Linear mixed models for genome-wide association studies

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Christoph Lippert

aus Diez

Tübingen

2013

Tag der mündlichen Qualifikation:	29.11.2013
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Karsten Borgwardt
2. Berichterstatter:	Prof. Dr. Bertram Müller-Myhsok

Abstract

Genome-wide association studies aim at uncovering genetic loci that regulate a phenotype of interest by performing statistical tests for association between observed genetic variants and the phenotype. However, confounding factors like population structure, family relatedness, and cryptic relatedness often lead to false positive findings, if not accounted for in the analysis. Linear mixed models are among the richest class of models used today for genome-wide association studies, and in contrast to other methods have been shown to be capable of to capture all of these forms of relatedness simultaneously, without knowledge of which are present and without the need to tease them apart. Despite their benefits the use of linear mixed models so far has been limited to smaller studies, due to the large computational burden. In this thesis, we investigate linear mixed models for genome-wide association studies and present new algorithms to scale up linear mixed model computations that thereby enable their use for the analysis of extremely large genome-wide association studies for the first time. Besides algorithmic contributions we also present improvements to the statistical modeling part, that lead to an increase in power and better calibration over the traditional use of linear mixed models. Based on these improvements, we investigate association tests for single as well as multiple genetic variants and a phenotype. Finally, we conclude by with a multivariate version of the linear mixed model that allows simultaneous analysis of multiple related traits.

Zusammenfassung

Das Ziel von genomweiten Assoziationsstudien ist es, genetische Loci zu finden, die einen Phänotyp regulieren, indem man statistische Tests zwischen gemessenen genetischen Varianten und dem Phänotyp durchführt. Allerdings ziehen Störgrößen, wie Populationsstruktur, Verwandtschaftsverhältnisse innerhalb Familien, oder unbekannte Verwandtschaften zwischen scheinbar unverwandten Individuen, wenn diese nicht in Betracht gezogen werden, die Gefahr von falsch positiven Ergebnissen in der Studie nach sich. Lineare gemischte Modelle gehören zu den komplexesten Modellen, die heutzutage in genomweiten Assoziationsstudien angewandt werden, da diese, im Gegensatz zu anderen Korrekturmethoden, in der Lage sind für all diese Störgrößen aufzukommen, ohne das explizite Wissen, welche davon vorkommen, und ohne diese auseinanderzudröseln zu müssen. Trotz der klaren Vorteile durch die Anwendung von linearen gemischten Modellen, war diese wegen des hohen Rechenaufwandes bisher auf kleinere Datensätze beschränkt. Diese Arbeit setzt sich mit linearen gemischten Modellen für genomweite Assoziationsstudien auseinander und stellt neue Algorithmen vor, die lineare gemischte Modelle hochskalieren und somit mit zum ersten mal die Analyse von extrem großen Datensätzen mit diesen Modellen ermöglichen. Neben diesen algorithmischen Beiträgen werden auch Verbesserungen auf der Seite der statistischen Modellierung von genomweiten Assoziationsstudiesn vorgestellt, welche im Vergleich zur traditionellen Anwendung von linearen gemischten Modellen zu mehr statistischer Power bei gleichzeitig besserer Kontrolle des Typ 1 Fehlers führen. Aufbauend auf diese Verbesserungen werden Assoziationstests von einzelnen als auch von mehreren genetischen Varianten vorgestellt und analysiert. Zum Abschluss der Arbeit wird eine multivariate Version von linearen gemischten Modellen zur Analyse von mehreren verwandten Phänotypen vorgestellt.

Acknowledgements

First and foremost I want to express my gratitude to my advisor Prof. Karsten Borgwardt for his excellent support and advice, not just on my research, but also on my professional development and on the writing of my thesis.

Thanks to the great interdisciplinary setup of his research group between two affiliated departments at the Max Planck Institute for Intelligent Systems and at the Max Planck Institute for Developmental Biology, I was in the lucky position to learn from and work with leading researchers in both machine learning as well as biology. These departments are led by Prof. Bernhard Schölkopf and Prof. Detlef Weigel, who I also want to thank for the advice they, as part of my PhD advisory committee, gave on my research.

Thank you Prof. Bertram Müller Myhsok for acting as an external reviewer on this thesis and Prof. Daniel Huson for acting as an examiner in my oral PhD examination.

I also want to thank all the other people whom I was lucky to work with very during the course of my thesis. These people include Oliver Stegle and Barbara Rakitsch whom I worked and still work on many projects. Oliver has also proofread parts of this thesis. Another thank you goes to Jennifer Listgarten and Prof. David Heckerman who I have worked with during my time as an intern at Microsoft Research in Los Angeles. I enjoyed working with these two so much that I have in the meantime moved over to Los Angeles.

Thank you to all the other students and postdocs in our group, especially Theofanis Karaletsos, Dominik Grimm, Nino Shervashidze, Chloé Azencott, for all those lively discussions.

Other people I interacted and worked with during the work on this thesis are Joris Mooij, Beth Rowan, George Wang, Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Carl Kadie, Bob Davidson, Eun Yong Kang, and Prof. Eleazar Eskin.

At the Max Planck Institute we also were supported by a great administrative team. Here I want to highlight the secretaries of the two affiliated departments, Sabrina Rehbaum and Hülya Wicher, as well as the system administrators Sebstian Stark, Martin Lang and Thomas Helle.

Funding for my PhD studies was provided by the Max Planck society.

In this thesis I made use of publicly funded data. A full list of the investigators who contributed to the generation of the Wellcome Trust Case-Control Consortium data is available from `http://www.wtccc.org.uk/`. Funding for the project was provided by the Wellcome Trust (076113 and 085475). The GAW14 data were provided by the members of the Collaborative Study on the Genetics of Alcoholism (US National Institutes of Health grant U10 AA008401).

Finally, I want to thank my family, my parents Ingrid and Hans-Wilhelm, and my beautiful wife Nora, for all their love and support. I love all of you.

Contents

1. Introduction

1.1. A chronological overview of genetics and association studies

1.1.1. The beginnings of genetics and Mendel's laws for simple traits

Since about hundred and fifty years geneticists try to understand the mechanisms that shape variation in heritable traits. The field started with Gregor Mendel, who observed how trait alleles are transmitted across related individuals. He soon understood, how simple discrete traits caused by a single mutation are inherited in diploid organisms, that is organisms having two distinct sets of chromosomes. From his observations he deduced two rules that described these "Mendelian" traits. The rule of "Segregation" states that for crosses of parents, each being homozygous carriers of one of the two trait alleles, then all direct descents will have the same trait as they are all heterozygous between the two parent alleles. The distribution of trait characteristics among descendent of heterozygous individuals like these on the other hand is given by a probability distribution over each of up to three distinct states for the trait[1]. Finally, the rule of "Independent Assortment" states that any two (unlinked) traits are inherited independent of each other [Mendel, 1866].

1.1.2. Development of statistical models for complex traits

Mendel's laws were able to explain inheritance of a wide range of traits and *rare diseases*. Douglas Galton, a cousin of Charles Darwin, analyzed many *quantitative* traits, including human height, intelligence and a range of behavioral traits by linear regression [Galton, 1869, Visscher et al., 2011]. The heritable nature of these traits has been known and utilized by animal and plan breeders for thousands of years. But for these traits, in contrast to Mendel's law of Segregation, Galton observed continuous blending inheritance and large environmental influences [Galton, 1897, 1898]. During his studies he invented a range of important concepts like statistical correlation or the standard deviation. He also came up with study designs that could be used to estimate heritability of a trait and the effects of environment. He measured the shared variation between twins to study the heritable variation in a trait. In order to control for the effects of shared environment he performed studies of adopted children [Burbridge, 2001, Bulmer, 2003].

Ronald Fisher correctly speculated that quantitative traits are influenced by a large number of unobserved mutations, each of which following Mendelian inheritance patterns and having a small additive effect on the trait. He showed that in the limit of an infinite number of unobserved random loci, each with an infinitesimally small effect on the trait, the trait would be normally distributed and the degree of trait covariation between

[1] For alleles A and a, these states are homozygous AA, homozygous aa and heterozygous Aa

related individuals is identical to the amount of their genetic material that is identical by descent. In this seminal work he unified the genetics of quantitative traits with Mendelian inheritance [Fisher, 1918].

Also, many heritable *common diseases*, despite being qualitative and showing familial aggregation, could not be explained in terms of Mendel's laws [Lobo, 2008]. Similarly to quantitative traits these could be explained by a *complex* polygenic inheritance of a large number of causal loci, each contributing a small amount to an unobserved normally distributed liability, which affects disease susceptibility and severeness of the disease [Wright, 1934a,b].

In the 1950s, building on Fisher's random effects model, Charles Henderson stated a system of equations that would subsequently revolutionize animal breeding. The solution of these "mixed model equations" provided an efficient criterion for artificial selection, as it yields an unbiased prediction of the unknown genetic component of a quantitative trait that is independent of environmental covariates [Henderson, 1950, 1984].

1.1.3. Linkage studies

Recombination [Morgan et al., 1922] (see Figure 1.1) leads to a reduction in genetic linkage between two loci that are far apart on the chromosome [Morgan et al., 1922]. This understanding allowed to build linkage maps by tracing the co-inheritance between known genetic markers across pedigrees [Griffiths et al., 2004]. Initially, such markers where phenotypes that followed Mendelian inheritance. Later, simple non-coding DNA sequences like microsatellites were used as markers.

In *linkage studies* use the insight that a causal locus cane mapped by tracing linked markers shared by affected related individuals over a pedigree [Morton, 1955]. Such studies successfully determined the genetic cause of many Mendelian traits including rare human diseases like Huntington's disease or Cystic Fibrosis [Chial, 2008].

Interventional gene knockouts allow to validate strong effects found by mapping studies by disrupting the functionality of a target gene and observing the resulting phenotype. These were a first steps towards the ultimate goal of genetic analyses: to gain functional knowledge that would allow for targeted intervention at the genomic level.

For many common diseases as well as quantitative traits, even though these showed strong heritability among relatives, linkage studies are usually underpowered in order to map a larger number of causal loci with small individual effects [Risch et al., 1996].

1.1.4. Genome-wide association studies

With the advent of high-throughput genotyping technologies, GWAS have recently emerged as a novel study design. Microarrays or extremely low-coverage sequencing are used to screen hundreds of thousands to millions of common single nucleotide polymorphism (SNP) markers at a population scale. The genetic variation is subsequently used to detect regions in the genome that are linked to causal variants by testing for statistical association between the individual SNP markers and the phenotype on a population level (see Figure 1.2). In contrast to linkage studies, which study related individuals to detect linked loci, in GWAS the individuals are usually assumed to be unrelated [Astle and Balding, 2009].

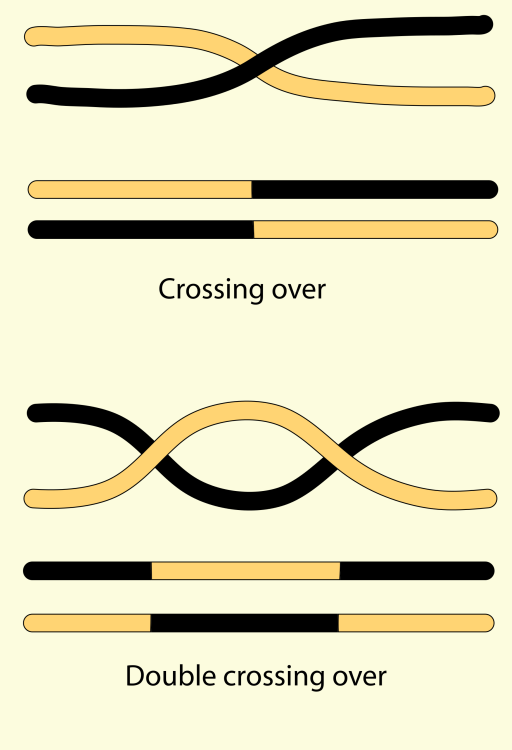


Figure 1.1. Recombination due to crossing over. During Meiosis, random crossing over of chromosomes lead to recombination of the chromosomes. The further the distance between two loci, the higher is the chance that a recombination event happens during meiosis. The resulting genetic distance is measured in centimorgans (cM) and refers to one meiosis out of one hundred causing recombination. The effect leads to a reduction of genetic linkage between loci that are far apart on the chromosome until they are in equilibrium [Morgan et al., 1922].[2]

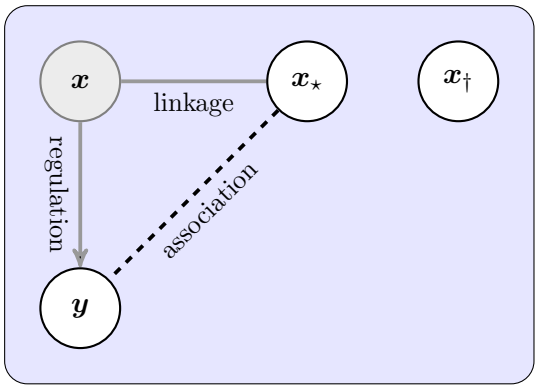


Figure 1.2. Graphical model representation of an idealized genome-wide association study involving a single causal variant. The state of the phenotype, denoted by $\boldsymbol{y}$ is caused by the state of a single unknown variant $\boldsymbol{x}$. While most of the markers $\boldsymbol{x}^\dagger$ are statistically independent of the phenotype, markers $\boldsymbol{x}^\star$ that are in in close proximity to the causal variant $\boldsymbol{x}$ assert an indirect association with the phenotype $\boldsymbol{y}$ due to linkage.

The population-based study design of GWAS has substantial practical advantages over family-based designs [Pritchard et al., 2000b]. Genotyping the relatives of individuals carrying a causal allele poses limitations on the size of a study compared to sampling unrelated individuals [Risch et al., 1996]. This limitation is especially true for studies of late-onset diseases. The use of unrelated individuals also allows to reduce genotyping costs by sharing individuals between studies of multiple quantitative traits or by re-using unaffected control individuals in studies of common diseases [Atwell et al., 2010, Huang et al., 2010, Burton et al., 2007].

The SNPs included on an array are chosen in a way that they most densely tag regions that are likely to be functional, including exonic regions and the regions upstream and downstream of genes. For low-density or low-quality samples the resolution can be increased further by imputation of missing loci based on the distribution that has been observed before in data that has been genotyped at a higher-resolution [de Bakker et al., 2008]. At this end population-scale re-sequencing projects [Altshuler et al., 2010, Autosomes Chromosome, 2012, Cao et al., 2011] aim at providing a complete view of the genetic variation that is present in a large sample, including SNPs, rare variants, insertions/delitions and copy-number variations. The data from these projects also helps to build maps of linkage-disequilibrium at the highest resolution possible.

GWAS of quantitative traits like human height, growth traits in plants and animals, or studies of molecular traits like gene expression have identified a large number of loci [Visscher, 2008, Atwell et al., 2010, Tian et al., 2011, Bolormaa et al., 2011, Kim et al., 2013]. Due to their study design, they have proven especially useful for detecting associations of common variants. Due to the polygenicity of common human diseases the selective pressure on individual disease-causing mutations is assumed to be reduced, leading to the *common disease, common variant hypothesis* [Lander et al., 1996, Reich and Lander, 2001, Pritchard and Cox, 2002] and the belief, that GWAS should be especially useful for these kinds of diseases (see Figure 1.3).

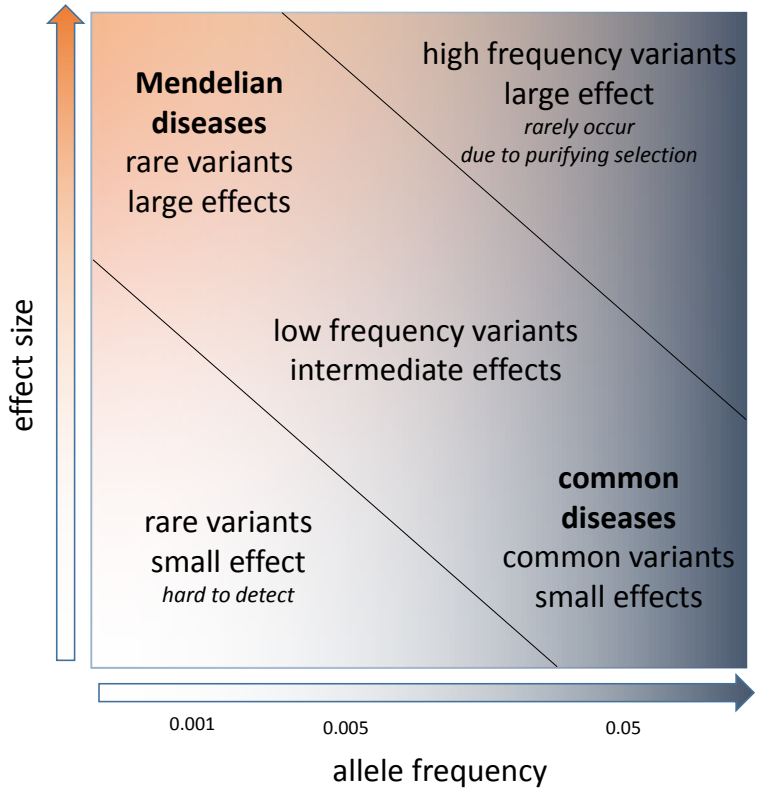


Figure 1.3. The allelic spectrum of human diseases and implications for genetic studies. The area between the two diagonal lines defines the targets of genetic studies of human disease. Mendelian diseases are usually caused by a single variant with large effect sizes that are rare due to strong selective pressures. Many of these have been detected by family studies. Variants with large effects on disease are usually rare due to strong selective pressures. Common diseases are assumed to be caused by multiple variants having smaller effect sizes due to reduced selective pressures. Common variants are valuable targets of GWAS. Rare variants with small effect sizes are hard to detect for realistic sample sizes.[3]

The number of loci that have been reliably associated with heritable human diseases is close to nine thousand. Since the publication of the first GWAS, the number of publications of human GWAS has constantly been growing every year [Hindorff et al., 2009, Manolio, 2010], already surpassing 1.5 thousand[4].

1.2. Challenges in modern genome-wide association studies

The sheer quantity of reported loci of course is an obvious success. And even though loci with weak effects reach genome-wide significance levels and reliably replicate, the practical value of these markers is in doubt. Personalized medicine based on genetics is already clinical practice for a small fraction of common diseases like breast cancer [Palma et al., 2006], but most of the currently available genetic tests are based on a small number of rare genetic variants with large effects on disease risk. These tests are generally only helpful for a small fraction of the affected individuals. One has to admit that the amount of functional knowledge that has been gained is limited, leading to justified criticism of the GWAS design [Visscher et al., 2012, Wade, 2010].

1.2.1. Missing heritability

Heritability is defined as the fraction of phenotypic variance that is caused by genetics and has traditionally been estimated from studies of families or twins and yields an upper bound on the amount of information that can be gained by genetic studies. A distinction is made between *narrow-sense* heritability (h^2) due to linear additive effects and the less commonly used *broad sense* heritability, which also measures non-additive effects like gene-gene interactions or interactions between genes and environment. Even though many common traits are highly heritable ($h^2 = 30\% - 80\%$) [Visscher et al., 2012], the genetic variants identified to date generally explain only a small fraction of the heritable trait variability as estimated from family studies, giving rise to the "missing heritability problem" [Maher, 2008]. Possible sources for missing heritability include effects not covered in the standard GWAS study design, like gene-gene or gene-environment interactions or effects of rare variants. On the other hand it is also hypothesized that estimates of heritability might be inflated due to confounding environmental effects in twin studies [Manolio et al., 2009].

By estimating narrow-sense heritability from genome-wide SNPs, Visscher and colleagues convincingly argued that most common diseases have a highly polygenic genetic architecture [Yang et al., 2010, 2011b, Stahl et al., 2012]. As a result of high polygenicity and the huge number of tests performed, GWAS need extremely large sample sizes to yield enough power to meet genome-wide significance for smaller effects [Manolio, 2010, Park et al., 2010]. The hypothesis that given a large enough sample size the complete heritable portion of a phenotype could be explained has already been shown to be true for the model organism yeast [Bloom et al., 2013].

[4]The numbers reported are taken from http://www.genome.gov/gwastudies (as of 3/15/2013).

1.2.2. Handling large volumes of data

In order to increase statistical power, researchers collect ever-increasing studies, in the tens of thousands to over a hundred thousand samples [Do et al., 2011]. Large international consortia are forming and combine their data into meta-analyses with total sample-sizes ranging well into the hundreds of thousands [Speliotes et al., 2010, Allen et al., 2010, Teslovich et al., 2010, Ehret et al., 2011]. These tremendous volumes of data impose huge technological challenges and requires the development of efficient algorithms and tools that enable experimental scientists to accurately analyze the data in a timely manner.

1.2.3. Confounding by population structure

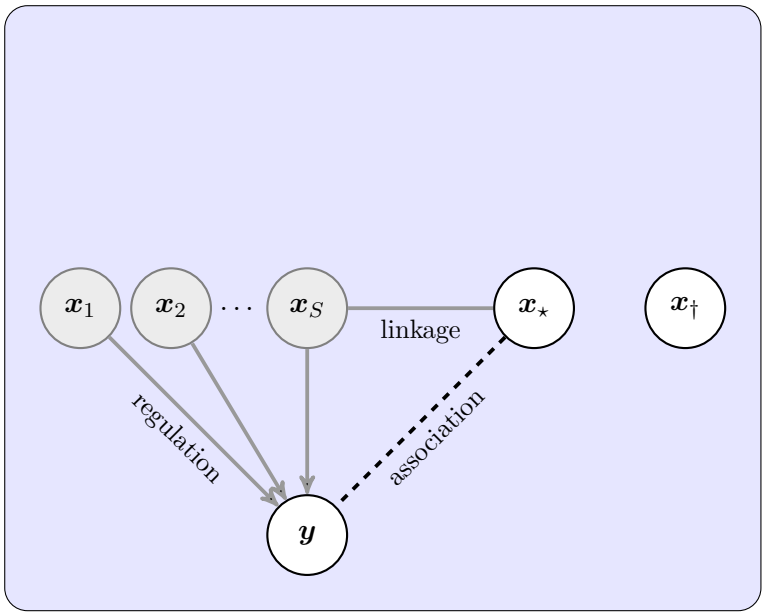
Another problem in GWAS are confounding factors like population structure, shared environment or technical artifacts, which have lead to false positive associations [McClellan and King, 2010, Lambert and Black, 2012, Devlin and Roeder, 1999, Pritchard et al., 2000b, Price et al., 2006].

In GWAS the individuals are assumed unrelated. Population structure, family structure and cryptic relatedness cause correlations between genome-wide SNPs. GWAS look for correlations between the phenotype and a marker linked to the causal variant. However, relatedness causes correlations between genome-wide loci including causal and non-causal variants, thereby causing spurious associations between the phenotype and unlinked markers (see Figure 1.4(b).). For example for a phenotype that is regulated by a single variant only, common inheritance of the causal variant and other SNPs yields statistical association between the phenotype and unlinked SNPs all over the whole genome [Ewens and Spielman, 1995, Pritchard and Rosenberg, 1999]. In extreme cases even for traits that are not genetically heritable cryptic relatedness can create false positive associations. An example would be a phenotype that depends on sociographic or geographic influences that correlate with population structure [Mathieson and McVean, 2012].

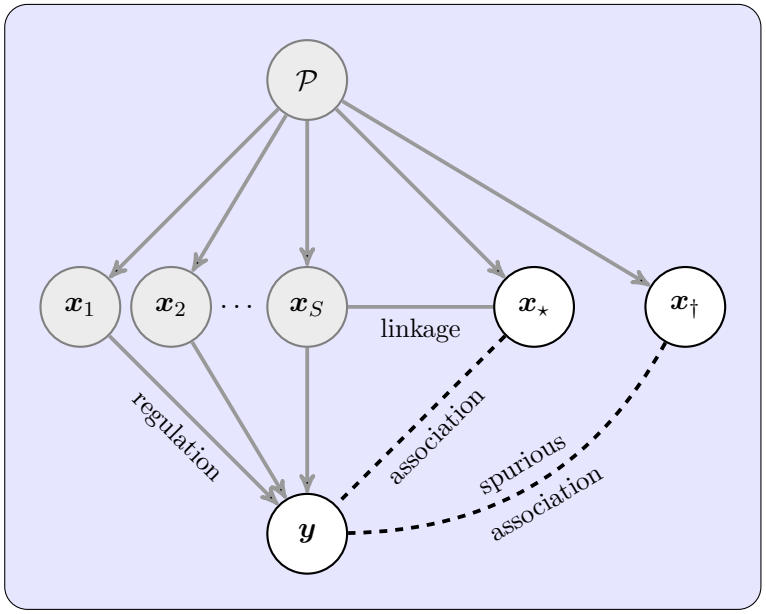
As a result of confounding by population structure some associations reported in the literature could be explained by differences in allele frequencies between populations and do not replicate [Freedman et al., 2004, McClellan and King, 2010, Wang et al., 2009]. For example in a study of native North Americans, an association has been found between immunoglobulin and type II diabetes, that disappeared when the authors stratified by admixed European ancestry [Knowler et al., 1988]. A popular example of stratification in European-derived populations is an association between the Lactase gene and height that has been found significant in European Americans. Though by ancestry-matching, the authors showed that the association was purely due to stratication [Campbell et al., 2005].

Improvements in study design and exclusion of individuals based on ethnicity help to somewhat alleviate the problem of population structure, but the problem of cryptic relationships remains. This approach also puts limitations on the data that can be collected. Close relatives could be detected based on genotype data and removed from the study, but such removal reduces statistical power.

A complementary way to account for confounding structure is by ways of statistical



(a) Multiple independent causes.



(b) External confounders cause spurious associations

Figure 1.4. Graphical model representation of a genome-wide association study involving multiple causes. In panel (a) the phenotype $\boldsymbol{y}$ is affected by multiple independent unknown causal variants $\boldsymbol{x}_1, \ldots \boldsymbol{x}_S$. A marker $\boldsymbol{x}_\star$ that lies within a region of linkage to a causal variant $\boldsymbol{x}_S$ becomes indirectly associated with the phenotype $\boldsymbol{y}$. In panel (b) an unknown confounding variable $\mathcal{P}$ like population structure causes dependence between causal variants and markers $\boldsymbol{x}_\dagger$ that are not in the proximity to any causal variant, leading to spurious associations.

modeling, an approach that is becoming more important as larger data sets are used to increase power. A range of methods have been proposed to correct for confounding in GWAS, including genomic control, principle components analysis and linear mixed models. Most of these account for confounding variation within the model.

