universität wien

MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

A comparison of Bayesian Model Selection Methods for the Analysis of Genome Wide Association Studies

verfasst von / submitted by

Michael Hagmann BSc. BSc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften
(Mag. rer. soc. oec.)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt/ degree programme code as it appears on the student record sheet:	A 066 951
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Magisterstudium Statistik
Betreut von / Supervisor:	Dipl.-Ing. Dr. Florian Frommlet, Privatdoz.

Contents

List of Figures

List of Tables

1 INTRODUCTION

Since the completion of the human genome project the capability of genotyping devices has made an impressive progress. With current state technology it is possible to collect millions of measures from a subject within a single experiment simultaneously. Consequently, this has led to experimental designs where the number of unknown features p massively outsizes the number of experimental units n. The data produced by such designs is usually called *high dimensional*. The possibility to conduct such experiments sounds like a gift, but actually the analysis of high dimensional data is extremely demanding. Developing statistical methods that are able to separate useful information from noise is a key challenge and a necessity to prove the utility of such experiments.

In the following I will present different methods and strategies to analyze data from such experiments and compare them on a real data example and in a simulation study. For this purpose, I limit the scope of my inquiry to a class of experimental designs called genome wide association study (abbreviated as GWAS). The following sections of this chapter will contain a brief exposition of the biological and statistical concepts as well as the terminology necessary to understand the purpose of a GWAS and the means and pitfalls in the involved data collection and analysis process.

1.1 Basic Biology

The exposition in Section 1.1 and 1.2 mainly follows the corresponding chapters of [15].

1.1.1 Phenotypes and Genotypes

A *phenotype* is any observable characteristic of an organism, e.g. height, muscular mass, eye color, disease status. Typically we are able to observe a significant amount of variation between individuals of the same species for a phenotype. Biologists attribute this variation to two general classes of causes. The first class of causes are dispositions within an organism. These dispositions are called *genetic*. The totality of these dispositions is called the *genotype* of an organism. The second class comprises sources that are external to an organism. Causes in this class are called environmental. The relative amount of

variability of the phenotype that can be accounted to the genotype is called the *heritability* of a trait. The way genetic and environmental causes interact in general and for specific phenotypes is a field of ongoing research, but most of the current biological research is focused on the genetic aspect.

In eukaryotes[1] most of the material that codes the genetic information is located in the cell nucleus. This material is organized in structures of deoxyribonucleic acid (DNA) which are called *chromosomes*. The DNA consists of two strands that are twisted around each other and form a double helix. Each strand is a biopolymer composed of simpler units called *nucleotides*. Each nucleotide contains a nucleobase – either cytosine (C), guanine (G), adenine (A), or thymine (T) –, a monosaccharide sugar called deoxyribose and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next. The end of a strand is called 5' if it corresponds to the phosphate group and 3' else[2]. This fact allows to introduce the notion of direction for a DNA strand. For DNA one strand runs in the direction 3'-5' and the other in the direction 5'-3'. The two strands of DNA are joined together via hydrogen bonds between the nucleobases according to the base pair rule which says that A ties with T and C pairs with G. These facts are summarized by the phrase that the two strands of DNA are antiparallel (run in opposite directions) and complementary (due to the base pair rule). For the purpose of statistical analysis a chromosome is simply represented as a sequence composed of the four letters A,C,G,T corresponding to the nucleobases of one of the complementary strings[3].

Usually these sequences vary between different subjects of the same species. Given a specific population of individuals of the same species, positions of the DNA sequence where a difference between two or more individuals can be observed represent potential sources of genetic variability that could explain phenotypic variation. These positions along with the deviations that are observed on these positions are called *genetic markers*. For the purpose of explaining a phenotype, it is sufficient to characterize the *genotype* of individuals by genetic markers. This concept is further detailed in the next paragraph.

1.1.2 Genetic Variation

The DNA sequence is almost identical between different individuals of a given species, e.g. for humans 99.9% of all base pairs match. Nonetheless, there are a large number

[1] A eukaryote is any organism whose cells contain a nucleus and other organelles enclosed within membranes. This taxon also applies to humans.

[2] This notation is derived form the carbon molecule numbering of the sugar molecule.

[3] Henceforth, I assume that the direction is the same for all individuals when I make statements involving more than one subject.

of spots in the genotype where one observes differences between individuals of the same species. These spots are called *polymorphic* loci and can extend from a single base pair location to large stretches of consecutive base pair positions. The observed variants at such a locus are called *alleles*. The most prominent examples of genetic variation are *single nucleotide polymorphisms* (SNP), *microsatellites* and *copy number variations* (CNV). A SNP is the result of a so called point mutation of a single base pair, e.g. the two sequences GATTACA and GATTATA show a SNP at their 6th position. Almost all SNPs have only two alleles, where the more frequent one is called the major allele or wild type and serves as a reference while the less frequent one is called the variant or minor allele. A microsatellite is a very short pattern of DNA ranging in length from 2 to 5 base pairs that is repeated a different number of times for individuals of the same species. They are also called short tandem repeats (STRs) or simple sequence repeats (SSRs). Copy number variations refer to long stretches of DNA (typically they range in length from 10^3 to 10^6 base pairs) which are repeated a different number of times (this is the variant of a CNV) in different individuals of the same species. In particular insertions, deletions and duplications of DNA stretches are counted as CNV.

The number of *homologous* chromosomes[4] differs between species. *Diploid* organisms (like humans) have two homologous chromosomes. So, for a diploid organism, the genotype at a given locus is represented by a pair of alleles. For example, let us consider a biallelic locus (e.g. a SNP), with alleles A and a. Then there exist three possible genotypes AA, Aa and aa. An individual carrying two identical alleles is called *homozygous*, and one that carries two different alleles is called *heterozygous*.

1.1.3 Relationship between Genotype and Phenotype

First and foremost we distinguish between causal and noncausal markers. A noncausative marker does not influence the phenotype under consideration. This means that the expected value of the phenotype given a certain manifestation of the marker equals the expectation of the phenotype for all levels of the marker M, or in short $\mathbb{E}[Y|M = m] = \mathbb{E}[Y]$ for all $m \in L$ where L denotes the set of all possible marker manifestations[5].

For causative makers geneticists distinguish between two modes of influence called *additive* and *dominant*. Let us consider a simple example to illustrate them. Assume, that we observe a quantitative trait Y and a biallelic locus M, then we define the following

[4] These are sets of chromosomes that pair up together and contain genetic information for the same feature on the same locus.

[5] We will see in Section 1.1.4, that this distinction must be slightly accommodated to be correct in reality.

expectations given a certain genotype $\mu_1 = \mathbb{E}[Y|M = AA]$, $\mu_2 = \mathbb{E}[Y|M = Aa]$ and $\mu_3 = \mathbb{E}[Y|M = aa]$. Because we assume that the marker is causative, we know that there exits a pair $i, j = 1, 2, 3$, for which $\mu_i \neq \mu_j$.

If $\mu_2 = (\mu_1 + \mu_3)/2$, which means that $\mathbb{E}[Y|M = AA] - \mathbb{E}[Y|M = Aa] = \mathbb{E}[Y|M = Aa] - \mathbb{E}[Y|M = aa]$, then the effect of the causal SNP is called *additive*. Otherwise, the marker is said to have a *dominance effect* quantified (and defined) as $\gamma := \mu_2 - (\mu_1 + \mu_3)/2$. If $d(\mu_1, \mu_2) < d(\mu_1, \mu_3)$, than A dominates a and vice versa if otherwise. For instance, in the extreme case that $\mu_1 = \mu_2$ (which means that the expected value given the genotypes AA and Aa is identical), allele A completely dominates over a. A is then called the dominant and a the recessive allele.

1.1.4 Markers and Genetic Linkage

Genetic linkage is the tendency of alleles that are close[6] together on a chromosome to be inherited together. As a result of this tendency one can observe a correlation structure between genetic markers within a population. This correlation structure is a well defined function of the distance between two loci for different kinds of experimentally produced populations but rather complicated for outbred populations. In the latter case biologist usually speak of *linkage disequilibrium* to describe the nonrandom association between two markers. Figure 1.1 on the following page illustrates a typical LD pattern.

The important consequence of linkage disequilibrium for statistical considerations is that two neighboring markers cannot be treated as stochastically independent. Therefore one might observe that a noncausal marker is associated with a phenotype due to linkage disequilibrium with a causal marker. Let us consider the following example to illustrate this.

Let Y be a quantitative trait with a continuous distribution which is causally influenced by the biallelic marker C, and let M be a noncausal biallelic marker in linkage disequilibrium with C. Then for all $m \in \{AA, Aa, aa\}$ by definition

$$\mathbb{E}[Y|M = m] = \int y f_{Y|M=m} dy.$$

The conditional density can be rewritten so that

$$f_{Y|M=m} = \frac{f_{Y,M=m}}{\mathbb{P}(M = m)} = \frac{1}{\mathbb{P}(M = m)} \sum_{c \in \{AA,Aa,aa\}} f_{Y,M=m|C=c} \mathbb{P}(C = c).$$

[6]The physical distance between two markers is expressed by the number of base pairs between them.

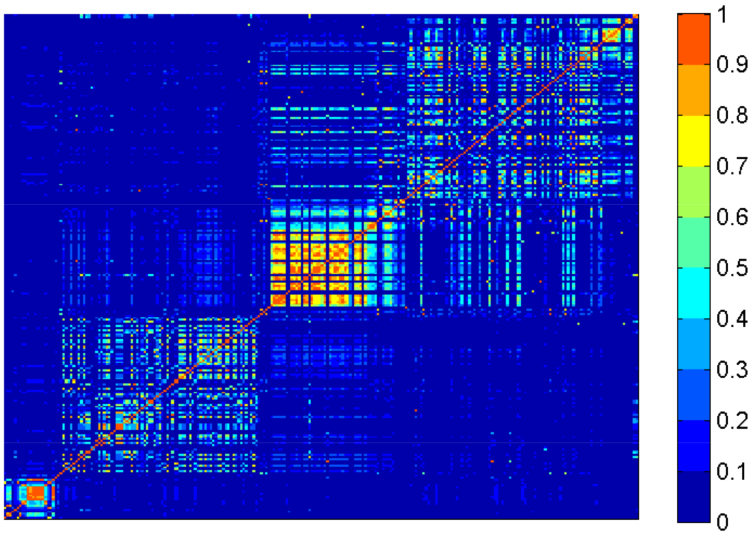

Figure 1.1: HeatMap illustrating the LD pattern for 90 individual of the CEU HapMap population. Color is coding the LD measure R^2 between the first 250 adjacent SNPs of ENCODE region ENm010 (after removing identical SNPs). This plot is taken from [15].

Because Y and M are stochastically independent given C, this simplifies to

$$\begin{aligned} f_{Y|M=m} &= \frac{1}{\mathbb{P}(M=m)} \sum_c f_{Y|C=c} \mathbb{P}(M=m|C=c)\,\mathbb{P}(C=c) \\ &= \sum_c f_{Y|C=c} \mathbb{P}(C=c|M=m)\,. \end{aligned}$$

So we get

$$\begin{aligned} \mathbb{E}[Y|M=m] &= \int y \left(\sum_c f_{Y|C=c} \mathbb{P}(C=c|M=m) \right) dy = \\ &= \sum_c \mathbb{E}[Y|C=c]\,\mathbb{P}(C=c|M=m)\,. \end{aligned}$$

This equivalence neatly displays the indirect mechanism we assume. We see that the indirect effect of M is a weighted average of the causal marker effects, so that for the considered scenario M mediated threw C could appear to be a causal SNP.

This observation combined with the roughly block-like structure of linkage disequilibrium (shown in Figure 1.1) establishes the core principle of association studies in outbred populations. The aim of genome wide association studies is not to track down a causal marker explicitly but rather to detect a DNA region which hosts such a marker. As we

can now see, it is sufficient to operate with a (well chosen) subset of all possible markers in order to pursue this goal. This observation is important from an experimental point of view, because even nowadays array based genotyping devices are only capable to cover about 20% of the human genome within a single experiment. This has changed with next generation sequencing, but this is currently still much more expensive compared to array based genotyping.

1.2 Genome Wide Association Studies

A fundamental genetic premise in the study of phenotype genotype relationships is the so called *common disease*[7]*-common variant* assumption. This postulate states that the markers causing a disease will be found in all populations of a species which manifest that disease and that each marker influencing a disease will have a small additive (or multiplicative) effect on the expression of the disease. A usual addendum to this postulate is the assumption that the number of causal markers is moderate. This means that the number is larger than one, but significantly less than the total number of markers. A trait with this property is called a *complex trait*. This is the kind of phenotype usually addressed by GWAS.

Genome wide association studies have become feasible with the development of SNP array technology. With a contemporary array it is possibly to measure up to 4 million markers of an individual within a single experiment simultaneously. In order to perform this task, SNP arrays utilize the mechanism of hybridization[8]. In principle an array consists of a carrier (e.g. a glass slide) with probes of short single stranded DNA samples (20-60 nucleotides long) mounted on it's surface. With current manufacturing technology is possible to place these probes very close together. The distance between two neighboring probes is usually just a few micrometers. To be able to determine which allele of a marker is present in an individual, probes[9] for every allele are placed on the device. The quantity which is measured is the relative amount of sample DNA (target) that hybridize on the alleles of a given locus[10]. From this information, one can infer which type of allele

[7] In the current context disease is synonymous with phenotype.

[8] Which means that complementary strains of DNA bind together according to the base pair rule.

[9] Usually the magnitude of replicates of a probe are of size 10^3. Identical probes are located in clusters on the array. The terminus technicus for these clusters are spots. The spatial location of these spots is well defined for an array, so that we have a fixed correspondence between spatial coordinates and markers.

[10] This measurement is usually done indirectly with the following method. Before a sample of DNA is exposed to an array, the DNA sample taken from an individual underwent a special preparation. During this preparation the DNA is broken in small pieces, such that it fits the length of the chip probes, and amplified, which means that replicates of these pieces are produced. Finally a fluores-

is present at a locus. For example, consider a biallelic locus. If all DNA binds at probe A the conclusion would be that the individual carries the homozygous genotype AA. If half of the DNA binds at probe A and the other half at probe a the individual would be heterozygous at that locus. In practice this process, called *genotype calling*, involves a lot of statistical techniques due to imperfections in the technical process and the process of hybridization itself. Even very sophisticated algorithms are usually not able to decide unambiguously upon a genotype of a SNP for all markers and for all array chips in a study. As a consequence the analyst of such data have to deal with the problem of missing values. This problem is of special importance if the analyst wants to apply more advanced methods which assume complete data.

1.2.1 Imputation

It is a common place within Imputation Theory to distinguish between missing values due to random causes and missing values due to systematic causes. A systematic cause could be for example, that the genotype calling failed for a particular marker for a large percentage of individuals in a study. Then this marker will be excluded from further analysis. A systematic missing pattern is also given if the calling quality of all markers on a particular chip is poor due to problems during the processing of the array in the lab. Then the entire data from this chip will be excluded from further investigations. On the other hand if missing values seem to be sporadic, the application of imputation algorithms permits inference of the missing genotype states.

Imputation methods base their estimates on two sources of information, namely the genotype data available from the study sample and data published in reference panels. The conceptual complexity of these methods ranges from very simple to highly complex. More advanced methods also go beyond the task of simply filling up random gaps in the data, and try to infer the genotype of markers not even present in the study data.

1.2.2 Models and Strategies for Analyzing GWAS Data

Currently the most common strategy applied to analyze GWAS data are single marker tests [18]. There exists a variety of tests developed in this context for both quantitative and dichotomous[11] phenotypes. But for convenience and flexibility it is useful to treat

cent dye is attached to one end of the probes. After the sample was presented to the array and hybridization took place, the dye is excited and the emitted light is collected via a photo sensitive chip. This image is the source data for genotype inference. Usually array manufactures provide a software together with a chip so that this image data can be easily transformed into probe intensities.

[11] An example of such a phenotype is the presence or absence of a certain disease. Such studies are usually called case-control study.

both cases under the unifying frame of generalized linear regression models [21, 25].

In its broadest generalization a regression function links a particular deterministic feature of the random variable Y to some determining factors X. Mathematically this is expressed by

$$C\left(Y|X\right) = g(X).$$

Restricting the regression function $g\left(X\right)$ to be a function η of a linear combination of X we arrive at the class of generalized linear regression models (GLM) with the mathematical representation

$$C\left(Y|X\right) = \eta\left(X\beta\right),$$

where η is called the link function. Examples for this class are the well known linear and logistic regression models. The former is commonly expressed as

$$\mathbb{E}\left[Y|X\right] = X\beta$$

with the additional assumption that $Y \sim N\left(X\beta, \sigma^2 I\right)$. For the latter, let $Y \in \{0,1\}$ be a Bernoulli distributed random variable with $p \coloneqq \mathbb{P}\left(Y = 1|X\right)$ and η the logit function, then we arrive at the familiar binary logistic regression model

$$\mathbb{E}\left[Y|\mathrm{X}\right] = \mathbb{P}(Y = 1|X) = \mathrm{logit}\left(X\beta\right).$$

In the context of association studies it is clear that Y represents the phenotype under investigation and X the genotype. In this case the restriction to a linear model correspond to the notion expressed in the common disease - common variant postulate that the individual markers cause the phenotype independent of each other and that their contribution is additive with the additional assumption that there is no epistasis[12].

In Section 1.1.3 we pointed out that geneticists distinguish between several types of relationship between genotype markers and phenotypes. To incorporate these patterns into statistical models it is necessary to accommodate the coding of markers. For illustrative purposes let us consider a biallelic marker, where A represents the major allele. First assume that the marker has an additive effect. As already pointed out this means that $\mathbb{E}\left[Y|M = AA\right] - \mathbb{E}\left[Y|M = Aa\right] = -c$ and $\mathbb{E}[Y|M = aa] - \mathbb{E}\left[Y|M = Aa\right] = c$ where c is some constant. This means that every occurrence of an allele a adds c to the effect

[12] Following [11] epistasis describes the situation that the effect of one marker depends on the manifestation of another maker. This kind of effect regulation could be modeled as a (multiplicative) interaction term in a regression model. So given m markers there are $C(m,2)$ possible two way interaction terms, where $C(n,k)$ denotes the binomial coefficient n over k. The effect regulation could be even more complex and include k makers, then there are $C\left(m,k\right)$ interaction terms.

of the marker. We can easily express an additive effect in a linear model by introducing the following coding $G \coloneqq \{AA = -1, Aa = 0, aa = 1\}$. The resulting model (including an intercept term) will be

$$\mathbb{E}[Y|M] = \beta_0 + \beta_1 G(M).$$

It follows that $\beta_0 = \mathbb{E}[Y|M = Aa]$ and $\beta_1 = c$. Thus additive effects can be seamlessly expressed in a GLM, but what about dominance?

As we have seen in Section 1.1.3 dominance is defined as a directed shift γ of the location of $\mathbb{E}[Y|M = Aa]$ away from the midpoint between $\mathbb{E}[Y|M = AA]$ and $\mathbb{E}[Y|M = aa]$. Formally this means that $\mathbb{E}[Y|M = AA] - \mathbb{E}[Y|M = Aa] = -c + \gamma$ and $\mathbb{E}[Y|M = aa] - \mathbb{E}[Y|M = Aa] = c + \gamma$ where c and γ are some constants. In order to adjust our GLM for dominance effects we can incorporate another term $D \coloneqq G^2 = \{AA = 1, Aa = 0, aa = 1\}$. The resulting model will be

$$\mathbb{E}[Y|M] = \beta_0 + \beta_1 G(M) + \beta_2 G^2(M).$$

Following our previous considerations it is obvious that $\beta_2 = \gamma$.

Now, we have seen that GLMs provide all the means to model the phenotype genotype relation in a very straight forward way. Besides, thanks to the well developed statistical theory for GLMs [25], we can also construct statistical tests for formally testing whether a marker influences a phenotype.

An advantage of GLMs over single marker tests is that they allow multimarker analysis. Because most GWAS are conducted to study complex traits, which are by definition influenced by a large number of markers, GLMs open the possibility of a more powerful way to analyze GWAS data.

