. The proof of Theorem 5.5 follows directly from the theory and computations of this section.

**Theorem 5.5 (Asymptotic representation of score statistic)** Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model and let the assumptions of Theorems 5.1 and 5.4 be satisfied. Denote the true value of the parameter by $\hat{\theta}$. For $\Theta_0 \subset \Theta$ let $H_0 : \theta \in \Theta_0$ be the null hypothesis and define the local parameter space corresponding to the null hypothesis $H_0$ as $\mathcal{H}_{n,0} = \sqrt{n}(\Theta_0 - \hat{\theta})$. Let $\mathcal{H}_{n,0}$ converge to a nonempty convex set $\mathcal{H}_0$ as $n \to \infty$. Finally, let $S_{n,0}$ be defined by (5.1) and $\hat{S}_{n}^0$ by (5.7). Then, under $H_0$ it holds $\hat{S}_{n}^0 = AS_{n,0} + o_P(1)$ as $n \to \infty$, for $A = \mathcal{I}_H^{1/2}(1 - \Pi_{\mathcal{I}_H})I_H^{1/2}$, where $\mathcal{I}_H^{1/2}$ is a square-root matrix of the Fisher information matrix $\mathcal{I}_H$, $\Pi_{\mathcal{I}_H}$ is an orthogonal projection linear operator onto the space $\mathcal{I}_H^{1/2}H_0$ and $I$ is the unit matrix. Consequently, $S_{n}^0$ is asymptotically normally distributed under $H_0$ with mean zero and variance $A\mathcal{I}_H A'$.

### 5.1.4 Slope of score statistic

For the purposes of finding an optimal pre-test level, we formulate a result concerning the asymptotic power of a test based on the score statistic. The following theorem is a special case of Theorem 15.4 in van der Vaart (1998) combined with Addendum 15.5 therein. The original Theorem 15.4 in van der Vaart (1998) assumes locally asymptotically normal experiments, while our formulation in Theorem 5.6 assumes a model $\{P_\theta : \theta \in \Theta\}$ differentiable in quadratic mean, which is sufficient for local asymptotic normality of the corresponding random sample (iid) based statistical experiment $\{P_\theta^n : \theta \in \Theta\}$.

**Theorem 5.6 (Limiting power functions)** Assume a statistical model $\{P_\theta : \theta \in \Theta\}$ and the corresponding sequence of experiments $\{P_\theta^n : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is an open subset of $\mathbb{R}^k$ and $P_\theta$ is a probability distribution for every $\theta \in \Theta$. Let the model be differentiable in quadratic mean with nonsingular Fisher information matrix $\mathcal{I}_H$. Consider the problem of testing a null hypothesis $H_0 : \psi(\theta) \leq 0$ against the alternative $H_1 : \psi(\theta) > 0$, where $\psi : \Theta \to \mathbb{R}$ is differentiable at some $\theta_0 \in \Theta$, with $\psi(\theta_0) = 0$ and the (row vector) gradient
\[ \dot{\psi}_{\theta_0} \] is non-zero. Put \( s_{\theta_0} = (\dot{\psi}_{\theta_0}^{-1} \dot{\psi}_{\theta_0}^{-1})^{1/2} \) and let \( V_n \) be a sequence of statistics such that
\[
V_n = s_{\theta_0}^{-1} \dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \Delta_n \theta_0 + o_{P_{\theta_0}}(1),
\]
for a sequence of statistics \( \Delta_n \theta_0 \) that converges in distribution under \( \theta_0 \) to the normal distribution \( N(0, I_{\theta_0}) \). Then, under the local alternative \( \theta = \theta_0 + h/\sqrt{n} \) for any \( h \in \mathbb{R}^k \), the probabilities \( P_0(V_n \geq \Phi^{-1}(1 - \alpha)) \) converge to \( 1 - \Phi(\Phi^{-1}(1 - \alpha) - s_{\theta_0}^{-1} \dot{\psi}_{\theta_0} h) \), where \( \Phi \) is the standard normal distribution function.

The quantity \( s_{\theta_0}^{-1} \dot{\psi}_{\theta_0} \) in Theorem 5.6 is called the slope of the statistic \( V_n \) for testing the null \( H_0 : \psi(\theta) \leq 0 \) against the alternative \( H_1 : \psi(\theta) > 0 \). The assertion of the theorem is that the slope determines the asymptotic power of the test which rejects \( H_0 \) for large values of \( V_n \) by determining the shift of the location of \( V_n \) under the local alternative.

Suppose we want to test \( H_0 : \theta_i = 0 \) against \( H_1 : \theta_i > 0 \) using as the test statistic the \( i \)-th coordinate of the score statistic with estimated parameters \( \hat{S}_n \) defined in (5.7). Denoting the \( i \)-th coordinate of \( \hat{S}_n \) by \( \hat{S}_{ni} \), Theorem 5.5 yields that under \( H_0 \) the statistic \( \hat{S}_{ni} \) is normally distributed with mean zero and variance \( a_i I_{\theta_0} a_i \), where \( a_i \in \mathbb{R}^k \) equals the \( i \)-th row of \( A \). Therefore, we can test \( H_0 \) by comparing \( \hat{V}_n = (a_i I_{\theta_0} a_i)^{-1/2} \hat{S}_{ni} \) with the appropriate standard normal quantile. For the purposes of power investigation of the test based on \( \hat{V}_n \), we apply Theorem 5.6 to the statistic \( \hat{V}_n \) and get the following result, which is formulated as a lemma for easy reference.

**Lemma 5.7** Under the assumptions of Theorem 5.5, the slope of \( \hat{V}_n \) is \( (e_i^T \hat{I}_{\theta_0}^{-1} e_i)^{-1/2} e_i \). Consequently, the asymptotic power function of \( \hat{V}_n \) is \( 1 - \Phi(\Phi^{-1}(1 - \alpha) - (e_i^T \hat{I}_{\theta_0}^{-1} e_i)^{-1/2} e_i h) \).

**Proof.** We need to show that \( \hat{V}_n \) has the form of (5.12). We start by noting that the hypotheses \( H_0 : \theta_i = 0 \) and \( H_1 : \theta_i > 0 \) are obtained by putting \( \psi(\theta) = \theta_i \) in Theorem 5.6. Consequently, the gradient \( \dot{\psi}_{\theta_0} \) is equal to the \( i \)-th canonical basis vector\(^5 \) \( e_i \). This makes \( s_{\theta_0} \) of Theorem 5.6 equal to the square root of the \( i \)-th diagonal element of \( \hat{I}_{\theta_0}^{-1} \), or in other words \( s_{\theta_0} = (e_i^T \hat{I}_{\theta_0}^{-1} e_i)^{1/2} \). On the other hand, by Theorem 5.5 it also holds \( \hat{V}_n = (a_i I_{\theta_0} a_i)^{-1/2} (a_i^T \hat{S}_{ni} + o_{P_{\theta_0}}) \), where \( \hat{S}_{ni} \) converges in distribution to \( N(0, I_{\theta_0}) \) under \( \theta_0 \). Therefore, in order for \( \hat{V}_n \) to have the form of (5.12), it is sufficient to show that
\[
a_i^T \hat{I}_{\theta_0}^{-1} e_i = (a_i^T \hat{I}_{\theta_0}^{-1} e_i)^{-1/2} = e_i.
\]
Denoting the left hand-side in the above display by \( L \) and plugging \( a_i = e_i^T A = e_i^T \hat{I}_{\theta_0}^{-1} (I - \Pi_{I^H}) \hat{I}_{\theta_0}^{-1} \) into \( L \) yields
\[
\begin{align*}
L &= e_i^T \hat{I}_{\theta_0}^{-1/2} (I - \Pi_{I^H}) \hat{I}_{\theta_0}^{-1/2} (e_i^T \hat{I}_{\theta_0}^{-1} e_i)^{1/2} (e_i^T \hat{I}_{\theta_0}^{-1/2} (I - \Pi_{I^H}) \hat{I}_{\theta_0}^{-1/2} e_i)^{-1/2} \\
&= e_i^T (\hat{I}_{\theta_0} - I_{\theta_0}) (J_{\theta_0} J_{\theta_0}^T) \hat{I}_{\theta_0}^{-1} e_i e_i^T \hat{I}_{\theta_0}^{-1} e_i (e_i^T (\hat{I}_{\theta_0} - I_{\theta_0}) (J_{\theta_0} J_{\theta_0}^T) \hat{I}_{\theta_0}^{-1} e_i)^{-1/2} \\
&= e_i^T M (e_i^T \hat{I}_{\theta_0}^{-1} e_i)^{1/2} (e_i^T M e_i)^{-1/2},
\end{align*}
\]
\(^5\)Canonical basis of \( \mathbb{R}^k \) are the column vectors of the identity matrix \( I_k \).
where we used $J_i = \text{diag}(e_i)$ to express the projection matrix $\Pi_{\mathcal{I}}$ derived in Section 5.4.3 and denoted $M = \mathcal{I}_{\theta_0} - \mathcal{I}_{\theta_0}(J_i, \mathcal{I}_{\theta_0})^{-1} \mathcal{I}_{\theta_0}$. Since $J_i$ is a block matrix, its pseudoinverse has the same block structure and the values in the blocks are obtained inverting $(\mathcal{I}_{\theta_0})^{-i,-i}$, which denotes $\mathcal{I}_{\theta_0}$ without the $i$-th row and column. Consequently, with the exception of the $i$-th diagonal element all elements of $M$ are equal to zero, which means that we only need to show that $e_i' \mathcal{I}^{-1}_{\theta_0} e_i e_i'M e_i = 1$. However, this again follows from the special block form of $M$, which results in the $i$-th diagonal element being equal to $1/e_i' \mathcal{I}^{-1}_{\theta_0} e_i$. Therefore, (5.13) holds, Theorem 5.6 applies, and the slope of $\tilde{V}_n$ follows by evaluating $s_{\theta_0}^{-1} \dot{\psi}_{\theta_0}$. 

5.2 Post-test for pooled-sample pre-test: Proofs

In this section we provide a proof of Theorem 3.3, which postulates joint asymptotic normality and asymptotic independence of the score vector with estimated parameters $\hat{S}_n$ and the pooled-sample chi-square statistics $T_{\theta_0}^{po}$ and $R_{\theta_0}^{po}$. As stated in Section 3.2.1, the statistic $T_{\theta_0}^{po}$ has the same asymptotic distribution as the squared Euclidean norm of the vector $T_{\theta_0}^{po}$ defined in (3.7). Additionally, as showed at the end of Section 3.2.1, the statistic $R_{\theta_0}^{po}$ is asymptotically distributed as the squared Euclidean norm of a linear transformation of $T_{\theta_0}^{po}$. Therefore, in light of the continuous mapping theorem, we focus on showing the asymptotic independence of $\hat{S}_n$ and $T_{\theta_0}^{po}$ by proving their joint asymptotic normality with block-diagonal asymptotic covariance matrix. Due to the asymptotic representation of the score statistic given by Theorem 3.1, we can show joint asymptotic normality of $\hat{S}_n$ and $T_{\theta_0}^{po}$ by showing that the joint asymptotic distribution of $\hat{S}_n$ and $T_{\theta_0}^{po}$ is normal. But first, we calculate the expectation of $T_{\theta_0}^{po}$, which in turn applies (with the appropriate parameter replacement) also to pre-test generating vectors $T_{\theta_0}^{co}$ of (3.12), $T_{\theta_0}^{ca}$ of (3.16), $T_{\theta_0}^{\delta}$ of (3.21) and (3.27), and $T_{\theta_0}^{\delta C}$ of (3.23) and of (3.29).

Expectation of the pre-test generating vectors

We start by calculating the expectation of a general pre-test generating vector. While we formulate the explicit result for the case of the pooled-sample pre-test only, the expression is the same for the other pre-test generating vector, provided that the appropriate probabilities and sample sizes are plugged into the expression.

Lemma 5.8 It holds

$$\mathbb{E} T_{\theta_0}^{po} = \frac{(n+1)/\sqrt{n}}{\sqrt{\pi_{kl} - \pi_{k} \pi_{l}}} = \sqrt{n}((\pi_{kl} - \pi_{k} \pi_{l})/\sqrt{\pi_{k} \pi_{l}})_{k,l} + O\left(\frac{1}{\sqrt{n}}\right).$$

Proof. We rewrite $\pi_{k,l} - \pi_{k} \pi_{l}$ as

$$(\pi_{k,l} - \pi_{k} \pi_{l}) - \pi_{k} (\pi_{l} - \pi_{l}) - \pi_{l} (\pi_{k} - \pi_{k}) + (\pi_{k} - \pi_{k} \pi_{l}) + (\pi_{k} - \pi_{k} \pi_{l})(\pi_{l} - \pi_{l}).$$
Taking the expectation of the above display yields
\[
E(\tilde{\pi}_{kl} - \tilde{\pi}_{k,j}) = (\pi_{kl} - \pi_{k,j}) + E(\tilde{\pi}_{k} - \pi_{k})(\tilde{\pi}_{j} - \pi_{j}).
\]  
(5.14)

Using the genotypes variables \(X_i, Y_i\) defined in Section 3.2, we write \(n_{kl} = \sum_{i=1}^{n} I_{X_i=k, Y_i=l}\), \(n_k = \sum_{i=1}^{n} I_{X_i=k}\) and \(n_l = \sum_{i=1}^{n} I_{Y_i=l}\). This then implies
\[
E(n_{kl}, n_{l}) = E\sum_{i=1}^{n} I_{X_i=k, Y_i=l} + E\sum_{i=1}^{n} I_{X_i=k}I_{Y_i=l} = n\pi_{kl} + n(n-1)\pi_{k}\pi_{j}.
\]
which in turn yields \(E(\tilde{\pi}_{k} - \pi_{k})(\tilde{\pi}_{j} - \pi_{j}) = n^{-2}E(n_{kl} - \pi_{kl}) - \pi_{k}\pi_{j} = n^{-1}(\pi_{kl} - \pi_{k}\pi_{j})\). Plugging this into (5.14) and standardizing the result by \((n/(\pi_{k}\pi_{j}))^{1/2}\) yields \(E\Psi_n^{\beta_0}\).

\[\square\]

**Normality of \(S_n\) and \(T_n^{\beta_0}\)**

Before focusing on \(S_n\) and \(T_n^{\beta_0}\), we formulate a supporting lemma. In order to avoid confusion, we note that the vectors \(T_n^1, T_n^2, T_n^3\) all have equal length, hence the indexing by \(k,l\).

This means that the components of \(T_n^2\) and \(T_n^3\) are appropriately repeated to achieve the same length as \(T_n^1\).

**Lemma 5.9** Let \(S_n\) be defined by (3.3) and put \(T_n^1 = \sqrt{n}(\pi_k - \pi_{kl})k,l, T_n^2 = \sqrt{n}(\pi_{k} - \pi_{k,l})k,l, T_n^3 = \sqrt{n}(\pi_{l} - \pi_{k,l})k,l\). If \(H_0^c\) holds, then the random vector \(\Psi_n^{\beta_0} = (S_n, T_n^1, T_n^2, T_n^3)'\) has zero expectation and it converges in distribution as \(n \to \infty\) to the centered normal distribution with variance matrix
\[
\mathbf{W}_{123} = \begin{pmatrix} \mathbf{I}_\beta & 0 \\ 0 & \mathbf{V}_{123} \end{pmatrix},
\]
(5.15)

where \(\mathbf{I}_\beta\) is the Fisher information matrix, \(\mathbf{V}_{123}\) is the asymptotic variance matrix of the vector \((T_n^1, T_n^2, T_n^3)\)'.

**Proof.** The vectors \(T_n^1, T_n^2, T_n^3\) have zero mean and under \(H_0^c\) the same holds for \(S_n\). Moreover, again using \(n_{kl} = \sum_{i=1}^{n} I_{X_i=k, Y_i=l}, n_k = \sum_{i=1}^{n} I_{X_i=k}\) and \(n_l = \sum_{i=1}^{n} I_{Y_i=l}\), we can write \(\Psi_n^{\beta_0} = n^{-1/2}\sum_{i=1}^{n} \mathbf{Z}_i\), where we define for \(i = 1, \ldots, n\)
\[
\mathbf{Z}_i = \left( (\Delta_i - \Psi_i^{\beta_0})' \mathbf{z}(X_i, Y_i), (I_{X_i=k,Y_i=l} - \pi_{kl})'k,l, (I_{X_i=k} - \pi_{k})'k,l, (I_{Y_i=l} - \pi_{l})'l,l \right)',
\]
where \(X_i, Y_i, \Delta_i\) are the genotypes at the two loci and the phenotype of the \(i\)-th individual and \(\Psi_i^{\beta_0} = \Psi(\beta_0^0 \mathbf{z}(X_i, Y_i))\) all defined in Section 3.1. It follows from the independence and identical distribution (iid) of the vectors \((X_i, Y_i, \Delta_i)\)'s, \(i = 1, \ldots, n\), that \(\mathbf{Z}_1, \ldots, \mathbf{Z}_n\) are also iid.

Denote the variance matrix of \(\mathbf{Z}_i\) by \(\mathbf{W}_Z\). By the multivariate central limit theorem it follows that \(\Psi_n^{\beta_0}\) is asymptotically zero-mean normal under the null hypothesis \(H_0^c\) with variance matrix \(\mathbf{W}_Z\). To conclude the proof we only need to show that the covariance matrix \(\mathbf{W}_Z\) has the form of \(\mathbf{W}_{123}\) in (5.15). Since \(\mathbf{I}_\beta\) is the variance matrix of \(S_n\) and since we make no claims about the shape of \(\mathbf{V}_{123}\), save for its obvious finiteness, we only need to show that
cov \left( S_n, T_n^1 \right) = 0, \ cov \left( S_n, T_n^2 \right) = 0 \ and \ cov \left( S_n, T_n^3 \right) = 0. \ Since \ \mathbb{E}(\Delta_1 - \Psi_{1}' | X_1, Y_1) = 0 \ under \ H_0^c, \ we \ get \nabla
\mathbb{E}(\Delta_1 - \Psi_{1}' | X_1, Y_1) I_{\left\{ X_1 = k, Y_1 = l \right\}} = \mathbb{E}(\mathbb{E}(\Delta_1 - \Psi_{1}' | X_1, Y_1) I_{\left\{ X_1 = k, Y_1 = l \right\}} | X_1, Y_1)) = 0. \ This \ yields \ cov \left( S_n, T_n^1 \right) = 0. \ The \ other \ two \ equalities \ follow \ analogously. \ \blackslug

With the above lemma we can easily show joint asymptotic normality and asymptotic independence of \( S_n \) and \( T_n^p \).

**Lemma 5.10** Let \( S_n \) be defined by (3.3) and \( T_n^p \) by (3.7). If \( H_0^c \) holds, then the random vector \( (S_n', (T_n^p - \mathbb{E}T_n^p))' \) has zero expectation and it converges in distribution, as \( n \to \infty \), to the centered normal distribution with variance matrix
\[
\mathbb{W}_p = \begin{pmatrix}
I_{\beta} & 0 \\
0 & \mathbb{V}_T^p
\end{pmatrix},
\]
where \( I_{\beta} \) is the Fisher information matrix, \( \mathbb{V}_T^p \) is the asymptotic variance matrix of \( T_n^p \).

**Proof.** In light of Lemma 5.9, we can use the delta method (Theorem A.9) on \( \Psi_n^p \) defined in Lemma 5.9 to obtain the asymptotic normality of \( (S_n', Z_n^p)' \), where
\[
Z_n^p = n^{1/2} \left( \left( \frac{n-k}{n} \right) - \pi_k, \frac{n-k}{n} \right) / \sqrt{\frac{n-k}{n} \frac{n-k}{n} k,l}.
\]
Since subtracting the term \( (\pi_k, \pi_k) / \sqrt{n-k/n} \) inside \( Z_n^p \) makes its expectation zero for all \( n \), the asymptotic distribution of \( Z_n^p \) also has zero expectation. Moreover, since the mapping within the delta method argument that transforms \( (S_n', T_n^1', T_n^2', T_n^3') \) into \( (S_n', Z_n^p)' \) is obviously differentiable and it acts separately on \( S_n \) and \( (T_n^1', T_n^2', T_n^3') \), the joint asymptotic normality and zero asymptotic covariance of \( S_n \) and \( Z_n^p \) follow. By Lemma 5.8 the expectation of \( T_n^p \) is \( (n^{1/2} + n^{-1/2}) (\pi_k - \pi_k, \pi_k) / \sqrt{n-k/n} \frac{n-k}{n} k,l \). Therefore, \( (S_n', (T_n^p - \mathbb{E}T_n^p))' = (S_n', Z_n^p)' + O(n^{-1/2}) \), which combined with Slutsky’s lemma concludes the proof. \ \blackslug

**Proof of Theorem 3.3 (Normality and independence with pooling)**

By Theorem 3.1, the random vector \( \bar{S}_n \) is asymptotically equal to a linear transformation of \( S_n \) given by matrix \( A \), i.e. \( \bar{S}_n = AS_n + o_p(1) \). Combining this with Lemma 5.10, Slutsky’s lemma and the fact that linear transformations preserve normality, we get
\[
\begin{pmatrix}
\bar{S}_n \\
T_n^p - \mathbb{E}T_n^p
\end{pmatrix} = \begin{pmatrix}
A & 0 \\
0 & 1
\end{pmatrix} \begin{pmatrix}
S_n \\
T_n^p - \mathbb{E}T_n^p
\end{pmatrix} + o_p(1) \to \mathcal{D} \mathbb{N}(0, \mathbb{W}_p).
\]

The asymptotic independence of \( \bar{S}_n \) and \( T_n^p - \mathbb{E}T_n^p \) follows from the asymptotic normality and asymptotic uncorrelatedness of \( \bar{S}_n \) and \( T_n^p - \mathbb{E}T_n^p \) by application of Slutsky’s lemma and the continuous mapping theorem with the mapping
\[
f(x_1, \ldots, x_d, y_1, \ldots, y_{ab}) = (x_1, \ldots, x_d \sum_{i=1}^a y_i - (a - 1)(b - 1)),
\]
where \( a = |G_X| \) and \( b = |G_Y| \). The independence of \( \bar{S}_n \) and \( R_n^p \) follows analogously. \ \blackslug
5.3 Post-tests based on adjusting: Proofs

In this section we focus on the post-test statistics that are based on adjusting for pre-test. We provide proofs of the theorems and corollaries formulated in Section 3.5. Those are Theorems 3.4 (Normality I), 3.5 (Independence I), 3.6 (Normality II), 3.7 (Independence II), 3.8 (Normality III), 3.9 (Independence III). These results concern the setup in which the pre-tests are based on the controls-only vectors $T_n^{co}$ and $R_n^{co}$ defined in (3.12) and (3.14), or $T_n^{ca}$ and $R_n^{ca}$ defined in (3.16) and (3.18). We used the pre-test generating vectors standardized by true probabilities instead of estimators, since that does not affect the asymptotic distribution. The reader will also recall that throughout the entire Part I we always assume that all vectors that represent reordering of elements of a $k \times l$ matrix are ordered row-wise, that is according to formula $i = 3k + l$, where $i$ is the index of the elements in the vector representation.

5.3.1 Adjusted post-test centered by true expectation: Proofs

First we provide a proof of Theorem 3.4. We start by showing joint asymptotic normality of $S_n$ and $T_n$. Then, similarly to the proof of Theorem 3.3, we use the asymptotic representation of the score statistic given by Theorem 3.1 to obtain joint asymptotic normality of $S_n$ and $T_n$.

Normality of $S_n$ and $T_n^{co}$, $R_n^{co}$, $T_n^{ca}$ or $R_n^{ca}$

The following lemma is the analogue of Lemma 5.10 combined with Lemma 5.9.

Lemma 5.11 Let $S_n$ be defined by (3.3) and put $T_n = T_n^{co}$ of (3.12), $T_n = R_n^{co}$ of (3.14), or $T_n = T_n^{ca}$ of (3.16), or $T_n = R_n^{ca}$ of (3.18), and let $H_0^f$ hold. Then, the random vector $(S_n', T_n - E T_n')'$ has zero expectation and as $n \to \infty$ it converges in distribution to the centered normal distribution with variance matrix

$$W_0 = \begin{pmatrix} I_{Z_0} & C_{ST} \\ C_{ST} & V_T \end{pmatrix},$$

where $I_{Z_0}$ is the Fisher information matrix, $V_T$ is the asymptotic variance matrix of $T_n$ and $C_{ST}$ is the asymptotic covariance matrix of $S_n$ and $T_n$.

Proof. The proof is analogous to the proofs of Lemmas 5.9 and 5.10. We only show the case $T_n = T_n^{co}$, since the other cases are completely analogous. We write the cell and marginal counts among controls as $N_{kl} = \sum_{p=1}^n (1 - \Delta_i) I_{x[i] = k, y[i] = l}$, $N_k = \sum_{i=1}^n (1 - \Delta_i) I_{x[i] = k}$ and $N_l = \sum_{i=1}^n (1 - \Delta_i) I_{y[i] = l}$. Recalling the assumption (3.2) and using the central limit theorem, we have the asymptotic normality of $n^{-1/2} \sum_{i=1}^n Z_i$, where we put

$$Z_i = \begin{pmatrix} (\Delta_i - \Psi^0_i) z(X_i, Y_i) \\ ((1 - \Delta_i) I_{x[i] = k, y[i] = l} - p_k)_{k,l}^l \\ ((1 - \Delta_i) I_{x[i] = k} - p_k)_{k,l} \\ ((1 - \Delta_i) I_{y[i] = l} - q_l)_{k,l} \end{pmatrix}.$$
Then, using the delta method we get the asymptotic normality of \((S_n', Z_n^{co'})',\) where
\[
Z_n^{co} = (\tau n)^{1/2}\left( (\hat{p}_{kl} - p_k q_l - p_k q_l + p_k q_l) / \sqrt{p_k q_l} \right)_{k,l}', \tag{5.18}
\]
with \(\tau\) being the limit of \(\tau_n\) as in (3.2). By adding \((\hat{p}_k - p_k)(q_l - q_l)\) we made the expectation of \(Z_n^{co}\) zero, since then for \(Z_n^{co}\) it also holds
\[
Z_n^{co} = (\tau n)^{1/2}\left( (\hat{p}_{kl} - p_k q_l - p_k q_l + p_k q_l) / \sqrt{p_k q_l} \right)_{k,l}', \tag{5.19}
\]
This implies that the asymptotic expectation of \(Z_n^{co}\) is also zero. Moreover, since \((S_n', (T_n^{co} - ET_n^{co})')' = (\tau_n/\tau)^{1/2} (S_n', Z_n^{co'})' + O_p(n^{-1/2}),\) Slutsky’s lemma yields that the asymptotic distributions of the vectors \((S_n', (T_n^{co} - ET_n^{co})')'\) and \((S_n', Z_n^{co'})'\) are the same, hence the asymptotic normality with zero mean of \((S_n', (T_n^{co} - ET_n^{co})')'\). The upper-left block term of the asymptotic variance matrix follows from the variance matrix of \(S_n\) being \(I_p\), while the rest of the terms are only a matter of notation. \(\blacksquare\)

In the light of the asymptotic representation of \(\tilde{S}_n\) given by Theorem 3.1 and the joint asymptotic normality of \(S_n\) and \(T_n\) given by Lemma 5.11, the proof of Theorem 3.4 is quite straightforward.

**Proof of Theorem 3.4 (Normality I) of Section 3.5**

First recall the definition of \(W_{ST}^\gamma\) in Theorem 3.4, then use the result of Theorem 3.1, i.e. \(\tilde{S}_n = AS_n + o_p(1)\). For either \(T_n = T_n^{co}\) or \(T_n = T_n^{ca}\) the combination of the asymptotic representation of \(\tilde{S}_n\) with Lemma 5.11 yields an analogous formula to (5.17), which is
\[
\left( \begin{array}{c} \tilde{S}_n \\ T_n - ET_n \end{array} \right) = \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix} \left( \begin{array}{c} S_n \\ T_n - ET_n \end{array} \right) + o_p(1) \rightarrow D N(0, W_{ST}). \tag{5.20}
\]
This shows that the vectors \((\tilde{S}_n', (T_n^{co} - ET_n^{co})')'\) as well as \((\tilde{S}_n', (T_n^{ca} - ET_n^{ca})')'\) are both also asymptotically zero-mean normal under \(H_0^g\). \(\blacksquare\)

Next we prove Theorem 3.5. The proof relies on the joint asymptotic normality of \((\tilde{S}_n', (T_n^{co} - ET_n^{co})')'\) and \((\tilde{S}_n', (T_n^{ca} - ET_n^{ca})')'\) provided by Theorem 3.4.

**Proof of Theorem 3.5 (Independence I) of Section 3.5**

Put either \(T_n = T_n^{co}\) or \(T_n = T_n^{ca}\). Since \(B^* = AC_{ST}V_T\) is a constant, we can combine Theorem 3.4 and the continuous mapping theorem and get
\[
\begin{pmatrix} \gamma_n^s \\ T_n - ET_n \end{pmatrix} = \begin{pmatrix} 1 & -B^* \\ 0 & 1 \end{pmatrix} \left( \begin{array}{c} \tilde{S}_n \\ T_n - ET_n \end{array} \right) + o_p(1). \tag{5.21}
\]
This implies that \(\gamma_n^s\) and \(T_n - ET_n\) are also jointly asymptotically normal. As shown by Theorem 3.4, the asymptotic variance matrix of \((S_n', T_n^{co} - ET_n^{co})'\) is
\[
W_{ST} = \begin{pmatrix} A I_p A' & AC_{ST} \\ C_{ST} A' & V_T \end{pmatrix}.
\]
In light of (5.21) it follows that the asymptotic variance matrix of \((\gamma_n^*, T_n - \mathbb{E}T_n)')'\) is

\[
\begin{pmatrix}
1 & -\mathbf{B}^* \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
\mathbb{A}C_{ST} \mathbb{A}' & \mathbb{A}C_{ST}^T \\
C_{ST}^T \mathbb{A}' & \mathbb{V}_T
\end{pmatrix}
\begin{pmatrix}
1 & 0 \\
-\mathbf{B}^* & 1
\end{pmatrix}
\begin{pmatrix}
\mathbf{V}_n^* & 0 \\
0 & \mathbb{V}_T
\end{pmatrix}
\]

where \(\mathbf{V}_n^* = \mathbb{A}(\mathbb{I}_n - C_{ST} \mathbb{V}_T^T C_{ST}) \mathbb{A}'\), which implies that \(\gamma_n^*\) and \(T_n\) are asymptotically independent. Additional usage of the continuous mapping theorem and Slutsky’s lemma implies the same for \(\gamma_n^*\) with \(T_n^{co}\) and \(\gamma_n^*\) with \(R_n^{co}\) if \(\gamma_n^*\) is based on \(T_n^{co}\). If \(\gamma_n^*\) is based on \(T_n^{ca}\), the asymptotic independence holds for \(\gamma_n^*\) with \(T_n^{ca}\) and \(\gamma_n^*\) with \(R_n^{ca}\). Finally, the marginal asymptotic normality of \(\gamma_n^*\) and its asymptotic variance immediately follow, while the asymptotic zero expectation of \(\gamma_n^*\) comes from \(\mathbb{S}_n\) having asymptotically zero expectation under \(H_0^c\) and \(T_n - \mathbb{E}T_n\) having zero expectation always.

\[\Box\]

### 5.3 Post-tests based on adjusting: Proofs

In this section we provide proofs of Theorems 3.6 and 3.7. But before we do that, we calculate the asymptotic covariance matrix \(\mathbb{S}_n\) and the pre-test generating vectors in the lemma below.

The results and the calculations leading to them are useful when formulating the proofs of Theorems 3.6 and later also 3.8. Lemma 5.12 also shows how exactly the covariance matrices differ from zero, thus showing the need for adjusting of the score statistic to account for dependence with the pre-test.

**Lemma 5.12.** Let \(C_{ST}^{co}, C_{SR}^{co}, C_{ST}^{ca}, C_{SR}^{ca}\) respectively be the asymptotic covariance matrices under \(H_0^c\) of \(\mathbb{S}_n\) with \(T_n^{co}\), \(R_n^{co}\), \(T_n^{ca}\) and \(R_n^{ca}\), and let \(\tau\) be defined in (3.2). The columns \(c_{3k+1}^{co}, k, l = 0, 1, 2,\) of \(C_{ST}^{co}\), and the columns \(d_{3k+1}^{co}, k, l = 0, 1, 2,\) of \(C_{SR}^{co}\) are

\[
c_{3k+1}^{co} = \tau^{-1/2}(p_k q_l)^{-1/2} c_{kl} - p_k \sum_j c_{kj} (p_k q_l)^{-1/2} - q_l \sum_i c_{il} (p_k q_l)^{-1/2},
\]

\[
d_{3k+1}^{co} = \tau^{-1/2}(c_{kl} - p_k \sum_j c_{kj} - q_l \sum_i c_{il}),
\]

where \(c_{kl} = -\Psi'_i(p_0 + \beta_1 k + \beta_2 l),\) with \(p_k, q_l\) defined in (3.9) and \(\pi_{kl}\) in (3.5), \(z(a,b) = (1, a, b, z(a, b))'\) and \(\Psi'_i(x) = x/(1 + e^{-x})^2\). Moreover, if \(r_{kl} = p_{kl}\) for all \(k, l\), then the columns of \(C_{ST}^{ca}\) satisfy \(c_{3k+1}^{ca} = -((\tau/(1-\tau))^{1/2} d_{3k+1}^{co}\) and the columns of \(C_{SR}^{ca}\) satisfy \(d_{3k+1}^{ca} = -(\tau/(1-\tau))^{1/2} d_{3k+1}^{ca}\). Finally, in all cases the finite sample covariance matrices differ from the asymptotic ones by terms of order \(O(n^{-1/2})\).

**Proof.** We only explicitly calculate the result for \(S_n\) and \(T_n^{co}\). Recall \(Z_{n}^{co}\) defined in (5.18). Since \((S_n', (T_n^{co} - \mathbb{E}T_n^{co})))'\) and \((S_n', Z_{n}^{co})')'\) have the same asymptotic distribution, we can determine \(C_{ST}^{co}\) by calculating \(\text{cov}(S_n, Z_{n}^{co})\), which in turn is determined by \(\text{cov}(S_n, \overline{p}_{kl})\), \(\text{cov}(S_n, \overline{q}_{kl})\) and \(\text{cov}(S_n, \overline{q}_{kl})\). We again write \(S_n = n^{-1/2} \sum_{i=1}^{n} (\Delta_i - \Psi'_i) z(X_i, Y_i)\) and \(\overline{p}_{kl} = n^{-1} \sum_{i=1}^{n} (1 - \Delta_i) I_{X_i=k, Y_i=l}\). When calculating the covariance, the cross-product terms with \(i \neq j\) drop out due to independence of individuals. Since also \(\mathbb{E}S_n = 0\) under \(H_0^c\), the calcula-
tion essentially boils down to
\[
C_{kl} = \mathbb{E}((\Delta_1 - \Psi_1(x_1, y_1)) z(x_1, y_1) (1 - \Delta_1) I_{[x_1 = k, y_1 = l]}) \\
= \mathbb{E}(\mathbb{E}((\Delta_1 - \Psi_1(x_1, y_1)) (1 - \Delta_1) I_{[x_1 = k, y_1 = l]} | x_1, y_1)) \\
= \mathbb{E}(\mathbb{E}((\Delta_1 - \Psi_1) (1 - \Delta_1) | x_1, y_1) z(x_1, y_1) I_{[x_1 = k, y_1 = l]}) \\
= -\mathbb{E}(\Psi_1^0 (1 - \Delta_1) z(x_1, y_1) I_{[x_1 = k, y_1 = l]}) \\
= -\mathbb{E}(\Psi_1^0 (\beta_0 + \beta_1 x_1 + \beta_2 y_1) z(x_1, y_1) I_{[x_1 = k, y_1 = l]}) \\
= -\Psi_1^0 (\beta_0 + \beta_1 k + \beta_2 l) z(k, l) \mathbb{E}I_{[x_1 = k, y_1 = l]} = -\Psi_1^0 (\beta_0 + \beta_1 k + \beta_2 l) z(k, l) \pi_{kl},
\]

where \(\pi_{kl}\) is defined in (3.5) and \(z(x, y) = (1, x, y, z(x, y))'\) and \(\Psi_1^0(x) = e^{-x^2}(1 + e^{-x^2})^2\) satisfies \(\Psi_1^0(x) = \Psi_1^0(x)(1 - \Psi_1^0(x))\). Moreover, since \(\hat{p}_k = \sum_l \hat{p}_{kl}\) and \(\hat{q}_l = \sum_k \hat{p}_{kl}\), the covariances \(\text{cov}(S_n, Z_{n}^{co})\) and \(\text{cov}(S_n, \hat{q}_l)\) have respective columns
\[
c_k = \mathbb{E}((\Delta_1 - \Psi_1^0) z(X_1, Y_1) (1 - \Delta_1) I_{[X_1 = k]}) = \sum_j c_{kj},
\]
\[
c_l = \mathbb{E}((\Delta_1 - \Psi_1^0) z(X_1, Y_1) (1 - \Delta_1) I_{[Y_1 = l]}) = \sum_i c_{il}.
\]
Combining these results with (5.19) yields that the matrix \(\text{cov}(S_n, Z_{n}^{co})\) has column vectors
\[
c_{kl}^{co} = c_{kl}(\tau p_k q_l)^{-1/2} - p_l \sum_j c_{kj}(\tau p_k q_l)^{-1/2} - q_l \sum_i c_{il}(\tau p_k q_l)^{-1/2}.
\]
Since these terms do not depend on \(n\), they give the asymptotic covariance matrix \(C_{ST}^{co}\).

As far as the covariance matrices \(C_{ST}^{ca}\) and \(C_{SR}^{ca}\) go, we only note that they can be calculated completely analogously by focusing on \(\mathbb{E}((\Delta_1 - \Psi_1^0) z(X_1, Y_1) \Delta_1 I_{[X_1 = k, Y_1 = l]}))\), from which it is also easy to see where the reversed signs relative to \(C_{ST}^{co}\) and \(C_{SR}^{co}\) come from.

Finally, the fact that the finite sample and the asymptotic covariance matrices differ only by an \(O(n^{-1/2})\) term can be shown directly by calculating the finite sample covariance of \(S_n\) with \(N^{1/2}(\hat{p}_k - p_k)(\hat{q}_l - q_l)\) for \(T_n^{co}\) and \(R_n^{co}\), or \(S_n\) with \(M^{1/2}(\hat{r}_k - r_k)(\hat{s}_l - s_l)\) for \(T_n^{ca}\) and \(R_n^{ca}\). These are the terms by which the pre-test generating vectors \(T_n^{co}, R_n^{co}, T_n^{ca}\) and \(R_n^{ca}\) differ (up to scaling by constants) from an iid sequence (see (5.19)), and which therefore determine the difference between the finite sample size and asymptotic covariance matrices.

**Proof of Theorem 3.6 (Normality II) of Section 3.5**

The proof is again analogous to the proof of Lemma 5.10 combined with the proof of Theorem 3.3. We define random variables \(B_i \in [0, 1], i = 1, \ldots, n\) as indicators whether or not the \(i\)-th individual was selected into the subsample from which we compute \(T_n^{co}\). This means that \(B_i = 0\) for all cases \((\Delta_1 = 1)\) and \(\sum_i B_i = N^{\delta}\). The rest of the proof is formulated only for the control-only pre-test based on \(T_n = T_n^{co}\) with \(T_n^{co}, T_n^{SC}\) defined correspondingly. The other cases follow analogously.