Among these, only linear mixed models have been shown to be capable of correcting for population structure, family structure, and cryptic relatedness, while retaining sufficient power to detect true associations [Yu et al., 2005a, Zhao et al., 2007, Malosetti et al., 2007, Kang et al., 2008, 2010, Price et al., 2010b]. Despite their benefits, linear mixed models have so far seen relatively little use on large data sets due to their tremendous computational cost.

1.2.4. Aggregating weak signals

Another important strategy that has been proposed to detect associations to complex traits is to perform aggregate analyses of sets of rare or common variants within a gene or pathway [Price et al., 2010a, Bansal et al., 2010, Wu et al., 2011], or to use regularization-based approaches to enable joint analysis of a large number of markers. In particular, such methods allow for aggregation of weak signals within the group if markers analyzed, enable interplay among variants to be captured, and reduce the burden of multiple hypothesis testing. Unfortunately, until now, these approaches mostly did not address confounding by family relatedness and population structure.

1.3. Thesis structure

A main topic of this thesis are enhancements to linear mixed models, covering novel efficient algorithms that break the aforementioned computational barriers and to allow joint analysis of GWAS containing hundreds of thousands of samples (see Chapter 3), as well as improvements in modeling of confounding effects that yield larger power and better correction by selection of phenotype-specific sets of markers for confounder correction (see Chapter 4). In these two chapters we also point out and investigate the so far unappreciated problem that use of target markers for correction leads to a loss in power and propose two efficient ways to avoid the problem by excluding physically linked markers.

In order to overcome limitations in power to detect weak effects while accounting for confounding structure, we present two powerful and efficient methods based on linear mixed models that aggregate smaller effects in a joint analysis and correct for population structure (see Chapter 5).

Finally, in Chapter 6 we give an outlook in how recent multivariate machine-learning methods could be used to speed up joint analyses of multiple related traits, which gained attention as a way to increase increase power in GWAS, but are constrained by computational complexity.

1.3.1. Publications covered and individual contributions

- **Christoph Lippert**[⋆,†], Jennifer Listgarten[⋆,†], Ying Liu, Carl M Kadie, Robert I Davidson, David Heckerman[⋆,†]:

FaST linear mixed models for genome-wide association studies,
Nature Methods 8 (10), 833–835.

Christoph Lippert, Jennifer Listgarten and David Heckerman developed the method. Christoph Lippert derived the mathematical tricks for the algorithm. Christoph Lippert, Jennifer Listgarten, Carl M Kadie and David Heckerman designed and performed the experiments. Christoph Lippert, Robert I Davidson and Carl M Kadie wrote the source code. Christoph Lippert, Jennifer Listgarten, Ying Liu and David Heckerman wrote the manuscript.

- Jennifer Listgarten[⋆,†], **Christoph Lippert**[⋆,†], Carl M Kadie, Robert I Davidson, Eleazar Eskin, David Heckerman[⋆,†]:
Improved linear mixed models for genome-wide association studies,
Nature Methods 9 (6), 525–526.

  Jennifer Listgarten, Christoph Lippert and David Heckerman developed the method. Christoph Lippert derived the mathematical tricks for the algorithm. Jennifer Listgarten, Christoph Lippert and David Heckerman designed and performed the experiments. Christoph Lippert, Carl M Kadie and Robert I Davidson wrote the source code. Jennifer Listgarten, Christoph Lippert, Eleazar Eskin and David Heckerman wrote the manuscript.

- Jennifer Listgarten[⋆,†], **Christoph Lippert**[⋆,†], Eun Y Kang, Jing Xiang, Carl M, Kadie, David Heckerman[⋆,†]:
A powerful and efficient set test for genetic markers that handles confounding,
Bioinformatics 29 (12), 1526–1533.

  Jennifer Listgarten, Christoph Lippert and David Heckerman developed the method. Jennifer Listgarten, Christoph Lippert and David Heckerman designed the experiments. Jennifer Listgarten and David Heckerman performed the experiments. Christoph Lippert, Eun Y Kang, Carl M Kadie, Jing Xiang, and Jennifer Listgarten wrote the source code. Jennifer Listgarten, Christoph Lippert and David Heckerman wrote the manuscript.

- Barbara Rakitsch[†], **Christoph Lippert**[†], Oliver Stegle[†], Karsten Borgwardt:
A Lasso multi-marker mixed model for association mapping with population structure correction,
Bioinformatics 29 (2), 206–214.

  Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt developed the method. Barbara Rakitsch, Christoph Lippert and Oliver Stegle developed the mathematical tricks for the algorithm. Barbara Rakitsch, Christoph Lippert and Oliver Stegle designed the experiments. Barbara Rakitsch performed the experiments. Barbara Rakitsch wrote the source code. Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt wrote the manuscript.

- Oliver Stegle[⋆], **Christoph Lippert**[⋆], Joris Mooij, Neil Lawrence, Karsten Borgwardt:
Efficient inference in matrix-variate Gaussian models with iid observation noise,
NIPS 2011: 25$^{\text{th}}$ Annual Conference on Neural Information Processing Systems.

Oliver Stegle, Christoph Lippert, Joris Mooij, Neil Lawrence and Karsten Borgwardt developed the method. Christoph Lippert and Oliver Stegle derived the mathematical tricks for the algorithm. Oliver Stegle and Christoph Lippert designed and performed the experiments. Oliver Stegle, Christoph Lippert, Joris Mooij, Neil Lawrence and Karsten Borgwardt wrote the manuscript.

($\star$): equal contributions, (†): corresponding author

2. Statistical methods for genome-wide association studies

In association testing, the strength of a potential relationship between a SNP and a phenotype is quantified by a statistical model. The models used in this work are all *linear*, both for continuous, as well as for case-control phenotypes (See Section 2.4 for a discussion).

To introduce linear models, we start out in Section 2.1 by the most basic linear model, *linear regression*, which models a phenotype by linear-additive effects of a fixed set of regressors that can include causal or linked variants, confounders, and other covariates. In the context of this model we give an introduction to parameter inference by maximum likelihood and statistical testing for GWAS.

The problem of confounding by population structure, family structure, and cryptic relatedness in GWAS is now widely appreciated [Balding, 2006, Kang et al., 2008, Price et al., 2006, Yu et al., 2006, Kang et al., 2010, Price et al., 2010b]. Statistical methods for correcting these types of confounders have progressed through the years and include a variety of approaches, that will be discussed in the remainder of the chapter. Linear mixed models are considered to be the best method for confounder correction and are introduced in Section 2.2. In GWAS testing using linear mixed models the phenotype is typically modeled as the sum of a *fixed* linear regression, containing the effects of the marker to be tested, and a *random* linear-additive term that accounts for unwanted confounding structure. In other GWAS applications, as in the example of Section 5.1, the random effects may also be used to model joint genetic effects of a set of genetic markers.

Linear mixed models have been shown numerous times to account for all levels of genetic structure, including population structure, that is differences in allele frequencies between populations, family structure and cryptic relatedness, both of which introduce confounding due to close relatedness [Yu et al., 2005a, Zhao et al., 2007, Malosetti et al., 2007, Kang et al., 2008]. For completeness, in Section 2.3, other commonly used methods like genomic control and principle components analysis are reviewed.

2.1. Linear regression

Linear regression is introduced based on a model, where the phenotype is given as the sum of a single linear marker effect and random noise. In order to evaluate this model on data, an appropriate set of parameters has to be determined. The method of maximum likelihood estimation of linear regression parameters is introduced in Section 2.1.3. Given the maximum-likelihood estimate of the marker effect, the significance of the association between the marker and a phenotype can be evaluated by a P value. In Section 2.1.4 we give short introduction on how P values can be computed by comparing a test statistic

obtained for the SNP to a hypothetical distribution of test statistics that would be expected for SNPs that have no association to the phenotype.

Under the simple linear regression model all other effects present in a study including environment or other genetic effects are modeled as noise, which is assumed to be independent between samples. As argued in Section 1.1 this assumption is overly simplistic for most of the traits studies in GWAS and in most cases this model fails to account for the complex structure of other influences on the phenotype. In Section 2.1.2 the model is extended to include known genetic and non-genetic influences.

2.1.1. A linear model of simple traits

In the simple linear regression model, the phenotype y_n of an individual with index n is given by as the sum of a bias μ that is constant across all individuals, a linear effect of a marker of interest $x_{n,\star}$ and random environmental noise $\epsilon_n \in \mathbb{R}$.

$$\underbrace{y_n}_{\text{Phenotype}} = \underbrace{\beta_0}_{\text{bias}} + \underbrace{x_{n,\star} \cdot \beta_\star}_{\text{SNP effect}} + \underbrace{\epsilon_n}_{\text{noise}}.$$

Assuming that the noise are independent samples of a Gaussian distribution with variance $\sigma^2 \in \mathbb{R}^+$, then for a dataset of N samples, the probability of the data given a set of parameters defines the *likelihood* $\mathcal{L}$ of the model.

$$\mathcal{L}\left(\{y_1, \ldots, y_N\} \,|\, \beta_0, \beta_\star, \sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \beta_0 + x_{n,\star} \cdot \beta_\star \,;\, \sigma^2\right). \tag{2.1}$$

As the data is observed and fixed, the likelihood $\mathcal{L}(\boldsymbol{\theta})$, and also the *log likelihood* denoted as $\log \mathcal{L}(\boldsymbol{\theta})$ are a good way to score a set of parameters $\boldsymbol{\theta}$ and are usually written as a function of these parameters alone.

2.1.2. A linear model of complex traits

Most quantitative traits take on a continuous range of values and can be considered complex, in the sense that the variation can not merely be traced back to a single polymorphic locus, but rather is the result of a variety of influences, including other SNP-effects and influences like race, gender, or environment.

In the single SNP linear regression model all such additional influences are accumulated in the noise variable ϵ_n. If these influences are not independent between individuals, then the assumption of independent noise is violated.

An easy way to account for such influences is to include variables with known effects into the model as regression covariates. The likelihood of a linear model containing C covariates $x_{n,c}$ with effects β_c, is defined by

$$\mathcal{L}\left(\beta, \beta_\star, \{w_1, \ldots, w_D\}, \sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \mu + x_\star \cdot \beta_\star + \sum_{c=1}^{C} x_{n,c} \cdot \beta_c \,;\, \sigma^2\right). \tag{2.2}$$

To unclutter notation, the N-dimensional column vector $\boldsymbol{y}$, holding the phenotype values of all samples, the $N \times D$-dimensional design matrix $\boldsymbol{X}$ holding all regressors plus

a constant vector of ones as columns, and the D-dimensional column vector $\boldsymbol{\beta}$ holding the corresponding regression coefficients are introduced.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,C} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,C} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,C} \end{bmatrix}; \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_C \end{bmatrix};$$

Without loss of generality, the SNP vector $\boldsymbol{x}_\star$ and the weight $\beta_\star$ are also assumed to be contained in $\boldsymbol{X}$ and $\boldsymbol{\beta}$ whenever appropriate, unless we want to highlight these explicitly. Using this notation, linear regression defines a multivariate normal distribution in the phenotype vector $\boldsymbol{y}$.

Given the N-by-1 vector of target values $\boldsymbol{y}$ and the N-by-D design matrix $\boldsymbol{X}$, the log-likelihood of linear regression is

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2\right) = -\frac{N}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{2.3}$$

2.1.3. Estimation in the linear regression model

The likelihood of the data under a set of parameters is a way to evaluate the quality of such parameters. A good estimator of the parameters σ^2 and $\boldsymbol{\beta}$ can be obtained by maximizing the likelihood. For the linear regression model in Equation (2.3), such an estimate can be obtained, by jointly equating the derivatives of the log-likelihood with respect to all parameters to zero.

Score function

The *score* is defined as the derivative of the log-likelihood with respect to a parameter and is important for parameter estimation as well as statistical testing. For the linear model, the score function for the effects $\boldsymbol{\beta}$ is given by the gradient

$$\frac{\nabla \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2\right)}{\nabla \boldsymbol{\beta}} = \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}. \tag{2.4}$$

And the score of the residual variance parameter σ^2 is given by

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma_\mathrm{e}^4} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{2.5}$$

Maximum likelihood estimation

The maximum-likelihood parameters of linear regression can easily be found by jointly equating the score with respect to all parameters to zero.

$$\frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{0}. \tag{2.6}$$

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma_\mathrm{e}^4} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) = 0. \tag{2.7}$$

Equation (2.6) can readily be solved for $\boldsymbol{\beta}_{\mathrm{M}}$ that does not depend on σ^2.

$$\boldsymbol{\beta}_{\mathrm{M}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.8}$$

Plugging the expression for maximum likelihood weights $\boldsymbol{\beta}_{\mathrm{M}}$ back into the equation for derivative with respect to σ^2, we obtain

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma_{\mathrm{e}}^4} \left(\boldsymbol{y} - \boldsymbol{X} \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}}_{\boldsymbol{\beta}_{\mathrm{M}}}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X} \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}}_{\boldsymbol{\beta}_{\mathrm{M}}}\right) = 0. \tag{2.9}$$

Solving for ${\sigma^2}_{\mathrm{M}}$ gives the maximum likelihood estimate for the variance.

$${\sigma^2}_{\mathrm{M}} = \frac{1}{N} \left(\boldsymbol{y} - \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right). \tag{2.10}$$

2.1.4. Association testing using linear regression

When testing for association between a SNP $\boldsymbol{x}_\star$ and a phenotype using the linear regression model in Equation (2.3), the null hypothesis $\mathcal{H}_0$ that the SNP has no association is compared to the alternative hypothesis $\mathcal{H}_1$ that the SNP is associated to the phenotype.

From the linear regression likelihood it is easy to see that the amount of phenotypic variation explained by the SNP $\boldsymbol{x}_\star$ is a quadratic function of the regression effect $\beta_\star$. So the magnitude of the regression effects is a measure of association between SNP and phenotype. If $\mathcal{H}_0$ would be true, the regression effect $\beta_\star$ would be zero, whereas if $\mathcal{H}_1$ would be true, the regression effect would deviate from zero. In a statistical test for association, these two hypotheses are compared to each other, where a SNP would be called associated to the phenotype, if the test rejects the null hypothesis $\mathcal{H}_0$ in favor of the alternative-hypothesis $\mathcal{H}_1$.

Likelihood-ratio test

Using the likelihood as a measure to evaluate the quality of the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$, a test statistic follows as the ratio of the maximum of the likelihood under $\mathcal{H}_1$ and the maximum of the likelihood under $\mathcal{H}_0$.

We define the likelihood-ratio statistic LRT as twice the logarithm of the ratio of the maxima of the respective likelihood functions, or equivalently the difference in their logarithms.

$$\mathrm{LRT} = 2 \left(\sup_{\Theta_1 \,|\, \mathcal{H}_1} \log \mathcal{L}\left(\theta_1\right) - \sup_{\Theta_0 \,|\, \mathcal{H}_0} \log \mathcal{L}\left(\Theta_0\right) \right). \tag{2.11}$$

If $\mathcal{H}_0$ and $\mathcal{H}_1$ are nested in the sense that $\mathcal{H}_0$ is fully contained in $\mathcal{H}_1$, as in the case of the tests considered here, this statistic is always larger than or equal to zero.

For two-sided tests in nested hypotheses, where the parameters are not bounded, the asymptotic distribution of LRT under $\mathcal{H}_0$ is approximately distributed as a Chi-square

random variable. The degrees of freedom of the distribution equal the difference in the number of free parameters between the null and the alternative hypothesis [Wilks, 1938].

So when testing for association of the $N \times 1$ SNP-vector $\boldsymbol{x}_\star$ while conditioning on the effect of the covariates contained in $\boldsymbol{X}$, the null distribution of the likelihood-ratio statistic for linear regression $\mathrm{LRT}_{\mathrm{LR}}$ is approximated by a chi-square distribution with one degree of freedom, as the alternative model contains a single extra parameter:

$$2 \log \frac{\max_{\boldsymbol{\beta},\beta_\star,\sigma^2} \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{x}_\star \cdot \beta_\star \,;\, \sigma^2 \boldsymbol{I}\right)}{\max_{\boldsymbol{\beta},\sigma^2} \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{x}_\star \cdot 0 \,;\, \sigma^2 \boldsymbol{I}\right)} \sim \chi^2_1. \tag{2.12}$$

P values are computed from the survival function of the distribution.

2.2. Linear mixed models

This section gives an introduction to linear mixed models and their use in genome-wide association studies. When introducing linear regression, effects of known variables are included in the model as covariates whose effects are estimated using maximum likelihood. The central idea behind confounder correction in GWAS using linear mixed models is that while it is hard to get reliably give point estimates for the effects of confounding genetic structure, it is often possible to describe these in terms of random effects, for which covariation can be quantified in terms of the degree of genetic relatedness between the samples.

The linear mixed model is introduced for the effects of a number of causal variants in Section 2.2. The most commonly used measures of relatedness are introduced in Section 2.2.1. Methods for parameter estimation in in linear mixed models are introduced in Section 2.2.3 and statistical testing is introduced in Section 2.2.4.

A linear mixed model of complex traits

In linear mixed models the phenotype $\boldsymbol{y}$ is written as the *mixed* sum of a *linear* term in the *fixed* effects $\boldsymbol{\beta}$, that as in the linear regression model include a bias term as well as the effects of known covariates and the marker of interest, and linear *random* effects $\boldsymbol{u}$.

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\bar{\boldsymbol{G}}\boldsymbol{u}}_{\text{random}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}, \tag{2.13}$$

where the N-by-S matrix $\bar{\boldsymbol{G}}$ is the design matrix holding S causal loci.

When testing a marker for association with the phenotype, the standard application of linear mixed models for genome-wide association studies, the variables of interest are modeled as fixed, whereas the random effects account for nuisance variation and are integrated out. If the causal loci are confounded by population structure, then including these in a test for association corrects for confounding variation in the phenotype, similar to covariates in a standard linear regression model (see Figure 2.1(a)).

For many complex traits it has been observed that the contribution of each of the S causal loci to the total level of *genetic variance* σ^2_{g} is approximately equal, with an effect size distribution that is inversely proportional to the corresponding minor allele frequencies f_s [Park et al., 2010, 2011]. Under this model the random effects are treated

as independent Gaussian variables, each contributing an equal fraction of $\frac{1}{S}\sigma_{\mathrm{g}}^2$ to the total variance σ_{g}^2. The S loci contained in the design matrix $\bar{\boldsymbol{G}}$ are assumed to have a mean of zero and unit variance.

If we define the total random genetic effect as $\boldsymbol{v} = \bar{\boldsymbol{G}}\boldsymbol{u}$, then $\boldsymbol{v}$ follows a multivariate normal distribution:

$$\boldsymbol{v} \sim \mathcal{N}\left(\mathbf{0}\,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K}\right), \tag{2.14}$$

where the covariance is proportional to $\boldsymbol{K} = \frac{1}{S}\bar{\boldsymbol{G}}\bar{\boldsymbol{G}}^\top$, a matrix that quantifies the genetic relationship between individuals based on the causal loci.

Under this commonly used model the marginal likelihood of $\boldsymbol{y}$ follows from marginalization of $\boldsymbol{v}$:

$$\int \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}\,;\, \sigma^2 \boldsymbol{I}\right) \cdot \mathcal{N}\left(\boldsymbol{v} \,\middle|\, \mathbf{0}\,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K}\right) \mathrm{d}\boldsymbol{v} = \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right). \tag{2.15}$$

Equivalently, the log (marginal) likelihood is a function of fixed effects $\boldsymbol{\beta}$, and the variance parameters $\boldsymbol{\theta} = \left[\sigma^2, \sigma_{\mathrm{g}}^2\right]$, namely the level of environmental noise σ^2 and the genetic variance σ_{g}^2.

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) = -\frac{N}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|\boldsymbol{V}_{\boldsymbol{\theta}}\right| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right), \tag{2.16}$$

where we defined the complete covariance term of the distribution as $\boldsymbol{V}_{\boldsymbol{\theta}} = \sigma_{\mathrm{g}}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}$.

In this form, the causal variants enter the model only in the genetic relatedness matrix $\boldsymbol{K}$, which directly represents the confounding variation in the phenotype (See Figure 2.1(b)).

2.2.1. Measures of relatedness

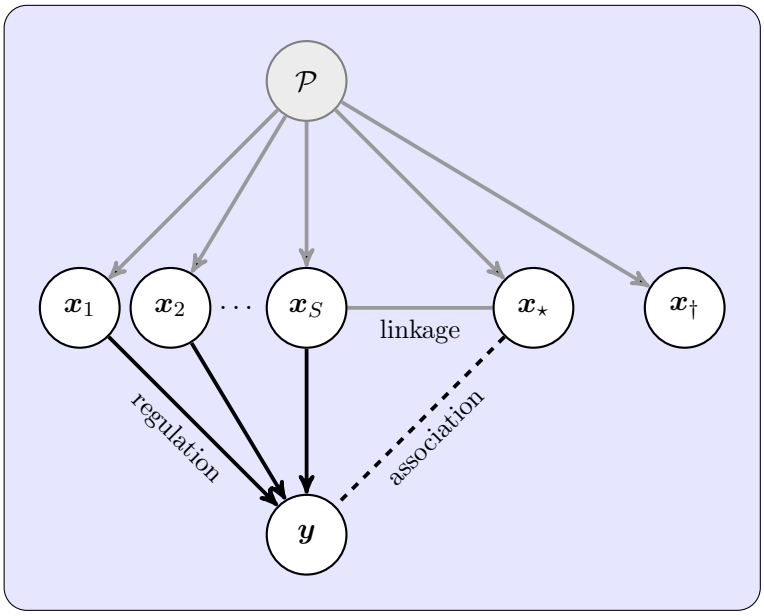
So far we have assumed that the matrix of genetic relatedness $\boldsymbol{K}$ is computed from the causal loci. In practice these are unknown and other measures of genetic relatedness have to be used instead. Here we review the most commonly used measures of genetic similarity, including identity by descent computed from pedigrees and the realized relationship matrix.

We also provide a brief introduction to kernel methods, that can be used to model other types of covariance structures.

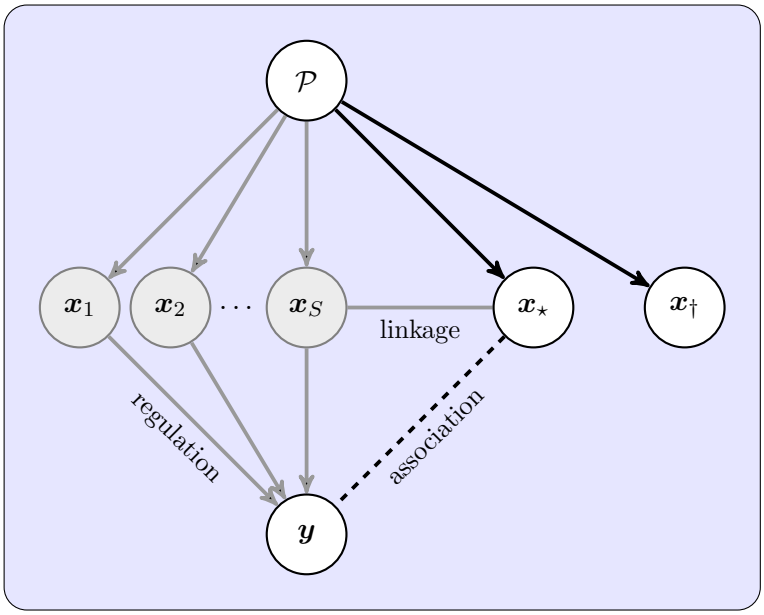
Kinship Matrix

In Fisher's infinitesimal model the distribution of a phenotype is derived for the case of an infinite number of causal variants in Equation (2.13). In his model a quantitative phenotype $\boldsymbol{y}$ with total genetic variance σ_{g}^2 be given by the sum of a large number of genetic effects of individual variances $\frac{1}{S}\sigma_{\mathrm{g}}^2$.

$$\boldsymbol{y} = \sum_{j=1}^{S} \bar{\boldsymbol{G}}_j u_j + \boldsymbol{e}, \tag{2.17}$$



(a) Conditioning on background SNPs



(b) Conditioning on population structure

Figure 2.1. Graphical model representations of two related concepts to account for confounders in GWAS. In both panels the phenotype $\boldsymbol{y}$ is affected by multiple unknown genetic causes $\boldsymbol{x}_1, \ldots \boldsymbol{x}_S$. There is a close relationship between (a) a model that conditions on possibly confounded background variants and (b) a model that corrects for unknown confounding influences $\mathcal{P}$ by either an estimate of the confounding variable or the variance it induces.

then in the limit of an infinite number of causal loci ($S \to \infty$), that independently follow Mendelian inheritance, the phenotypic covariation between two individuals is proportional to the amount of genetic material at the causal loci that is identical by descent (IBD).

Introducing additional fixed effects $\boldsymbol{\beta}$, then the distribution of the phenotype is given by a linear mixed model.

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta};\, \sigma_{\mathrm{g}}^2 \boldsymbol{K}_{\mathrm{IBD}} + \sigma^2 \boldsymbol{I}\right), \tag{2.18}$$

where $\boldsymbol{K}_{\mathrm{IBD}}$ is the matrix of IBD coefficients between pairs of individuals.

Kinship coefficients can be computed from known pedigrees [Fisher, 1918] and should be corrected for an increase in relatedness due to inbreeding [Wright, 1922, Malécot, 1948]. For the case, where the pedigree is not known, the kinship matrix $\boldsymbol{K}_{\mathrm{IBD}}$ from genetic markers [Abecasis et al., 2001, Hardy and Vekemans, 2002]. Such marker-based estimates of IBD have been used in mixed model applications to GWAS in maize and *A. thaliana* [Yu et al., 2005a, Zhao et al., 2007].

Realized Relationship Matrix

Estimates of *realized relationships* between individuals are obtained by counting average number of shared marker alleles between two individuals and have been shown to improve prediction of the genetic component of a trait over predictions using pedigree-based kinship estimates [Nejati-Javaremi et al., 1997]. These predictions have been further improved by the use of dense genome-wide markers that tag causal loci due to linkage [Meuwissen et al., 2001, Villanueva et al., 2005]. Also for GWAS the use of relationships estimated from genome-wide markers have been shown to improve correction for confounders over relationship based on kinship [Kang et al., 2008].

Let the N-by-S matrix $\boldsymbol{G}$ be a matrix holding S genotyped markers for N individuals. We assume that each marker in $\boldsymbol{G}$ is mean centered and is normalized to have unit variance. We define the *realized relationship matrix* (RRM) as the empirical covariance matrix

$$\boldsymbol{K}_{\mathrm{RRM}} = \frac{1}{S}\boldsymbol{G}\boldsymbol{G}^{\top}. \tag{2.19}$$

Similar to the linear mixed model that uses the causal variants, the linear mixed model using the RRM can be written as a linear regression model where some regressors are fixed and some regressors are random.