That this assertion is indeed true has been proven by Frommlet et al. [18]. I will briefly repeat this argument. Let y_i, $i \in \{1, ..., n\}$ denote measurements of a normally distributed quantitative trait with $\left(\mu, \tau^2\right)$ taken from n individuals. Furthermore, let x_{ij} represent the genotype of SNP j from individual i, where $j \in \{1, \ldots, p\}$. Assume that $p \gg n$, and that $J^* = \{j_1, \ldots, j_k\}$ denotes the set of causal SNPs for this phenotype. If we further assume that the joint effect of this SNPs is a simple linear combination of the individual effects, then the true model for the phenotype expression of individual i will be

$$M_{J^*} : y_i = \beta_0 + \sum_{l \in J^*} \beta_l x_{il} + \epsilon_i,$$

where $\epsilon_i \sim N\left(0, \sigma^2\right)$ summarizes the effect of environmental causes. If this data is analyzed with single marker tests, then each marker j is analyzed without considering

any of the other ones. Hence, this test strategy limits the scope of explanatory models to those of the form

$$M_j : \; y_i = \beta_0^{(j)} + \beta_1^{(j)} x_{ij} + \epsilon_i^{(j)}.$$

The usual F-test for M_j compares the proportion of genotype variance explained by SNP j compared to the proportion left unexplained, thus it could be used to detect causal SNPs. Based on least squares regression the F-statistic for SNP j is calculated as

$$F_j = (n-2)\,\frac{MSS_j}{RSS_j},$$

where $RSS_j = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0^{(j)} + \hat{\beta}_1^{(j)} x_{ij}\right)\right)^2$ and $MSS_j = \sum_{i=1}^{n} (y_i - \bar{y})^2 - RSS_j$. For both statistics Frommlet et al. derived the distribution under the causal model. This allows to calculate the power of the tests based on F_j. So far the setting of the example requires just the common assumptions that underlie genome wide association studies. To simplify things for exposition, let us additionally assume that the SNPs in M_{J^*} are orthogonal[13] and that the squared scaled effect size $\tau = \frac{n\beta_l^2}{\sigma^2}$ is equal for all $l \in J^*$. With this additional assumptions the distributions for RSS_j and MSS_j are independent and

$$RSS_j \sim \sigma^2 \chi^2\,(n-2, (k-1)\,\tau)$$

$$MSS_j \sim \sigma^2 \chi^2\,(1, \tau)$$

Where $\chi^2\,(df, ncp)$ denotes the noncentral chi-squared distribution with df degrees of freedom and noncentrality parameter ncp. Large values of F_j indicate a significant deviation from the null hypothesis that $\beta_1^{(j)} = 0$. Given a sample size n, in order that F_j is large either RSS_j must be small or MSS_j must be large. For any τ the distribution of MSS_j is fixed but the distribution of RSS_j depends on the true number of causal SNPs. The larger k the larger will be the probability for large RSS_j values. Hence, the power of a test based on F_j decreases uniformly over τ with increasing k. The results of an illustrative simulation study are summerized in Figure 1.2 on the next page, which clearly demonstrate the poor power of single marker tests in the face of a complex trait even when the (scaled quadratic) effect size is rather large. Even if the assumption of orthogonality is hardly met in a GWAS and it is also quite artificial to assume that all causal SNPs have equally strong influence the main conclusion of these deliberations should hold true, anyway.

Therefore single marker tests are not particularly well suited to analyze GWAS for

[13] This means that $[\mathbf{1}, \mathbf{x}_{\cdot j_1}, \ldots, \mathbf{x}_{\cdot j_k}]^{'}\,[1, \mathbf{x}_{\cdot j_1}, \ldots, \mathbf{x}_{\cdot j_k}] = n\mathbf{I}_{k+1}$

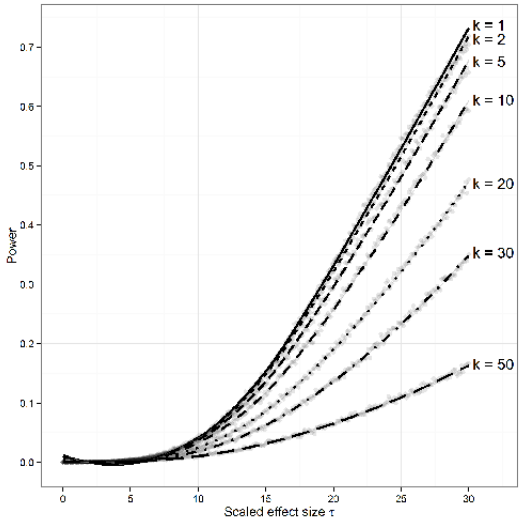

Figure 1.2: Power to detect a causal SNP with single marker tests, when k SNPs are causal. Other simulation parameters are $n = 2000$ and $\alpha = 10^{-6}$.

complex traits and in order to increase the power of the involved statistical tests one should consider multi marker models. Given that the analyst of a GWAS usually doesn't know which SNPs are good proxies for the causative SNPs of a trait, this naturally raises the problem to find the causal SNPs. Hence, we have to deal with a model selection problem.

1.2.3 Sample Considerations and Population Structure

Ideally GWAS are based on a large number of unrelated individuals, where unrelated means that the relationship between the individuals is distant enough that no linkage due to relatedness between them is observed. This assumption is important for many statistical tests. If this assumption is violated the properties of these tests can not be guaranteed.

Another fundamental assumption in association studies is random mating. This means that all individuals in a population are potential partners. In reality this assumption is hardly met and one is confronted with a *population structure*. This fact can have severe impact on the conclusions drawn from a GWAS, if the analysis is not appropriately adjusted. To illustrate the distorting impact of population structure on OLS-based inference I will first discuss the consequences in general followed by presenting a remedy for regression models.

1.2.3.1 The Impact of Population Structure on OLS Estimator based Inference

Let us assume the standard situation for estimating a multiple linear regression model, where Y is a n-vector, X a $(n \times p)$-matrix and β a p-vector. Further let $\mathbb{E}[Y|X] = X\beta$ and $Var(Y|X) = \sigma^2 I_n$ where I_n is the $(n \times n)$-identity matrix. Then we know that the OLS estimators for β and σ^2 are

$$\hat{\beta}_{ols} = \left(X^t X\right)^{-1} X^t Y$$

respectively

$$\hat{\sigma}^2_{ols} = \frac{Y^t Y - Y^t X \left(X^t X\right)^{-1} X^t Y}{n-p} = \frac{1}{n-p} \operatorname{tr}\left(\left(I_n - X\left(X^t X\right)^{-1} X^t\right) Y Y^t\right),$$

with conditional expected values of

$$E\left[\hat{\beta}_{ols}|X\right] = \left(X^t X\right)^{-1} X^t \mathbb{E}[Y|X] = \beta$$

and

$$\mathbb{E}\left[\hat{\sigma}^2_{ols}|X\right] = \frac{1}{n-p} \operatorname{tr}\left(\left(I_n - X\left(X^t X\right)^{-1} X^t\right) Var(Y|X)\right) = \sigma^2.$$

The conditional bias of an estimator $\hat{\theta}$ for the true parameter θ is defined as

$$\mathbb{B}\left[\hat{\theta}|X\right] := \mathbb{E}\left[\hat{\theta}|X\right] - \theta.$$

Hence, both estimators are unbiased under these assumptions. Thus,

$$\hat{Var}\left(\hat{\beta}_{ols}|X\right) = \hat{\sigma}^2_{ols}\left(X^t X\right)^{-1}$$

is also an unbiased estimator of the variance of $\hat{\beta}_{ols}$ in this setting.

So far we have only considered the case that the outcomes are uncorrelated. Now, we want to investigate the behavior of $\hat{\beta}_{ols}$, $\hat{\sigma}^2_{ols}$ and $Var\left(\hat{\beta}_{ols}|X\right)$ when the outcomes are correlated but we ignore this while we construct our estimators. Following Stram [34], we assume that the covariance between the observations is $Var(Y|X) = \sigma^2 I_n + \gamma^2 K$, where K is positive definite. Then $\hat{\beta}_{ols}$ is still conditional unbiased.

Now, let us calculate the conditional bias of the variance of $\hat{\beta}_{ols}$. Under the actual

model the true variance of our estimator is

$$\begin{aligned} Var\left(\hat{\beta}_{ols}|X\right) &= \left(X^tX\right)^{-1}X^tVar\left(Y|X\right)X\left(X^tX\right)^{-1} \\ &= \left(X^tX\right)^{-1}X^t\left(\sigma^2I_n+\gamma^2K\right)X\left(X^tX\right)^{-1} \\ &= \sigma^2\left(X^tX\right)^{-1}+\gamma^2\left(X^tX\right)^{-1}X^tKX\left(X^tX\right)^{-1} \end{aligned}$$

and the conditional expectation of the variance estimator will be

$$\begin{aligned} \mathbb{E}\left[\hat{Var}\left(\hat{\beta}_{ols}|X\right)|X\right] &= \mathbb{E}\left[\hat{\sigma}^2_{ols}|X\right]\left(X^tX\right)^{-1} \\ &= \frac{1}{n-p}\operatorname{tr}\left(\left(I_n-X\left(X^tX\right)^{-1}X^t\right)Var\left(Y|X\right)\right)\left(X^tX\right)^{-1} \\ &= \left[\sigma^2+\frac{\gamma^2}{n-p}\left(\operatorname{tr}\left(K\right)-\operatorname{tr}\left(\left(X^tX\right)^{-1}X^tKX\right)\right)\right]\left(X^tX\right)^{-1}. \end{aligned}$$

So, $\mathbb{B}\left[\hat{Var}\left(\hat{\beta}_{ols}|X\right)|X\right]$ is

$$\gamma^2\left(X^tX\right)^{-1}\left[\frac{1}{n-p}\left(\operatorname{tr}\left(K\right)-\operatorname{tr}\left(\left(X^tX\right)^{-1}X^tKX\right)\right)I_p-X^tKX\left(X^tX\right)^{-1}\right].$$

For illustrative purposes let us now consider the special case that $p=2$ and that $X=[\mathbf{1},\mathbf{x}]$, where $\mathbf{1}$ denotes the n-vector whose components are all 1 and $\mathbf{x}$ is a n-vector that represents the variable of interest. Without loss of generality let $\mathbf{x}$ be centered, this means that $\mathbf{1}'\mathbf{x}=0$. Under this additional assumptions the conditional bias for the slope parameter, which is the (2,2)-element of $\mathbb{B}\left[\hat{Var}\left(\hat{\beta}_{ols}|X\right)|X\right]$, reduces to

$$\frac{\gamma^2}{\mathbf{x^t}\mathbf{x}}\left[\frac{\operatorname{tr}(K)}{n-2}-\frac{\mathbf{1}^tK\mathbf{1}}{n\left(n-2\right)}-\frac{\mathbf{x}^tK\mathbf{x}}{\mathbf{x^t}\mathbf{x}}\left(1+\frac{1}{n-2}\right)\right].$$

For large n this expression is approximately

$$\frac{\gamma^2}{\mathbf{x^t}\mathbf{x}}\left[\frac{\operatorname{tr}(K)}{n}-\frac{\mathbf{1}^tK\mathbf{1}}{n^2}-\frac{\mathbf{x}^tK\mathbf{x}}{\mathbf{x^t}\mathbf{x}}\right].$$

If we look at the bias as a function of $\mathbf{x}$ we see that the expected bias is most severe when the quadratic form $\mathbf{x}^tK\mathbf{x}$ (s.t $\mathbf{x}^t\mathbf{x}=const$) reaches an extremum. The maximizer for this problem is the eigenvector which correspondents to the largest eigenvalue and the minimizer is the eigenvector that corresponds to the lowest eigenvalue. We can also see, that the bias is negative – which means that the estimator on average underestimates the true variance – when $\frac{\mathbf{x}^tK\mathbf{x}}{\mathbf{x}'\mathbf{x}}>\frac{tr(K)}{n}-\frac{\mathbf{1}^tK\mathbf{1}}{n^2}$. This systematic under-

estimation of $Var\left(\hat{\beta}_{ols}|X\right)$ will lead to anti-conservative tests and confidence intervals. One consequence of this is an increased probability to detect false positive signals. On the other hand if $\frac{\mathbf{x}^t K\mathbf{x}}{\mathbf{x}'\mathbf{x}} < \frac{tr(K)}{n} - \frac{\mathbf{1}^t K\mathbf{1}}{n^2}$, then the estimator systematically overestimates $Var\left(\hat{\beta}_{ols}|X\right)$ which in consequence leads to over-conservative tests. A result of this is a decreased power to detect true positive signals. Obviously, neither one is a favorable situation.

So far we have only looked at a specific realization of $\mathbf{x}$. Let us now consider the case that $\mathbf{x}$ is a random vector with $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] = \tau^2 K$[14] where all entries of K are greater or equal to 0 and let λ denote the vector whose components are the eigenvalues of K. Then the expected value of the bracket multiplied by $\mathbf{x}^\mathbf{t}\mathbf{x}$ of the approximate slope parameter bias will be

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\frac{\mathrm{tr}(K)}{n}\mathbf{x}^\mathbf{t}\mathbf{x} - \frac{\mathbf{1}^t K\mathbf{1}}{n^2}\mathbf{x}^\mathbf{t}\mathbf{x} - \mathbf{x}^t K\mathbf{x}\right)\right] &< \frac{\mathrm{tr}(K)}{n}\,\mathrm{tr}\left(E\left[\mathbf{x}\mathbf{x}^\mathbf{t}\right]\right) - \mathrm{tr}\left(K\mathbb{E}\left[\mathbf{x}\mathbf{x}^\mathbf{t}\right]\right) \\ &= \tau^2\left(\frac{tr(K)^2}{n} - \mathrm{tr}\left(KK\right)\right)\end{aligned}$$

Because the trace is equal to the sum of the eigenvalues of a matrix, and the eigenvalues of KK are the squared eigenvalues of K the last expression equals

$$\tau^2\left[\frac{\|\lambda\|_1^2}{n} - \|\lambda\|_2^2\right].$$

Using the fact $\|\lambda\|_1 \leq \sqrt{n}\|\lambda\|_2$, we can see that the expected value of this condition is smaller than zero. Hence, we are expecting that the estimator underestimates the variance of $\hat{\beta}_{ols}$ and we are in a situation where we no longer control the Type I Error probability and must fear potentially many false positive results.

In this section I have only examined the effect of a misspecified dependence structure of the outcome for single marker tests based on OLS regression, but the results carry over to multimarker models and GLMs [34].

At first the assumption that the dependence structure of Y could be written as $Var(Y|X) = \sigma^2 I_n + \gamma^2 K$ seems arbitrary, but it is actually highly appropriate to model deviations from the homogenous population and random mating assumptions such as relatedness, hidden non-mixing populations, incomplete admixture or other factors that act as a confounder for genetic association [34]. But this assumption also makes sense from a genetic point of view. As mentioned in Section 1.1.1 geneticists postulate that the

[14] This assumption is met for instance for hidden non-mixing strata when the Balding-Nichols model holds true.

variation of a phenotype can be explained either by environmental factors or by genotype variation. If we add the common assumption that these sources are stochastically independent, then the proposed dependence structure (where K represents the genetic similarity between the sampled individuals) reflects those considerations.

1.2.3.2 Regression based Correction for Population Structure

One consequence of the results of the last section is that the distortion induced by a population structure is not homogeneous among all SNPs. SNPs that are located in the span of the largest[15] eigenvectors of K will show the most severe distortion of their test statistics. One of the most common methods to deal with this problem was suggested by Price [28]. Essentially he proposed the following five step procedure

1. Find an estimator $\hat{K}$ for K.
2. Compute the eigenvector/eigenvalue pairs of $\hat{K}$.
3. Select the l eigenvectors with the l largest eigenvalues.
4. Transform the phenotype data vector y and all marker data vectors x_k such that they are orthogonal to the l selected eigenvectors. This is usually achieved by calculating the residuals of a linear regression of these vectors on the selected eigenvectors.
5. Calculate the statistics based on the residual vectors.

In the context of linear multiple regression step 4 can be incorporated in the OLS-estimator of the parameter vector β [26] by considering the model

$$\mathbb{E}\left[Y|X,E\right] = X\beta + E\gamma,$$

where E is the matrix whose columns are the l selected eigenvectors of $\hat{K}$. In this way the resulting test statistics for $\hat{\beta}$ are adjusted for the population structure. So it is easy to incorporate Prices' approach in the context of regression models. What is left is the specification of an estimator for K, a computational efficient method to compute the eigenvalue/eigenvector pairs of it and a method to choose l. Price [28] proposed the correlation matrix between individuals based on the observed marker phenotypes as an estimator for K. Hence

$$\hat{K} \coloneqq \frac{1}{M} X_s X_s^t,$$

[15] Here large refers to an ordering induced by the eigenvalues of the eigenvectors.

where X_s is a $(N \times M)$-matrix whose columns are the standardized genotypes for each individual. Because $\hat{K}$ is a correlation matrix the tasks stated in step 2 an 3 are identical with the computations needed to perform principal components analysis (PCA). This is a standard procedure implemented in a lot of software packages. Once the eigenvalue/eigenvector pairs for the largest eigenvalues are computed one can for instance use a screeplot to choose l.

1.3 Statistical Methods for Highdimensional Data Analysis

1.3.1 Multiple Testing

The theory of multiple testing, or multiple comparison as it is sometimes called, is concerned with the problem of testing $m > 1$ hypotheses in a sample simultaneously. To formalize the problem let us consider the situation where m tests are performed for a given sample X with corresponding pairs of null hypothesis $H_0^{(j)}$ and alternative hypothesis $H_A^{(j)}$ for $j = 1, \ldots, m$. Let $m_0 \leq m$ denote the number of tests for which the null hypothesis is correct. Further, let the test decisions for each of the m tests be based on the corresponding test statistic $T_1(X), T_2(X), \ldots, T_j(X), \ldots, T_m(X)$. Statistical tests are constructed in such a way that the null and the alternative hypothesis are mutually exclusive. Consequently a statistical test can yield one of the following four possible results:

1. H_0 is true and the test accepts H_0.
2. H_0 is true but the test rejects H_0 (this is called a Type I error).
3. H_A is true but the test accepts H_0 (this is called a Type II error).
4. H_A is true and the test rejects H_0.

Table 1.1 provides the standard notation to summarize the outcome of m test results. In this table V denotes the number of Type I errors that have happened and T the number of Type II errors. Naturally we would like both numbers to be as small as possible, but unfortunately this optimal situation is not achievable with a finite sample size because the probability of a Type I error and the probability of a Type II error are antagonistically related for a statistical test. In concreto this means that a low Type I error probability necessarily leads to an increased Type II error probability and vice versa. Usually the Type I error probability is fixed at a certain level $\alpha \in (0, 1)$ (named the α-level of the test) and one chooses or constructs a test for a certain situation such

	H_0 accepted	H_0 rejected	Total
H_0 is true	U	V	m_0
H_A is true	T	S	$m - m_0$
Total	$m - R$	R	m

Table 1.1: Notation for multiple testing

that the Type II error probability is as small as possible, or equivalently that the power of the test – defined as the probability to reject the null hypothesis when the alternative is true – is maximized. So the most important property of a statistical test is its ability to control the Type I error probability at α, which means that $\mathbb{P}(reject\, H_0) \leq \alpha$ where $\mathbb{P}$ denotes the probability measure under H_0. An important question is how to generalize this property, the ability to control the probability of a certain misjudgment, to the multiple testing situation.

A large number of measures have been suggested for this purpose [23]. Common generalizations are the *per-family error rate* $PFER := \mathbb{E}[V]$, the *per-comparison error rate* $PCER := {}^{\mathbb{E}[V]}/_{m}$ and most important the *family wise error rate* $FWER := \mathbb{P}(V > 0)$ which is the probability to make at least one Type I error within the family of m tests. Unfortunately, for very large m, controlling the FWER at an acceptable level leads to procedures with very low power to detect true signals in the data. This implies that such a procedure will miss a lot of true signals and therefore have a high probability to generate Type II errors.

One of the most influential innovations in multiple testing was the rediscovery and popularization of the *False Discovery Rate* (FDR) as a generalized measure of Type I error by Bejamini and Hochberg [3]. The FDR is formally defined as

$$FDR := \mathbb{E}\left[{}^{V}/_{R}\right] \text{ with } {}^{V}/_{R} = 0 \text{ if } R = 0$$

or equivalently $FDR = \mathbb{P}(R > 0)\,\mathbb{E}\left[{}^{V}/_{R} | R > 0\right]$. So, as we can easily see, the FDR is designed to control the expected proportion of incorrectly rejected null hypotheses among the rejected null hypotheses. Under the total null, which means that $m_0 = m$, it holds that ${}^{V}/_{R} = 1$ and consequently $\mathbb{E}\left[{}^{V}/_{R} | R > 0\right] = 1$ whenever $R > 1$, so that the FDR coincides with FWER in this situation. If $m_0 < m$, than $0 < \mathbb{E}\left[{}^{V}/_{R} | R > 0\right] < 1$. Therefore the FWER is strictly bigger than the FDR. So FDR is a less stringent generalized Type I error rate than the FWER, which allows for a potential gain in power. In the subsequent sections the most prominent procedures for controlling the FWER and the FDR are presented.

1.3.1.1 Controlling the Family Wise Error Rate

There exists a huge variety of procedures that guarantee control of FWER in different multiple testing situation. I limit my exposition to the most popular procedure, namely the *Bonferroni correction* and two of its variants the *Bonferroni-Holm step-up* and the *Hochberg step-down* procedure.

Let us first define the event $B_j \coloneqq \left\{H_0^{(j)} \; rejected\right\}$. For the corresponding test statistic T_j and an associated critical value T_{crit,α_j} an equivalent characterization is $B_j = \left\{T_j \geq T_{crit,\alpha_j}\right\}$. By definition the critical value for a test statistic is chosen such that $\mathbb{P}(B_j) = \alpha_j$ where α_j is the α-level of the jth-test. Without loss of generality we assume that hypotheses are ordered, so that the null hypothesis is true for the first m_0 hypotheses. Thus, we can write

$$FWER = \mathbb{P}(V > 0) = \mathbb{P}\left(\bigcup_{j=1}^{m_0} B_j\right).$$

It follows from the sub-additivity of the probability measure that

$$FWER \leq \sum_{j=1}^{m_0} \mathbb{P}(B_j) = \sum_{j=1}^{m_0} \alpha_j \leq \sum_{j=1}^{m} \alpha_j.$$

Therefore, whenever $\sum_{j=1}^{m} \alpha_j$ is bounded by some $\alpha \in (0,1)$ then also the FWER for the m simultaneous tests is bounded by α. This fact is exploited by the Bonferroni procedure. The Bonferroni correction suggests to choose the individual $\alpha_j \in (0,1)$ such that $\sum_{j=1}^{m} \alpha_j = \alpha$ for some predefined $\alpha \in (0,1)$ and reject $H_0^{(j)}$ whenever $p_j \coloneqq \mathbb{P}(T_j \geq t_{j,obs}) \leq \alpha_j$ where $t_{j,obs}$ is the observed test statistic for test j. The standard choice is $\alpha_j = \alpha/m$.