We start by writing the pre-test sample counts as \(N_{kl}^{\delta} = \sum_{i=1}^{n} B_i (1 - \Delta_1) I_{[X_i = k]}\) and \(N_{jl}^{\delta} = \sum_{i=1}^{n} B_i (1 - \Delta_1) I_{[Y_i = l]}\), while analogous equalities holds for
N^iδC, N^hδC, and N^lδC. Using the central limit theorem we get the asymptotic normality of \( n^{-1/2} \sum_{i=1}^n Z^i_\delta \), where

\[
Z^i_\delta = \begin{cases} 
B_i(1-\Delta_i)(I_{[X_i=k,Y_i=l]}-p_{kl}, I_{[Y_i=l]}-q_l)^{k,l}_i, \\
(1-B_i)(1-\Delta_i)(I_{[X_i=k,Y_i=l]}-p_{kl}, I_{[X_i=k]}-p_k, I_{[Y_i=l]}-q_l)^{k,l}_i.
\end{cases}
\]

Then, using the delta method, we get the asymptotic normality of \( (S'_n, Z^{β'}, Z^{δC})' \), where

\[
Z^{β} = (δτn)^{1/2}(p_{kl}^δ - p_{kl}(q_l^δ - q_l) - q_l(p_k^δ - p_k)/√p_kq_l)_k,l',
\]

\[
Z^{δC} = ((1-δ)τn)^{1/2}(p_{kl}^{δC} - p_{kl}(q_l^{δC} - q_l) - q_l(p_k^{δC} - p_k)/√p_kq_l)_k,l',
\]

with \( p_{kl}^δ, p_{kl}^{δC}, q_l^δ, q_l^{δC} \) defined in (3.21) and (3.23). Since

\[
(S'_n, (T_n^δ - E T_n^δ)', (T_n^{δC} - E T_n^{δC})')' = a_n(S'_n, Z^{β'}, Z^{δC})' + O_p(n^{-1/2}),
\]

where \( a_n = (δτn)/δN \rightarrow 1 \), Slutsky’s lemma yields the asymptotic normality of \( (S'_n, (T_n^δ - E T_n^δ)', (T_n^{δC} - E T_n^{δC})')' \) with zero mean. Therefore, the rest of the argument for asymptotic normality follows from the delta method, Slutsky’s lemma and the asymptotic representation \( S_n = AS_n + o_p(1) \) given by Theorem 3.1 using the same type of argument as in (5.17) and (5.20). As far as the asymptotic variance matrix goes, the diagonal block elements come from the asymptotic variance of \( S_n \) and the fact that \( T_n^δ \) and \( T_n^{δC} \) have the same asymptotic variance as \( T_n \) (see Lemmas 5.16 and 5.17). The zero covariance blocks of \( T_n^δ \) and \( T_n^{δC} \) come from the fact that \( T_n^δ \) and \( T_n^{δC} \) are independent, and finally the covariance blocks of \( S_n \) with \( T_n^δ \) and \( T_n^{δC} \) follow since \( cov(S_n, Z^{δ}) \) and \( cov(S_n, Z^{δC}) \) have column vectors \( e^{δ}_{3k+l} = \sqrt{δ}c^{δ}_{3k+l} \) and \( e^{δC}_{3k+l} = \sqrt{1-δ}c^{δ}_{3k+l} \), respectively, with \( k, l = 0, 1, 2 \) and \( c^{δ}_{3k+l} \) defined in Lemma 5.12.

**Proof of Theorem 3.7 (Independence II) of Section 3.5**

Under \( H_0^δ \), by Theorem 3.6, for any \( δ ∈ (0,1] \) the vectors \( \hat{S}_n, T_n^δ - E T_n^δ \) and \( T_n^{δC} - E T_n^{δC} \) are jointly asymptotically normal with zero mean and variance matrix \( W^δ \) defined in Theorem 3.6. By Lemma 5.8 the vectors \( T_n^δ \) and \( \sqrt{δ/(1-δ)}T_n^{δC} \) have equal distributions up to a term \( O_p(n^{-1/2}) \), therefore \( γ^γ_δ \) has vanishing expectation. It follows, by an argument similar to (5.21), which is based on the continuous mapping theorem and Slutsky’s lemma. Using \( B^δ = √δACSTV_T \), the argument goes

\[
\left(\begin{array}{c}
γ^γ_δ \\
T_n^δ - E T_n^δ
\end{array}\right) = \left(\begin{array}{cc}
1 & -B^δ \\
0 & I
\end{array}\right) \left(\begin{array}{c}
\hat{S}_n \\
T_n^δ - E T_n^δ \\
T_n^{δC} - E T_n^{δC}
\end{array}\right) + o_p(1).
\]

This shows that \( γ^γ_δ \) and \( T_n^δ - E T_n^δ \) are jointly asymptotically normal with zero mean and variance matrix \( MW^δM' \), where \( W^δ \) is defined in Theorem 3.6 and we denoted the transformation
matrix in the display above by $M$. Since

$$MW^\delta = \begin{pmatrix} \frac{A\hat{r} A'}{\sqrt{\delta}} & 0 & \sqrt{1-\delta}AC_{ST} + (\delta/\sqrt{1-\delta})AC_{ST}\end{pmatrix},$$

then

$$MW^\delta M^\prime = \begin{pmatrix} V_0^\delta & 0 \\ 0 & V_T \end{pmatrix},$$

where $V_0^\delta = A\hat{r} A' + \frac{\delta}{\sqrt{\delta}}AC_{ST}V_T C_{ST}' A'$. This means that $\gamma_n^\delta$ and $T_n^\delta - ET_n^\delta$ are asymptotically independent. Using the continuous mapping theorem and Slutsky's lemma once more yields the asymptotic independence for $\gamma_n^\delta$ and $T_n^\delta$ as well as $\gamma_n^\delta$ and $R_n^\delta$. Marginal asymptotic normality of $\gamma_n^\delta$ with zero expectation and variance matrix $V_0^\delta$ simply follows. □

5.3.3 Adjusted post-test based on centering by the other sample: Proofs

In this section we formulate the proofs of Theorems 3.8 and 3.9. In line with Section 3.5.3 being analogous to Section 3.5.2, the following proofs are analogous to the proofs of Theorems 3.6 and 3.7. We sketch these proofs mainly for reasons of completeness.

Proof of Theorem 3.8 (Normality III) of Section 3.5

The proof is again analogous to the proof of Lemma 5.10 combined with the proof of Theorem 3.3 and the asymptotic normality of $n^{-1/2}\sum_i Z_i^cc$, where

$$ Z_i^cc = \left( (1 - \Delta_i)\left( I_{X_i=k, Y_i=l} - p_{kl} I_{X_i=k} - p_{k, I_{Y_i=l} = q_l} \right) \right)_{k,l}, $$

implied by the central limit theorem. Delta method then provided the asymptotic normality of $(S_n^c, Z_n^{ca}, Z_n^{ca})', \text{ where } Z_n^{ca}$ is the same as in (5.19), and

$$ Z_n^{ca} = ((1 - \tau) n)^{1/2}\left( (\hat{r}_k - r_k) - r_k (\hat{s}_l - s_l) - s_l (\hat{r}_k - r_k)/\sqrt{\hat{r}_k s_l} \right)_{k,l}. $$

Continuous mapping theorem, Slutsky's lemma and the asymptotic representation $\bar{S}_n = AS_n + o_P(1)$ provide the rest of the argument for the asymptotic distribution claims about $(\bar{S}_n, (T_n^{co} - ET_n^{co}))', (T_n^{ca} - ET_n^{ca})'$, where the argument is essentially the same as in the proofs of the Normality I and II theorems. The asymptotic variance matrix can be derived analogously to that in Theorem 3.6. The diagonal block elements come from the asymptotic variance of $\bar{S}_n$ and the fact that $T_n^{co}$ and $T_n^{ca}$ have the same asymptotic variance if $p_{kl} = r_{kl}$ for all $k, l$ (see Lemmas 5.16 and 5.17). The zero covariance blocks of $T_n^{co}$ and $T_n^{ca}$ come from their independence, the covariance blocks of $\bar{S}_n$ with $T_n^{co}$ and $\bar{S}_n$ with $T_n^{ca}$ follow under $r_{kl} = p_{kl}$ for all $k, l$ from Lemma 5.12.
Finally, the asymptotic normality results for the vectors \((\hat{S}_n^e, Y_n^{co'}, (T_n^{co} - \mathbb{E}T_n^{co})')\) and \((\hat{S}_n^o, Y_n^{co'}, (T_n^{oa} - \mathbb{E}T_n^{oa})')\) follow from the above result using the same type of argument as in (5.21) and (5.23), the zero asymptotic expectations are obvious, and the asymptotic covariances in Propositions 5.14 and 5.15. We formulate the results below as asymptotic independence of both \(T_n^{co}\) and \(T_n^{oa}\) only by different standardization. \(\square\)

**Proof of Theorem 3.9 (Independence III) of Section 3.5**

For either of the considered choices of \(T_n\) and \(Y_n\), Theorem 3.8 provided asymptotic normality of both \((\hat{S}_n^e, Y_n', (T_n - \mathbb{E}T_n)')\). From that the asymptotic normality of \((\gamma_n^Y, (T_n - \mathbb{E}T_n)')\) follows by an argument similar to (5.21) and (5.23). Clearly, the linear transformation which turns \((\hat{S}_n^e, Y_n', (T_n - \mathbb{E}T_n)')\) into \((\gamma_n^Y, (T_n - \mathbb{E}T_n)')\) (up to \(o_p(1)\) terms) preserves the desired asymptotic independence of \(\gamma_n^Y\) with \(T_n - \mathbb{E}T_n\). The asymptotic independence of \(\gamma_n^Y\) with \(T_n^{co}\), or \(R_n^{co}\), or \(T_n^{oa}\) or \(R_n^{oa}\) follows by continuous mapping theorem and Slutsky’s lemma. Finally, the asymptotic variance of \(\gamma_n^Y\) in all cases are matters of simple calculation similar to those performed in the proof of Theorem 3.7. \(\square\)

### 5.3.4 Post-tests based on adjusting for parallel pre-tests: Extensions

In this subsection we formulate several propositions that extent the results of Section 3.5.4, which focuses on adjusting for several parallel pre-test vectors. The idea is to obtain a post-test statistic that is independent of multiple parallel pre-tests. We first formulate asymptotic normality result in Proposition 5.13, which is then followed by two additional results about asymptotic covariances in Propositions 5.14 and 5.15. We formulate the results below as propositions since we do not provide their proofs. However, a closer look at the results shows that the arguments required for these proofs would be analogous to those presented in the preceding sections.

**Proposition 5.13 (Normality IV)** Assume that for the \(i\)-th pair of loci, \(i \in \{1, \ldots, K\}\), the null hypothesis \(H_{0i}^e\) holds. Let \(\hat{S}_n^{(i)}\) be the corresponding full-sample score statistic defined by (3.4) and let \(X_{n1}^\delta, \ldots, X_{nR_i}^\delta\) be defined by (3.39) for fixed values of \(R_i\) and \(\delta \in (0, 1)\) with the corresponding pre-test samples the same in all vectors \(X_{n1}^\delta, \ldots, X_{nR_i}^\delta\). Then, as \(n \to \infty\), the random vector \(\Psi_n^{(i)} = (\hat{S}_n^{(i)'}, X_{n1}^{oR_i}, \ldots, X_{nR_i}^{oR_i})\) converges in distribution to a zero-mean normal distribution with variance

\[
W^{(R_i)}_{SSX} = \begin{pmatrix}
(1 - \delta)A \bar{T}_g A' & 0 & 0 & \cdots & 0 \\
0 & V_{(1)}^{(1)} & C_{TT}^{(12)} & \cdots & C_{TT}^{(1R_i)} \\
0 & C_{TT}^{(21)} & V_{TT}^{(22)} & \cdots & C_{TT}^{(2R_i)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & C_{TT}^{(R_i1)} & C_{TT}^{(R_i2)} & \cdots & V_{TT}^{(R_iR_i)}
\end{pmatrix}
\]
Proposition 5.14

Fix a normal distribution with variance $\mathbf{B}$ by (3.41) for suitable matrices $\mathbf{B}_r$. Moreover, denote the loci of the genotype variables that all of the genotype variables take values in $S$. Clearly, $C_{\mathbf{1}_{j,k;L,L}}$ are the asymptotic variance matrices of $\mathbf{B}_r$. Then, under $H_{0j}'$, it holds $\gamma_{ij}$ converges to zero-mean normal distribution with variance

$$\text{var} \gamma_{ij} = A_T A' + \frac{1}{t} \sum_{r=1}^{K} \mathbf{B}_r \mathbf{B}_r' + \frac{1}{t} \sum_{r=1}^{K} \mathbf{B}_r \mathbf{C}_{TT} \mathbf{B}_r' + o(1).$$

(5.24)

Next, we compute $\text{cov}(S_{ij}, T_{ij})$ and $\text{cov}(T_{ij}, T_{ij})$, where $S_{ij}$ and $T_{ij}$ is the full-sample score statistic of the $i$-th loci pair with true values of the parameter and $T_{ij}$ is the partial sample pre-test generating vector of the $j$-th loci. In light of the asymptotic representation of the score statistic $S_{ij}$, given by Theorem 3.1, the covariance $\text{cov}(S_{ij}, T_{ij})$ is asymptotically linked to $\text{cov}(S_{ij}, T_{ij})$.

**Covariance of $S_{ij}$ and $T_{ij}$**

$T_{ij}$ is defined as an analogue of $T_{ij}$ defined in (3.21), which in formula means

$$T_{ij} = \left((N^\delta)^{1/2} \left( \mathbf{P}_k^{\delta,ij} - \mathbf{P}_k^{\delta,ij} \mathbf{q}_l^{\delta,ij} \right) \left( \mathbf{P}_k^{\delta,ij} \mathbf{q}_l^{\delta,ij} \right)^{1/2} \right)_{k,l=0,1},$$

where $\mathbf{P}_k^{\delta,ij} = N_k^{ij}/N^\delta$, $\mathbf{P}_k^{\delta,ij} = N_k^{ij}/N^\delta$ and $\mathbf{q}_l^{\delta,ij} = N_l^{ij}/N^\delta$ with $N_k^{ij}$, $N_k^{ij}$, $N_l^{ij}$ the cell and marginal counts based on the $j$-th pair of loci and the chosen subsample of $N^\delta$ controls. Clearly, $\mathbf{P}_k^{\delta,ij}$, $\mathbf{P}_k^{\delta,ij}$, $\mathbf{q}_l^{\delta,ij}$ are the ML estimators of $p_k^{ij}$, $p_k^{ij}$, $q_l^{ij}$. Proposition 5.14 gives an expression for the covariance of $S_{ij}$ and $T_{ij}$.

**Proposition 5.14** Fix $i, j \in \{1, \ldots, K\}$, $j \neq i$. Denote $T_{ij}$ the pre-test generating vector for the $j$-th pair of loci computed on the full sample of $N$ controls and denote by $C_{ST}^{ij}$ the asymptotic covariance matrix of $S_{ij}$ and $T_{ij}$. Then, $\text{cov}(S_{ij}, T_{ij}) = C_{ST}^{ij} + o(1)$ and for any $\delta \in [0,1]$ under $H_{0j}$ it holds $\text{cov}(S_{ij}, T_{ij}) = \sqrt{N} C_{ST}^{ij} + o(1)$ and $\text{cov}(S_{ij}, T_{ij}^{\gamma_c}) = \sqrt{1 - \delta} C_{ST}^{ij} + o(1)$. Moreover, denote the loci of the $i$-th and $j$-th pairs respectively by $L_i, L_j$ and $L_i, L_j$. Assume that all of the genotype variables take values in $[0,1,2]$. Further assume that the phenotype $\Delta_{ij}$ is fully determined by genotypes $X_{ij}, Y_{ij}$ at $L_i, L_j$, meaning that

$$\mathbb{P}(\Delta_{ij} = 0 | X_{ij}, Y_{ij}) = \mathbb{P}(\Delta_{ij} = 0 | X_{ij} = x, Y_{ij} = y).$$

(5.25)

for all $k,l,x,y$, where $W_{ij}, Z_{ij}$ are the genotypes at loci $L_i, L_j$, of the $v$-th individual. Then, $C_{ST}^{ij}$ is given by its columns $a_{3k+l}^{ij}$, which are equal to

$$a_{3k+l}^{ij} = c_{3k+l}^{ij} - \sum_{s} \sqrt{r_{n,s} \pi_r} c_{3k+l}^{ij} - \sum_{t} \sqrt{r_{n,k,r} \pi_l} c_{3k+l}^{ij},$$

(5.26)
where
\begin{align*}
\psi^{(ij)}_{3k+1} &= -\sum_x \sum_y \psi_1'(\beta_0 + \beta_1 k + \beta_2 y) z(x, y) \pi_{xk}/\sqrt{\tau_n \omega_{k_l}}, \quad \text{for } L_a \neq L_c, L_b \neq L_d \\
\psi^{(ij)}_{3k+1} &= -\sum_x \sum_y \psi_1'(\beta_0 + \beta_1 k + \beta_2 y) z(x, l) \pi_{xk}/\sqrt{\tau_n \omega_{k_l}}, \quad \text{for } L_a \neq L_c, L_b = L_d \\
\psi^{(ij)}_{3k+1} &= -\sum_y \sum_y \psi_1'(\beta_0 + \beta_1 k + \beta_2 y) z(k, y) \pi_{yk}/\sqrt{\tau_n \omega_{k_l}}, \quad \text{for } L_a = L_c, L_b \neq L_d,
\end{align*}
with \( \tau_n = N/n, z(a, b) = (1, a, b, z(a, b))^T, \psi_1'(x) = e^{-x}/(1 + e^{-x})^2, \) and
\begin{align*}
\pi_{xk} &= \mathbb{P}(X_1 = x, Y_1 = y, W_1 = k, Z_1 = l), \quad \omega_{k_l} = \mathbb{P}(W_1 = k | \Delta_1 = 0), \\
\pi_{xk} &= \mathbb{P}(X_1 = x, W_1 = k, Z_1 = l), \quad \omega_{k_l} = \mathbb{P}(Z_1 = l | \Delta_1 = 0), \\
\pi_{yk} &= \mathbb{P}(Y_1 = y, W_1 = k, Z_1 = l).
\end{align*}

**Proof.** Even though it is more involved, the proof of the proposition is essentially analogous to the combined calculations that yielded the results in Lemma 5.12.

We note that the assumption (5.25) is not essential and it could be dropped. In that case the terms \( \psi_1'(\beta_0 + \beta_1 x + \beta_2 y) \) in the formulas for \( \psi^{(ij)}_{3k+1} \) would need to be replaced by \( \psi_1(\beta_0 + \beta_1 x + \beta_2 y) \mathbb{P}(\Delta_1 = 0 | W_1 = k, Z_1 = l, x, y) \).

**Covariance of \( T_{ni}^3 \) and \( T_{nj}^3 \)**

As a conclusion to this chapter, we focus on the covariance of \( T_{ni}^3 \) and \( T_{nj}^3 \). For \( i = j \) the covariance follows from Lemma 5.16 where the probabilities \( p_{kl}, p_k, q_l \) correspond to the \( i \)-th pair of loci. The covariance matrix for \( i \neq j \) is provided by the following proposition, whose proof is analogous to the proof of Lemma 5.16 and is omitted.

**Proposition 5.15** Fix \( i, j \in \{1, \ldots, K\}, j \neq i \). Using an obvious extension of the notation used for instance in (3.12), the elements of the asymptotic covariance matrix of \( T_{ni}^3 \) and \( T_{nj}^3 \), denoted as \( C_{ij}^{(3)} = \{u_{ab}^{(ij)}\} \), where \( a = 3k + l \) and \( b = 3r + s \) and \( k, l, r, s = 0, 1, 2 \), are
\begin{align*}
u_{ab}^{(ij)} &= \left( p_k, q_l, p_{kl}, q_{ls}, q_{j}^{(i)} \right)^{-1/2} \left[ p_{kl} - p_k p_{ls} - q_l p_{ij} - p_k q_{ij} - p_{ij} q_{ij} - p_{ij} q_{ij} - p_k q_{ij} - q_l p_{ij} - q_l q_{ij} - q_l q_{ij} + q_l q_{ij} + q_l q_{ij} + p_k q_{ij} - p_k q_{ij} + q_l p_{ij} - q_l p_{ij} + q_l q_{ij} - q_l q_{ij} \right] \\
&+ O_p\left( \frac{1}{\sqrt{N}} \right).
\end{align*}

### 5.4 Auxiliary results: Formulas for variance matrices

In this section we provide explicit formulas for asymptotic variance and covariance matrices of the pre-test and post-test vectors.
5 Two-stage testing for epistasis: Proofs

5.4.1 Variance of pre-test generating vectors

In Lemma 5.8 we calculated the expectation of the pooled-sample pre-test generating vector $T_{n}^{po}$. The result we formulated in that lemma can also be modified to apply to the other pre-test generating vectors. For instance, taking the expectation provided by Lemma 5.8 and replacing $n$ by $N$ and $\pi_{kl}, \pi_{k},$ and $\pi_{l}$ by $p_{kl}, p_{k}$ and $q_{l}$ yields the expectation of $T_{n}^{co}$. The following lemma adds to these results by providing a formula for the asymptotic variance matrix of $T_{n}^{co}$.

Lemma 5.16 Let $T_{n}^{co}$ be defined by (3.12) and denote by $V_{T}^{n} = (v_{ij}^{n})$ the variance matrix of $T_{n}^{co}$ and assume its elements are ordered inside the vector according to numbering given by $3k+l$, where $k, l = 0, 1, 2$. Then, up to terms of order $o(1)$, the elements of the matrix $V_{T}^{n} = (v_{ij})$, where $i = 3k+l$ and $j = 3r+s$ for $k, l, r, s = 0, 1, 2$, are equal to

\[
v_{ii} = \frac{p_{kl}(1 + p_{k}q_{l} - p_{k} - q_{l}) + (p_{k} + q_{l})(p_{k}q_{l} - p_{kl}) + 4p_{k}q_{l}(p_{kl} - p_{k}q_{l})}{p_{k}q_{l}}, \quad \text{for } k = r, l = s, \quad (5.27)
\]

\[
v_{ij} = \frac{p_{kl}(2p_{r}q_{s} - p_{rs}) + 2p_{k}q_{l}(p_{rs} - p_{r}q_{s}) + p_{r}q_{l}(p_{ks} - p_{k}q_{s}) + p_{k}q_{s}(p_{rl} - p_{r}q_{l})}{\sqrt{p_{k}q_{l}p_{r}q_{s}}}, \quad \text{for } k \neq r, l \neq s, \quad (5.28)
\]

\[
v_{ij} = \frac{p_{r}(2p_{k}q_{l} - p_{kl}) + p_{r}q_{l}(3p_{kl} - p_{k}q_{l}) + p_{k}q_{l}(p_{rl} - 3p_{r}q_{l}) + p_{k}(p_{rl} - p_{r}q_{l}) - p_{r}p_{kl}}{\sqrt{p_{k}q_{l}p_{r}q_{s}}}, \quad \text{for } k \neq r, l = s, \quad (5.29)
\]

\[
v_{ij} = \frac{p_{k}(2p_{k}q_{l} - p_{kl}) + p_{k}q_{s}(3p_{kl} - p_{k}q_{l}) + p_{k}q_{l}(p_{ks} - 3p_{k}q_{s}) + q_{s}(p_{kl} - p_{k}q_{l}) - q_{l}p_{ks}}{p_{k}\sqrt{q_{l}q_{s}}}, \quad \text{for } k = r, l \neq s. \quad (5.30)
\]

Before we give the calculation that proves Lemma 5.16 we note that the result above does not require that independence assumption $p_{kl} = p_{k}q_{l}$ for all $k, l = 0, 1, 2$ holds. If we additionally adopted this equality, then $E T_{n, co}^{0} = 0$ and the expressions for elements of the variance matrix simplify considerably to

\[
v_{ii} = 1 - p_{k} - q_{l} + p_{k}q_{l}, \quad \text{for } k = r, l = s, \quad v_{ij} = \sqrt{p_{k}q_{l}p_{r}q_{s}}, \quad \text{for } k \neq r, l \neq s,
\]

\[
v_{ij} = -(1 - q_{l})\sqrt{p_{k}q_{l}}, \quad \text{for } k \neq r, l = s, \quad v_{ij} = -(1 - p_{k})\sqrt{q_{l}q_{s}}, \quad \text{for } k = r, l \neq s.
\]

We also point out that while the assertions of Lemma 5.16 are formulated for $T_{n, co}^{i}$, which is based on the full sample of controls, the directly also applies to $T_{n}^{po}$, $T_{n}^{ca}$, $T_{n}^{0}$, $T_{n}^{co}$, as well as any other vector of this type, provided of course the probabilities in the lemma are replaced with the corresponding probabilities for the population on which the given pre-test generating vector is based. Hence Lemma 5.17 below. In addition to this, since $R_{n, po}^{i}$ and $R_{n}^{ca}$ only differ from the above vector by different scaling, their variances are closely related. They are treated separately in Section 5.4.2.
Lemma 5.17  Let $T_n$ be defined by any of (3.7), (3.16), (3.21), (3.27), (3.23), (3.29). The variance matrix of $T_n$ is given by Lemma 5.16 up to a vanishing term, provided we replace in (5.27)–(5.30) the probabilities $p_{kl}$, $p_k$ and $q_l$ by the probabilities that correspond to $T_n$.

Proofs of Lemmas 5.16 and 5.17

The elements of $T_n^c$ are $t_{kl} = (N/(p_k q_l))^{1/2} (\tilde{p}_{kl} - \tilde{p}_k \tilde{q}_l)$, where

$$
\tilde{p}_{kl} - \tilde{p}_k \tilde{q}_l = (\tilde{p}_{kl} - p_{kl}) - q_l (\tilde{p}_k - p_k) + (p_{kl} - p_k q_l) + O_p(N^{-1}).
$$

Let us denote the variance matrix of $T_n^c$ by $V_n^c$ and its elements by $v_{ij}^n$, where the relationship between indices $i, j$ and the indices $k, l, p, r, s$ of the elements of the vector $T_n^c$ is $i = 3k + l$ and $j = 3r + s$, with $k, l, r, s = 0, 1, 2$. Just like any other variance terms, the elements $v_{ij}^n$ have the form $v_{ij}^n = E(t_{kl} - E(t_{kl}))(t_{rs} - E(t_{rs}))$. Therefore, we are interested in the expectations of the cross products $(t_{kl} - E(t_{kl}))(t_{rs} - E(t_{rs}))$. Putting $t_{kl}^0 = t_{kl} - E(t_{kl})$, the products of random variables $t_{kl}^0$ and $t_{rs}^0$ are equal to

$$
\left[(\tilde{p}_{kl} - p_{kl})(\tilde{p}_{rs} - p_{rs}) - p_k(\tilde{q}_l - q_l)(\tilde{p}_{rs} - p_{rs}) - p_l(\tilde{p}_{kl} - p_{kl})(\tilde{q}_s - q_s) - q_l(\tilde{p}_k - p_k)(\tilde{p}_{rs} - p_{rs}) - q_s(\tilde{p}_{kl} - p_{kl})(\tilde{p}_r - p_r) + p_k q_l(\tilde{q}_l - q_l)(\tilde{q}_s - q_s) + q_l q_s(\tilde{p}_k - p_k)(\tilde{p}_r - p_r) + p_l q_s(\tilde{p}_{kl} - p_{kl})(\tilde{q}_s - q_s) + p_k q_s(\tilde{q}_l - q_l)(\tilde{p}_r - p_r)\right]
N/\sqrt{pq_0 q_r q_s} + O_p(N^{-1/2}).
$$

We need to compute the expectations of the terms $t_{kl}^0 t_{rs}^0$, which we do by determining the expectations of the products of individual terms in the previous display. These expectations are

$$
E(\tilde{p}_{kl} - p_{kl})(\tilde{p}_{rs} - p_{rs}) = E(N_{kl} N_{rs})/N^2 - p_{kl} p_{rs},
$$

$$
E(\tilde{p}_k - p_k)(\tilde{p}_{rs} - p_{rs}) = E(N_k N_{rs})/N^2 - p_k p_{rs},
$$

$$
E(\tilde{q}_l - q_l)(\tilde{p}_{rs} - p_{rs}) = E(N_l N_{rs})/N^2 - q_l p_{rs},
$$

$$
E(\tilde{p}_{kl} - p_{kl})(\tilde{q}_s - q_s) = E(N_{kl} N_s)/N^2 - p_{kl} q_s,
$$

$$
E(\tilde{p}_k - p_k)(\tilde{q}_s - q_s) = E(N_k N_s)/N^2 - p_k q_s,
$$

$$
E(\tilde{q}_l - q_l)(\tilde{q}_s - q_s) = E(N_l N_s)/N^2 - q_l q_s,
$$

$$
E(\tilde{p}_{kl} - p_{kl})(\tilde{p}_r - p_r) = E(N_{kl} N_r)/N^2 - p_{kl} p_r,
$$

$$
E(\tilde{p}_k - p_k)(\tilde{p}_r - p_r) = E(N_k N_r)/N^2 - p_k p_r,
$$

$$
E(\tilde{q}_l - q_l)(\tilde{p}_r - p_r) = E(N_l N_r)/N^2 - q_l p_r.
$$

Similarly to the proof of Lemma 5.8, we recall the variables $\Delta_i, X_i, Y_i$ defined in Section 3.2 and note that for any $k, l = 0, 1, 2$ we can write $N_{kl} = \sum_{i=1}^n (1 - \Delta_i) I_{[X_i = k]} I_{[Y_i = l]}$, $N_k = \sum_{i=1}^n (1 - \Delta_i) I_{[X_i = k]}$ and $N_l = \sum_{i=1}^n (1 - \Delta_i) I_{[Y_i = l]}$. For the sake of simplicity of the formulas in the following we assume that the first $N$ individuals of our pooled sample $(X_i, Y_i)$, $i = 1, \ldots, m$, are controls, that is $\Delta_i = 0$ for $i = 1, \ldots, N$ and $\Delta_i = 1$ for $i = N + 1, \ldots, n$. This
obviously causes no loss of generality and with this assumption we can rewrite the counts as
\( N_{kl} = \sum_{i=1}^{N} I_{\{X_i=k, Y_i=l\}} \), \( N_{k} = \sum_{i=1}^{N} I_{\{X_i=k\}} \) and \( N_{l} = \sum_{i=1}^{N} I_{\{Y_i=l\}} \).

Next, we calculate the cross-product expectations in (5.33), where the sum indices therefore range over \( 1, \ldots, N \).

**Step A1 - Diagonal Elements** \((k = r, l = s)\):

- \( E N_{kl}^2 = E \sum_{i} I_{\{X_i=k, Y_i=l\}} E \sum_{i} I_{\{X_i=k, Y_i=s\}} I_{\{X_i=k\}} I_{\{X_i=k\}} = N p_{kl} + N(N-1) p_{kl}^2 \),
- \( E N_{kl}^2 = E \sum_{i} I_{\{X_i=k\}} E \sum_{i} I_{\{X_i=k\}} I_{\{X_i=k\}} = N p_{k} + N(N-1) p_{k}^2 \),
- \( E N_{kl}^2 = E \sum_{i} I_{\{Y_i=l\}} E \sum_{i} I_{\{Y_i=l\}} I_{\{Y_i=l\}} = N q_{l} + N(N-1) q_{l}^2 \),
- \( E N_{kl}^2 = E \sum_{i} I_{\{X_i=k, Y_i=l\}} E \sum_{i} I_{\{X_i=k, Y_i=l\}} I_{\{X_i=k\}} I_{\{Y_i=l\}} = N p_{kl} + N(N-1) p_{kl} q_{l} \),
- \( E N_{kl}^2 = E \sum_{i} I_{\{X_i=k, Y_i=l\}} E \sum_{i} I_{\{X_i=k, Y_i=l\}} I_{\{X_i=k\}} I_{\{Y_i=l\}} = N p_{kl} + N(N-1) p_{kl} q_{l} \).

**Step A2 - Off-Diagonal Elements with** \( k = r, l = s \):

- \( E N_{kl} N_{ks} = E \sum_{i} I_{\{X_i=k, Y_i=l\}} I_{\{X_i=k, Y_i=s\}} = N(N-1) p_{kl} p_{ks} \),
- \( E N_{kl} N_{ks} = E \sum_{i} I_{\{Y_i=l\}} I_{\{X_i=k, Y_i=s\}} = N(N-1) p_{kl} q_{l} \),
- \( E N_{kl} N_{ks} = E \sum_{i} I_{\{X_i=k\}} I_{\{Y_i=l\}} I_{\{X_i=k\}} = N(N-1) q_{l} q_{s} \).

**Step A3 - Off-Diagonal Elements with** \( k \neq r, l = s \):

- \( E N_{kl} N_{rl} = E \sum_{i} I_{\{X_i=k, Y_i=l\}} I_{\{X_i=r, Y_i=l\}} = N(N-1) p_{kl} p_{rl} \),
- \( E N_{kl} N_{rl} = E \sum_{i} I_{\{X_i=k\}} I_{\{X_i=r\}} I_{\{Y_i=l\}} = N(N-1) p_{kl} p_{rl} \),
- \( E N_{kl} N_{rl} = E \sum_{i} I_{\{X_i=k\}} I_{\{X_i=r\}} I_{\{X_i=r\}} = N(N-1) p_{k} p_{r} \).

**Step A4 - Off-Diagonal Elements with** \( k \neq r, l \neq s \):

- \( E N_{kl} N_{rs} = E \sum_{i} I_{\{X_i=k, Y_i=l\}} I_{\{X_i=r, Y_i=s\}} = N(N-1) p_{kl} p_{rs} \).

**Step A5 - Expectations (5.33):** Using the results of Step A1 - A4 we can compute the expectations (5.33). The diagonal elements are

\[
E(\bar{p}_k - p_k)^2 = N p_{kl} + (N-1) p_{kl}^2 - N p_k^2 = p_k(1 - p_k),
\]

\[
E(\bar{p}_k - p_k)^2 = N p_k + (N-1) p_k^2 - p_k^2 = p_k(1 - p_k),
\]

\[
E(\bar{q}_l - q_l)^2 = N q_{l} + (N-1) q_{l}^2 - q_{l}^2 = q_{l}(1 - q_{l}),
\]

\[
E(\bar{p}_k - p_k)(\bar{p}_k - p_k) = p_k + (N-1) p_k^2 - N p_k p_{kl} = p_k(1 - p_k),
\]

\[
E(\bar{p}_k - p_k) = p_k + (N-1) p_k q_{l} - N p_k q_{l} = p_k(1 - q_{l}),
\]

\[
E(\bar{p}_k - p_k)(\bar{q}_l - q_l) = p_{kl} + (N-1) p_k q_{l} - N p_k q_{l} = p_{kl}(1 - q_{l}),
\]

\[
E(\bar{p}_k - p_k)(\bar{q}_l - q_l) = p_{kl} + (N-1) p_k q_{l} - N p_k q_{l} = p_{kl}(1 - q_{l}).
\]
while the off-diagonal elements are ($k \neq r, l \neq s$)

\[ N \mathbb{E}(\overline{p}_{kl} - p_{kl})(\overline{p}_{ls} - p_{ls}) = (N - 1)p_{kl}p_{ks} - Np_{kl}p_{ks} = -p_{kl}p_{ks}, \]

\[ N \mathbb{E}(\overline{p}_{kl} - p_{kl})(\overline{p}_{rl} - p_{rl}) = (N - 1)p_{kl}p_{rl} - Np_{kl}p_{rl} = -p_{kl}p_{rl}, \]

\[ N \mathbb{E}(\overline{p}_{kl} - p_{kl})(\overline{p}_{rs} - p_{rs}) = (N - 1)p_{kl}p_{rs} - Np_{kl}p_{rs} = -p_{kl}p_{rs}, \]

\[ N \mathbb{E}(\overline{p}_{kl} - p_{kl})(\overline{q}_{s} - q_{s}) = (N - 1)p_{kl}q_{s} - Np_{kl}q_{s} = -p_{kl}q_{s}, \]

\[ N \mathbb{E}(\overline{q}_{l} - q_{l})(\overline{q}_{s} - q_{s}) = (N - 1)q_{l}q_{s} - Nq_{l}q_{s} = -q_{l}q_{s}. \]

Using the expectations computed in steps A1 - A4 we can get the following expressions for the elements of the variance matrix $V^n_T = (v^n_{ij})$, where $v^n_{ij} = \mathbb{E}(t_{ij} - \mathbb{E}t_{ij})(t_{ks} - \mathbb{E}t_{ks})$.

**Step B1 - Diagonal Elements ($k = r, l = s$):**

\[ v^n_{kk} = \left[ p_{kl}(1 - p_{kl}) - 2p_{kl}q_{kl}(1 - q_{kl}) - 2q_{kl}p_{kl}(1 - p_{kl}) + 2q_{kl}p_{kl}(1 - q_{kl}) + p_{kl}^2 q_{kl}(1 - q_{kl}) \right] \mathbb{E}(t_{kl} - \mathbb{E}t_{kl})(t_{kl} - \mathbb{E}t_{kl})^{-1} + o(1). \]

**Step B2 - Off-Diagonal Elements with $k = r, l \neq s$:**

\[ v^n_{kl} = \left[ -p_{kl}p_{ks} + p_{kl}p_{kl}q_{s} - q_{l}p_{kl}(1 - p_{kl}) - q_{l}p_{kl}(1 - p_{kl}) + p_{kl}q_{s}p_{kl} - p_{kl}q_{s}p_{kl} - p_{kl}^2 q_{kl}q_{s} + q_{l}q_{s}p_{kl}(1 - p_{kl}) \right] \mathbb{E}(t_{kl} - \mathbb{E}t_{kl})(t_{kl} - \mathbb{E}t_{kl})^{-1} + o(1). \]

**Step B3 - Off-Diagonal Elements with $k \neq r, l = s$:**

\[ v^n_{kl} = \left[ -p_{rl}p_{rl} + p_{rl}p_{kl}q_{l} + p_{rl}p_{kl}q_{l} - p_{rl}p_{rl}(1 - q_{l}) - p_{rl}p_{rl}(1 - q_{l}) - p_{rl}q_{l}(p_{rl} - q_{l}) \right] \mathbb{E}(t_{kl} - \mathbb{E}t_{kl})(t_{kl} - \mathbb{E}t_{kl})^{-1} + o(1). \]

**Step B4 - Off-Diagonal Elements with $k \neq r, l \neq s$:**

\[ v^n_{kl} = \left[ -p_{kl}p_{rs} + p_{kl}p_{rs}q_{l} + p_{kl}p_{rs}q_{l} + p_{kl}p_{rl}q_{s} + p_{kl}q_{l}(p_{ks} - q_{s}) \right] \mathbb{E}(t_{kl} - \mathbb{E}t_{kl})(t_{kl} - \mathbb{E}t_{kl})^{-1} + o(1). \]
To get the elements of the matrix $V_T = (v_{ij}) = \lim_{n \to \infty} V^n_T$ we take the limit with $n \to \infty$ of the previously computed expressions for $v^n_{ij}$, which makes the $o(1)$ terms vanish, which concludes the proof of Lemma 5.16.

It is clear that the calculations we performed to prove Lemma 5.16 could have easily been done using the counts and probabilities that correspond to any of the vectors $T^o_n$, $T^c_n$, $T^o_n$, $T^c_n$, or any other such vector, which means that an explicit proof of Lemma 5.17 is not necessary.