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\boldsymbol{G}\boldsymbol{u}}_{\text{random}} + \boldsymbol{\epsilon}. \tag{2.20}$$

In this view the variants contained in the RRM are used as random regressors or covariates that capture the genetic variation in the phenotype by being linked to the unknown causal variants or by ways of confounding. Overfitting due to the large number of covariate effects is avoided by integrating the regressors over independent normal distributions with variance σ_{g}^2/S (See Figure 2.1(a)).

Kernel methods

Random effects can be interpreted as a Gaussian random process, whose covariance is given by the genetic relatedness [Rasmussen and Williams, 2005]. After integration of the genetic effects in the linear mixed model likelihood (Equation (2.15)), the random effects only appear implicitly as a function of their covariance matrix $\boldsymbol{K}$. Any features contained in the original design matrix are used only implicitly in the form of dot-products.

For these kinds of models, it has been shown that in principle any symmetric semi-positive-definite *kernel* matrix could be used for $\boldsymbol{K}$. While in the standard linear mixed model these dot products are computed directly on the features, resulting in a model that is linear in the features, kernel functions may represent non-linear dot-products and thus can yield models that are non-linear in the original features [Kimeldorf and Wahba, 1970, Schölkopf et al., 2001, Schölkopf and Smola, 2001].

Kernel methods have been used to come up with covariance structures that do not only cover genetic effects, but also effects of hidden environment.

For example in the context of expression quantitative trait locus (eQTL) studies, covariance structures based on latent variable models [Lawrence, 2004, 2005] representing shared hidden influences can be estimated jointly from all expression phenotypes, and has been shown to yield improved correction and a gain in power to detect novel associations [Stegle et al., 2010, Listgarten et al., 2010, Fusi et al., 2012, Stegle et al., 2012, Fusi et al., 2013].

2.2.2. Best linear unbiased prediction

The best linear unbiased predictor (BLUP) is a minimum variance predicted value of the random effects $\boldsymbol{v}$ in a linear mixed model. Predictions of random effects are a means to predict the phenotype of an individual from genotyped SNP-data [Lee et al., 2008]. For example in animal breeding these predictions are utilized as a genomic *selection index* or *breeding value* to increase gain in breeding experiments [Henderson, 1950, 1984, Meuwissen et al., 2001, Villanueva et al., 2005],.

The BLUP $\hat{v}_\star$ of an individual of interest indexed by $\star$ is obtained by maximizing the joint distribution of the vector of all observed phenotypes $\boldsymbol{y}$ and the random genetic effect $v_\star$ of that individual of interest. Let $\boldsymbol{V_\theta}$ be the total covariance term of $\boldsymbol{y}$, the 1-by-N dimensional vector of genetic relatedness between the individual of interest and all observed individuals be $\boldsymbol{k}_{\star,:}$ and the genetic relatedness of the individual of interest with itself be $k_{\star,\star}$, then the joint distribution of $\boldsymbol{y}$ and $v_\star$ is given as

$$\begin{bmatrix} \boldsymbol{y} \\ v_\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X\beta} \\ 0 \end{bmatrix}; \begin{bmatrix} \boldsymbol{V_\theta} & \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:}^\top \\ \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} & \sigma_\mathrm{g}^2 k_{\star,\star} \end{bmatrix} \right)$$

The BLUP is now equal to the mean of the the conditional distribution of $v_\star$ given $\boldsymbol{y}$.

$$v_\star \,|\, \boldsymbol{y} \sim \mathcal{N}\left(\underbrace{\sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} \boldsymbol{V_\theta}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})}_{\text{BLUP}}; \sigma_\mathrm{g}^2 k_{\star,\star} - \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} \boldsymbol{V_\theta}^{-1} \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:}^\top \right). \tag{2.21}$$

Given the vector of covariates for the individual of interest $\boldsymbol{x}_\star$, then the conditional distribution of the phenotype of the individual $y_\star$ follows by adding the covariates effects

and accounting for the environmental variance.

$$y_\star \,|\, \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{x}_\star\boldsymbol{\beta} + \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) ;\, \sigma_\mathrm{g}^2 k_{\star,\star} + \sigma^2 - \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} \left(\sigma_\mathrm{g}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:}^\top \right). \tag{2.22}$$

Equivalent predictors of random or latent quantities in Gaussian models have been developed multiple times in other fields, these are for example known as Wiener-Kolmogorow filters, smoothing-spline models, Kriging and Gaussian-process regression [Robinson, 1991, Matheron, 1963, Wahba, 1990, Rasmussen and Williams, 2005]. Many tricks developed for these methods are directly applicable to linear mixed models with applications in genetics.

2.2.3. Parameter estimation in linear mixed models

Starting with the linear mixed model with random effects integrated out with log likelihood equal to (2.16), the goal is to infer the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = \left[\sigma^2, \sigma_\mathrm{g}^2\right]$ and any additional covariance parameters if these are present.

Score

The gradient of the log-likelihood as given in Equation (2.16) with respect to fixed effects $\boldsymbol{w}$ defines the score of $\boldsymbol{w}$.

$$\frac{\nabla \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \sigma_\mathrm{g}^2\right)}{\nabla \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \boldsymbol{X}\boldsymbol{\beta}. \tag{2.23}$$

The score of a variance parameter is the partial derivative of the log-likelihood with respect to a variance parameter θ_i:

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right)}{\partial \theta_i} = -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{V}_{\boldsymbol{\theta}}}{\partial \theta_i} \right) + \frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{V}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{2.24}$$

The matrix derivative of the covariance $\boldsymbol{V}_{\boldsymbol{\theta}} = \sigma_\mathrm{g}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}$ with respect to the environmental variance $\theta_1 = \sigma^2$ equals

$$\frac{\partial \boldsymbol{V}_{\boldsymbol{\theta}}}{\partial \sigma^2} = \boldsymbol{I},$$

and the matrix derivative with respect to the genetic variance $\theta_2 = \sigma_\mathrm{g}^2$ equals

$$\frac{\partial \boldsymbol{V}_{\boldsymbol{\theta}}}{\partial \sigma_\mathrm{g}^2} = \boldsymbol{K}.$$

Maximum likelihood estimation

As in the case of the linear regression, the likelihood is maximized by equating the gradient with respect to all parameters to zero and jointly solving the resulting equations.

Though, while for linear regression the maximum likelihood parameters can be found in closed form from the gradient equations, this is not the case for linear mixed models. Moreover, the log marginal likelihood function is not jointly convex in the variance parameters, rendering it hard to ensure global maximization of the likelihood.

A straightforward way to obtain a local optimum of the parameter values is to use gradient descent methods. For most GWAS applications, though, naive use of gradient descent techniques is not well suited, as these involve repeated computation of the log-likelihood function as well as of the gradients and for second-order methods like Fisher scoring or Newton-Raphson also of the Fisher or observed information matrix [Demidenko, 2004].

Maximum likelihood estimation can be simplified by writing the log likelihood in Equation (2.16) as a function of the ratio $\gamma = \frac{\sigma_g^2}{\sigma^2}$ of the genetic variance σ_g^2 over the environmental variance σ^2 [Hartley and Rao, 1967].

$$\log \mathcal{L}\left(\gamma, \sigma^2, \boldsymbol{\beta}\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\log|\boldsymbol{H}_\gamma| - \frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_\gamma^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right), \quad (2.25)$$

where we defined the matrix $\boldsymbol{H}_\gamma = \boldsymbol{I} + \gamma\boldsymbol{K}$. In this formulation the maximum likelihood solutions for all parameters other than γ ($\boldsymbol{\beta}$, and σ^2) follow in closed form for any positive value of γ.

The maximum likelihood value $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ of the fixed effects as a function of γ is found by taking the gradient of the log-likelihood in Equation (2.25) with respect to $\boldsymbol{\beta}$ and jointly setting all entries of the gradient to zero.

$$\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{y} - \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \boldsymbol{0}.$$

By bringing the part involving $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ to one side and after cancelling σ^2 from the equation, this becomes

$$\boldsymbol{X}^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \boldsymbol{X}^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{y}.$$

Multiplication of both sides by the inverse of the factor on the left side yields the maximum likelihood solution of the fixed effects given a value of γ as

$$\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \left(\boldsymbol{X}^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{y}. \quad (2.26)$$

To find the maximum likelihood value of the genetic variance σ^2 as a function of γ, the maximum likelihood values of the fixed effects $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ from Equation (2.26), which do not depend on σ^2, are substituted into the log likelihood, Equation (2.25). The derivative with respect to σ^2 is set to zero, giving

$$-\frac{N}{2{\sigma^2}_{\mathrm{M}_\gamma}} + \frac{1}{2{\sigma^4}_{\mathrm{M}_\gamma}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^\top\boldsymbol{H}_\gamma^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right) = 0.$$

Both sides are multiplied by $2\hat{\sigma}^4$, and the result is solved for ${\sigma^2}_{\mathrm{M}_\gamma}$, such that the maximum likelihood solution of the residual variance given γ is

$${\sigma^2}_{\mathrm{M}_\gamma} = \frac{1}{N}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^\top\boldsymbol{H}_\gamma^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right).$$

After further simplification, this becomes

$${\sigma^2}_{\mathrm{M}_\gamma} = \frac{1}{N}\boldsymbol{y}^\top\boldsymbol{P}_\gamma^\top\boldsymbol{H}_\gamma^{-1}\boldsymbol{P}_\gamma\boldsymbol{y}, \quad (2.27)$$

where we defined $\boldsymbol{P}_\gamma$ as the matrix

$$\boldsymbol{P}_\gamma = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1}. \tag{2.28}$$

Plugging the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 back into the likelihood, a profile log likelihood is obtained as

$$\log \mathcal{L}(\boldsymbol{\gamma}) = -\frac{N}{2}\log\left(2\pi {\sigma^2}_{\mathrm{M}_\gamma}\right) - \frac{1}{2}\log|\boldsymbol{H}_\gamma| - \frac{1}{2{\sigma^2}_{\mathrm{M}_\gamma}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^\top \boldsymbol{H}_\gamma^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right).$$

Using the maximum likelihood expressions from Equations (2.26) and (2.27) and simplifying, this profile log likelihood becomes a function of γ alone

$$\log \mathcal{L}(\gamma) = -\frac{N}{2}\left(1 + \log\frac{2\pi}{N}\right) - \frac{1}{2}\log|\boldsymbol{H}_\gamma| - \frac{N}{2}\log \boldsymbol{y}^\top \boldsymbol{P}_\gamma^\top \boldsymbol{H}_\gamma^{-1}\boldsymbol{P}_\gamma \boldsymbol{y}. \tag{2.29}$$

In principle, a local optimum with respect to γ of this profile log-likelihood could be obtained by the use of gradient descent methods. Alternatively, derivative-free methods like a grid search can be used to find an optimum for γ.

Restricted maximum likelihood estimation

On finite data the maximum likelihood estimate has been found to underestimate the variances in the Gaussian model. This can be attributed to the fact that under maximum likelihood estimation, the estimate of variances depends on a distribution that has been profiled for the fixed effects and exerts a loss in degrees of freedom.

Restricted maximum likelihood[1] estimation has been proposed to overcome this loss on degrees of freedom by estimating variance components of the model only on a projection of the target variable (i.e., the phenotype) into an $N - D$-dimensional subspace, that is orthogonal to the fixed effects. Intuitively, the variance components are estimated from residuals of the target variable, after the fixed effects have been regressed out. The fixed effects on the other hand are estimated from another projection, which under the model is statistically independent to the former projection. More formally, for $N > D$, two suitable projection matrices $\boldsymbol{S}$ and $\boldsymbol{Q}_\gamma$ are chosen such that they fulfill the four following criteria [Patterson and Thompson, 1971]:

1. $\operatorname{rank}(\boldsymbol{S}) = N - D$.
   $\operatorname{rank}(\boldsymbol{Q}_\gamma) = D$.

2. The two projections are statistically independent under the model.
   $\Leftrightarrow \mathbb{C}\mathrm{ov}(\boldsymbol{S}\boldsymbol{y}, \boldsymbol{Q}_\gamma \boldsymbol{y}) = \boldsymbol{0}$.
   $\Leftrightarrow \boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{Q}_\gamma^\top = \boldsymbol{0}$.

3. The expected value of $\boldsymbol{S}\boldsymbol{y}$ under the model is zero.
   $\Leftrightarrow \mathbb{E}(\boldsymbol{S}\boldsymbol{y}) = \boldsymbol{0}$.
   $\Leftrightarrow \boldsymbol{S}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$.
   $\Leftarrow \boldsymbol{S}\boldsymbol{X} = \boldsymbol{0}$.

[1] also: residual maximum likelihood

4. $\operatorname{rank}(\boldsymbol{Q}_\gamma \boldsymbol{X}) = D$.

From these conditions it follows that the likelihood can be written the product of likelihood functions of two independent projection of the data, one on $\boldsymbol{Sy}$ and one on $\boldsymbol{Q}_\gamma \boldsymbol{y}$ [Patterson and Thompson, 1971].

$$\mathcal{L}\left(\boldsymbol{\beta}, \gamma, \sigma^2\right) \propto \mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right) \cdot \mathcal{L}\left(\boldsymbol{\beta} | \boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right), \tag{2.30}$$

where $\mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$ is also called the *restricted likelihood*, for which Harville [1974] proposed suitable matrices for $\boldsymbol{S}$ and ${\boldsymbol{Q}_\gamma}^2$, namely the N-by-N orthogonal projector for the fixed effects $\boldsymbol{X}$

$$\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \tag{2.31}$$

and the D-by-N matrix

$$\boldsymbol{Q}_\gamma = \left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1}. \tag{2.32}$$

Parameter estimation is then performed in a two-step procedure. First $\log \mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$ is maximized with respect to the variance parameters γ and σ^2. Then, the solutions obtained are plugged into $\log \mathcal{L}\left(\boldsymbol{\beta} | \boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right)$ which subsequently is maximized with respect to $\boldsymbol{\beta}$.

Estimation of variance parameters by restricted maximum likelihood In order to find a suitable expression for $\log \mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$ one has to account for the fact that the covariance of $\boldsymbol{Sy}$ is $\sigma^2 \boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{S}$, a matrix that is rank deficient due to a projection to the space orthogonal to $\boldsymbol{X}$. A way to do so is the use of the pseudo-determinant and the Moore-Penrose pseudo-inverse of $\boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{S}$.

$$\log \mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right) = -\frac{N-D}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} \log |\boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{S}|_\dagger - \frac{1}{2\sigma^2} \boldsymbol{y}^\top \boldsymbol{S} \left(\boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{S}\right)^\dagger \boldsymbol{Sy}.$$

Both, the pseudo-determinant as well as the pseudo-inverse can be computed from the economy spectral decomposition $\boldsymbol{V}_{\boldsymbol{S}} \boldsymbol{\Sigma}_\gamma \boldsymbol{V}_{\boldsymbol{S}}^\top$ of $\boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{S}$, where $\boldsymbol{\Sigma}_\gamma$ is an $(N-D)$-by-$(N-D)$ diagonal matrix, holding the non-zero eigenvalues of $\boldsymbol{S}\boldsymbol{H}_\gamma \boldsymbol{S}$ and $\boldsymbol{V}_{\boldsymbol{S}}$ is an N-by-$(N-D)$ matrix, holding the corresponding eigenvectors as columns. As shown in Lemma C.32 $\boldsymbol{S}$ can be written as $\boldsymbol{V}_{\boldsymbol{S}} \boldsymbol{V}_{\boldsymbol{S}}^\top$. Also using $\boldsymbol{V}_{\boldsymbol{S}}^\top \boldsymbol{V}_{\boldsymbol{S}} = \boldsymbol{I}$, we get

$$-\frac{N-D}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} \log |\boldsymbol{\Sigma}_\gamma| - \frac{1}{2\sigma^2} \boldsymbol{y}^\top \boldsymbol{V}_{\boldsymbol{S}} \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{V}_{\boldsymbol{S}}^\top \boldsymbol{y}.$$

It follows, that $\mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$ equals to the regular multivariate normal distribution on $\boldsymbol{V}_{\boldsymbol{S}}^\top \boldsymbol{y}$ with covariance matrix $\boldsymbol{\Sigma}_{gamma}$.

$$\mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right) = \mathcal{N}\left(\boldsymbol{V}_{\boldsymbol{S}}^\top \boldsymbol{y} \,\middle|\, \boldsymbol{0}\,;\, \sigma^2 \boldsymbol{\Sigma}_\gamma\right). \tag{2.33}$$

The restricted maximum likelihood estimators of the variance parameters ${\sigma^2}_{\mathrm{R}_\gamma}$ and $\gamma_{\mathrm{R}_\gamma}$ are found by applying maximum likelihood estimation to $\mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$, given by

[2]Note, that Patterson and Thompson [1971] originally used $\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$ and $\boldsymbol{Q}_\gamma = \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1}$.

Equation (2.33). Taking the derivative of the logarithm of $\mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$ with respect to σ^2 and setting this to zero, we get

$$0 = -\frac{N-D}{2\sigma^2{}_{\mathrm{R}_\gamma}} + \frac{1}{2\sigma^2{}_{\mathrm{R}_\gamma}} \boldsymbol{y}^\top \boldsymbol{V_S} \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{V_S}^\top \boldsymbol{y}.$$

The solution to this Equation is

$$\sigma^2{}_{\mathrm{R}_\gamma} = \frac{1}{N-D} \boldsymbol{y}^\top \boldsymbol{V_S} \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{V_S}^\top \boldsymbol{y}. \tag{2.34}$$

Profile restricted likelihood When plugging the restricted maximum likelihood estimator for the environmental noise $\sigma^2{}_{\mathrm{R}_\gamma}$ back into the log restricted likelihood, a log restricted likelihood, which is profiled over σ^2, is derived as

$$\log \mathcal{L}\left(\gamma, \sigma^2{}_{\mathrm{R}_\gamma} | \boldsymbol{Sy}\right) = -\frac{N-D}{2}\left(1 - \log\frac{2\pi}{N-D}\right) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\gamma| - \frac{N-D}{2}\log R_{\boldsymbol{S}}, \tag{2.35}$$

where the residual term is

$$R_{\boldsymbol{S}} = \boldsymbol{y}^\top \boldsymbol{V_S} \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{V_S}^\top \boldsymbol{y}. \tag{2.36}$$

The derivative of this log likelihood with respect to the remaining free parameter γ is

$$\frac{\partial \log \mathcal{L}\left(\gamma, \sigma^2{}_{\mathrm{R}_\gamma} | \boldsymbol{Sy}\right)}{\partial \gamma} = -\frac{1}{2}\mathrm{tr}\,\boldsymbol{\Sigma}_\gamma^{-1}\frac{\partial \boldsymbol{\Sigma}_\gamma}{\partial \gamma} - \frac{N-D}{2}\frac{\frac{\partial R_{\boldsymbol{S}}}{\partial \gamma}}{R_{\boldsymbol{S}}}. \tag{2.37}$$

Herein, the derivative of the matrix $\boldsymbol{\Sigma}_\gamma$ of the $N-D$ non-zero eigenvalues of $\boldsymbol{SH}_\gamma\boldsymbol{S}$ is given by

$$\frac{\partial \boldsymbol{\Sigma}_\gamma}{\partial \gamma} = \frac{\boldsymbol{\Sigma}_\gamma - \boldsymbol{I}_{N-D}}{\gamma}. \tag{2.38}$$

As can easily be verified using Lemma C.15, the derivative of the residual term is given by

$$\frac{\partial R_{\boldsymbol{S}}}{\partial \gamma} = \boldsymbol{y}^\top \boldsymbol{V_S} \boldsymbol{\Sigma}_\gamma^{-1} \frac{\partial \boldsymbol{\Sigma}_\gamma}{\partial \gamma} \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{V_S}^\top \boldsymbol{y}. \tag{2.39}$$

Estimation of fixed effects by restricted maximum likelihood An expression for the logarithm of $\mathcal{L}\left(\boldsymbol{\beta} | \boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right)$ can be found as

$$\log \mathcal{L}\left(\boldsymbol{\beta} | \boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right) = -\frac{D}{2}\log\sigma^2 - \frac{1}{2}\log|\boldsymbol{Q}_\gamma \boldsymbol{H}_\gamma \boldsymbol{Q}_\gamma^\top|^{-\frac{1}{2}} - \frac{1}{2\sigma^2} R_{\boldsymbol{Q}_\gamma}, \tag{2.40}$$

where

$$R_{\boldsymbol{Q}_\gamma} = \left(\boldsymbol{Q}_\gamma \boldsymbol{y} - \boldsymbol{Q}_\gamma \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{Q}_\gamma \boldsymbol{H}_\gamma \boldsymbol{Q}_\gamma^\top\right)^{-1} \left(\boldsymbol{Q}_\gamma \boldsymbol{y} - \boldsymbol{Q}_\gamma \boldsymbol{X}\boldsymbol{\beta}\right). \tag{2.41}$$

So, $\mathcal{L}\left(\boldsymbol{\beta} | \boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right)$ is equal to a multivariate Normal distribution of $\boldsymbol{Q}_\gamma \boldsymbol{y}$.

$$\mathcal{L}\left(\boldsymbol{\beta} | \boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{Q}_\gamma \boldsymbol{y} \,\middle|\, \boldsymbol{\beta};\, \sigma^2 \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X}\right). \tag{2.42}$$

From this, the maximum with respect to $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ is found in closed form as the general least squares estimator:

$$\boldsymbol{\beta}_{\mathrm{R}_\gamma} = \left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{y}. \tag{2.43}$$

The estimator has the same form as the maximum likelihood estimate, but differs in the estimate of the parameter γ, as for REML γ is estimated by maximizing the profile restricted likelihood (Equation (2.35)).

Bayesian interpretation of restricted maximum likelihood estimation While restricted maximum likelihood estimation might seem heuristic, there is the following equivalent Bayesian interpretation. When instead of maximizing the likelihood over the fixed effects, these are integrated over a prior distribution, then, as the prior variance $\sigma_{\boldsymbol{\beta}}^2$ goes to infinity, the resulting marginal likelihood is proportional to the restricted likelihood. It follows, that the restricted maximum likelihood covariance parameters coincide with the maximum likelihood parameters of the marginal likelihood. Also, the posterior expectation of the fixed effects coincide with the restricted maximum likelihood estimator [Harville, 1974, Dempster et al., 1984].

As the integral over the fixed effects $\boldsymbol{\beta}$ of the posterior distribution $p\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}\right)$ of $\boldsymbol{\beta}$ equals one, the restricted likelihood $\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right)$ can be written as

$$\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) = \mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) \underbrace{\int p\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}\right) \mathrm{d}\boldsymbol{\beta}}_{1}. \tag{2.44}$$

$\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right)$ is not affected by the fixed effects and thus can be moved inside the integral.

$$\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) = \int \mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) p\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}\right) \mathrm{d}\boldsymbol{\beta}. \tag{2.45}$$

Assuming that the prior distribution over the fixed effects is an isotropic normal distribution with variance $\sigma_{\boldsymbol{\beta}}^2$. Then, the posterior distribution can be identified by completing the squares as

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\, \boldsymbol{m}_{\boldsymbol{\beta}} ;\, \boldsymbol{V}_{\boldsymbol{\beta}} \,\right), \tag{2.46}$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2} \left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X} \right)^{-1} + \frac{1}{\sigma_{\boldsymbol{\beta}}^2} \boldsymbol{I} \right)^{-1} \tag{2.47}$$

and

$$\boldsymbol{m}_{\boldsymbol{\beta}} = \frac{1}{\sigma^2} \boldsymbol{V}_{\boldsymbol{\beta}} \left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{Q}_\gamma \boldsymbol{y}. \tag{2.48}$$

It is easy to see, that in the limit of this distribution, as $\sigma_{\boldsymbol{\beta}}^2$ goes to infinity

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{Q}_\gamma \boldsymbol{y} ;\, \sigma^2 \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X} \right),$$

which equals $\mathcal{L}\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right)$.

For this case, we can write

$$\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) = \int \mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) \mathcal{L}\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right) \mathrm{d}\boldsymbol{\beta} \tag{2.49}$$

As shown before, the product of $\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right)$ and $\mathcal{L}\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right)$ is proportional to the full likelihood

$$\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right) = C \cdot \int \mathcal{L}\left(\boldsymbol{\beta}|\boldsymbol{Q}_\gamma \boldsymbol{y}, \gamma, \sigma^2\right) \mathrm{d}\boldsymbol{\beta}, \tag{2.50}$$

using Lemma C.8, the constant C can be identified as $|\boldsymbol{X}^\top \boldsymbol{X}|^{\frac{1}{2}}$ [Harville, 1974]. Solving the integral analytically, the restricted likelihood is obtained as being proportional to the marginal distribution of $\boldsymbol{y}$, when the fixed effects are integrated over a prior distribution with infinite variance.

$$\left(2\pi\sigma^2\right)^{-\frac{N-D}{2}} |\boldsymbol{X}^\top \boldsymbol{X}|^{\frac{1}{2}} |\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X}|^{-\frac{1}{2}} |\boldsymbol{H}_\gamma|^{-\frac{1}{2}} \exp -\frac{1}{2\sigma^2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{R}_\gamma}\right)^\top \boldsymbol{H}_\gamma^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\mathrm{R}_\gamma}\right), \tag{2.51}$$

where the posterior expectation of the fixed effects equals the restricted maximum likelihood estimator $\boldsymbol{\beta}_{\mathrm{R}_\gamma} = \left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{y}$.

2.2.4. Statistical testing using linear mixed models

Here we provide the likelihood ratio test for testing a fixed effects in a GWAS.

2.2.5. Likelihood ratio test

Also for linear mixed models the null distribution of the likelihood-ratio statistic for linear regression $\mathrm{LRT}_{\mathrm{LR}}$ can be approximated by a chi-square distribution with one degree of freedom, as the alternative model contains a single extra parameter, when testing for association of the $N \times 1$ SNP-vector $\boldsymbol{x}_\star$ while conditioning on the effect of any covariates contained in $\boldsymbol{X}$ [Hartley and Rao, 1967].

$$\mathrm{LRT}_{\mathrm{LR}} = 2 \log \frac{\max_{\boldsymbol{\beta},\beta_\star,\sigma^2,\sigma_{\mathrm{g}}^2} \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{x}_\star \cdot \beta_\star \,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)}{\max_{\boldsymbol{\beta},\sigma^2,\sigma_{\mathrm{g}}^2} \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{x}_\star \cdot 0 \,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)} \sim \chi_1^2. \tag{2.52}$$

As for linear regression P-values are computed from the survival function of the distribution.