We have seen that the argument behind the Bonferroni correction makes no use of the actual distribution of $(T_1, T_2, \ldots, T_m)$. Therefore this procedure has the favorable property that it guarantees FWER control for all possible joint distributions of the test statistics, but on the other hand the actual FWER may be much smaller than the nominal α. For instance this is the case when $m_0 \ll m$ or when the test statistics are positively correlated. A multiple testing procedure with this property is called *conservative*. This property is typically associated with a reduced power to detect true signals in the data.

An improvement (in terms of a power gain) of the Bonferroni correction is the Bonferroni-Holm procedure, which results from applying the *closed testing principle* [24]. Let $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[j]} \leq \cdots \leq p_{[m]}$ be the ordered sequence of p-values obtained by

the m individual tests and $H_0^{[j]}$ the corresponding sequence of null hypotheses. Let k be the smallest index j such that $p > \alpha/(m+1-j)$. Then the Bonferroni-Holm procedure rejects all m hypotheses if no such k exists and rejects all $H_0^{[j]}$ with $j < k$ otherwise.

In contrast to the Bonferroni adjustment the acceptance or rejection of a particular hypothesis $H_0^{(j)}$ depends on the value of all other test statistics T_i for $i \neq j$. The benefit of this is an enlarged rejection region and thus an increased power compared to the Bonferroni adjustment.

The Bonferroni-Holm procedure is an example of a so called *step-down procedure*. Stepwise procedures are characterized by the fact, that they make test decisions based on an ordered sequence of p-values. Step-down procedures start from the smallest one, each time checking if a condition is satisfied and stop the first time this condition is met. Then all null hypotheses with a smaller index than the stopping index are rejected. For each step-down procedure one can define a corresponding *step-up procedure*. In contrast to a step-down procedure the step-up procedure starts with the highest p-value and compares p-values with a certain criterion in a descending order. The procedure stops when the condition is not satisfied for the first time and rejects all null hypotheses with an index smaller or equal to the stopping index. Let k_{down} and k_{up} denote the corresponding stopping indices for both procedures, then it always holds that $k_{up} \geq k_{down}$. Therefore a step-up procedure is at least as powerful as its corresponding step-down procedure, but usually it is more difficult for a step-up procedure to control the FWER at a nominal level [15].

The step-up procedure which corresponds to the Bonferroni-Holm procedure is called *Hochberg procedure*. Hochberg demonstrated [22] that this procedure controls the FWER at a nominal level whenever the Simes inequality (see [31] for definition) holds, which is not always the case.

1.3.1.2 Controlling the False Discovery Rate

In a seminal paper Benjamini and Hochberg [3] discussed the following step-up procedure to control the FDR which was introduced by Simes in [33]. Named after the former authors this method is called the *Benjamini-Hochberg procedure*, and is one of the most commonly applied procedures to control the FDR.

Let $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[j]} \leq \cdots \leq p_{[m]}$ be the ordered sequence of p-values obtained by m individual tests and $H_0^{[j]}$ the corresponding sequence of null hypotheses. Further, let k denote the largest index j such that $p_{[j]} \leq j\alpha/m$, and reject all $H_0^{[j]}$with $j \leq k$.

Benjamini and Hochberg proved under the assumption of independent test statistics that this procedure controls the FDR at a level of $(m_0/m)\,\alpha$. Benjamini and Yekutieli

[4] also proved that this procedure controls the false discovery rate at the same nominal level when the test statistics T_j are positive regression dependent. In the same article the authors also propose a conservative modification of the Benjamini-Hochberg procedure that controls the FDR at a given nominal level for any form of dependence structure between the test statistics.

1.3.1.3 Bayes Oracle

A different perspective on multiple testing is offered when Bayesian statistical decision theory is employed to formalize and analyze this problem.

Basically we can partition the simultaneous tests in two classes, the class C_0 of tests for which the null hypothesis is true and the class C_A of tests for which the alternative hypothesis is true. Let p denote the fraction of tests that belongs to C_A. Let F_0 denote the distribution function for the test statistics when the test belongs to C_0 and F_A denote the distribution function for the test statistic when the test belongs to C_A. So the test statistics T_j of the simultaneous tests are distributed according to the following mixture

$$T_j \sim (1-p)\, F_0 + pF_A$$

Let us also define a loss-function that assigns the loss δ_I to a Type I Error, the loss δ_{II} to a Type II Error and loss 0 in case of a correct test decision. The total loss for the m simultaneous tests is simply defined as the sum of losses over all tests. If $\delta_I = \delta_{II} = 1$, then the total loss is identical to the number of misjudgments. The function that minimizes the expected total loss R (which is called *risk*) is called the Bayes classifier or *Bayes oracle*. In practice this optimal test procedure can not be applied, because it is based on unknown quantities and thus the risk of any feasible procedure is always bigger than the Bayes risk R_{opt} (which is the risk of the Bayes oracle). But it can serve as a theoretical benchmark measure.

Frommlet et al. [16] called a procedure *asymptotically Bayes optimal under sparsity* (ABOS) if the ratio R/R_{opt} converges to 1 under an asymptotic scheme satisfying $m \to \infty$, $n_m \to \infty$, $p_m \to 0$ and $2\log(p)/n \to C \in [0, \infty)$ (for fixed losses). Because there exists more than one possible asymptotic regime that satisfies the previous conditions a procedure is always ABOS with respect to a specific regime.

In the same article Frommelt et al. derive conditions under which the Bonferroni correction with FWER level α and the Benjamini-Hochberg procedure with FDR level α are ABOS. They proved that α could be kept constant only when $n \propto \log(m)$. When n increases faster, than α_n must converge to zero at $\mathcal{O}(1/\sqrt{n})$. Frommlet et al also

demonstrated that given $n \propto \log(m)$ the Bonferroni correction is ABOS just in the case that $p \propto 1/m$, while the FDR is ABOS when $p \propto m^{-\beta}$ for any $\beta \in (0, 1]$.

1.3.2 Model Selection

In Section 1.2.2 we have seen that regression models are very well suited for modelling GWAS data. We have also seen that single marker tests are a poor strategy for analyzing GWAS data because of the low power of marginal tests. So instead of looking at each SNP individually and select them according to the results of multiple testing adjusted single marker tests one can switch to a multi marker perspective and reformulate the task so that we face a model selection problem.

One way to perform model selection is to construct a measure to rank the models and decide based upon this ranking which model is the best. This measure is called a *model selection criterion*. The two most prominent examples are *Akaike's Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC). Both criteria are presented in the subsequent sections together with some modifications to fit the needs of high dimensional statistics. But in advance, I would like to make a general comment on model selection in a high dimensional setting and in particular in its application to GWAS data.

Both AIC and BIC require to calculate the criterion for every candidate model in order to select the best one. However, in a usual GWAS the number of markers[16] astronomically exceeds the number of individuals. Exacerbating this already unfavorable fact geneticists introduce even more complexity in the model space by their desire to study dominance effects or epistatis. Summarizing the problem, in a GWAS the analyst is confronted with a model universe that is so vast that it is infeasible to calculate the model selection criterion for each model on modern desktop workstations. Even more problematic is the fact that it is usually not possible to calculate these criteria when the number of regressors m entering a model exceeds the sample size n.

On a first glance it seems that model selection could not be applied to GWAS at all, but what helps is the fact that only a comparable small number of markers are actually causative. Abstracting the GWAS context, this idea plays a prominent role in high dimensional statistics and is known as *sparsity* [19, 15]. For the problem sketched in the previous paragraph this means that we can usually limit the set of candidate models to models of size k (number of regressors in the model) with $k \ll n$. But even then this subspace can be so enormous that it is infeasible to calculate the criterion for all

[16] For a typical GWAS the number of genetic markers p is between 10^5 to 10^7 and usually between 100 to 1000 times the number of sampled individuals. This numbers are already impressive, but the number of potential models is 2^p!

models. Therefore complete enumeration is not an option and most software packages and programs that perform model selection for GWAS have some sort of search strategy implemented to find the best model (in the sense of some criterion).

1.3.2.1 The Likelihood Function

The likelihood function plays a central role in a lot of branches of contemporary statistics and it also does in model selection. Assume that n observations are sampled independently out of a homogenous population so that $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed with density function f_θ and parameter $\theta \in \Theta \subset \mathbb{R}^k$ where Θ is an open subset. Then the likelihood function $\mathcal{L}(\theta|Y_1, \ldots, Y_n) : \Theta \to \mathbb{R}^+$ is defined as

$$\mathcal{L}(\theta|Y_1, \ldots, Y_n) \coloneqq f(Y_1, \ldots, Y_n; \theta) = \prod_{j=1}^{n} f_\theta(Y_j).$$

For example let us consider the linear regression model with variable selection, which will be very important for the following discussion. To fix notation, let $S_r \subset \{1, 2, \ldots, m\}$ denote a set of indices that characterizes the subset of regressors which are included in a given model M_r. So each observation Y_i follows the model

$$M_r : Y_i = \beta_0 + \sum_{j \in S_r} \beta_j x_{ij} + \epsilon_i$$

where $\epsilon_i \sim N\left(0, \sigma^2\right)$ independently for $i = 1, \ldots, n$ and $\sigma^2 > 0$ is known. For this model $\Theta = (\beta_0, \beta_{S_r}) \subset \mathbb{R}^{1+|S_r|}$ with dimension $k = |S_r| + 1$. Then, the likelihood function for this model is

$$\mathcal{L}(\theta|Y_1, \ldots, Y_n) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{RSS_r}{2\sigma^2}\right),$$

where

$$RSS_r \coloneqq \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j \in S_r} \hat{\beta}_j x_{ij}\right)^2$$

is the residual sum of square for the fitted model M_r and $\hat{\theta}_r = \left(\hat{\beta}_0, \hat{\beta}_{S_r}\right)$ denotes the *maximum likelihood estimator* (MLE), which is defined as $\hat{\theta}_r \coloneqq \operatorname{argmax} \mathcal{L}(\theta|Y_1, \ldots, Y_n)$. In this model the maximum likelihood estimator is identical to the ordinary least squares estimator (OLS).

1.3.2.2 Akaike's Information Criterion

Fundamental for the derivation of the AIC is the specification of a quasi-distance measure between two different densities. As we have seen in the previous sections a model can be characterized by the density function f_θ. Assume that f^* is the true model. Obviously we would like to select a model f_θ that is close to f^*. In the derivation of the AIC [2] Akaike measured the distance between two models by the *Kullback-Leibler divergence*

$$I\left(f^*, f_\theta\right) \coloneqq \int_{x\in\Omega} f^*\left(x\right) \log\left(\frac{f^*\left(x\right)}{f_\theta\left(x\right)}\right) dx.$$

The Kullback-Leibler divergence measures the information that is lost when f^* is approximated by f_θ. Consequently we want to choose θ such that $I\left(f^*, f_\theta\right)$ is as small as possible. In practice this is not possible, because f^* is unknown. Instead Akaike demonstrated that it is possible to estimate the expected Kullback-Leibler divergence $\mathbb{E}_Y\left[I\left(f^*, f_{\hat{\theta}}\right)\right]$, where $\hat{\theta}$ denotes the maximum likelihood estimator, up to a constant. Under certain technical conditions it holds that

$$\mathbb{E}_Y\left[I\left(f^*, f_{\hat{\theta}}\right)\right] = -\log\mathcal{L}\left(\theta|Y_1, \ldots, Y_n\right) + k_r + const,$$

where k_r is the dimension of the model. Utilizing this fact Akaike defined the AIC as

$$AIC \coloneqq -2\log\mathcal{L}\left(\theta|Y_1, \ldots, Y_n\right) + 2k_r.$$

To perform model selection one has to choose the model with the lowest AIC out of a set of candidate models. Looking at the definition of the AIC, we note the interesting fact that the AIC belongs to the class of *penalized log-likelihood* selection criteria.

Criteria of this class are composed of two terms. The first term is proportional to $\log\mathcal{L}\left(\theta|Y_1, \ldots, Y_n\right)$ and measures the fit of the model to the observed data. Typically the model fit increases, this is expressed by an increased likelihood, with the number of model parameters (model dimension). The second term is a penalty term that counteracts this effect so that more complex models receive a higher penalty.

1.3.2.3 Bayesian Information Criterion

In order to express the key ideas leading to the BIC we first introduce the following notation. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random n-vector with density function corresponding to model M_r denoted by $f_r\left(Y|\theta_r\right)$. When we fix M_r and Y then $f_r\left(Y|\theta_r\right) : \Theta_r \to \mathbb{R}^+$ is again a likelihood function. To achieve a full specification of the model in a Bayesian

sense we also need to define the prior distribution for the k_r-vector θ_r denoted by $g_r(\theta_r)$ and a prior distribution for M_r denoted by $\pi(M_r)$. Using Bayes theorem the posterior probability for a model after observing Y is

$$\mathbb{P}(M_r|Y) \propto \mathbb{P}(Y|M_r)\,\pi(M_r),$$

where

$$\mathbb{P}(Y|M_r) = \int_{\Theta_r} f_r(\theta_r|Y)\, g_r(\theta_r)\, d\theta_r.$$

In Bayesian model selection one chooses the model with the highest posterior probability, which is the key idea behind the BIC.

In practice the computation of $\mathbb{P}(M_r|Y)$ is a time consuming task that requires computational intensive techniques like Markov Chain Monte Carlo (MCMC). Schwarz's [32] idea was to drastically reduce the computational burden by approximating the integral $\mathbb{P}(Y|M_r)$ with the Laplace-Approximation for large samples. In the rest of this section I will heuristically sketch the arguments involved in the derivation of BIC for sufficiently nice behaving densities.

Let us start by looking at the following quantity

$$S(M_r|Y) \coloneqq -2\log\mathbb{P}(M_r|Y) = -2\log\mathbb{P}(Y|M_r) - 2\log\pi(M_r) + const.$$

In order to simplify this expression Schwarz uses the Laplace-Approximation to approximate $\mathbb{P}(Y|M_r)$. By taking the second order Taylor expansion of $\log f_r(Y|\theta_r)$ around its maximizer $\hat{\theta}_r$ (which is identical to the maximum likelihood estimator) one obtains

$$\log f_r(\theta_r|y) \approx \log f_r\left(\hat{\theta}_r|y\right) - \frac{1}{2}\left(\theta_r - \hat{\theta}_r\right)^t \left[n\bar{J}\left(\hat{\theta}_r, y\right)\right]\left(\theta_r - \hat{\theta}_r\right),$$

where $\bar{J}\left(\hat{\theta}_r, y\right) \coloneqq -\frac{1}{n}\frac{\partial^2 \log f_r\left(\hat{\theta}_r|y\right)}{\partial\theta_r \partial\theta_r^t}$ is the average observed Fisher information matrix. Thus

$$f_r(\theta_r|y) \approx f_r\left(\hat{\theta}_r|y\right)\exp\left(-\frac{1}{2}\left(\theta_r - \hat{\theta}_r\right)^t \left[n\bar{J}\left(\hat{\theta}_r, y\right)\right]\left(\theta_r - \hat{\theta}_r\right)\right),$$

and for $\theta_r \approx \hat{\theta}_r$

$$\mathbb{P}(Y|M_r) \approx f_r\left(\hat{\theta}_r|y\right)\int_{\Theta_r}\exp\left(-\frac{1}{2}\left(\theta_k - \hat{\theta}_k\right)^t \left[n\bar{J}\left(\hat{\theta}_r, y\right)\right]\left(\theta_k - \hat{\theta}_k\right)\right) g_r(\theta_r)\, d\theta_r.$$

We know that $\bar{J}\left(\hat{\theta}_r, y\right)$ is positive definite for every n, because $\hat{\theta}_k$ is the maximizer of $\log f_r(Y|\theta_r)$. Thus, for large n, the exponential term of the integrand rapidly decreases

towards zero for any point in the neighborhood of $\hat{\theta}_k$. Hence, the exact form of $g_r(\theta_r)$ is almost irrelevant. Substituting $g_r(\theta_r)$ by a constant (and for simplicity setting $g_r(\theta_r) = 1$) the integral can be easily solved and one obtains

$$\mathbb{P}(Y|M_r) \approx f_r\left(\hat{\theta}_r|y\right)(2\pi)^{\frac{k_r}{2}} \det\left(n\bar{J}\left(\hat{\theta}_r, y\right)\right)^{-\frac{1}{2}}.$$

So for large n one has up to a constant

$$S(M_r|Y) \approx -2\log f_r\left(\hat{\theta}_r|y\right) - k_r \log 2\pi + k_r \log n + \log\det\left(\bar{J}\left(\hat{\theta}_r, y\right)\right) - 2\log\pi(M_r).$$

If we assume that $\det\left(\bar{J}\left(\hat{\theta}_r, y\right)\right)$ is bounded for $n \to \infty$, then we can ignore it as well as the other terms that are bounded, including $2\log\pi(M_r)$, in a large sample setting. This final approximation yields the BIC definition

$$BIC := -2\log f_r\left(\hat{\theta}_r|y\right) + k_r \log n.$$

For model selection one chooses the model with the smallest BIC out of a set of candidate models.

We want to make two remarks about BIC. First, the BIC of a model is independent of the priors $g_r(\theta_r)$ and $\pi(M_r)$. Second, BIC also belongs to the class of penalized log-likelihood criteria. If we compare it to AIC we can see that for $n > 8$ BIC puts a higher penalty on the model complexity than AIC and thus has the tendency to favor smaller models than AIC. A highly relevant property of BIC is that it is consistent for a fixed collection of candidate models, which means that if the correct model is among the candidate models than BIC will select it with probability converging to 1 as $n \to \infty$. The same is not the case for AIC which has the tendency of overfitting even in the classical setting where $p \ll n$ [27].

1.3.2.4 Modifications of BIC for Highdimensional Data under Sparsity

Broman and Speed [7] were among the first researches who explored the model selection approach to analyze QTL[17] data. They observed that both AIC as well as BIC have the tendency to select too big models for moderate sample sizes. Bogdan et al. [5, 6] offered the following explanation for this behavior.

In the previous section I sketched the derivation of the BIC under an asymptotic

[17] In principle the aims of a QTL study and a GWAS are the same. The difference between them is that a GWAS draws samples from a natural population whereas QTL studies use experimentally breeded individuals.

regime that holds the set of candidate models constant while $n \to \infty$. A consequence of this setting was that the model prior $\pi(M_r)$ could be ignored in the definition of BIC. Bogdan et al. [5, 6] argue that this is equivalent with choosing a uniform prior $\pi(M_r) = 1/R$, where R is the cardinality of the set of candidate models. While this choice is uninformative for the models, which means that no model is preferred a priori over the other, it implies an informative prior on the model dimension. For regression models with m potential regressors there are $C(m,k)$ possible models of size k. Hence, a noninformative prior on the set of all possible models implies the prior $p(k) \propto C(m,k)$ on the model dimension . To asses the preference that this prior expresses let us note that $C(m,k+1) = (n-k)/(k+1) C(m,k)$ and that the factor $(n-k)/(k+1)$ is strictly decreasing for $k = 0, \ldots, n$ from n to 0 and is larger than 1 only for $k < n/2$, so $p(k)$ is growing fast for small k and has a maximum at $m/2$, due to the symmetry of the binomial coefficient $g(k)$ around $m/2$. Thus this "noninformative" prior expresses a strong preference for models with dimension $k \approx m/2$. The effect of this on the BIC is a strong bias towards models with a sparsity level of $p \approx 1/2$ for small and moderate sample sizes. Thus, the BIC will tend to select too large models in situations where we expect that the sparsity p of the true model is much smaller than $1/2$. This is the case in GWAS where we expect at most a few hundreds out of hundred thousands of markers to be causative.

In the same articles Bogdan et al. suggest the following remedy for this ill behavior of the BIC in high dimensions. Instead of a uniform prior they specified the model prior

$$\pi(M_r) = \omega^{k_r} (1-\omega)^{m-k_r},$$

which induces a binomial prior $p(k) = C(m,k)\,\omega^k (1-\omega)^{m-k}$ on the model dimension. The parameter ω of the model prior distribution can be interpreted as the expected proportion of causative SNPs in the true model, or in other words our expected sparsity of the true model. The a priori expected number of causative SNPs is given by $c \coloneqq \mathbb{E}[K] = m\omega$. So instead of neglecting the term $-2\log\pi(M_r)$ as Schwarz did in the derivation of BIC, Bogdan et al. suggest to amend the BIC with this term. After some simple approximation the resulting criterion is called the *modified Bayesian Information Criterion* ($mBIC$) and is defined as

$$mBIC \coloneqq BIC + 2k_r \log\left(\frac{1}{\omega}\right) = -2\log f_r\left(\hat{\theta}_r | y\right) + k_r \log\frac{nm^2}{c^2}.$$

If there is no prior knowledge on the expected number of SNPs in the true model then Bogdan et al. [6] suggest $c = 4$ as a default choice. This choice guarantees control of

FWER at a level of 0.1 for $n \geq 200$ and $m \geq 10$. As Bogdan et al. demonstrated there is a close relationship (see Section 1.3.2.5 for details) between $mBIC$ and the Bonferroni procedure for multiple testing .