### 5.4.2 Variance of the pooled-sample statistic $R^p_n$

In Section 3.2.1 we discussed the statistic $R^p_n$ defined by Lewinger et al. (2013) as $R^p_n = (r - e_0)^2 / s_0^2$, where $r = n^{-1/2} \sum_{k,l=0}^2 z(u_{kl}, v_1) n_{k,l}$, $e_0 = \bar{E}_0 r$ and $s_0^2 = \text{var} r$. It is straightforward to see that $e_0 = n^{-1/2} \sum_{k,l=0}^2 z(u_{kl}, v_1) n_{k,l} / n_1$ is a suitable unbiased estimator of the expectation of $r$. Assuming $z(x, y) = xy$, Lewinger et al. (2013) suggest to standardize $r - e_0$ by $s^2 = n^{-1} \sum_{k=0}^2 u_k^2 n_k. \sum_{l=0}^2 v_l^2 n_l$. However, such value grossly overestimates the actual variance of $r - e_0$ by ignoring the fact that $e_0$ is an estimator, which means that it has non-zero variance. Thus, their conclusion about the asymptotic distribution of $R^p_n$ is not valid.

In order to obtain an expression for $\text{var}(r - e_0)$, we recall that the numerator term can be written as $r - e_0 = n \sum_{k,l=0}^2 z(u_{kl}, v_1)(\pi_{k,l} - \bar{\pi}_k \bar{\pi}_l)$. Using Lemmas 5.16 and 5.17 we calculate the variance of the vector $(\bar{\pi}_{k,l} - \pi_{k,l})$ by plugging probabilities $\pi_{k,l}$, $\pi_k$, and $\pi_l$ into (5.27) − (5.30) and simultaneously dropping the denominators in those expressions. Denoting the resulting variance matrix by $V_R$ and defining $v = (z(u_{kl}, v_1))_{k,l}$, it immediately follows that $\text{var}(r - e_0)$ is equal to $v^T V_R v$. Replacing the unknown probabilities in $v^T V_R v$ by their ML estimators then yields $s_0^2$. Assuming that $\pi_{k,l} > 0$ and that $v$ is neither a vector with coordinates all equal to 1, nor all equal to 0, it follows from central limit theorem that $(r - e_0)/s_0$ is asymptotically standard normally distributed distributed, which makes $R^p_n$ asymptotically chisquare distributed with one degree of freedom. These results also apply to $R^c_n$, $R^{o,c}_n$, $R^{o,s}_n$ and $R^{c,o,s}_n$, while the variances of the corresponding pre-test generating vectors have the same the form of $\text{diag}(v) V_R \text{diag}(v)$, except instead of $\pi_{k,l}$, $\pi_k$, and $\pi_l$ the matrix $V_R$ is calculated using the probabilities corresponding probabilities to the given pre-test generating vector.

### 5.4.3 Estimating asymptotic representation matrix

The results of Chapter 3 and this appendix make extended use of the asymptotic representation result given by Theorem 3.1. Many of our theoretical arguments utilized the asymptotic representation matrix $A = \bar{T}^{1/2}_p (1 - \bar{\Pi}_n) I_{p-1}^{1/2}$ formulated in Theorem 3.1. In order to put the derived results to practical use, we need to be able to approximate the matrix $A$. We conclude this chapter by deriving an estimator for $A$.

---

6Our definition of $R^p_n$ may seem different than that of Lewinger et al. (2013), because we added a standardization by $n^{-1/2}$ inside $r$ and $e_0$, but the effect of this is cancelled out by also adding $n^{-1}$ into the variance term $s^2$. 

---

7We have chosen to use this term because it allows us to express the asymptotic distribution of $R^p_n$ in terms of normal distribution and chi-square distribution. This makes the calculations easier and more intuitive for practitioners. 

---

8Our definition of $R^p_n$ is based on the fact that Lewinger et al. (2013) suggested a standardization by $n^{-1/2}$ inside $r$ and $e_0$, but the effect of this is cancelled out by also adding $n^{-1}$ into the variance term $s^2$. 

---
Recall that $\mathbf{I} - \Pi_{I\mathcal{H}}$ is a projection onto the orthogonal complement space of $I_{\beta}^{1/2}\mathcal{H}_0$. It is well known that for a space generated by a matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$, which we denote by $\mathcal{S}(\mathbf{X})$, the projection matrix $\mathbf{H}$ of a vector $\mathbf{y} \in \mathbb{R}^m$ onto $\mathcal{S}(\mathbf{X})$ can be written using the matrix $\mathbf{X}$ as $\mathbf{H} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$. To find the projection matrix onto the space $I_{\beta}^{1/2}\mathcal{H}_0$ we have to find a base of the limiting null hypothesis local parameter space $\mathcal{H}_0 = \{((\beta_0, \beta_1, \beta_2, \beta_3)' \in \mathbb{R}^3 \times \{0\})$. It is easy to see that $\mathcal{H}_0$ is the column space of the matrix $\mathbf{J}_4 = \text{diag}((1, 1, 1, 0))$. Thus, the space $I_{\beta}^{1/2}\mathcal{H}_0$ is generated by the matrix $\mathbf{X} = I_{\beta}^{1/2}\mathbf{J}_4$. Since both $I_{\beta}$ and $\mathbf{J}_4$ are symmetric, plugging $\mathbf{X}$ into the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ yields

$$
\Pi_{I\mathcal{H}} = I_{\beta}^{1/2}\mathbf{J}_4(I_{\beta} I_{\beta} \mathbf{J}_4)^{-1} I_{\beta}^{1/2} = I_{\beta}^{1/2}(I_{\beta} I_{\beta} \mathbf{J}_4)^{-1} I_{\beta}^{1/2}.
$$

(5.34)

The second equality above follows from the fact that a (pseudo)inverse of a block-diagonal matrix is obtained by (pseudo)inverting the individual blocks. Consequently, the multiplication of $(I_{\beta} I_{\beta} \mathbf{J}_4)^{-1}$ by $\mathbf{J}_4$ from left and right does not modify $(I_{\beta} I_{\beta} \mathbf{J}_4)^{-1}$. Assuming we have an estimator $\hat{I}_{\beta}$ for $I_{\beta}$, we use (5.34) to estimate $\Pi_{I\mathcal{H}}$ by $\hat{\Pi}_{I\mathcal{H}} = \hat{I}_{\beta}^{1/2}(I_{\beta} \hat{I}_{\beta} \mathbf{J}_4)^{-1} \hat{I}_{\beta}^{-1/2}$. This in turn suggests to estimate $\mathbf{A}$ using $\hat{\mathbf{A}} = \hat{I}_{\beta}^{1/2}(\mathbf{I} - \hat{\Pi}_{I\mathcal{H}}) \hat{I}_{\beta}^{-1/2}$, where $\hat{I}_{\beta}^{-1/2}$ is a square-root matrix of a (pseudo)inverse matrix of $\hat{I}_{\beta}$. We note that estimating $\hat{I}_{\beta}$ was already discussed in Section 3.1.1.
PART II

ROBIN HOOD: TWO-STAGE DESIGN AND ANALYSIS
Robin Hood: A method for two-stage experimental design and analysis

In Part II of the thesis, i.e. this chapter, we present a novel variant of a cost-efficient two-stage experimental design and analysis procedure designed to be used for a large-scale simultaneous inference problems in a cost-restricted setting. The design builds up on existing methods published in the literature and improves on a number of these methods especially in application areas with non-homogeneous sparse effects. This makes the method particularly useful in many areas of life sciences such as genomics. The contents of Part II have been submitted for publication in Pecanka and Goeman (2016).

6.1 Introduction

The traditional approach to a large scale genetic study is to collect all of the data beforehand and subsequently analyze it using a single-stage procedure which tests the hypotheses of interest for all available individuals. While having advantages in terms of reusability of the collected data, such approach also results in an unnecessary loss of power to detect false hypotheses especially in a cost-restricted setting. In an effort to mitigate such problems over the past decade and a half considerable attention has been paid to multi-stage analysis designs (Satagopan et al. (2002), Satagopan and Elston (2003), Satagopan et al. (2004), Zehetmayer et al. (2005), Wen et al. (2006), Goll and Bauer (2007), Zehetmayer et al. (2008), Pahl et al. (2009), Grünwald and Hössjer (2012), Zehetmayer and Posch (2012)). A common idea of many of these approaches is that the available budget can be more efficiently utilized if data is collected and analyzed in stages. In each stage the researcher uses only portion of the overall budget to gather additional data for as many subjects (individuals) as possible (affordable), tests the hypotheses of interest and advances to the next stage for additional data gathering
those hypotheses that have \( p \)-values below a suitable threshold selected for that stage in a way that the overall type I error of the procedure is controlled. After a pre-selected number of stages the hypotheses with significant \( p \)-value in every stage are declared as rejected by the procedure.

It has been shown that such multi-stage procedures can be much more powerful than single-stage procedures, where by power we understand the probability that a false hypothesis is identified among many by the complete procedure in the context of testing multiple hypotheses (Satagopan et al. (2002)). The improved performance by the multi-stage design is primarily driven by the increased efficiency of resource allocation among the tested hypotheses, which allows for an overall much larger sample size to be gathered for analysis compared to the single-stage design. This is a crucial aspect of the multi-stage design. In order to realize that power performance potential, however, it is necessary to optimize several input parameters the methods rely on. The input parameters are generally of three types. First choice required of the user is to decide on the number of stages. In many applications it can be impractical to use a setup with a high number of stages. Pahl et al. (2009) investigated the optimality of different stage counts and showed that using no more than four stages is sufficient in practice and that two-stages typically suffice for near optimal performance. The second type of input concerns the optimal budget allocation between the stages of the procedure. In a two-stage design this amounts choosing the portion of the budget used in the initial stage. Finally, the last input decision concerns the levels of significance used in each stage. Below we address all of these aspects of the multi-stage design.

Historically, multi-stage procedures have been studied especially in the context of GWAS around a decade ago. Since then, however, reusability of data is now given higher priority than efficiency of analysis design with respect to a specific phenotype, which caused the multi-stage designs to fall into disuse. We argue, however, that cost-effective multi-stage designs have a much wider applicability than just GWAS, and they should be considered whenever measuring a limited number of features is potentially cheaper than measuring all features. This is true for many types of omics data, for instance in classical transcriptomics, where PCR measurements are cheaper than full microarray or RNA-seq experiments, or in methylation, where pyrosequencing of a limited number of CpGs can be cheaper than full genome methylation measurements. In this chapter we consider a two-stage procedure in a general setting of a large omics study which aims to find a link between individuals’ features and a response variable of interest.

Two-stage designs should be optimized for specific application settings in terms of the optimality of input parameters. Many of the optimization issues have been addressed to some degree in the GWAS-focused literature, where the focus has generally been toward a setting with homogeneous effects, i.e. effects of (almost) equal sizes. In the general multi-stage design described above it is required that a hypothesis be tested in every single stage before it can be declared as rejected by the procedure regardless of how strong the available interim evidence against the hypothesis is. Such design is generally fine when the non-zero effects...
are homogeneous. However, if some of the non-zero effects are very large, which makes it likely for the interim evidence against the corresponding hypotheses to be very strong, it is undesirable to direct precious resources towards gathering of additional data for such effects in the later stages of the procedure. Having the ability to make an early rejection of such hypotheses would mean that the available resources can be redirected towards gathering of the more valuable data for other hypotheses and thus improving the ability to detect moderately sized effects. In this chapter we present a novel implementation of precisely such feature within a two-stage procedure. Inspired by the ability of our method to shift available budget away from the very large ("rich") effects to the moderately sized or even small effects ("poor"), we refer to the method as Robin Hood.

An important factor for the use of any multi-stage procedure are the data gathering costs, which must be included in the consideration if a fair comparison with a single-stage procedure is desired. In addition to an overall budget, the performance of a multi-stage method is heavily driven by the data gathering unit measurement costs in each stage, which are defined as feature measurement costs per-feature per-individual. With current data collection technologies in many fields (e.g. genetics) it is typical that the unit costs increase substantially after the initial stage often as much as several thousand-fold. Generally, if unit costs increase between stages it creates an obstacle for a multi-stage procedure to perform well. However, in settings where a large portion of the considered hypotheses are true (or "nearly true"), even with a large unit cost increase between stages, a multi-stage procedure can convincingly outperform the traditional single-stage procedure.

In this chapter we focus on the Gaussian (normal) location model with (asymptotically) normally distributed statistics. While this is likely the most relevant application setup, the presented methodology is far from limited to this setup. On the contrary, the two-stage methodology in general and the Robin Hood method in particular can be employed in many other contexts. In fact, unlike many of the methods available in literature, the Robin Hood method is intentionally designed to operate on the p-values, which can in principle be based on any meaningful statistic in any model of interest as long as the p-values are uniformly distributed under the corresponding null hypotheses in that model. Below we employ specific assumptions, such as the Gaussian model and a few other technical assumptions relating to measurement and sampling cost functions, in order have a framework for the optimization of input parameters and a numerical comparison of the Robin Hood method with existing methods. However, this could easily be done under various alternative sets of assumptions.

The chapter is structured as follows. In Section 6.2 we provide a formal definition of the model and the considered single- and multi-stage procedures including the Robin Hood method. In Section 6.3 we focus on various theoretical properties of the two-stage methods relevant to their usage including the control of type I error and the power performance. In Section 6.4 we perform a numerical study in order to compare the performances of the methods under both homogeneous and heterogeneous effect scenarios and illustrate the power superiority of the Robin Hood method. Finally, in Section 6.5 we present a sensitivity study.
of the power performance of the methods with respect to violations of the assumptions made during optimization of input parameters.

### 6.2 Method setup

The technical setup we consider here is the following. Suppose that a given number of $N$ hypotheses are to be investigated using statistical tests with budget $B$ available for data collection. We assume that the budget affords us for each of the $N$ hypotheses to collect an $n$-sized random sample of (iid) variables $Z_{i,1}, \ldots, Z_{i,n}$, where $i = 1, \ldots, N$. We further assume that these random variables follow the unit-variance\(^\dagger\) normal distribution $N(\theta_i, 1)$ and that the null hypotheses of interest are $H_{0i} : \theta_i = 0$ with the corresponding alternative $H_{1i} : \theta_i > 0$ (one-sided). We refer to $\theta_i$ as effects and as usual our aim is to identify as many of the non-zero effects (i.e. false null hypotheses) as possible, while minimizing the number of falsely identified zero effects (i.e. true null hypotheses). While the performance in terms of the former relates to power, the latter we want to control in terms of the family-wise error rate (FWER) (i.e. the probability of making a non-zero number of rejections of true hypotheses) by a pre-specified value $\gamma \in (0, 1)$ such as $\gamma = 0.05$.

In order to avoid confusion, we stress that throughout this chapter we denote by $\gamma$ the desired type I error over the whole family of $N$ tests (e.g. $\gamma = 0.05$), while we use $\alpha$ to denote as the level of significance used for a test of single hypothesis. The value of $\alpha$ is assumed to be such that the overall type I error is controlled by $\alpha$, where the simplest way to account for multiplicity of testing $N$ hypotheses is to use the Bonferroni correction (i.e. level $\alpha_B = \gamma/N$) or the Holm-Bonferroni step-down correction (Goeman and Solari (2014)). Both of these methods are simple and they both guarantee strong control of FWER for any possible dependence structure among the test statistics. Alternatively, for independent test statistics strong control of FWER is guaranteed also by the Sidak’s corrected significance level $\alpha_S = \frac{1}{N} - \gamma / N$, or the Hochberg’s step-up procedure, which allows also some forms of positive dependence (Goeman and Solari (2014)). In the interest of providing explicit and simple formulas we focus on the Bonferroni correction.

In many contexts it is of interest to focus on the situation when only a relatively small fraction of the $N$ null hypotheses is false, meaning that only a relatively small part of the tests should in fact result in rejection of the corresponding null hypothesis. We will refer to such situation as the rare alternative scenario (RAS). While RAS drove our investigation and is used in the numerical investigation of Section 6.4, we do not use it explicitly in the design of the two-stage testing procedure nor is it in principle necessary. Consequently, the applicability of the methods described here is not necessarily limited to RAS.

\(^\dagger\)We make the assumption of unit variance only for notational convenience. Formulating the methods for the case of normal distribution with estimated variance is still possible and relatively straightforward though slightly more cumbersome. Generally speaking in such case the results presented here become asymptotic (with $n_1, n_2 \to \infty$).
Regarding the inclusion of budget restrictions and costs in the framework of this chapter, we generally only focus on the feature measurement costs while not including the costs of obtaining the values of the continuous response variable, or the sampling costs, or the costs of the actual statistical analysis. In principle, all of these could be included in the framework in a straightforward way, but for the sake of clarity of presentation we choose not to do that here. In areas of application such as genomics this is also justified by the fact that the feature measurement costs (genetic markers, PCR measurements, CpGs, etc.) generally account for the biggest portion of the budget by far and therefore are the main determinant of the relative performances of the methods outlined in this chapter.

Finally, for the sake of clarity of presentation in the theoretical part we focus on the one-sided alternative hypotheses, while re-formulating the results for the two-sided case would be completely straightforward. On the other hand, since the two-sided case is probably more relevant in practice, the numerical results presented in Section 6.4 focus on the case of two-sided alternatives.

### 6.2.1 Classical single-stage approach (M1)

Classically, the individual null hypotheses $H_{0i}$ can be tested using the exact (single-stage) $z$-test statistic $T_{i,n} = \frac{n-1}{2} \sum_{j=1}^{n} (Z_{ij} - \theta_i)$, which is distributed according to $N(\sqrt{n}\theta_i, 1)$. The hypothesis $H_{0i}$ is rejected in favor of $H_{1i}$ for large values of $T_{i,n}$, which are determined as the appropriate critical values of $N(0,1)$ at level $\alpha$, where $\alpha$ is a value sufficient for the FWER among the $N$ hypotheses to be controlled by the user selected value of $\gamma$. We refer to the $N$ parallel single-stage $z$-tests as M1. In M1 we use the entire budget $B$ to collect all of the data. We assume that the necessary budget can be calculated as $B = n N c_1$, where $c_1$ is the relevant unit cost function. In general, $c_1$ is a function of both $n$ and $N$. In areas where data for different subjects are collected separately $c_1$ generally does not depend on $n$. We refer to this as proportionality of the unit costs and employ it as a simplifying assumption throughout this chapter.

### 6.2.2 Robin Hood multi-stage method (RHM)

At the start of the analysis we allocate a partial budget $B_1 = \delta B$, where $\delta \in (0,1)$, and collect as much data as $B_1$ allows. Given the proportionality of unit costs, the affordable sample size with budget $B_1$ is $n_1 = \lfloor \delta n \rfloor$, hence we obtain a random sample $Z_{i,1}, \ldots, Z_{i,n_1}$ for each of the $N$ hypotheses. Using this data we calculate the test statistics $T_{i,n_1} = \frac{n_1-1/2}{\sum_{j=1}^{n_1} (Z_{ij} - \theta_i)}$ for each of the $N$ hypotheses $H_{01}, \ldots, H_{0N}$ and determine the corresponding $p$-values $p_1, \ldots, p_N$.

In the initial step of the procedure called stage 0 (S0) we identify the $N_1$ $p$-values that are below $\alpha_0$ for a user-chosen $\alpha_0 \in [0,\alpha)$. The corresponding $N_1$ hypotheses are rejected by the procedure without any further testing. In a subsequent stage 1 (S1) we identify all $p$-values in $(\alpha_0, \alpha_1)$ for a fixed $\alpha_1 \in (\alpha_0,1)$ and denote their count by $N_2$. If $N_2 = 0$ we stop.
and reject no more hypotheses. Otherwise, without loss of generality (wlog), we assume that it is $p_1,\ldots,p_{N_2}$ that take values in $(\alpha_0,\alpha_1)$. With those $N_2$ $p$-values we proceed to the next stage, which refer to as stage 2 (S2), we obtain additional data using the remaining budget $B_2 = (1-\delta)B$. Based on the independent data we calculate $q_1,\ldots,q_{N_2}$ using a chosen test statistic, which can but does not have to be the same statistic that was used in S1. We then combine $q_1,\ldots,q_{N_2}$ with $p_1,\ldots,p_{N_2}$ via a Fisher-type combination method to form the test statistic

$$F_i = -2u_i \log \overline{p} - 2v_i \log q_i, \quad i = 1,\ldots,N_2,$$

where $\overline{p} = (p_i - \alpha_0)/(\alpha_1 - \alpha_0)$ and $u_i, v_i$ are sample sized based weights such that $u_i + v_i = 1$ for all $i$. A good choice for the weights is for instance $u_i = n_1/(n_1 + n_2)$ and $v_i = n_2/(n_1 + n_2)$. Using the null hypothesis distribution of $F_i$ (specified below), we obtain $p$-values $r_1,\ldots,r_{N_2}$ and compare them with the level $\alpha_2 = (\alpha - \alpha_0)/(\alpha_1 - \alpha_0)$. Finally, all hypotheses corresponding to $p$-values $r_1,\ldots,r_{N_2}$ below $\alpha_2$ are rejected by the procedure in addition to those hypotheses rejected in S0.

### 6.2.3 Vanilla two-stage method (VM)

The Robin Hood method is an extension of the basic two-stage scheme proposed in literature (e.g., Satagopan and Elston (2003), Pahl et al. (2009)). For short we refer to the basic scheme as vanilla two-stage methods (VM). Essentially, the main difference between RHM and VM is that VM does not include S0. In the context of RHM, a vanilla version of our method is obtained when $\alpha_0 = 0$, which leads to using the level $\alpha_2 = \alpha/\alpha_1$ at S2. Consequently, the probability of the type I error for the $i$-th test simplifies to $F_i = P(p_i < \alpha_1, r_i < \alpha_2) = P(p_i < \alpha_1)P(r_i < \alpha_2 | p_i < \alpha_1) = \alpha_1 \alpha_2 = \alpha$.

In RHM in order for the S2 level to be meaningful, i.e. between 0 and 1, we required $\alpha_0 < \alpha$. Naturally, there is a price to be paid for the the additional S0 comparisons in RHM, which comes in the form of stricter S2 tests compared to VM. However, the price in terms of a more stringent $\alpha_2$ vanishes rather quickly as $\alpha_0$ decreases away from $\alpha$ as illustrated by Figure 6.1. Moreover, if some of the values of $\theta_i$ are large, RHM rejects the corresponding hypotheses already in S0, which results in larger budget (and sample size) available for the tests at S2. If the overall number of non-zero effects is small relative to $N$, while a non-negligible portion of the non-zero effects are large, RHM leads to a substantial power improvement over the original two-stage method.

In addition to the test statistics, RHM can be contrasted with the methods available in the literature also in terms of how the data collected in different stages are treated in the analysis. Essentially, two approaches have been suggested in the literature, where Skol et al. (2006) provides a comparison. The simpler option is to use a replication-based approach and keep the data employed in each stage disjoint, thus making the statistics in between the two-stages independent. We refer to this as the disjoint vanilla two-stage method (DVM). Alternatively,
one can take all of the data available in S2 and combine it into a single sample and base the S2 test statistic on it. We refer to this as the pooled vanilla two-stage method (PVM). Such pooling of the data requires the user to work out the dependence structure between the statistics used in the two stages in order to determine the highest possible level of significance in S2 that guarantees type I error control. Consequently, determining $\alpha_2$ under dependence is computationally non-trivial and requires two-dimensional integration of the bivariate normal density. Moreover, working out the dependence can be even more complicated when dealing with non-normally distributed statistics. We note that the designs with tests at each stage performed using either disjoint or pooled data are also known as pilot design and integrated design, respectively ([154]). While it can be argued that using the S1 data only to select hypotheses for S2 is a generally less powerful approach (Skol et al. (2006), Wen et al. (2006)), the conditional independence of the two stages under the null hypothesis allows the usage of $\alpha_2 = \alpha/\alpha_1$. This is the same level used by the RHM with $\alpha_0 = 0$. Since RHM does in fact reuse data at S2, it can be argued that RHM combines the advantages of efficient usage of data of PVM and the simplicity of determining the S2 level of significance in DVM.

### 6.3 Theoretical properties of the methods

Next we address various properties of the two-stage methods relevant to their usage such as the type I error control, the role of input parameters, the role of unit costs and the potential for budget savings as well as their relative power performances and their optimization.

#### 6.3.1 Type I error control by RHM

A crucial property utilized in the definition of $F_i$ is that under $H_{0i}$ the conditional distribution of $p_i$ conditioned on $p_i \in (\alpha_0, \alpha_1)$ is uniform on $(\alpha_0, \alpha_1)$, which means that $\overline{p}_i = (p_i - \alpha_0)/(\alpha_1 - \alpha_0)$ is conditionally standard uniform. Consequently, for $u_i = v_i = 1$ the
null hypothesis conditional distribution of $F_i$ is the chisquare-four distribution by Fisher’s classical result, while for general weights $u_i \neq v_i$ the null hypothesis distribution is given in Theorem 2.4 in Box (1954).

Moreover, given fixed values of $\alpha_0, \alpha_1, \alpha_2$ the probability of the type I error corresponding to the $i$-th hypothesis $F_i$ equals

$$F_i = P(p_i < \alpha_0) + P(\alpha_0 \leq p_i < \alpha_1, r_i < \alpha_2)$$

$$= \alpha_0 + P(\alpha_0 \leq p_i < \alpha_1)P(r_i < \alpha_2 | \alpha_0 \leq p_i < \alpha_1)$$

$$= \alpha_0 + (\alpha_1 - \alpha_0)\alpha_2 = \alpha_0 + (\alpha_1 - \alpha_0)(\alpha - \alpha_0)/(\alpha_1 - \alpha_0) = \alpha,$$

where we used the conditional uniformities of the $p$-values $p_i$ and $r_i$ (derived from the distribution of $F_i$) and their conditional independence under the null hypothesis. Since $\alpha$ is assumed to be such that the FWER is (strongly) controlled, for instance the Bonferroni level $\alpha = \gamma/N$, it follows that the RHM properly (strongly) controls FWER with such $\alpha$.

### 6.3.2 Choice of levels

The motivation behind the two-stage approach is to improve the overall power by better utilizing the total budget by focusing the resources towards the more promising hypotheses while ignoring the least promising ones. The power of the two-stage methods can be maximized by optimizing the effective levels of significance in each stage but it must be done in a way that maintains control of type I error, namely FWER in our case. Conveniely, the (conditional) independence and standard uniformity of the $S_1$ and $S_2$ $p$-values under the null hypothesis in RHM (and DVM) lead to the maximum permissible $S_2$ level $\alpha_2$ being simple function of $\alpha_1$ and $\alpha$, namely $\alpha_2 = (\alpha - \alpha_0)/(\alpha_1 - \alpha_0)$ for RHM (and $\alpha_2 = \alpha/\alpha_1$ for DVM). In other words, it is possible to choose the $S_1$ level $\alpha_1$ completely arbitrarily from $(\alpha_0, 1)$ without affecting strong control of FWER. However, for good power performance it is essential to choose the two levels carefully. By balancing the allotment of type I error between the two stages we can achieve a substantial improvement of power relative to the single-stage method $M_1$.

We note that (6.2) implies that it is possible to use different (a priori fixed) levels in stages 0 and 1 for each of the $N$ tests, provided that the $S_2$ levels are chosen accordingly. This can be useful in certain applications where prior knowledge is available and can be injected into the procedure via the choice of $S_0$ and $S_1$ levels, thus further optimizing its performance.

Besides $\alpha_1$, RHM requires the user to also specify $\alpha_0$. Crucially, RHM controls the type I error for any value of $\alpha_0$ below $\alpha$, although obviously not all choices lead to the same power performance. Addressing the optimality of a choice of $\alpha_0$ theoretically is difficult in general since the answer strongly depends on the true distribution of effects which is unknown. A viable strategy is to target a reasonable balance between the power benefits and the cost in terms of decreased $S_2$ level that the presence of $S_0$ in RHM brings. Figure 6.1 shows that trade-off and one can base the choice of $\alpha_0$ on it. Setting $\alpha_0$ to say $\alpha/10$ or $\alpha/100$ both correspond to only a modest decrease of $S_2$ level. In that sense $S_0$ comes almost "for free".
6.3.3 Special property of Fisher’s method

Naturally, a Fisher-type $p$-value combination method is not the only possibility for defining a valid S2 test in a two-stage procedure. Even when the goal is to formulate a method using $p$-values there are other methods that can be relied on such as Stouffer’s $z$-score method (Stouffer et al. (1949)) or Chen’s Gamma combination method (Chen et al. (2014)). However, Fisher’s method has a useful "special" property which distinguishes it from the others, namely, it is the property that Fisher’s method has the "worst case" S2 outcome, namely a S2 $p$-value equal to 1. In such case we can use Fisher’s combination method to determine how small does the corresponding S1 $p$-value need to be so that when it is combined with 1 the resulting statistic $F_1$ of (6.1) is significant. In fact, the ability of Fisher’s method to assess the overall significance in multi-stage procedure without having to gather data for all stages was the motivation behind the introduction of S0 in RHM. In our formulation of the method we did not use this "special" property directly, but instead we opted to introduce a more general formulation based on the parameter $\alpha_0$ and we argued for setting $\alpha_0$ by considering the trade-off in terms of $\alpha_2$. While we do not go into details here, it would be also possible to use the "special" property to arrive at an alternative way for setting choose a value for $\alpha_0$.

6.3.4 Affordable sample sizes and unit costs

By the definition of the budget ratio $\delta$, the size of the sample underlying $p_1, \ldots, p_N$ is $n_1 = \delta B/(N c_1)$. With S2 unit costs denoted as $c_2$, the S2 affordable sample size $n_2$ satisfies $B_2 = n_2 N c_2$, where $B_2 = (1 - \delta)B$. Consequently, $n_2$ is a function of the random variable $N_2$, whose distribution depends on the values of $\theta_1, \ldots, \theta_N$, $\alpha_1$, $\delta$ and the dependence structure among the test statistics. Since our framework allows for dependence of unit cost $c_2$ on the sample size $n_2$, we denote it as $c_2(n_2)$. Typically, $c_2$ is a decreasing function in $n_2$ due to economies of scale, although the exact shape of $c_2(n_2)$ depends on the technology used to gather the data. Crucially, however, the two-stage methods can be vastly superior to the single-stage analysis even with variable unit cost function and a between-stage unit cost ratios in the order of thousands or even more, as shown both in the literature (Goll and Bauer (2007)) and in our numerical analysis in Section 6.4. Under the proportionality of S1 unit costs the relative performance of single-stage and two-stage methods depends only on the ratio of unit costs $c_2/c_1$ since changing the value of $c_1$ is equivalent to changing the monetary units in which the costs are measured. Therefore, in the rest of this chapter we fix $c_1 = 1$.

6.3.5 Lower expected costs

Besides the main advantage of the two-stage approach which is its potential for superior power performance, there is an additional benefit to such analysis scheme. Namely, the possibility of budget savings, which are especially desirable if the complete null hypothesis (no non-zero effects) is true. In a two-stage procedure the total budget $B_{12}(\theta) = B_1(\theta) + B_2(\theta)$
6.3 Theoretical properties of the methods

required by the two-stage analysis is at most $B$, while $E B_{12}(\theta)$ is strictly less than $B$ for any combination of $\theta$, $\delta$ and $\alpha_1$ that yields $P_0(\theta) = P_0(N_2 = 0) > 0$. For a given $\delta$ the expected budget is $E B_{12}(\theta) = \delta B_0(\theta) + B(1 - P_0(\theta)) = B(1 - P_0(\theta)(1 - \delta))$. Since a single-stage analysis always exhausts the entire budget $B$ regardless of what the values in $\theta$ are, the fact that a two-stage procedure allows budget savings is potentially quite beneficial.

6.3.6 Power functions

We already discussed how to achieve type I error control in RHM as well as in VM. Next we formulate several theoretical results concerning power functions within the two approaches. Regarding the vanilla methods, we formulate the power function explicitly only for DVM, while that of PVM for a pair of jointly normal statistics can be found in Satagopan and Elston (2003).

Power of M1

In the normal shift model $N(\theta, 1)$ the power function of the one-sided z-test at level $\alpha \in (0, 1)$ with sample size $n$ is $\Pi(\sqrt{n} \theta, a) = 1 - \Phi(\sqrt{n} \theta, 1 - a)$. Consequently, since each M1 test is performed at level $\alpha = \gamma/N$, the power of M1 to reject $H_{0i}$ is $\Pi(\sqrt{n} \theta, 1, \alpha)$.

Power of VM

In the vanilla methods where rejection occurs only for hypotheses with significant tests in all stages the power function can be written as $\Pi_{12}^{VM}(\theta_i, \alpha_1, \alpha_2, \delta) = P(A_i) P(B_i | A_i) = \Pi(\sqrt{n} \theta_i, \alpha_1) \Pi_{2}^{VM}(\theta_i, \alpha_1, \alpha_2, \delta)$, where $A_i$ and $B_i$ denote the rejection events of the $i$-th hypothesis at stages 1 and 2, respectively, and $\Pi_{2}^{VM}(\theta_i, \alpha_1, \alpha_2, \delta) = \Pi(\sqrt{n} \theta_i, \alpha_2)$ is the conditional power of the $i$-th test in S2 conditioned on rejection by the $i$-th test in S1. Since $n_2$ is random, $\Pi_{12}^{VM}$ is an expectation taken over the distribution of $n_2$ (or equivalently, over the distribution of $N_2$) conditioned on $N_2 \geq 1$. We denote the power functions of DVM by $\Pi_{12}^{DVM}$.

Power of RHM

When testing $H_{0i}$ in RHM the probability of rejection at S0 is $\Pi(\sqrt{n} \theta_i, a_0)$. Similarly, the probability of advancing to S2 is $\Pi_{10}(\sqrt{n} \theta_i, a_1, a_0)$, where we put $\Pi_{10}(x, a_1, a_0) = \Pi(x, a_1) - \Pi(x, a_0)$. In terms of the distribution of the $p$-value, this is equal to the integral over $(a_0, a_1)$ of the density $h_{\mu_j}$ given by (6.3) with $\mu_j = \sqrt{\pi} \theta_i$. Analogously to the S2 in DVM, the probability of rejection at S2 (alone) of RHM denoted as $\Pi_{2}^{RHM}(\theta_i, a_0, a_1, \alpha_2, \delta)$ is the expected conditional power over the distribution of $n_2$ or $N_2$, where the latter is a sum of (possibly dependent) Bernoulli distributed variables with success probabilities that depend on $\theta_1, \ldots, \theta_N$. Consequently, the overall power of RHM to reject $H_{0i}$ is

$$\Pi_{12}^{RHM}(\theta_i, a_0, a_1, \alpha_2, \delta) = \Pi(\sqrt{n} \theta_i, a_0) + \Pi_{10}(\sqrt{n} \theta_i, a_1, a_0) \Pi_{2}^{RHM}(\theta_i, a_0, a_1, \alpha_2, \delta).$$

We note that for $a_0 = 0$ the power functions of RHM and VM become identical.
In order to calculate $\Pi_{2}^{RHM}$ we need to start by looking at the distribution of the Fisher-combined $p$-values. Under the normal model $N(\theta,1)$, while denoting the standard normal density and distribution function as $\varphi$ and $\Phi$, a $p$-value $p$ based on a sample of size $n$ is distributed according to density

$$h_\mu(p) = \frac{\varphi(\Phi^{-1}(1-p) - \mu)}{\varphi(\Phi^{-1}(1-p))}, \quad p \in (0, 1),$$

(6.3)

where $\mu = \sqrt{n}\theta$. This follows directly from the transformation theorem for random variables with the transformation in question being $\tau(x) = 1 - \Phi(x)$. The inverse transformation is $\tau^{-1}(y) = \Phi^{-1}(1-y)$, which has derivative $t(y) = -1/\varphi(\Phi^{-1}(1-y))$, hence (6.3). Consequently, $X = -2\log p$ has density $g_{\mu,\nu}(x) = 0.5w^{-1}e^{-0.5xw^{-1}}h_\mu(e^{-0.5xw^{-1}})$ supported on $x \in (0, \infty)$. Consequently, for two independent $p$-values $p$ and $q$ from the model $N(\theta,1)$ it holds that $F = -2\log(p) - 2\log q$ has density $f_{\mu,n,v} = \mathcal{g}_{\mu,n} \ast \mathcal{g}_{\nu,v}$, where $\ast$ denotes convolution. Written explicitly, the density $f_{\mu,n,v}(x)$ supported on $(0, \infty)$ is equal to

$$f_{\mu,n,v}(x) = (\mathcal{g}_{\mu,n} \ast \mathcal{g}_{\nu,v})(x) = \int_{\mathbb{R}} g_{\mu,n}(x-y)g_{\nu,v}(y)dy = (4uv)^{-1/2}e^{-0.5xuv^{-1}}\int_{\mathbb{R}} e^{0.5y(u^{-1} - v^{-1})}h_\mu(e^{-0.5yuv^{-1}})dy.$$

(6.4)

In order to get $\Pi_{2}^{RHM}$ we need the conditional density of the $i$-th $p$-value $p_i$ conditioned on $p_i \in (\alpha_0, \alpha_1)$, which is $h_{\mu_i}^p(p) = c_i^{-1}h_{\mu_i}(p)$, $p \in (\alpha_0, \alpha_1)$ with $c_i = \int_{\alpha_0}^{\alpha_1} h_{\mu_i}(t) dt = \int_{\alpha_0}^{\alpha_1} q(t - \mu_i) dt$, where $\alpha_i = \Phi^{-1}(1 - \alpha_i)$. The latter integral is probably more convenient for numerical integration. Thus, the conditional density of the scaled $p$-value $\overline{p}_i$ is $h_{\mu_i}^p(p) = (\alpha_1 - \alpha_0) h_{\mu_i}^p(\alpha_0 + p(\alpha_1 - \alpha_0))$ for $p \in (0, 1)$. With a fixed sample size $n_2$, the conditional density (conditioned on $p_i \in (\alpha_0, \alpha_1)$) of statistic $F_i$ is $f_{\mu,n_2,n_1,v_1}$, which is the analogue of $f_{\mu,n_2,n_1,v_1}$ defined in (6.4) with $h_{\mu_i}$ replaced by $h_{\mu_i}^p$. Finally, taking the expectation of the conditional distribution of $F_i$ with respect to the distribution of $n_2$ (or $N_2$) yields $\Pi_{2}^{RHM}$.

### 6.3.7 Distribution of $N_2$ under independence

A common theme to the two-stage methods with random sample size at S2 is that the corresponding power functions are the expectations over the distribution of the random sample size $n_2$, which in our setting can be equivalently obtained by taking the expectation with respect to the conditional distribution of $N_2$ conditioned on $N_2 > 0$. Below we specify the distribution of $N_2$ under independence. It is crucial to keep in mind that none of the methods presented here in fact require independence in order to be valid. It is only the optimization of input parameters that is simplified under independence. And as we showed in Section 6.3.1 RHM controls the type I error for all values of $\alpha_1$ (and $\delta$) as soon $\alpha$ is properly adjusted for multiplicity of testing (e.g. $\alpha = \gamma/N$ for strong error control under arbitrary dependence). It is only the power performance that can be influenced by inappropriately assuming independence during optimization of input parameters. However, in Section 6.7 we show that such
effect is only relatively mild when independence is severely violated, while for weak and moderate (positive) dependence the effect is virtually negligible.