2.3. Other methods for population structure correction

Apart from linear mixed models a range of methods have been proposed to correct for population structure. Even though linear mixed models have been shown to improve confounder correction over these alternative methods in a number of GWAS in maize, *A. thaliana*, potato and human [Yu et al., 2005a, Zhao et al., 2007, Malosetti et al., 2007, Kang et al., 2008, 2010], there might be benefits from combining mixed models with these other methods to get a more stringent correction.

Genomic control estimates the amount of inflation in a GWAS by comparing quantiles of the observed distribution of test statistics to the theoretical unconfounded distribution and corrects for inflation by simple matching of the median [Devlin and Roeder, 1999].

While genomic control corrects for inflation in a standard analysis, the other methods presented in this section correct for genetic structure by ways of modeling. The idea underlying these methods is that population structure is summarized by a small number of features that subsequently are included as covariates in a standard regression analysis. Similar to the different variants of the genetic relatedness matrix, these methods can be interpreted as either trying to estimating the confounding variable, or conditioning on other confounded variables (see Figure 2.1).

2.3.1. Genomic control

The idea underlying genomic control is simple. In order to correct for inflation of P values, it is possible to compare the distribution of test statistics obtained to their theoretical null distribution.

Genomic control defines the genomic inflation factor λ as the ratio of the observed median test statistic over the theoretical test statistic under an the null hypothesis in a theoretical unconfounded analysis.

$$\lambda = \frac{\text{median}\,(\text{LRT})}{\text{median}\,(Hnull)}. \tag{2.53}$$

So for the likelihood ratio test of a fixed effect in a linear model (like linear regression or linear mixed models) λ equals the median of twice the observed LRT over the median of a Chi-square distribution with one degree of freedom. Another common variant uses quantiles of the base ten logarithmic distribution of P values. In this case, λ is given by the median of the observed $-\log_{10}(P)$ over $-\log_{10}(0.5)$

Correction by genomic control is performed by dividing all test statistics by λ and can be shown to yield a conservative test.

From an intuitive standpoint the reasoning behind genomic control is that the vast majority if not all tested markers are not linked to causal loci and for this reason their test statistics should follow the distribution under the null hypothesis. Differently than methods that account for population structure by ways of modeling, genomic control uniformly affects the test statistics of unlinked as well as linked SNPs and does not change the order of test statistics. In experiments Price et al. [2006] show that such uniform adjustment is on the one side insufficient for markers showing stronger than average differentiation between ancestral populations and leads to a loss in power at markers having weaker differentiation. While approaches that model population structure can in some cases lead to an increase in power compared to an uncorrected analysis, the use of genomic control always reduces power.

Due to its simplicity though, correction by genomic control can be applied in conjunction with any model or statistical test, as long as the distribution of the test statistics is known or can be reliably estimated. For example it would be possible to apply genomic control to correct for residual inflation in an analysis using mixed model.

Besides for correction λ is also a commonly used measure of the calibration of the Type 1 error in a GWAS. A value of λ larger than one is an indicator of anti-conservativeness, or *inflation* of Type-1 errors, a value that is smaller that one indicates loss of power due to *deflation*.

Note however, that while values of λ larger than 1.05 and above in studies of human have usually been attributed to confounding [Burton et al., 2007], for studies of highly polygenic traits like body mass index or human height much larger values of λ have been shown to occur due to broad linkage to causal loci alone, without the presence of confounding [Speliotes et al., 2010, Allen et al., 2010, Yang et al., 2011b]. In this case correction by genomic control would yield overly conservative estimates.

2.3.2. Ancestry informative markers

Similar to the use of SNP makers to estimate relatedness in linear mixed models, such markers have also be used as covariates in a linear or logistic regression analysis. A concern when introducing covariates in a standard regression analysis is that the model suffers from a loss in degrees-of-freedom due to fitting the covariate effects in the null model. Only effects that are orthogonal to the set of covariates can show any association. To avoid a strong loss of power, the number of covariates should be kept small, instead of using all genome-wide markers. Still, the use of a relatively small number of widely spaced, randomly selected markers ($\approx 10^2$) has been shown to correct for population stratification [Setakis et al., 2006]. Also the use of a single marker that has been selected to be ancestry informative has been proposed [Wang et al., 2005].

Another approach that uses ancestry informative markers to place the individuals in subgroups of varying degree of admixture has been proposed by [Epstein et al., 2007]. Instead of including these markers as covariates in a regression analysis, the markers where used as a score that allows a stratified analysis of the individual subgroups.

2.3.3. Structured association

Instead of directly including markers that implicitly reflect population structure by differences in allele frequencies between populations, markers can also be used estimate explicit estimates of shared ancestry. From genetic markers, the software package STRUCTURE estimates a number of latent variables representing ancestry using Markov-chain Monte Carlo sampling [Pritchard et al., 2000a]. The model can be interpreted as clustering, where the membership variables represent shared ancestry between cluster members. These latent variables are then used as covariates in an association study [Pritchard et al., 2000b].

Compared to use of markers, summarizing the genetic variation in a small number of latent variables has the advantage that typically a fewer number of covariates are required to correct for population structure, yielding a smaller loss in power. The Markov-chain Monte Carlo algorithm, though, has a runtime that makes application of structured association infeasible on larger numbers of markers and individuals. Another problem is to properly determine the correct number of latent variables. Even though the STRUCTURE program outputs the likelihoods achieved for a number of latent variables, repeated runs of the algorithm would further increase the runtime of the method.

As latent variables capture differences in variation on a population scale, structured association is useful for correcting for population structure but unlikely to correct for cryptic relatedness present in the data.

2.3.4. Principal components analysis

Another latent variable method that has been applied to correct for population structure genetic studies is principal components analysis (PCA) [Zhang et al., 2002, Price et al., 2006].

Principal components (PCs) are estimated from a genome-wide covariance matrix sim-

ilar to the realized relationship matrix

$$\boldsymbol{K}_{\text{PCA}} = \frac{1}{S}\boldsymbol{G}\boldsymbol{G}^{\top}. \tag{2.54}$$

While PCA is computationally more efficient than structure association, as is requires computation of the first k Eigenvectors of $\boldsymbol{K}_{\text{PCA}}$, which can be performed in $O(N^2k)$ runtime.

As for structured association it is unclear how to best chose the number of principal components to use. By default EIGENSTRAT uses the first ten principal components. Identifying the correct number of components to use can be cumbersome. It was proposed to select the number of components such that the total genomic variation is significantly captured by the PCs [Patterson et al., 2006, Price et al., 2006]. In practice, however the number of components is typically chosen by comparing values of λ [Tian et al., 2008]. The first principle components tend to be dominated by large regions of strong linkage. As a result these components give little information on population structure [Astle and Balding, 2009]. Two to fifteen PCs ave been reported to by sufficient in practice [Astle and Balding, 2009].

It has also been shown that the number of PCs required for correction could be reduced by selecting the PCs by correlation to the phenotype [Novembre and Stephens, 2008, Lee et al., 2011].

2.4. Application of linear models to case-control phenotypes

Theoretically, linear models as linear regression or linear mixed models are not appropriate for modeling case-control phenotypes. These would ideally be modeled using a logistic or probit model. Use of linear models, however, severely reduces computational burden and avoids assessment of statistical significance of approximations to logistic or probit mixed models, for which no exact solution can be computed [Rasmussen and Williams, 2005, Agresti, 2002]. The linear approximation of case-control phenotypes finds broad use in practice as it has been shown to work well in practice for tests of sufficiently common SNPs and intermediate ratios of cases and controls [Price et al., 2006, Astle and Balding, 2009, Agresti, 2002]. Note however, that we observed skewed distributions of test statics when testing rare variants ($f < 0.01$) using a linear model [Listgarten et al., 2013b].

3. FaST linear mixed models for genome-wide association studies

Linear mixed models are among the richest class of models used today for genome-wide association studies, and have been shown to be capable of correcting for population structure, family structure, and cryptic relatedness [Astle and Balding, 2009, Price et al., 2010b]. In contrast to other methods that were discussed in Section 2.3, linear mixed models can capture all of these forms of relatedness simultaneously, without knowledge of which are present and without the need to tease them apart.

Despite of the benefits of linear mixed models, their widespread use on contemporary data sets has long been limited. The main reason for this is that statistical inference in linear mixed models involves computations that in terms of runtime scale cubic in the number of samples N. Even on studies involving a moderate number of samples, naive evaluation of the model for every single SNP is infeasible, as the typical number of SNPs in a genome-wide association study ranges from the hundreds of thousands to millions. Another bottleneck, when applying linear mixed models to large cohorts is that the memory requirements to store the complete relationship matrix is quadratic in the number of samples.

The situation has changed due to a recent focus on adapting linear mixed models to make them scalable to larger and larger studies [Aulchenko and de Koning, 2007, Kang et al., 2008, 2010, Zhang et al., 2010].

We start by an introduction to the Efficient Mixed Model Association (EMMA) algorithm. EMMA makes smart use of linear algebra to avoid repeated cubic operations on the covariance matrix in the mixed model when estimating the variance parameters in a test [Hartley and Rao, 1967, Patterson and Thompson, 1971, Kang et al., 2008]. Even though the computational savings over naive evaluation are tremendous, a spectral decomposition of an N-by-N matrix has to be computed for every marker that is tested, such that the cubic runtime requirements per test remain. Due to this runtime bottleneck this approach is practically limited to the analysis of genome-wide association studies on no more than several hundred samples.

As exact mixed model computations have commonly been considered too expensive to be applicable to even moderately sized cohorts, various approximations have been proposed, that aim at faster computations at the possible cost of reduced accuracy [Aulchenko and de Koning, 2007, Kang et al., 2010, Zhang et al., 2010, Svishcheva et al., 2012]. An overview over these methods can be found in Section 3.2. The most widely used approach, which has been shown to work well on many data sets, is to make the simplifying assumption that variance parameters are fixed for every SNP tested and can be estimated on the null model [Kang et al., 2010, Zhang et al., 2010]. Due to this simplification, cubic computations in the form of two spectral decompositions of N-by-N matrices have to be performed only once, for the null-model. The computations that are

required per SNP are reduced from cubic to quadratic in the number of samples. The storage requirements remain quadratic, as the algorithm still requires the full genetic relatedness matrix. Even though this approach has successfully been applied to cohorts of over ten thousand samples, together with the quadratic storage, the remaining cubic computations, which are hard to parallelize efficiently, are still a considerable bottleneck. In practice, the approach is not applicable to studies on extremely large cohorts, that are produced nowadays in order to gain sufficient power to get new insights on complex phenotypes, detect weak SNP effects, or effects of rare alleles [Do et al., 2011, Speliotes et al., 2010, Allen et al., 2010, Teslovich et al., 2010, Ehret et al., 2011].

With the new FaST-LMM algorithm [Lippert et al., 2011], presented in Section 3.3, we demonstrate, that exact mixed model computations are feasible on data sets of more than ten thousand samples, without making any simplifying assumptions. In contrast to earlier algorithms FaST-LMM requires only a single initial cubic spectral decomposition, while the computations that have to be performed per SNP tested are only *quadratic* in the number of samples. Thus, the runtime is N times faster than previous exact algorithms [Kang et al., 2008] and has the same runtime as when variance parameters are assumed to be fixed [Kang et al., 2010, Zhang et al., 2010][1].

Finally, in Section 3.4 an extension of the FaST-LMM algorithm is presented, that breaks the quadratic barrier by use of a reduced set of SNPs to measure genetic similarity, thereby achieving both *linear* runtime and *linear* memory use. Thus, FaST-LMM enables application to extremely large data sets. The computational gains rely on the number of markers used to estimate genetic similarity being smaller than the number of individuals in the study. On real data sets we show that a set of only a few thousand SNPs sampled linearly along the chromosome provides a good measure of genetic similarity and is sufficient to correct for population structure in a genome-wide association study. In Chapter 4 we show that by selecting a small number of markers by their association to the phenotype FaST-LMM even yields a consistent increase in power and better correction for genetic relatedness compared to use of genome-wide markers. Consequently, FaST-LMM provides extraordinary speedups when tens of thousands of individuals or more are analyzed, which we demonstrate by analyzing a dataset containing more than 120,000 individuals [Lippert et al., 2011].

3.1. Efficient mixed model association

The EMMA algorithm [Kang et al., 2008] builds on the insight that the maximum likelihood, or alternatively, the restricted maximum likelihood, of a linear mixed model can be rewritten as a function of just a single parameter, γ, the ratio of the environmental noise variance σ^2 to the genetic variance σ_g^2, rather than as a function of all of the model parameters [Hartley and Rao, 1967, Patterson and Thompson, 1971, Kang et al., 2008]. Given a value of γ the (restricted) maximum likelihood values for all of the model param-

[1] An algorithm similar to the full-rank FaST-LMM algorithm was proposed earlier in an unpublished PhD thesis that we obtained by personal correspondence [Astle, 2009] and is implemented in the R package GenAbel [Aulchenko et al., 2007] as well as in the MMM package [Pirinen et al., 2012]. An algorithm that is almost identical to the full-rank FaST-LMM algorithm has also been proposed and implemented as Genome-Wide Efficient Mixed Model Association (GEMMA) [Zhou and Stephens, 2012].

eters (i. e. the genetic and environmental variances along with the fixed-effects) follow in closed form. Consequently, the identification of the optimal parameters becomes an optimization problem over this single variable γ.

Additionally, EMMA makes clever use of spectral decompositions to reduce the cost of evaluating the log-likelihood for any value of γ, which is ordinarily cubic in the number of individuals, to linear in the number of individuals, once the two spectral decompositions are performed [Patterson and Thompson, 1971, Kang et al., 2008].

3.1.1. Maximum likelihood estimation

For maximum likelihood estimation, EMMA uses the formulation of the likelihood using the ratio $\gamma = \frac{\sigma_g^2}{\sigma^2}$ of the variance parameters σ_g^2 and σ^2, for which the maximum likelihood estimators for all other parameters follow in closed form[2]. By plugging the maximum likelihood estimators $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$, as given in Equation (2.26) and the maximum likelihood estimator ${\sigma^2}_{\mathrm{M}_\gamma}$ as given in Equation (2.27) back into the log-likelihood, the profile likelihood $\log \mathcal{L}(\gamma)$ is obtained as given in Equation (2.29):

$$\log \mathcal{L}(\gamma) = -\frac{N}{2}\left(1 + \log \frac{2\pi}{N}\right) - \frac{1}{2}\log |\boldsymbol{H}_\gamma| - \frac{N}{2}\log \boldsymbol{y}^\top \boldsymbol{P}_\gamma^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{P}_\gamma \boldsymbol{y},$$

where

$$\boldsymbol{P}_\gamma = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{H}_\gamma^{-1}.$$

According to Lemma C.27, $\boldsymbol{P}_\gamma^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{P}_\gamma$ equals the Moore-Penrose pseudoinverse of $\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}$, where $\boldsymbol{S} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right)$.

$$\log \mathcal{L}(\gamma) = -\frac{N}{2}\left(1 + \log \frac{2\pi}{N}\right) - \frac{1}{2}\log |\boldsymbol{H}_\gamma| - \frac{N}{2}\log \boldsymbol{y}^\top \left(\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}\right)^\dagger \boldsymbol{y}.$$

As shown in Lemma C.15, the economy spectral decomposition of $\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}$ can be obtained efficiently from $\boldsymbol{U}_{\boldsymbol{S}}\left(\boldsymbol{\Sigma} + \boldsymbol{I}\right)\boldsymbol{U}_{\boldsymbol{S}}$, the economy spectral decomposition of $\boldsymbol{S}\left(\boldsymbol{K} + \boldsymbol{I}\right)\boldsymbol{S}$, where $\boldsymbol{\Sigma}$ is obtained by subtracting one from each non-zero eigenvalue of $\boldsymbol{S}\left(\boldsymbol{K} + \boldsymbol{I}\right)\boldsymbol{S}$. The pseudoinverse of $\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}$ can be solved from this economy spectral decomposition, by inverting the non-zero eigenvalues:

$$\left(\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}\right)^\dagger = \boldsymbol{U}_{\boldsymbol{S}}\left(\underbrace{\gamma\boldsymbol{\Sigma} + \boldsymbol{I}_{N-D}}_{\boldsymbol{\Sigma}_\gamma}\right)^{-1} \boldsymbol{U}_{\boldsymbol{S}}^\top. \tag{3.1}$$

Let the spectral decomposition of $\boldsymbol{K}$ be $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. As shown in Lemma C.9, the spectral decomposition of $\boldsymbol{H}_\gamma = \gamma\boldsymbol{K} + \boldsymbol{I}$ is given by $\boldsymbol{U}\left(\gamma\boldsymbol{\Lambda} + \boldsymbol{I}\right)\boldsymbol{U}^\top$. By using the equalities $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}| \cdot |\boldsymbol{B}|$, for full rank matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and $|\boldsymbol{U}| = 1$, the logarithm of $|\boldsymbol{H}_\gamma|$ can be written as the logarithm of $|\gamma\boldsymbol{\Lambda} + \boldsymbol{I}|$. Plugging in these terms, we get

$$\log \mathcal{L}(\gamma) = -\frac{N}{2}\left(1 + \log \frac{2\pi}{N}\right) - \frac{1}{2}\log |\gamma\boldsymbol{\Lambda} + \boldsymbol{I}| - \frac{N}{2}\log \boldsymbol{y}^\top \boldsymbol{U}_{\boldsymbol{S}}\left(\gamma\boldsymbol{\Sigma} + \boldsymbol{I}\right)^{-1} \boldsymbol{U}_{\boldsymbol{S}}^\top \boldsymbol{y}. \tag{3.2}$$

[2] For keeping the presentation consistent with earlier literature [Hartley and Rao, 1967, Patterson and Thompson, 1971], we prefer to use γ instead of $\delta = \frac{1}{\gamma}$, as used in the original publication of EMMA [Kang et al., 2008]. All derivations are analogous and the runtime is the same.

In order to make efficient evaluation efficient, we write Equation (3.2) using only the entries of these matrices.

$$\log \mathcal{L}(\gamma) = -\frac{N}{2}\left(1 + \log\frac{2\pi}{N}\right) - \frac{1}{2}\sum_{n=1}^{N}\log\left(\gamma\left[\boldsymbol{\Lambda}\right]_{n,n} + 1\right) - \frac{N}{2}\log\left(\sum_{i=1}^{N-D}\frac{\left[\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}\right]_i^2}{\gamma\left[\boldsymbol{\Sigma}\right]_{i,i} + 1}\right).$$

The derivative with respect to γ is then given by

$$\frac{\partial \log \mathcal{L}(\gamma)}{\partial \gamma} = -\frac{1}{2}\sum_{n=1}^{N}\frac{\left[\boldsymbol{\Lambda}\right]_{n,n}}{\gamma\left[\boldsymbol{\Lambda}\right]_{n,n} + 1} - \frac{N}{2}\cdot\frac{\sum_{i=1}^{N-D}\frac{\left[\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}\right]_i^2\left[\boldsymbol{\Sigma}\right]_{i,i}}{\left(\gamma\left[\boldsymbol{\Sigma}\right]_{i,i}+1\right)^2}}{\sum_{j=1}^{N-D}\frac{\left[\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}\right]_j^2}{\gamma\left[\boldsymbol{\Sigma}\right]_{j,j}+1}}. \tag{3.3}$$

3.1.2. Restricted maximum likelihood estimation

EMMA maximizes the log restricted likelihood $\log \mathcal{L}\left(\gamma, {\sigma^2}_{\mathrm{R}_\gamma} | \boldsymbol{S}\boldsymbol{y}\right)$ in the form given in Equation (2.35) with σ^2 profiled out. As in the case of maximum-likelihood estimation described in Section 3.1.1 the economy spectral decomposition of $\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}$ is obtained efficiently from the economy spectral decomposition $\boldsymbol{U}_{\boldsymbol{S}}\left(\boldsymbol{\Sigma} + \boldsymbol{I}_{N-D}\right)\boldsymbol{U}_{\boldsymbol{S}}$ of $\boldsymbol{S}\left(\boldsymbol{K} + \boldsymbol{I}_N\right)\boldsymbol{S}$.

$$\log \mathcal{L}\left(\gamma, {\sigma^2}_{\mathrm{R}_\gamma} | \boldsymbol{S}\boldsymbol{y}\right) = -\frac{N-D}{2}\left(1 + \log\frac{2\pi}{N-D}\right) - \frac{1}{2}\log\left|\underbrace{\gamma\boldsymbol{\Sigma} + \boldsymbol{I}_{N-D}}_{\boldsymbol{\Sigma}_\gamma}\right| - \frac{N-D}{2}\log R_{\boldsymbol{S}},$$

where the residual is given by

$$R_{\boldsymbol{S}} = \boldsymbol{y}^{\top}\boldsymbol{U}_{\boldsymbol{S}}\left(\underbrace{\gamma\boldsymbol{\Sigma} + \boldsymbol{I}_{N-D}}_{\boldsymbol{\Sigma}_\gamma}\right)^{-1}\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}.$$

Again, this log-likelihood can be evaluated efficiently for any value of γ in $O(N)$ as

$$-\frac{N-D}{2}\left(1 + \log\frac{2\pi}{N-D}\right) - \frac{1}{2}\sum_{n=1}^{N-D}\log\left(\gamma\left[\boldsymbol{\Sigma}\right]_{n,n} + 1\right) - \frac{N-D}{2}\log\left(\sum_{i=1}^{N-D}\frac{\left[\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}\right]_i^2}{\gamma\left[\boldsymbol{\Sigma}\right]_{i,i} + 1}\right).$$

The same is true for the derivative with respect to γ given by

$$\frac{\partial \log \mathcal{L}\left(\gamma, {\sigma^2}_{\mathrm{R}_\gamma} | \boldsymbol{S}\boldsymbol{y}\right)}{\partial \gamma =} - \frac{1}{2}\sum_{n=1}^{N-D}\frac{\left[\boldsymbol{\Sigma}\right]_{n,n}}{\gamma\left[\boldsymbol{\Sigma}\right]_{n,n} + 1} - \frac{N-D}{2}\cdot\frac{\sum_{i=1}^{N-D}\frac{\left[\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}\right]_i^2\left[\boldsymbol{\Sigma}\right]_{i,i}}{\left(\gamma\left[\boldsymbol{\Sigma}\right]_{i,i}+1\right)^2}}{\sum_{j=1}^{N-D}\frac{\left[\boldsymbol{U}_{\boldsymbol{S}}^{\top}\boldsymbol{y}\right]_j^2}{\gamma\left[\boldsymbol{\Sigma}\right]_{j,j}+1}}. \tag{3.4}$$

3.1.3. Optimizing the ratio of variances

In order to solve the non-convex optimization over the ratio of variances γ, EMMA applies a combination of grid search and a derivative based method. To bracket local minima, the derivative of the likelihood as in Equation (3.3) or the derivative of the restricted likelihood as in Equation (3.4) is evaluated on one hundred equally spaced points on the logarithm of γ ranging from -5 to 5. For every two consecutive points, where the derivative changes, a root finder based on Brent's algorithm is applied to equate the derivative to zero within the respective interval and retrieve the local optimum.

3.1.4. Runtime and memory footprint

Once $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{U}^\top \boldsymbol{y}$ are computed, both the log-likelihood as well as the derivative with respect to γ can be evaluated in $O(N)$ for any value of γ. Assuming that the number of evaluations of the derivative is given by a constant C, the cost of finding an optimal γ for a single SNP is $O(C \cdot N)$. The required upfront computations are computation of the eigenvalues $\boldsymbol{\Lambda}$ of $\boldsymbol{K}$, the economy spectral decomposition $\boldsymbol{U_S} (\boldsymbol{\Sigma} + \boldsymbol{I}) \boldsymbol{U_S}$ of $\boldsymbol{S}\boldsymbol{H}_1\boldsymbol{S}$, and multiplication of the phenotype by $\boldsymbol{U_S}^\top$. A problem that arises when this algorithm is applied to GWAS is, that for every SNP tested, the matrix $\boldsymbol{X}$ of fixed effects is a different one. It follows that the matrix $\boldsymbol{S} = \left(\boldsymbol{I} - \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right)$ is a different one for each SNP. As a result a new economy spectral decomposition of an N-by-N matrix $\boldsymbol{S}\boldsymbol{H}_1\boldsymbol{S}$ is required. As the rank of $\boldsymbol{S}\boldsymbol{H}_1\boldsymbol{S}$ equals $N - D$, the economy spectral decompositions could be computed in $O(N^2 \cdot (N - D))$ for example using iterative methods. In practice though, the number of fixed effects D used in genome-wide association studies, is not more than a one digit integer and can be treated as a constant. It follows, that the required computations for testing all SNPs are in $O(C \cdot N + S \cdot N^3) = O(S \cdot N^3)$, where S is the number of all SNPs tested. If each SNP only is into loaded to memory while being tested, the memory footprint is dominated by the cost of storing the genetic similarities $\boldsymbol{K}$, given by $O(N^2)$.

3.2. Efficient approximations to the mixed model

In order to apply the linear mixed model to the analysis of larger data several approximations have been proposed. The earliest such approximation was the Genomewide Rapid Association using Mixed Model and Regression (GRAMMAR) algorithm [Aulchenko and de Koning, 2007], which uses a mixed model only in a single upfront to compute a population-structure corrected version of the phenotype which can be analyzed by standard linear regression. The EMMAX and P3D algorithms avoid repeated cubic computations by estimating the ratio γ of variance parameters in the mixed model only once, keeping it fixed across all tests.

3.2.1. Generating stratified pseudo-phenotypes by prediction

The idea of the GRAMMAR algorithm is to use a linear mixed model to generate stratified pseudo-phenotypes, which can be analyzed efficiently by a linear regression and is implemented in the GenAbel package [Aulchenko and de Koning, 2007, Aulchenko et al., 2007]. From a linear mixed model without including a SNP, the pseudo-phenotypes are obtained by subtracting the BLUP as derived in Section 2.2.2 of the random effects from the phenotype.

$$\boldsymbol{y}_{\text{strat}} = \boldsymbol{y} - \underbrace{\sigma_{\text{g}}^2 \boldsymbol{k}_{:,:} \left(\sigma_{\text{g}}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}_{\text{BLUP}}. \tag{3.5}$$

The remaining free parameters σ^2, σ_{g}^2 and $\boldsymbol{\beta}$ can be found by either maximum likelihood or restricted maximum likelihood.