Another adaption of the BIC to a sparse high dimensional setting was introduced by Frommlet et al. [14, 18]. This modification is based on a discussion by Abramovich et al.[1] on penalized model selection schemes that control the FDR at a given level α. For a linear regression model M_r with size k_r Abramovich et al. defined their selection criterion as

$$S(M) \coloneqq \frac{RSS_r}{\sigma^2} + \sum_{l=1}^{k_r} q^2\left(\frac{\alpha l}{2m}\right),$$

where RSS_r denotes the residual sum of squares of the fitted model and $q(x)$ denotes the $(1-x)$-quantile function of a standard normal distribution. They furthermore assume that σ^2 is known. Under the assumption that the regressors are orthogonal this criterion is closely related to the Benjami-Hochberg procedure.

Frommlet et al. started their adoption by approximating the penalty terms of the sum

$$q^2\left(\frac{\alpha l}{2m}\right) \approx 2\log\left(\frac{m}{l}\right) + \log\left(\frac{2}{\pi}\right) - 2\log(\alpha).$$

This leads to the following approximation for the sum

$$\sum_{l=1}^{k_r} q^2\left(\frac{\alpha l}{2m}\right) \approx 2k_r\log(m) - 2\log(k_r!) + k_r\log\left(\frac{2}{\pi}\right) - 2k_r\log(\alpha).$$

For theoretical reasons presented in [14] Frommlet et al. choose $\alpha \propto 1/\sqrt{n}$, which implies that the FDR of this procedure should decrease towards 0 when $n \to \infty$. With this choice the approximation simplifies to

$$\sum_{l=1}^{k_r} q^2\left(\frac{\alpha l}{2m}\right) \approx k_r\log\left(m^2 n\frac{2}{\pi}\right) - 2\log(k_r!).$$

For large m and n the factor $2/\pi$ could be replaced with $1/c^2$ where $c \coloneqq \mathbb{E}[k] = m\omega$.

Finally the criterion suggested by Frommlet et al. is named *modified Bayesian Information Criterion version 2* ($mBIC2$) and is defined as

$$mBIC2 \coloneqq -2\log f_r\left(\hat{\theta}_r|y\right) + k_r\log\left(\frac{m^2 n}{c^2}\right) - 2\log(k_r!).$$

We can easily see that the $mBIC2$ complexity penalty is smaller than the penalty

applied by $mBIC$ for models which include at least two regressors. This fact is not surprising because $mBIC2$ is designed to control the FDR (a rather liberal multiple comparison control) and $mBIC$ is designed to control the FWER control (a fact which is presented in the next section).

1.3.2.5 Relation of BIC, mBIC and mBIC2 to multiple testing procedures

In this section I will start with the discussion of the relationship between BIC, $mBIC$ and the Bonferroni procedure for multiple testing in the special case of a linear regression model with orthogonal regressors. At this place I reproduce the argument presented by Bodan et al. in [6]. Then I will mention some asymptotic results for $mBIC$ and $mBIC2$ without discussing the corresponding arguments.

Let us start by considering models of the form

$$M_r : Y \sim N\left(X_r\beta_r, \sigma^2 I_n\right)$$

with orthogonal regressors

$$X_r^t X_r = nI_{k_r+1}.$$

Here $Y = (Y_1, Y_2, \ldots, Y_n)$ is a n-vector of observations and $X_r = (\mathbf{1}, x_{.1}, x_{.2}, \ldots, x_{.k_r})$ is a $n \times (k_r + 1)$-matrix of k_r orthogonal regressor variables and $\mathbf{1}$ denotes a n-vector whose components are all 1. Let us further assume that σ^2 is known. For this model

$$-2\log f_r\left(\hat{\theta}_r|y\right) = n\log\left(2\pi\sigma^2\right) + \frac{RSS_r}{\sigma^2},$$

where

$$RSS_r = Y^tY - n\hat{\beta}_r^t\hat{\beta}_r$$

and

$$\hat{\beta}_r = {}^1\!/_n X_r^t Y.$$

Thus, the BIC for this model is

$$BIC\left(M_r\right) = n\log\left(2\pi\sigma^2\right) - \frac{RSS_r}{2\sigma^2} + k_r\log\left(n\right).$$

In the set of candidate models the BIC selects that model which maximizes

$$S\left(M_r\right) \coloneqq \sum_{i=1}^{k_r}\left(\frac{n\hat{\beta}_r^2}{\sigma^2} - \log\left(n\right)\right).$$

This quantity is maximized if and only if the chosen model M_r includes only those regressors for which

$$Z_j^2 := \left(\frac{\sqrt{n}\hat{\beta}_r}{\sigma} \right)^2 > \log(n).$$

Now, observe that under the null hypothesis $H_0^{(j)} : \beta_j = 0$, $Z_j = \frac{\sqrt{n}\hat{\beta}_r}{\sigma}$ has a standard normal distribution. Thus the probability of a Type I Error, which means that a chosen regressor is actually not present in the true model, is

$$\alpha_n = 2\mathbb{P}\left(Z > \sqrt{\log(n)} \right).$$

Using the fact $\mathbb{P}(c) = {}^{\phi(c)}/_{c}\,(1 + o(c))$, where $\phi(.)$ denotes the density of the standard normal distribution, gives

$$\alpha_n \approx \sqrt{\frac{2}{\pi n \log(n)}}.$$

If we assume that the number of true regressors is k then the expected number of incorrectly selected regressors is $(m-k)\,\alpha_n$. In an asymptotic scheme where k is fixed and $m, n \to \infty$ the expected number of incorrectly selected regressors is of order $\frac{m}{\sqrt{n \log(n)}}$ which is not converging towards 0. This illustrates, that under such a scheme the BIC is not consistent. One remedy is to have α_n depending on m as well, for instance set $\alpha_{n,m} := {}^{\alpha_n}/_{m}$. This modification is nothing else but the well known Bonferroni adjustment in multiple testing. In order to find the modification of the BIC such that the selection is consistent under this asymptotic scheme we do some reverse engineering.

In particular we are looking for the quantity c_{Bon} such that

$$2\mathbb{P}\left(Z > \sqrt{c_{Bon}} \right) = \frac{\alpha_n}{m}.$$

Using the above approximation again and considering approximations for large m and n (which implies a large c_{Bon})

$$c_{Bon} \approx \log(n) + 2\log(m).$$

If we repeat the argument that gave us the critical value $c_{BIC} := \sqrt{log(n)}$ for $mBIC$ we obtain the critical value

$$c_{mBIC} = \log(n) + 2\log(m) - 2\log(c).$$

We see that $c_{mBIC} \approx c_{Bon}$ for large m and n. Hence, under this asymptotic scheme

the expected number of incorrectly selected regressors converges to 0 for $mBIC$ based selection.

Bogdan et al. also derived the following approximation for the FWER of $mBIC$ based selection for large m and n

$$FWER \approx \sqrt{\frac{2}{\pi}} \frac{c}{\sqrt{n\left(\log\left(n\right) + 2\log\left(m\right) - 2\log\left(c\right)\right)}},$$

and demonstrated that the probability to select the true model rapidly converges to 1 when $n \to \infty$. In summary this leads to the conclusion that $mBIC$ based selection is consistent for the special case of orthogonal regressors under the assumed asymptotic scheme.

Frommlet et al. [14] proved that under orthogonality the $mBIC$ is ABOS for the linear regression model only in case of extreme sparsity ($p \propto 1/m$). These results were obtained for both known and unknown σ^2. In the same article Frommlet et al. also showed that the $mBIC2$ is ABOS for the linear regression model with known σ^2 in a much wider range of sparsity levels ($p \propto m^{-\beta}$ for $\beta \in (0, 1]$). These findings are in accordance with those of the related multiple testing procedures.

2 METHODS

In this chapter I will present the actual details of the compared algorithms and the setup of the simulation study and the real data example that I used for this comparison. To this end I start with a brief description of the data set which I use for both parts of my inquiry. Because the genotype data of this data set contains missing values, which must be imputed in order to apply model selection procedures, I will also sketch the principles of the imputation algorithm. This will be followed by a description of the Bayesian variable selection method implemented in the PiMass software package and the algorithmic details of the two search strategies implemented in the MOSGWA software package, which allows to perform $mBIC2$ based model selection. Finally, I will give a detailed account on the technicalities of the calibration, conducted to find suitable parameters for the memetic search strategy implemented in MOSGWA, and the simulation study as well as the real data example.

2.1 The Data

The dataset STAMPEED: Northen Finland Birth Cohort 1966 (NFBC1966)[1] serves as a real data analysis example and the corresponding genotype data was used to perform a simulation study.

The Northern Finland Birth Cohorts program (NFBC) was initiated in the 1960s in the two northernmost provinces of Finland to study risk factors involved in preterm birth and intrauterine growth retardation, and the consequences of these early adverse events on subsequent morbidity and mortality. The data of the cohort is obtained from early fetal life (including maternal health during pregnancy) to adulthood. After birth, the offspring was examined and then again underwent clinical evaluation at ages $1y$, $14y$ and $31y$. At each visit, a wide range of phenotypic data was gathered by questionnaires and clinical examinations. DNA samples were obtained from 5402 subjects.

Deidentified genome wide genotype data and a selected list of phenotype data including triglycerides (TG), high density lipoprotein (HDL), low density lipoprotein (LDL), c-

[1]This dataset is available at dbGaP with the accession number phs000276.v2.p1.

reactive protein (CRP), glucose (GLU), insulin (INS), body mass index (BMI), systolic (SYS) and diastolic (DIA) blood pressure measured at the 31y examination are available at dbGaP. Details about the measurement of these nine variables can be found in [30].

2.1.1 Preprocessing of Genotype Data

From a statistical point of view the preprocessing of genotype data is equivalent with the imputation of missing values. It is a common practice in the GWAS community to preselect markers and individuals before the application of an imputation algorithm to ensure that the imputation is not distorted by systematic missing patterns. I preselected[2] the SNPs according to the following widely agreed criteria:

- *Minor Allele Frequency* (MAF) larger than .01.
- *Calling frequency* of at least .975.
- The SNP must pass[3] a *Hardy-Weinberg Equilibrium* test[4] at a significance level $\alpha = 10^{-8}$.
- The marker must be a SNP (so CNV etc. are excluded).

Individuals are only considered for analysis if their total call rate is larger than .95 and if they are not related to any other subject in the sample (IBD smaller than .2). This preselection strategy lead to no reduction of the sample size, but the number of markers decreased from 370404 to 324310 SNPs. Missing values for this selected dataset were than imputed per chromosome with the BEAGLE4[5] software package (for details of the algorithm see Section 2.2).

2.1.2 Population Structure in the Data

In order to account for the population structure (as described in Section 1.2.3.2) the first twenty eigenvector-eigenvalue pairs were calculated and presented in Figure 2.1. Based on the screeplot, the first five eigenvectors were used to adjust for the population structure in the data.

[2] All data management was performed with PLINK (v1.90b3s).
[3] For this test pass means that the null hypothesis is accepted.
[4] This is basically a Fisher exact test.
[5] I used the default values of the parameters and no reference panel data.

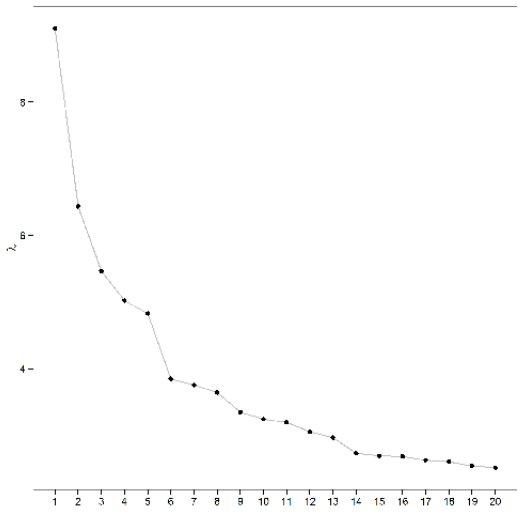

Figure 2.1: Screeplot of the twenty biggest eigenvalues of XX^t where X is the scaled genotype matrix.

2.1.3 Preprocessing of the Phenotype Data

Sabatti et al. [30] present a GWAS of a subsample (in terms of markers and individuals) of the NFBC 1966 where they analyzed the nine phenotypes previously stated. They defined sensible exclusion criteria for most of these phenotypes so that potentially biased measurements are excluded from the analysis. The criteria defined by Sabbati et al. served as a guideline for the definition of the following exclusion criteria[6].

For GLU, INS and the lipid phenotypes (TG, HDL and LDL) individuals were excluded if they had not fastened before the blood sample was taken or if they were diabetic. Subjects were excluded for BMI if the weight measure results were self-reported or if the subject was pregnant. No exclusion criterion was defined for CRP, SYS and DIA.

2.2 BEAGLE

In this section I am going to explain the principles guiding the imputation method described in [9] which is implemented in the BEAGLE[7] program. The algorithmic details are only presented up to a degree of detail that is required for understanding the principle mechanisms of the algorithm. Further details can be found in [8, 9, 10].

[6]Subjects were excluded from analysis by setting their phenotype value to missing.
[7]Retrievable at: https://faculty.washington.edu/browning/beagle/beagle.html [03/20/16]

The BEAGLE algorithm links the problem of imputing missing values to the problem of (computationally) inferring the allele sequence of two homologous chromosomes from (in this respect) unsorted genotype data, as for example obtained from GWAS array experiments. This problem is commonly phrased as *haplotype*[8] *phasing* or just *phasing*.

A central concept for this algorithm is the notion of a *localized haplotype-cluster model*. In order to understand this device let us first assume, that we possess a collection of N haplotypes[9] of length M. This collection forms the population for further considerations. Further, we assume that these sequences contain no missing values.

Now, we partition this population at a position $t \in \{1, 2, \ldots, M\}$ such that all haplotypes within an equivalence class (called a cluster) have a similar conditional probability $\mathbb{P}(a_{t+1}a_{t+2}\ldots a_M | a_1 a_2 \ldots a_t)$. In other words, given the cluster membership at position t the exact pattern $a_1 a_2 \ldots a_t$ is irrelevant for the probability to observe a subsequence $a_{t+1}a_{t+2}\ldots a_M$ right of t. The collection of all partitions is the localized haplotype-cluster model. A localized haploytpe-cluster model can also be represented by a directed acyclic graph with the following properties:

1. The graph has one root node at $m = 0$ with no incoming edges, and one terminal node at $m = M$ with no outgoing edges. The root node represents all of the N haplotypes before any maker is processed, and the terminal node represents all haplotypes after all markers are processed.

2. The graph is leveled with $M + 1$ levels. Each node A has a level m. All incoming edges at a node A at level m originate in a parent node at level $m - 1$, and all outgoing edges from A have a child node at level $m + 1$.

3. The level of an edge corresponds to the level of its child node. An edge is labeled with an allele for the m-th marker. Two edges originating from the same node cannot be labeled with the same allele.

4. For each hapoltype in the population, there is a path from the root node to the terminal node, such that the m-th allele of the haplotype is the label of the m-th edge of the path. Conversely, each edge of the graph has at least one haplotype in the sample whose path traverses the edge, and so represents this group of haplotypes.

An illustration of a localized haplotype-cluster model based on the data presented in Table 2.1 is given in Figure 2.2. For each edge e we define the edge count $n(e)$ to be the

[8] In our usage, this term refers to one specific copy of a homologous chromosome. So a diploid organism possesses two haplotypes for each chromosome.

[9] From our abstract point of view these are sequences of symbols built of a two letter alphabet coding the different alleles.

number of haplotypes in the population whose path traverses the edge, and we define the parent node count $n_p(e)$ to be the number of haplotypes in the population whose path traverses the parent node of the edge.

An algorithm for fitting localized haploytpe-cluster models to a given collection of haplotypes is described in [8, 10]. This algorithm starts by constructing a rooted directed tree graph which encompass all haplotypes in the sample and then merges two nodes A and B at level t if

$$\max_{1\leq k\leq M-t}\max_{a_{t+1}\ldots a_{t+k}}\left|\frac{n_A(a_{t+1}\ldots a_{t+k})}{n_A}-\frac{n_A(a_{t+1}\ldots a_{t+k})}{n_A}\right|<\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$$

where n_X denotes the number of haplotypes whose path traverses node X and $n_X(a_{t+1}\ldots a_{t+k})$ denotes the number of haplotypes whose path traverses node X (at level t) and the sequence $a_{t+1}\ldots a_{t+k}$ at the positions right of t. Obviously $n_X(a_{t+1}\ldots a_{t+k})$ is monotonically decreasing with k and the value of the calculated difference is bounded by the maximum of both relative frequencies, thus long patterns $a_{t+1}\ldots a_{t+k}$ will have no influence on the criterion. Loosely speaking, this criterion merges two nodes based on the similarity of the conditional distributions of short sequences right of t. Thus, the haplotypes are grouped according to local patterns of no specific length, a behavior that is well suited to model linkage disequilibrium patterns.

Haplotype	**Count**
aaaa	21
aaaA	79
aaAA	95
aAAa	116
Aaaa	25
AaaA	112
AaAA	152

Table 2.1: Example of haplotype data

In order to utilize this concept for the purpose of imputation we first recognize that a localized haplotype-cluster model determines a *Hidden Markov Model* (HMM) for which the status space comprises the edges of the graph representing the model. A HMM is specified when the following objects are defined [29]:

1. The space of hidden states H of the HMM.

2. The distinct observable symbols S given a state.

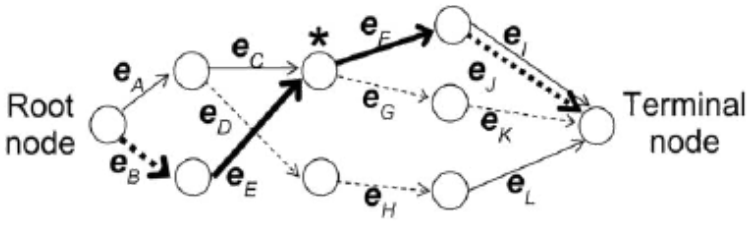

Figure 2.2: Example of a directed acyclic graph representing the localized haplotype-cluster model for four markers, with the haplotype counts given in Table 2.1. For each marker, allele a is represented by a solid line, and allele A by a dashed line. The boldline edges from the root node to the terminal node represent the haplotype AaaA. The node marked by an asterisk (*) is the parent node for edge e_F. This example is taken from [9].

3. The state transition probabilities.

4. The emission probabilities of a symbol given a state.

5. The initial distribution.

As already mentioned the state space of the HMM comprises the edges of the localized haploytpe-cluster model. The emitted symbol for each state is the allele that labels the edge. Thus this symbol (allele) is emitted with probability 1 given the state. The state transition probabilities and the initial distribution are calculated from the edge counts. The initial distribution (for the edges at level 1) is estimated by $\mathbb{P}(e_1 = A) = \frac{n(e_1=A)}{N}$ and $\mathbb{P}(e_1 = a) = \frac{n(e_2=a)}{N} = 1 - \mathbb{P}(e_1 = A)$ for a biallelic locus. The transition probabilities are estimated by $\mathbb{P}(e_i|e_j) = \frac{n(e_i)}{n_p(e_i)}$ if the parent node of e_i is the child node of e_j and $\mathbb{P}(e_i|e_j) = 0$ else. We should also note, that the graph representing the localized haplotype-cluster model of the data is leveled, thus we can partition the state space into classes L_m, where $m = 1, \ldots, M$, so that all edges with level m and only these edges are in the set L_m.

So far we have only specified a haploid HMM, but because we have to deal with data obtained from diploid organisms, we need a diploid HMM. We specify the state space of a diploid HMM as ordered pairs of edges for each level of the localized haplotype-cluster model. Thus the state space for the diploid HMM is $\cup_{i=1}^{M} L_i \times L_i$. The emitted symbol for each state is the observed unordered allele pair at locus i, where $i = 1, \ldots, M$. The emitting probability is either 0 or 1 depending whether or not the state is consistent with the genotype. We assume that the loci in our experiment are in Hardy-Weinberg equilibrium, so the initial probabilities are $\mathbb{P}((e_i, e_j)) = \mathbb{P}(e_i)\,\mathbb{P}(e_j)$ for $e_i, e_j \in L_1$ and

the transition probabilities are $\mathbb{P}\left((e_i, e_j) \,|\, (e_k, e_l)\right) = \mathbb{P}\left(e_i|e_k\right)\mathbb{P}\left(e_i|e_l\right)$ for $e_i, e_j \in L_{t+1}$ and $e_k, e_l \in L_t$ and $\mathbb{P}\left((e_i, e_j) \,|\, (e_k, e_l)\right) = 0$ else.

Now, that the diploid HMM is fully specified we can use the forward-backward algorithm to sample hidden states conditional on the observed genotypes, and the Viterbi-algorithm to find the most likely sequence of hidden states given the observed genotypes [9, 29].