Let the values of $\delta \in (0, 1)$ and $\alpha_0, \alpha_1 \in (0, 1)$ be given with $\alpha_2 \in (0, 1)$ chosen accordingly. Denoting $R_i = I(p_i \in (\alpha_0, \alpha_1))$, we get $N_2 = \sum_{i=1}^{N} R_i$. In general the distribution of $N_2$ can be quite complicated due to the potential unknown dependence structure among $R_1, \ldots, R_N$. However, in the special case of independent test statistics $N_2$ is a sum of $N$ independent Bernoulli distributed variables with possibly different success probabilities. The resulting Poisson-binomial distribution is governed by probabilities $p_k(\theta) = \mathbb{P}_\theta(N_2 = k)$ for $k = 0, 1, \ldots, N$ and $\theta = (\theta_1, \ldots, \theta_N)'$. Recalling that we defined $\Pi_{10}(x, \alpha_1, \alpha_0) = \Pi(x, \alpha_1) - \Pi(x, \alpha_0)$, we can write $p_k(\theta)$ explicitly as

$$p_k(\theta) = \sum_{S \in \mathcal{S}_k} \prod_{i \in S} \Pi_{10}(\sqrt{n\theta_i}, \alpha_1, \alpha_0) \prod_{j \notin S}[1 - \Pi_{10}(\sqrt{n\theta_j}, \alpha_1, \alpha_0)],$$  
(6.5)

with $\theta = (\theta_1, \ldots, \theta_N)'$ and $\mathcal{S}_k$ denoting a set of all $k$-sized subsets of $\mathcal{K} = \{1, \ldots, N\}$.

### 6.3.8 Distribution of effects and optimization of input parameters

A somewhat complicating factor for the use of the two-stage approach is the fact that the optimization of $\delta$ and $\alpha_1$ requires certain assumptions regarding the distribution of effects. While these assumptions generally do not need to be exactly correct, the necessity to make them might be seen as unpleasant for the user. However, the possibility to make choices also means that the procedure can be tuned to maximize performance for specific scenarios, which can in fact be seen as an advantage. In the numerical analysis in Section 6.4 we compare the methods under the following two scenarios.

**m homogeneous effects (notation E|m|N)** The simplest scenario is one with $m$ equal (i.e. homogeneous) effects among $N$ hypotheses, which we denote as $E|m|N$. Under $E|m|N$ with independence the distribution of $N_2$ is relatively simple. Without loss of generality let us assume that the effects are $\theta_1 = \cdots = \theta_m > 0$ while $\theta_i = 0$ for $i > m$. Then, the distribution of $N_2$ is equal to the distribution of the sum of two binomial distributions with success probabilities $\pi_{10} = \Pi_{10}(\sqrt{n\theta_1}, \alpha_1, \alpha_0)$ and $\alpha_1$, that is

$$p_k(\theta_1) = \sum_{u=0}^{k} \binom{m}{u} \binom{N-m}{k-u} \pi_{10}^u (1 - \pi_{10})^{m-u} \alpha_1^{k-u} (1 - \alpha_1)^{N-m-k+u},$$

where we used the fact that $\binom{a}{b} = 0$ for any integer $a, b$ such that $a < b$ or $b < 0$. With $m$ homogenous effects the expectation of $N_2$ equals $\mathbb{E}N_2 = m\pi_{10} + \alpha_1(N - m)$.

**m heterogeneous effects (notation U|m|N)** The other scenario of interest is one where the non-zero effects have different sizes, i.e. the effects are heterogenous. We denote such scenario with $m$ heterogeneous effects among $N$ hypotheses as $U|m|N$. For many applications such scenario is in fact much more realistic than the homogeneous effect scenario above. Under $U|m|N$, especially if some of the effects are very large, the inclusion of $S_0$ in RHM can
result in substantially better power performance. Under independence an explicit formula for \( p_k(\theta) \) can be easily derived from (6.5), though with \( m \) possibly different effects the formula can be very lengthy and therefore we omit it here. Again, assuming that \( \theta_1, \ldots, \theta_m \) are the non-zero effect, the expectation of \( N_2 \) equals \( \mathbb{E}N_2 = \sum_{i=1}^m \prod_{j \neq i} (\sqrt{\delta n \theta_j}) + \alpha_1 (N - m) \).

### Optimization in practice

In many areas of application the user has an approximate idea of how many and how large non-zero effects to expect. In such case perhaps the simplest strategy is to target the \( \mathbb{E}|m|N \) scenario with a reasonably selected value for \( m \). Regarding the effect size, either a fixed target effect size can be assumed or the user can specify the desired power (e.g. 80%), which in turn yields the effect size that corresponds to the selected power under the \( \mathbb{E}|m|N \) scenario. With such assumptions in place, the input parameters \( \alpha_1 \) and \( \delta \) can be easily optimized.

In Section 6.7 we focus on the sensitivity of the such optimization towards the miss-specification of the value of \( m \). We show that choosing a value for \( m \) that is larger than the true number of non-zero effects generally yields only a negligible decrease in performance. Consequently, in order to achieve a near-optimal power performance in applications it is generally enough for the user to give a reasonable upper bound on the number of non-zero effects without having to specify it exactly.

### 6.3.9 Approximate power functions

As discussed above, for the evaluation of the power functions of the two-stage procedures it is essential to specify the (conditional) distribution of \( N_2 \). Given that this distribution is unknown and can be quite complicated in general, a natural idea is to approximate the exact power function by the power function corresponding to the expected number of tests in \( S_2 \) (Goll and Bauer (2007)). An advantage of such an approximation is that unlike the exact distribution of \( N_2 \), the value of \( \mathbb{E}N_2 \) does not depend on the dependence structure among the \( N \) tests. Instead, it only depends on the values of \( \theta_1, \ldots, \theta_N, \alpha_0, \alpha_1, \delta \) which fully determine the success probabilities of the Bernoulli variables that yield \( N_2 \). The affordable sample size \( n_2 \) corresponding to \( \mathbb{E}N_2 \) satisfies \( B_2 = n_2 \mathbb{E}N_2 c_2 \), where \( B_2 = (1 - \delta)B \). Denoting as \( \bar{n}_2 \) the solution of that equation, and assuming that the unit cost function \( c_2 \) is constant in \( n_2 \), we get \( \bar{n}_2 = (1 - \delta)B/(\mathbb{E}N_2 c_2) \). In the more general case without the assumption of sample size independent \( c_2 \) the sample size equation can be solved iteratively.

It should be noted that the word *advantage* in previous paragraph should perhaps be put into quotation marks. While it is true that using the approximation simplifies things technically and the resulting approximate function is indeed the same with or without independence, the *quality* of the approximation is still very much influenced by the dependence structure among the tests. In fact, a slightly more suitable approximation of the power function would be obtained by considering \( \mathbb{E}(N_2|N_2 > 0) \) instead of \( \mathbb{E}N_2 \), since \( \mathbb{E}(N_2|N_2 > 0) \) is a better proxy for the actual conditional distribution of \( N_2 \) especially under strong dependence. How-
ever, since the calculation of $E(N_2|N_2 > 0)$ requires the knowledge of $P(N_2 = 0)$, which itself changes with dependence, the expectation $E(N_2|N_2 > 0)$ is generally unknown. Therefore, in the numerical analysis we focus on the approximation based on $EN_2$ instead.

### 6.4 Numerical results

In this section we look at the relative performance of the M1, RHM and the vanilla methods represented by DVM and PVM. Collectively, we refer to the two-stage methods as M2. Below we compare the methods under the homogeneous and heterogeneous effect scenarios described in the previous section. We fix the number of tests $N$ and look directly at the power functions and other relevant quantities as functions of the shift parameter $\theta_i$. The goal of such comparison is to illustrate the additional power performance gains that RHM delivers relative to the two two-stage vanilla methods and the single-stage method M1 under the heterogeneous effect scenarios while at the same time providing equal or superior power performance to the two vanilla methods under the homogeneous effect scenario. In addition to the two perspectives we also focus on the so-called break-apart points, i.e. the number of tests for which a two-stage method outperforms M1. The goal is to illustrate the minimal magnitude of the multiple testing burden for which it is beneficial to utilize the two-stage design. Unlike in the formulation of the theoretical power functions, for the numerical analyses we considered two-sided alternatives. All numerical analyses presented in this chapter were performed using the statistical software R (R Core Team (2015)). The presented results are based on numerical evaluation of the underlying theoretical quantities instead of simulation.

#### 6.4.1 Cost and budget assumptions

Regarding the unit costs, we focus on two scenarios for $c_2$ while setting $c_1 = 1$ (in suitable units). First we consider inflated unit costs of $c_2 = 1000$, which represents a steep increase of unit costs between stages and therefore a difficult scenario for the two-stage approach to perform well. As an alternative, we also consider equal unit costs when $c_2 = c_1$, a borderline case which represents a setting perfectly suited for the use of two-stage analysis. For a fixed $n$, $N$ and $c_1$ the equation $B = nNc_1$ yields the corresponding budget necessary to perform M1. This is the budget we assume to have available for the two-stage analysis as well.

#### 6.4.2 Sample size assumptions

For the numerical analysis we fixed the baseline sample size (i.e. affordable sample size when using method M1) at $n = 100$, however, the specific choice here is not crucial. In the normal model with the assumption of proportional S1 unit costs and fixed S2 unit costs $c_2$ the specific value of $n$ is completely inconsequential in terms of the relative performance of the considered methods, which is driven not by the baseline sample size $n$ directly but by the
ratio of sample sizes affordable by each method, namely \( \kappa = \frac{n_1 + n_2}{n} \) (henceforth referred to as the *sample size amplification factor*). In such a setting, \( \kappa \) is invariant in \( n \) since both the numerator and denominator scale equally with \( n \), and so do the effective expectations of all of the \( z \)-test statistics involved. In other words, changing the value of \( n \) merely re-scales the effective location parameter, i.e. the values on the \( x \)-axes in the presented figures, without altering the values of the presented measures and parameters on the \( y \)-axes in the figures.

### 6.4.3 Budget savings

A way to measure the expected budget savings is to consider the expected saved budget ratio 

\[
R^+ = 1 - \frac{EB_{12}(\theta)}{B}
\]

Below we focus on the homogeneous effect scenario \( \mathbb{E}[|\theta|]N \), which is parametrized by the single effect \( \theta_1 \), and consider the budget savings from two perspectives. First, \( R^+ \) is calculated for each \( \theta_1 \) using the optimal values of \( \alpha_1 \) and \( \delta \) derived for each value of \( \theta_1 > 0 \). This represents the expected savings for an "unlucky" researcher that suffered a false negative finding with respect to \( \theta_1 \). In addition, we also calculated the expected saved budget ratio under the complete null hypothesis, i.e. when \( \theta_i = 0 \) for all \( i \), at which point any budget savings are in fact desirable since there are no non-zero effects to be found.

### 6.4.4 Perspective 1: Homogeneous effects

Assuming \( N = 10^6 \) tests, for each value of the shift parameter \( \theta_1 \) we calculated power functions of both M1 and the considered two-stage methods where we used the optimal combination of \( \alpha_1 \) and \( \delta \) for each M2 method. The results are plotted in Figure 6.2, which shows as functions of \( \theta_1 \) the M1 and M2 powers, the logit powers, the ratios of powers, the optimal values of \( \delta, \alpha_1 \), the corresponding \( \alpha_2 \) and the expected number of tests in S2, the expected affordable sample sizes in each stage, the expected budget savings, and finally the expected sample size amplification. The presented results are based on the approximate power functions discussed in Section 6.3.9, since the calculation of the exact power functions can be computationally quite intensive (especially for RHM and to a lesser degree also for PVM). The power approximations proved to work extremely well under the considered settings and the differences in power functions and other relevant measures were negligible.

**Power performance with inflated unit costs**

Looking at Figure 6.2, the top row of plots show a clear superiority of M2 over M1. In Figures 6.2a – 6.2d we notice a substantial positive gap between the powers of M1 and RHM, PVM and DVM in favor of the two-stage methods. Figure 6.2d shows the power of M2 to be as much as 65 times that of M1, while the relative power advantage remains several-fold for over half of the relevant effect sizes. The power logit functions in Figure 6.2c also show a consistent power advantage in favor of RHM, PVM and DVM. Judging by the power ratio the biggest relative improvement by the two-stage approach occurs for small effects, which
are the most difficult ones to find. Especially when the power ratio plot is viewed together with the power logit plot the superiority of the two-stage methods over M1 is clear.

**Power performance with equal unit costs**

As an alternative to the setting with substantially increasing unit costs, in Figure 6.3 we consider the comparison with equal unit cost, i.e., \( c_1 = c_2 \). In such a setting, we expected the two-stage scheme to work well and Figure 6.3c confirms that. It shows that with equal unit costs...
costs the power ratio can be nearly 20000, while the power logit plot shows a consistently and vastly superior performance by all three two-stage methods. Moreover, Figure 6.3a shows that for the effect size where M2 achieves power above 0.8 the power of M1 is still virtually zero. This shows that in areas with non-increasing unit costs and small number of non-zero effects the two-stage scheme can be incredibly powerful.

Optimal parameters
Next we focus on the optimal values of $\delta$ and $\alpha_1$ shown in Figures 6.2ef, 6.3de for the two scenarios. The figures show how the optimal budget allocation towards the initial stage increases rather rapidly with effect size reaching a split of about 90/10 for the largest of effects. Similarly, the optimal $\alpha_1$ (on the log-scale) increase with effect size for all three methods with a slight decrease for DVM for the largest of considered effects. Given that the optimal $\alpha_2$ is a deterministic function of $\alpha_1$ for all three two-stage methods, Figures 6.2f and 6.2g are perfect complements of each other. Moreover, Figure 6.2h shows the expected number of S1 rejections corresponding to the optimal values of $\delta$ and $\alpha_1$.

Expected sample sizes and budget
The source of power superiority of the two-stage methods is their ability to achieve increased sample size at S2 by eliminating a vast majority of the hypotheses in the initial stage. Since for the smallest effects the affordable sample sizes can be unrealistically large, we limit the sample size amplification factor $\kappa$ to at most 100. This means that the S2 sample size cannot exceed the baseline sample size $n$ by more than that value. Figure 6.2i shows the expected sample size amplification factors within the two-stage methods. Crucially, the overall sample size by the two-stage methods is in fact as much as hundred-fold for the near-zero effect size, while it consistently several-fold throughout most of the effective effect size range. For instance, at the effect size where the two-stage methods reach power of 80% the affordable overall sample size and the corresponding power are both approximately twice those of M1. Figure 6.2j shows the corresponding expected budget savings described in Section 6.3.5. The plot shows the potential for substantial budget savings using the two-stage methods especially under the complete null hypothesis (i.e. with no non-zero effects in the data).

Sensitivity to optimal parameters
Using two-stage methods requires optimization of $\alpha_1$ and $\delta$, which means that the sensitivity of the methods towards the values of these input parameters is a relevant question. In order to answer that question we plotted a power landscape of DVM for various combinations of the input parameters $\delta$ and $\alpha_1$. We chose DVM because it is the simplest to evaluate. The resulting power landscape plots are shown in Figures 6.4a–6.4f. We focused on the the homogeneous effect scenario $E|1|10^6$ and provide the landscapes for both the inflated unit costs $c_2 = 1000c_1$ (top row) and the equal unit costs $c_2 = c_1$ (bottom row). For the landscape plots it was necessary to specify an effect size, i.e. the value of $\theta_1$. We considered three
Figure 6.4: Power function landscapes for $\Pi_{12}^{DVM}$ at various combinations of $\delta$ and $\alpha_1$ under the homogeneous effect scenario $E||1|0^\delta$ with $c_2 = 1000c_1$ (top row) and $c_2 = c_1$ (bottom row) and $\theta$ such that M1 power is $\pi$ (i.e., $\Pi(\sqrt{\theta}, \gamma/N) = 0.8$), where $\pi = 0.2$ (left), $\pi = 0.5$ (middle), $\pi = 0.8$ (right). In each plot the maximum of $\Pi_{12}^{DVM}$ is denoted by a solid circle, while the power of M1 (located at $\delta = 1$, $\alpha_1 = \gamma/N$) is denoted by a solid square.

options, which were determined based on M1. From left to right in Figure 6.4 the plots correspond to effects for which M1 attains powers $\pi = 0.2$, $\pi = 0.5$ and $\pi = 0.8$. In each plot we denote the point of maximum of $\Pi_{12}^{DVM}$ by a circle and the power of M1 by a square.

First focusing on the top row of plots in Figure 6.4, perhaps the most interesting aspect of the power landscape under the inflated unit cost setup is the steep drop-off of the power for large values of $\alpha_1$. A similar but milder drop-off occurs also if $\delta$ is excessively small. Such behavior is intuitively clear since large $\alpha_1$ and/or small $\delta$ mean that a much larger than optimal fraction of the tests passes the initial stage. As a consequence, in the extreme case the second stage does not have the necessary sample size to deliver any meaningful power. As evidenced by the bottom row of plots in the figure, the power drop-off for large $\alpha_1$ is linked to the ratio of the unit cost functions $c_1$ and $c_2$. Contrasting the two sets of plots, it is clear that
Power advantage by RHM over DVM and PVM with \( m = 30 \) non-zero effects and \( N = 10^5 \)

**Ratios of ENDE among moderate effects only**

- (a) 10% huge effects
- (b) 20% huge effects
- (c) 30% huge effects
- (d) 50% huge effects
- (e) 70% huge effects
- (f) 90% huge effects

**Ratios of ENDE among all effects**

- (g) 10% huge effects
- (h) 20% huge effects
- (i) 30% huge effects
- (j) 50% huge effects
- (k) 70% huge effects
- (l) 90% huge effects

**Difference of ENDE**

- (m) 10% huge effects
- (n) 20% huge effects
- (o) 30% huge effects
- (p) 50% huge effects
- (q) 70% huge effects
- (r) 90% huge effects

Figure 6.5: Relative power improvements by RHM over DVM and PVM for various fractions of moderately sized effects out of the \( m = 30 \) non-zero effects. In each plot moderate effects equal \( \theta \) which varies according to \( \theta = \lambda \theta_0 \), where \( \theta_0 \) is such that \( M_1 \) yields power of 80% at \( \theta_0 \), while the non-moderate (huge) effects are such that \( M_1 \) detects them with power \( \approx 1 \). For each method (regardless of the value of \( \lambda \) and the fraction of huge effects) the values of \( \alpha_1 \) and \( \delta \) were optimized for target power of \( \pi = 0.8 \) under the scenario \( E[30|10^5] \) with \( c_2 = 1000c_1 \).

when there is no unit cost increase the drop-off all but disappears and the method becomes even less insensitive to the choice of \( \alpha_1 \). Interestingly, the degree of sensitivity of power with respect to \( \delta \) does not seem to differ substantially between the two unit cost scenario.

### 6.4.5 Perspective 2: Heterogeneous effects

In order to illustrate the degree of power improvement that RHM offers over the vanilla methods DVM and PVM we focus on the heterogeneous effect scenario \( U|m|N \) with \( m = \ldots \)
30 and \( N = 10^5 \). Unlike under the homogeneous effect scenario, here we must make an additional choice about nature of non-homogeneity of the distribution of effects. To that end, we assume that the \( m = 30 \) non-zero effects come from a mixture of two homogeneous effect scenarios, where a given fraction of effects are so-called moderate effects and the remainder are huge effects. Let \( \theta_0 \) be such that M1 (a Bonferroni corrected z-test) with \( N \) hypotheses yields power of 80% at level \( \gamma = 0.05 \). For such \( \theta_0 \) let the moderate effects be all equal to \( \lambda \theta_0 \), where \( \lambda \) ranges over \((0,1)\) and let the huge effects be such that M1 detects them with power \( \approx 1 \), say \( 2\theta_0 \). Consequently, the huge effects are trivially detectable by any of the considered methods, while only the moderate effects set RHM apart.

Given that in practice the actual distribution of effects is unknown, in the analysis the input parameters \( \alpha_1 \) and \( \delta \) for each method were optimized for power \( \pi = 0.8 \) under the homogeneous effect scenario \( E[30|10^5] \). In other words, we used the actual value of \( m \) in the optimization, but as we show in the sensitivity analysis of Section 6.5 specifying the value of this parameter exactly is not crucial. The additional parameter \( \alpha_0 \) within RHM was not optimized and instead, motivated by Figure 6.1, we simply put \( \alpha_0 = \alpha/50 \).

Under our heterogeneous effect scenario a suitable measure of comparative performance of two methods is the expected number of detected effects (ENDE). In Figure 6.5 we compare RHM with PVM and DVM using the ratios of ENDE and the differences of ENDE, which we plotted as functions of \( m \) ranging over \((0,1)\). The top row of plots in the figure shows the relative improvement of ENDE among the moderate effects only, the middle row shows the power improvement among all \( m = 30 \) non-zero effects, while the bottom row compares the methods in terms of differences of ENDE. Given that the huge effects are trivially detectable by any method, it seems more relevant to focus our attention primarily towards the top and bottom rows of plots in Figure 6.5. In the ratio of ENDE plots (top two rows) the \( y \)-axis shows relative improvement, where the baseline level of 100% means equal ENDE for the two methods and any value above that means that RHM improves over the corresponding vanilla method. Similarly, in the difference of ENDE plots (bottom) a positive value of the \( y \)-axis corresponds to superiority of RHM. In terms of the relative increase of ENDE among the relevant effects, the biggest improvement by RHM occurs when the fraction of moderate effects is small (i.e. many huge effects) as evidenced by the increasing maxima in Figures 6.5a – 6.5e. The inverse trends observed in both Figures 6.5f – 6.5j and Figures 6.5k – 6.5o do not contradict this conclusion. In fact, the reversal is merely a somewhat deceptive consequence of dissolving the improvement in an ever larger pool of trivially detectable effects and an ever diminishing space for improvement by RHM over the other two methods.

Based on the results of this section taken together with the comparisons under the homogeneous effect scenarios of Section 6.4.4 it is clear that RHM provides substantial power improvement over the existing vanilla two-stage methods. The message of Figure 6.5 is that RHM can yield a substantial improvement over both PVM and DVM. Moreover, the clear difference of performance under these settings illustrates the superiority of PVM over DVM under certain circumstances, which agrees with the conclusions by Skol et al. (2006).
### 6.4.6 Break-apart points

Consider a two-stage method which reduces to a single-stage method for certain parameter choices. A break-apart point (BAP) \( N_π \) for the two methods is defined as the biggest number of tests for which the single-stage special case yields optimal power. In general, BAPs depend on the distribution of effect sizes. For simplicity, we again focus on DVM under scenario E|1|N, where the BAPs can be equivalently characterized by the optimal power \( π \) corresponding to the single effect \( θ \), hence the indexing by \( π \). The relationship that links the two is \( \Pi(\sqrt{nθ}, γ/N_π) = π \). Due to the nature of the two-stage design the value of \( N_π \) depends on the expected affordable sample size at stage 2. With \( α_1 \) and \( δ \) optimized for DVM to yield maximum power, the BAP is determined by only three factors, namely the total budget \( B \), the distribution of \( N_2 \), and the unit cost function \( c_2 \).

#### Finding break-apart points

Recall that DVM uses level \( α_2 = α/α_1 \) at stage 2, where \( α = γ/N \) (Bonferroni correction). \( N_π \) is a BAP if it is the biggest number of tests that satisfies \( \Pi_{12}^{DVM}(θ_1^*, α_1^*, γ/(α_1^*N_π), δ^*) = π \), where \( α_1^* \), \( δ^* \) maximize \( \Pi_{12}^{DVM} \) at \( θ_1^* \), which is the value of the shift parameter for which M1 has power \( π \) (given \( n \)). In other words, \( θ_1^* \) satisfies \( \sqrt{nθ_1^*} = Φ_0^{-1}(1 − γ/N_π) − Φ_0^{-1}(1 − π) \). Since the power function \( Π \) depends on the shift parameter \( θ_1 \) only via \( \sqrt{nθ_1} \), the value of \( \sqrt{nθ_1^*} \) is invariant in \( n \), which makes \( N_π \) invariant in \( n \) as well, thus making it a function of \( π \) only. From the definition of \( N_π \) it follows that \( δ^* = 1 \) and \( α_1^* = γ/N_π \) (for which DVM turns into M1), since otherwise the two powers could not be equal. Consequently, for any combination of \( α_1 \) and \( δ \) the power \( \Pi_{12}^{DVM} \) at \( θ_1^* \) is at most \( π \), thus

\[
\Pi_{12}^{DVM}(n^{-1/2}(Φ_0^{-1}(1 − γ/N_π) − Φ_0^{-1}(1 − π)), α_1, γ/(N_π α_1), δ) ≤ π, \tag{6.6}
\]

for all \( α_1, δ ∈ (0, 1] \). By definition \( N_π \) is the largest value for which (6.6) holds, which allows identifying \( N_π \) for each \( π \) by simply looking for the smallest \( N \) which violates (6.6) for some \( α_1, δ ∈ (0, 1] \).

#### Results

Figure 6.6 shows the values of \( N_π \) as functions of \( π \) for various unit cost functions \( c_2 \). The plots show the minimum numbers of tests required to achieve superiority of the two-stage approach under the scenario E|1|N for four different unit cost functions \( c_2 \). From left to right in Figure 6.6 the four plots correspond to fixed ratio \( c_2/c_1 \) with \( c_2 = c_1 \), \( c_2 = 100c_1 \), \( c_2 = 1000c_1 \) and \( c_2 = 10000c_1 \), respectively. What we can is that \( N_π \) is generally a slowly varying increasing function for majority of the considered range of powers \( π \). Only for the most extreme powers very near 1 is there a steep increase in number of tests required to differentiate between the two methods. It is easy to see that such behavior is a direct consequence of the fact that the power functions basically coincide for such powers as evidenced by all of the power function plots in this chapter and an extreme scenario (in terms of \( N \)) is required.
6.5 Sensitivity analysis

Finally, we investigate the sensitivity of the two-stage methods to violations of the assumptions underlying the calculation of the distribution of $N_2$. Such question is relevant in application of the methods to real data since the assumptions are necessary for optimization of $\alpha_1$.
and $\delta$. Throughout this section we focus on DVM due to its numerical simplicity. The fact that the difference in performance between DVM and PVM is generally small (as evidenced by the results of this chapter presented above), qualitative conclusions based on DVM also apply to PVM and RHM. Throughout this section we again employ the simplifying assumption that the unit cost function $c_2$ does not depend on sample size $n_2$ and put $c_2 = 1000c_1$.

The two necessary assumptions underlying the input parameter optimization concern the non-zero effects count $m$ and independence of the initial stage test statistics. Both of these are relevant especially for optimization of the exact power functions, which are expectations with respect to $N_2$ (conditioned $N_2 > 0$), and for the optimization of the approximate power functions described in Section 6.3.9, since violations of the assumptions decrease the quality of such approximation. Here we focus on the sensitivity of the exact power functions.

First we focus on the problem of misspecified $m$. We compare the power performance of the methods under three homogeneous effect scenarios, which include both underestimation and overestimation of $m$. Based on the results we observed that the consequences of misspecifying $m$ are relatively limited under both overestimation and underestimation of $m$ by a reasonable factor. Regarding the assumption of independence, we investigate the power performance decline due to violations of independence. We focus on a multivariate normal model with band covariance matrix with geometrically vanishing off-diagonal elements and show that mild dependence leads to only very slightly suboptimal input parameters. We also observe that for extremely strong positive dependence among the tests the decrease of power is more severe, however, even then the two-stage procedure yields a several-fold improvement of the corresponding single-stage method within the relevant range of effect sizes.

### 6.5.1 Influence of misspecified false hypothesis count $m$

We investigated the influence of misspecified $m$ during optimization of the input parameters $\delta$ and $\alpha_1$ under five different homogeneous effect scenarios, namely $E[5]N$, $E[40]N$, $E[50]N$, $E[70]N$ and $E[250]N$ with $N = 10^6$. Using the parameters determined under each scenario we calculated the power function $\Pi_{12}^{DVM}$ under the scenario $E[50]N$. In other words, when the optimization was performed under $E[m]N$ with $m = 5, 40, 70, 250$, the value of $m$ was misspecified, while for $E[50]N$ the utilized parameters were actually optimal. Thus the results under $E[50]N$ serve a benchmark for judging the consequences of misspecifying $m$.

We present the results in Figure 6.7, where in 6.7a–b we plot the power functions and power function ratios corresponding the input parameter values of $\delta$ and $\alpha_1$ shown in Figures 6.7c–d. It is interesting to see that while the actual values of $\delta$ and $\alpha_1$ are influenced by misspecification of $m$, and their degree of deviation from the actually optimal parameters for $m = 50$ increased with the degree of misspecification, the corresponding power functions in fact change only slightly overall. Especially for the smaller effect sizes they remain virtually unaffected as Figure 6.7b makes clear. This behavior can be explained by the fact that none of the misspecification scenarios leads to utilization of a grossly inflated value of $\alpha_1$. Especially
6.5 Sensitivity analysis

when $m$ is only mildly misspecified (such as with $m = 40, 70$), the resulting value of $\alpha_1$ is almost exactly optimal, and so is the value of $\delta$ and consequently the power functions for those two scenarios virtually overlap with the ideal power function corresponding to $E|50|N$. The importance of not "over-shooting" the optimal value of $\alpha_1$ appears clear when looking at the power landscapes in Figures 6.4a – 6.4a, which correspond to the current unit cost setting. The figures show that, with $\delta$ near the optimum, there is a steep decline of power due to larger than optimal $\alpha_1$. On the other hand, the figures also show that the decline of power associated with suboptimal $\delta$ is much less severe. Consequently, given that misspecification of $m$ (by a reasonable degree) during parameter optimization under the considered scenarios seems to lead to neither excessively large $\alpha_1$ nor severely suboptimal $\delta$, the power loss due to misspecified false hypothesis count appears generally negligible.

6.5.2 Influence of dependence among tests

Next we investigate the sensitivity of the two-stage methods to violations of the assumption of independence, which can affect the performance by leading to sub-optimal values of the input parameters $\delta$ and $\alpha_1$. Using a model of dependence we compare the power functions corresponding to values of the parameter optimized under dependence with those that correspond to the actually optimal combination of parameters. We show that the differences between the two sets of parameters and the resulting power functions are negligible. This result greatly increases the practical usability of the two-stage methods, since lack of knowledge of the underlying dependence structure among the tests does not seem to result in vastly suboptimal values of $\alpha_1$ and $\delta$ even if the independence assumption is utilized during optimization.

Dependence model

We focus on a multivariate normal model of dependence with $N$-dimensional covariance matrix $\Sigma$. For a general $\Sigma$ in the high-dimensional setting it can be challenging to calculate the
probabilities either analytically or numerically even if the values in $\Sigma$ are known. Therefore, we consider a block-diagonal structure for $\Sigma$ with each block being a $b \times b$ band matrix. Denoting the bandwidth of the blocks by $w$, such covariance structure corresponds to the statistics within each block being a weighted sliding sum of $2w + 1$ independent normally distributed variables. In other words, the dependence between the statistics is induced via the overlap in such sliding sum. The block-diagonal covariance structure is numerically convenient, since the distribution of $N_2$ can be obtained by determining the distribution of rejections counts separately for each block and subsequent convolution.

The sizes of the covariance matrix diagonal blocks were $b = 200$, $b = 500$ and $b = 2000$. Paired up with these block sizes were bandwidths $w = 15$, $w = 50$ and $w = 100$. The vector of weights within each window of size $2w + 1$ was $q_w, q_w^{-1}, \ldots, q^2, q, 1, q, q^2, \ldots, q_w^{-1}, q_w$ with quotients $q = 0.75$, $q = 0.85$, $q = 0.95$ accompanying the three pairs of block sizes and bandwidths respectively. In other words, each block of the variance matrix is $CI_{2w+1}C'$, where $I_k$ is the $k$-dimensional unit matrix and $C$ is an $b \times (2w + 1)$ matrix defined as

$$C = \begin{pmatrix}
q_w & q_w^{-1} & \ldots & q & 1 & q & \ldots & q_w^{-1} & q_w \\
q_w & q_w^{-1} & \ldots & q & 1 & q & \ldots & q_w^{-1} & q_w \\
q_w & q_w^{-1} & \ldots & q & 1 & q & \ldots & q_w^{-1} & q_w \\
& & \ddots & & & & & & \\
q_w & q_w^{-1} & \ldots & q & 1 & q & \ldots & q_w^{-1} & q_w
\end{pmatrix},$$

where the blank elements are all zero. We refer to the three dependence settings as mild positive dependence ($b = 200$, $w = 15$, $q = 0.75$), moderate positive dependence ($b = 500$, $w = 50$, $q = 0.85$), and strong positive dependence ($b = 2000$, $w = 100$, $q = 0.95$). In the simulation scenarios of this section we focus on DVM and use $N = 10^5$ tests, $m = 30$ non-zero effects, baseline sample size $n = 100$ and 5000 replications.

**Results**

Using the positive dependence model we simulated the distribution of $N_2$ under the three dependence scenarios. In Figure 6.8 we contrast the actual distribution of $N_2$ (for $\alpha_1 = 10^{-3}$ and $\alpha_1 = 10^{-4}$) under each dependence scenario with the Poisson-binomial distribution which governs $N_2$ under independence. The plots show how dependence leads to over-dispersion of the distribution of $N_2$ with inflated probabilities of zero rejections. Unsurprisingly, the degree to which both of these effects occur increases with the strength of dependence.

In Figure 6.9 we plotted the performance measures and the corresponding input parameters under the three dependence scenarios contrasted in each plot the power measures and input parameters corresponding to the independence case. For comparison we also included the power measures for M1. The plots illustrate the amount of power loss due to the usage of suboptimal input parameters determined under independence. The power plots (firth three columns of Figure 6.9) show that while there is some loss, it is generally limited, especially under the mild and moderate dependence scenarios, where both the raw powers and their pro-
6.6 Conclusion

In this chapter we proposed a novel extension to the multi-stage design called Robin Hood. Unlike the existing methods, which mainly focus on the homogeneous effect scenarios, our Robin Hood method goes beyond these. We showed that with heterogeneous effects our extension substantially improves on the existing multi-stage methods in terms of power per-
Figure 6.9: Sensitivity of power performance and optimal parameter value to violations of the assumption of independence during parameter optimization for mild (top row), moderate (middle row) and strong dependence (bottom row). The parameters were optimized using the correct value of $m = 30$ both with the assumption of independence and under the actual distribution of $N_2$. 
formance, while requiring essentially zero efficiency trade-off in the original homogeneous effect realm. This makes our method very useful in many practical areas, where heterogeneous effects are the rule rather than the exception.

As already mentioned, in the field of genetics the multi-stage approach was originally proposed and utilized for genome-wide association studies over a decade ago. Though initially recognized as a potentially fruitful way of designing and analyzing experiments in GWAS, the popularity of the multi-stage approach faded in the recent years due to the occurrence of large genetic consortia, where data sets from numerous sources are pooled and jointly analyzed. As a consequence of this shift, the ability to pool data sets originally collected for different purposes has gained precedence over the efficiency of design and analysis within the individual constituent experiments. Although there are numerous reasonable arguments for prioritizing reusability of data in some areas (e.g. GWAS), it also needs to be stressed that such strategy comes at a cost. As showed by us in this chapter and others in literature, such costs in terms of efficiency can be very high, which suggests that maximizing the potential for reusability of data should not always be seen as an obvious all-important priority when experiments are designed.

Moreover, for other types of genomic data reusability is not as much of an issue to begin with. In many areas with large-scale sparse-effect problems and budget constraints, when additional subjects are potentially plentiful, even with highly increased unit measurement costs, the efficiency of the multi-stage design makes it a superior strategy. Therefore, we argue that it should be considered more widely by researchers seeking to maximize their chances to discover relevant effects.

**Acknowledgements**

This work was supported by the NWO VIDI Grant 639.072.412 and the Hypatia Tenure Track Grants of Radboud University Medical Center Nijmegen, Netherlands.
PART III

ASSOCIATING GENES AND MULTIVARIATE PHENOTYPES
Modeling multivariate phenotypes

In Part III of this thesis we focus on the problem of modeling and discovering association between genotypes, or in principal any covariates of interest, and multivariate correlated phenotypes. We focus on exploratory analysis based on regression, where genetic loci are used as explanatory variables for a multivariate phenotype, which is the dependent variable. Although we primarily formulate the problem and the methods within a genetic setting where the number of covariates of interest is larger than the dimension of the response, the methodology and the results can be readily applied in many other areas with similar characteristics.

An application area of interest to us is the analysis of behavioral data, where the phenotype (response) is assumed to be generated via a factor model from the underlying genetic variants (covariates). The phenotype typically consists of several related quantitative measurements which capture a psychological or behavioral condition, while the goal is to find a link between the condition and a subset of the genetic variants. Since the observed quantitative measurements all relate to a single underlying medical condition, they often exhibit correlation (dependence). Therefore, they should be treated simultaneously in a way that reflects this correlation in order to increase the statistical efficiency of the analysis.

Part III of the thesis is structured in the following way. In this chapter we provide a description of the problem of modeling multivariate phenotypes using genetic data. We introduce a multivariate multiple linear regression approach, where "multivariate" refers to the plurality of regressors, while "multiple" indicates that the response is also multivariate. The task at hand then becomes regression parameter estimation, hence we conclude the chapter with a brief overview and a discussion of available methods suitable for such task. In Chapter 8, we formulate a partly novel method called adaptive simultaneous variable selection, or adaptive SVS, which falls into the rich landscape of penalized regression methods such as the lasso (Tibshirani (1996)), ridge regression (Hoerl and Kennard (1970)), and the group lasso (Zou and Hastie (2005)). The adaptive SVS method is directly linked to two existing statistical methods. On the one hand, it is an extension of the simultaneous variable selec-
tion method introduced by Turlach et al. (2005). On the other hand, it is closely linked with the adaptive group lasso method of Wang and Leng (2008). The idea of adaptation in this context means the ability of the estimation objective function to better reflect the information within the data to which it is applied. Due to both penalization and adaptation, the use of the method requires choosing values for several input parameters. We describe how to choose these parameters under the genetic setting of interest, where we investigate the performance of the adaptive SVS method and compare it with several other suitable methods. Finally, in Chapter 9 we apply the adaptive SVS method to an eQTL analysis of an existing expression data set, where we look for SNP-driven gene expression regulation. The analyzed data comes from the Geuvadis RNA sequencing project for 1000 Genomes samples (Lappalainen et al. (2013)) and is publicly available. One of the results of our analysis is a comparison of the adaptive SVS method with a global testing procedure recently developed by Chaturvedi et al. (2015) specifically for such data.

The contents of Part III are being prepared for two publications. The real data analysis presented in Chapter 9 will be published in the upcoming article by de Menezes et al. (2016). It is also our plan to submit the contents of Chapter 8 for publication as a separate paper.

7.1 Setting of the problem

Our task is to model multiple correlated continuously varying phenotypes using a large number of genetic loci as regressors. Due to the sheer size of the genome the available sample size (the number of individuals) in such data is typically much smaller than the number of available regressors (loci), which means that we face a high-dimensional statistical inference problem. A popular method suitable for modeling such phenotypes is the multivariate multiple linear regression model, where the observed numerically represented genotypes at each locus are the regressors. Since the number of parameters in the full model far exceeds the number of available observations, a reliable inference about the model’s parameters is a daunting task which demands additional assumptions about the modeled phenomena. A popular approach to dealing with such systems is the concept of sparsity, which is the notion that the observed response is in fact influenced only by a subset of the available explanatory variables. Such assumption seems appropriate especially in GWAS, where the goal is to find the truly associated SNPs among tens or hundreds of thousands of candidates. This remains true even for phenotypes with thousands of truly associated SNPs, which might include many common complex diseases such as schizophrenia (Austin et al. (2013)).