GRAMMAR has been shown to lead to overly conservative correction as upfront stratification ignores a possible linear-additive interactions between the BLUP and the effects

of the SNPs tested [Aulchenko and de Koning, 2007]. In order to correct for conservativeness, GRAMMAR typically yield genomic control values (λ, see Section 2.3.1) smaller than one. It has been proposed to use correction by genomic control to account for this conservativeness [Amin et al., 2007].

Runtime and memory footprint

The computations required for testing a single SNP on the pseudo-phenotype by linear regression is linear in the number of individuals. As GRAMMAR uses a standard linear mixed model to generate the pseudo phenotype, the runtime required for optimizing the parameters on the null model and computing the best linear unbiased prediction is $O(N^3)$ and the memory requirement is $O(N^2)$, dominated by the size of the genetic similarity matrix.

3.2.2. Linear mixed models with fixed ratio of variances

A practical approximation that leads to a considerable speedup over exact linear mixed model computations was obtained by estimating the variance parameters only once, rather than re-estimating these per SNP [Kang et al., 2010, Zhang et al., 2010]. For many studies of interest, this approximation is expected to work nearly as well as the exact model, yet it made problems that were computationally infeasible, now feasible. The algorithm has successfully been applied to the analysis of genome-wide association studies containing five thousand samples [Kang et al., 2010, Burton et al., 2007]. But on a study in Mouse it has been shown that fixing γ results in a loss in power compared to an exact mixed model [Zhou and Stephens, 2012].

The algorithm performs maximum likelihood or restricted maximum likelihood estimation on the null model using the EMMA algorithm, as shown in Section3.1 to obtain an estimate γ_0 for the ratio of variances is obtained. Given this value of γ_0, the inverse and the determinant of $\boldsymbol{H}_{\gamma_0} = (\gamma \boldsymbol{K} + \boldsymbol{I})$, and in case of REML estimation the determinant of $\boldsymbol{X}^\top \boldsymbol{H}_{\gamma_0}^{-1} \boldsymbol{X}$ can be computed once and used to test all SNPs.

Runtime and memory footprint

The runtime to find γ_0, and compute all terms involving $\boldsymbol{H}_{\gamma_0}$ is given by $O(N^3)$. Evaluation of the likelihood for testing a single SNP requires computation of a matrix vector product with runtime of $O(N^2)$. Storage of the inverse of $\boldsymbol{H}_0$ requirement $O(N^2)$ memory. The total asymptotic runtime for testing S markers in a GWAS follows as $O(N^2 S)$.

3.2.3. Compressed mixed models

Another way to speed up linear mixed models is by approximating the relationship matrix by a matrix that can be inverted more efficiently than standard relationship matrices. Compressed mixed models [Zhang et al., 2010] perform a clustering of the individuals present in a study to come up with a simpler covariance structure, where the cubic dependency in the number of individuals is reduced to a cubic dependency on the number of clusters G of genetically similar individuals.

In this approach, individuals are first grouped into G genetically similar groups. Then, for the purposes of confounder correction, members of a group were assumed to be genetically identical. A G-by-G between group genetic similarity matrix $\boldsymbol{K}_{\mathrm{G}}$ is obtained by averaging the genetic similarities over all group members. Finally, a LMM-analysis is performed, using the log likelihood

$$\log \mathcal{L}\left(\sigma^2, \sigma_{\mathrm{g}}^2, \boldsymbol{\beta}, \boldsymbol{Z}\right) = \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{Z}\boldsymbol{K}_{\mathrm{G}}\boldsymbol{Z}^\top + \sigma^2 \boldsymbol{I}\right), \tag{3.6}$$

where $\boldsymbol{Z}$ is an N-by-G binary indicator matrix, that assigns each of N individuals to exactly one of the G groups.

Using the Woodbury-Sherman Lemma as well as the matrix determinant lemma, all expensive computations involving the genetic similarities can be computed in $O(G^3)$[3].

$$\begin{aligned}\left(\gamma \boldsymbol{Z}\boldsymbol{K}_{\mathrm{G}}\boldsymbol{Z}^\top + \boldsymbol{I}\right)^{-1} &= \boldsymbol{I} - \boldsymbol{Z}\left(\frac{1}{\gamma}\boldsymbol{K}_{\mathrm{G}}^{-1} + \boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top. \\ \left|\gamma \boldsymbol{Z}\boldsymbol{K}_{\mathrm{G}}\boldsymbol{Z}^\top + \boldsymbol{I}\right| &= \left|\frac{1}{\gamma}\boldsymbol{K}_{\mathrm{G}}^{-1} + \boldsymbol{Z}^\top\boldsymbol{Z}\right| \cdot |\gamma \boldsymbol{K}_{\mathrm{G}}| \cdot |\boldsymbol{I}|\,.\end{aligned}$$

To determine the number clusters, a hierarchical clustering is computed based on genetic relatedness. Then, for a range of distance thresholds , the likelihood of the model is computed on the null model. The final analysis is performed on the distance threshold maximizing the likelihood [Zhang et al., 2010].

Runtime and memory footprint

Compression reduces the runtime of the expensive computations to cubic in the number of groups G. When the number of groups is much smaller than the number of individuals, the computational savings are tremendous. Determining the similarities and clustering the individuals, however, is a quadratic operation in the number of individuals. Finding an appropriate number of clusters for a data set is an important issue, as it greatly influences the results . For this purpose, the authors propose to determine the appropriate number of clusters by maximizing the model likelihood over the distance threshold in the hierarchical clustering. As such an approach requires repeated solutions of the mixed model, compression can hardly be considered a speedup in this case [Zhang et al., 2010].

3.3. FaST-linear mixed models

Here we describe our approach called FaST-LMM, which stands for *fa*ctored *s*pectrally *t*ransformed *l*inear *m*ixed *m*odels. The algorithm computes the same exact linear mixed model as the EMMA algorithm described in 3.1. Similarly to the EMMA algorithm, the spectral decomposition of the genetic similarity matrix is used to cache expensive computations. But unlike the EMMA algorithm FaST-LMM provides a mathematical reformulation of the likelihood that allows to re-use of a single spectral decomposition

[3]The original publication does not contain details about the exact computations performed [Zhang et al., 2010].

for all tests. As a result, the FaST-LMM algorithm performs exact linear mixed model inference in a runtime that is N times faster than EMMA.

A key insight behind our approach is that the spectral decomposition of the genetic similarity matrix allows the likelihood of the linear mixed model to be refactored in such a way that it is directly analogous to the likelihood of a linear regression model. Intuitively, our algorithm algebraically transforms/rotates the target data (the phenotypes) and the input data (the SNPs and covariates) in such a way that that this rotated data effectively contains pseudo-individuals that are uncorrelated, and hence can be analyzed with a linear regression model that is linear in the number of individuals.

3.3.1. Maximum likelihood estimation

In what follows, we derive formulas that allow for efficient evaluation of the log likelihood, and the maximum likelihood parameters.

Linear-time evaluation of the log likelihood

Applying the formula for the N-variate Normal distribution to the log-likelihood parameterized by the ratio of variance parameters γ, as in Equation (2.25), we obtain

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \gamma, \sigma^2\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\log\left|\underbrace{\gamma\boldsymbol{K} + \boldsymbol{I}}_{\boldsymbol{H}_\gamma}\right| - \frac{1}{2\sigma^2}R,$$

where the residual term equals

$$R = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\underbrace{\gamma\boldsymbol{K} + \boldsymbol{I}}_{\boldsymbol{H}_\gamma}\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right).$$

Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \boldsymbol{K}$ be the spectral decomposition of $\boldsymbol{K}$, and noting that $\boldsymbol{I} = \boldsymbol{U}\boldsymbol{U}^\top$.

$$R = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{U}\left(\gamma\boldsymbol{\Lambda} + \boldsymbol{I}\right)\boldsymbol{U}^\top\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right).$$

The inverse can be rewritten using the property that $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, the fact that $\boldsymbol{U}^{-1} = \boldsymbol{U}^\top$ and $\boldsymbol{U}^{-\top} = \boldsymbol{U}$. Thus, after additionally pushing $\boldsymbol{U}$ out from the covariance term so that it now acts as a rotation matrix on the inputs $\boldsymbol{X}$ and targets $\boldsymbol{y}$

$$R = \left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\gamma\boldsymbol{\Lambda} + \boldsymbol{I}\right)^{-1} \left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}\right). \tag{3.7}$$

Also the determinant can be written using the spectral decomposition of $\boldsymbol{K}$, where the property that $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$ and the fact that $|\boldsymbol{U}| = \left|\boldsymbol{U}^\top\right| = 1$ are used.

$$|\gamma\boldsymbol{K} + \boldsymbol{I}| = |\gamma\boldsymbol{\Lambda} + \boldsymbol{I}|.$$

Using this determinant and the expression for R in Equation (3.7) the log likelihood, we obtain

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \gamma, \sigma^2\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\log|\gamma\boldsymbol{\Lambda} + \boldsymbol{I}| - \frac{1}{2\sigma^2}R. \tag{3.8}$$

As the covariance matrix of the Normal distribution is now a diagonal matrix $(\gamma \boldsymbol{\Lambda} + \boldsymbol{I})$, the log likelihood can be rewritten as the sum over N terms, yielding

$$-\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{n=1}^{N}\log\left(\frac{1}{\gamma\lambda_n + 1}\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\frac{\left([\boldsymbol{U}^\top \boldsymbol{y}]_n - [\boldsymbol{U}^\top \boldsymbol{X}]_{n,:}\boldsymbol{\beta}\right)^2}{\frac{1}{\gamma\lambda_n+1}}. \tag{3.9}$$

Note that this expression implies, that the model likelihood equals the product of N single-variate Normal distributions, on the data transformed by $\boldsymbol{U}^\top$.

$$\mathcal{L}\left(\gamma, \sigma^2, \boldsymbol{\beta}\right) = \prod_{n=1}^{N}\mathcal{N}\left([\boldsymbol{U}^\top \boldsymbol{y}]_n \,\middle|\, [\boldsymbol{U}^\top \boldsymbol{X}]_{n,:}\boldsymbol{\beta};\, \sigma^2\frac{1}{\gamma\lambda_n + 1}\right).$$

The "Fa" in FaST-LMM gets its name from this factored likelihood.

Having pre-computed the spectral decomposition of $\boldsymbol{K}$, we can rotate the phenotype and all SNPs once to get $\boldsymbol{U}^\top \boldsymbol{X}$ and $\boldsymbol{U}^\top \boldsymbol{y}$. Given the parameters γ, σ^2 and $\boldsymbol{\beta}$ each evaluation of the likelihood is now linear in the number of individuals N, as compared to cubic for direct evaluation of the log likelihood.

Finding the maximum likelihood fixed effect weights efficiently

We take the gradient of the log likelihood in Equation (3.8) with respect to $\boldsymbol{\beta}$ and set it to zero, giving

$$\frac{1}{\sigma^2}\left(\boldsymbol{U}^\top \boldsymbol{X}\right)^\top (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\left(\boldsymbol{U}^\top \boldsymbol{y}\right) - \frac{1}{\sigma^2}\left(\boldsymbol{U}^\top \boldsymbol{X}\right)^\top (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\left(\boldsymbol{U}^\top \boldsymbol{X}\right)\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \boldsymbol{0}.$$

Solving for $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$, we obtain

$$\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \left(\left(\boldsymbol{U}^\top \boldsymbol{X}\right)^\top (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\left(\boldsymbol{U}^\top \boldsymbol{X}\right)\right)^{-1}\left(\boldsymbol{U}^\top \boldsymbol{X}\right)^\top (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\left(\boldsymbol{U}^\top \boldsymbol{y}\right). \tag{3.10}$$

As $(\gamma\boldsymbol{\Lambda} + \boldsymbol{I})$ is a diagonal matrix, the matrix products again can be written as a sum over N independent terms, yielding

$$\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \left(\sum_{n=1}^{N}\frac{1}{\frac{1}{\gamma\lambda_n+1}}[\boldsymbol{U}^\top \boldsymbol{X}]_{n,:}{}^\top [\boldsymbol{U}^\top \boldsymbol{X}]_{n,:}\right)^{-1}\left(\sum_{n=1}^{N}\frac{1}{\frac{1}{\gamma\lambda_n+1}}[\boldsymbol{U}^\top \boldsymbol{X}]_{n,:}{}^\top [\boldsymbol{U}^\top \boldsymbol{y}]_n\right), \tag{3.11}$$

which is analogous to linear regression estimates for $\boldsymbol{\beta}$ on the rotated data. Assuming that all the terms involving the spectral decomposition of $\boldsymbol{K}$ are precomputed, this equation can be evaluated in $O(N)$.

Finding the maximum likelihood environmental variance efficiently

We start by substituting $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ from the previous section into the log likelihood, Equation (3.9), and set the derivative with respect to σ^2 to zero, giving

$$-\frac{1}{2}\left(\frac{N}{{\sigma^2}_{\mathrm{M}_\gamma}} - \frac{1}{{\sigma^4}_{\mathrm{M}_\gamma}}\sum_{n=1}^{N}\frac{\left([\boldsymbol{U}^\top \boldsymbol{y}]_n - [\boldsymbol{U}^\top \boldsymbol{X}]_{n,:}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^2}{\frac{1}{\gamma\lambda_n+1}}\right) = 0. \tag{3.12}$$

Multiplying both sides by $2{\sigma^4}_{\mathrm{M}_\gamma}$ and solving for ${\sigma^2}_{\mathrm{M}_\gamma}$, we get

$${\sigma^2}_{\mathrm{M}_\gamma} = \frac{1}{N}\sum_{n=1}^{N}\frac{\left([\boldsymbol{U}^\top\boldsymbol{y}]_n - [\boldsymbol{U}^\top\boldsymbol{X}]_{n,:}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^2}{\frac{1}{\gamma\lambda_n+1}}. \tag{3.13}$$

This equation also can be evaluated in $O(N)$.

Efficient evaluation of the maximum likelihood

Plugging in ${\sigma^2}_{\mathrm{M}_\gamma}$ and $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ into Equation (3.9), the log likelihood becomes a function only of γ:

$$\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{M}_\gamma}, \gamma, {\sigma^2}_{\mathrm{M}_\gamma}\right) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{n=1}^{N}\log\left(\frac{1}{\gamma\lambda_n+1}\right) - \frac{N}{2} - \frac{N}{2}\log\left(\frac{1}{N}R_{\mathrm{M}_\gamma}\right), \tag{3.14}$$

where the residual term includes the maximum likelihood weights.

$$R_{\mathrm{M}_\gamma} = \left(\sum_{n=1}^{N}\frac{\left([\boldsymbol{U}^\top\boldsymbol{y}]_n - [\boldsymbol{U}^\top\boldsymbol{X}]_{n,:}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^2}{\frac{1}{\gamma\lambda_n+1}}\right).$$

As described next, we optimize this function of γ using a one-dimensional numerical optimizer to find the maximum likelihood value of γ, from which the maximum likelihood values of all the parameters can be directly computed.

3.3.2. Restricted maximum likelihood estimation

So far the derivations have been limited to maximum likelihood parameter estimation. However, it is straightforward to extend these results to the restricted log likelihood, which comprises the log likelihood with $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ plugged in, plus three additional terms as can easily be seen from the restricted likelihood $\log\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right)$ given in Equation (2.51) [Kang et al., 2008, Harville, 1974].

$$\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{R}_\gamma}, \gamma, \sigma^2\right) + \frac{1}{2}\left(D\log\left(2\pi\sigma^2\right) + \frac{1}{2}\log\left|\boldsymbol{X}^\top\boldsymbol{X}\right| - \frac{1}{2}\log\left|\boldsymbol{X}^\top\left(\gamma\boldsymbol{K}+\boldsymbol{I}\right)^{-1}\boldsymbol{X}\right|\right), \tag{3.15}$$

where $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ has the same form as $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$.

Again, using the spectral decomposition of $\boldsymbol{K}$, the restricted log likelihood $\log\mathcal{L}\left(\gamma, \sigma^2|\boldsymbol{S}\boldsymbol{y}\right)$ becomes

$$\begin{aligned}\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{R}_\gamma}, \gamma, \sigma^2\right) + \frac{D}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2}\log\left|\boldsymbol{X}^\top\boldsymbol{X}\right| \\ - \frac{1}{2}\log\left|\sum_{n=1}^{N}\frac{1}{\frac{1}{\gamma\lambda_n+1}}[\boldsymbol{U}^\top\boldsymbol{X}]_{n,:}{}^\top[\boldsymbol{U}^\top\boldsymbol{X}]_{n,:}\right|.\end{aligned} \tag{3.16}$$

Neglecting the cubic dependence on d for computing the determinants, these additional terms can be evaluated in time complexity $O(N)$.

The restricted maximum likelihood variance estimate is given by

$${\sigma^2}_{\mathrm{R}_\gamma} = \frac{1}{N-D}\sum_{N=1}^{N}\frac{\left([\boldsymbol{U}^\top\boldsymbol{y}]_n - [\boldsymbol{U}^\top\boldsymbol{X}]_{n,:}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^2}{\frac{1}{\gamma\lambda_n+1}}. \tag{3.17}$$

3.3.3. Optimization of the ratio of variances

As we've just shown, finding the maximum likelihood solution or the restricted maximum likelihood solution of the mixed model ($\log \mathcal{L}\left(\sigma^2, \sigma_g^2, \boldsymbol{\beta}\right)$) is equivalent to finding the value of γ that maximizes the log likelihood $\log \mathcal{L}\left(\boldsymbol{\beta}_{M_\gamma}, \gamma, {\sigma^2}_{M_\gamma}\right)$ or the log restricted likelihood $\log \mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{Sy}\right)$, a non-convex optimization problem. To avoid local maxima in FaST-LMM, a quasi-exhaustive one dimensional optimization scheme similar to the one used by EMMA in Section 3.1.3 is applied. In order to bracket local minima, we evaluate the maximum of the log likelihood for 100 equidistant values of $\log \gamma$, ranging -10 to 10. Note that setting the number of grid-values to 100 is a conservative choice that in our experience can considerably reduced (in typical settings 10 should be sufficient) to obtain a mild speedup. Then, we apply the derivative-free version of Brent's method (a 1D numerical optimization algorithm) to find the locally optimal γ in each bracket where the middle log likelihood is higher than the log likelihoods of the neighboring evaluations.

3.3.4. Time and space complexity

Given γ and having pre-computed the spectral decomposition of $\boldsymbol{K}$, each evaluation of the likelihood has time complexity that is linear in N. Consequently, when testing S SNPs in a genome-wide association study, the time complexities are $O(N^3)$ for finding all eigenvalues ($\boldsymbol{\Lambda}$) and eigenvectors ($\boldsymbol{U}$) of $\boldsymbol{K}$, $O(N^2 S)$ for rotating the phenotype vector $\boldsymbol{y}$, and all of the SNP and covariate data (i. e. computing $\boldsymbol{U}^\top \boldsymbol{y}$ and $\boldsymbol{U}^\top \boldsymbol{X}$), and $O(CNS)$ for performing C evaluations of the log likelihood during the one-dimensional optimization over γ. The total time complexity of FaST-LMM, given $\boldsymbol{K}$, is therefore $O(N^3 + N^2 S + CNS) = O(N^2 S)$. If optionally γ is kept fixed to its value from the null model (as done in EMMAX/P3D), this complexity reduces slightly without reducing the asymptotic complexity to $O(N^3 + N^2 S + CN) = O(N^2 S)$. The size of both $\boldsymbol{K}$ and $\boldsymbol{U}$ is $O(N^2)$, which dominates the space complexity, as each SNP can be processed independently so that there is no need to load all SNP data into memory at once. In most applications, the number of fixed effects per test, D, is a single digit integer and is omitted in these expressions because its contribution is negligible.

3.4. FaST-linear mixed models in linear time

In general, obtaining the required rotation matrix (i. e. it via a spectral decomposition) for FaST-LMM is a cubic operation in the number of individuals. When the number of SNPs used to construct the genetic similarity matrix is less than the number of individuals, however, the required rotation matrix can be obtained in time linear in the number of individuals (and quadratic in the number of SNPs). Intuitively, these savings can be achieved because the intrinsic dimensionality of the space of the SNPs and individuals can never be higher than the smaller of these two values (i. e. the rank of the data matrix used to construct the similarities is at most the smaller of these two values). Thus, we can always choose to perform operations in the smaller space without any loss of information. That is, the computations remain exact. Once the rotation matrix has been computed, performing the rotations is linear in the size of the matrix and the number of SNPs tested. On one or a small number of processors, these rotations require the

most time. However, these rotations are easily parallelized, making construction of the rotation matrix the dominant computation.

This linear-time speed-up requires use of a particular type of genetic similarity matrix—in particular, the realized relationship matrix (see Section 2.2.1). We show that performance of the linear mixed model on two data sets with this genetic similarity is comparable to that of a linear mixed model with an identity by state matrix. Additionally, it was reported that use of an identity by state matrix often outperforms indentity by descent in a linear mixed model [Kang et al., 2010]. As shown in Section 3.4.1, when realized relationships are used, the spectral decomposition required by the linear mixed model can be obtained directly from the data bypassing explicit computation of the realized relationship matrix. The required time complexity is linear in the number of individuals. This computation is possible because of the well-known relationship between the spectral decomposition of a covariance matrix and the singular value decomposition of the data from which a covariance matrix is estimated (e. g. [Berrar et al., 2003]). Consequently, an exact linear mixed model analysis remains linear in the number of individuals.

Our approach using spectral transformations offers speedups beyond those of EMMA and EMMAX even when the realized relationship matrix is not used, provided the rank of the matrix is low—that is, less than the number of individuals. The resulting speed-up, however, is then quadratic in the number of individuals and linear in the rank of the matrix, because even when the matrix is low rank, the starting point of the computation is the matrix itself, an object that is quadratic in the number of individuals.

3.4.1. Relating spectral decomposition and singular value decomposition

Before we discuss the low-rank version of FaST-LMM, it will be useful to review the relationship between spectral decomposition and singular value decomposition for matrices, for which the factorization $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{G}^\top$ is known, such as the realized relationship matrix or the Eigenstrat covariance matrix [Price et al., 2006]. In this section, we shall refer to a matrix $\boldsymbol{K}$ that has this form as being *factored*.

The spectral decomposition of the genetic similarity matrix, $\boldsymbol{K}$, given by $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \boldsymbol{K}$, yields the eigenvectors ($\boldsymbol{U}$) and eigenvalues ($\boldsymbol{\Lambda}$) of $\boldsymbol{K}$. In general, this decomposition can be determined by first computing the genetic similarity matrix ($\boldsymbol{K}$), and then taking the spectral decomposition of it. For many measures of genetic similarity, including realized relationship matrix, the time complexity of computing $\boldsymbol{K}$ is $O(N^2 S_\mathrm{c})$, where S_c is the number of SNPs used to compute $\boldsymbol{K}$. Given the genetic similarity matrix, the eigenvalues and eigenvectors of $\boldsymbol{K}$ can then be found solving the spectral decomposition at a time complexity of $O(N^3)$ and space complexity of $O(N^2)$. If only the first k eigenvectors are desired, the computation can be achieved with other algorithms that have time complexity of $O(N^2 k)$ and a space complexity of $O(N^2)$.

When $\boldsymbol{K}$ if factored, however, one can bypass explicit computation of $\boldsymbol{K}$, obtaining the required eigenvectors and eigenvalues by direct application of an singular value decomposition to the $S_\mathrm{c} \times N$ data matrix of SNP markers at a time complexity of $O(N S_\mathrm{c}^2)$ (or $O(N S_\mathrm{c} k)$ for only the top k eigenvectors using, for example, [Tipping and Bishop, 1999]) and space complexity of $O(N S_\mathrm{c})$. Construction of $\boldsymbol{K}$ can be bypassed because (1) the eigenvectors (equivalently, singular vectors) of the factored matrix are the same as the singular vectors of the data matrix, and (2) the eigenvalues (equivalently singular

values) of the factored matrix are the square of the singular values of the data matrix. This relationship is widely-known (e. g. [Berrar et al., 2003]) and is demonstrated below. In our experiments, FaST-LMM bypasses computation of the factored matrix to obtain the required spectral decomposition whenever $S_c < N$.

Note that, when the rank of $\boldsymbol{K}$ is less than the number of individuals N (such as occurs when the data matrix used to compute the factored genetic similarity matrix contains fewer SNPs than individuals), the singular value decomposition with time cost $O(NS_c^2)$ is actually an *economy* singular value decomposition, that is, it yields only the first S_c eigenvectors. This set of eigenvectors is denoted $\boldsymbol{U}_1$ in Section 3.4.2 and referred to as the economy spectral decomposition.

We now demonstrate the relationship just noted. Let $\boldsymbol{G} \in \mathbb{R}^{N \times S_c}$ be the matrix containing the set of SNPs used to compute the factored matrix, $\boldsymbol{K}$, defined as

$$\boldsymbol{K} \equiv \boldsymbol{G}\boldsymbol{G}^\top. \tag{3.18}$$

Let $\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{G}$. Then Equation (3.18) can be rewritten as

$$\boldsymbol{K} = \left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^\top\right)^\top = \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^\top. \tag{3.19}$$

Because $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$, we obtain

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \tag{3.20}$$

where $\boldsymbol{\Lambda}_{ii} \equiv \boldsymbol{\Lambda}_{ii}^{\frac{1}{2}}\boldsymbol{\Lambda}_{ii}^{\frac{1}{2}}$. By definition, $\boldsymbol{U}$ consists of the eigenvectors of $\boldsymbol{K}$ (because it satisfies the properties of a spectral decomposition of $\boldsymbol{K}$, namely that $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{\Lambda}$ is diagonal and $\boldsymbol{U}$ contains orthonormal vectors). Furthermore, the eigenvalues of $\boldsymbol{K}$ are clearly given by $\boldsymbol{\Lambda}_{ii}^{\frac{1}{2}}\boldsymbol{\Lambda}_{ii}^{\frac{1}{2}}$. Consequently, we can obtain the spectral decomposition of $\boldsymbol{K}$ by computing the singular value decomposition of $\boldsymbol{G}$, which has time cost $O(NS_c^2)$.

3.4.2. Low rank linear mixed models

Next we consider the case where the rank of $\boldsymbol{K}$, k, is low ($k < N$) (i. e. $\boldsymbol{K}$ is not full rank). This case will occur when the realized relationship matrix is used and the number of SNPs used to estimate it, $S_c = k$, is smaller than N, or when we use a rank k approximation of the genetic similarity matrix as mentioned in the Discussion.

Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \boldsymbol{K}$ be the complete spectral decomposition of $\boldsymbol{K}$. Thus, $\boldsymbol{\Lambda}$ is an N-by-N diagonal matrix containing the k non-zero eigenvalues on the top-left of the diagonal, followed by $N - k$ zeros on the bottom-right, and $\boldsymbol{U}$ is an $N \times N$ matrix of eigenvectors. Now, write the N-by-N orthonormal matrix $\boldsymbol{U}$ as $\boldsymbol{U} = [\boldsymbol{U}_1, \boldsymbol{U}_2]$, where $\boldsymbol{U}_1 \in \mathbb{R}^{n \times k}$ contains the eigenvectors corresponding to non-zero eigenvalues, and $\boldsymbol{U}_2 \in \mathbb{R}^{N \times N-k}$ contains the eigenvectors corresponding to zero eigenvalues. Thus, we have

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = [\boldsymbol{U}_1, \boldsymbol{U}_2]\begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 \end{bmatrix}[\boldsymbol{U}_1, \boldsymbol{U}_2]^\top = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2^\top.$$

As $\boldsymbol{\Lambda}_2 = [\boldsymbol{0}]$, $\boldsymbol{K}$ can be recovered by $\boldsymbol{K} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top$, the economy-spectral decomposition of $\boldsymbol{K}$, so-called because it contains only eigenvectors corresponding to k non-zero

eigenvalues and arises from taking the spectral decomposition of a matrix of rank k. The expression $(\gamma \boldsymbol{K} + \boldsymbol{I})$ appearing in the LMM likelihood, however, is always of full rank, as it is the sum of the positive semi-definite matrix $\gamma \boldsymbol{K}$ and the positive definite matrix $\boldsymbol{I}$:

$$\gamma \boldsymbol{K} + \boldsymbol{I} = \boldsymbol{U} (\gamma \boldsymbol{\Lambda} + \boldsymbol{I}) \boldsymbol{U}^\top = \boldsymbol{U} \begin{bmatrix} \gamma \boldsymbol{\Lambda}_1 + \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \boldsymbol{U}^\top.$$

Therefore, it is not possible to simply ignore $\boldsymbol{U}_2$ while using the FaST-LMM in section 3.3, as it enters the expression for the log likelihood. However, directly computing the complete spectral decomposition does not exploit the low rank of $\boldsymbol{K}$.

3.4.3. Linear time evaluation of the likelihood

To exploit the low rank of $\boldsymbol{K}$ to evaluate the log likelihood efficiently, one possible approach would be to augment the spectrum using $N - k$ vectors that are orthogonal to the first k. Unfortunately, this strategy has a time complexity of $O((N-k)N^2)$. Consequently, we take the following alternative approach.

We begin with Equation (2.25):

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \gamma\right) = -\frac{N}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} \log |\gamma \boldsymbol{K} + \boldsymbol{I}| - \frac{1}{2\sigma^2} R,$$

where the quadratic form is given by

$$R = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\gamma \boldsymbol{K} + \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The two terms involving $\gamma \boldsymbol{K} + \boldsymbol{I}$ will be treated separately in the following.

Efficient evaluation of the log determinant

As in Equation (3.9), the log-determinant of the genetic similarity matrix can be efficiently computed using the economy spectral decomposition of $\boldsymbol{K}$:

$$\log |\gamma \boldsymbol{K} + \boldsymbol{I}| = \sum_{n=1}^{N} \log\left(\frac{1}{\gamma \lambda_n + 1}\right).$$

As the last $N - k$ eigenvalues equal zero, the last $N - k$ terms in the sum are equal to zero.

$$\log |\gamma \boldsymbol{K} + \boldsymbol{I}| = \sum_{N=1}^{k} \log\left(\frac{1}{\gamma \lambda_n + 1}\right). \tag{3.21}$$

Efficient evaluation of the quadratic form

Also, as we show in Section D.1, the residual quadratic form R can be evaluated using the low-rank decomposition:

$$\underbrace{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\gamma \boldsymbol{K} + \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}_{R} = R_k + R_{N-k},$$

where R_k is a quadratic form on data transformed by the first k eigenvectors of the genetic similarity.

$$R_k = \left(\boldsymbol{U}_1^\top \boldsymbol{y} - \boldsymbol{U}_1^\top \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\gamma \boldsymbol{\Lambda}_1 + \boldsymbol{I}_k\right)^{-1} \left(\boldsymbol{U}_1^\top \boldsymbol{y} - \boldsymbol{U}_1^\top \boldsymbol{X}\boldsymbol{\beta}\right). \tag{3.22}$$

Further, R_{N-k} is a quadratic form computed on the residuals obtained from regressing out the first k eigenvectors from the data.

$$R_{N-k} = \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1\boldsymbol{U}_1^\top\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right)^\top \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1\boldsymbol{U}_1^\top\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right). \tag{3.23}$$

Furthermore, both expressions can be written as sums.

$$R = \underbrace{\sum_{n=1}^{k} \frac{\left(\left[\boldsymbol{U}_1^\top \boldsymbol{y}\right]_n - \left[\boldsymbol{U}_1^\top \boldsymbol{X}\right]_{n,:}\boldsymbol{\beta}\right)^2}{\frac{1}{\gamma\lambda_n + 1}}}_{R_k} + \underbrace{\sum_{n=1}^{N} \left(\left[\boldsymbol{y} - \boldsymbol{U}_1\left(\boldsymbol{U}_1^\top \boldsymbol{y}\right)\right]_n - \left[\boldsymbol{X} - \boldsymbol{U}_1\left(\boldsymbol{U}_1^\top \boldsymbol{X}\right)\right]_n \boldsymbol{\beta}\right)^2}_{R_{N-k}}. \tag{3.24}$$

Finding the maximum likelihood and parameters efficiently

Plugging both the determinant (Equation (3.21)) and the quadratic form (Equation (3.24)) into the log likelihood, we obtain

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \gamma, \sigma^2\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) + \sum_{n=1}^{k} \log\left(\frac{1}{\gamma\lambda_n + 1}\right) - \frac{1}{2\sigma^2}\left(R_k + R_{N-k}\right). \tag{3.25}$$

Setting the gradient of $\log \mathcal{L}\left(\boldsymbol{\beta}, \gamma, \sigma^2\right)$ in Equation (3.25) with respect to $\boldsymbol{\beta}$ to zero, we obtain

$$\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{c}_{\boldsymbol{X},\boldsymbol{y}}, \tag{3.26}$$

where the D-by-D matrix $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}}$ equals

$$\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}} = \sum_{n=1}^{k} \frac{\left[\boldsymbol{U}_1^\top \boldsymbol{X}\right]_{n,:}^\top \left[\boldsymbol{U}_1^\top \boldsymbol{X}\right]_{n,:}}{\frac{1}{\gamma\lambda_n + 1}} + \sum_{n=1}^{N} \left[\left(\boldsymbol{I} - \boldsymbol{U}_1\boldsymbol{U}_1^\top\right)\boldsymbol{X}\right]_{n,:}^\top \left[\left(\boldsymbol{I} - \boldsymbol{U}_1\boldsymbol{U}_1^\top\right)\boldsymbol{X}\right]_{n,:}, \tag{3.27}$$

and the D-by-1 vector $\boldsymbol{c}_{\boldsymbol{X},\boldsymbol{y}}$ equals

$$\boldsymbol{c}_{\boldsymbol{X},\boldsymbol{y}} = \sum_{n=1}^{k} \frac{\left[\boldsymbol{U}_1^\top \boldsymbol{X}\right]_{n,:}^\top \left[\boldsymbol{U}_1^\top \boldsymbol{y}\right]_n}{\frac{1}{\gamma\lambda_n + 1}} + \sum_{n=1}^{N} \left[\left(\boldsymbol{I} - \boldsymbol{U}_1\boldsymbol{U}_1^\top\right)\boldsymbol{X}\right]_{n,:}^\top \left[\left(\boldsymbol{I} - \boldsymbol{U}_1\boldsymbol{U}_1^\top\right)\boldsymbol{y}\right]_n$$

Plugging $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ into the log likelihood and setting the derivative with respect to σ^2 to zero, we get

$$-\frac{1}{2}\left(\frac{N}{{\sigma^2}_{\mathrm{M}_\gamma}}\right)-\frac{1}{{\sigma^4}_{\mathrm{M}_\gamma}}\left(R_k(\boldsymbol{\beta}_{\mathrm{M}_\gamma})+R_{N-k}(\boldsymbol{\beta}_{\mathrm{M}_\gamma})\right)=0,$$

where we made the dependence of $R_k(\boldsymbol{\beta}_{\mathrm{M}_\gamma})$ and $R_{N-k}(\boldsymbol{\beta}_{\mathrm{M}_\gamma})$ on the maimum likelihood estimator of the weights $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ explicit. Consequently, the maximum likelihood estimator is

$${\sigma^2}_{\mathrm{M}_\gamma}=\frac{1}{N}\left(R_k(\boldsymbol{\beta}_{\mathrm{M}_\gamma})+R_{N-k}(\boldsymbol{\beta}_{\mathrm{M}_\gamma})\right). \tag{3.28}$$

Plugging Equations (3.26) and (3.28) into (3.25) yields an expression for the logarithm of the likelihood profiled for the fixed effects and the environmental noise variance σ^2. $\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{M}_\gamma},\gamma,{\sigma^2}_{\mathrm{M}_\gamma}\right)$, which can be evaluated in $O(N+k)$, as

$$-\frac{N}{2}\log(2\pi)-\frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{1}{\gamma\lambda_n+1}\right)-\frac{N}{2}-\frac{N}{2}\log\frac{R_k(\boldsymbol{\beta}_{\mathrm{M}_\gamma})+R_{N-k}(\boldsymbol{\beta}_{\mathrm{M}_\gamma})}{N}. \tag{3.29}$$

3.4.4. Restricted maximum likelihood

Here we extend the derivations so far to restricted maximum likelihood, in a similar fashion as was done in Section 3.3.2 for full rank genetic similarities. We start with the form of the log restricted likelihood from Equation (2.51) that equals the log likelihood with the restricted maximum likelihood estimator of the fixed effects $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ plugged in, plus three additional terms [Kang et al., 2008, Harville, 1974]:

$$\log\mathcal{L}\left(\gamma,\sigma^2|\boldsymbol{Sy}\right)=\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{R}_\gamma},\gamma,\sigma^2\right)+\frac{1}{2}\left(D\log\left(2\pi\sigma^2\right)+\log\left|\boldsymbol{X}^\top\boldsymbol{X}\right|-\log\left|\boldsymbol{X}^\top\left(\gamma\boldsymbol{K}+\boldsymbol{I}\right)^{-1}\boldsymbol{X}\right|\right),$$

where the restricted maximum likelihood estimator of the fixed effects $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ equals the form of the maximum likelihood estimator $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$.

Neglecting the cubic dependence on D for computing the determinants, these additional terms can be evaluated in time complexity $O(N+k)$, using the economy spectral decomposition $\boldsymbol{K}=\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top$. For this purpose, we re-use the results from Section D.1, substituting $\boldsymbol{X}$ for $\boldsymbol{a}$, to get

$$\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{R}_\gamma},\gamma,\sigma^2\right)=\log\mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{R}_\gamma},\gamma,\sigma^2\right)+\frac{D}{2}\log\left(2\pi\sigma^2\right)+\frac{1}{2}\log\left|\boldsymbol{X}^\top\boldsymbol{X}\right|-\frac{1}{2}\log\left|\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}}\right|, \tag{3.30}$$

where $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}}$ is given by Equation (3.27).

The restricted maximum likelihood (*REML*) variance component estimate is given by

$${\sigma^2}_{\mathrm{R}_\gamma}=\frac{1}{N-D}\left(R_k+R_{N-k}\right), \tag{3.31}$$

where R_k is given in Equation (3.22) and R_{N-k} in Equation (3.23). The formulas for the remaining parameters remain unchanged.

Time and space complexity

Given the economy spectral decomposition of $\boldsymbol{K}$, the likelihood of the model can be evaluated in a time complexity of $O(NSk)$ for the required rotations and $O(C(N+k)S) = O(CNS)$ for the C evaluations of the log likelihood during the one-dimensional optimization over γ. By keeping γ fixed to its value from the null model, as in EMMAX/P3D, $O(C(N+k)S)$ can be reduced to $O(C(N+k))$. In general, as discussed, the economy spectral decomposition can be computed from $k = S_\mathrm{c}$ SNPs by first computing the genetic similarity matrix with a time complexity of $O(N^2 S_\mathrm{c})$ and a space complexity of $O(N^2)$, and then finding its first k eigenvalues and eigenvectors with a time complexity of $O(N^2 k)$. When the realized relationship matrix is used, however, we can perform the economy spectral decomposition more efficiently by circumventing the computation of $\boldsymbol{K}$, because the singular vectors of the data matrix are the same as those of the realized relationship matrix constructed from that data (e. g. [Berrar et al., 2003]). In particular, we can obtain the economy spectral decomposition of $\boldsymbol{K}$ from the economy singular value decomposition of the $N \times S_\mathrm{c}$ SNP matrix directly, which is an operation with a time complexity of $O(NS_\mathrm{c}k)$ and requires space $O(NS_\mathrm{c})$. For testing S variants the total asymptotic runtime follows as $O(NS_\mathrm{c}S)$

However, we note that, for both the normal and low-rank versions of FaST-LMM, the rotations and the search for γ for each test are easily parallelized. Consequently, the runtime of the LMM analysis is dominated by the spectral decomposition (or singular value decomposition for the low-rank version). Although parallel algorithms for singular-value decomposition exist, improvements to such algorithms should lead to even greater speedups.

3.4.5. Compressed FaST-LMM

The ideas behind FaST-LMM can be applied to compressed linear mixed models to improve their computational efficiency. Here, we demonstrate this application for a compressed linear mixed model similar to the one in Section 3.2.3 [Zhang et al., 2010].

In the spirit of FaST-LMM, we look for an efficient way of computing the spectral decomposition of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$. This spectral decomposition can then be plugged into Formulas (3.25)–(3.29) as a means to evaluate Equation (3.6), in runtime and memory that are linear in the number of individuals N. In Section D.2, we consider the case where genetic similarity is defined by an realized relationship matrix. We show that, given a $G \times S_\mathrm{c}$ matrix $\boldsymbol{G}$ of S_c SNPs, obtained by averaging the SNP data for individuals over the members of each group, the economy spectral decomposition of the realized relationship matrix $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{Z}^\top$ can be computed from the singular value decomposition of the $G \times S_\mathrm{c}$ matrix $(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}\boldsymbol{G}$ in $O(\min(G, S_\mathrm{c})GS_\mathrm{c})$ time and $O(GS_\mathrm{c})$ memory. (It is easy to verify that the same $\boldsymbol{G}\boldsymbol{G}^\top$ would be obtained if instead we used a group-wise average of the $N \times N$ realized relationship matrix.) In Section D.2.2, we consider arbitrary genetic similarity. We prove that, given any $G \times G$ positive semi-definite group similarity matrix $\boldsymbol{K}$, the spectral decomposition of the $N \times N$ matrix $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$ can be computed from the spectral decomposition of the much smaller $G \times G$ matrix $(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}\boldsymbol{K}(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}$ using $O(G^3)$ time and $O(G^2)$ memory.

3.5. Experiments

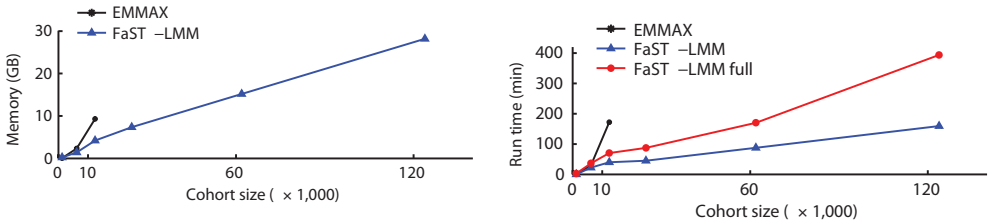
As we have just discussed, FaST-LMM reduces the computational effort from cubic to linear in the number of samples, when the realized relationship matrix is used as the measure of genetic similarity between individuals and when the number of SNPs used to estimate these similarities is substantially less than the number of individuals in the data. We explored these conditions using two publicly available, real data sets: the Genetic Analysis Workshop (GAW) 14 for smoking (see Section A.2) [Edenberg et al., 2005] and the Wellcome Trust Case Control Consortium (WTCCC) 1 data for seven common diseases (see Section A.1) [Burton et al., 2007], of which the latter has been previously analyzed using linear mixed models [Kang et al., 2010]. In contrast to previously published analyses of WTCCC, we included non-white individuals and close family members so as to produce a dataset with greater potential confounding structure—structure that LMMs have been shown to be able to handle well [Price et al., 2010b].

We obtained P values from our linear mixed model analyses using a likelihood ratio test (see Section 2.2.5). The calibration of P values was assessed, in part, using the λ statistic (see Section 2.3.1). In addition to this summary statistic, we assessed differences in two P value distributions using a two-sample Kolmogorov-Smirnov test on P values near the level of genome-wide significance (5×10^{-7} as in Burton et al. [2007]).

Proximal contamination While investigating the benefits of FaST-LMM, we encountered a phenomenon that substantially affected our evaluation. In particular, we found that λ was consistently lower when the genetic similarity matrix was constructed from the same SNPs tested for association than when the genetic similarity matrix was constructed from SNPs not tested for association. In order to avoid this effect caused by linkage between the marker being tested and the markers used for estimating similarity, when testing a given chromosome we estimated genetic similarity always from all but this chromosome (see Section 4.1 for an in-depth analysis of the effect of proximal contamination.)

3.5.1. Comparison of computational cost

We compared memory footprint and run time for non-parallelized implementations of the FaST-LMM and EMMAX algorithms (Figure 3.1). (The EMMAX implementation was no less efficient in terms of run time and memory use than that of P3D in the trait analysis by association, evolution and linkage (TASSEL) package). In the comparison, we used Genetic Analysis Workshop 14 data to construct synthetic datasets with the same number of SNPs (8,000 SNPs) and roughly 1, 5, 10, 20, 50 and 100 times the cohort size of the original data (see Section A.3). The largest such dataset contained data for 123,800 individuals. We tested all SNPs and used them all to estimate genetic similarity. EMMAX would not run on the 20×, 50× or 100× datasets because the memory required to store the large matrices exceeded the 32 gigabytes available. In contrast, FaST-LMM, which did not require these matrices (because it bypassed their computation, using them only implicitly), completed the analyses using 28 gigabytes of memory on the largest dataset. Runtime results highlight the linear dependence of



(a) Memory footprint of FaST-LMM and EMMAX.

(b) Runtime of FaST-LMM and EMMAX.

Figure 3.1. Computational costs of FaST-LMM and EMMAX. Memory footprint (a) and run time (b) of the algorithms running on a single processor as a function of the cohort size in synthetic datasets based on GAW14 data. In each run, we used 7,579 SNPs both to estimate genetic similarity (realized relationship matrix for FaST-LMM and identity by state for EMMAX) and to test for association. In the FaST-LMM full analysis, the variance parameters were re-estimated for each test, and in the FaST-LMM analysis these parameters were estimated only once for the null model, as in EMMAX. FaST-LMM and FaST-LMM full had the same memory footprint. EMMAX would not run on the datasets that contained 20 or more times the cohort size of the GAW14 data because the memory required to store the large matrices exceeded the 32 GB available.

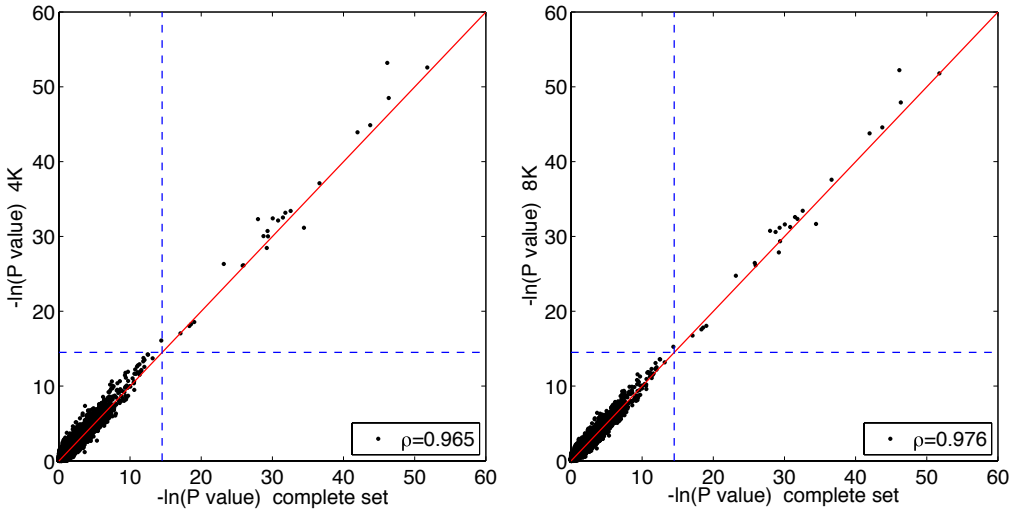
the computations on the cohort size when that size exceeded the 8,000 SNPs used to construct the realized relationship matrix. Also, computations remained practical using our approach even when we re-estimated the variance parameters for each test.

3.5.2. Assessing the accuracy of SNP sampling

As shown in Section 2.2.1, the linear mixed model with no fixed effects using a realized relationship matrix constructed from a set of SNPs is equivalent to a linear regression of the SNPs on the phenotype, with weights integrated over independent normal distributions with the same variance. In this view, sampling SNPs for construction of the realized relationship matrix can be seen as the omission of regressors and hence an approximation. Nonetheless, SNPs could be sampled uniformly across the genome so that linkage disequilibrium would diminish the effects of sampling.

To examine this issue, we compared association P values with and without sampling on the Wellcome Trust Case Control Consortium (WTCCC) data for Crohn's disease.

Specifically, we tested all SNPs on chromosome 1 while, in order to avoid proximal contamination (see Section 4.1), using SNP sets of various sizes from all but this chromosome (the complete set (340,000 SNPs) and uniformly distributed samples of 8,000 SNPs and 4,000 SNPs) to compute the realized relationship matrix. The 4,000 and 8,000 SNP sets were created by including every forty-eighth and every twenty-forth SNP, respectively, along each chromosome. The P values resulting from the complete and sampled sets were similar (see Figure 3.2). The different SNP sets led to nearly identical calls of significance, using the genome-wide significance threshold of 5×10^7. When we used the complete set, the algorithm called 24 SNPs significant, and the 8,000-SNP and 4,000-SNP analyses labeled only one additional SNP significant and missed none. By comparison, the Armitage trend test labeled seven additional SNPs significant and



(a) Correlation 4,000-SNP sample. **(b)** Correlation 8,000-SNP sample.

Figure 3.2. Accuracy of association P values resulting from SNP sampling on WTCCC data for the Crohn's disease phenotype. Each point in the plot shows the negative ln P values of association for a particular SNP from a linear mixed model using (a) 4,000-SNP and (b) 8,000-SNP samples and all SNPs to compute the realized relationship matrix. The complete set used all 340,000 SNPs from all but chromosome 1, whereas the 4,000-SNP and 8,000-SNP samples used equally spaced SNPs from these chromosomes. All 28,000 SNPs in chromosome 1 were tested. Dashed lines show the genome-wide significance threshold (5×10^{-7}). The correlation ρ for the points in the plots are 0.965 for 4,000 SNPs and 0.976 for 8,000 SNPs.

missed none. Furthermore, the λ statistic was similar for the complete, 8,000-SNP and 4,000- SNP analyses (1.132, 1.173 and 1.203, respectively) in contrast to $\lambda = 1.333$ for the ATT. We show corresponding quantile-quantile (Q-Q) plots in Figure 3.3. Finally, using these SNP samples to construct genetic similarity, FaST-LMM ran an order of magnitude faster than EMMAX: 23 min and 53 min for the 4,000-SNP and 8,000-SNP FaST-LMM analyses compared with 260 min and 290 min for the respective EMMAX analyses.

3.5.3. Materials and Methods

All analyses assumed an additive effect of SNP on phenotype. To normalize the SNP data and impute missing SNPs, we used the approach reported in [Price et al., 2006]. Runtimes were measured on a dual AMD six core Opteron machine with a 2.6GHz clock and 32GB of RAM. We restricted computations to a single core. FaST-LMM used the AMD Core Math Library.

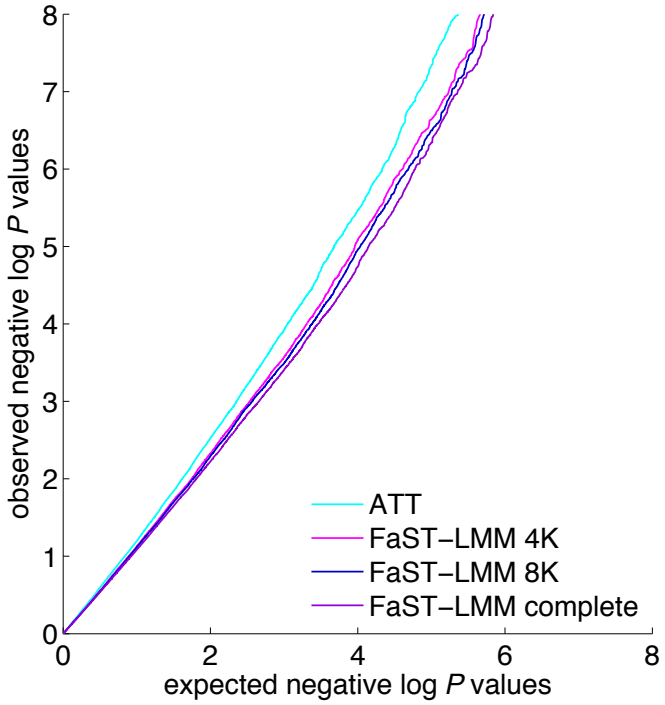


Figure 3.3. Q-Q plot comparison for FaST-LMM analyses of the WTCCC data. Shown are observed versus expected negative log P values for the association analyses on the Chrohn's disease phenotype. We used FaST-LMM to test all SNPs on chromosome 1, and SNP sets of various sizes from all but this chromosome—the complete set (340K), 8K, and 4K—to compute the realized relationship matrix. We also used the Armitage trend test to compute P values.

3.6. Chapter summary and discussion

In this chapter we have introduced and demonstrated FaST-LMM, a new, computationally efficient approach to linear mixed models for the analysis of genome-wide association studies.

Our new approach is linear in the number of individuals. Moreover, when parallelized, the slowest computation in our approach is linear in the number of individuals and quadratic in the number of SNPs used to estimate genetic similarity. The dramatic speedups are realized provided (1) the number of SNPs used to estimate the genetic similarities between individuals is substantially less than the number of individuals in the dataset and (2) these genetic similarities are determined with the realized relationship matrix. Using two real data sets, we showed that the realized relationship matrix computed from only thousands of linearly spaced SNPs provides a good measure of genetic similarity. Consequently, FaST-LMM breaks the current computational barrier of linear mixed model application to data sets with tens of thousands or more individuals.