In principle we are now able to determine the most likely ordered pairs of haplotypes conditional on the genotype data and the diploid HMM. But, if we look back at the beginning, we started with the assumption that we possess a sample of complete haplotypes, so we are still not able to perform our desired task, namely, to impute missing genotype data. Instead we have to work with the incomplete unordered genotype samples. The BEAGLE algorithm solves this problem by the following iterative procedure:

1. Set iteration counter to 1.

2. For each of the N observed genotypes impute the missing values by randomly selecting an unordered pair of alleles according to the allele frequencies determined in the sample at the position of the missing value.

3. Phase each of the N genotypes by randomly ordering the alleles for each unordered pair. Now build $2N$ haplotypes out of these randomly ordered allele pairs.

4. Build a localized haplotype-cluster model for the haplotypes obtained in [3] and the corresponding diploid HMM.

5. If the iteration counter equals a predefined number of iterations, use the Viterbi-algorithm to sample the most likely sequence of ordered allele pairs given each of the N genotypes and the diploid HMM specified at [4], then stop the procedure. If the iteration counter is less than the predefined number of iterations, increment the iteration counter by 1 and use the forward-backward algorithm to sample a sequence of ordered allele pairs for each of the N genotypes. Then generate a sample of $2N$ haplotypes out of the sampled ordered pairs. With this sample go back to step [4].

The final result of this procedure are complete sequences of ordered allele pairs which we can use to impute missing values in our genotype data by replacing our missing genotype data with corresponding loci of these sequences.

2.3 PiMass

In the next few paragraphs I am going to sketch the Bayesian Variable Selection Regression (BVSR) proposed by Guan and Stephens [20] and implemented in the PiMass[10] program. For my cursory exposition I will focus on the model specification, as this is the most crucial part in all applications of Bayesian data analysis and on the most important posterior statistics reported by PiMass. I will give no account of the implemented Monte Carlo Markov Chain (MCMC) and other technical aspects of the actual algorithm implemented in PiMass.

The specification begins by considering the following model, which relates a phenotype $\mathbf{y}$ to the genotypes X

$$\mathbf{y}|\mu, \beta, X, \tau \sim N_n\left(\mu\mathbf{1} + X\beta, \tau^{-1}\mathbf{I_n}\right).$$

Here $\mathbf{y}$ denotes a n-vector of observed phenotypes for n individuals, μ is a real valued parameter, $\mathbf{1}$ a n-vector whose components are all 1, X a $(n \times p)$-matrix of p observed markers for n individuals, β a p-vector of regression coefficients, $\tau^{-1} \in \mathbb{R}^+$ denotes the variance and $\mathbf{I_n}$ denotes the $(n \times n)$-identity matrix. μ, β and τ^{-1} are in principal unobservable quantities which we want to infer. For convenience, let us also assume that the columns of X are centered, thus $\sum_{i=1}^{n} x_{ij} = 0$ for all $j = 1, \ldots, p$.

So far we have only specified a linear regression model. In order to incorporate model selection in this model Guan and Stephens define the p-vector $\gamma \in \{0, 1\}^p$ which serves as a selection indicator, following the convention that the marker j is selected if and only if the jth component of γ equals 1. The adapted model can be written as

$$\mathbf{y}|\gamma, \mu, \beta, X, \tau \sim N_n\left(\mu\mathbf{1} + X_\gamma\beta_\gamma, \tau^{-1}\mathbf{I_n}\right),$$

where X_γ denotes the design matrix that is restricted to the selected markers and β_γ denotes the regression weights for the selected markers. To complete the specification for this model Guan and Stephens define the following prior distributions

$$\tau \sim Gamma\left(\lambda/2, \kappa/2\right)$$

$$\mu|\tau \sim N\left(0, \sigma_\mu^2/\tau\right)$$

$$\gamma_j \sim Bernoulli\left(\pi\right)$$

[10] Retrievable at: http://www.haplotype.org/pimass.html [03/20/16]

$$\beta_\gamma|\tau,\gamma \sim N_{|\gamma|}\left(0,\left(\sigma_a^2/\tau\right)\mathbf{I}_{|\gamma|}\right)$$

$$\beta_{-\gamma}|\gamma \sim \delta_0$$

where $|\gamma| \coloneqq \sum_{j=1}^{p}\gamma_j$, $\beta_{-\gamma}$ denotes the vector of coefficients for the nonselected markers and δ_0 denotes a point mass at 0. $\pi, \sigma_a, \lambda, \kappa$ and σ_μ are hyperparameters, among which λ, κ and σ_μ are of minor importance. For the calculation of the posterior distributions Guan and Stephens (only) consider the limiting case $\sigma_\mu^2 \to \infty$ and $\lambda, \kappa \to 0$,which means that they make use of improper priors.

In their paper Guan and Stephens put special emphasis on the choice of the hyperparameters π and σ_a. They remark that these parameters have a specific interpretation, namely that π reflects the sparsity of the model and that σ_a reflects the typical size of the nonzero coefficients. Because these are crucial and usually unknown model features in GWAS, Guan and Stephens put prior distributions on both parameters. In the Bayesian literature π and σ_a are usually chosen such that they are stochastically independent, whereas Guan and Stephens argue, that such a choice implies that more complex models (bigger selection probability π) are expected to have a substantially larger proportion of explained variance (PVE). In a more genetic parlance, this means that traits with a more complex genetic architecture are expected to have higher heritability. Guan and Stephen doubt this implication, and suggest that the priors for π and σ_a should not be modeled as independent. First, they choose the following prior distribution for $log\left(\pi\right)$

$$log\left(\pi\right) \sim U\left(log\left(1/p\right), log\left(M/p\right)\right)$$

where M is a predefined constant. This prior implies that $\pi \in [1/p, M/p]$, which means that the expected number of SNPs in the model ranges from 1 to M. It also expresses a preference towards more sparse models, due to the fact that the prior distribution of π puts more weight on smaller values. Instead of modeling the prior of σ_a explicitly Guan and Stephens do it indirectly by putting an approximately uniform prior on the conditional expected PVE. They start by defining the following quantity

$$V\left(\beta,\tau\right) \coloneqq \frac{\tau}{n}\sum_{i=1}^{n}\left[\left(X\beta\right)_i\right]^2.$$

Given that the columns of X are centered, this quantity is the variance of $X\beta$ divided by τ^{-1}. Hence the proportion of variance in y explained by X given β is

$$PVE(\beta,\tau)=\frac{V\left(\beta,\tau\right)}{1+V\left(\beta,\tau\right)}.$$

Guan and Stephens further note the fact that

$$v(\gamma, \sigma_a) \coloneqq \mathbb{E}[V(\beta, \tau) | \gamma, \sigma_a, \tau] = \sigma_a^2 \sum_{j:\gamma_j=1} Var(x_{\cdot j})$$

and use this to define the quantity

$$h(\gamma, \sigma_a) \coloneqq \frac{v(\gamma, \sigma_a)}{1 + v(\gamma, \sigma_a)}$$

as an approximation for the conditional expected PVE[11]. Guan and Stephens place a uniform prior $h \sim U(0, 1)$ on this quantity. This prior is independent of γ, but it implies that the pdf[12] and the cdf of σ_a^2 given γ are

$$pdf_{\sigma_a^2|\gamma}(z) = \frac{c_\gamma}{(1 + c_\gamma z)}$$

respectively

$$cdf_{\sigma_a^2|\gamma}(t) = 1 - \frac{1}{1 + c_\gamma t},$$

where $t, z \geq 0$ and $c_\gamma \coloneqq \sum_{j:\gamma_j=1} Var(x_{\cdot j})$. In order to better understand the relationship between γ and σ_a^2 we look at the cdf of $\sigma_a^2|\gamma$ as a function of c_γ. We can see that for every t the cdf is monotonically increasing with c_γ. Thus the pdf puts more mass on lower values as c_γ increases. c_γ is big if the selected markers have a high variance or if a lot of markers are selected. In summary the consequence of this modeling is that the effect of single markers tends to be smaller in bigger models.

The most important output quantities of PiMass are the *posterior inclusion probabilities* of the considered markers and a *heritability* estimate for the investigated phenotype. For the former PiMass calculates a crude estimator (the proportions of models in the Markov Chain that include the specific SNP) and a Rao-Blackwellized refinement of it. In order to select SNPs one chooses one of these estimators and selects all SNPs with an estimated posterior probability bigger than .5. PiMass provides two estimators for the heritability, one based on MCMC results for the sampled h (called h) and the other one using the derived sequence for PVE that is calulated from the sampled MCMC sequences of β and τ (called hh).

[11] Actually $h(\gamma, \sigma_a)$ is an upper bound for the expected PVE, due to Jensen's inequality.

[12] We should also note, that this distribution doesn't possess a first moment. In other words, the distribution is heavy tailed.

2.4 MOSGWA

The software package MOSGWA[13] allows to perform model selection based on different model selection criteria (including mBIC2 and mBIC) for case-control and continuous phenotype GWAS. As mentioned in Section 1.3.2 one of the intricacies of the model selection approach in GWAS is the enormous number of candidate models so that complete enumeration is not feasible even on modern computers. The only way to cope with this problem is the implementation of optimization heuristics. Currently, MOSGWA has implemented two search strategies to find the optimal model. One is a variant of a greedy algorithm and is called *fast stepwise search* (FSS) and the other one is a *memetic algorithm*.

2.4.1 Fast Stepwise Search

In order to find an improved (in the sense of *a criterion*) model M starting from an *initial model* M_{init} FSS iteratively applies a three-step procedure [13]. The procedure starts by ordering all the $m - |M_{init}|$ candidate Markers (excluding the SNPs already in M_{init}) according to single marker test p-values (starting with the lowest). Then two groups of markers G_1 and G_2 are derived from this sequence. For this purpose two numbers m_1 and m_2 are chosen such that $0 < m_1 < m_2 \leq m - |M_{init}|$ (default $m_1 = 350$ and $m_2 = 500$). Now G_1 is defined as the set of the first m_1 elements of the ordered sequence and G_2 as the set of the first m_2 elements of this sequence. Thus $G_1 \subset G_2$. These tuples are a preselection of candidate markers based on their "marginal explanatory power" and thus limit the search space to the most promising models. Now, the initial model is set as the current model and the iteration begins.

The first step is called *directed*[14] *forward step*. This step itself is a loop over G_1 where in each step the current model is enhanced by including the current SNP of G_1 in the model. If the criterion of this enhanced model is smaller than the one of the active model the loop is broken and the enhanced model is set as the current model.

The next step is called *exchange step*. In this step all SNPs in the current model are tested whether exchanging them with suitable candidates decreases the criterion. Suitable candidates for a model SNP are the d nearest (the distance is measured in absolute base pairs between locations) SNPs of G_2. The exchange step starts with the

[13] Retrievable at: https://sourceforge.net/projects/mosgwa/ [03/20/16]. I used MOSGWA v1.2.10 for the experiments presented in this thesis. After I finished my calculations a new version of MOSGWA (v1.2.11) which implemented a major modification of the memetic search algorithm was available at scourceforge.

[14] In the sense that the search is performed along the p-value sorted sequence G_1.

first SNP in the model. If there is at least one replacement that improves the current model then the model with the lowest criterion is set as the current model. Then the second SNP is replaced in the same fashion, then the third and so forth until every SNP in the model has been considered once.

The last step is an extended *backward elimination step*. First all models where one SNP is removed from the current model are evaluated. If this yields a model with a lower criterion then the model with the lowest criterion is set as the current model and this step is finished. If this yields no improvement than among the narrowed models the one with the smallest criterion is fixed and the process is repeated. This iteration is repeated up to three times.

The result of the last step is passed on to a directed forward step and the next iteration is started. This loop stops when no improvement of the criterion is achieved within an iteration.

This procedure is the major building block of the actual search that is performed when MOSGWA is running in this mode. For convenience let us denote this process as a function called FSS with input parameters M_{init}, *test* and *criterion*. The result of FSS is a model M, which is an element of the search space. Let us further define a new model selection criterion

$$mBIC_{60} \coloneqq -2 \log f_r \left(\hat{\theta}_r | y \right) + k_r \log \frac{nm^2}{c^2 60^2}.$$

The complexity penalty of this criterion is smaller than the one applied by $mBIC2$ for sufficiently small models. Hence, this criterion has the tendency to select bigger models than $mBIC2$ in this subclass of models.

One of the problems of search heuristics in general is that they can get stuck in a local optimum. For model selection this means that too small (compared with the true model) models are selected. To avoid this MOSGWA has implemented the following three-step strategy to avoid getting trapped in a local minimum. Starting with the empty model M_0 it computes

1. $M^* = FSS\left(M_0, \text{Cochran Armitrage}, mBIC_{60}\right)$
2. $M^{**} = FSS\left(M^*, \text{Score Test}, mBIC_{60}\right)$
3. $M^{fin} = FSS\left(M^{**}, \text{Score Test}, mBIC2\right)$

The rational behind this strategy is that the application of the milder criterion $mBIC_{60}$ in the first two rounds should yield models which are far too large and hopfully include a large proportion of relevant causal SNPs. This model is then trimmed back in the third

step when the relevant (and stricter) criterion is applied. Thus it should be guaranteed that the algorithm does not get stuck in a poor local minimum for the relevant criterion. The main difference between steps 1 and 2 is the test which is used to introduce the ordering of candidate SNPs. The Cochran Armitrage test evaluates each candidate SNP independent of all other SNPs, whereas the Score Test evaluates the candidate SNPs conditional on the SNPs which are already included in M^*.

2.4.2 The Memetic Search Algorithm

Previous studies [18, 13] have shown that the stepwise search presented in the previous section works very well. However it is not guaranteed that the result obtained by this search strategy is really the best solution. Memetic algorithms are a possible approach to boost the results obtained by FSS and find models with a smaller criterion value.

However the motive to find a better solution than FSS is not the only consideration that motivates the application of a memetic algorithm to the model selection problem addressed by MOSGWA. One of the shortcomings of classical model selection criteria is the fact that they only provide one "best" model, and do not allow to estimate the uncertainty tied to this choice. If we look back at the derivation of $mBIC2$ (Section 1.3.2.3 and 1.3.2.4) we can see that

$$\mathbb{P}(M|Y) \propto \mathbb{P}(Y|M)\,\pi(M) \approx \exp\left(-mBIC2(M)/2\right).$$

In order to exploit this fact to calculate model posterior probabilities we need to compute the normalizing constant $\mathbb{P}(Y)$. By the law of total probability we know that

$$\mathbb{P}(Y) = \sum_{M \in \mathscr{M}} \mathbb{P}(Y|M)\,\pi(M) \approx \sum_{M \in \mathscr{M}} \exp\left(-mBIC2(M)/2\right),$$

where $\mathscr{M}$ denotes the set of all possible models. Unfortunately $|\mathscr{M}| = 2^m$, m denoting the number of candidate markers, is usually so large that it is simply unfeasible to calculate this sum exactly. But some reflections on the individual terms of the sum show us a way for a good and feasible approximation of this sum.

Without doubt it is reasonable to assume that most of the models in $\mathscr{M}$ will fit the data poorly, resulting in very large $mBIC2$ values for these models. Thus, the individual contribution of these models to the sum will be approximately zero. Hence, they can be neglected for this calculation and it is possible to approximate this sum with

a significantly smaller subset $\mathscr{M}^*$. Therefore

$$\mathbb{P}(Y) \approx \sum_{M \in \mathscr{M}^*} \exp\left(-mBIC2(M)/2\right).$$

One can then further define an estimator for the *posterior inclusion probability of a SNP* j as

$$\hat{\mathbb{P}}(j|Y) := \sum_{M \in \mathscr{M}^* : j \in M} \mathbb{P}(M|Y).$$

So the crucial point is to find a good set $\mathscr{M}^*$. In the next paragraphs I will sketch the design of a memetic algorithm[15] that performs this task by an extensive search over models which have a large posterior probability.

The memetic algorithm that is proposed here for GWAS analysis is closely related to an algorithm developed and applied by Frommlet et al. [17] in the context of QTL mapping. The term memetic algorithm refers to a synergy of population based evolutionary search strategies (known as genetic algorithms) with separate local improvement strategies for individuals. Genetic algorithms work with an initial population of models which evolves over time (whereby the number of individuals is kept constant), so that the fitness (measured by a fitness function that maps an individual to a real number) of the whole population (defined as the sum of the individual's fitness values) is non-decreasing over time. To that end genetic algorithms apply the random operations of *selection*, *recombination* and *mutation* to the population to generate offspring. If the offspring proves to be fitter than the least fit individual in the population this individual is replaced by the offspring and the process is repeated. So when applying a genetic algorithm to solve an optimization problem one has to define these operations and construct a suitable initial population.

Obviously the *fitness function* that has to be optimized in our application is $mBIC2$[16], but we also want that the algorithm visits a lot of models in $\mathscr{M}^*$ so that we can get good posterior probability estimates. To achieve both goals (and keep the runtime within reasonable bounds) we need to generate the *initial population* so that on the one hand the algorithm visit mainly models in $\mathscr{M}^*$ but on the other hand provides sufficient diversity among the individuals of the initial population to avoid getting stuck in a local optimum.

To this end a first set of v models is obtained by applying a greedy selection procedure. The first model of the population is obtained by running the procedure described in Section 2.4.1. Then we remove all SNPs with a single marker p-value larger than .1

[15] The algorithm I present here is the memetic algorithm implemented in MOSGWA v1.2.11.
[16] In our case a higher fitness is indicated by a lower $mBIC2$ value.

from the set of candidate markers. We also exclude the SNPs that are included in the first model from this set. In order to obtain a second member we apply the procedure termed FSS in the previous section starting with an empty model and BIC as criterion to search over the set of models that can be build out of the reduced set of candidate SNPs. Then we remove the selected SNPs from the candidate set and perform the search again. Thus we obtain a third member. This procedure is repeated until v models are selected. Because the complexity penalty of BIC is smaller than the one of $mBIC2$ this step selects models that are way to large. To avoid unrealistically large models we restrict the maximum size of a model to $k_{max} = 150$ SNPs.

Now, a second set of $(u - v)$ models is created by random selection from the remaining set of candidates. Again, we apply an iterative procedure which starts by calculating a selection probability $\pi_i \propto 1/p_i$, where p_i denotes the single marker test p-value for SNP i, for each candidate SNP. Now, k_{max} SNPs are randomly selected and combined to form a model. These SNPs are removed from the set of candidates and the procedure is repeated until $(u - v)$ models are generated.

Finally we put both sets together. Hence, we achieved to construct an initial population of size u where all models are disjoint.

We implement tournament selection as the *selection* operator to choose parent models. Therefore two models are randomly drawn from the population and the fitter one is chosen as a parent. This process is repeated until we obtain two distinct parent models.

Recombination between the parents is performed with probability $p_r = .9$. If no recombination is performed than the fitter parent is chosen as offspring. The rational behind this option is to allow direct mutation of population individuals without previous recombination. The recombination step itself consists of a modified forward respectively backward selection which always includes the markers that are present in both parent models in the offspring model. Let I denote the set of all markers that are present in both parents and D denote the set of all markers that are present in one and only one parent model. The forward selections starts from I and in a greedy fashion adds one SNP at a time from D to the model. The backward elimination starts with the model that includes all SNPs from both parents and in a greedy fashion eliminates one SNP after the other. During this step only SNPs in D are allowed to be eliminated. The fittest model obtained by these two procedures is chosen as offspring.

This step generates a child that is at least as fit as both parents, so this is not really a random recombination but rather an implicit optimization step. Thus we have a memetic algorithm and not a pure genetic algorithm.

Next, *mutation* of the child model is performed mandatory if no recombination hap-

pened or with probability $p_m = .25$ otherwise. In the mutation step either the action of adding a SNP to the model or the action of removing one SNP from the model is performed. Which sort of action happens is randomly determined, whereby both actions are equally likely. In the case of deletion one randomly selected marker is removed from the model, except when the child model is of size one, in which case the marker is substituted by another randomly selected marker from the set of candidates. In the case of addition the child model is appended by a randomly selected marker that satisfies the condition that its absolute correlation with ever SNP already in the child model is less than .5.

After selection and mutation a local improvement step for the child model is performed. This step resembles the exchange step described in section 2.4.1. During this step a randomly selected fraction p_{local} of child model SNPs (one at a time) are tested whether exchanging them with SNPs from their neighborhood improves the fitness of the model. The neighborhood of a SNP is defined to consist of those SNPs among the 100 closest (50 SNPs on each side) which are correlated at a level $|r| > .3$ with the considered SNP. In a greedy manner we always keep the SNP that results in the best model fitness. Compared with the algorithm of Frommlet et al. [17] we put less emphasis on local improvement. The reason for this choice is that – keeping in mind our second goal to estimate posterior probabilities – we want the algorithm to visit a lot of good models and therefore we are willing to sacrifice some algorithmic efficiency (in terms of convergence rate).

For the purpose of calculating posterior probability estimates and also to improve the computational efficiency we keep track of all visited models and store the associated $mBIC2$ value. This inventory is called the *pool* (of visted models). The algorithm stops if (A) within a certain number of iterations no new model was found which is among the B best models of the population or (B) if the pool size exceeds a specified number. As it is always the case with memetic algorithms the actual performance of the algorithm strongly depends on the choice of the involved parameters. Therefore attention has to be paid to this problem.