7.1.1 Statistical model

Throughout Part III we assume an \( n \times p \)-dimensional random matrix of responses \( Y = (Y_{ij}) \) with independent rows and (possibly) dependent columns. Further assume we have an \( n \times q \)-
7 Modeling multivariate phenotypes

A dimensional regression matrix $X = (x_{ij})$ and a $q \times p$-dimensional (non-random) matrix of regression parameters $B = (\beta_{ij})$ such that

$$Y = XB + E,$$  \hspace{1cm} (7.1)

where $E$ is an $n \times p$-dimensional random matrix with zero mean and independent rows and variance matrix $\Sigma$. To eliminate the need for intercept in model (7.1), we center (on a per-column basis) both the response matrix $Y$ and the regression matrix $X$. In the intended genetic application the matrix $X$ contains genotype information at $q$ loci and $Y$ is a collection of $p$ phenotypes\footnote{The use of $p$ to denote the column dimension of $Y$ is precisely motivated by the fact, that $Y$ contains phenotypes.} for each of the $n$ individuals in a given sample. The rows of $Y$ can be viewed as a collection of $p$ phenotypes of a single individual, or they can be seen as a single $p$-dimensional multivariate phenotype. We take the latter view throughout this text.

In the classical regression model the parameters $p$ and $q$ are assumed to be fixed and asymptotic results are formulated for $n \to \infty$. However, in our context the number of loci $q$ is (much) larger than the number of individuals $n$, which means that the regression model is under-determined. We refer to this as the “large $q$” setting\footnote{In the literature this is often called “large $p$” setting, but here we denote by $p$ the column-dimension of the response matrix.}. From an asymptotic perspective the classical and the "large $q$" settings are conceptually very different, since the latter requires that $q$ stays above $n$ as $n \to \infty$, which leads to asymptotic theory where both $n \to \infty$ and $q \to \infty$ and we briefly discuss it further in Section 7.2. In any case, in the regression model (7.1) penalized regression methods have been successfully used to identify relevant groups of variables with good prediction accuracy (Efron et al. (2004), Zhang and Huang (2008), Kyung et al. (2010)).

Regression matrix assumptions

In order to reflect the nature of our target data we make two assumptions about the regression parameter matrix $B = (\beta_{ij})$ of model (7.1). The two assumptions are:

A1 **Sparsity**: There exists a combination of $i \in \{1, \ldots, q\}$, $j \in \{1, \ldots, p\}$ such that $\beta_{ij} = 0$.

A2 **Common association**: For any $i \in \{1, \ldots, q\}$, $j \in \{1, \ldots, p\}$ it holds $\beta_{ij} = 0$ if and only if $\beta_{ik} = 0$ for all $k \neq j$.

In essence, sparsity requires that not all of the $q$ available loci are actually associated with the multivariate phenotype, while common association specifies where the sparsity lies. Specifically, common association assumes that if one of the components of the multivariate phenotype is associated with a locus, then all other components of the phenotype are also associated with that locus, and vice-versa. Common association is related to but stronger than the narrow-sense sparsity condition (NSC) of Wei and Huang (2010), where the latter only requires that at least one regressor is unnecessary in the model, i.e. there exists $i \in \{1, \ldots, q\}$, $j \in \{1, \ldots, p\}$ such that $\beta_{ij} = 0$.
Setting of the problem

\{1, \ldots, q\} \text{ such that } \max_i |\beta_i| = 0. \text{ Both of these assumptions seem reasonable in our intended application and we rely on them (to a degree) in Chapter 9, where we attempt to identify genes involved in what is known as alternative splicing.}

Stacked-up model

Although many existing regression estimation methods are formulated for the univariate response, they can be easily applied to the multi-response model (7.1) by transforming the multi-response matrix \( Y \) and the regression parameter matrix \( B \) into vectors by "stacking up" their columns. This leads to

\[
\begin{pmatrix}
Y_\bullet_1 \\
\vdots \\
Y_\bullet_p
\end{pmatrix}
= 
\begin{pmatrix}
X \\
\vdots \\
X
\end{pmatrix}
\begin{pmatrix}
B_\bullet_1 \\
\vdots \\
B_\bullet_p
\end{pmatrix}
+ 
\begin{pmatrix}
E_\bullet_1 \\
\vdots \\
E_\bullet_p
\end{pmatrix},
\]  

(7.2)

where \( Y_\bullet_i, B_\bullet_i, E_\bullet_i \) are the \( i \)-th columns of matrices \( Y, B, E \) of (7.1), respectively. We denote the matrices in (7.2) from left to right by \( \tilde{Y}, \tilde{X}, \tilde{B}, \tilde{E} \). In other words \( \tilde{X} \) is a block-diagonal matrix with \( X \) on the diagonal repeated \( p \) times.

Utilizing correlation within the response

In the motivating setting of behavioral data, the measured responses are typically a set of closely correlated characteristics of a patient and the problem at hand is to identify association of these characteristics with genetic loci, which provides the need to consider correlated response columns. If considered separately, these characteristics are typically only loosely associated with the patient’s genotype, which makes identifying the association a very difficult task. However, considering the phenotypes as part of a correlated network and performing a joint association analysis has been shown to lead to more potent procedures. Available methods that provide exactly this kind of analysis are the graph-guided fused lasso (GFlasso) by Kim et al. (2009) and simultaneous variable selection (SVS) by Turlach et al. (2005) and Tropp et al. (2006), both of which will serve as benchmarks for the performance of the adaptive SVS method formulated in Chapter 8.

7.1.2 Desired properties of methods

When answering a scientific question by fitting a regression model to a sparse data set, in addition to efficient parameter estimation, it is often desirable to identify the true explanatory variables and discard the rest. This is known as model selection.Trimming down the effective number of regressors is a possible strategy for avoiding the problem of over-fitting, which refers to a situation when the selected model has many regressors and consequently fits the current data set quite well, but it fails miserably when predicting the response in an independent data set. Over-fitted solutions with large numbers of non-zero estimates tend to have undesirably high sensitivity to changes of the values of the regressors, which often
leads to highly variable predictions. Moreover, interpretability of such solutions within the context of the original scientific question tends to be very problematic. To summarize, a well-performing regression estimation method should

- correctly identify relevant regressors (loci), i.e., those regressors that are truly associated with at least one of the phenotypes have non-zero estimates,
- discard regressors with negligible or no association with the phenotypes, i.e. irrelevant regressors have exact zero estimates,
- provide quality estimates for the degree of association for each regressor,
- provide accurate and stable predictions of future phenotypes,
- allow for reasonably easy interpretation of the fitted model.

The first three criteria relate to the ability of an estimator to correctly identify and approximate the underlying nature of the data. In more technical terms they can be identified with the concepts of estimation consistency, selection consistency and oracle selection property.

Estimation consistency

For a general investigation of the asymptotic properties of estimators in the under-determined setting, both the value and the row-dimension of the true regression matrix $B$ from (7.1) is allowed to vary with $n$, which we denote by adding the lower index $n$ into the notation and denoting the true and estimated parameter matrices and the design matrix $B_n$, $\hat{B}_n$ and $X_n$, respectively. An estimator $\hat{B}_n$ is said to be estimation consistent (EC) for $B_n$, if the difference between $\hat{B}_n$ and $B_n$ vanishes in probability as the number of observations grows to infinity, that is $\hat{B}_n - B_n = o_p(1)$, as $n \to \infty$.

Selection consistency

An estimator $\hat{B}_n = (b_{ij}^n)$ is said to be selection consistent (SC) for $B_n = (\beta_{ij}^n)$ if, with probability approaching 1, it correctly assigns zero and non-zero estimates to all zero and non-zero parameters in $B_n$, respectively. Denoting the sets of non-zero coefficients in $B_n$ and $\hat{B}_n$ as $A_n = \{(i,j): \beta_{ij}^n \neq 0\}$ and $\hat{A}_n = \{(i,j): b_{ij}^n \neq 0\}$, the estimator $\hat{B}_n$ is SC if $\lim_{n \to \infty} P(\hat{A}_n = A_n) = 1$.

Rate of convergence

If, additionally, it holds $\hat{B}_n - B_n = O_p(n^{-1/2})$, we speak of $\sqrt{n}$-(estimation) consistency. Unfortunately, the $\sqrt{n}$ rate of convergency is only possible in the classical setting with fixed $q$ (or with $q < n$ growing slowly enough with $n$). In the "large $q$" setting, provided the penalty parameter is chosen "correctly" (i.e. $\lambda \approx \sqrt{n^{-1} \log q}$), a typical rate of convergence is $r_{n,q} = \sqrt{s^* \log q/(q_{n,q} \sqrt{n})}$, where $s^*$ is the (fixed) number of non-zero parameters in $B_n$ and $q_{n,q}$ is a measure of "sparse invertibility of the design matrix", where $q_{n,q} \gg 0$ is desirable.
Moreover, for consistent model selection it is required that every non-zero value in $B_n$ is away from zero by at least $r_{n,q}$, a property known as $\beta$-min. For a detailed overview we refer the reader to Buhlmann and van de Geer (2011), a standard reference in this area.

**Oracle selection property**

While both estimation and selection consistencies are desirable for an estimator in the sparse regression model, it is often difficult to achieve them simultaneously. And it becomes even more challenging to design a consistent estimator that is also required to have the oracle property, that is converge at an optimal rate. For a recent overview of results concerning oracle-type properties for various estimators in high-dimensional setting we refer the reader to Huang et al. (2012), Buhlmann and van de Geer (2011) or van de Geer (2016).

One of the original proposals for a framework for addressing the question of rate of convergence was made by Fan and Li (2001). They define an estimator $\hat{B}_n$ for $B_n$ to possess the oracle selection property if it is both estimation and selection consistent and asymptotically normal at the rate of $\sqrt{n}$. Unfortunately, although seemingly quite reasonable, the framework of Fan and Li (2001) turns out to have severe limitations. The property suggests that, without knowing which components of the true parameter are zero, an estimator can perform asymptotically just as well as if the correct zero components were in fact known. However, as Leeb and Potscher (2008) point out, there is a danger of inherent disconnect between the point-wise asymptotic behavior described by the properties and the actual finite sample performance of an estimator that possesses them. Moreover, Leeb and Potscher (2008) showed that an estimator performing consistent model selection must have an unbounded (scaled) risk function even though the risk function of the optimal estimator in the true model is bounded (Wagener and Dette (2013)). Based on this, Leeb and Potscher (2008) argue that the reasoning underlying the oracle concept of Fan and Li (2001) is misguided because it identifies estimators as optimal although their properties do not hold uniformly over the parameter space. It is clear that the oracle property of Fan and Li (2001) is not the appropriate concept of asymptotic optimality of estimators and a more appropriate oracle framework has been developed. The example of the oracle property illustrates, however, that the concept of asymptotic optimality in the high-dimensional sparse setting is complex. Besides, most asymptotic frameworks are highly idealized, which makes it strongly advisable to evaluate the finite sample behavior of an estimator in a realistic setting via a simulation study instead of solely relying on asymptotic optimality properties.

### 7.2 A selective overview of sparse data regression methods

In this section we list several popular regression methods and briefly discuss some of their properties. This section provides definitions of the methods included in the simulation study of Chapter 8, where we compare their performance with that of the adaptive SVS.
Classical regression methods

By far the most popular estimation method in the linear regression model is the classical method of ordinary least squares (OLS), which estimates $B$ by minimizing the sum of squared residuals. The estimator is $\widehat{B}_{ols} = \arg \min_B \| Y - XB \|_2^2$, where $\| \cdot \|_2$ is the Frobenius norm (i.e., $\|A\|_2 = (\sum_{ij} a_{ij}^2)^{1/2}$). Unfortunately, OLS is not suitable for the sparse data analysis since it yields only non-zero estimates for all parameters and thus lacks any model selection property. With the OLS estimator model selection is classically performed using subset selection methods such as Mallow’s $C_p$, Akaike’s information criterion (AIC) and the Bayes information criterion (BIC). These classical methods share common theme of penalization of the negative log-likelihood in a normal experiment, i.e., the regression sum of squared residuals, by the number of regressors actually included in the model. This can be seen as penalization by the $\ell_0$ norm of the vector of regression parameters, where the $\ell_0$ norm of a vector $x \in \mathbb{R}^n$ is defined as $\|x\|_0 = \sum_i \text{sgn} |x|$, i.e. it counts the non-zero elements in $x$. However, if the number of regressors is even moderately high, these classical methods become computationally infeasible, since the number of candidate models grows exponentially with the numbers of regressors (Yuan and Lin (2007)). Another problem with the subset selection methods is their extreme variability because of their inherent discreteness (Zou (2006)).

Putting aside the lack of model selection behavior of OLS, it can be directly applied to the multivariate multi-response model (7.1) only if the number of regressors $q$ is smaller than the number of individuals $n$, otherwise the design matrix $X$ (and consequently also $\tilde{X}$) is not of full rank ($X'X$ is singular). Consequently, there are infinitely many OLS solutions, which leads to instability of the estimator and makes the estimates difficult to interpret. This behavior can be remedied by treating sufficiently small subgroups of regressors separately, thus making the subgroup linear model well-determined. However, in general, applying OLS to these subgroup models can yield only limited insight into the association of response and the regressors. Nonetheless, a surprisingly popular method of analysis in certain GWAS studies (Austin et al. (2013)) is to focus on subgroups of size one, measuring association of each regressor with the phenotype separately via parallel univariate OLS solutions. We refer to such approach as naive OLS and discuss it in more detail in Section 7.2.1.

Penalized regression methods

The concerns with the classical methods have led to the development of a wide range of alternative regression methods, which typically penalize via higher order norms. Popular methods are the least absolute shrinkage and selection operator (lasso) by Tibshirani (1996), ridge regression by Hoerl and Kennard (1970), least angle regression selection (LARS) by Efron et al. (2004), among others. The benefits of penalization by higher order norms include possible automatic model selection behavior, improved prediction accuracy, as well as avoiding the computational and stability problems of the classical methods. The lasso is arguably the most popular penalized regression method (Breheny and Huang (2011)), which penalizes by the $\ell_1$ norm of the regression parameters. Changing the norm to $\ell_2$ leads to ridge
7.2 A selective overview of sparse data regression methods

regression. LARS on the other hand is conceptually related to a classical method known as forward model selection. Roughly speaking, LARS starts with an empty model, from which it "builds up" by incrementally adding regressors based on their correlation with the response by following an equiangular direction among the already included regressors. Efron et al. (2004) showed that LARS and lasso share the same geometry and are closely related with the solutions being nearly equal. In fact, a simple modification of LARS yields an entire solution path of the lasso estimates at different values of penalties. Remarkably, it is extremely efficient and the computation takes the same order amount of time as a simple least squares fit (Breheny and Huang (2011)). For this reason it is often used for the computation of the lasso. One reason for the popularity of the lasso is its ability to yield sparse solutions, which means that the method simultaneously performs parameter estimation and model selection. This led to the development of various extensions of the lasso suitable for various specialized application. Ridge regression on the other hand does not have this sparsity property, which made it much less appealing in many applications.

In the rest of this chapter we present relevant regression estimation methods within the multi-response regression model (7.1). The list does not aspire to be a complete overview of available methods and a much more complete overview can be found for instance in Fan and Lv (2010). Instead, our list focuses on methods that provide a build-up to the adaptive SVS method of the next chapter. The first item on the list is the method of ordinary least squares, which is actually not a penalized regression method. We then move to other methods that are based on the aforementioned penalization such as the lasso and its extensions such as the bridge regression, the group lasso and the adaptive group lasso, the graph-guided fused lasso and simultaneous variable selection. The adaptive SVS method formulated in Chapter 8 is an extension of the simultaneous variable selection, and, in a special case, it has the same objective function as the adaptive group lasso.

For the sake of notational simplicity, we do not always explicitly denote the dependence of the penalized estimators on their tuning penalty parameter (typically denoted as $\lambda$), although we do denote it in each penalized estimator’s definition. Similarly, we do not explicitly denote the dependence of estimators of $B$ on the sample size $n$.

### 7.2.1 Univariate approach: Naive OLS

Naive OLS is a very simple approach to association analysis, in which a single-response univariate linear regression model is applied to a pre-defined (usually linear) function of the phenotypes that transforms the multivariate phenotype into a univariate response. Limiting ourselves to the linear transformations, a vector $Z$ is defined as

$$Z = Yc$$

(7.4)

using a user selected vector of constants $c \in \mathbb{R}^p$. Then, instead of $Y$, the vector $Z$ is modeled using $q$ univariate regression models with the genotypes at a single locus as the regressor in
each of the \(q\) models. The \(q\) parameters are estimated using

\[
b_{k}^\text{naiveols} = \arg \min_b \| Z - bX_{\cdot k} \|, \quad k = 1, \ldots, q,
\]

(7.5)

where \(X_{\cdot k}\) is the \(k\)-th column of \(X\). The \(q\) estimates express the degree of association (linear correlation) of each genetic locus with the transformed univariate response. The hope is that the information obtained this way for \(Z\) translates also to the original multivariate phenotype \(Y\). The simplest choice for \(c\) in (7.4) is \(c = (1, \ldots, 1)'\), which we use throughout this text.

**Multiple testing correction**

Under the usual linear model assumptions such as normality, the association of a regressor (locus) is assessed using the corresponding \(p\)-value with a suitable multiple testing correction (MTC). We must stress the need to correct for multiple testing which is necessary to avoid many false selections when there are many regressors. In our application in Chapter 8 we use the Bonferroni correction within the naive OLS, where we also include the naive OLS method without MTC in order to illustrate its poor performance.

**Criticism of the naive OLS**

Despite its popularity, usefulness of the univariate approach for estimation and/or model selection seems limited. First problem with the naive OLS approach is its inefficiency. The modeling of association for each regressor separately effectively reduces to calculating the Pearson correlation coefficients between the single locus genotypes and the transformed phenotype \(Z\). While computationally appealing, such analysis can be obviously very simplistic. Another major problem is the choice of values in \(c\), where there is a danger that the transformed phenotype might not preserve the information about the links between the regressors and the phenotype that user aims to discover. Inevitably, any specific value imposes possibly strong and unjustified assumptions about the exact relationship between the individual components of the phenotype and the genotypes. For instance, selecting \(c = 1\) amounts to the assumption that all components of the multivariate phenotype have the same direction (sign) of association with the causal genetic loci. Moreover, as soon as one selects a transformation, applies it to the multivariate phenotypes and uses the \(q\) univariate models to estimate coefficients for each regressor, it is not possible to make predictions for the original components of the multivariate phenotype using these parameters. A possible workaround is to use the \(p\)-values associated with the \(q\) estimates to perform model selection by discarding loci with multiple testing corrected \(p\)-value above selected level \(\alpha \in (0, 1)\). If this results in elimination of sufficiently many regressors so that the design matrix with the remaining regressors is full rank, one can build a model for the stacked-up multivariate phenotype around such design matrix and use OLS to fit the model. However, such cumbersome analysis is far from optimal and a more complete and at the same time simpler analysis method is clearly preferable. Fortunately, a number of better alternatives exist.
7.2 A selective overview of sparse data regression methods

7.2.2 Lasso

Unlike the non-penalized least squares, the lasso is directly applicable in the context of multivariate multiple regression. The lasso estimator of the regression matrix $B$ is defined as

$$
\hat{B}_{\text{lasso}}(\lambda) = \arg\min_B \|Y - XB\|_2^2 + \lambda \|B\|_1,
$$

(7.6)

where $\lambda > 0$ is a suitably chosen tuning parameter and $\| \cdot \|_1$ is the matrix $\ell_1$ norm, i.e., $\|A\|_1 = \sum_{ij} |a_{ij}|$ for a matrix $A = (a_{ij})$. Due to the $\ell_1$-type penalty the lasso yields sparse solutions whenever $\lambda$ is chosen small enough. This is a very useful property whenever the sparsity assumption A1 of Section 7.1.1 holds. However, the shape of the penalty in (7.6) does not utilize the common association assumption A2 of the same section, which makes it less suitable for the situations when A1 is appropriate and could be exploited.

**Consistency of the lasso**

The lasso estimator is a special case of the bridge regression estimator (Frank and Friedman (1993)), which is defined as

$$
\hat{B}_{\text{bridge}}(\lambda) = \arg\min_B \|Y - XB\|_2^2 + \lambda \|B\|_\gamma^\gamma,
$$

(7.7)

for a given $\lambda > 0$ and $\gamma > 0$, where $\|A\|_\gamma^\gamma = (\sum_{ij} |a_{ij}|^\gamma)^{1/\gamma}$ for a matrix $A = (a_{ij})$. For the fixed $B$ setting (i.e. "small $q$"), under certain regularity conditions on $X$, the bridge estimators with $\gamma \geq 1$ were shown by Knight and Fu (2000) to be estimation consistent if the penalty parameter $\lambda$ is sample size dependent and $\lambda_n = o(n)$. Since the bridge estimator turns into the lasso for $\gamma = 1$, these results immediately apply to the lasso under the same setup and regularity assumptions. However, Knight and Fu (2000) also show that if $\lambda_n/n \rightarrow \lambda_0$ and $\lambda_0 > 0$, the bridge estimator for the non-zero parameters is asymptotically biased. Moreover, as Zhang and Huang (2008) points out, for certain configurations of the design matrix and the parameters the asymptotic point mass probability at 0 of the lasso estimator for the truly zero coefficients is less than one. This suggests that the lasso is not selection consistent under certain conditions even in the fixed $q$ setting. Consistency of the lasso was further studied by Zhao and Yu (2006), who formulated a condition of irrerepresentability, which, under properly behaving error terms in the fixed $B$ setup, is sufficient and almost necessary for both types of consistency. Under additional regularity conditions, they also show that their irrerepresentability condition implies selection consistency of the lasso under the setup of $B$ dependent on $n$, provided that the number of non-zero elements in $B$ grows "not too fast". Further consistency results for the lasso in the "large $q$" setting were obtained by Zou (2006) but especially by Meinshausen and Bühlmann (2006) and Zhang and Huang (2008). A general message about the lasso is that under many scenarios there is a trade-off between estimation consistency and selection consistency. This trade-off occurs because the value of the penalty parameter required to achieve selection consistency causes the parameter estimates of the non-zero parameters to be asymptotically biased by over-shrinking them.
Moreover, if the lasso is selection consistent, then it is not efficient for estimating the non-zero parameters.

**Limitations of the lasso**

In many applications of the linear regression model, we must deal with **collinearity**, i.e., high correlation among the regressors. Among genetic loci, correlation occurs in the form of **linkage disequilibrium** (LD) (see Section 2.3) typically as a consequence of common inheritance of loci that are physically close to each other in the genome. A somewhat undesirable property of the lasso is that in a group of highly correlated regressors lasso tends to select only one variable and it chooses it (essentially) arbitrarily. Therefore, any conclusion about locus-phenotype association based on the lasso estimate must be made with regard to all loci (highly) correlated with the locus selected by the lasso. Additionally, under collinearity, the prediction performance of the lasso can be dominated by the ridge estimator, as evidenced empirically by Tibshirani (1996) for the case of \( q < n \), and by the works of Zou and Hastie (2005) and Aseervatham et al. (2011) when \( q \geq n \).

An often cited limitation of the lasso in the case of \( q > n \) is that it can never select more than \( n \) variables because of the nature of the convex optimization problem (Zou and Hastie (2005)). This, however, does not appear to be a substantially troubling feature in an exploratory GWAS setting, when typically \( n \) is in the order of thousands, while the hope is to identify a reasonably small group of candidate loci for further research. In such case one might actually welcome such "limited selection" by the lasso, and in fact even desire a substantially fewer than \( n \) selected loci.

### 7.2.3 Adaptive lasso

As an extension to the lasso, Zou (2006) suggested the **adaptive lasso** defined as

\[
\overline{B}_{\text{lasso}}(\lambda) = \arg\min_{B = (\beta_{ij})} \|Y - XB\|^2_2 + \lambda \sum_{i,j} w_{ij} |\beta_{ij}|.
\] (7.8)

The difference between the penalty of the lasso and that of the adaptive lasso is the presence of weights \( w_{ij} \) in the latter. As we can see from (7.8), employing the weights inside the \( \ell_1 \)-type penalty allows the user to differentiate the amount penalization each regressor receives, thus permit some of them to obtain larger estimates (in absolute value sense) compared to the lasso while forcing the rest closer to zero. In the fixed \( B \) setup, Zou (2006) shows that if the weights are data-dependent and appropriately chosen, the estimator can achieve efficient estimation in the sense of the oracle property of Fan and Li (2001). Huang et al. (2008) extended this result to the case of sample size dependent \( B_n \) with increasing number of non-zero coefficients. These results give the idea of adaptation a lot of appeal, which is further enhanced by the fact that, similarly to the lasso, finding the adaptive lasso estimate is a convex optimization problem, which can be efficiently solved.

The idea of improving on the lasso by using adaptation has been applied in a Bayesian...
setting by Sun et al. (2010), who proposes a method called *iterative adaptive lasso*. Similarly to our problem Sun et al. (2010) in their investigation focus on mapping association of multiple loci with quantitative phenotypes in a GWAS setting and they show that their adaptive procedure performs well in such a setting. In Chapter 8 we utilize the idea of adaptation inside the adaptive SVS method.

### 7.2.4 Group lasso

In the *sparse common association* setting it is often beneficial to use an estimator that utilizes the common association aspect of the data. Since both the lasso and the adaptive lasso are invariant to reordering of the coefficients in the penalty (assuming the associated weights in the adaptive lasso are reordered in the same way), they do not exploit the common association assumption. In an effort to improve on this, Zou and Hastie (2005) and Yuan and Lin (2006) introduced a generalization of the lasso called the *group lasso*. For the univariate "stacked-up" model of (7.2), assuming the regressors are grouped into $K$ groups according to a predefined criterion, it is defined as

$$
\hat{B}_{\text{glasso}}(\lambda) = \arg\min_{\hat{B}} \|\hat{Y} - \hat{X}\hat{B}\|_2^2 + \lambda \sum_{k=1}^{K} \|\hat{B}_k\|_{G_k},
$$

where $\hat{B}_k$, with $k = 1, \ldots, K$, are sub-vectors of $\hat{B}$ containing parameters of those regressors that belong to the $k$-th group and $G_1, \ldots, G_K$ are data-independent positive definite matrices which generate weighted norms according to $\|\beta\|_{G_k} = (\beta'G_k\beta)^{1/2}$. At the group level this procedure acts similarly to the lasso by enforcing sparsity, when entire groups of predictors may drop out of the model depending on $\lambda$. At the within-group level the group lasso does not enforce sparsity, which, however, does not present a troubling aspect for our analysis under the common association assumption with parameters grouped per locus.

### 7.2.5 Adaptive group lasso

It has been shown that the penalty inside the group lasso can be excessive, which can both deteriorate estimation efficiency (Fan and Li (2001)) and affect selection consistency (Leng et al. (2006), Zou (2006), Yuan and Lin (2007)). In order to avoid this undesirable behavior, Wang and Leng (2008) borrowed the idea of adaptive estimation employed by Zou (2006) in the case of the adaptive lasso, and proposed the *adaptive group lasso* estimator

$$
\hat{B}_{\text{aglasso}}(\lambda) = \arg\min_{\hat{B}} \|\hat{Y} - \hat{X}\hat{B}\|_2^2 + \lambda \sum_{k=1}^{K} w_k \|\hat{B}_k\|_{G_k},
$$

Analogously to the adaptive lasso they allow the groups of regressors to be treated differentially through the weights calculated using an initial estimate of $B$. Wang and Leng (2008) studied the consistency and oracle properties of the adaptive group lasso estimator in the fixed $B$ setting and concluded that the method can possess all three of the properties, provided both the penalization parameter $\lambda$ and the weights $w_1, \ldots, w_K$ are chosen suitably.
Denoting \( \tilde{\mathcal{A}} \) the set of indices corresponding to the non-zero elements in \( \tilde{\mathcal{B}} \) and putting \( u_n = \max\{\lambda w_k : k \in \tilde{\mathcal{A}}\} \), \( v_n = \min\{\lambda w_k : k \notin \tilde{\mathcal{A}}\} \), they showed the following. For \( \sqrt{n} \)-estimation consistency it is sufficient that \( u_n = o_P(n^{-1/2}) \), as \( n \to \infty \), while for selection consistency and oracle property it is sufficient if both \( u_n = o_P(n^{-1/2}) \) and \( \sqrt{n} v_n \to_p \infty \). In the high-dimensional setting with varying regression matrix \( \mathcal{B}_n \) the adaptive group lasso has been studied by Wei and Huang (2010) and Wei et al. (2011). Under certain conditions with suitably selected weights \( w_k \) the method has been shown to be consistent in group selection (Wei and Huang (2010)) and to possess the oracle selection property (Wei et al. (2011)). It is also worth pointing that the adaptive group lasso can be viewed as a special case of the group lasso, since the adaptation is numerically equivalent to selecting suitable \( G_1, \ldots, G_K \). Therefore, an algorithm for finding a solution of (7.9) for an arbitrary \( G_1, \ldots, G_K \) can be used to find the solution of (7.10).

### 7.2.6 Graph-guided and graph-weighted fused lasso (GFlasso)

In the effort to bind parameter estimates closer together Tibshirani et al. (2005) introduced the fused-lasso, which for the univariate "stacked-up" model of (7.2) solves

\[
\tilde{\mathcal{B}}_{\text{lasso}}(\lambda, \mu) = \arg \min_{\tilde{\mathcal{B}} = (\tilde{\beta}_k)} \|\tilde{Y} - \tilde{X} \tilde{\mathcal{B}}\|_2^2 + \lambda \|\tilde{\mathcal{B}}\|_1 + \mu \sum_{k=1}^q |\tilde{\beta}_{i+1} - \tilde{\beta}_i|_1,
\]

where \( \lambda, \mu \) are tuning parameters and \( \tilde{\mathcal{B}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_q)' \). The fused lasso essentially binds together parameter estimators of neighboring regressors via an \( \ell_1 \) penalty.

The idea of fusion can be quite useful in the multivariate multi-response regression model. Kim et al. (2009) and Kim and Xing (2009) use it to introduce a method for penalized parameter estimation in a multivariate multi-response regression model. Their method is called graph-guided fused lasso (GFlasso) and it is based on the lasso with a secondary fusion penalty introduced to bind the estimates of parameters of the same regressor when highly correlated responses are modeled. Their graph-guided fusion penalty is guided by a phenotype graph. More specifically, using correlations of the phenotypes a graph of phenotypes as nodes is constructed in which two phenotypes are connected by an edge if their correlation exceeds a preset bound. Whenever two nodes are connected by an edge the parameter estimates for the corresponding components of the phenotype are fused together via the \( \ell_1 \) norm. For more adaptability, the terms inside the fusion penalty are weighted by the amount of correlation between the components of the phenotype, which results in the graph-weighted fused lasso (GwFlasso) defined as

\[
\tilde{\mathcal{B}}_{\text{Gw}}(\lambda, \mu) = \arg \min_{\tilde{\mathcal{B}}} \|\tilde{Y} - \tilde{X} \tilde{\mathcal{B}}\|_2^2 + \lambda \|\tilde{\mathcal{B}}\|_1 + \mu \sum_{k=1}^p w(r_{kl}) \|\tilde{\beta}_{k\bullet} - \text{sgn}(r_{kl}) \tilde{\beta}_{\bullet l}\|_1.
\]

The tuning parameters \( \lambda \) and \( \mu \) determine the amount of penalization by each penalty, while \( \tilde{\beta}_{k\bullet} \), \( \tilde{\beta}_{\bullet l} \) are the \( k \)-th and \( l \)-th columns of \( \tilde{\mathcal{B}} \), respectively, and \( r_{kl} \) is the Pearson correlation coefficient of the response vectors \( Y_k \) and \( Y_l \) (columns of \( Y \)) and \( w(r) \) is a weight function,
i.e., a non-negative function on \((-1, 1)\). Additionally, Kim et al. (2009) require \(w(r)\) to be equal to 0 on \((-\rho, \rho)\) where \(\rho \in (0, 1)\) is a suitably chosen cut-off value so that the pairs of phenotypes with correlation coefficient below \(\rho\) (in absolute value) do not enter the fusion penalty in (7.12) at all.

As a simplification of GwFlasso Kim et al. (2009) also define a non-weighted estimator called graph-constrained fused lasso (GcFlasso) estimator, which is obtained from GwFlasso by selecting the weight function \(w(r) = I(|r| > \rho)\). Through simulation Kim and Xing (2009) show that the performance of GcFlasso is dominated by GwFlasso with the weight function either \(w(r) = |r|\), in which case it is called GwFlasso(1), or \(w(r) = r^2\), which is referred to as GwFlasso(2). They also show that GwFlasso(1) and GwFlasso(2) provide very similar performance. A more detailed discussion can be found in Chen et al. (2012), Kim et al. (2009) and Kim and Xing (2009). Since the graph-weighted fused lasso is specifically designed for the type of data that we aim to analyze the adaptive SVS of the next chapter, this method is the performance benchmark for our method. Based on the empirical evidence by Kim and Xing (2009), we only consider GwFlasso(1) in the simulation study, which we refer to simply as GFlasso.

### 7.2.7 Simultaneous variable selection (SVS)

An alternative penalized regression approach specifically designed for the multi-response multivariate regression model setting is called simultaneous variable selection (SVS). It is based on the idea to penalize the sum of squared residuals by the sum of \(\ell_\alpha\) norms of the rows of the parameter matrix \(B\), thereby forcing the resulting estimates closer together. The SVS estimator is defined by

\[
\hat{B}_{\text{SVS}}(\lambda) = \arg \min_B \|Y - XB\|_2^2 + \lambda \sum_{k=1}^q \|\beta_{k\bullet}\|_\alpha,
\]

where \(\lambda > 0\) is again a tuning constant and \(\beta_{k\bullet}\) is the \(k\)-th row of the matrix \(B\) and \(\alpha\) is a positive constant determining the type of norm that is applied to the rows of the parameter matrix \(B\) before these norms are summed up to create a penalty. Using the notation of Obozinski et al. (2010), the penalty in (7.13) is a combined \(\ell_1/\ell_\alpha\) type norm. The case with \(\alpha = 1\) turns SVS into the lasso, while \(\alpha = \infty\) is referred to as the \(L_\infty\)-SVS and was proposed and studied by Turlach et al. (2005) and Negahban and Wainwright (2011). The case of \(\alpha = 2\) is referred to as \(L_2\)-SVS and it was also mentioned by Turlach et al. (2005), but they did not pursue it in detail because of computational efficiency considerations. It was later studied in detail by Malioutov et al. (2005), Similä and Tikka (2007) and Obozinski et al. (2011) among others.

A quick look at the group lasso estimator of (7.9) reveals that the \(L_2\)-SVS can be viewed as a special case of the group lasso estimator with \(q\) groups applied to the "stacked-up" variables of (7.2), where each of the groups corresponds to one coordinate of the multivariate phenotypes, and the weight matrices \(G_1, \ldots, G_q\) are all equal to the \(p\)-identity matrix \(I_p\). As such, \(L_2\)-SVS has the attractive property that it performs variable selection at the group level.
and is invariant under (group-wise) orthogonal transformations, making it similar to ridge regression in that regard (Meier et al. (2008)). Moreover, the asymptotic consistency results formulated by Bach (2008) for the group lasso directly apply to $L_2$-SVS.

Examining the shape of the penalty in (7.13), we notice that its fusion effect is more subtle when compared for instance with the graph-guided and graph-weighted fused lasso methods. The effect of taking a combination of $\ell_1$ and $\ell_\alpha$ norms lead to between-group sparsity achieved by the $\ell_1$ norm, while the higher order norm leads to fusion of the estimates corresponding to the same regressor. For instance, with $\alpha = 2$ (i.e. $L_2$-SVS) the degree of fusion will be strong, since all estimators in the same row in $B$ will be pushed close together to minimize the sum of areas of the corresponding squares. On the other hand, $\alpha = \infty$ (i.e. $L_\infty$-SVS) penalizes only the maximum value of the estimator in a given row, while the remaining terms are allowed to roam free (in absolute value) between zero and that maximum. As we will show in the next chapter, this balancing act between the two norms within the penalty can result in high quality estimates in the sparse common association setting, while further improvement can be achieved by adding adaptation into the SVS method.

### 7.2.8 Other regression methods

The area of penalized regression methods has flourished for the last two decades and much knowledge has been accumulated both in the classical and especially in the high-dimensional setting (Bickel et al. (2008), Zhang and Huang (2008), Castillo and van der Vaart (2012), Bühlmann et al. (2013), Castillo et al. (2015)). Existing methods are being studied both from the frequentist and the Bayesian perspectives (Castillo et al. (2015)), while novel methods are being proposed. Examples of existing methods suitable for high-dimensional regression analysis we did not discuss in the overview include the Dantzig selector (Candes and Tao (2007)), SQRT Lasso (Belloni et al. (2011), LAD Lasso (Wang (2013)), $L_q$ Lasso (which combines SQRT and LAD lassos), and others.
Adaptive simultaneous variable selection

In the previous chapter we provided a list of methods suitable for estimation of the regression matrix $B$ in model (7.1). However, only some of these methods properly reflect the nature of the current problem of modeling multivariate phenotypes with correlated coordinates. While the problems associated with the use of OLS are obvious and need not be further discussed, the design of the lasso does not reflect the multivariate correlated nature of the response variable, where additional information is left unutilized. On the other hand, GFlasso represents a natural extension of the lasso, which seeks to improve on the lasso in the current model by utilizing the covariance structure of the multivariate response by introducing a second penalty, although this comes at a price of higher computational complexity as well as a more difficult choice of tuning penalty parameters $\lambda$ and $\mu$. The adaptive group lasso and the SVS methods, on the other hand, require the choice of only one penalty parameter, while also reflecting the multivariate nature of the response to some extent. The standard (non-adaptive) SVS works with a general $\ell_1/\ell_\alpha$ norm combination, while the adaptive group lasso works with the special case of $\ell_1/\ell_2$ norm combination. Unlike a simple $\ell_\alpha$ norm, the combined $\ell_1/\ell_\alpha$ norm allows for a more efficient utilization of the correlation between the components of the phenotypes. We can apply the $\ell_1$ part to achieve sparsity between regressors while applying the $\ell_\alpha$ norm over the estimates that correspond to the same component of the phenotype thereby binding them closer together. Through the weights and the possibility of choosing matrices $G_k$ inside the $\ell_1/\ell_2$ penalty the adaptive group lasso further adds the ability to provide more “freedom” to those estimates of parameters corresponding to those regressors that appear to be more relevant for explaining the response, while at the same time being potentially more stringent with those regressors that do not show evidence of association with the phenotypes.

Looking at the SVS and adaptive group lasso side by side, it seems potentially beneficial to take the general $\ell_1/\ell_\alpha$ norm combination of the (non-adaptive) SVS (with $\alpha > 1$) and combine it with the idea of adaptation by introducing additional weighting of the terms inside the $\ell_1$ part of the penalty similarly to the adaptive group lasso. Below we propose such
8.1 Method

We define the adaptive SVS (aSVS) estimator of the parameter matrix $B$ of (7.1) as

$$
\hat{B}_{aSVS}(\lambda) = \arg\min_B \|Y - XB\|_2^2 + \lambda \sum_{k=1}^{q} \pi_k \|\beta_k\|_\alpha,
$$

where $\lambda > 0$ is a tuning parameter and $\pi_k, k = 1, \ldots, q$, are non-negative penalization weights.