Furthermore, even when we use all available SNPs, use of any kind of genetic similarity, and re-learn the variance parameters for each SNP tested, our approach requires only a single cubic operation in the number of individuals rather than one cubic operation per SNP tested, as is required by EMMA [Kang et al., 2008].

While random selection of markers may be seen as an approximation to using all available SNPs for correction, in the next chapter we are going to demonstrate how careful selection of such markers can be used to improve correction in the linear mixed model.

While investigating the benefits of FaST-LMM, we observed that λ was consistently lower when the genetic similarity matrix was constructed from the same SNPs tested for association than when the genetic similarity matrix was constructed from SNPs not tested for association. Here we avoided this problem by leaving out the whole chromosome being tested from computation of genetic similarities, but this approach ignores confounding due to linked causal markers on that chromosome. A more thorough exploration of the problem and an efficient algorithm for excluding markers only in a linked region is presented in Section 4.1.

Another important computational saving that results from our approach is that once the expensive upfront computation needed for univariate SNP tests has been performed (i. e. computation and application of the rotation matrix to individual SNPs), joint association tests of any number of SNPs can be achieved just as efficiently as univariate tests (i. e. linear in the number of individuals). In Chapter 5 we are going to propose two variants of such linear-additive mixed models involving multiple markers, one proposes association tests of pre-defined sets of markers, like SNPs in a gene or pathway (Section 5.1), and one that allows to automatically detect groups of relevant markers from a genome-wide panel using shrinkage estimation (Section 5.2).

4. Modeling phenotype-specific relatedness by selection of genetic markers

Linear mixed models tackle confounding by using a matrix of pairwise genetic similarities to model the relatedness among subjects. Until now, the consensus has been that, ideally, all available SNPs should be used in the determination of these similarities. As demonstrated in Section 3.5, thousands of linearly-spaced SNPs could be used to achieve good performance in practice with a greatly reduced computational burden [Lippert et al., 2011]. Initially, this approach was proposed as an approximation in order to scale linear mixed models to larger data sets. In this chapter, however, we argue that a small number of carefully selected SNPs should be used for determining genetic relatedness. Based on these insights, we propose the FaST-LMM-Select algorithm, that comprises two methods presented in this chapter. On real and synthetic data we demonstrate, that FaST-LMM-Select achieves systematically increased power (joint reduction of false positives and false negatives), improved calibration (the avoidance of inflation or deflation of the test statistic), and a lower computational cost compared to the traditional use of linear mixed models.

When using realized relationships the genetic relatedness at unknown causal variants is approximated by genome-wide markers utilizing linkage to the causal variants (see Section 2.2.1). We noted that the linear mixed model using genetic similarities constructed from a set of SNPs is mathematically equivalent to linear regression of the SNPs on the phenotype, where the weights of the SNPs used to determine the genetic similarities have been integrated over independent Normal distributions having the same variance. That is, a linear mixed model using a given set of SNPs for genetic similarity is equivalent to linear regression using those SNPs as random covariates to correct for confounding. It follows that by using a linear mixed model to test a given SNP for an association with the phenotype, we are in effect adjusting for the effect of *background SNPs* (see Figure 2.1(a)).

In the light of this equivalence, it becomes evident that the common approach of including all SNPs in the computation of genetic similarities has two potential flaws: *proximal contamination* by inclusion of nearby SNPs and *dilution* by inclusion of SNPs that are unrelated SNPs.

The first of these problems means that whenever a SNP is tested for association, all SNPs that lie in close proximity to this SNP should be excluded from computation of genetic similarities. This becomes evident from the fact, that inclusion of such SNPs is equivalent to using these SNPs as covariates in a linear regression model. As genetic linkage causes strong correlation among SNPs nearby, the induced null model that only includes such covariates should have a reduced potential of improving its model fit by inclusion of the SNP tested and thus smaller likelihood ratios. While inclusion of SNPs that are correlated due to confounding is used to stratify for such confounding, correlation

due to physical linkage is expected to cause a loss in power to detect associations. We call this effect *proximal contamination*. A demonstration of the deflation due to proximal contamination is given in Section 4.1.

An ideal procedure that avoids proximal contamination requires exclusion of all SNPs in a window of linkage disequilibrium to the SNP tested from computation of the genetic similarity matrix. Such an approach involves many thousand different sets of SNPs used to estimate genetic similarity. If done naively, even using the efficient algorithms presented in Sections 3.3 and 3.4, this would require the computation of a new spectral decomposition and subsequent transformation of the data each time a new set is considered. As these computations that dominate the runtime of mixed model evaluation, such a naive procedure is infeasible in practice.

In Section 4.1.3 we present an efficient algorithm to correct for proximal contamination, that avoids computation of new spectral decompositions. We prove that by computing of corrective terms, that take the implied difference in genetic similarities between the complete (contaminated) genetic similarity matrix a genetic similarity matrix that avoids proximal contamination into account, the algorithm only requires the spectral decomposition of the full matrix and yields exactly the same result as the naive approach. Asymptotically, the runtime of the algorithm is identical to the runtime, when proximal contamination is not taken into account.

The second problem is due to the observation, that in a linear regression model one would ideally include only such SNPs that are either related to the phenotype or tag a latent confounding variable. If on the other hand, SNPs that are unrelated to the phenotype were included, these only add random noise to the model and for this reason could result both in a loss of power as well as worse stratification for true confounding. We call this effect *dilution*. In Section 4.2, we propose a heuristic to carefully select the SNPs used to determine genetic similarities in order to combat dilution, that is easy to use and works well in practice.

Together with the efficient algorithm to avoid proximal contamination, this comprises our new method, FaST-LMM-Select, an algorithm for genome-wide association studies that avoids systematic loss of power and inaccurate test statistics arising in the traditional use of linear mixed models.

In Section 4.3, we demonstrate empirically, that by avoiding proximal contamination and dilution FaST-LMM select not only improves on the traditional use of linear mixed models in terms of power and P value calibration, but in combination with the algorithm presented in Section 3.4 also yields large computational savings as typically only a few hundred SNPs are selected to determine genetic similarity.

4.1. Proximal contamination

As stated in Section 3.5, while investigating the benefits of FaST-LMM (see Chapter 3), we encountered a phenomenon that substantially affected our evaluation. In particular, we found that λ was consistently lower when the genetic similarity matrix was constructed from the same SNPs tested for association (*in-sample matrix*) than when the genetic similarity matrix was constructed from SNPs not tested for association (*out-of-sample matrix*). A likely explanation for this effect is that the analysis with the out-of-sample

matrix is the correct approach and that the analysis with the in-sample matrix is deflated with respect to it. In particular, if there is a true association between a SNP and the phenotype (or a spurious one due to residual confounding), then if that SNP is included in the construction of the genetic similarity matrix, the random effects may predict the phenotype too well. Consequently, the log likelihood of the null model in the likelihood ratio test may be too high, leading to a P value that is too low. We refer to this effect as *proximal contamination*. A similar theoretical concern has been made about stratification by the use of principal components constructed from genome-wide SNPs, but the authors did not find empirical evidence for the problem [Price et al., 2006].

An alternative explanation for this phenomenon is that P values obtained using the out-of-sample matrix are *inflated* with respect to that using the in-sample matrix due to *local* confounding structure not captured by the linear mixed model, wherein genetic similarities determined from SNPs for one chromosome do not adequately capture local confounding structure in other chromosomes. Evidence against this hypothesis and for null-model contamination, however, is that when SNPs having a large apparent association with the phenotype were removed from the set used to construct the genetic similarity matrix, the values of λ for in-sample analyses increased. Experiments described in Section 4.1.1 and Section 4.1.2 also suggest that the alternative explanation is unlikely.

4.1.1. Testing proximal contamination on real data

In order to check for apparent null model contamination on the data from the Wellcome Trust Case Control Consortium (See Section A.1), we partitioned the SNPs by even and odd chromosome numbers. The X chromosome was included in the even group. The resulting partitions had 184,559 and 176,098 SNPs in the even and odd chromosomes, respectively. using either the same (in-sample) or different (out-of-sample) partitions for testing and construction of the genetic similarity matrix. With WTCCC, we computed P values using a genetic similarity matrix constructed from SNPs on even chromosomes and then tested SNPs on odd chromosomes only. The in-sample analyses on several of the phenotypes yielded substantially lower values for λ (see Table 4.1).

With GAW14 [Edenberg et al., 2005] (see Section A.2), when we performed an out-of-sample analysis, wherein we computed P values using a genetic similarity matrix constructed from the 3,769 SNPs on even chromosomes and then tested on the 3,810 SNPs on odd chromosomes, and vice-versa, λ for the combined distribution of P values was 1.005. In contrast, when we performed an in-sample analysis, wherein we computed P values using a matrix constructed from SNPs on even (odd) chromosomes and testing SNPs on even (odd) chromosomes, λ was 0.951.

4.1.2. Proximal contamination by distance to the SNP tested

In this section, we show that, on the WTCCC data for the Chrohn's disease phenotype, this effect produces substantially deflated P values as measured by the λ statistic, and quantify the degree to which linkage disequilibrium plays a role.

We used an approach where the SNPs used to construct the realized relationship matrices were chosen to be systematically further and further away from a set of test SNPs,

genetic relatedness	bd	cad	ht	Cd	ra	t1d	t2d
Out-of-sample IBS	1.156	1.083	1.072	1.12	1.069	1.071	1.105
Out-of-sample RRM	1.151	1.091	1.069	1.111	1.076	1.071	1.109
In-sample RRM	0.987	0.798	0.982	0.962	0.953	0.98	0.991
Uncorrected	1.185	1.099	1.103	1.304	1.087	1.081	1.13

Table 4.1. Effect of proximal contamination on genomic control λ in WTCCC data. "Out-of-sample IBS" refers to use of a linear mixed model with an identity by state matrix computed from genome-wide SNP markers from even chromosomes when testing odd chromosomes and vice versa. "Out-of-sample RRM" refers to use of a linear mixed model with an RRM computed from genome-wide SNP markers from even chromosomes when testing odd chromosomes and vice versa. "In-sample RRM" refers to use of a linear mixed model using an RRM computed from genome-wide SNP markers from even chromosomes when testing even chromosomes and from odd chromosomes when testing odd chromosomes. "Uncorrected" refers to an uncorrected analysis using the Armitage trend test.

while holding the number of SNPs used to construct the realized relationship matrices (i.e., the number of background SNPs in the equivalent linear regression) constant. In particular, after ordering SNPs by their position, we used every thirty-second SNP starting from the i^{th} SNP in each chromosome to form a set of test SNPs. In addition, we created six sets of SNPs to construct realized relationship matrices, each set lying further away from the set of test SNPs. In a given set, we included every thirty-second SNP starting at the $i+j^{\text{th}}$ SNP in each chromosome, $j = 0, 1, 2, 4, 8$, and 16. This experiment was performed for $i = 1, 2, 3, 4$, and 5. Each set of SNPs contained approximately 11K SNPs. As shown in Figure 4.1, λ generally increased with j for $j \leq 8$, (i.e. distance between SNPs tested and SNPs used to estimate genetic similarity), beyond which linkage disequilibrium presumably had little effect. Note that the values for λ for the experiments having the greatest amount of proximal contamination ($j = 0$) were quite similar to those when all 367K SNPs were used to construct the realized relationship matrix (differences were less than 0.027 over all values of i), suggesting that our experiment did not deviate substantially from the idealized one.

These experiments show that null-model contamination can be a substantial effect. Consequently, when using a linear mixed model to test whether a given SNP is associated with the phenotype, the realized relationship matrix should be computed from all SNPs except for those in close proximity to the test SNP.

4.1.3. Efficient algorithm to avoid proximal contamination

As discussed, when using a linear mixed model to test for the association between a given SNP and phenotype, the SNPs used to construct the realized relationship matrix should exclude that test SNP and those that lie in close proximity to it. A naive approach to this problem would involve a new spectral decomposition each time some SNPs were removed or added back in to the computation for the matrix. As it is this spectral decomposition that is the computational bottleneck of linear mixed model analysis, such an approach would not be feasible for testing for association on a genome-wide scale [Lippert et al.,

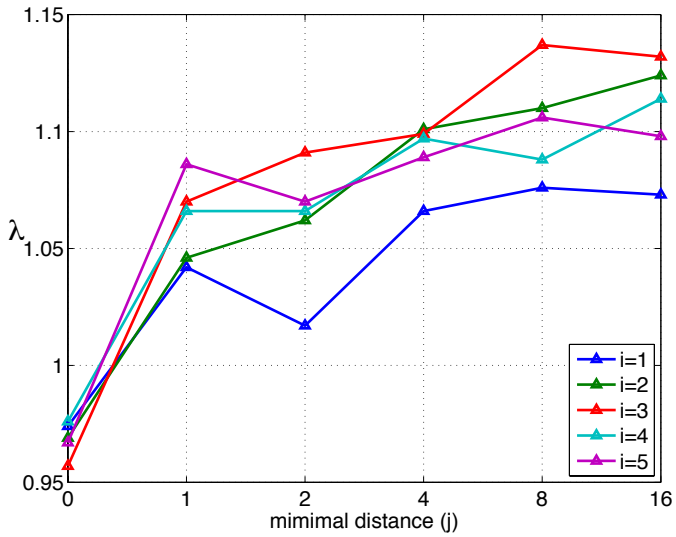


Figure 4.1. Strength of proximal contamination as a function of distance between markers used to compute genetic similarities and markers tested. The λ statistic as a function of the minimum distance between a SNP in the test set and a SNP used to construct the realized relationship matrix. For a given curve, the set of test SNPs was selected by incorporating every thirty-second SNP along each chromosome starting at position i. λ increases with distance between the sets of tested markers and markers used to compute genetic similarities.

2011]. Here we present an algorithm that enables us to use just a single spectral decomposition, and then cheaply add corrective terms into the log-likelihood to exactly account for having used the spectral decomposition of the uncorrected realized relationship matrix. We prove that result is the same as though we had actually computed the spectral decomposition of the corrected matrices for each test. We obtain an efficient algorithm for performing our desired association analysis that has identical asymptotic runtime and memory footprint as the uncorrected version of FaST-LMM presented in Sections 3.3 and 3.4.

The algorithm uses the property that the realized relationship matrix, given by $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{G}^\top$, where $\boldsymbol{G}$ denotes the matrix of SNP data to be used in the RRM and is of dimension $N \times S_\mathrm{C}$ (for N individuals), decomposes into a sum of contributions from S_C single SNPs.

$$\boldsymbol{G}\boldsymbol{G}^\top = \sum_{s=1}^{S_\mathrm{C}} [\boldsymbol{G}]_{:,s} [\boldsymbol{G}]_{:,s}^\top ,$$

where $[\boldsymbol{G}]_{:,s}$ denotes the s-th column of $\boldsymbol{G}$.

It follows that the realized relationship matrix with a subset $\mathcal{A}$ of SNPs removed can be written as the difference between the full realized relationship matrix and the sum over contributions from the SNPs in the set $\mathcal{A}$. With a slight abuse of notation, where

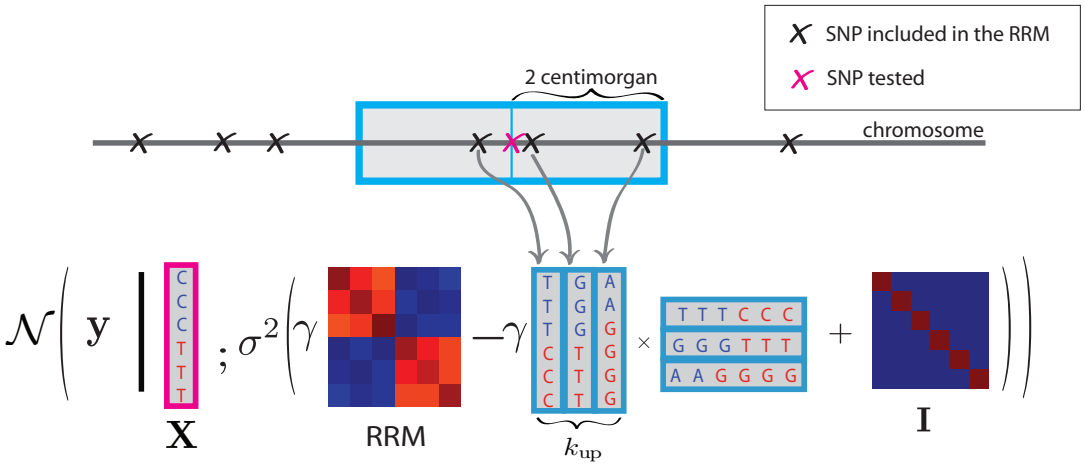


Figure 4.2. Schematic illustration of an efficient algorithm for avoiding proximal contamination. For every SNP tested, we exclude all SNPs in a window (e.g., 2 centimorgans) around that SNP from the realized relationship matrix used in the likelihood calculations, by subtracting the product of the corresponding columns of the SNP matrix used to construct the realized relationship matrix from the covariance term in the linear mixed model likelihood.

$\mathcal{A}$ denotes the set of indices of SNPs in the set $\mathcal{A}$, this difference becomes

$$\boldsymbol{K}' \equiv \boldsymbol{G}\boldsymbol{G}^\top - \underbrace{\sum_{l \in \mathcal{A}} [\boldsymbol{G}]_{:l} [\boldsymbol{G}]_{:l}^\top}_{\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top},$$

where $\tilde{\boldsymbol{G}}$ is the $N \times k_{\text{up}}$ matrix containing the k_{up} SNPs to be removed. In most practical circumstances,k_{up} will be smaller than both the number of individuals N and the number of SNPs in the realized relationship matrix S_{C}, and thus, as will show, it would be wasteful to compute the spectral decomposition of $\boldsymbol{K}$. Instead our algorithm uses the spectral decomposition of the full realized relationship matrix to efficiently evaluate the maximum likelihood function (or alteratively the restricted maximum likelihood function) and treats the removal of SNPs from the realized relationship matrix as low-rank updates at evaluation time. This approach is described in Section 4.1.3 for the case where the RRM is full rank ($S_{\text{C}} \geq N$) and in Section 4.1.3 for the case where the RRM is low rank ($S_{\text{C}} < N$). A schematic overview of our new algorithm is given in Figure 4.2.

Low-rank updates to full-rank genetic similarity matrices

Let $\boldsymbol{G}\boldsymbol{G}^\top$ be a factored genetic similarity matrix, as defined by Equation (3.18). Let $\tilde{\boldsymbol{G}} \in \mathbb{R}^{N \times k_{\text{up}}}$ be a matrix containing a subset of k_{up} columns of $\boldsymbol{G}$. Given the spectral decomposition of $\boldsymbol{G}\boldsymbol{G}^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ we can evaluate the likelihood of a linear mixed model with the updated genetic similarity matrix $\left(\boldsymbol{G}\boldsymbol{G}^\top - \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)$ in $O(N{k_{\text{up}}}^2 + {k_{\text{up}}}^3)$ as follows:

In this section, we treat the case where $\boldsymbol{G}$ is an $N \times S_{\mathrm{C}}$ matrix with $N \leq S_{\mathrm{C}}$, resulting in a full-rank genetic similarity matrix. The log-likelihood can be written as

$$\log \mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma^2\left(\gamma \boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{I} - \gamma \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)\right).$$

Replacing $\boldsymbol{G}\boldsymbol{G}^\top$ by its spectral decomposition $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we get

$$\log \mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma^2\left(\gamma \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \boldsymbol{I} - \gamma \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)\right).$$

In contrast to the approach taken in Chapter 3, rotating the data by the matrix of Eigenvectors $\boldsymbol{U}^\top$ of $\boldsymbol{G}\boldsymbol{G}^\top$ does not yield a diagonal covariance term in the log-likelihood, but rather a full $N \times N$ matrix,

$$\log \mathcal{N}\left(\boldsymbol{U}^\top\boldsymbol{y} \,|\, \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma^2\left(\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\right)\right).$$

When applying the logarithm to the formula of the multivariate Normal distribution, we get

$$-\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\log\left|\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\right| - \frac{1}{2\sigma^2}\tilde{R}, \tag{4.1}$$

where

$$\tilde{R} = \left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}\right)^\top\left(\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\right)^{-1}\left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}\right). \tag{4.2}$$

To evaluate the maximum of this log-likelihood efficiently, we have to solve for the maximum-likelihood parameters, and evaluate the squared form of the Normal distribution and the determinant of the covariance term. In Sections 4.1.3–4.1.3, we provide efficient solutions for each of these steps.

Maximum likelihood parameters

Given γ, the maximum likelihood weight parameters of the log likelihood in Equation (4.1) are given by the generalized least squares estimator

$$\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{c}_{\boldsymbol{X},\boldsymbol{y}}, \tag{4.3}$$

where

$$\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{X}} = \left(\left(\boldsymbol{U}^\top\boldsymbol{X}\right)^\top\left(\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\right)^{-1}\left(\boldsymbol{U}^\top\boldsymbol{X}\right)\right)$$

and

$$\boldsymbol{c}_{\boldsymbol{X},\boldsymbol{y}} = \left(\boldsymbol{U}^\top\boldsymbol{X}\right)^\top\left(\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\right)^{-1}\left(\boldsymbol{U}^\top\boldsymbol{y}\right).$$

Given γ and ${\sigma^2}_{\mathrm{M}_\gamma}$, the maximum likelihood environmental variance parameter is given by

$${\sigma^2}_{\mathrm{M}_\gamma} = \frac{1}{N}\left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right)^\top\left(\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\right)^{-1}\left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{M}_\gamma}\right). \tag{4.4}$$

Both the weight vector $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$ as well as the environmental variance parameter ${\sigma^2}_{\mathrm{M}_\gamma}$ involve quadratic forms of the same form as in the log-likelihood function in Equation (4.3). An efficient solution for these quadratic forms is provided in Section 4.1.3.

Determinant update

To compute the log likelihood of the linear mixed model we need to compute the determinant of the covariance,

$$\log \left| \gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right) \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)^\top \right|.$$

To do so efficiently, we make use of the matrix determinant lemma, $\left| \boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}^\top \right| = |\boldsymbol{A}| \cdot \left| \boldsymbol{I} + \boldsymbol{C}^\top \boldsymbol{A}^{-1} \boldsymbol{B} \right|$. In particular, we plugin $\boldsymbol{A} = (\gamma \boldsymbol{\Lambda} + \boldsymbol{I})$, $\boldsymbol{B} = -\gamma \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)$ and $\boldsymbol{C} = \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)$, yielding

$$\log |\gamma \boldsymbol{\Lambda} + \boldsymbol{I}| \cdot \left| \boldsymbol{I}_{k_{\mathrm{up}}} - \gamma \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)^\top (\gamma \boldsymbol{\Lambda} + \boldsymbol{I})^{-1} \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right) \right|.$$

Finally, applying the logarithm to this expression, we obtain the sum of two log determinants,

$$\log |\gamma \boldsymbol{\Lambda} + \boldsymbol{I}| + \log |\boldsymbol{M}|,$$

where the k_{up}-by-k_{up} matrix $\boldsymbol{M}$ is given by

$$\boldsymbol{M} = \boldsymbol{I}_{k_{\mathrm{up}}} - \gamma \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)^\top (\gamma \boldsymbol{\Lambda} + \boldsymbol{I})^{-1} \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right) \tag{4.5}$$

The log determinant of $(\gamma \boldsymbol{\Lambda} + \boldsymbol{I})$ is merely the sum of the logs of its diagonal entries. The right side is a full $k_{\mathrm{up}} \times k_{\mathrm{up}}$ matrix whose computation has runtime $O(N {k_{\mathrm{up}}}^2)$. Computing its log determinant is an $O({k_{\mathrm{up}}}^3)$ operation, resulting in a runtime of $O(N {k_{\mathrm{up}}}^2 + {k_{\mathrm{up}}}^3)$ to compute the determinant.

Squared form update In all three Equations (4.1), (4.3), and (4.4) needed to evaluate the maximum-likelihood, we must evaluate squared forms such as

$$\boldsymbol{a}^\top \left(\gamma \boldsymbol{\Lambda} + \boldsymbol{I} - \gamma \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right) \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)^\top \right)^{-1} \boldsymbol{b},$$

for different values of $\boldsymbol{a}$ and $\boldsymbol{b}$. We note that the term $\gamma \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right) \left(\boldsymbol{U}^\top \tilde{\boldsymbol{G}} \right)^\top$ is a rank-k_{up} update on the genetic similarity matrix. It follows that we can use the Sherman-Morrison-Woodbury identity (also called the Matrix inversion lemma) to efficiently evaluate these squared forms. The lemma states that

$$(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}\boldsymbol{D}) = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B} \left(\boldsymbol{C}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B} \right)^{-1} \boldsymbol{D}\boldsymbol{A}^{-1}. \tag{4.6}$$

We apply the Sherman-Morrison-Woodbury identity to our case by plugging in $\boldsymbol{A} = (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})$, $\boldsymbol{B} = -\gamma\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)$, $\boldsymbol{C} = \boldsymbol{I}_{k_\mathrm{up}}$ and $\boldsymbol{D} = \left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top$, yielding

$$\boldsymbol{a}^\top(\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{b} + \gamma\boldsymbol{a}^\top(\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\boldsymbol{M}^{-1}\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top(\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{b}. \tag{4.7}$$

The bracketing

$$\boldsymbol{a}^\top(\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{b} + \gamma\left(\left(\boldsymbol{a}^\top(\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\right)\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)\right)\boldsymbol{M}^{-1}\left(\left(\boldsymbol{U}^\top\tilde{\boldsymbol{G}}\right)^\top\left((\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{b}\right)\right)$$

allows for evaluation of these squared forms in $O(N{k_\mathrm{up}}^2 + {k_\mathrm{up}}^3)$.

Updates for low-rank similarity matrices

In this section, we treat the case where $\boldsymbol{G}$ is an $N \times S_\mathrm{C}$ matrix with $N > S_\mathrm{C}$, resulting in a low-rank genetic similarity matrix. The log-likelihood is

$$\log\mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma^2\left(\gamma\boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{I} - \gamma\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)\right).$$

Let $\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top$, with $\boldsymbol{U}_1 \in \mathbb{R}^{N\times S_\mathrm{C}}$ and $\boldsymbol{\Lambda}_1 \in \mathbb{R}^{S_\mathrm{C}\times S_\mathrm{C}}$, be the economy spectral decomposition of $\boldsymbol{G}\boldsymbol{G}^\top$ as in Section 3.4. Replacing $\boldsymbol{G}\boldsymbol{G}^\top$ by its spectral decomposition and writing out the formula for the logarithm of a Normal distribution yields an expression for the log likelihood of

$$-\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\log\left|\gamma\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{I} - \gamma\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right| \tag{4.8}$$

$$-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top\left(\gamma\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{I} - \gamma\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{4.9}$$

As in the full rank case, we have to solve for the maximum-likelihood parameters, and evaluate the squared form of the Normal distribution and the determinant of the covariance term. In Sections 4.1.3–4.1.3, we provide efficient solutions for each of these steps.