When MOSGWA performs the memetic search, it offers two ways to select a "best" model. One is simply to choose the model with the minimal selection criterion found by the memetic algorithm. The other choice is based on the estimator of the posterior probabilities of the markers. This procedure starts with adding the estimated posterior probabilities of all makers within a region[17]. If this cumulative estimated posterior probability is bigger than .5, the marker with the highest estimated posterior probability

[17] In order to be in the same region two SNPs have to pass three criteria. They must be located on the same chromosome and the physical distance between them must be smaller than 1MBP. The third criterion is that the Pearson correlation between them must be bigger than .3.

Scenario	k	h^2	β_{min}	β_{max}
I	0	0		
II	20	.18	.05	.24
III	30	.38	.05	.34
IV	50	.69	.1	1

Table 2.2: Characteristics of the four simulation scenarios, where k denotes the number of causal SNPs and h^2 the heritability. β_{min} and β_{max} are the smallest and the largest effect size.

within a region is reported as a selected SNP.

If MOSGWA performs memetic search it also calculates a heritability estimator, which is simply the proportion of explained variance of the selected model and an associated credibility interval.

2.5 The Experiments

The main purpose of this thesis is to investigate the properties of $mBIC2$ based model selection and the effect of different search strategies for a typical quantitative trait GWAS and compare the results to the Bayesian variable selection model implemented in PiMass[18]. Furthermore, I would like to compare the heritability estimator offered by MOSGWA with those obtained by PiMass. To this end I conducted a simulation study and reanalyzed a real data example.

An integral part of my work was the fine tuning of certain parameters of the memetic algorithm. The resulting parameters will be used as default values in MOSGWA.

2.5.1 Simulation Study

The simulations are based on the real SNP data of the first chromosome ($p = 24622$) of the $n = 5402$ subjects of the NFBC66 study. I consider four different scenarios, whose characteristics are presented in Table 2.2. In the first scenario simulations are performed under the total null model whereas scenarios II to IV consider different numbers of causal SNPs, effect sizes and heritability. The causal SNPs are chosen to be approximately equidistant (with distance larger than 3 MBP) in such a way that the whole chromosome is covered and that the minor allele frequency (MAF) of each SNP is bigger than 0.3. The Pearson correlation between any two causal SNPs is smaller than 0.1. The choice of

[18] Choice of PiMass parameters used for the rest of this thesis: 1000 burn in steps and total chain length of 300000 which is recorded every 10 steps.

causal SNPs and their corresponding effect size for each scenario is listed in Table 2.4 and Table 2.6.

For each scenario a linear regression model with a standard normally distributed error term was used to generate 100 artificial phenotypes. For scenarios II and III, which correspond to relatively small and intermediate heritability, the true regression coefficients were equally spaced between the minimum and maximum value. For the fourth scenario I consider 41 SNPs with relatively small effect size and 9 SNPs with a large effect size, consequently the heritability in this scenario is fairly large.

The causal SNPs for each scenario have been eliminated from the set of candidate markers before analysis in order to account for the fact that in a real GWAS only a fraction of all possible SNPS are measured and so most of the time the causative SNP can only be detected indirectly.

As mentioned in Section 1.1.4 the aim of a GWAS is to detect regions that contain at least one causative marker. In order to define suitable performance measures, I introduce the following terminology. The set of all candidate SNPs which have an absolute Pearson correlation of at least .3 with a causal SNP and are not more than 1.5 MBP away is defined as the *region around a causal SNP*. A causal SNP is called *detected*, when at least one SNP of its region is also a selected SNP. A causal SNP is a *false negative*, if it is not detected. A selected SNP which does not belong to a region of a causal SNP is termed a *false positive*. If the converse is true, we call the SNP a *true positive*.

The results of an algorithm will be evaluated according to the following performance measures

$$\begin{aligned}
Power &:= \frac{\#\{detected\, SNPs\}}{\#\{true\, model\, SNPs\}} \\
False\, Positive &:= \#\{selected\, SNPs\} - \#\{detected\, SNPs\} \\
FDR &:= \frac{\#\{False\, Positve\}}{\#\{selected\, SNPs\}} \\
False\, Negative &:= \#\{true\, model\, SNPs\} - \#\{detected\, SNPs\} \\
Miss &:= \#\{False\, Positive\} + \#\{False\, Negative\}
\end{aligned}$$

where $\#X$ denotes the cardinality of the set X. These measures are calculated for each of the 100 replications. The final comparison of the algorithms will be based on the mean values of these observed measures.

SNPId	BP	MAF	S1	S2	S3	S4
rs3766178	1468043	0.3173	0	0.05	0.05	0.1
rs277672	6694194	0.4802	0	0	0.06	0.1025
rs2982376	11312923	0.4869	0	0.06	0	0.105
rs848212	16136039	0.3799	0	0	0.07	0.1075
rs12070677	21209909	0.4684	0	0	0	0.11
rs2095426	26090671	0.4822	0	0.07	0.08	0.1125
rs4949294	31008523	0.4526	0	0	0.09	0.115
rs6702202	36869181	0.3143	0	0.08	0	0.1175
rs4660429	40794563	0.3592	0	0	0.1	0.12
rs3013595	45925790	0.4444	0	0	0	0.1225
rs1875645	50562467	0.3959	0	0.09	0.11	0.125
rs2767503	55925313	0.4834	0	0	0.12	0.1275
rs6587887	60781874	0.436	0	0.1	0	0.13
rs913199	65643650	0.4778	0	0	0.13	0.1325
rs6424388	70703376	0.3964	0	0	0	0.135
rs6656537	75558471	0.3603	0	0.11	0.14	0.1375
rs1012455	80512131	0.4393	0	0	0	0.14
rs6660237	85468592	0.332	0	0.12	0.15	0.1425
rs12035196	90502332	0.3388	0	0	0.16	0.145
rs10458508	95415812	0.304	0	0	0	0.1475
rs3766600	100160527	0.3277	0	0.13	0.17	0.15
rs6695731	105309419	0.4479	0	0	0	0.1525
rs7545139	110389479	0.3482	0	0	0.18	0.155
rs360622	115208584	0.3733	0	0.14	0.19	0.1575
rs539708	120010026	0.388	0	0	0	0.16

Table 2.4: First 25 of 50 SNPs from chromosome 1 selected to be causal for the simulation study. The consecutive columns contain: SNPId, position (in base pairs), minor allele frequency (MAF) and the regression coefficients for Scenarios 1, 2, 3 and 4.

SNPId	BP	MAF	S1	S2	S3	S4
rs12724816	144350681	0.3955	0	0	0.2	0.1625
rs1498308	148561755	0.3964	0	0.15	0	0.165
rs884618	152867781	0.4305	0	0	0.21	0.1675
rs2106092	157099068	0.3736	0	0	0.22	0.17
rs2841959	161324818	0.4857	0	0.16	0	0.1725
rs1021621	165465160	0.406	0	0	0.23	0.175
rs6656814	170007569	0.4793	0	0	0	0.1775
rs12033847	174294601	0.4083	0	0.17	0.24	0.18
rs4076449	178263641	0.3375	0	0	0	0.1825
rs7549909	182512254	0.31	0	0.18	0.25	0.185
rs10754227	187001959	0.4469	0	0	0	0.1875
rs1234722	191121675	0.3412	0	0	0.26	0.19
rs10754210	195278734	0.3313	0	0.19	0	0.1925
rs2782581	199770355	0.3666	0	0	0.27	0.195
rs823096	203946510	0.4507	0	0.2	0.28	0.1975
rs6661316	208162150	0.4794	0	0	0	0.2
rs1391553	212540742	0.3041	0	0	0.29	0.25
rs6684205	216676325	0.301	0	0.21	0.3	0.3
rs35746652	220779831	0.3714	0	0	0	0.4
rs3738725	225240833	0.4043	0	0.22	0.31	0.5
rs531592	229463257	0.4168	0	0	0	0.6
rs291388	233714648	0.4996	0	0	0.32	0.7
rs2278642	237933766	0.3783	0	0.23	0	0.8
rs2047137	242263033	0.3032	0	0	0.33	0.9
rs11204620	246495261	0.4135	0	0.24	0.34	1

Table 2.6: Last 25 of 50 SNPs from chromosome 1 selected to be causal for the simulation study. The consecutive columns contain: SNPId, position (in base pairs), minor allele frequency (MAF) and the regression coefficients for Scenarios 1, 2, 3 and 4.

2.5.1.1 Calibration of the memetic search

The performance and runtime of the memetic search algorithm are highly dependent on the choice of its main parameters[19] population size ($modelsNo$), maximum number of visited models ($maxPoolSize$), the degree of exhaustion during local search ($pLocalExchangeTrial$), the maximum number of iterations ($maxNoProgressIter$), the reset condition (B) and the composition of the initial population ($fastForwardModelsNo$).

In order to find a balanced setup between runtime and the prospect to find a globally optimal model I conducted a small simulation study based on five repetitions for each parameter setting, with phenotypes generated according to scenario III. I decided to fix the values for the maximum number of visited models to 200000 and the population size to 100 to keep the runtime within reasonable bounds. In a first step I focused my interest on three aspects, namely, the composition of the initial population, the degree of localization of the search strategy and the iteration reset condition to get a good performing combination of parameters. For this, I considered all combinations of the following parameter values $fastForwardModelsNo \in \{10, 25, 50\}$, $pLocalExchangeTrial \in \{.05, .1, .25, .5\}$, $B \in \{10, 20, 30, 40\}$ and $maxNoProgressIter = 5000$. The poolsize was nearly 200000 for every replication and all combinations. I observed that the choice for B had no substantial impact on the performance or the runtime and that the combination $fastForwardModelsNo = 50$, $pLocalExchangeTrial = .25$ showed the best (and very satisfying) performance. In the next step I wanted to reduce the runtime while not losing too much in terms of performance. So I fixed $fastForwardModelsNo = 50$ and $pLocalExchangeTrial = .25$, and investigated all combinations of $B \in \{10, 20, 30, 40\}$ and $maxNoProgressIter \in \{500, 1000, 2500, 5000\}$. It turned out that the combination $B = 10$ and $maxNoProgressIter = 1000$ clearly reduced the runtime while the loss in the performance parameters was only minor.

The final parameter setting which was used throughout the rest of the thesis and which now serves as the default setting of MOSGWA is $maxPoolSize = 200000$, $modelsNo = 100$, $fastForwardModelsNo = 50$, $B = 10$, $pLocalExchangeTrial = .25$, and $maxNoProgressIter = 1000$.

[19] The names of the parameters controlling these aspects of the memetic search algorithm in MOSGWA are given in brackets.

2.5.2 Real Data Analysis

As a real data example I will reanalyze the NFBC66 dataset. A subset of this dataset has already been analyzed by Sabbati et al. [30]. I tried to mimic the analysis of Sabbati et al. as closely as possible, but due to the fact that the datasets are not identical and that some data preprocessing steps could not be replicated with the data available at dbGaP a direct comparison between my results and those of Sabbati et al. is not possible. Instead I will focus on comparing the results obtained by MOSGWA with those of PiMass and linear regression based single marker tests applying Benjamini-Hochberg and Bonferroni adjustment with the typical choice of .05 for $FWER$ respectively FDR. Because the memetic search algorithm of MOSGWA and the MCMC sampler implemented in PiMass contain random elements I repeat the analysis with those algorithms five times to assess the stability of the results. I used both PiMass and MOSGWA with the same parameter setup as in the simulation study, with the only exception of $maxPoolSize$ for MOSGWA's memetic search which I set to 100000 to bound the runtime at a reasonable level.

As mentioned in Section 2.1.2 there is clear evidence for population structure in the data. In order to control the results for this I applied the strategy described in Section 1.2.3.2. Sabbati et al. also suggested that three other variables (use of oral contraception (OC), sex of a subject (SEX) and pregnancy status) should be considered as covariates for analysis. Therefore all regression models include these three variables and the first five eigenvectors as mandatory regressors. The only exception to this is BMI where PG is omitted as covariate. This adjustment can be applied directly for MOSGWA and the single marker tests. Unfortunately PiMass is not able to handle mandatory regressors. So, I had to use the residuals of the phenotype regressed on these variables instead of the observed phenotype as the regressand for PiMass. This remedy, which is not equivalent with the intended procedure, is suggested by Guan and Stephens [20]. The inability of PiMass to include mandatory covariates is certainly a disadvantage.

3 RESULTS

3.1 Simulation Study

Let us start by reviewing the performance results of the competing model selection algorithms (see Table 3.1 and Figure 3.1). MOSGWA regardless of the search strategy does an excellent job in controlling the FWER under the total null hypothesis (scenario I). The same is true for PiMass (we observed only one run for scenario I were PiMass reported a SNP).

	MA_Best	**MA_Post**	**Greedy**	**PiMass**	**BH**
	Scenario I: k=0				
FDR	<.01	<.01	0.08	0.01	<.01
	Scenario II: k=20				
Power	0.756	0.76	0.743	0.612	0.762
FDR	0.004	0.004	0.002	0.03	0.0874
FP	0.06	0.06	0.03	0.42	14.87
Mis	4.87	4.86	5.17	8.18	19.63
	Scenario III: k=30				
Power	0.855	0.854	0.837	0.744	0.853
FDR	0.0085	0.0086	0.0023	0.0249	0.1586
FP	0.23	0.23	0.06	0.61	50.6
Mis	4.58	4.61	4.93	8.27	55.02
	Scenario IV: k=50				
Power	0.932	0.932	0.924	0.833	0.722
FDR	0.011	0.0107	0.0095	0.0315	0.4086
FP	0.53	0.51	0.45	1.44	161.58
Mis	3.93	3.92	4.24	9.77	175.47

Table 3.1: Comparison of estimated *power*, *false discovery rate* (FDR), *false positives* (FP) and *number of misclassifications* (Mis) for PiMass, MOSGWA in FSS mode (Greedy), MOSGWA in memetic search mode with posterior inclusion probability based selection (MA_Post) and best criterion based selection (MA_Best) as well as Benjamini Hochberg adjusted single marker tests (BH) with a nominal FDR level of .0085. This choice is based on the approximate theoretical FDR level of mBIC2 calculated by formula 16 presented in [6].

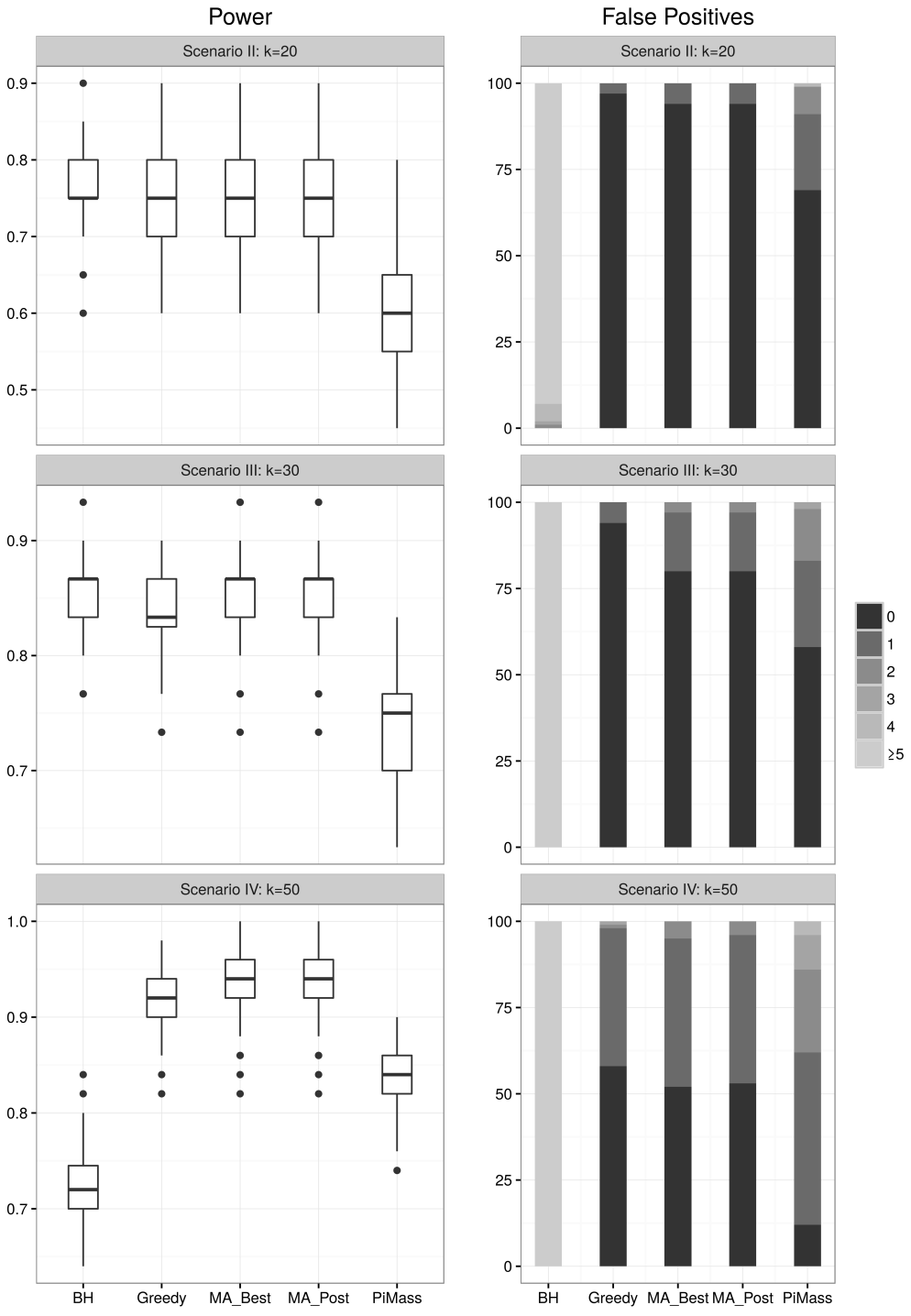

Figure 3.1: Comparison of estimated power (right column) and number of false positives (left column) for all methods and scenarios II-IV. For FP dark gray equals zero false positives and the lighter the higher the number of FP (see legend on the right).

The dark gray portion in the left column of Figure 3.1 on the preceding page shows the number of runs where no FP was reported. This number decreases for both methods with increasing scenario complexity, but the performance loss of PiMass is much more pronounced compared to MOSGWA. A detailed look at this figure shows that for all scenarios $mBIC2$ based algorithms reported no false positive signals in more than 50 out of 100 replications, this number exceeds 75 for scenario III and is about 90 for scenario II. In all scenarios PiMass shows a much stronger tendency to report false positive signals (this is also supported by the numbers presented Table 3.1). Especially in scenario IV PiMass reported at least one wrong SNP in 90 out of 100 runs. When it comes to the total average number of misclassified SNPs $mBIC2$ based methods again clearly outperform PiMass in every scenario. These results are in line with the fact that $mBIC2$ based selection is asymptotically a Bayes optimal classifier [14]. They further suggest that this property holds at typical GWAS sample sizes.

Based on the considerations presented in [6] we would expect that $mBIC2$ based methods should control the FDR at a level of .0085. For scenarios II, III and IV all $mBIC2$ based methods show a FDR of less than .011 which is roughly what we would expect. Compared to PiMass the FDR control of MOSGWA is much stricter. In scenario II and III the observed FDR of PiMass is roughly an order larger than the one observed for MOSGWA and in scenario IV it is nearly three times as big. But in absolute terms PiMass performed quite well with an FDR below .032 for all scenarios. In general the FDR of each method increases with the complexity of the scenario.

Now, let us turn our attention to the power to detect causal SNPs. Given that $mBIC2$'s ABOS property holds for moderate sample sizes, we would expect that MOSGWA reports less false negative results than PiMass, which in turn should yield a better power. That is exactly what we have observed. $mBIC2$ based methods dramatically outperform PiMass in every scenario in terms of power. This finding is even more apparent when we look at the selection pattern of MOSGWA and PiMass for scenario II (the patterns are similar for scenarios III and IV) as presented in Figure 3.2. MOSGWA shows a very regular, stable and desirable selection pattern. Highly influential SNPs were detected in nearly all replications, the biggest effect size with a detection rate below .95 was $\beta = .13$. The selection pattern shown by PiMass is quite different. For PiMass the selection frequency of a region shows no direct correspondence with the effect size of the causative SNP. For example, the selection frequency of a region with a causative SNP effect size of $\beta = .22$ (the third biggest effect in this scenario!) was practically zero, which is indisputable much lower than the detection frequency of 90% for a region with a causal SNP effect size of $\beta = .1$.