We will additionally require that the penalization weights satisfy the condition $\sum \pi_k = q$ (which can be easily achieved by re-scaling, provided $\pi_k$ are all finite) so that the tuning parameter $\lambda$ better corresponds to the tuning parameter of the (non-adaptive) SVS in (7.13). Reasonable choices for the norm parameter $\alpha$ in (8.1) are the same as in the case of the (non-adaptive) SVS. Putting $\alpha = 1$ leads to a version of the adaptive lasso by Zou (2006), while $\alpha = 2$ and $\alpha = \infty$ will be referred to as $L_2$-aSVS and $L_\infty$-aSVS, respectively. The case of $\alpha = 2$ is also known as the adaptive group lasso by Wang and Leng (2008).

Adaptation by weighing

The motivation behind the penalization weights is to change the degree by which each regressor (locus) contributes to the value of the penalty, thus making the estimate adapt to additional information available to the user as expressed through $\pi_k$. A large value of $\pi_k$ means that the $k$-th regressor (locus) is heavily penalized relative to the regressors with small penalization weights, thus restricting the freedom of the $k$-th regressor parameter estimate and making it deflated compared to their (non-adaptive) SVS counterparts. On the other hand, a small value of $\pi_k$ gives more freedom to the estimates for the $k$-th regressor thus allowing them to inflate, while $\pi_k = 0$ leads to no penalization for the $k$-th regressor.

As we will show, applying this simple modification to the form of the penalty can significantly improve the performance of the SVS method. In order to achieve this, however, the penalization weights $\pi_k$ in (8.1) must be suitably selected. One way to do this is to employ an initial data-based estimator of $B$ and use it to define $\pi_k$ and let the method adapt to the initial estimate of $B$, hence the name adaptive SVS. If the initial estimator is estimation consistent, the adaptation should result in an efficiently performing method. In fact, with $\alpha = 1$ the adaptive SVS turns into the adaptive lasso and the consistency and oracle properties derived by Huang et al. (2008) immediately translate to adaptive SVS with $\alpha = 1$. Moreover, for $\alpha = 2$ the adaptive SVS turns into the adaptive group lasso within the stacked-up model (7.2), thus the consistency and oracle properties derived by Wei and Huang (2010) and Wei et al. (2011) translate to the adaptive SVS with $\alpha = 2$. Consequently, it appears reasonable to expect that other adaptive variants of the adaptive SVS might also perform well under similar conditions.
Novelty of the method

As far as the question of novelty of the adaptive SVS is concerned, clearly, the shape of the objective function is not completely new, since special choices of \( \alpha \) turn it into well-established methods. However, as far as we are aware, the choice of \( \alpha = \infty \) together with the adaptation is novel, and so is the use of both \( L_2 \)-aSVS and \( L_\infty \)-aSVS in the multivariate multiple regression with correlated responses setting. The fact that the adaptive SVS overlaps in special cases with other methods is actually a benefit that allows utilization of existing efficient numerical solvers such as the R package \texttt{glmnet} (see Section 8.3.2).

8.2 Determining weights in adaptive methods

For a good performance by an adaptive method the weights must be chosen suitably, which can be based on a reasonable initial estimate for \( B \). For the adaptive (group) lasso the usual choice for such initial estimate is determined using the (group) lasso. We will focus on two different approaches to obtaining the initial estimate. Primarily, we investigate the behaviour of the adaptive SVS using weights based on the OLS estimates within a univariate regression model for each component of the phenotype. As an alternative, which we utilize in only a part of our investigation, we obtain the initial estimate for \( B \) using the lasso.

Univariate OLS based weights

The univariate OLS approach is similar to using the naive OLS, although instead of transforming the multivariate phenotype into a univariate response, we regress each component of the phenotype on each regressor separately and apply a chosen function to obtain a single weighting factor for each regressor. This effectively boils down to using the Pearson correlation coefficient of each regressor with each component of the phenotype. Such approach is quite advantageous because of its computational simplicity, which contributes to fast analysis. As we illustrate in this chapter, the resulting adaptive methods can perform quite well in the context of genotype-phenotype data. In formulas, we compute

\[
\hat{B}_{\text{ols}} = \left( b_{\text{ols}}^{kl} \right)_{k=1}^{q} \text{ according to }
\]

\[
b_{\text{ols}}^{kl} = \arg\min_{b \in \mathbb{R}} \| Y_{\bullet l} - b X_{\bullet k} \|_{2}^{2}, \quad k = 1, \ldots, q, \quad l = 1, \ldots, p,
\]

and use it as the initial estimate for \( B \). Then, since weighing in the adaptive SVS of (8.1) is done per regressor, for each \( k \) we need to transform \( b^{\text{ols}}_{kl} \), \( l = 1, \ldots, p \) into a single value \( \pi_k \). To that end, we suggest to first apply a suitable function \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) such as the mean or the maximum to \( b^{\text{ols}}_k = (b^{\text{ols}}_{k1}, \ldots, b^{\text{ols}}_{kp}) \). Since it is desirable to have large penalization weights for regressors that are weakly correlated with the responses and small weights for the regressors with high correlation with the responses, the straightforward option is to use the inverse transform and put \( \pi_k = 1/f(b^{\text{ols}}_k) \). Besides the mean and the maximum other choices for \( f \) might be suitable. Generally, for any fixed \( f \) the influence of the initial estimates on the adaptive SVS estimator can be strengthened or weakened by adding for instance a polynomial
transformation on top of $f$.

An advantage of using the univariate OLS for determining the adaptation weights $\pi_k$ is the straightforwardness of such approach. It relies on linear correlation between regressors and responses, which also makes it simple and fast to implement. However, since the univariate correlation is used only in a limited way compared to the naive OLS, the criticisms of the former do not directly apply here. Furthermore, after a fixed $f$ is chosen, there are no more tuning parameters involved in determining the adaptation weights. This increases the appeal of the univariate OLS approach in comparison with for instance using the non-adaptive SVS or the lasso as the basis for determining the weights, which would require additional analysis to determine their penalty parameters, thereby adding to the already relatively heavy computational burden for large data sets.

The idea to base the weights on the univariate OLS is similar to that of Huang et al. (2008), who investigate it under certain conditions including partial orthogonality of the design matrix and show the consistency of the univariate estimates for estimating $B$. Partial orthogonality in this case requires that the covariates with zero coefficients and those with non-zero coefficients are sufficiently weakly correlated (see Huang et al. (2008)). In our application, we must keep in mind that using univariate regression as a proxy for a multivariate model can be tricky, unless the design matrix is (near) orthogonal, since with highly correlated regressors many estimates might end up large while in reality only a handful of the corresponding parameters are in fact non-negligible. However, the univariate estimates are used only as basis for the subsequent estimation by the adaptive SVS, which limits the severity of this caveat. Additionally, as we illustrate in Section 8.3, when the design matrix is based on real genotypes, the typical degree of correlation observed in such data does not severely hinder the univariate approach. Consequently, it should come as no surprise that, despite its simplicity, for an association analysis of such data the adaptive SVS procedure with univariate OLS based weights can yield substantially improved performance over both the (non-adaptive) SVS, the GFlasso methods and the lasso.

**Lasso based weights**

The lasso based weights are obtained by replacing $\hat{B}_{\text{ols}}$ with the lasso estimator $\hat{B}_{\text{lasso}}(\lambda) = (b_{kl}^{\text{lasso}}(\lambda))$ as the initial estimate in $B$. However, since the lasso contains the penalty parameter $\lambda$, which creates additional computational burden, we only used the lasso based weights in the adaptive lasso method of (7.8), while sticking with the univariate OLS based weights for the adaptive SVS. With $\hat{B}_{\text{lasso}}(\lambda)$ we followed the same method of transforming the initial estimates into the weighing factors by taking a suitable transformation $f : \mathbb{R}^p \to \mathbb{R}$, such as the average, here applied to $b_{kl}^{\text{lasso}}(\lambda)$, which we subsequently invert to obtain $\pi_k$. For those $k$, where $\pi_k$ would not be well defined due to $\sum_l b_{kl}^{\text{lasso}}$ being zero one can put $\pi_k$ equal to $1$.

---

1 In effect, adding the power transformation is equivalent to taking a different $f$. For notation convenience we keep the two transformations separate, however, since later we choose $f$ to be the mean, and we apply various power transformations to it.
chosen large value, thus shrinking the corresponding adaptive lasso estimates close to zero. Applying $f$ to obtain a single weighing term for parameters corresponding to the same locus was motivated by the common association assumption (Section 7.1.1). Naturally, if this assumption is not applicable in a given context, the transformation by $f$ can be skipped and each element of $B$ can be weighted separately by the adaptive lasso of (7.8) using weights $w_{kl} = 1/b_{kl}^{\text{lasso}}(\lambda)$ whenever $b_{kl}^{\text{lasso}}(\lambda) \neq 0$ and $w_{kl}$ equal to a chosen large value otherwise.

### 8.3 Simulation study

Next we investigate the performance of the adaptive SVS method on data sets generated from real genetic data. We use the multivariate multiple regression model to generate correlated multivariate phenotypes and we deploy the adaptive SVS and several other methods and compare their relative performances using various measures.

#### 8.3.1 Used estimation methods

In the analysis we considered both versions of the adaptive SVS method, namely $L_\infty$-aSVS and $L_2$-aSVS, each in several different variants based on different choices for the adaptation weights. As performance benchmarks for the adaptive SVS we use the non-adaptive SVS of (7.13), the naive OLS of (7.5), the lasso of (7.6), the adaptive lasso of (7.8), and the GFlasso of (7.12). In order to negate the need for an intercept in the models we centered (per-column) all of the phenotype and genotype matrices.

Regarding the GFlasso we used several values for the cut-off parameter $\rho$, namely 0.05, 0.1, 0.2, 0.3. Based on the maximum observed correlations within the data larger values for $\rho$ would be redundant. For the weighing function inside the fusion penalty we selected $w(r) = |r|$. In the adaptive SVS methods we used the univariate OLS based weights obtained with $f$ equal to the mean and applied the power transformations to them via

$$
\pi_k^\nu = d_k (\sum_l b_{kl}^{\text{ols}}/p)^{-\nu},
$$

where $\nu = 1,2,0.5$ with $d_k$ chosen so that $\sum_k \pi_k = q$. We respectively denote the resulting methods corresponding to each value of $\nu$ as $L_2$-aSVS(1), $L_2$-aSVS(2), and $L_2$-aSVS(0.5) for $\alpha = 2$, and $L_\infty$-aSVS(1), $L_\infty$-aSVS(2), and $L_\infty$-aSVS(0.5) for $\alpha = \infty$.

In the comparison of methods we also include the lasso and the adaptive lasso methods. We used four different variants of the adaptive lasso of (7.8), which differed by the choices of weights $w_{kl}$. The first three variants utilized the same univariate OLS approach as the adaptive SVS methods, which was achieved by putting $w_{kl} = \pi_k^\nu$ for all $l = 1, \ldots, p$ with the same choices for $\nu = 1,2,0.5$. The corresponding adaptive lasso methods are referred to as alasso(1), alasso(2), and alasso(0.5). The fourth variant of the adaptive lasso utilized weights based on the lasso with $\lambda$ determined through cross-validation (see Section 8.4.1). We refer
to the resulting method as alasso(lasso).

### 8.3.2 Software

The GFLasso estimates were computed using a MATLAB (www.matlab.com) script, which was graciously provided to us by Seyoung Kim, one of the authors of the method. Estimates of regression parameters by the $\ell_2$ norm based adaptive and non-adaptive SVS methods were obtained using the R package glmnet (R Core Team (2015)), while the $\ell_\infty$ norm based estimates were determined using the large-scale optimization software MOSEK\(^2\) and the R interface package Rmosek available from CRAN. Finally, the lasso and the adaptive lasso estimates were calculated using glmnet.

### 8.3.3 Genotype data

The genotypic data set used as basis for the simulation contained a total 3000 SNPs for 2000 individuals. The genotypes for each locus were numerically represented as the number of minor alleles at that locus, i.e., each genotype was a sequence of 2000 values from the set \{0, 1, 2\}. Figure 8.1 shows the LD heat plot for the full data set.

We used 8 of the available 3000 SNPs (chosen randomly) to simulate additional 8 simulated SNP genotypes as their empirical neighbors. The SNPs were simulated in a way that resulted in the desired amount of LD between the simulated SNP and its neighbor, which was either 0.2 or 0.75. The MAFs of the simulated SNPs were either 0.1 or 0.5, while all of the neighboring SNPs had MAFs of at least 0.2. We merged the simulated SNP genotypes with the real life SNP genotypes to obtain a realistic data set, which would serve as the design matrix in the simulation. Repeating this process of selecting SNPs and simulating genotypes as their neighbors 25 times we obtained a total of 25 design matrices. For the sake of notational simplicity, we numbered the simulated SNPs as 3001,...,3008 and put them in the last 8 of the total of 3008 columns of each design matrix irrespective of their "actual" locations (determined by the locations of the empirical neighbor SNPs) in the genome. Figure 8.2 shows the observed correlations between the simulated SNPs and the rest of the full genotype data set for a randomly chosen (out of the 25 sets) collection of 8 simulated SNPs.

### 8.3.4 Phenotype data

Starting with $X = (x_{ij})$ we repeatedly simulated $p$-dimensional phenotypes for each of the $n$ individuals using the linear regression model (7.1) with $q \times p$ regression matrix $B$. For the simulation of phenotypes we considered several different scenarios (settings of parameters) in order to investigate different aspects of the performances of the considered methods. For given choices of $n$ and $p$ we start by generating a factor variable for each of the $n$ individuals.

\(^2\)MOSEK is a commercial high performance software for large-scale optimization. A free academic licence can be obtained at http://www.mosek.com.
Figure 8.1: Heat map of LD patterns as measured by squared correlation coefficient within the real life genetic data set with $q_0 = 3000$ SNPs and $n = 2000$ individuals used as a basis for the simulation study in this chapter. SNPs are numbered $1, 2, \ldots, 3000$ from left to right and the empirical SNPs used as neighbours in simulation are selected from SNPs $2001, \ldots, 3000$.

Figure 8.2: Typical observed sample correlation coefficients between a set of 8 simulated SNPs and the rest of the real life 3000 SNPs. Positions of the simulated SNPs (next to their empirical neighbor SNPs) are denoted by vertical lines, where the heights of these lines express the observed maximum and minimum correlations.

Through a univariate-response multivariate linear regression model where only the simulated SNP genotypes have non-zero coefficients, henceforth referred to as the causal SNPs. Denoting by $Q$ the set of all SNPs (i.e. columns of $X = (x_{ij})$), we simulated a vector of $n$ factor variables $F = (F_1, \ldots, F_n)'$ according to

$$F_i = \sum_{k \in Q} \alpha_k x_{ik} + e_i, \quad i = 1, \ldots, n,$$

(8.4)

where $\alpha_k, k \in Q$ were selected in several different ways specified below (see scenarios A, B, C) and $e_i$ are independent zero-mean normally distributed errors with fixed variance $0.2$.

The observed factor variables $F_1, \ldots, F_n$ were entered into $p$ univariate linear regression models with normally distributed independent random errors and regression coefficients ranging over predefined set of values, which produced an $n \times p$ matrix of row-wise independent and column-wise correlated responses $Y = (Y_{ij})$. Written in formula the model is

$$Y_{ij} = \gamma_j F_i + e_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p.$$

(8.5)
Note that (8.5) implies

\[ E Y_{ij} = \sum_{k \in Q} \beta_{kj} x_{ik}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p, \]

where \( \beta_{kj} = \alpha_k \gamma_j \). The idea in (8.4) is to set only a small fraction of values of \( \alpha_k \), \( k \in Q \) non-zero. Respectively denote the subsets of zero and non-zero coefficients by \( Q_0 \) and \( Q_1 \) and their sizes by \( q_0 \) and \( q_1 \). In the three scenarios below, we have \( q = 3008 \) and \( q_1 = 8 \) (scenario A) and \( q = 1004 \) and \( q_1 = 4 \) (scenarios B and C) with only the simulated SNPs in \( Q_1 \). Summary statistics for each scenario can be found in Table 8.1 and the details of the simulation settings are described below.

Simulating data according to the scheme described by (8.4) and (8.5) is designed to induce dependence between the components of the multivariate phenotypes in a way that adheres to the assumptions of sparsity and common association of Section 7.1.1. We paid close attention to simulating the data in a realistic way using real life genotypes, although we admit that the setup is still somewhat idealized. The phenotypes in \( Y \) were simulated using the same model that is assumed by the estimation methods, while the amount of noise was relatively low, making \( Y \) very much determined by the genetic regressors in \( X \).

**Simulation scenario A**

Under scenario A we took the available 25 genotype data sets with \( n = 2000 \) individuals and \( q = 3008 \) SNPs (both observed and simulated) and used the simulated \( q_1 = 8 \) SNPs as causal ones. Causality here means that precisely the values of \( \alpha_{3001}, \ldots, \alpha_{3008} \) in the factor model (8.4) were non-zero, where we used \( \alpha_k = 0.1 \) for \( k = 3001, \ldots, 3004 \), and \( \alpha_k = 0.15 \) for \( k = 3005, \ldots, 3008 \). The loading parameters \( \gamma_j, j = 1, \ldots, 5 \) ranged over \( 0.1, 0.2, 0.3, 0.4, 0.5 \), i.e. \( \gamma_j = j/10 \). This resulted in the total of 25 data sets generated under scenario A.

Under scenario A we focussed on evaluating the dependence of performances of the considered penalized methods on the choices of penalty parameters. For each of the 25 data sets we determined the estimates by each of the considered methods for a number of different penalty choices and calculated various performance measures for each method. The primary performance measure in such investigation is the squared expectation prediction error (see below), which uses the true expectation of the responses to measure how well a method performs. This allows us to determine the optimal value of penalty parameter for each of the considered penalized methods within each data set. By looking at the average optimal values we can illustrate the potential each method has in terms of prediction and how sensitive each method is to the choice of penalty parameters. Among the methods under scenario A we included the adaptive and non-adaptive SVS, GFlasso, the lasso and the naive OLS. Under scenarios B and C we also investigate the behavior of the adaptive lasso.

Figure 8.1 uses the squared correlation coefficient as the measure of LD, which illustrates the overall pattern of (pair-wise) dependence among the SNPs, while Figure 8.2 shows the dependence between the simulated SNPs with non-zero effects and the rest of the SNPs. The observed patterns of overall dependence in terms LD are typical for genetic data. We expect
that under scenario A, where all \( \alpha_k, k \in Q_1 \) are positive and all of the loadings \( \gamma_j, j = 1, \ldots, p \) have equal sign, the naive OLS with the transformed response \( Z \) of (7.4) equal to the sum of components of the phenotypes (i.e., \( c = (1, \ldots, 1)' \)) should perform reasonably well particularly for the SNPs with LD 0.2 with its empirical neighbor (i.e \( k = 3001, 3002, 3005, 3006 \)).

**Simulation scenarios B and C**

Additionally, we simulated a number of data sets under two other settings of parameters, which we refer to as scenarios B and C. Under scenarios B and C we allowed the sample sizes to vary, which is intended to yield insight into the dynamics of performance by the compared methods in terms of sample size. Like before, we started with the full genotype data sets with \( n = 2000 \) individuals and 3008 SNPs. Unlike under scenario A, we kept only the 1000 SNPs indexed by 2001, \ldots, 3000 for scenarios B and C. Among the selected SNPs we included all of the empirical SNPs used in simulation. Then, a randomly selected group of 4 simulated SNPs were used as causal in the simulation of phenotypes via the factor model (8.4), where we put \( \alpha_{1001} = 0.05, \alpha_{1002} = 0.1, \alpha_{1003} = 0.15, \alpha_{1004} = 0.2 \) and \( \alpha_k = 0 \) for \( k = 1, \ldots, 1000 \). Additionally putting \( p = 10 \), we used the factor model and the model (8.5) to generate 25 genotype-phenotype data sets under each of the two scenarios B and C.

The way the two scenarios B and C differ is in the choice of the values of the loading parameters \( \gamma_1, \ldots, \gamma_{10} \). Under scenario B we put \( \gamma_j = j/10 \) where \( j = 1, \ldots, 10 \), which means that all of the loadings take non-zero values and all of them are positive, which should favor the use of naive OLS approach with the summed-up phenotypes in \( Z \). On the other hand, under scenario C we used different loadings, namely \( \gamma_j = j/10 - 0.5 \) for \( j = 1, \ldots, 10 \), in which the first 4 components have negative values of the loadings and the last 5 components have positive values for the loadings. Such setup is supposed to emulate a situation in which the type of relationship between the genotypes and the considered phenotypes is not favourable for the naive OLS with the summed-up phenotypes because the relationship between the individual components of phenotypes and genotypes does not translate into \( Z \). It also allowed us to investigate the performance of the methods under slight deviation from the common association assumption, which is violated due to \( \gamma_5 \) being zero under scenario C. Under scenarios B and C we included all of the methods described in Section 8.3.1 with the exception of the naive OLS without MTC.

**8.3.5 Measures of performance**

In order to judge the performances of each method we calculated and plotted several measures, which focus on the quality of an estimator with respect to three different perspectives, namely the response expectation prediction, the correct model selection, and the parameter estimation error. Each of these aspects is addressed by one or more of the following quality measures.


8.3 Simulation study

<table>
<thead>
<tr>
<th>scenario</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>phenotypes</td>
<td>( p = 5 )</td>
<td>( p = 10 )</td>
<td>( p = 10 )</td>
</tr>
<tr>
<td>sample size</td>
<td>( n = 2000 )</td>
<td>( n = 100, 200, \ldots, 2000 )</td>
<td>( n = 100, 200, \ldots, 2000 )</td>
</tr>
<tr>
<td>number of SNPs</td>
<td>( q = 3008 )</td>
<td>( q = 1004 )</td>
<td>( q = 1004 )</td>
</tr>
<tr>
<td>non-causal SNPs</td>
<td>( q_0 = 3000 )</td>
<td>( q_0 = 1000 )</td>
<td>( q_0 = 1000 )</td>
</tr>
<tr>
<td>causal SNPs</td>
<td>( q_1 = 8 )</td>
<td>( q_1 = 4 )</td>
<td>( q_1 = 4 )</td>
</tr>
<tr>
<td>( \alpha_k )</td>
<td>( \alpha_k = 0 ) for ( k = 1, \ldots, 3000 )</td>
<td>( \alpha_k = 0 ) for ( k = 1, \ldots, 1000 )</td>
<td>( \alpha_k = 0 ) for ( k = 1, \ldots, 1000 )</td>
</tr>
<tr>
<td></td>
<td>( \alpha_k = 0.1 ) for ( k = 3001, \ldots, 3004 )</td>
<td>( \alpha_k = 0.05 ) for ( k = 1001 )</td>
<td>( \alpha_k = 0.1 ) for ( k = 1002 )</td>
</tr>
<tr>
<td></td>
<td>( \alpha_k = 0.15 ) for ( k = 3005, \ldots, 3008 )</td>
<td>( \alpha_k = 0.15 ) for ( k = 1003 )</td>
<td>( \alpha_k = 0.2 ) for ( k = 1004 )</td>
</tr>
<tr>
<td>( \gamma_j )</td>
<td>( \gamma_j = j/10 ) for ( j = 1, \ldots, 5 )</td>
<td>( \gamma_j = j/10 ) for ( j = 1, \ldots, 10 )</td>
<td>( \gamma_j = j/10 - 0.5 ) for ( j = 1, \ldots, 10 )</td>
</tr>
<tr>
<td>error variance</td>
<td>( \sigma^2 = 0.2 )</td>
<td>( \sigma^2 = 0.2 )</td>
<td>( \sigma^2 = 0.2 )</td>
</tr>
</tbody>
</table>

Table 8.1: Summary of settings used during phenotype simulation via (8.4) and (8.5) under scenarios A, B and C.

Selection rates (FSR and TSR)

Our first measure of quality of an estimate is the fraction of non-zero values among estimates of true zero coefficients, also known as the false selection rate (FSR). Under the factor model used in simulation, keeping in mind that each \( b_{kj} \) is an estimate for \( \beta_{kj} = \alpha_k \gamma_j \), the rate of falsely selected SNPs is defined as

\[
\text{FSR} = (pq_0)^{-1} \sum_{k,j: \hat{\beta}_{kj} \neq 0} I(b_{kj} \neq 0) = (pq_0)^{-1} \sum_{k \in Q_0} \sum_{j=1}^p I(b_{kj} \neq 0).
\]

A complementary measure is the ratio of non-zero values among estimates of the true non-zero coefficients, also known as the true selection rate (TSR). For the current regression model it is defined as

\[
\text{TSR} = (pq_1)^{-1} \sum_{k,j: \hat{\beta}_{kj} \neq 0} I(b_{kj} \neq 0) = (pq_1)^{-1} \sum_{k \in Q_1} \sum_{j=1}^p I(b_{kj} \neq 0).
\]

When evaluating the performance in terms FSR and TSR, one must keep in mind that there exists a natural trade-off between FSR and TSR performances, where a method that fails to identify any structure in the data might exhibit a near perfect performance in terms of FSR, while failing miserably in terms of TSR, and vice-versa.

Expectation prediction error (SEPE)

The overall quality of an estimate \( \hat{B} = (\hat{b}_{ij}) \) can be measured by the quality of the predictions it yields. In a simulation where we know the true values of the regression parameters the quality of the prediction of the response’s true expectation by an estimator \( \hat{B} \) can be measured. Using \( \hat{B} \), the expectation of all \( n \times p \) observed components of phenotypes can be predicted using \( X\hat{B} \). Comparing such prediction with the known true expectation \( XB \) using the summed squared error leads to the squared expectation prediction error \( \text{SEPE} = ||X\hat{B} - XB||^2 \). Since such definition of prediction error is sample size dependent, for comparison of prediction

\[
\text{SEPE} = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (\hat{b}_{ij} - b_{ij})^2
\]

Where \( b_{ij} \) is the true value of the \( ij \)th component of the phenotype vector, \( \hat{b}_{ij} \) is the predicted value using the estimator \( \hat{B} \), and \( n \) is the sample size.
errors over various sample sizes it is useful to scale SEPE by \( n \), which yields the average squared prediction error per individual. We use such rescaled SEPE to compare the performances under the scenarios B and C.

**Estimation errors**

Additionally, we use the total estimation error (TEE), which is the \( \ell_1 \) distance between the estimates and the true values of the regression matrix \( B \). Focusing on all SNPs in \( Q \), the causal SNPs in \( Q_1 \), or the non-causal SNPs in \( Q_0 \) respectively leads to

\[
TEE^\beta_z = \sum_{k \in Q_0} \sum_{j=1}^p |b_{kj}|, \quad TEE^\beta_{nz} = \sum_{k \in Q_1} \sum_{j=1}^p |b_{kj} - \hat{\beta}_{kj}|, \quad TEE^\beta_{sum} = TEE^\beta_z + TEE^\beta_{nz}.
\]

The choice of the \( \ell_1 \) distance for the estimation error is motivated by the sparsity of \( B \) and our desire to have estimation errors for the true zero coefficients contribute meaningfully to the value of the estimation error measure. Had we chosen the \( \ell_2 \) distance, we would have effectively assigned much smaller weights to those types of estimation errors.

Instead of focusing on estimation \( b_{kj} \) directly, we can also choose to scale the estimates \( b_{kj} \) and the true values \( \hat{\beta}_{kj} \) using the known values of \( \gamma_1, \ldots, \gamma_p \) in order to bring the estimation errors for different components of phenotypes on a similar scale. Putting \( a_{kj} = b_{kj}/\gamma_j \) we define analogous quantities to those above, except calculated using \( a_{kj} \) and \( \alpha_{kj} \) instead of \( b_{kj} \) and \( \hat{\beta}_{kj} \), which we denote as \( TEE^a_z, TEE^a_{nz} \) and \( TEE^a_{sum} \).

**8.4 Investigation under scenario A**

We first focus on the performance aspects of the methods under scenario A. We start with the results concerning the choice of penalty parameters and show that cross-validation (CV) is a viable strategy. We then turn to the best-case (in terms of penalty parameters) overall average performances by each method as measured by SEPE, FSR and TSR and compare them with the average overall performances at the cross-validated penalty parameters.

**8.4.1 Determining penalty parameters by cross-validation**

Any estimation procedure that requires a choice or tuning parameters performs only as well as the particular choice of those parameters allows. In our comparison we consider the adaptive SVS, the lasso, the GFlasso. All of these require parameter input in the form of penalty parameters \( \lambda \) and \( \mu \) (only GFlasso). In order to determine suitable values for these penalties we turn to the following \( k \)-fold CV scheme. For a given method and a fixed value of \( k \) we randomly split the available data into \( k \) disjoint portions of (approximately) equal size. In the estimation phase we systematically select \( k - 1 \) portions of the data and denote the corresponding design and response matrices by \( X_e \) and \( Y_e \). Using \( X_e \) and \( Y_e \) we obtain estimates \( \hat{B} \) by the given method for a range of penalty parameter values. For the validation phase we
denote the design and the response matrices corresponding to the remaining \( k \)-th portion of the data by \( X_v \) and \( Y_v \). Using \( \hat{B} \) and \( X_v \), we calculate predictions for \( Y_v \) based on which we calculate a chosen prediction error measure for these predictions. The estimation-validation steps are repeated \( k \) times each time with a different portion of the data used in the validation phase. The value of the penalty parameter that minimizes the average CV criterion function is then returned as the result of the CV. As the CV criterion function we use the square observation prediction error SOPE, which is an \( L_2 \) measure of the difference between the predictions and the observed values of the validation data set, that is

\[
SOPE = \|X_v\hat{B} - Y_v\|^2.
\] (8.6)

Ideally, instead of choosing the penalty that minimizes prediction error of the observed responses \( Y_v \), we would want to minimize the prediction error of the true expectation \( \mathbb{E}Y \). In other words, instead of SOPE we would want to use SEPE as the CV criterion. Unfortunately, in practice this is hardly possible, since \( \mathbb{E}Y \) is unknown, thus SEPE cannot be evaluated. Although using SEPE as the prediction error measure is not an option in real data analysis, it is possible to calculate it for simulated data, where \( B \) is known.

Somewhat unpleasantly, the \( k \)-fold CV itself has a type of tuning parameter in the form of a \( k \), which effectively determines the sample size ratio between the estimation and validation phases of the procedure. Generally, the leave-one-out CV (LOOCV) with \( k = n \) is considered ideal. Unfortunately, for large-scale data the LOOCV is often infeasible and a much smaller value for \( k \) must used. For a fixed \( k \) the sample size fraction used in the estimation phase of the CV scheme is equal to \( (k - 1)/k \). We can re-parameterize the scheme using this sample size ratio, which we denote as \( \tau \). In the CV analysis we investigated the influence of \( \tau \) on the resulting penalties. We considered a range of values for \( \tau \) including those that do not correspond to any particular choice of \( k \). For a sample of size \( n \) the largest sensible value for \( \tau \) is \( (n - 1)/n \), which yields the LOOCV. Although it is expected that the bigger the value of \( \tau \) the better, we also considered values of \( \tau \) smaller than 0.5 in order to get an idea of sensitivity of the estimates of \( B \) to the choice of \( \tau \). This is motivated by the fact that computational speed is inversely linked with \( \tau \), where smaller values of \( \tau \) yield faster procedures due to smaller sample size during the computationally demanding estimation phase and also a smaller number of estimation re-runs.

**Results of cross-validation analysis**

Using the \( N = 25 \) simulated data sets under scenario A we performed CV for each of the selected variants of methods \( L_2 \)-aSVS, \( L_{\infty} \)-aSVS, GFlasso and the lasso for various values of \( \tau \) ranging between 0.2 and 0.9 in steps of 0.1. The largest value of \( \tau = 0.9 \) effectively corresponds to 10-fold CV, while \( \tau = 0.5 \) yields 2-fold CV. The resulting cross-validated penalties averaged over the 25 data sets are presented in Figure 8.3. The plots show the cross-validated penalties as functions of \( \tau \) plotted as a solid black line with circles at the cross-validated penalties, which are denoted as \( \lambda_{cv}^r \) and \( \mu_{cv}^r \). Also shown are the optimal
Figure 8.3: Results of CV analysis with various values of the sample split ratio $\tau$. Cross-validated penalties for several variants of methods $L_2$–aSVS, $L_2$–aSVS, $L_2$–aSVS, $L_2$–aSVS, GFlasso and lasso were averaged over 10 reruns for each of 25 data sets under scenario A. For each method the average observed cross-validated penalty parameters ($\lambda^*_C$ for all methods and $\mu^*_C$ for GFlasso) are plotted against $\tau$. In each plot the dashed horizontal line represents the average SEPE-optimal penalty values $\lambda_{opt}$ and $\mu_{opt}$, which were determined using the known true expectation. The full line represents the average values of cross-validated penalties $\lambda_{CV}$ and $\mu_{CV}$, where the average is taken over the value of $\tau$ between 0.3 and 0.8. The cross-validated penalties corresponding to $\tau = 0.5$ (denoted as $\lambda^{0.5}$ and $\mu^{0.5}$) are also marked. The whiskers in each plot represent the range containing 95% of the optimized values of each penalty parameter.
penalties $\lambda_{\text{opt}}$ and $\mu_{\text{opt}}$, which were determined for each of the 25 data sets using $SEPE$. The values $\lambda_{\text{opt}}$ and $\mu_{\text{opt}}$ are averaged over the 25 data sets and represented by red dashed lines in the plots of Figure 8.3. Additionally, the plots also show as solid blue lines the values $\lambda_{\text{mean}}$ and $\mu_{\text{mean}}$, which are the averages of cross-validated penalties taken over those cross-validated penalties that correspond to $\tau$ between 0.3 and 0.8. Finally, the 2-fold cross-validated penalties ($\tau = 0.5$) are denoted by purple filled circles, which we singled out since it is the borderline case between the usual CV scheme parameterized by $k$ and our more generalized scheme parameterized by $\tau$.

As we can see in Figure 8.3, for the (adaptive) SVS methods CV provides a fairly precise and stable determination of the optimal input parameters for a wide range of sample split ratios $\tau$. For all eight variants of the two methods this is evidenced by the almost horizontal black lines in the plots and their closeness to the optimal value indicated by a red dashed line, which was determined as the point of minimum prediction error in Figure 8.4. It seems that in the idealized simulation setting the ratio between the number of estimated parameters and the sample size available for estimation can be quite large without harming the estimation procedures in terms of $SEPE$. As far as the variants of GFLasso and the lasso are concerned, it appears that CV works quite well for determining the penalty parameters $\lambda$ of these procedures, perhaps with a slight caution for GFLasso ($\rho = 0.05$).

On the other hand, the CV procedure seems to provide somewhat less favourable results for the fusion penalty parameter $\mu$, where the CV averages appear to be both relatively imprecise and less stable compared to the results for $\lambda$. This suggests caution when using methods that utilize multiple penalty parameters, which might be more sensitive to the size of samples used in estimation. Nonetheless, as shown in Figure 8.4, this apparent instability especially in $\mu$ does not result in dramatically different values of $SEPE$ at the cross-validated values of $\lambda$ and $\mu$ compared to the optimal values of the penalties in terms of $SEPE$. Furthermore, similar conclusion can be made also for the observed $TSR$ at the cross-validated values of penalty parameters. The measure that appears to be influenced the most by the CV is $FSR$ especially for $\rho = 0.05$, where the slope of $FSR$ as a function of $\mu$ is the steepest.

Overall, our analysis provides evidence that CV is a reliable procedure for determining suitable penalty values under scenarios similar to A. The obtained penalties generally do not seem to differ substantially from the optimal values in terms of prediction measure $SEPE$ for a wide range of $\tau$. Nonetheless, for practical use of the (adaptive) SVS methods during CV it is strongly recommended to use as large values for $\tau$ as computationally possible. Moreover, the results suggest that CV works slightly better for the (adaptive) SVS than the GFLasso.

### 8.4.2 Overall performance under scenario A

Figure 8.4 and Table 8.2 illustrate the performance of the estimation methods under scenario A. Analyzing the plots in the figure, we first focus on the performance of the adaptive variants of $L_2$–aSVS and $L_{\infty}$–aSVS relative to their non-adaptive counterparts $L_2$–SVS and
8 Adaptive simultaneous variable selection 197

Figure 8.4: Comparison of performance measures SEPE, FSR and TSR and average cross-validated penalties under scenario A for methods $L_2$–SVS, $L_2$–aSVS(1), $L_2$–aSVS(2), $L_2$–aSVS(0.5) (left column), $L_\infty$–SVS, $L_\infty$–aSVS(1), $L_\infty$–aSVS(2), $L_\infty$–aSVS(0.5) (middle column), four variants of GFlasso and the lasso (right column). The plots are based on 25 data sets. In the top row plots the observed values of SEPE are plotted as functions of penalty parameters $\lambda$ and $\mu$ for the considered methods (with horizontal axes on a logarithmic scale for the (adaptive) SVS methods). For each method the penalties $\lambda_{\text{min}}$ and $\mu_{\text{min}}$ corresponding to the minimum SEPE are denoted by bold lines in the color of each method. Additionally, the values of SEPE at the average cross-validated penalty parameters $\lambda_{\text{cv}}$ and $\mu_{\text{cv}}$ are denoted by thin vertical lines in each color. The plots for the (adaptive) SVS methods also show denoted by the grey areas the empirical 95% confidence intervals for the cross-validated penalties, which are based on 10 for each data set. Respectively, the middle and bottom row plots illustrate FSR and TSR performance for variants of each method as functions of penalty parameters with the observed values of FSR and TSR at the SEPE-optimal penalties $\lambda_{\text{min}}$ and $\mu_{\text{min}}$ (again denoted by bold vertical lines in each color) and FSR and TSR at the cross-validated penalties $\lambda_{\text{cv}}$ and $\mu_{\text{cv}}$ (denoted by thin vertical lines in each color). For the GFlasso only integer values $\lambda$ were considered, which remain fixed between the ticks, while $\mu$ varies over $(0, 200)$ between the ticks.

Then, we compare these methods with the GFlasso variants and the lasso. For each method, we considered a range of penalty values for $\lambda$ and $\mu$ in the GFlasso, where the considered ranges were similar to those seen on the $y$-axes in Figure 8.3. For each method and each combination of penalty parameter(s) we determined the corresponding estimate of $B$ in each data set under scenario A. We combined the obtained results over the 25 data sets to calculate the average values of SEPE, FSR and TSR. Plotting these values against the $L_\infty$–SVS.
198

8.4 Investigation under scenario A

<table>
<thead>
<tr>
<th>Method</th>
<th>(\lambda_{\text{min}}(\mu_{\text{min}}))</th>
<th>SEPE</th>
<th>FSR</th>
<th>TSR</th>
<th>(\lambda_{\text{cv}}(\mu_{\text{cv}}))</th>
<th>SEPE</th>
<th>FSR</th>
<th>TSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_2)–aSVS(2)</td>
<td>205</td>
<td>2.72</td>
<td>0.0033</td>
<td>1.00</td>
<td>190</td>
<td>2.73</td>
<td>0.0038</td>
<td>1.00</td>
</tr>
<tr>
<td>(L_2)–aSVS(2)</td>
<td>350</td>
<td>3.02</td>
<td>0.028</td>
<td>1.00</td>
<td>320</td>
<td>3.03</td>
<td>0.028</td>
<td>1.00</td>
</tr>
<tr>
<td>(L_2)–aSVS(1)</td>
<td>43</td>
<td>3.00</td>
<td>0.0054</td>
<td>1.00</td>
<td>34</td>
<td>3.26</td>
<td>0.011</td>
<td>1.00</td>
</tr>
<tr>
<td>(L_2)–aSVS(1)</td>
<td>75</td>
<td>3.64</td>
<td>0.022</td>
<td>1.00</td>
<td>55</td>
<td>4.12</td>
<td>0.039</td>
<td>1.00</td>
</tr>
<tr>
<td>(L_2)–aSVS(0.5)</td>
<td>25</td>
<td>3.99</td>
<td>0.010</td>
<td>1.00</td>
<td>20</td>
<td>4.47</td>
<td>0.021</td>
<td>1.00</td>
</tr>
<tr>
<td>(L_2)–aSVS(0.5)</td>
<td>45</td>
<td>5.01</td>
<td>0.025</td>
<td>1.00</td>
<td>35</td>
<td>5.50</td>
<td>0.046</td>
<td>1.00</td>
</tr>
<tr>
<td>(L_2)–SVS</td>
<td>17</td>
<td>7.58</td>
<td>0.031</td>
<td>0.97</td>
<td>17</td>
<td>7.58</td>
<td>0.031</td>
<td>0.97</td>
</tr>
<tr>
<td>GFlasso((\rho = 0.2))</td>
<td>8 (150)</td>
<td>9.57</td>
<td>0.092</td>
<td>0.81</td>
<td>8 (65)</td>
<td>9.70</td>
<td>0.073</td>
<td>0.80</td>
</tr>
<tr>
<td>GFlasso((\rho = 0.3))</td>
<td>9 (100)</td>
<td>8.92</td>
<td>0.058</td>
<td>0.98</td>
<td>33</td>
<td>8.97</td>
<td>0.047</td>
<td>0.97</td>
</tr>
<tr>
<td>GFlasso((\rho = 0.1))</td>
<td>7 (60)</td>
<td>9.80</td>
<td>0.045</td>
<td>0.75</td>
<td>11 (40)</td>
<td>10.43</td>
<td>0.024</td>
<td>0.71</td>
</tr>
<tr>
<td>GFlasso((\rho = 0.05))</td>
<td>6 (50)</td>
<td>10.00</td>
<td>0.130</td>
<td>0.89</td>
<td>6 (50)</td>
<td>10.24</td>
<td>0.150</td>
<td>0.90</td>
</tr>
<tr>
<td>lasso</td>
<td>10</td>
<td>11.41</td>
<td>0.019</td>
<td>0.70</td>
<td>11</td>
<td>11.47</td>
<td>0.015</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 8.2: Average observed SEPE, FSR, TSR at SEPE-optimal (left) and cross-validated penalties (right) under scenario A. The averages are taken over the 25 data sets and the rows are ordered by the minimum SEPE with the best values in each column marked by bold italic.

considered penalty values resulted in the plots in Figure 8.4.

For the comparison of the methods we consider SEPE to be the "baseline" measure. This means that we first judge the methods based on how low their minimum observed values of SEPE are. We also compare the methods in terms of the selection rates FSR and TSR at the SEPE-optimal penalty values \(\lambda_{\text{min}}\) and \(\mu_{\text{min}}\) and at the observed cross-validated penalty values \(\lambda_{\text{cv}}\) and \(\mu_{\text{cv}}\). Comparing the methods in terms of the selection rates at parameter values determined by SEPE is sensible at least for two reasons. First, in application the penalty parameters are usually determined via cross-validation by minimizing a prediction error measure such as SOPE of (8.6), which makes it interesting to see what the selection rates are at these points. The second reason has to do with the fact that for any of our penalized methods it is always possible to make FSR equal to zero by taking a sufficiently high value for the penalty parameter, which means that judging the methods based on whether they can achieve zero FSR is pointless. Analogously, for any penalized method based on minimizing the sum of squared residuals it is also possible to make TSR equal to one by putting the penalty parameter equal to zero. Therefore, the ability to achieve maximum TSR does not yield a reasonable comparison either.

Performance of \(L_2\)–aSVS methods

In the left-most column of plots in Figure 8.4 and in Table 8.2 we see that, when compared to the non-adapted \(L_2\)–SVS, all variants of \(L_2\)–aSVS provide noticeably better performances in terms of both minimum \(L_2\)–SVS (top plot) and FSR (middle plot) and TSR (bottom plot) at the SEPE-optimal penalties. The improvement over the non-adaptive SVS is most visible for \(L_2\)–aSVS(2), while \(L_2\)–aSVS(1) is a very close second. The observed values of minimum prediction error are \(\text{SEPE}_{\text{min}} = 2.72\) at \(\lambda_{\text{min}} = 205\) for \(L_2\)–aSVS(2), \(\text{SEPE}_{\text{min}} = 3.00\) at \(\lambda_{\text{min}} = 43\) for \(L_2\)–aSVS(1), \(\text{SEPE}_{\text{min}} = 3.99\) at \(\lambda_{\text{min}} = 25\) for \(L_2\)–aSVS(0.5), and
$SEPE_{\min} = 7.58$ at $\lambda_{\min} = 17$ for $L_2$–SVS. Respectively, these values represent reductions of approximately 64%, 60% and 47% by the $\ell_2$ based adaptive SVS methods over the non-adaptive $L_2$–SVS method. In order to provide a scale for the observed values of $SEPE$ we point out that the all-zero estimate ($\hat{B}_0 = 0$) yields $SEPE = 49.97$.\(^3\)

Besides good prediction performance, the adaptive SVS variants yield similarly flattering results in terms of FSR at the $SEPE$-optimal value $\lambda_{\min}$. The observed minimum false selection rates in Figure 8.4 are $FSR_{\min} = 0.33\%$ for $L_2$–aSVS(2), and $FSR_{\min} = 0.57\%$ for $L_2$–aSVS(1), and $FSR_{\min} = 0.98\%$ for $L_2$–aSVS(0.5), while the corresponding (non-adaptive) SVS method yields a much less favourable value of $FSR_{\min} = 2.8\%$. This represents at minimum an almost three-fold improvement by the adaptive SVS over the (non-adaptive) SVS, while the best performing adaptive SVS method $L_2$–aSVS(2) provides as much as ten-fold improvement of the false discovery rate at $\lambda_{\min}$.

In addition, the comparison of the true selection rates of $L_2$–aSVS(2), $L_2$–aSVS(1), $L_2$–aSVS(0.5) against $L_2$–SVS in the left-most plot in the third row of Figure 8.4 also shows appreciable improvements by the adaptive SVS methods over the (non-adaptive) SVS at the $\lambda_{\min}$. On average, all three variants of the adaptive SVS successfully identified 100% of the 8 causal SNPs for all 5 phenotypes in all data set, while the (non-adaptive) SVS managed to do so in "only" about 97% of the cases at $\lambda_{\min}$.

**Performance of $L_{\infty}$–aSVS methods**

Turning to methods $L_{\infty}$–SVS and $L_{\infty}$–aSVS, we focus on the middle column of Figure 8.4 and Table 8.2, where we observe a similar rate of improvement in terms of $SEPE$, $FSR$ and $TSR$ at the $SEPE$-optimal $\lambda_{\min}$ by the adaptive SVS methods over the non-adaptive $L_{\infty}$–SVS (relative to the $\ell_2$ methods). The top plot shows that $L_{\infty}$–aSVS(2) yields $SEPE_{\min} = 3.02$ at $\lambda_{\min} = 350$, $L_{\infty}$–aSVS(1) yields $SEPE_{\min} = 3.64$ at $\lambda_{\min} = 75$, $L_{\infty}$–aSVS(0.5) yields $SEPE_{\min} = 5.01$ at $\lambda_{\min} = 45$, and $L_{\infty}$–aSVS yields $SEPE_{\min} = 8.92$ at $\lambda_{\min} = 30$. Similarly to the $\ell_2$ methods, the $\ell_{\infty}$ adaptive SVS methods provide substantial improvement of prediction errors over their non-adaptive counterpart.

A differentiation between the methods is provided by the observed values of $FSR$ at $\lambda_{\min}$ and $\lambda_{cv}$. Generally speaking, the $\ell_{\infty}$ methods show noticeably higher $FSR$ both at $\lambda_{\min}$ and $\lambda_{cv}$. At $\lambda_{\min}$ the observed values were $FSR_{\min} = 2.8\%$ for $L_{\infty}$–aSVS(2), $FSR_{\min} = 2.2\%$ for $L_{\infty}$–aSVS(1), and $FSR_{\min} = 2.5\%$ for $L_{\infty}$–aSVS(0.5), while the corresponding (non-adaptive) SVS method yields $FSR_{\min} = 5.8\%$. The results at $\lambda_{cv}$ were quite similar as can be seen in Table 8.2. In terms of $TSR$ the $\ell_{\infty}$ methods exhibit as good a performance as their $\ell_2$ counterparts, when they successfully identify all non-zero parameters in all data sets. The non-adaptive SVS methods appear to be inferior to the adaptive SVS in this respect.

Overall, a comparison between the first two columns of Figure 8.4 suggests that the overall better performers in terms of the considered measures are the $\ell_2$ adaptive SVS methods.

\(^3\)For the all-zero estimate $SEPE$ equals the squared norm of the true expectation $||XB||^2$, where $B$ is the true parameter matrix.
Among the $\ell_2$ methods it appears to be $L_2$-aSVS(2) and $L_2$-aSVS(1) that have the upper hand over the other $\ell_2$ based methods. Moreover, the clear dominance of the adaptive SVS methods over their non-adaptive counterpart makes a persuasive case in favor of adaptation.

**Comparison with GFlasso and lasso**

Next, we focus on the performances of the GFlasso and the lasso relative to the adaptive SVS methods. For that we turn to the right-most column in Figure 8.4, which illustrates the relative performance of the lasso and variants of GFlasso with different values of the correlation cut-off parameter $\rho$. The plots again show the observed values of $SEPE$, $FSR$ and $TSR$ as functions of the penalty parameters $\lambda$ and $\mu$. In order to show the simultaneous dependence on two parameters within a single plot for each measure, we limit the consideration to integer values for $\lambda$ between 1 and 20 each accompanied by an appropriate range of values for $\mu$, which explains the somewhat unique look of the plots in the right-most column of the figure.

From the top plot showing the observed values of $SEPE$ it is clear that all of the methods in that plot provide more or less similar performances at or near the point of their respective optimal penalty combination. The minimum observed values of $SEPE$ are between 9.57 and 10.32 for the GFlasso, while for the lasso we observed $SEPE_{\min} = 11.41$. Among the considered variants of the GFlasso it appears to be GFlasso($\rho = 0.2$) with $SEPE_{\min} = 9.57$ that yields the best prediction performance. An interesting even if largely expected result is that the classical lasso with $SEPE_{\min} = 11.41$ trails behind all of the variants of GFlasso. This is easily explainable by the fact the GFlasso better exploits the correlation structure within the response matrix through the fusion penalty, which the lasso does not do. On the other hand, a somewhat unpleasant aspect of the GFlasso is the presence of the double penalty, which has a direct influence on its performance. It seems that careful choices of the values for the penalty parameters must be made in order to achieve the right balance between the effects of two penalties. Especially if the fusion penalty is allowed to play a strong role in the method, which occurs when $\rho$ is small, the influence of an ill-chosen value for either of the penalty parameters can be very detrimental in terms of $SEPE$, and, as evidenced by the bottom two plots in the right-most column of Figure 8.4, even more so in terms of the selection rates $FSR$ and $TSR$. As far as the two rates are concerned, we notice that even the lasso, which is the best performing method in terms of $FSR$ and $TSR$ at the $SEPE$-optimal penalties among the methods in the right-most column, comes clearly behind the adaptive SVS methods. While the observed $FSR_{\min} = 1.9\%$ for the lasso is quite competitive, the true selection rate at the $SEPE$-optimal $\lambda$ is only $TSR_{\min} = 70\%$, which does not come even close to the near-100% success rates by the adaptive SVS methods.

Turning to the GFlasso, we notice a worryingly inferior performance of its variants at the $SEPE$-optimal values $\lambda_{\min}$ and $\mu_{\min}$ when compared to the lasso and the non-adaptive SVS methods, but especially relative to the adaptive SVS. There appears to be a fairly strong negative influence of the fusion penalty on the false selection rates in Figure 8.4. Unsurprisingly, it seems that this influence is strongest if $\rho$ is relatively small. In such case the fusion penalty af-
fects more parameters which results in the estimates to be non-zero more likely. Focussing on GFLasso($\rho = 0.05$) at its $\lambda_{\text{min}}$, shows that varying $\mu$ can increase the observed FSR as much as four-fold from about 5% to about 20%. Moreover, even for GFLasso($\rho = 0.3$), which is the least affected among the GFLasso variants, the increase of FSR caused by changing $\mu$ is still as much as two-fold from roughly 2.5% to roughly 5%. Since the same behavior occurs also vice-versa when we use $\mu_{\text{min}}$ and vary $\lambda$, it illustrates the potentially detrimental influence of the fusion penalty GFLasso’s FSR which can lead to poor model selection performance.

As far as true selection rates by the GFLasso methods are concerned, the bottom-right plot in Figure 8.4 shows monotone marginal influences of the two penalty parameters on TSR. Unsurprisingly, larger values of $\lambda$ result in smaller FSR, while the influence of larger $\mu$ goes in the opposite direction, where larger $\mu$ (with $\lambda$ fixed) results in higher TSR. Also unsurprisingly, the influence of $\mu$ on TSR decreases for larger values of $\rho$. As for the actual values of TSR by GFLasso at the SEPE-optimal penalty parameters, the highest value of $\text{TSR}_{\text{min}} = 97\%$ is achieved by GFLasso($\rho = 0.05$), which, however, goes together with a rather underwhelming $\text{FSR}_{\text{min}} = 16\%$. Conversely, a relatively acceptable false rejection rate $\text{FSR}_{\text{min}} = 4.3\%$ by GFLasso($\rho = 0.3$) is duly paid for by $\text{TSR}_{\text{min}}$ of only 75%.

In conclusion, judging from the plots in Figure 8.4 under scenario A, it appears that the adaptive SVS methods quite convincingly outperform the rest of the considered methods. Compared to any of the adaptive SVS methods, both the lasso and the variants of the GFLasso yield substantially inferior performance in terms of the three measures. This suggests that the exploitation of the correlation in the response by the fusion penalty of the GFLasso is somewhat limited and generally performs worse than what is achieved by the $\ell_2$ and $\ell_\infty$ penalties used by the adaptive SVS methods. Overall, it seems that $L_2$–aSVS(1) and $L_2$–aSVS(2) are the champions among all of the considered methods under scenario A.

### 8.4.3 Performance for a single data set under scenario A

Having judged the potential performance of the methods in terms of SEPE, FSR and TSR, we emulated a real life applied setting. For a single data set under scenario A we obtained estimates for $B = (\beta_{ij})$ using 2-fold cross-validated ($\tau = 0.5$) values of the penalty parameters. We present the results in Figure 8.5, where instead of plotting the actual estimates of the parameters in $B$, we focus on $\alpha_1, \ldots, \alpha_{3008}$ in (8.4), which we were able to determine from the estimated values of $B$ since we know the actual values of loading parameters $\gamma_1, \ldots, \gamma_5$ in (8.5) from the simulation. Since this way we get a different value of $\alpha_i$ for each of the $p = 5$ phenotypes, we average the values for each locus, which simplifies the already quite busy plots in Figure 8.5. Focusing on $\alpha_i$’s instead of $\beta_{ij}$’s allows us to bring the estimates for different phenotypes on par with each other before they are averaged. Averaging $\beta_{ij}$ would not be nearly as appropriate, plus the plots would become counterproductively busy.

In Figure 8.5 we plotted the estimates of $\alpha_j, j = 1, \ldots, 3008$ (averaged over phenotypes) obtained by each variant of the adaptive methods $L_2$–aSVS and $L_\infty$–aSVS, the non-adaptive
Figure 8.5: Parameter estimates by various estimation methods under scenario A. Estimates of $\alpha_1, \ldots, \alpha_{3008}$ averaged over the 5 phenotypes obtained by variants of (adaptive) SVS, GFlasso, lasso and naive OLS with and without MTC. Cross-validated penalties $\lambda_{cv}$ and $\mu_{cv}$ were used. Only non-zero estimates for SNPs 1, ... , 3000 are plotted as small circles. True values and estimates of parameters $\alpha_{3009}, \ldots, \alpha_{3008}$ are shown at the top of each plot (with indices shifted by 3000). They are denoted by rotated squares (true values) black triangles (estimates) and marked by L1, ... , L8 and placed at their “actual” position within the genome (between SNPs 2001 and 3000).
Table 8.3: Observed values of performance measures SEPE, FSR, TSR, TEE$^\beta_z$, TEE$^\beta_{nz}$ and TEE$^\beta_{sum}$ for a single data set under scenario A. Each row in the table corresponds to a single plot in Figure 8.5. The rows are ordered according to observed SEPE and the best value in each column is marked by bold italic font.

$L_2$–SVS, $L_\infty$–SVS, the GFlasso, the lasso and the naive OLS. In Table 8.3 we included the observed values of $\text{SEPE}$, $\text{FSR}$, $\text{TSR}$, $\text{TEE}^\beta_z$, $\text{TEE}^\beta_{nz}$ and $\text{TEE}^\beta_{sum}$ with each row corresponding to a plot in Figure 8.5. In this comparison we included the naive OLS (both with and without MTC) in order to compare this simplistic method with the more sophisticated methods and see whether the criticism of the naive OLS outlined in Section 7.2 was justified.

**False and true selection rates**

From the plots in Figure 8.5 it is clear that the six variants of the adaptive SVS are the best performers among the considered methods. The adaptive SVS methods yield the smallest number of purple points, i.e. the non-zero estimates. The underlying numerical values are presented in Table 8.3. Comparing the adaptive SVS under cross-validated penalties among themselves, it seems to be $L_2$–aSVS(2) and $L_\infty$–aSVS(2) that are ahead of the other methods in terms of $\text{FSR}$. Both of these methods falsely select only 21 out of the 3000 non-causal SNPs, which corresponds to $\text{FSR}$ of 0.7%. Additionally, the other variants of the adaptive SVS are not trailing too much behind in this respect judging by their $\text{FSR}$ of 1.17%, 1.6%, 1.87%, 2.17% for $L_2$–aSVS(1), $L_\infty$–aSVS(1), $L_2$–aSVS(0.5), $L_\infty$–aSVS(0.5), respectively. The non-adaptive SVS methods $L_2$–SVS and $L_\infty$–SVS yield $\text{FSR}$ of 2.7% and 4%. The values observed for the adaptive SVS are genuinely quite impressive, especially when compared with the GFlasso and the lasso methods, the best of which is the lasso with $\text{FSR}$ of 5.93%, whereas the considered GFlasso method exhibit $\text{FSR}$ between 9.4% by GFlasso($\rho = 0.3$) and 68.9% by GFlasso($\rho = 0.05$). The impressiveness of the low $\text{FSR}$ by $L_2$–aSVS(2) and $L_\infty$–aSVS(2) is further enhanced by the fact that virtually all of the
falsely selected SNPs have estimates that are smaller than all of the estimates for the causal SNPs. Moreover, all of the causal SNPs were correctly selected by the adaptive SVS methods, which is a crucial point that should be stressed.

As far as the performance of the GFlasso is concerned, we notice a striking number of false non-zero estimates produced by the GFlasso. Compared with the adaptive SVS, the only variant of the GFlasso that bears any comparison is that with $\rho = 0.3$, for which the number of falsely selected parameters is 282, while the rest, and especially GFlasso($\rho = 0.05$), perform much worse than would be expected based on Figure 8.4. This seems to be related to the difficulty of cross-validating the two penalty parameters. Once $\rho$ becomes large enough to render the fusion ineffective, the problem seems to be somewhat mitigated as evidenced by the results for the lasso, which is a special case of GFlasso. Judging from Figure 8.5 and Table 8.3, the lasso with the cross-validated penalty performs relatively better than the GFlasso in terms of $FSR$. Nonetheless, it does not seem to reach its full potential as evidenced by the observed $FSR$ of almost 6%. Comparing this value with the results in Figure 8.4, where we saw $FSR$ below 2% at $SEPE$-optimal penalty, the current value of 6% is quite unimpressive.

**Prediction error**

In terms of prediction error, it is again $L_2$-aSVS(2) and $L_\infty$-aSVS(2) that perform best among all considered methods with $SEPE = 2.78$ and $SEPE = 3.02$, respectively. The other adaptive methods yield $SEPE$ between 3.25 and 5.29, which means that they also provide considerably improved performance over the two non-adaptive SVS methods and especially over the GFlasso and the lasso, where the best performing GFlasso($\rho = 0.2$) possesses $SEPE = 9.7$. Reassuringly, all of these values of $SEPE$ in Table 8.3 are near the observed minimums in Figure 8.4, which shows that cross-validation works well for this data.

**Estimation error**

Looking in more detail at the numerical characteristics contained within each plot, possibly with the exception of the lasso and GFlasso($\rho = 0.3$), all of the considered penalized regression methods perform similarly well in terms of estimating the true non-zero parameters as measured by $TEE_{nz}^a$. The total estimation error for all parameters $TEE_{sum}^a$, on the other hand, suggests that the adaptive SVS methods have the upper hand over the variants of the GFlasso including the lasso, which agrees with the observed behavior in terms of $FSR$.

**Performance of naive OLS**

Finally, we turn to the results for the naive OLS with and without MTC, which are in the bottom row of both Figure 8.5 and Table 8.3. First focusing on the naive OLS without MTC, we notice a large number of false detections resulting in $FSR$ of 8%. It is clear, however, that $FSR$ alone does not provide a complete picture since it ignores the fact that many of the estimates for truly zero parameters are severely biased. Numerically, this is expressed by the total estimation error $TEE_{sum}^b$, which here equals 36.6. In fact, the corresponding prediction
error $SEPE = 63.67$ is worse for this method than for the trivial all-zero estimate $\hat{B}_0 = 0$ with $SEPE$ of about 50 under scenario A. Given the nature of the naive OLS without MTC such poor performance is hardly surprising. Turning to the naive OLS with MTC, we notice a considerably better performance with $SEPE = 15.85$ and $FSR = 1.6\%$. However, both $TSR = 75\%$ and the estimation error $TEE_{null}^\beta = 8.7$ are unimpressive. Perhaps due to the repeated selection of SNPs that are correlated with the causal SNPs, the naive OLS with MTC estimates yield roughly the same non-zero estimation error $TEE_{null}^\beta = 4.5$ as the estimation errors obtained by the naive OLS without MTC, for which we observed $TEE_{null}^\beta = 4.6$. It seems fair, and completely unsurprising, that the naive OLS method both with and without MTC performs relatively poorly at both estimation and model selection.

### 8.5 Investigation under scenarios B and C

The data sets simulated under scenarios B and C were analysed by the methods employed under scenario A and by the additional four variants of the adaptive lasso. The aspect of performance we investigated under scenarios B and C was the sensitivity of each method to sample size. We used the $N = 25$ simulated genotype-phenotype data sets under each scenario, which we repeatedly analyzed using $k \leq 2000$ of the 2000 individuals, where $k$ ranged between 400 and 2000 in steps of 100. When increasing $k$, a randomly selected additional 100 individuals were added to the previously selected individuals in each step.

#### 8.5.1 Performance results under scenarios B and C

For each sample size, scenario and method we calculated an estimate of $B$. We used these to obtain average values of the prediction error $SEPE$, selection rates $FSR$, $TSR$, and total
estimation errors $\hat{T}_{z}^\beta$, $\hat{T}_{nz}^\beta$, $\hat{T}_{\text{sum}}^\beta$, which we plotted in Figure 8.7. Since there are large differences between the performance of different methods, the plots are on logarithmic scales, which means that fixed vertical distances between lines in the lower are indicative of a smaller difference between methods than those in the upper parts of the plots. We also note that we plotted smoothed versions of the actually observed lines.\footnote{With a single exception of $\hat{F}_{z}$ line for naive OLS under scenario C, which was too erratic and needed to be addressed separately. The smoothing was again done using the R function smooth.spline() with parameter $\text{spar}=0.65$.} Even though the original data resulted in relatively stable relationships, smoothing the lines substantially improved the readability. Otherwise, due to the relatively high number of compared methods, the raw data plots would have been considerably less readable. However, we must stress that we paid close attention to preserving the patterns observed in the raw results even after smoothing.

The considered variants of the adaptive SVS methods were the same as under scenario A, where we used $\nu = 1, 2, 0.5, 0$ with the two norms $\ell_2$ and $\ell_\infty$. Unlike under scenario A, we additionally used the adaptive lasso methods alasso(1), alasso(2), alasso(0.5) and alasso(lasso). In the GFLasso methods we put $\rho$ equal to 0.05, 0.1, 0.2, 0.3 and 0.4 under scenario B, whereas under scenario C we only considered values 0.05, 0.1 and 0.2. These choices were motivated by the maximum absolute values of correlation between the components of the phenotypes in each of these data sets, which under scenario B were between 0.5 and 0.6, while under scenario C they were between 0.2 and 0.3 for all considered sample sizes. The observed correlation coefficients are plotted in Figure 8.6, which shows histograms of the correlations and the maximum absolute value of the correlations as function of sample size. The plots clearly show the effect of the parameter choices we made under the two scenarios. Under scenario B (left), the observed correlations produce an asymmetric right-skewed histogram with a clearly positive mean correlation, while under C the histogram is symmetric with a near zero mean correlation. In the second and fourth plots (from left) we can see the maxima of absolute correlations, which decrease with sample size. We can also clearly see where the maxima lie, which justifies the chosen values for $\rho$ under the two scenarios.

The results of our analyses under scenarios B and C are shown in Figure 8.7, where the top six plots correspond to scenario B and the bottom six plot to scenario C. Based on the selected values for loading parameters $\gamma_i$, relatively speaking, we expected that scenario B favors the use of the naive OLS, since in this case the summing-up of the components of the phenotype into a single response does not destroy the information about association of regressors (SNPs) and the phenotype. Conversely, we expect scenario C to have a negative effect on the performance of the naive OLS.

### Selection rates

Judging from the plots in Figure 8.7, while keeping in mind the natural trade-off between $\hat{F}_{z}$ and $\hat{T}_{z}$, it is quite clear that the variants of the adaptive SVS perform very well in terms of both $\hat{F}_{z}$ and $\hat{T}_{z}$ under both scenarios. Especially $L_2-aSVS(1)$, $L_2-aSVS(2)$,
Figure 8.7: Observed average measures of performance FSR, TSR, SEPE and $\text{TEE}_\beta$, $\text{TEE}_{\text{sum}}$, for various methods (including the all-zero estimate $\hat{\beta}_0 = 0$) as functions of sample sizes under scenarios B (top) and C (bottom). For easier reading the values of SEPE were scaled up by the base line sample size of 2000. The y-axes are on natural logarithmic scales.
\(L_2–\)aSVS(0.5) and \(L_\infty–\)aSVS(1) exhibit FSR below 1\% and TSR generally well above 50\% for (almost) all sample sizes (note the y-axis log-scales). Additionally, with TSR increases up to about 80\% with sample size for these methods, which under scenario C makes them "catch up" to the best performing method in terms of TSR, while under scenario B they are not trail far behind the champions either. The most positive aspect of the good TSR performance by the adaptive SVS methods is the fact that it is not paid for by a lousy FSR showing. For all adaptive SVS methods FSR remains well controlled for all considered sample sizes.

As far as the non-adaptive SVS methods are concerned, both of them appear to be very good performers in terms of FSR, lacking only slightly behind the adaptive SVS methods under both scenarios. In terms of TSR both the non-adaptive SVS methods are performing well under scenario B, while under scenario C the method \(L_2–\)SVS appears to be noticeably inferior compared to the other SVS methods. However, when compared with the other methods its performance is quite reasonable even under scenario C.

Turning to the GFlasso, we notice large differences among the variants of GFlasso, where the performances in terms of FSR and TSR hugely depend on the value of \(\rho\). Looking at the FSR plots, it is clear that for small values of \(\rho\) the performance of GFlasso in terms of FSR is quite awful under both scenarios, where especially under scenario C two out of the three of the GFlasso variants have an unacceptably high FSR. Among the GFlasso variants, it is the ones with the highest values of \(\rho\) that perform best under each scenario. Under scenario B, it is GFlasso(\(\rho = 0.4\)), which exhibits FSR around 10\%, while under scenario C, it is GFlasso(\(\rho = 0.2\)) with FSR around 2\%. On the other hand, GFlasso(\(\rho = 0.05\)) is the weakest in both cases with FSR around 40\% (B) and 35\% (C). Translated into parameter counts such high FSR means that GFlasso(\(\rho = 0.05\)) yields more than 2000 non-zero estimates out of the 10000 truly zero parameters in B, which is a rather excessively large number. This is attributable to the effects of the fusion penalty which forces many estimates to be falsely non-zero. In fact, it is the lasso, a special case of the GFlasso without the fusion penalty, that shows the positive effects of removing the fusion penalty on FSR. Without the fusion penalty the lasso is actually among the overall best performers in terms of FSR together with the SVS methods. However, while it is true that the negative behavior of the GFlasso can be somewhat managed by increasing the value of \(\rho\) to diminish the effect of the fusion penalty, GFlasso’s performance in terms of FSR is never better than that of the lasso. And while the FSR performance of the lasso appears to be very good, as soon as we turn the attention to the TSR plots in Figure 8.7, we see the substantial negative trade-off between FSR and TSR performance of these methods. In terms of TSR the lasso is among the worst performers under both scenarios, while GFlasso(\(\rho = 0.05\)) is the overall best performer under both scenarios. However, this is negated by the worst showing in terms of FSR.

Turning to the adaptive lasso, Figure 8.7 shows that in terms of FSR the differences between the four variants of the adaptive lasso are rather small. Keeping in mind the log-scales of the y-axes, there seems to be virtually no difference between the methods. Especially under scenario C the observed values of FSR are all low. In terms of TSR we notice the more
pronounced differences. Under scenario B, it seems that alasso(1) and alasso(2) both lack behind the other adaptive lasso variants. Moreover, it is quite clear that the adaptive lasso does not substantially improve on the lasso. In some cases the adaptations even seem to have a detrimental effect on TSR. In either case, all of the adaptive lasso variants show an overall inferior performance to the adaptive SVS methods, perhaps with the exception of FSR with small sample sizes under scenario B.

**Prediction errors**

Looking at the SEPE plot (top) in Figure 8.7, which we scaled up by the baseline sample size of 2000, we can clearly see that it is again the adaptive methods that provide the superior performances in terms of SEPE, where $L_2-aSVS(2)$, $L_2-aSVS(1)$, $L_\infty-aSVS(2)$ and $L_\infty-aSVS(1)$ particularly stand out under both scenarios. Under scenario B we observed as much as three-fold improvement of SEPE by the best performing adaptive SVS method over the GFlasso, the lasso and the adaptive lasso methods, while under scenario C we observed roughly two-fold improvement of SEPE. Additionally, under both scenarios the ratio of improvement was increasing with sample size suggesting that the adaptive SVS methods make more efficient use of the additional data. It is also interesting that in terms of the prediction error the variants of the GFlasso and the lasso all perform very similarly with only minor difference between them. Finally, the two non-adaptive SVS methods seem to perform quite differently under each scenario, where $L_2-SVS$ provides a substantial improvement over the GFlasso and the (adaptive) lasso under both scenarios, $L_\infty-SVS$ is only competitive under scenario C while under scenario B it lacks behind even $L_2-SVS$ in terms of SEPE.

**Estimation errors: Perspective 1**

In Figure 8.7 we present three total estimation error measures for each scenario, namely $TEE^\beta_z$, $TEE^\beta_{nz}$, $TEE^\beta_{sum}$. Unlike $TEE^\alpha_z$, $TEE^\alpha_{nz}$, $TEE^\alpha_{sum}$ under scenario A, which focused on the underlying values of $\alpha_i$, here we directly measure how well a given procedure estimates the true parameter matrix $B$ in (7.1). Looking at the overall estimation error $TEE^\beta_{sum}$ under scenario B we notice an ordering of the methods that is very similar to that given by SEPE.

It is again clear that the best performance is provided by the adaptive SVS methods, where especially those based on the $\ell_2$ norm provide impressively low estimation error, which is less than one third of the error exhibited by either variant of the GFlasso for the largest sample size. The plots with the constituent parts of $TEE^\beta_{sum}$, namely $TEE^\beta_z$ and $TEE^\beta_{nz}$, provide further confirmation of the superiority of the adaptive SVS methods, since the ordering of the methods remains almost unchanged over the three plots under scenario B. The only noticeable exceptions again seem to be the lasso and the adaptive lasso variants for the smallest considered sample sizes, where they seem to be in the lead. However, in light of the relatively lower performance in terms of TSR and $TEE^\beta_{nz}$ it seems that these methods generally struggle to identify the existing links between regressors and responses. Consequently, those low values of either FSR or $TEE^\beta_z$ do not really suggest strong overall performance.
Ratios of regression parameter estimates and true values as function of sample size under scenario B

The averages were calculated for each parameter over the 25 data sets. Colors with various degrees of "hotness" ranging from blue (cold) to red (hot) indicate the absolute value of the corresponding true parameter value. Average TSR for each method is superposed over the ratios as blue line and the "target" ratio of 1 is plotted as green line.

Figure 8.8: Average ratios of estimates and true values for the 40 truly non-zero regression parameters in B as functions of sample size under scenarios B (top) and C (bottom). Estimation methods are indicated in each plot.
Turning to scenario C, we start with the overall estimation error $\text{TEE}_{\text{sum}}^\beta$. The corresponding plot again shows a clear inferiority of the GFlasso methods, where especially $\text{GFlasso}(\rho = 0.05)$ provides the most biased estimates for all but the smallest sample sizes. For the largest considered sample sizes, the best choices seem to again be the three adaptive SVS methods $L_2-aSVS(1)$, $L_2-aSVS(2)$ and $L_2-aSVS(0.5)$, while $L_{\infty}-aSVS(1)$, $L_2-aSVS(2)$ and the non-adaptive $L_2-SVS$ are not far behind. However, for the smaller sample sizes there is a somewhat worrying performance by $L_{\infty}-aSVS(2)$, which has a relatively high $\text{TEE}_{\text{sum}}^\beta$ under scenario C. Looking at the other two estimation error plots for an explanation we notice that the poor $\text{TEE}_{\text{sum}}^\beta$ showing goes hand in hand with relatively high values of $\text{TEE}_{nz}^\beta$ by $L_{\infty}-aSVS(2)$ for small sample sizes. In light of the good showing in terms of FSR and SEPE by the method for all sample sizes, the larger values of $\text{TEE}_{nz}^\beta$ do not seem terribly worrying, especially when compared with the estimation errors exhibited by the GFlasso. Nonetheless, it seems that under scenarios similar to C it requires relatively larger sample size for the relatively strong adaptation given by $\nu = 2$ to appropriately identify the underlying model, since a relatively high value of $\nu$ effectively removes penalization from some of the regressors by making their weights $\pi_k$ in (8.1) close to zero. For the (adaptive) lasso methods under scenario C we notice low values of $\text{TEE}_{nz}^\beta$ and relatively weak performances in terms of $\text{TEE}_{nz}^\beta$, which is similar to what was seen under scenario B. This again suggests a struggle to identify the links between regressors and responses.

Estimation errors: Perspective 2

In addition to Figure 8.7, the estimation errors by the considered methods can be evaluated based on Figure 8.8, where we plotted the average observed ratios of estimates and true values for the truly non-zero regression parameters as functions of sample size under scenarios B (top) and C (bottom). Using the notation of (7.3) and defining $A = \{(i,j) : \beta_{ij} \neq 0\}$ for $B = \{\beta_{ij}\}$, the plotted ratios are $\hat{B}_{|A}/B_{|A}$. Consequently, the behavior that indicates good performance by a method is for all of the ratio lines to be converging to the target value of one as the sample size increases and the quicker the convergence the better. However, a complete convergence of all lines is unrealistic under either of the two scenarios. Therefore, it is reasonable to generally prefer to have at least the ratios for coefficients with large (absolute) values converge to one, while the estimates for the small valued coefficients can be partially neglected. In the plots we used colors with various degrees of "hotness" ranging from blue (cold) to red (hot) to indicate the absolute value of the corresponding true parameter. Consequently, we want to see especially the ratios with darker red lines to be as close to 1 as possible. Additionally, we also superposed the average $\text{TSR}$ (from Figure 8.7) over the ratios as blue line, which allows for an easy differentiation of the methods in terms of $\text{TSR}$.

Looking at the top set of plots in Figure 8.8, we first judge the individual performances under scenario B, which are shown in the top 19 plots in the figure. In the first column we plotted the ratios for the GFlasso estimates. These plots provide additional evidence for the relative volatility of the estimates by that method especially with small $\rho$. The plots in
the second column show the results for the lasso and the four variants of the adaptive lasso. Generally speaking, it seems quite difficult to identify the best performing methods in those plots, since they all seem to perform rather poorly. We again see that under scenario B the adaptation did not lead to superior procedures over the lasso. The third and fourth columns of the plots under scenario B clearly show additional evidence of superiority of the adaptive SVS methods. We see that most of the red and the blue lines persuasively approach the target ratio of 1, while the value of $TSR$ are also generally high for all the adaptive SVS methods.

Making life more difficult for the estimation methods, we turn to the results under scenario C, which are shown in the bottom 16 plots in Figure 8.8 and where we immediately see a noticeably poorer performance by most of the methods. However, even under scenario C the champion methods seem to be the adaptive SVS. In Figure 8.8 we also notice that scenario C inhibits $TSR$ across all methods compared to scenario B, while the adaptive SVS methods exhibit the smallest decrease of $TSR$.

**Naive OLS**

Having judged the performances of all of the other methods we now use Figures 8.7 and 8.8 to evaluate the naive OLS under the scenarios B and C. First focusing on Figure 8.7, we notice that in terms of $FSR$ the method actually performs quite well for small sample sizes under scenario B, where it shows $FSR$ as low as the best performers. However, the observed $FSR$ rises quickly with the sample size becoming as much as ten-fold compared to the best performers. This behavior can be attributed to the fact that the naive OLS method selects those SNPs that have significant multiple testing corrected $p$-values in a univariate regression model. Since those $p$-values are equivalent to the $p$-values of a linear correlation test between the genotype at each locus and the proxy phenotype $Z$, the $p$-values tend to decrease with sample size unless the genotype and the proxy phenotype are completely uncorrelated. However, complete uncorrelatedness is not the case in our data, where there is a non-zero correlation between the proxy phenotype and the genotypes at causal SNPs ($k = 1001, \ldots, 1004$). Since the causal genotypes are not orthogonal to the rest of the genotypes (see Figures 8.1 and 8.2), we expect the number of significant $p$-values in the univariate regression model to increase with sample size.

In terms of $TSR$ under scenario B, the naive OLS also shows an increasing trend with sample sizes, though unfortunately it seems to level off rather quickly and never exceed 70%. As the method can only be considered (partially) reasonable for model selection, it comes as no surprise that it performs very poorly in terms of both estimation and prediction under scenario B. In terms of $SEPE$, it is apparent that the naive OLS performs very badly relative to all of the other methods. In terms of the total estimation error under scenario B, the naive OLS is by far the worst performer, exhibiting off-the-charts terrible estimation errors as measured by $TEE_z^\beta$ and $TEE_{sum}^\beta$. On the other hand, the performance in terms of $TEE_{inz}^\beta$ is very good, suggesting that in situations where the transformation of the phenotype does not destroy the association information (e.g. scenario B) the method can actually yield good estimates of the
8 Adaptive simultaneous variable selection

Effect sizes. The evidence of good estimation of non-zero parameters is further strengthened by Figure 8.8 where under scenario B the naive OLS yields very accurate estimates. As we indicated before, this is not surprising since scenario B favours the method and here we only focus on a single aspect of performance, namely $\text{TEE}_\beta^{nz}$.