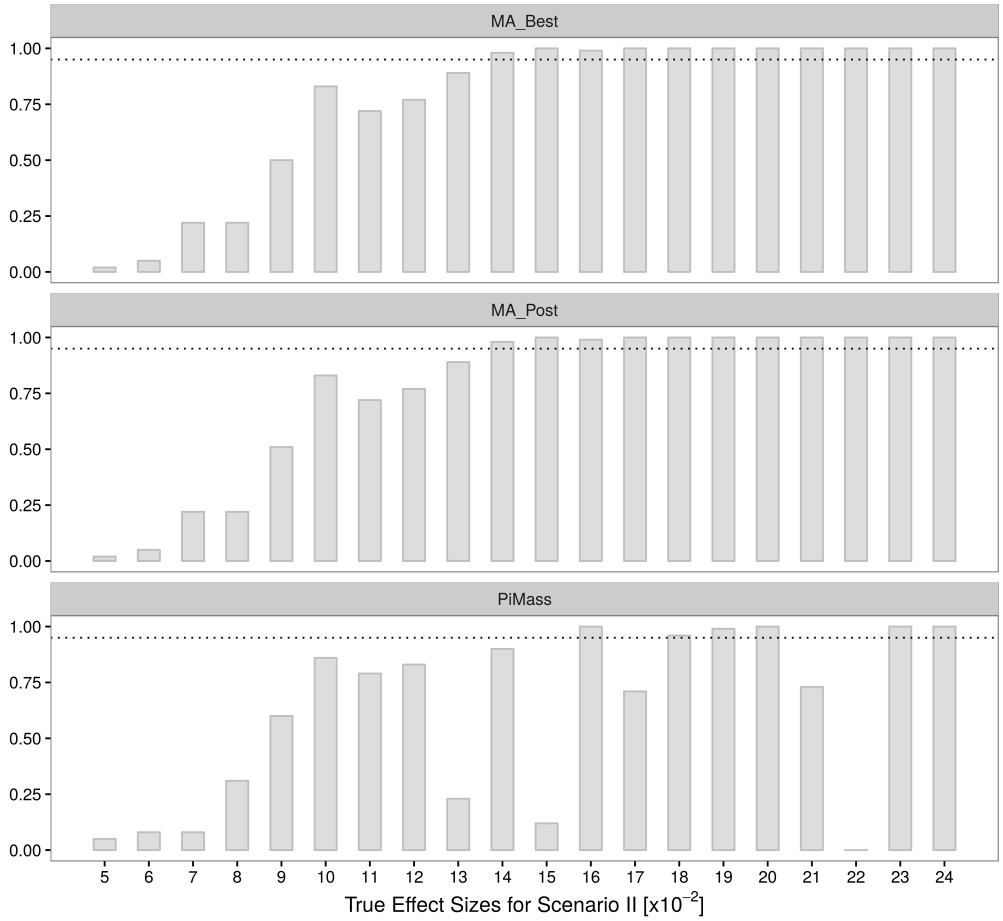

Figure 3.2: Illustration of the selection pattern of PiMass and MOSGWA. Columns display the selection frequency of a region around a causal SNP.

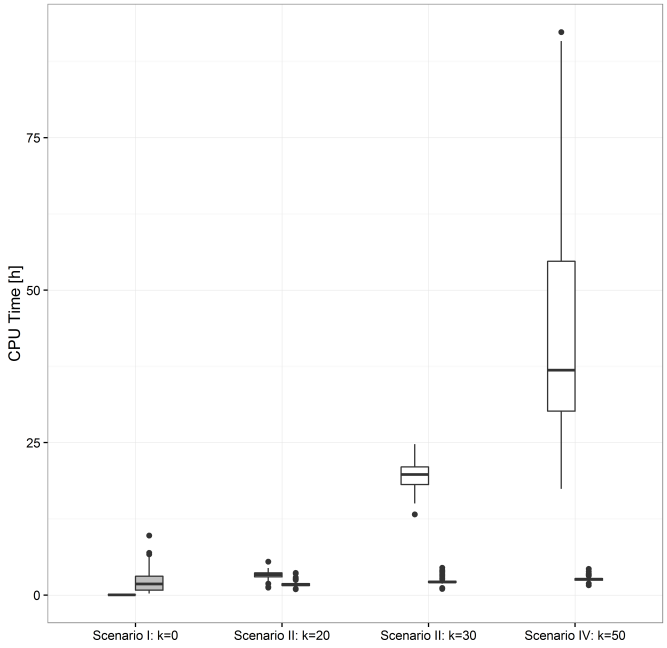

Figure 3.3: Runtime of PiMass (right, gray filled Boxplots) and MOSGWA in memetic mode (left, white filled Boxplots) for scenarios I-IV.

So far we have seen that MOSGWA clearly outperforms PiMass in any aspect. The only advantage that PiMass offers is a relatively moderate runtime (see Figure 3.3) which is nearly the same for all scenarios, whereas the runtime of MOSGWA's memetic algorithm grows nearly exponentially with the complexity of the considered scenario. However, limiting the maximal pool size will also bound the runtime of the memetic algorithm, which tends to be proportional to the number of visited models.

Finally we want to look at the heritability estimators. According to Figure 3.4 all heritability estimators have the tendency to underestimate the heritability in all scenarios. This is no surprise for MOSGWA h^2, which is basically the fraction of explained variance, because most of the selected models did not contain all causal regions. This bias increases with the complexity of the scenario for all methods. PiMass estimator h shows the most severe increase in bias, followed by PiMass hh which is nearly unbiased for scenario II but shows a strong bias for scenario IV. MOSGWA h^2 performs relatively best in the most complex scenario, nonetheless the estimator is severely biased.

In all three scenarios MOSGWA h^2 provides credibility intervals which are by a factor 5 to 10 smaller than the estimators offered by PiMass. In general the length of the credibility intervals of h^2 was very small with respect to the point estimator and approximately

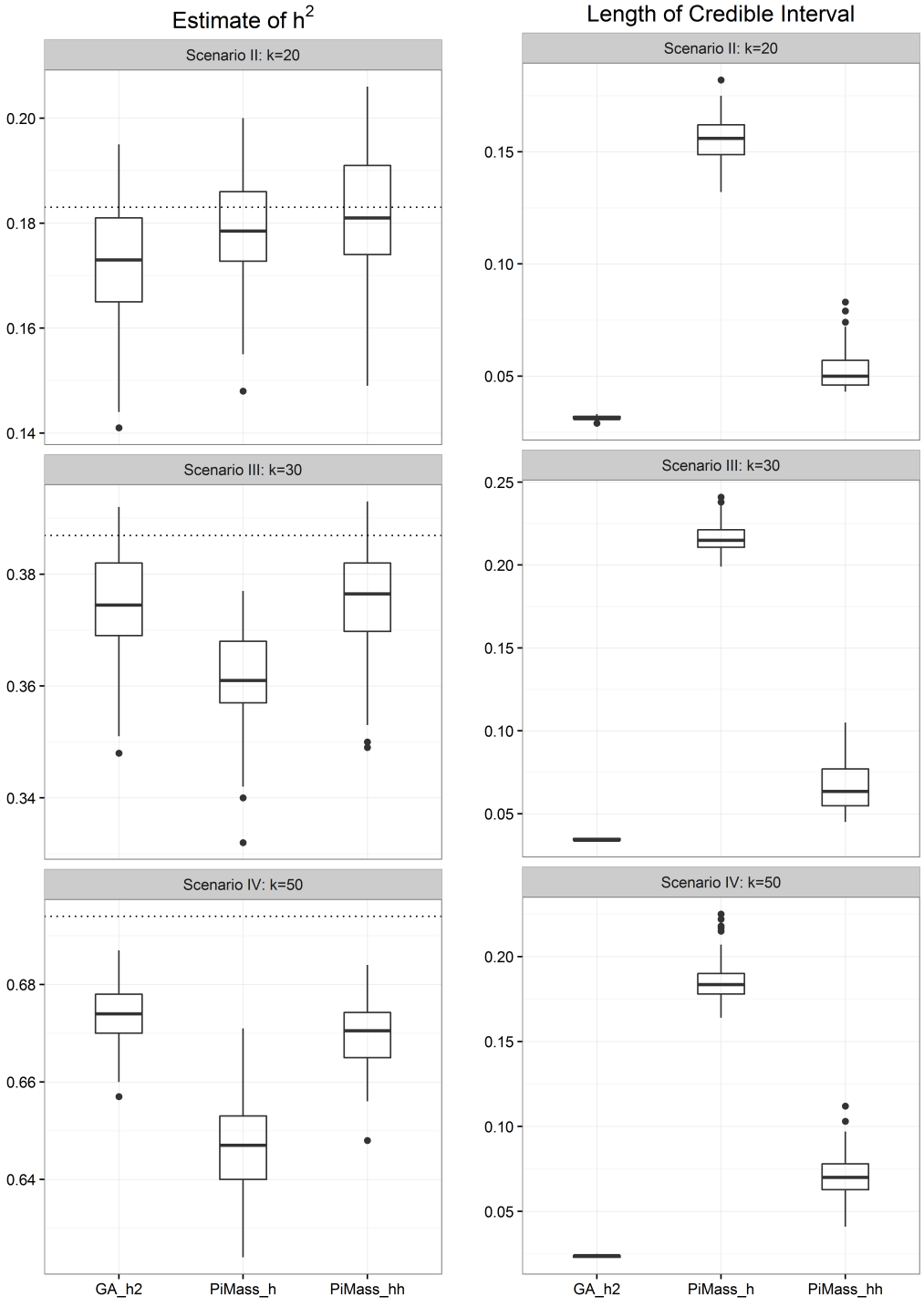

Figure 3.4: Estimates of heritability (left column) and length of credibility interval (CI) (right column). The dotted line represents the true heritability for a scenario.

equal in all scenarios. In combination with the bias described above this led to a lower coverage compared to the interval estimators offered by PiMass (see Table 3.2). The best coverage was observed for scenario III where the true value was in the credibility interval in 73 out of 100 runs. Still this is far below the nominal level of .95. The shortness of

Scenario	**MA_h^2**	**PiMass_h**	**PiMass_hh**
II	.67	1.00	.97
III	.73	1.00	.97
IV	.13	1.00	.68

Table 3.2: Frequency of the event that the true heritability value was in the credibility interval.

the credibility intervals obtained with the genetic algorithm is most likely due to the fact that the algorithm visits a too homogeneous or too small subpopulation of the model space. This finding will be revisited and related to other results in the Discussion section.

Now we want to compare the performance of MOSGWA with the very common strategy of Benjamini-Hochberg (BH) adjusted single marker tests. What is really striking is the fact that BH seems to completely fail to control the FDR at the nominal level (see Table 3.1) for the scenarios II to IV and that the average number of false positives is 100 to 1000 times bigger compared to $mBIC2$ based methods. In terms of our definition of FP this means, that BH selects a large number of SNPs that are not in the neighborhood of a causal SNP. This behavior of single marker based test strategies has already been observed by Frommlet et al.[18]. The key to understand this undesired behavior is the actual distribution of the single marker test statistics under a complex model (see section 1.2.2 for a principle sketch of the argument). Their important finding was that the selection probability is essentially controlled[1] by the noncentrality parameter ν of the model sum of squares (MSS) distribution. For a noncausal SNP j in a given finite sample this parameter is

$$\sqrt{\nu_j} = \left| \frac{1}{\sqrt{\hat{\mathrm{Var}}\left(X_j\right)}} \sum_{l \neq j} \beta_l \hat{\mathrm{Cov}}\left(X_l, X_j\right) \right|,$$

where β_l denotes the effect of SNP l on the phenotype under consideration. This fact reveals that noncausal SNPs are "charged" via sample correlations with causal SNPs.

Let us consider the case that the causal SNPs are statistically independent from each other and that the noncausal SNP j is not in linkage with any of them. Then the sum

[1] A larger noncentrality parameter implies a larger selection probability.

above is a random variable with expected value 0 and variance $\sum_{l \neq j} \beta_l^2 \operatorname{Var}\left(\hat{\operatorname{Cov}}\left(X_l, X_j\right)\right)$. Thus, the probability that the statistical test for a nonindicative SNP j has an associated noncentrality parameter that is large enough to give a significant selection probability for a nonindicative SNP is nonzero, and increases with trait complexity due to the increasing number of nonzero β_j. Given the huge number of nonindicative SNPs in a typical GWAS one can expect that a certain number of them will be selected, even if the probability for a single SNP is rather small. So, some of the false positives are weak linkage SNPs, but far more troubling, we can conjecture that a significant portion of false positives are entirely nonindicative SNPs.

The validity of this argument can be clearly seen in Figure 3.5, where the observed relationship between the noncentrality parameter and the selection frequency of false positive detections is depicted. Further, this argument also explains why the false positive rate is higher in more complex scenarios. In contrast to the single marker tests MOSGWA shows only a small increase of false positives with increasing trait complexity.

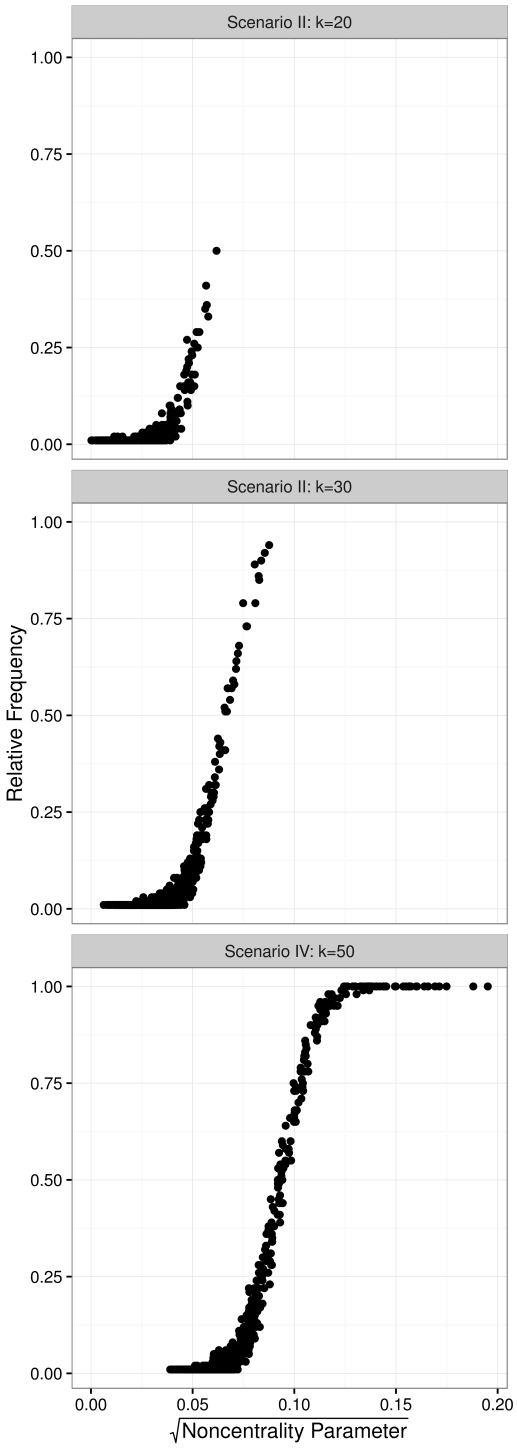

Figure 3.5: Illustration of the relationship between the noncentrality parameter of the single marker test statistic distribution of a noncausal SNP and the frequency of false positive occurrence in the 100 simulation runs for scenario II to IV.

3.2 Real Data Reanalysis

Rudimentary sample characteristics are presented in Table 3.3. We can see that the Genomic Inflation Factor *lambda*[2] (based on the population structure adjusted single marker tests) is approximately 1 for all phenotypes. This indicates that the p-values of the single marker based analysis for all phenotypes are not distorted by an unaccounted population structure. By virtue of the fact that we have applied the same correction strategy for all methods, these statistics also indicate that the obtained results are nearly unaffected by population structure effects for all methods. We will now discuss the results for each phenotype.

Pheno	Variants	N_tot	N_male	N_female	lambda
BMI	324310	4915	2491	2424	1.02921
CRP	324310	5257	2514	2743	1.02091
DIA	324310	5242	2511	2731	1.03531
GLU	324310	5239	2500	2739	1.01976
HDL	324310	4835	2282	2553	1.04597
INS	324310	5212	2482	2730	1.00841
LDL	324310	4821	2274	2547	1.05264
SYS	324310	5251	2512	2739	1.03794
TG	324310	4834	2281	2553	1.01447

Table 3.3: Sample characteristics for each phenotype.

The selected SNPs for each phenotype are tabulated in Table 3.4 whenever at least one method selected a SNP for a particular phenotype. We can see that for both blood pressure measures (*SYS* and *DIA*), Body Mass Index (*BMI*) and insulin level (*INS*) no method was able to detect a causative region.

We observed consensus about the causative regions among the methods for *CRP*, but we have to note that PiMass found the region indicated by SNP rs1169300 only four out of five times.

The results show a different picture for *GLU*. For this phenotype MOSGWA selected five regions which were also selected by Bonferroni adjusted single marker tests. Out of these five regions PiMass reported only four and no more. We also note that the results for PiMass are not identical for all replications and that the Benjamini-Hochberg adjusted single marker tests reported one additional region compared to MOSGWA.

[2]The Genomic Inflation Factor $\lambda := \text{median}\left(\chi^2\right)/0.456$ proposed by Devlin [12] tries to quantify to what extent a collection of χ^2 based test statistics suffer from distortions caused by population structure. A Genomic Inflation Factor near 1 indicates that no overall population structure effect is present.

SNPid	chr	pos	Post	Best	Greedy	PiMass	BH	Bon	Sabbati
				CRP					
rs12753193	1	65942267							5.1E-06
rs2794520	1	157945440	5	5	x	5	x (1)	x (1)	1.3E-19
rs1169300	12	119915608	5	5	x	4	x (5)	x (5)	1.1E-08
				GLU					
rs2025934	1	192001523	5	5	x	3	x (1)	x (1)	
rs560887	2	169471394	5	5	x	4	x (1)	x (1)	4.3E-10
rs3798004	5	9423945	5	5	x		x (1)	x (1)	
rs35781869	6	32661960	5	5	x	2	x (10)	x (6)	
rs10244051	7	15030358							1.5E-06
rs7858883	9	70481450	5	5	x	1	x (1)	x (1)	
rs1447352	11	92362409							8.4E-08
rs3794687	16	83684042					x (1)		
				HDL					
rs6728178	2	21047434					x (4)		
rs10096633	8	19875201				1	x (1)		
rs2740486	9	106706334				2	x (1)		
rs7120118	11	47242866	5	5	x	4	x (3)	x (1)	2.2E-07
rs1532085	15	56470658	5	5	x	5	x (3)	x (3)	2.7E-11
rs3764261	16	55550825	5	5	x	5	x (4)	x (4)	8.2E-29
rs255052	16	66582496	5	5	x	1	x (19)	x (9)	2.1E-07
rs1800961	20	42475778					x (1)		
				LDL					
rs207150	1	55579053				2	x (3)		
rs646776	1	109620053	5	5	x	5	x (2)	x (2)	7.3E-11
rs4844614	1	205941798				2	x (1)		1.1E-06
rs1713222	2	21124828	5	5	x	5	x (12)	x (10)	3.2E-12
rs945559	10	89813127				1			
rs174556	11	61337211				3	x (8)		1.3E-07
rs2228671	19	11071912	5	5	x	5	x (3)	x (2)	7.2E-09
rs157580	19	50087106	5	5	x	5	x (1)		1.1E-08
				TG					
rs3923037	2	21011755				1	x (1)		3.4E-07
rs1260326	2	27584444	5	5	x	5	x (2)	x (2)	5.2E-11
rs10096633	8	19875201	5	5	x	3	x (1)	x (1)	9.5E-09
rs12805061	11	116058235					x (2)		

Table 3.4: Indicated regions for all phenotypes with at least one reported region. Reported SNPs that are within 1.5 MBP have been summarized in a single region which is represented by the most frequent SNP. For deterministic algorithms x marks a selection followed by the number of selected SNPs in a region. For random algorithms the reported number indicated how many times this region was detected. For both adjustment procedures an adjusted p-value of .05 was regarded as significant. We also include the p-values reported by Sabbati et al. [30] if the p-value is of order 10^{-6} or smaller.

The results for *HDL* were also ambiguous. MOSGWA reported four causative regions which were found all the time. The same regions were also reported by Bonferroni adjusted single marker tests. PiMass indicated more regions than MOSGWA, but most of the findings could not be replicated in each of the five runs, especially those that are additional findings compared to MOSGWA. Benjamini-Hochberg adjusted single marker tests reported more regions than MOSGWA.

The results for *LDL* show a similar picture. MOSGWA reported the (same) four regions in all replications. Bonferroni corrected single marker tests have found only three of those four regions and no more. Again, the results for PiMass were very unstable. Of all the regions indicated by PiMass only those which were also reported by MOSGWA were replicated all the time. The replication rate for the additional four regions was very poor. Again, and as expected, Benjamini-Hochberg adjusted single marker tests reported more regions than MOSGWA.

MOSGWA indicated two regions for *TG* in all replications. These regions were also reported by Bonferroni adjusted single marker tests. PiMass reported an additional region, but this finding was reported in only one of the five replications. Benjamini Hochberg adjusted single marker tests reported more regions than MOSGWA.

Let us start our synopsis of the results with the observation that we have seen no difference in the performance of the three search strategies implemented in MOSGWA. In general the findings reported by MOSGWA have shown to be very robust and reproducible. This favorable behavior set MOSGWA apart from PiMass. The regions reported by PiMass have proven to be highly unstable and changed dramatically when the algorithm was applied repeatedly to the same data set. This converse behavior of the two procedures can also be seen in the estimated posterior probabilities obtained by PiMass and MOSGWA, which are depicted in Figure 3.6. In this figure we see that estimated posterior probabilities are much more pronounced and pointed for MOSGWA. This observation could be an indicator that the Markov chain of the sampler implemented in PiMass is not in equilibrium (or at least not long enough) and requires additional runtime.