Under scenario C, our expectations about its unfavorability for the naive OLS are confirmed. The method fails miserably at identifying almost any of the underlying structure within the data, which is clear from all of the corresponding plots in Figures 8.7. For lack of space, we did not include the naive OLS estimates under scenario C in Figure 8.8, since the plot was essentially empty with virtually all of the estimates equal to zero. Unlike under scenario B, where the naive OLS method actually differentiated itself from the all-zero solution $\hat{B}_0 = 0$, under scenario C the naive OLS estimates largely coincide with $\hat{B}_0$. As we pointed out, this is not surprising under scenario C, where the transformed phenotype loses most of the relevant information about the association of phenotypes and regressors. This causes the naive OLS estimates to essentially coincide with the all-zero solution $\hat{B}_0$, yielding terrible performances in terms of $\text{TSR}$, $\text{SEPE}$ and $\text{TEE}_\beta^{nz}$. The fact that the method might appear relatively competitive in terms of $\text{FSR}$, $\text{TEE}_z^{\beta}$ and $\text{TEE}_\sum^{\beta}$ provides only little solace, since that is not an indicator of quality performance, but rather further evidence of near complete lack of ability to model this kind of data.

Finally, we address the single exception that we made in terms of smoothing, which we omitted for $\text{TEE}_z^{\beta}$ for the naive OLS. The exhibited much more erratic behavior (compared to the other methods) serves to further illustrate the problems with this method. There seems to be an inherent "discontinuity" of selection based on MTC corrected $p$-values, which none of the other methods exhibited. On the one hand, the naive OLS provided all-zero estimates most of the time under scenario C while giving relatively large values of estimates for the truly zero parameters as soon as the $p$-value threshold is exceeded, which leads to a negative behavior in terms of the associated estimation error measure.

8.6 Conclusions

In this chapter we presented the adaptive SVS method, which is a method for estimating parameters in the multivariate multiple regression model of (7.1) which exploits the assumptions of sparsity and common association. We put the method under thorough scrutiny in a realistic simulation study in the context of genotype-phenotype data, where we compared it from numerous perspectives with several other methods under several different simulation scenarios. A conclusion that can be drawn from our investigation is that the adaptive SVS method is a powerful tool which yields accurate estimates while providing good performance in terms of both selection rates and prediction errors. An overall impression is that the adaptive SVS method performs persuasively better than the other considered methods, which included methods that were specifically tailored for the considered multivariate phenotypes.
Due to the popularity of the naive OLS, we paid special attention to the performance of the method under several different scenarios. We showed that although the method can work quite well in certain respects such as $\text{TEE}_n^{\beta}$ under scenario B, it often fails miserably in other respects ($\text{TEE}_n^{\beta}$ and $\text{SEPE}$) and/or under other scenarios (A and C). However, the adaptive SVS methods seem to be on par with the naive OLS even under the scenario favourable to the naive OLS, while the naive OLS is clearly inferior to the adaptive SVS methods in all of the other considered measures. It seems therefore clearly unwise to use the naive OLS as the method of choice for such analysis.

The adaptive SVS method requires user input in two distinct ways. On the one hand, it requires the user to select a suitable way to determine the adaptation weights, where we showed that an approach as simple as univariate regression can already yield very favourable behavior. This suggests that there is still room for improvement of the performance of the adaptive SVS method by using a more sophisticated way of determining adaptation weights although probably at the cost of increased computational burden. On the other hand, like any penalized regression method, the adaptive SVS also requires a good choice of value of the tuning penalty parameter. Usual ways to determine the tuning parameters include cross-validation and information criteria such as AIC, BIC and GIC (Fan and Tang (2013)). In our analysis we showed that the non-adaptive approach based on cross-validation works quite well towards allowing the method to maximize its potential. In summary, the adaptive SVS is a strong method that in our opinion should become the workhorse for analysis of association between a large number of regressors and correlated multivariate phenotypes.
Application to data: Alternative splicing

In this chapter we present the results of an eQTL analysis of the expression data generated by the Geuvadis RNA sequencing project for 1000 Genomes samples (Lappalainen et al. (2013)). The goal of our analysis is to identify SNPs involved in alternative splicing, which is a gene expression regulation process that allows for a single gene to code for multiple proteins (see below). The complete Geuvadis data is publicly available and there are currently several ongoing projects dedicated to the analysis this data. We note that at the time of writing the project is still ongoing and a complete report of the analysis of Geuvadis data will soon be available in de Menezes et al. (2016).

In this chapter we present only a sample of results of the pilot study of the Geuvadis data performed by a team at VU University Medical Center Amsterdam. The pilot study included only genes located on chromosome 1 and among other things included the usage of the adaptive SVS estimation method for the modeling of association between gene expression levels and SNP data. Here we aim to illustrate the type of behavior the considered penalized regression methods exhibit on the type of data analyzed in the study. Among other things, we compare the selection performance of the adaptive SVS method with the performance of the univariate OLS, the lasso, several variants of the adaptive lasso, and also a global testing procedure called G2, which was recently developed by Chaturvedi et al. (2015) specifically for analyzing such multivariate data. The comparison shows that all of the considered penalized regression methods including the adaptive SVS exhibit a desirable performance particularly in terms of sparsity of the estimates. Moreover, the comparison with the G2 test uncovered a high degree of agreement. We believe that the results justify the usage of these methods for application in the context of detecting alternative splicing as well as many other similar contexts. Since in the previous chapters we focused on the linear model, here we also focus on regression methods based on the linear model. The analysis in the paper was performed using the analogous regression methods for the multinomial regression model, which is more appropriate for the expression data at hand.
As stated above, the goal of our analysis is to identify a SNP-driven gene expression regulation process known as alternative splicing. Since our focus is mainly on the statistical methods, we provide only a brief description of the underlying genetic processes. For a more thorough treatment see for example Blencowe and Graveley (2007).

A gene is a sequence of genetic code (nucleotide bases) that contains instructions to make various protein molecules. Most human genes, about 94% in fact (Ward and Cooper (2010)), are so called interrupted genes, which means that they consist of several regions of different functional type referred to as exons and introns. The number of exons in human genes varies between 1 and 363 (the latter in the titin gene) and the average number of exons per gene is about 10 (Strachan and Read (2011)). During DNA transcription the genetic code undergoes a process called splicing, when introns are removed while exons are preserved and transcribed into RNA (i.e. expressed). Crucially, however, not all exons are always expressed, which means that the same genetic code in a gene can lead to different RNA transcripts. This occurs when, during RNA transcription, different subsets of exons are expressed. This phenomenon when a single gene produces different RNA transcripts is called alternative splicing. Interestingly, different RNA transcripts do not have to result in differential protein expression.

If expression data per exon is available, it is possible to look for evidence of alternative splicing. This can be done by checking if all exons are observed in equal proportion. There are technical limitations that need to be taken into account during such analysis. For instance, exon length may affect measurements since exons below a certain minimal length may be less efficiently handled both during sequencing and during alignment. This effect is not linear, which means that as exon length decreases, the number of reads mapping to the exon decreases with exon length faster than linearly.

**9.2 Data**

The Geuvadis project produced RNA-sequencing for 465 lymphoplastoid cell line (LCL) samples from 5 populations of the 1000 Genomes project: the CEPH (CEU), Finns (FIN), British (GBR), Toscanin (TSI) and Yoruba (YRI). Of these, 423 were part of the 1000 genomes Phase 1 data set with low-coverage whole genome and high-coverage exome sequencing data, and the remaining 42 are part of the later phases of 1000 Genomes with Omni 2.5M SNP array data at the time of this study. The complete Geuvadis data contains 148,002 exons spread over 15,480 genes. After quality control (QC) performed by the Geuvadis team there were 462 unrelated individuals from various cohorts remaining. In the pilot study we restricted the focus only to exons on chromosome 1 and used the usual QC criteria such as a MAF threshold of 5%, etc. The distribution of per-gene exon counts on chromosome 1 in the raw data is shown in Figure 9.1. The figure shows that while the maximum observed number
of exons in a single gene was 105, over 50% of the included genes have 8 or fewer exons, only 10% have more than 20 exons and only 1% genes contained more than 40 exons.

The raw data was pre-processed using the software tool PEER\(^1\). In order to correct for large trends typically produced by batch effects, the first 10 principal components were removed. However, in the pilot we did not correct for the effect of exon length on expression. Exons that were not expressed at all in the data were eliminated. Moreover, genes with only a single exon were also removed, since these cannot undergo alternative splicing. From the total of 14,758 exons in 1389 genes on chromosome 1, there were 14,656 exons in 1376 genes left in the data set after pre-processing. In order to improve the homogeneity of the data we additionally removed the data for the Yoruba population from the pilot study, after which there were 373 samples left. Besides removing the top 10 principal components, the exon expression data was further transformed to account for the fact that exon expressions are on an exponential scale. The transformation used was the variance-stabilizing transformation \( h(y) = \gamma \arcsin(h(a + by)) \) first proposed by Huber et al. (2002) in the context of high-dimensional data analysis for microarray data normalization. The constants \(a, b, \gamma\) are determined by fitting an assumed quadratic relationship between the mean and the variance of the data parameterized via \(a, b, \gamma\) (for details see Huber et al. (2002)).

\(^1\)See http://www.sanger.ac.uk/resources/software/peer
9.3 Analysis methods

In order to identify markers on chromosome 1 that might be involved in the regulation of alternative splicing we modeled the data for each gene in the following way. The transformed expression counts for each exon of the gene were entered as multivariate response matrix $Y$ into a linear model of (7.1) with individual SNP genotypes forming the columns of the design matrix $X$. For every gene the number of rows in both $Y$ and $X$ (i.e., the sample size) was the number of individuals in the data, which was $n = 373$. The number of columns in $Y$ and $X$ differed between the genes and equaled the counts showed in the top and bottom plots of Figure 9.1, respectively. For simplicity of analysis each column of the response and the design matrices were centered with the corresponding (per-column) sample means.

For the estimation itself we used several estimation methods. Of primary interest to us were the adaptive SVS methods detailed in Chapter 8. In order to put the estimates obtained by the adaptive SVS methods into a proper perspective, we also employed several secondary estimation methods (see below) and compared the obtained results. For each of the penalized methods the values of penalty parameter was optimized via 10-fold cross-validation. In the notation of Section 8.4.1) this corresponds to $\tau = 0.9$.

### Primary estimation methods

The estimation methods of primary interest were the variants of the $L_2$-aSVS method. For reasons of computational complexity we chose the $\ell_2$ based methods instead of the $\ell_\infty$ methods, since the available solver MOSEK is considerably less efficient than the glmnet package for R. For the analysis we used both the non-weighted ($L_2$-SVS) and weighted variants of the method with several different weighting schemes, which were all based on the univariate OLS approach described in Section 8.2. Three of them were identical to the ones used throughout Chapter 8, which we denoted by $L_2$-aSVS(1), $L_2$-aSVS(2) and $L_2$-aSVS(0.5). In addition to those we also considered a number of higher power transformations of the OLS based weights and denote the corresponding methods as $L_2$-aSVS($\nu$), where $\nu = 3, \ldots, 10$ (see (8.3)). The higher power transformations put more emphasis on the initial univariate OLS estimates. It seems reasonable to expect that there is an optimum transformation which strikes the right balance between the information contained within the univariate OLS estimates, which determine the weights, and the ability of penalized regression methods to uncover the association.

In order to illustrate the actual numerical scale of the weights in Figure 9.2 we show the various power transformed weights. In order to keep the plots readable we only show the weights for those SNPs selected by each corresponding variant of adaptive SVS method, since those are the most interesting to us.

### Secondary estimation methods

In order to better judge the performance of the $L_2$-aSVS methods on real data, we employed a total of five additional estimation methods. These were the lasso, three variants of the
Figure 9.2: Effective weights utilized by the adaptive methods. The power transformation parameter $\nu$ values are denoted in the upper left corner of each plot. Only the weights for SNPs selected by the adaptive SVS method are shown.
adaptive lasso (Section 7.2.3) with weights determined in the same way as in Chapter 8, and the naive OLS with MTC by the total number of tests (i.e. 53,286,710). Analogously to the adaptive SVS methods above, we denote the three adaptive lasso variants as alasso(1), alasso(2), alasso(0.5), where the number in brackets is the value for $\nu$ (see (8.3)).

**Testing method: G2 global test**

In Chaturvedi et al. (2015) the authors develop a testing procedure called G2, which allows to evaluate the presence of association between two groups of variables in the context of generalized linear models. The test treats the first group of variables as a multivariate response and models it using a multivariate design matrix, which contains the second group of variables. For each multivariate response and design matrix the test produces a single $p$-value, which allows the user to assess whether the two are associated or not. Since it works on the aggregate level G2 is a type of a global test. In our application the two groups of variables are the exon expression levels on the response side (first group) and the SNP genotypes of a single gene on the covariate side (second group). Regarding the test statistic here we provide the definition only for the linear model while referring the reader to Chaturvedi et al. (2015) for the motivation, its properties and the calculation of the $p$-values via a permutation scheme.

In the linear model (7.1) with $p$ denoting the phenotype dimension (i.e. the number of exons) and $q$ denoting the number of variables (i.e. SNP genotypes), the G2 test proceeds with the "stacked-up" model (7.2) with parameter vector $\tilde{B}$ of length $pq$. In such model the question of interest for a global test is whether $\tilde{B} = 0$ or not. Using a random effects framework G2 assumes that the parameter $\tilde{B}$ comes from a multivariate distribution with zero expectation and the covariance matrix $E(\tilde{B}\tilde{B}^\prime) = \tau^2 \Sigma_B$, where $\Sigma_B$ is a positive semi-definite $pq \times pq$ matrix. This leads to an equivalent null hypothesis $H_0: \tau^2 = 0$, which is tested against $H_1: \tau^2 > 0$. Relying on the high-dimensional local optimality derived in Goeman et al. (2006), the authors of Chaturvedi et al. (2015) propose a score-type test statistic

$$S^*_G2 = \frac{1}{2} \{\ell_n^\prime \Sigma_B \ell_n - \text{trace}(I_n \Sigma_B)\},$$

where $\ell_n$ is the score of the null hypothesis model for $\tilde{B}$ and $I_n$ is the corresponding sample Fisher information matrix. The matrix $\Sigma_B$ is a tuning parameter of the G2 test, which strongly influences the power of the test. Chaturvedi et al. (2015) suggest to use $\Sigma_B = I_p \otimes \Sigma_Y$, where $L \Sigma_Y$ is the sample covariance matrix of $\tilde{Y}$ and $\otimes$ denotes the Kronecker product. This leads to the test statistic

$$S_{G2} = \text{trace}((I_n - \frac{1}{n} 1_n 1_n^\prime)\tilde{X}\tilde{X}^\prime(I_n - \frac{1}{n} 1_n 1_n^\prime)\tilde{Y}\tilde{Y}^\prime).$$

We deployed the test statistic $S_{G2}$ to the Geuvadis data on a per-gene basis, where for each gene the exon expressions formed the multivariate response and the SNP genotypes of the SNPs in that gene were used as the regressors. Using the resulting corrected $p$-values we were able to reject the null hypothesis for 64 out of the 1376 genes on chromosome 1 at level $\alpha = 0.05$. The list of the rejected gene names can be found in Table 9.1.
9 Application to data: Alternative splicing

Table 9.1: List of the 64 genes on chromosome 1 identified as significant by the G2 test.

<table>
<thead>
<tr>
<th>genes (total of 1376)</th>
<th>SNPs (total of 495875)</th>
<th>G2 vs M</th>
</tr>
</thead>
<tbody>
<tr>
<td>method (M)</td>
<td>selected</td>
<td>ratio</td>
</tr>
<tr>
<td>naiveols</td>
<td>242</td>
<td>17.59 %</td>
</tr>
<tr>
<td>L2-SVS</td>
<td>512</td>
<td>37.21 %</td>
</tr>
<tr>
<td>L2-aSVS(0.5)</td>
<td>442</td>
<td>32.12 %</td>
</tr>
<tr>
<td>L2-aSVS(1)</td>
<td>368</td>
<td>26.74 %</td>
</tr>
<tr>
<td>L2-aSVS(2)</td>
<td>280</td>
<td>20.35 %</td>
</tr>
<tr>
<td>L2-aSVS(3)</td>
<td>235</td>
<td>17.08 %</td>
</tr>
<tr>
<td>L2-aSVS(4)</td>
<td>205</td>
<td>14.90 %</td>
</tr>
<tr>
<td>L2-aSVS(5)</td>
<td>200</td>
<td>14.53 %</td>
</tr>
<tr>
<td>L2-aSVS(6)</td>
<td>205</td>
<td>14.90 %</td>
</tr>
<tr>
<td>L2-aSVS(7)</td>
<td>111</td>
<td>15.33 %</td>
</tr>
<tr>
<td>L2-aSVS(8)</td>
<td>216</td>
<td>15.70 %</td>
</tr>
<tr>
<td>L2-aSVS(9)</td>
<td>231</td>
<td>16.79 %</td>
</tr>
<tr>
<td>L2-aSVS(10)</td>
<td>222</td>
<td>16.13 %</td>
</tr>
<tr>
<td>lasso</td>
<td>80</td>
<td>5.81 %</td>
</tr>
<tr>
<td>alasso(0.5)</td>
<td>82</td>
<td>5.96 %</td>
</tr>
<tr>
<td>alasso(1)</td>
<td>86</td>
<td>6.25 %</td>
</tr>
<tr>
<td>alasso(2)</td>
<td>108</td>
<td>7.85 %</td>
</tr>
<tr>
<td>G2 test</td>
<td>64</td>
<td>4.70 %</td>
</tr>
</tbody>
</table>

Table 9.2: Comparison of SNP and gene selection counts by various regression methods. The results of the G2 test are also included.

9.4 Results

The numerical results of the estimation by the various penalized regression methods and the G2 test are shown in Tables 9.2, 9.3 and 9.4, which also allow for comparison of the agreement between the methods. The SNP and gene selection counts of Table 9.2 are also depicted in Figure 9.2. In addition to Figure 9.2 we also depict the counts of SNPs selected in each gene (among genes with at least 1 selected SNP) in Figure 9.2, while the average
absolute values of individual estimates are shown in Figure 9.5, which provides the reader with an illustration of the degree of agreement between each regression method and the G2 test underneath each x-axis (see the caption of Figure 9.5).

**SNP and gene selection counts**

In Table 9.2 we present for each method the counts and rates of genes with at least 1 selected SNP, the number of selected SNPs in total, and the rates of agreement between the regression methods and the G2 test. In order to obtain a reasonably fair comparison of the considered penalized regression methods all of the estimates were obtained using the R package `glmnet`. Moreover, all penalty parameters were 10-fold cross-validated using the function `cv.glmnet()` out of a list of 100 candidates automatically selected by `cv.glmnet()`.

Looking at the results in Table 9.2 we immediately notice that all of the estimation methods yield very sparse estimates, which is a desirable outcome given the context of the data at hand. Unsurprisingly, the method that leads to the smallest dimensionality reduction is the naive OLS method, which selects as many as 6669 SNPs with a selection rate of 0.13%. These 6669 SNPs were located on 242 genes, which yields an average of over 27 selected SNPs per gene (again the average is computed only among genes that had at least 1 selected SNP). Given that the naive OLS is a univariate method which models each SNP separately from the others, and the fact that SNPs close together on the same chromosome tend to be correlated, even with the multiple testing correction the selected SNP counts by the naive OLS can be seen as kinds of upper bounds for the actual number of SNPs involved in alternative splicing (assuming the linear model is appropriate). The actual per-gene selected SNP counts can be seen in Figure 9.2, which shows the expected poor performance by the naive OLS and the seemingly much more reasonable SNP selection counts by the other methods.
most of which select no more than two SNPs in at least 50 percent of the selected genes.

For the SVS methods the table shows that the non-adaptive $L_2$-SVS yields the least sparse solution and selects 1782 SNPs in 512 genes. With the adaptation we obtain much sparser solutions and both the SNP and the gene selection counts decrease. Since the 11 adaptive SVS methods differ only by the degree to which the initial univariate OLS estimates shape the final solution it is not surprising that higher values of $p$ lead to more sparse solutions with for instance $L_2$-aSVS(1) and $L_2$-aSVS(2) selecting 1100 and 797 SNPs in 368 and 280 genes, respectively. The sparsity generally increases further with increasing $p$, although the maximum sparsity is not attained for the highest considered values of $p$. As illustrated by Figure 9.3, the minimum per-gene SNP selection count was observed for $p = 3$ with $p = 2$ a close second, the minimum overall SNP selection count was seen with $p = 4$, and the minimum gene selection count with $p = 5$. Given these values and the observed rates of
Table 9.3: Intersection counts of SNPs selected by various regression methods. For each cell the table shows the number of SNPs selected by both methods corresponding that cell.

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
<th>M8</th>
<th>M9</th>
<th>M10</th>
<th>M11</th>
<th>M12</th>
<th>M13</th>
<th>M14</th>
<th>M15</th>
<th>M16</th>
<th>M17</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1: naiveols</td>
<td>6669</td>
<td>275</td>
<td>35</td>
<td>257</td>
<td>234</td>
<td>187</td>
<td>156</td>
<td>140</td>
<td>127</td>
<td>104</td>
<td>100</td>
<td>93</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M2: L2-SVS</td>
<td>275</td>
<td>1782</td>
<td>481</td>
<td>275</td>
<td>178</td>
<td>481</td>
<td>799</td>
<td>421</td>
<td>241</td>
<td>143</td>
<td>87</td>
<td>67</td>
<td>37</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>M3: L2-aSVS(0.5)</td>
<td>35</td>
<td>481</td>
<td>1399</td>
<td>390</td>
<td>208</td>
<td>113</td>
<td>59</td>
<td>37</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>M4: L2-aSVS(1)</td>
<td>257</td>
<td>799</td>
<td>1100</td>
<td>571</td>
<td>336</td>
<td>211</td>
<td>141</td>
<td>109</td>
<td>57</td>
<td>42</td>
<td>24</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>M5: L2-aSVS(2)</td>
<td>234</td>
<td>421</td>
<td>208</td>
<td>571</td>
<td>336</td>
<td>211</td>
<td>141</td>
<td>109</td>
<td>57</td>
<td>42</td>
<td>24</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>M6: L2-aSVS(3)</td>
<td>187</td>
<td>241</td>
<td>113</td>
<td>59</td>
<td>37</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M7: L2-aSVS(4)</td>
<td>156</td>
<td>143</td>
<td>59</td>
<td>37</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M8: L2-aSVS(5)</td>
<td>140</td>
<td>87</td>
<td>37</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M9: L2-aSVS(6)</td>
<td>127</td>
<td>67</td>
<td>24</td>
<td>109</td>
<td>57</td>
<td>42</td>
<td>24</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M10: L2-aSVS(7)</td>
<td>104</td>
<td>50</td>
<td>20</td>
<td>84</td>
<td>57</td>
<td>42</td>
<td>24</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M11: L2-aSVS(8)</td>
<td>100</td>
<td>40</td>
<td>16</td>
<td>67</td>
<td>57</td>
<td>42</td>
<td>24</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M12: L2-aSVS(9)</td>
<td>93</td>
<td>39</td>
<td>16</td>
<td>62</td>
<td>57</td>
<td>42</td>
<td>24</td>
<td>16</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M13: L2-aSVS(10)</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>11</td>
<td>18</td>
<td>29</td>
<td>47</td>
<td>65</td>
<td>101</td>
<td>114</td>
<td>146</td>
<td>996</td>
<td>114</td>
<td>146</td>
<td>996</td>
<td>114</td>
</tr>
<tr>
<td>M14: alasso</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M15: alasso(0.5)</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M16: alasso(1)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M17: alasso(2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Agreement with G2 it can be argued that either $p = 2$ or $p = 3$ are suitable choices.

All the selection counts for the adaptive SVS methods correspond to averages of about 3 selected SNPs per gene, whereas the average was about 3.5 selected SNPs per gene for the non-adaptive L2-SVS. As far as the average number of selected SNPs is concerned, similarly to the adaptive SVS methods the lasso and the three adaptive variants of the lasso yield similar results of around 3. However, the number of genes with at least 1 selected SNP is noticeably lower with these methods which select only 80-108 genes, or about 6-8%.

Agreement of SNP and gene selection among the methods

Besides the selection counts one might be interested to judge the degree of agreement among the various methods including the G2 test. From Table 9.2 we see that the lasso and the adaptive lasso methods selected at least 1 SNP in just under half of the genes found significant by the G2 test. On the other hand, the top five of the (adaptive) SVS methods selected SNPs in 80 percent of more of the genes flagged by the G2 test and even the most extreme adaptive SVS methods select about 60 percent of the genes flagged by G2. Both the non-adaptive SVS method and the top three of the adaptive SVS methods also exceeded the naive OLS in terms of the agreement with the G2 test, while the next two methods selected SNPs in only slightly fewer genes than the naive OLS.

The agreement among the estimation methods can be judged based on Tables 9.3 and 9.4, which respectively show the counts of SNPs and genes selected simultaneously by both methods corresponding to each cell of the two tables. Looking at both tables we notice a high degree of selection agreement between the naive OLS and the (adaptive) SVS methods especially for the lower values of $p$. On the other hand, the two tables also show a lack
Figure 9.5: Average estimates of the regression parameters for each SNP on chromosome 1. Averages in absolute value are taken over the columns of estimated $\mathbf{B}$. Under the x-axis orange bars indicate genes with at least 1 SNP selected by the estimation method, while purple bars indicate genes declared significant by the G2 test.
of agreement between the first 13 methods (the naive OLS and the (adaptive) SVS) and the remaining 4 methods (the ordinary and adaptive lasso) as well as among the 4 (adaptive) lasso methods themselves. The observed lack of agreement in terms of SNP selection could perhaps be explained by the possibly strong LD among the SNPs, which causes collinearity within the design matrix. However, the relatively poor degree of agreement also in terms of gene selection makes the observed selection behavior by the (adaptive) lasso rather worrying. The results suggest that the (adaptive) SVS methods exhibit relatively more stability in their selection.

Estimates by different methods

In Figure 9.5 we plotted the estimates obtained by the various methods. Each plot shows the average absolute value estimates for the corresponding method, where the averages are taken over the exons within a single gene, i.e. over the columns of the estimates of $B$ for each gene. The plots also provide a comparison of the estimation methods and the G2 test, where under each x-axis the orange bars indicate genes with at least 1 SNP selected by the estimation method, while the purple bars indicate genes declared significant by the G2 test. The bars represent visualization of the three right-most columns in Table 9.2.

9.5 Conclusion

Our goal in this chapter was to provide an illustration of the relative performances of the penalized regression methods with an emphasis on the adaptive SVS. While we admit that the
illustration was limited in terms of scope and depth, we believe the results here show the usefulness of the considered penalized regression methods, where especially the adaptive SVS methods (with cross-validated parameters) are able to select a very sparse subset of genes and genetic markers. Such behavior is deemed highly desirable in the context of alternative splicing.

While the variable selection performance was the primary focus of our analysis we were also interested in the quality of effect size estimation. This is one of the strong points for using penalized regression methods in general, since they allow the user to avoid an undesirable reusing of the data, which stands in contrast with many of the classical statistical methods which perform model selection and estimation separately by reusing the available data. The latter would be a concern also in an analysis based on the G2 test, which in this chapter we used as a sort of baseline for the comparison of the selection performances of the considered penalized regression methods.

Finally, we also focused on the comparison of the individual penalized regression methods against each other in applied setting. It turned out, likely due to the more efficient exploitation of the multivariate nature of the responses, that the adaptive SVS methods provided SNP selection for a significantly larger portion of the considered genes compared to the lasso and the adaptive lasso variants. It seems that the latter are perhaps too restrictive during selection. While selecting a larger number of genes, the adaptive SVS variants simultaneously limit the number of selected SNPs to a manageably small number. This stands in contrast with the univariate OLS method, which appears to ignore both the multivariate nature of the response as well as the dependence among the SNPs and thus yields a much larger counts of selected SNPs. Overall, the initial results obtained from the pilot analysis of alternative splicing in Geuvadis data using the penalized regression methods were quite favorable. Naturally, a deeper statistical analysis and biological investigation are required to further our understanding of the genetic processes involved in alternative splicing.
Appendix: General Theory

For completeness here are several classical definitions and theorems utilized in this thesis.

**Definition A.1 (Convergence in probability)** Let \( \{X_n, n \in \mathbb{N}\} \) be a sequence of random variables which are defined on probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and let \(X\) be a random variable also defined on \((\Omega, \mathcal{F}, \mathbb{P})\). We say that \(X_n\) converge to \(X\) in probability as \(n \to \infty\), if, for any \(\varepsilon > 0\), it holds \(\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0\). We denote this by \(X_n \xrightarrow{p} X\).

**Definition A.2 (Convergence in distribution)** Let \(\{X_n, n \in \mathbb{N}\}\) be a sequence of \(m\)-dimensional random vectors which are defined on probability spaces \((\Omega_n, \mathcal{F}_n, \mathbb{P}_n)\) and let \(X\) be an \(m\)-dimensional random vector defined on \((\Omega, \mathcal{F}, \mathbb{P})\). We say that \(X_n\) converge to \(X\) in distribution as \(n \to \infty\), if the distribution functions \(\mathbb{P}_n(X_n \leq x)\) converge to \(\mathbb{P}(X \leq x)\), as \(n \to \infty\), for all \(x \in \mathbb{R}^m\), where \(\mathbb{P}(X \leq x)\) is continuous. We denote this by \(X_n \xrightarrow{D} X\).

**Definition A.3 (Order of convergence)** Let \(a_n, b_n\) be sequences of real numbers. We define \(a_n = o(b_n)\) to be equivalent to \(\lim_{n \to \infty} a_n/b_n = 0\).

**Definition A.4 (Order and boundedness in probability)** Let \(\{X_n\}, n \in \mathbb{N}\) be a sequence of random variables all defined on the probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and let \(\{a_n\}\) be a sequence of real numbers. We say that a sequence \(\{X_n\}\) is bounded in probability if for every \(\varepsilon > 0\) exists \(M > 0\) such that \(\mathbb{P}(|X_n| > M) < \varepsilon\) for all \(n \in \mathbb{N}\). We write \(X_n = o_p(a_n)\) if \(X_n/a_n \xrightarrow{p} 0\) as \(n \to \infty\) and \(X_n = O_p(a_n)\) if \(X_n/a_n\) is bounded in probability.

**Definition A.5 (Estimator consistency)** Let \(X_1, \ldots, X_n\) come from a distribution which depends on a parameter \(\theta\) and let, for each \(n \in \mathbb{N}\), \(T_n = T_n(X_1, \ldots, X_n)\) be an estimator of \(\theta\). We say that \(T_n\) is a consistent estimator of \(\theta\), if \(T_n \xrightarrow{p} \theta\) in probability, as \(n \to \infty\).

**Lemma A.6 (Slutsky’s lemma)** Let \(X_n, Y_n, Z_n, n \in \mathbb{N}\), and \(Y\) be \(m\)-dimensional random vectors and let \(a, b \in \mathbb{R}^m\) be constants. If \(X_n \xrightarrow{D} X, Y_n \xrightarrow{D} a\) and \(Z_n \xrightarrow{D} b\), as \(n \to \infty\), then \(Y_n X_n + Z_n \xrightarrow{D} aX + b\), as \(n \to \infty\).

**Proof.** For proof see for example Bickel and Doksum (1977), Theorem A.14.9.

**Theorem A.7 (Continuous mapping theorem)** Let \(X_n\) and \(X\) be \(m\)-dimensional random vectors and let \(f : \mathbb{R}^m \to \mathbb{R}^k\) be a continuous mapping on \(C_f \subset \mathbb{R}^m\) where \(\mathbb{P}(X \in C_f) = 1\). If \(X_n \xrightarrow{D} X\), as \(n \to \infty\), then also \(f(X_n) \xrightarrow{D} f(X)\), as \(n \to \infty\).
Proof. For proof see for example Billingsley (1968), Chapter 1, Section 5.

**Theorem A.8 (Joint convergence theorem)** Let \( X_n, X \) and \( Y_n, Y \), \( n \in \mathbb{N} \), be \( m \)-dimensional random vectors. If \( X_n \to_d X \) and \( Y_n \to_p c \) for some constant \( c \in \mathbb{R}^m \), as \( n \to \infty \), then also \((X_n, Y_n) \to_d (X, c)\), as \( n \to \infty \).

**Proof.** For proof see for example van der Vaart (1998), Theorem 2.7.

**Theorem A.9 (Delta method)** Let \( \phi : \mathbb{R}^k \to \mathbb{R}^m \) be a map defined on \( D_\phi \subset \mathbb{R}^k \) and differentiable in a neighborhood of \( \theta \in \mathbb{R}^k \). Let \( T_n \) be random vectors taking their values in \( D_\phi \). If \( r_n(T_n - \theta_n) \to_d T \) for vectors \( \theta_n \to \theta \) and numbers \( r_n \to \infty \), then \( r_n(\phi(T_n) - \phi(\theta_n)) \to_d T \).

**Proof.** For proof see for example van der Vaart (1998), chapter 4.

**Theorem A.10 (Multivariate Feller-Lindeberg CLT)** For each \( n \in \mathbb{N} \) let \( X_{nk1}, \ldots, X_{nk_k} \), \( k_n \in \mathbb{N}, k_n \to \infty \) as \( n \to \infty \), be independent random vectors with finite variances such that as \( n \to \infty \)

\[
\sum_{k=1}^{k_n} \mathbb{E}(\|X_{nk}\|^2 I_{\{\|X_{nk}\| > \varepsilon\}}) \to 0, \quad \text{for every } \varepsilon > 0, \quad \text{and} \quad \sum_{k=1}^{k_n} \text{var}X_{nk} \to \Sigma,
\]

for some finite matrix \( \Sigma \). Then, as \( n \to \infty \), the sequence \( \sum_{k=1}^{k_n} (X_{nk} - \mathbb{E}X_{nk}) \) converges in distribution to the normal distribution with zero mean and variance matrix \( \Sigma \).

**Proof.** For proof see for example Billingsley (1995).

**Lemma A.11 (Boole and Bonferroni inequalities)** Given a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) let \( A_k, k \in \mathbb{N} \) be a sequence of events \( A_k \in \mathcal{F} \). Then, for any \( n \in \mathbb{N} \) it holds

\[
\mathbb{P}\left( \bigcup_{k=1}^{n} A_k \right) \leq \sum_{k=1}^{n} \mathbb{P}(A_k), \quad (A.1)
\]

\[
\mathbb{P}\left( \bigcap_{k=1}^{n} A_k \right) \geq \sum_{k=1}^{n} \mathbb{P}(A_k) - n + 1. \quad (A.2)
\]

The proof of Lemma A.11 can be found for instance in Galambos and Simonelli (1996). The inequality in (A.1) is called the Boole inequality, while (A.2) is known as Bonferroni inequality. Confusingly, since the Boole inequality is often used to prove the validity of the Bonferroni multiple testing correction method, it is often also referred to as the Bonferroni inequality.
Bibliography


Summary

Statistical genetics is a scientific area at the intersection of genetics - the study of the genetic code inside cells of living organisms - and quantitative statistical analysis. With a strong emphasis on a mathematically rigorous formulation of the statistical methodology, the contents of this thesis lean heavily towards the latter of the two fields. In this thesis we address three general and important statistical problems highly relevant to modern genetics and numerous related fields. The common theme running through the thesis is the focus on efficient and reliable identification of rare signals (such as association of genes and phenotypes) in a large-scale setting via multi-step statistical analyses.

In Part I of the thesis we devise a novel statistical approach to simultaneous inference about the existence of links between binary phenotypes and several sections of genome that affect the phenotype as a group in a non-additive fashion, which is referred to as interaction. The methods of Part I are aimed at efficiently dealing with the multiple testing problem in a genome-wide search for interactions by employing a two-stage testing scheme which proceeds by eliminating non-promising candidates in the first step.

In Part II of the thesis we formulate a novel variant of a cost-efficient two-stage experimental design and analysis procedure designed to be used for a large-scale simultaneous inference problems in general cost-restricted settings. Gains of statistical efficiency are achieved by improved allocation of budget towards the more promising candidates. The design builds up on existing methods published in the literature and improves on a number of these methods especially in application areas with non-homogeneous sparse effects.

Finally, in Part III of the thesis we present a novel usage of penalized regression for efficient identifying of small groups of genetic loci (markers) within a much larger collection of loci based on common association with a group of numerical mutually correlated phenotypes. The idea is to use the correlation among the phenotype variables to improve the selection process over methods that treat these phenotypes as independent. For instance, such problems arise in application in behavioral data analysis with data generated via a factor model, where related quantitative measurements are used to capture a medical condition of individuals, and the goal is the discovery of links between the underlying condition and genetic loci. Since the quantitative measurements all relate to a single underlying medical condition, they often exhibit correlation and should be treated simultaneously for efficiency purposes. In addition to exploiting the phenotype correlation, via an initial stage assessment of the marginal association of each covariate with the response prior to the actual fitting of the penalized regression model the method prioritizes promising covariates at the expense of the rest.
Samenvatting

Statistische genetica is een wetenschap in zowel de genetica - de studie naar de genetische code in cellen van levende organismen - als de kwantitatieve statistiek. De nadruk in het proefschrift ligt op de laatste van de twee onderzoeksvelden. In het proefschrift besteden we aandacht aan drie statistische problemen die zowel relevant zijn voor de moderne genetica als aanverwante onderzoeksgebieden. Het belangrijkste thema van het proefschrift is de identificatie van zeldzame signalen in hoog-dimensionale vraagstellingen (zoals de associatie tussen genen en fenotype) met efficiënte en betrouwbare methoden die uit meerdere stappen bestaan.

In het eerste deel van het proefschrift beschrijven we een nieuwe methode om het bestaan van verbanden aan te tonen tussen binaire fenotypen en meerdere regio’s van het genoom die het fenotype op een niet additieve wijze beïnvloeden; zogenaamde interacties. Het doel van deze methode is om, bij de zoektocht naar interacties, op efficiënte wijze het "multiple testing" probleem toe te passen. Dit door middel van een twee-stappen toets die weinig belovende kandidaten in de eerste stap elimineert.

In het tweede deel van het proefschrift formuleren we een nieuwe variant van een kosten-efficiënte twee stappen experimenteel design voor hoog dimensionale vraagstukken, waarin de kosten beperkend zijn. Een grotere statistische efficiency ontstaat door het verbeteren van de toewijzing van budget aan veel belovende kandidaten. Het design bouwt voort op bestaande methoden uit de literatuur en verbetert een aantal van hen, met name in toepassingsgebieden met niet-homogene schaarse effecten.

In het derde deel van het proefschrift beschrijven we een nieuwe toepassing van "penalized regression" voor het efficiënt identificeren van kleine groepen loci (markers) binnen een grote groep loci, die een associatie hebben met meerdere onderling gecorreleerde fenotypen. De correlatie tussen de verschillende fenotypen wordt gebruikt om de selectie te verbeteren. Dergelijke vraagstukken komen voor in de gedragswetenschappen waarin meerdere gecorreleerde metingen worden gebruikt om een medische toestand van individuen te beschrijven, teneinde de relatie tussen de onderliggende toestand en loci op het genoom te ontdekken. Omdat de kwantitatieve metingen gerelateerd zijn aan een onderliggende medische conditie, zijn zij veelal gecorreleerd en moeten ze simultaan geanalyseerd worden vanuit een oogpunt van efficiency. Daarnaast wordt in een eerste stap de marginale associatie van elke covariaat met de uitkomst gebruikt om veel belovende covariaten meer gewicht te geven in de parameter schatting van het gepenaliseerde regressiemodel.