When both procedures nominally control their corresponding generalized Type-I error rate at a level of .05, we have observed that Bonferroni adjusted single marker tests in total report one region less than MOSGWA and that Benjamini-Hochberg adjusted single marker tests report more. Both findings are quite interesting, especially when we take into account what we have learned from the simulation experiments. At a first glance is seems quite obvious that $mBIC2$ based selection performs better than Bonferroni adjusted single marker tests because the $mBIC2$ is designed to control the FDR which is a less stringent generalized Type I error rate than FWER, and therefore must be

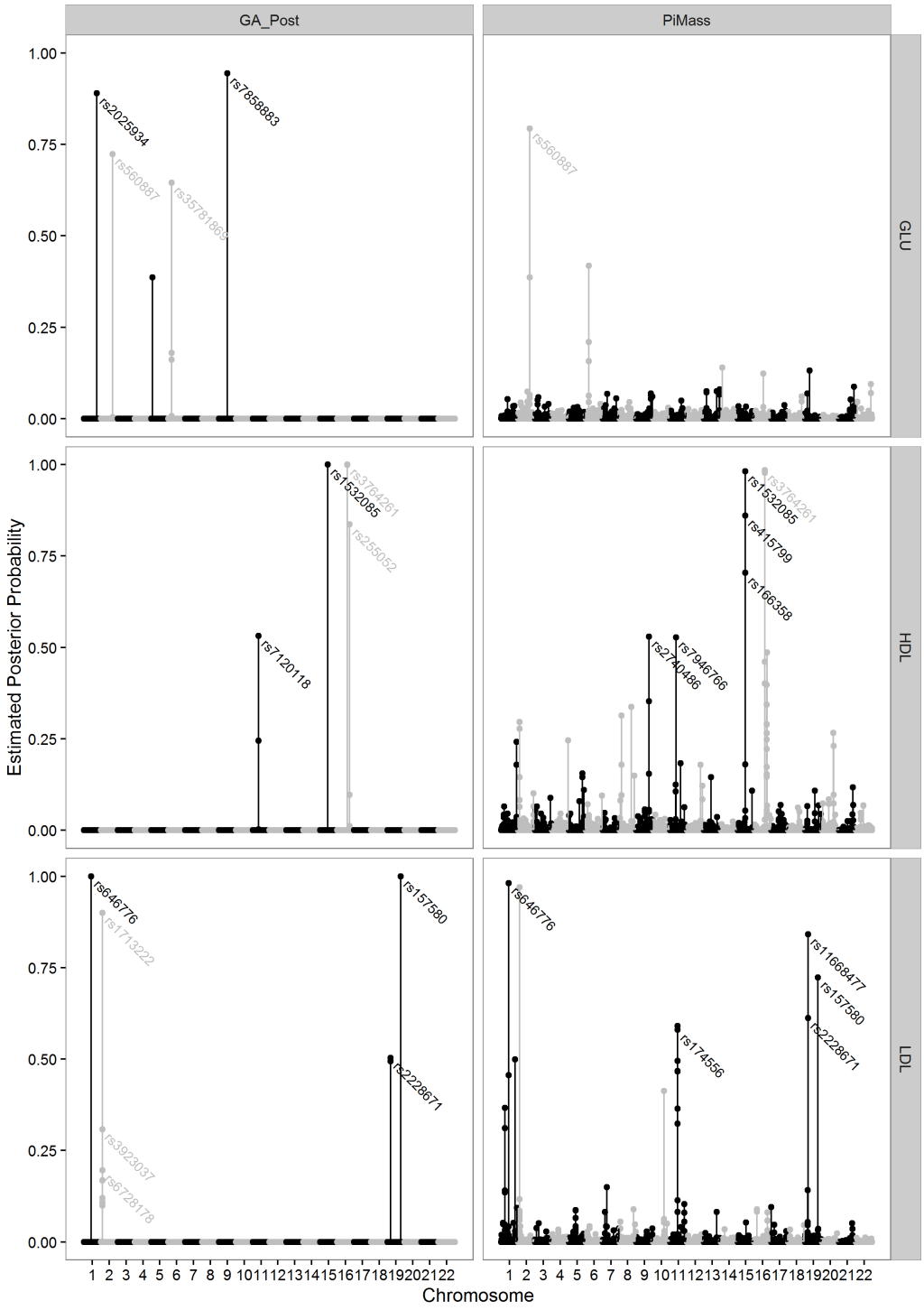

Figure 3.6: Estimated posterior SNP inclusion probabilities for selected phenotypes by PiMass and MOSGWA for the first replication.

more powerful. This statement would certainly be true, if MOSGWA and the Bonferroni procedure control the FDR respectively the FWER at approximately the same level. But, in the simulation experiment we have seen that MOSGWA controls the FDR at a much tighter level which is roughly ten times smaller than the nominal level applied for the Bonferroni adjustment in this analysis. Therefore, it is quite impressive that MOSGWA despite this stricter control shows a slightly higher power. If we set the nominal FWER level of the Bonferroni adjustment to .0082, which is according to [6] the approximate theoretical FDR level of $mBIC2$ based selection for this datasets, then MOSGWA reports four regions more then Bonferroni adjusted single marker tests (see column Bonf* of Table 3.5). It is also worthwhile to note that all of the additionally indicated SNPs are reported for LDL, HDL and GLU, which are apparently the most complex traits among the analyzed phenotypes. This is in line with the theoretical considerations presented in the previous sections which suggest that the benefits of $mBIC2$ based model selection compared to single marker based test strategies increase with the complexity of the underlying trait.

The reason for the second observation becomes clear in the light of the findings of the simulation study. There we have seen, that the Benjamini-Hochberg procedure dramatically fails to control the FDR at the nominal level. Even for a moderate complex trait (scenario II) the actual level was approximately 10 times larger than the nominal level. If we assume that the investigated traits are only influenced by a low to moderate number of genes (as it seems to be the case), we estimate that the actual FDR level of the Benjamini-Hochberg procedure is roughly .5. So we expect that nearly half of the detected signals are spurious. With this in mind, it seems reasonable to suspect that all the additional findings are actually false positives.

We can (not in a very strict sense) test this conjecture when we set the nominal FDR to a more adequate level, that facilitates a direct comparison of MOSGWA and Benjamini-Hochberg adjusted single marker tests results, and notice which regions are not reported anymore. If we set the nominal FDR level to .0082 which approximately corresponds to the theoretical FDR of $mBIC2$ based selection, we see that most of the additionally reported regions are gone (see column BH* of Table 3.5). Regarding the simulation experiment results, especially the excessive number of false positives reported by Benjamini-Hochberg adjusted single marker tests, we also considered the case of a nominal FDR level of .001 (see column BH** of Table 3.5). According to the simulation results presented in the previous section this level is clearly too strict for the low complex traits CRP and TG, but for the more complex traits LDL, HDL and GLU this nominal level will roughly yield an actual FDR level that corresponds to MOSGWA. At this level

all the additionally reported regions of the Benjamini-Hochberg adjusted single marker tests are gone. Actually MOSGWA reports more regions. But both procedures still agree on most of their shared findings. In toto these results support the hypothesis, that most (if not all) of the additional findings reported by Benjamini-Hochberg adjusted single marker tests are spurious.

SNPid	chr	pos	MOSGWA	Bon*	BH*	BH**
			CRP			
rs2794520	1	157945440	x (1)	x (1)	x (1)	
rs1169300	12	119915608	x (1)	x (2)	x (5)	x (2)
			GLU			
rs2025934	1	192001523	x (1)		x (1)	x (1)
rs560887	2	169471394	x (1)	x (1)	x (1)	x (1)
rs3798004	5	9423945	x (1)		x (1)	
rs35781869	6	32661960	x (1)	x (6)	x (7)	x (6)
rs7858883	9	70481450	x (1)	x (1)	x (1)	x (1)
rs3794687	16	83684042			x (1)	
			HDL			
rs6728178	2	21047434			x (4)	
rs7120118	11	47242866	x (1)		x (2)	
rs1532085	15	56470658	x (1)	x (1)	x (3)	x (2)
rs3764261	16	55550825	x (2)	x (4)	x (4)	x (4)
rs255052	16	66582496	x (1)	x (5)	x (12)	x (9)
rs1800961	20	42475778			x (1)	
			LDL			
rs207150	1	55579053			x (1)	
rs646776	1	109620053	x (1)	x (2)	x (2)	x (2)
rs1713222	2	21124828	x (3)	x (7)	x (12)	x (7)
rs174556	11	61337211			x (2)	
rs2228671	19	11071912	x (1)	x (2)	x (2)	x (2)
rs157580	19	50087106	x (1)		x (1)	
			TG			
rs1260326	2	27584444	x (1)	x (2)	x (2)	x (1)
rs10096633	8	19875201	x (1)	x (1)	x (1)	

Table 3.5: Indicated regions for all phenotypes with at least one reported region. Reported SNPs that are within 1.5 MBP have been summarized in a single region. x marks a selection followed by the number of selected SNPs in a region. For Bonf* and BH * an adjusted p-value of .0082, which is approximatly the theoretical FDR of $mBIC2$ based selection, was regarded as significant. For BH** an adjusted p-value of .001, which permits a direct comparison with MOSGWA, was regarded as significant.

4 DISCUSSION

In this thesis I have described a model selection approach to genome wide association studies (GWAS) using modifications of the Bayesian Information Criterion (BIC) which is based on sound theoretical considerations (see Section 1.3.2.4). Elementary statistical arguments (see Section 1.2.2) suggest that this approach should be a more powerful strategy to analyze GWAS data for complex traits than the frequently used single marker based test strategy. One of the problems of model selection in high dimensional datasets is the astronomical size of the potential model universe. Consequently, full enumeration is not a feasible option and model selection must be based on search heuristics. So, the actual performance of a model selection criterion is a compound of its theoretical properties and the behavior of the implemented search strategy for the required optimization.

For $mBIC2$ based selection I have described two different search algorithms which are implemented in the MOSGWA software package. One is a very elaborated greedy algorithm called fast stepwise search (see Section 2.4.1) and the other one is a variant of a memetic search algorithm (see Section 2.4.2). For the latter two distinct modes of final SNP selection exist. One is simply to consider the model with the smallest selection criterion visited by the memetic algorithm, and the other one is based on the estimated posterior inclusion probabilities of each candidate SNP.

In order to assess the performance of $mBIC2$ I conducted a simulation study and reanalyzed a real dataset (see Section 2.5 for details). For the simulation study I contrasted the $mBIC2$ based findings with those obtained by the Bayesian variable selection model implemented in PiMass and Benjamini-Hochberg adjusted single marker tests. For the real data example I additionally considered Bonferroni adjusted single marker tests. I omitted a more in depth comparison with the original analysis because the published data available at dbGaP deviates significantly from the data analyzed in the original publication [30].

The findings of the simulation study demonstrated an overall superior performance of $mBIC2$ based model selection compared to PiMass. $mBIC2$ based selection detects more causal SNPs and has a tighter FDR control resulting in a much lower average number of misclassified SNPs. Furthermore it has demonstrated a very desirable selection

pattern and the ability to control the false positive detection rate scales very well with trait complexity. These findings suggest that the asymptotic optimality properties of $mBIC2$ can be relied upon at a typical GWAS sample size. What was quite surprising was the fact that the observed results were practically unaffected by the applied search strategy. A possible reason for this could be some weakness in the construction of the initial population of the memetic algorithm that I discovered during my work. I have passed this finding on to the maintainer of MOSGWA who implemented an improved initial procedure whose performance is currently examined.

We have also seen that MOSGWA outperforms PiMass in terms of the reproducibility of the reported SNPs in the real data example. In both experiments we have seen that PiMass produces shaky and very unstable findings. A possible explanation for this contrasting behavior may be the different way the implemented search strategies walk through the search space. The memetic search in MOSGWA is strongly bound to sample models with high likelihoods and to move only through this subspace of the model universe. In contrast PiMass implements a MCMC sampler that walks through the model space more erratically and therefore the visited models are much more heterogeneous in terms of their likelihood. Because most of the models in the model universe possess an extremely low likelihood, the sampled Markov chain contains a high proportion of models that are practically neglectable for the estimation of model posterior probabilities. In consequence just a small fraction of models dominates the estimation. This leads to more fluctuation and stronger biased estimated posterior inclusion probabilities as we have seen, even if the number of visited models is nominally bigger (as it is the case in the presented experiments). For that reason on might conclude in general that a simple MCMC sampler is not well suited for model selection in high dimensional settings and that algorithms which are bound to search in the subset of "good" models, like the proposed memetic algorithm, should be superior to this approach.

The only thing left to discuss is the higher number of regions indicated by Benjamini-Hochberg adjusted single marker tests compared to MOSGWA in the real data example. There exist good arguments to assume that this additional findings are all false positives. In the simulation study we have seen that the FDR for $mBIC2$ based methods is roughly .004 for scenario II (which is most comparable to the complexity of the traits in the real data example). This is more than ten times as strict as the nominal FDR level of .05 we have chosen for the Benjamini-Hochberg adjusted single marker tests, and much stricter than the actual FDR level which can be assumed to be around .5. When the nominal FDR for the Benjamini-Hochberg procedure is set to the much stricter level that permits a direct comparison these additionally detected SNPs are all gone. However, we can not

be certain if these additionally indicated SNPs are true or false positives. But in theory $mBIC2$ based model selection performs asymptotically as good as a Bayes oracle for a wide range of sparse asymptotic regimes and we have seen in the simulations that this property holds for a typical GWAS. So it is not unreasonable to conjecture that $mBIC2$ based model selection is approximately a Bayes optimal classifier and therefore most of the additional findings by BH will be false positives.

Nonetheless this raises a point. In a discovery context where false positives are far less important than possible findings one could argue that the property of $mBIC2$ to control the FDR proportional to $1/\sqrt{n}$ is too strict, especially when the sample size is bigger than a few hundredth. So it seems desirable to create a selection criterion that controls the FDR at a nominal level. Such a criterion is currently developed and will be discussed in an upcoming publication.

Bibliography

[1] Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Special invited lecture: adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, pages 584–653, 2006.

[2] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[5] Malgorzata Bogdan, Jayanta K Ghosh, and RW Doerge. Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167(2):989–999, 2004.

[6] Malgorzata Bogdan, Jayanta K Ghosh, and Malgorzata Zak-Szatkowska. Selecting explanatory variables with the modified version of the bayesian information criterion. *Quality and Reliability Engineering International*, 24(6):627–641, 2008.

[7] Karl W Broman and Terence P Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656, 2002.

[8] Brian L Browning and Sharon R Browning. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic epidemiology*, 31(5):365–375, 2007.

[9] Sharon Browning and Brian Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5):1084–1097, July 2007.

[10] Sharon R Browning. Multilocus association mapping using variable-length markov chains. *The American Journal of Human Genetics*, 78(6):903–913, 2006.

[11] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[12] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[13] Erich Dolejsi, Bernhard Bodenstorfer, and Florian Frommlet. Analyzing genome-wide association studies with an fdr controlling modification of the bayesian information criterion. *PloS one*, 9(7):e103322, 2014.

[14] F. Frommlet, M. Bogdan, and A. Chakrabarti. Asymptotic bayes optimality under sparsity of selection rules for general priors. *Technical Report*, arXiv:1005.4753, 2010.

[15] Florian Frommlet, Malgorzata Bogdan, and Ramsey David. *Phenotypes and Genotypes: The Search for Influential Genes*. Springer, 2016.

[16] Florian Frommlet, Arijit Chakrabarti, Magdalena Murawska, and Malgorzata Bogdan. Asymptotic bayes optimality under sparsity for generally distributed effect sizes under the alternative. *arXiv preprint arXiv:1005.4753*, 2010.

[17] Florian Frommlet, Ivana Ljubic, Helga Björk Arnardóttir, Malgorzata Bogdan, et al. Qtl mapping using a memetic algorithm with modifications of bic as fitness function. *Stat Appl Genet Mol Biol*, 11(4):2, 2012.

[18] Florian Frommlet, Felix Ruhaltinger, Piotr Twarog, and Małgorzata Bogdan. Modified versions of bayesian information criterion for genome-wide association studies. *Computational Statistics & Data Analysis*, 56(5):1038–1051, 2012.

[19] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[20] Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, pages 1780–1815, 2011.

[21] Frank E Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media, 2013.

[22] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[23] Yosef Hochberg and Ajit C Tamhane. Multiple comparison procedures. 2009.

[24] Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[25] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[26] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[27] Ryuei Nishii et al. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

[28] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[29] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46, 2009.

[31] Sanat K Sarkar et al. On the simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, pages 231–242. Institute of Mathematical Statistics, 2008.

[32] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[33] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[34] Daniel Stram. *Design, Analysis, and Interpretation of Genome-Wide Association Scans*. Springer-Verlag New York, 2014.

APPENDIX

Abstract

Even nowadays multiple comparison adjusted single marker tests are the most commonly applied strategy to analyze genome wide association studies (GWAS). Elementary statistical considerations demonstrate that this strategy is highly suboptimal in terms of power to detect causative regions on the genome. Especially if the phenotype of interest is a complex trait. A potentially more powerful strategy is the application of model selection for multi marker regression models. I discuss a model selection criterion ($mBIC2$) which is an adaption of the Bayesian Information Criterion (BIC) to high dimensional statistics. This modification is based on sound statistical theory, and guarantees that model selection based on $mBIC2$ is asymptotically a Bayes optimal classifier for a wide range of relevant sparse asymptotic regimes. A prevailing difficulty for model selection in the context of high dimensional datasets is the tremendous size of the potential model universe. In fact this number is so big that enumeration is not a feasible option anymore. In order to find the optimal model search heuristics must be applied. I present two methods for $mBIC2$ based model selection which are currently implemented in the MOSGWA software package. One is a version of a greedy algorithm called fast stepwise selection and the other one is a novel memetic algorithm. Based on these search strategies I compare the performance of mBIC2 based model selection with a Bayesian variable selection model (implemented in PiMass) and typical single marker test strategies in a simulation study and a reanalyzed real data example.

The findings of these experiments demonstrated an overall much better performance of $mBIC2$ based model selection compared to PiMass and single marker tests. $mBIC2$ based selection succeeds in all relevant performance measures. It detects more causal regions, has a tighter FDR control, a much lower average number of misclassified signals and shows a very desirable selection pattern for all search strategies.

Zusammenfassung

Selbst heutzutage sind Single-Marker Tests nach wie vor die gängigste Analysestrategie zur Auswertung Genomweiter Assoziationsstudien (GWAS). Elementare statistische Argumente führen jedoch zu dem Urteil, dass diese Auswertestrategie hochgradig ineffektiv ist um kausale Marker aufzuspüren. Dies gilt im Besonderen dann, wenn der zugrundeliegende Phänotyp durch eine Vielzahl genetischer Marker bestimmt wird. Ein Weg die Effektivität der Auswertung zu steigern ist, das zugrundeliegende Problem als Modelselektion aufzufassen. Somit rückt also die Suche nach dem (im Sinne eines Kriteriums) besten Regressionsmodell zur Erklärung des zugrundeliegenden Phänotyps in den Fokus. Dabei werden die einzelnen Marker nicht mehr für sich getrennt, sondern gebündelt betrachtet, man spricht daher von einem Multi-Marker Ansatz. Zu diesem Zweck stelle ich das Selektionskriterium $mBIC2$ vor. Bei diesem Kriterium handelt es sich um eine theoretisch wohlfundierte Anpassung des Bayesian Information Critera (BIC) für hochdimensionale statistische Daten mit herausragenden asymptotischen Eigenschaften.

Eine nicht zu übersehende Schwierigkeit bei der Modellselektion in hochdimensionalen Daten ist die überwältigende Anzahl der möglichen Modelle. Deren Anzahl ist so groß, dass die vollständige Enumeration selbst mit zeitgenössischen Rechnern nicht möglich ist. Es bleibt daher keine andere Option als dieses Optimierungsproblem heuristisch zu lösen. In der vorliegenden Arbeit stelle ich zwei Heuristiken vor die es erlauben $mBIC2$ basierte Modellselektion in GWAS durchzuführen. Bei der einen Suchstrategie handelt es sich um eine bereits erprobte Variante eines Greedy-Algorithmus, die zweite beschreibt eine grundlegend neuartige Variante eines Memetischen-Algorithmus.

Zur Evaluation der Performanz $mBIC2$ basierter Selektion – mit einem Fokus auf den memetischen Algorithmus – führte ich eine Simulationsstudie und die Reanalyse eine bereits veröffentlichten GWAS durch. In beiden Fällen wurden die Ergebnisse der Heuristiken untereinander und mit alternativen Auswertungsmethoden verglichen. Die betrachteten Alternativen waren Bonferroni bzw. Benjamini-Hochberg adjustierte Single-Marker Tests und eine Bayesianisches Variablen Selektionsmodell (implementiert in PiMass).

Zusammengefasst zeigte sich eine deutliche Überlegenheit der $mBIC2$ basierten Selektion. Unabhängig von der Optimierungsheuristik zeigten diese eine mit Abstand höhere Power, eine niedrigere FDR sowie eine deutlich niedrigere Anzahl an falsch klassifizierten Markern.