Maximum likelihood parameters Given γ, the maximum likelihood weight parameters of the log likelihood in Equation (4.9) are given by the generalized least squares estimator

$$\boldsymbol{\beta}_{\mathrm{M}_\gamma} = \left(\boldsymbol{X}^\top\left(\gamma\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{I} - \gamma\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\left(\gamma\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{I} - \gamma\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)^{-1}\boldsymbol{y}. \tag{4.10}$$

Given γ and $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$, the maximum likelihood genetic variance parameter is given by

$${\sigma^2}_{\mathrm{M}_\gamma} = \frac{1}{N}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top\left(\gamma\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{I} - \gamma\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\top\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{4.11}$$

Analogously to the full rank update section earlier, here, quadratic forms are again need for evaluation of Equations (4.9), (4.10), and (4.11). An efficient solution for these quadratic forms is provided in Section 4.1.3.

Determinant update The determinant in the log-likelihood in Equation (4.9) that we have to evaluate is

$$\log \left| \gamma \boldsymbol{U}_1 \boldsymbol{\Lambda}_1 \boldsymbol{U}_1^\top + \boldsymbol{I} - \gamma \tilde{\boldsymbol{G}} \tilde{\boldsymbol{G}}^\top \right|.$$

Given the log determinant of $\left(\gamma \boldsymbol{U}_1 \boldsymbol{\Lambda}_1 \boldsymbol{U}_1^\top + \boldsymbol{I}\right)$ from Equation (3.21), we can apply the matrix determinant lemma to evaluate the log determinant of the updated linear mixed model covariance,

$$\sum_{n=1}^{S_\mathrm{C}} \log \left(\frac{1}{\gamma \lambda_n + 1} \right) + \log \left| \boldsymbol{I} - \gamma \tilde{\boldsymbol{G}}^\top \left(\gamma \boldsymbol{U}_1 \boldsymbol{\Lambda}_1 \boldsymbol{U}_1^\top + \boldsymbol{I} \right)^{-1} \tilde{\boldsymbol{G}} \right|. \tag{4.12}$$

Using the equivalence shown in Equation (D.10), we get the expression

$$\tilde{\boldsymbol{G}}^\top \left(\gamma \boldsymbol{U}_1 \boldsymbol{\Lambda}_1 \boldsymbol{U}_1^\top + \boldsymbol{I} \right)^{-1} \tilde{\boldsymbol{G}} = \boldsymbol{M}_k + \boldsymbol{M}_{N-k}, \tag{4.13}$$

where the k_up-by-k_up matrices $\boldsymbol{M}_k$ and $\boldsymbol{M}_{N-k}$ are obtained as

$$\boldsymbol{M}_k = \left(\boldsymbol{U}_1^\top \tilde{\boldsymbol{G}} \right)^\top (\gamma \boldsymbol{\Lambda}_1 + \boldsymbol{I}_k)^{-1} \left(\boldsymbol{U}_1^\top \tilde{\boldsymbol{G}} \right),$$

and

$$\boldsymbol{M}_{N-k} = \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1 \boldsymbol{U}_1^\top \right) \tilde{\boldsymbol{G}} \right)^\top \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1 \boldsymbol{U}_1^\top \right) \tilde{\boldsymbol{G}} \right).$$

This $S_\mathrm{C} \times S_\mathrm{C}$ matrix can be computed in $O((N + S_\mathrm{C}) k_\mathrm{up})$ time. Substituting the expression from Equation (4.13) into the determinant from Equation (4.12), we obtain

$$\sum_{n=1}^{S_\mathrm{C}} \log \frac{1}{\gamma \lambda_n + 1} + \log \left| \boldsymbol{I}_N - \gamma \boldsymbol{M}_k - \gamma \boldsymbol{M}_{N-k} \right|,$$

which can be evaluated in $O(S_\mathrm{C} + {k_\mathrm{up}}^3)$ time, resulting in a total runtime of $O({k_\mathrm{up}}^3 + (N + S_\mathrm{C}) k_\mathrm{up})$ to evaluate the log determinant.

Squared form update Here we derive efficient evaluations for the squared form

$$\boldsymbol{a}^\top \left(\underbrace{\gamma \boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{I}}_{\boldsymbol{H}_\gamma} - \gamma \tilde{\boldsymbol{G}} \tilde{\boldsymbol{G}}^\top \right)^{-1} \boldsymbol{b}, \tag{4.14}$$

that allows for efficient evaluation of Equations (4.9), (4.10), and (4.11) by plugging in the the appropriate values for $\boldsymbol{a}$ and $\boldsymbol{b}$.

Given the inverse of $\boldsymbol{H}_\gamma$, we can apply the Sherman-Morrison-Woodbury identity in Equation (4.6) to derive the inverse of the updated genetic similarity matrix as

$$\left(\boldsymbol{H}_\gamma - \gamma \tilde{\boldsymbol{G}} \tilde{\boldsymbol{G}}^\top \right)^{-1} = \boldsymbol{H}_\gamma^{-1} + \gamma \boldsymbol{H}_\gamma^{-1} \tilde{\boldsymbol{G}} \left(\boldsymbol{I}_{k_\mathrm{up}} - \gamma \tilde{\boldsymbol{G}}^\top \boldsymbol{H}_\gamma^{-1} \tilde{\boldsymbol{G}} \right)^{-1} \tilde{\boldsymbol{G}}^\top \boldsymbol{H}_\gamma^{-1}.$$

When plugging this expression into the squared form in Equation (4.14) that we need to evaluate, we obtain

$$\boldsymbol{a}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{b} + \gamma \boldsymbol{a}^\top \boldsymbol{H}_\gamma^{-1} \tilde{\boldsymbol{G}} \left(\boldsymbol{I}_{k_\mathrm{up}} - \gamma \tilde{\boldsymbol{G}}^\top \boldsymbol{H}_\gamma^{-1} \tilde{\boldsymbol{G}} \right)^{-1} \tilde{\boldsymbol{G}}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{b}. \tag{4.15}$$

Noting that there are now squared expressions in $\boldsymbol{H}_\gamma^{-1}$, we can use the solution for the low-rank quadratic form in Equation (D.10) to efficiently evaluate these expressions in $O((N + S_\mathrm{C})k_\mathrm{up})$. The additional required inversion of an $k_\mathrm{up} \times k_\mathrm{up}$ matrix has runtime $O({k_\mathrm{up}}^2)$. Finally, using the following ordering of computations, we can efficiently compute the required matrix products.

$$\boldsymbol{a}^\top \boldsymbol{H}_\gamma^{-1} \boldsymbol{b} + \gamma \left(\left(\boldsymbol{a}^\top \boldsymbol{H}_\gamma^{-1} \right) \tilde{\boldsymbol{G}} \right) \left(\boldsymbol{I}_{k_\mathrm{up}} - \gamma \tilde{\boldsymbol{G}}^\top \boldsymbol{H}_\gamma^{-1} \tilde{\boldsymbol{G}} \right)^{-1} \left(\tilde{\boldsymbol{G}}^\top \left(\boldsymbol{H}_\gamma^{-1} \boldsymbol{b} \right) \right).$$

The total runtime to evaluate this expression becomes $O((N + S_\mathrm{C})k_\mathrm{up} + {k_\mathrm{up}}^2)$.

Restricted maximum likelihood

So far the derivations have been limited to maximum likelihood parameter estimation. However, it is straightforward to extend these results to the restricted log likelihood $\log \mathcal{L}\left(\gamma, \sigma^2 | \boldsymbol{S}\boldsymbol{y}\right)$ from Equation (2.51), which comprises the log likelihood with the restricted maximum likelihood estimator of the fixed effects $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ plugged in, plus three additional terms [Kang et al., 2008, Harville, 1974]:

$$\log \mathcal{L}\left(\boldsymbol{\beta}_{\mathrm{R}_\gamma}, \gamma, \sigma^2\right) + \frac{D}{2} \log\left(2\pi\sigma^2\right) + \frac{1}{2} \log \left| \boldsymbol{X}^\top \boldsymbol{X} \right| - \frac{1}{2} \log \left| \boldsymbol{X}^\top \left(\gamma \boldsymbol{K}' + \boldsymbol{I} \right)^{-1} \boldsymbol{X} \right|,$$

where the restricted maximum likelihood estimator of the fixed effects $\boldsymbol{\beta}_{\mathrm{R}_\gamma}$ equals the form of the maximum likelihood estimator $\boldsymbol{\beta}_{\mathrm{M}_\gamma}$.

We observe that the only additional term involving the updated genetic similarity matrix is

$$\log \left| \boldsymbol{X}^\top \left(\underbrace{\gamma \boldsymbol{G}\boldsymbol{G}^\top - \gamma \tilde{\boldsymbol{G}} \tilde{\boldsymbol{G}}^\top}_{\gamma \boldsymbol{K}'} + \boldsymbol{I} \right)^{-1} \boldsymbol{X} \right|,$$

which again involves a squared form that can be solved efficiently using the efficient squared form update from Equation (4.7) for the case when $\boldsymbol{G}\boldsymbol{G}^\top$ has full rank and Equation (4.15) for the case where $\boldsymbol{G}\boldsymbol{G}^\top$ has low rank.
The restricted maximum likelihood estimator of the residual variance component, given by

$$\sigma^2_{\mathrm{g}\mathrm{R}_\gamma} = \frac{1}{N - D} \left(\boldsymbol{U}^\top \boldsymbol{y} - \boldsymbol{U}^\top \boldsymbol{X} \boldsymbol{\beta}_{\mathrm{R}_\gamma} \right)^\top \left(\gamma \boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{I} - \gamma \tilde{\boldsymbol{G}} \tilde{\boldsymbol{G}}^\top \right)^{-1} \left(\boldsymbol{U}^\top \boldsymbol{y} - \boldsymbol{U}^\top \boldsymbol{X} \boldsymbol{\beta}_{\mathrm{R}_\gamma} \right),$$

involves no additional expensive terms to be compute compared to the ML solution. The formulas for the remaining parameters remain unchanged. The space requirements for resrticted maximum likelihood estimation are the same as those for maximum likelihood estimation.

4.2. A simple heuristic to avoid dilution

As mentioned, we refer to the possible loss of power and worse stratification due to the use of SNPs, that are not associated to the phenotype in the computation of genetic relatedness, as dilution. We combat dilution by identifying SNPs for inclusion in the

genetic similarity matrix that are most strongly associated with the phenotype when no correction for confounding is performed. That is, we select those SNPs that are associated with the phenotype according to univariate linear regression. SNPs identified in this way are likely either to be indirectly associated with the phenotype (e.g., by way of population structure), to have a direct effect on the phenotype, or tag a hidden common cause of the SNP and the phenotype. These three categories represent precisely those SNPs that we want to include in computing the similarity matrix.

As for the threshold for inclusion, we have found the following to yield improved power and calibration. First, we order SNPs by their linear-regression P values. Then, we construct genetic similarity matrices with an increasing number of SNPs according to this ordering until we find the first minimum in the genomic control factor λ. We determined the first minimum in λ by a coarse grid search followed by a golden section search. For the Chrons disease genome-wide association studies, the grid search consisted of the SNP set sizes 0, 100, 200, 300, 400, and the golden section search consisted of the SNP set sizes 280, 340, 320, 290, and 310.

In practice, the number of SNPs selected is typically smaller than the number of individuals analyzed, a condition that can be exploited by the FaST-LMM algorithm presented in Section 3.4, to yield runtime, that is linear in the number of samples.

Others have explicitly used only a subset of available SNPs as covariates to correct for population structure [Setakis et al., 2006], and have included only a subset of SNP principal components that are predictive of phenotype so as to increase genome-wide association studies power [Novembre and Stephens, 2008, Lee et al., 2011].

4.3. Emprical assessment of FaST-LMM-Select

Together, the linear-regression scan to select SNPs for inclusion in the matrix along with removal of the test SNPs and those nearby constitute our new approach, FaST-LMM-Select.

4.3.1. Assessment of dilution and proximal contamination in simulations

We explored the detrimental effects of dilution and proximal contamination using synthetic data so as to have access to ground truth. As in other papers examining correction for population structure in genome-wide association studies, SNPs were generated with the Balding-Nichols model [Astle and Balding, 2009]. We used 3000 individuals consisting of two populations in a ratio of six to four. We chose 100 SNPs at random to be causal of the phenotype, half of which were differentiated between the two populations ($F_{ST} = 0.1$), and the other half not. We generated the phenotype by way of the linear mixed model, using the 100 causal SNPs in the genetic similarity matrix (realized relationship matrix), no fixed effects, and parameters that were comparable to what has been seen on real data when using a traditional linear mixed model approach (genetic variance=0.1, residual variance=0.1) [Kang et al., 2010].

FaST-LMM-Select yielded better calibration and more power than the traditional approach, which in turn yielded less deflation and more power than using a small number of equi-spaced SNPs (as in the original version of FaST-LMM). Furthermore, we saw

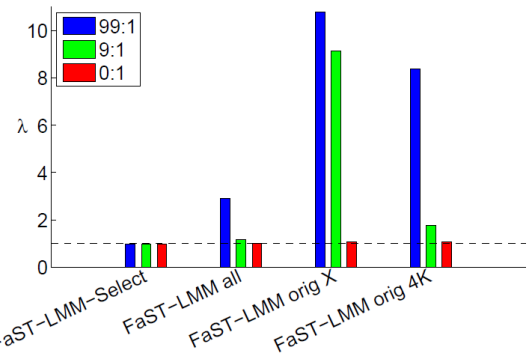


Figure 4.3. Synthetic experiments showing effects of dilution on calibration while avoiding proximal contamination. We generated 100,000 SNPs that could be used to construct the similarity matrix and varied the proportion of undifferentiated to differentiated SNPs (99:1, 9:1, and 0:1), with $F_{ST} = 0.1$ for the differentiated SNPs. The test set comprised another 5000 independently generated SNPs, of which twenty percent were differentiated ($F_{ST} = 0.1$). "FaST-LMM all" refers to use of all SNPs in the similarity matrix. "FaST-LMM orig X" refers to the random selection of X SNPs for the similarity matrix, where X was the number used by FaST-LMM-Select. FaST-LMM orig 4K refers to using 4,000 randomly selected SNPs to estimate genetic similarity. (4,000 equi-spaced SNPs were used in Chapter 3). A realized relationship matrix was used for genetic similarity.

that the deleterious effects of proximal contamination were lessened when dilution was greater.

Calibration under dilution in the absence of proximal contamination First, we examined how circumventing dilution in the absence of proximal contamination improved calibration (the avoidance of inflation or deflation of the test statistic). In particular, we generated 100,000 SNPs that could potentially be used in the genetic similarity matrix (only some of which would be selected by our method). We varied the proportion of undifferentiated to differentiated SNPs (99:1, 9:1, and 0:1), with $F_{ST} = 0.1$ for the differentiated SNPs. Although there is evidence that many SNPs are undifferentiated (e.g., the fact that Ancestry Informative Marker panels typically number in the hundreds [Kidd et al., 2011, Price et al., 2008, Nassir et al., 2009]) we wanted to examine how spurious associations change under a range of scenarios. We used a test set comprising another 5,000 independently generated SNPs, of which twenty percent were differentiated ($F_{ST} = 0.1$). We chose such a test set for three reasons: (1) we wanted the set to be constant across the different proportions of 99:1, 9:1 and 0:1, (2) we wanted a reasonably high proportion of SNPs to be differentiated as these are the ones that become spuriously associated due to confounding, and (3) we wanted the set to be independent from SNPs in the genetic similarity matrix so that proximal contamination could not occur. No SNP in the test set was causal, but we expected those that were differentiated to be spuriously associated with the phenotype if confounding was not corrected for, thus producing inflated test statistics. We also expected that, with a smaller and smaller proportion of differentiated

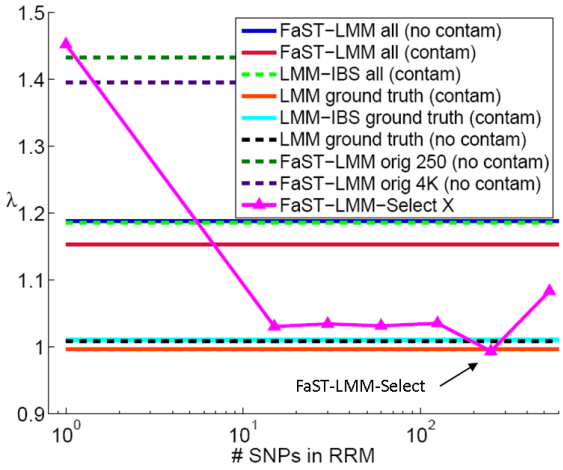


Figure 4.4. Synthetic experiments showing effects of both dilution and proximal contamination on calibration. We generated 100,000 SNPs that could be used to construct the similarity matrix and varied the proportion of undifferentiated to differentiated SNPs (99:1, 9:1, and 0:1), with $F_{ST} = 0.1$ for the differentiated SNPs. The test set comprised another 5000 independently generated SNPs, of which twenty percent were differentiated ($F_{ST} = 0.1$). "FaST-LMM orig X" refers to the random selection of X SNPs for the similarity matrix, where X was the number used by FaST-LMM-Select. "FaST-LMM orig 4K" refers to using 4,000 randomly selected SNPs to estimate genetic similarity. (4,000 equi-spaced SNPs were used in Chapter 3.) Variations in λ when both dilution and proximal contamination could occur. We limited ourselves to the 99:1 condition from Figure 4.3, and used the same 100,000 SNPs for possible inclusion in the similarity matrix. The test set comprised the true causal SNPs as well as a 5,000 SNP subset of the 100,000 SNPs allowed in the matrix (including the 1,000 that were differentiated). The genomic control factor λ is plotted as a function of number of SNPs used in the similarity matrix with our new approach when contamination was accounted for (line with triangular points). A first minimum in λ occurs when 250 SNPs were used. "FaST-LMM-Select X" refers to the use of the top X SNPs from linear regression to estimate genetic similarity. A realized relationship matrix was used for genetic similarity except for the conditions labeled "IBS all" and "IBS ground truth", wherein identity by state was used with all available and ground truth SNPs, respectively.

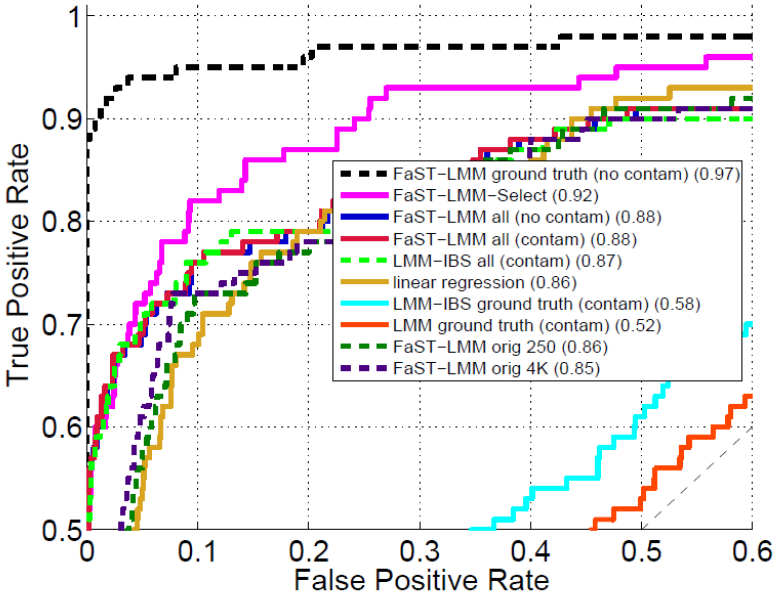


Figure 4.5. Synthetic experiments showing effects of both dilution and proximal contamination on power. We generated 100,000 SNPs that could be used to construct the similarity matrix and varied the proportion of undifferentiated to differentiated SNPs (99:1, 9:1, and 0:1), with $F_{ST} = 0.1$ for the differentiated SNPs. The test set comprised another 5000 independently generated SNPs, of which twenty percent were differentiated ($F_{ST} = 0.1$). "FaST-LMM orig X" refers to the random selection of X SNPs for the similarity matrix, where X was the number used by FaST-LMM-Select. "FaST-LMM orig 4K" refers to using 4,000 randomly selected SNPs to estimate genetic similarity. (4,000 equi-spaced SNPs were used in Chapter 3). Receiver operating characteristic curves and area under the curve (in parentheses) when both dilution and proximal contamination could occur. We limited ourselves to the 99:1 condition from Figure 4.3, and used the same 100,000 SNPs for possible inclusion in the similarity matrix. The test set comprised the true causal SNPs as well as a 5,000 SNP subset of the 100,000 SNPs allowed in the matrix (including the 1,000 that were differentiated). The genomic control factor λ is plotted as a function of number of SNPs used in the similarity matrix with our new approach when contamination was accounted for (line with triangular points). A first minimum in λ occurs when 250 SNPs were used. "FaST-LMM-Select X" refers to the use of the top X SNPs from linear regression to estimate genetic similarity. A realized relationship matrix was used for genetic similarity except for the conditions labeled "IBS all" and "IBS ground truth", wherein wherein identity by state was used with all available and ground truth SNPs, respectively.

SNPs used in the realized relationship matrix, dilution would lead to more and more inflated test statistics, because the differentiated SNPs were those that should be included in the matrix. Indeed, we saw these results (Figure 4.3). Only FaST-LMM-Select remained calibrated across all experimental conditions, whereas other approaches were calibrated only when all SNPs were differentiated (0:1). As expected, calibration for the other approaches became worse as fewer SNPs were differentiated. Linear regression, not shown in the figure, yielded extremely inflated test statistics ($\lambda = 10.2$).

Calibration and power under proximal contamination Next, we examined how dilution and proximal contamination together affected calibration and power. Here we limited ourselves to the 99:1 condition just described, using the same 100,000 SNPs for possible inclusion in the realized relationship matrix as in the previous experiment. The test set comprised the true causal SNPs as well as a 5,000 SNP subset of the 100,000 SNPs allowed in the genetic similarity matrix (including the 1,000 SNPs that were differentiated). When accounting for proximal contamination, we removed only the test SNP itself from the matrix (rather than using the 2 centimorgan rule that we apply on real data), because the synthetic SNPs are not in physical linkage disequilibrium. FaST-LMM-Select used 250 SNPs in the matrix, as this is where the first minimum in λ occurred (Figure 4.4), and yielded $\lambda = 0.99$, comparable to $\lambda = 1.01$ from the ground truth matrix (using only the causal SNPs) that accounts for proximal contamination. In contrast, when all SNPs were used in the matrix, λ was strongly inflated as in the previous experiment. Note that identity by state performed similarly to the realized relationship matrix, but does not have the required factored decomposition which allows FaST-LMM to run most efficiently, nor is it directly amenable to the efficient algorithm for removing SNPs to account for proximal contamination. Also note that using a random selection of SNPs in the matrix (in the experiments in Section 3.5, we used equi-spaced SNPs, which corresponds to a random selection in these synthetic experiments) did not perform well, either with 4,000 SNPs, or 250 SNPs, the number used by FaST-LMM-Select.

Turning to power (Figure 4.5), when proximal contamination was avoided with the ground truth genetic similarity matrix, the linear mixed model obtained nearly perfect power, whereas failing to avoid proximal contamination dramatically reduced power—no SNP signal remained. In contrast, when all available SNPs where used in the matrix, proximal contamination had little effect on power, illustrating the interaction between dilution and proximal contamination. FaST-LMM-Select obtained the most power among methods that did not have access to the ground truth. Note that whether the realized relationship matrix or identity by state were used with all, or ground truth SNPs, power and λ were about the same. Using a random selection of SNPs did not perform well, either with 4,000 SNPs or 250 SNPs, the number used by FaST-LMM-Select. Finally, note that although dilution and proximal contamination had opposite effects on λ (so that models having both effects appeared to perform well in terms of calibration), both effects reduced power.

4.3.2. Genome-wide association study of Crohn's disease

When applied to Wellcome Trust Case Control Consortium data for Crohns disease (see Section A.1) [Burton et al., 2007], including close family members and non-Caucasians,

Algorithm	λ	false positives	false negatives	Runtime without speedup (min)	Runtime with speedup (min)	Memory usage (GB)
FaST-LMM-Select	1.08	0	1	1.3×10^3	45	< 1
FaST-LMM (all)	1.09	2	2	4.0×10^5	4,567	86
FaST-LMM (orig 310)	1.26	9	1	1.1×10^3	6	< 1
FaST-LMM (orig 4,000)	1.17	5	1	2.1×10^5	30	2
Traditional	0.97	2	6	4.1×10^1	NA	45

Table 4.2. Algorithm performance on Chron's disease. The original version of FaST-LMM, which used equally spaced SNPs to estimate genetic similarity, was evaluated using 310 SNPs (the same number used by FaST-LMM-Select) and 4,000 SNPs (as used in the original version of FaST-LMM). The five algorithms yielded substantially different P values (Figure 4.6), which in turn led to different SNPs being deemed significant (using the P value threshold of 5×10^7 [Burton et al., 2007]). Previous studies were used to determine the gold standard in order to label the false positive and false negative loci (for a list of all SNPs found significant by at least one method see supplementary Table 1 in Listgarten et al. [2012].). "speedup" refers to the use of efficient low-rank updates to avoid recomputing the spectral decomposition of the genetic similarity matrix when correcting for proximal contamination.

FaST-LMM-Select performed well (see Table 4.2). Compared with the use of all SNPs (while still accounting for proximal contamination), FaST-LMM-Select had slightly less inflation and fewer false positives (due to lack of dilution), and used far less computer time and memory. Compared with the traditional approach, FaST-LMM-Select had better calibration, far more power, and better computational efficiency. Compared with the original version of FaST-LMM, wherein equi-spaced SNPs were used to reduce computational demands, FaST-LMM-Select had far better calibration and fewer false positives. Finally, the avoidance of proximal contamination alone (comparing "FaST-LMM all" with "Traditional" wherein all available SNPs are used) had a dramatic effect on calibration and false positives, even though only 516 SNPs on average were excluded from the genetic similarity matrix for the testing of a given SNP.

4.3.3. Genome-wide association study of LDL in a Finnish cohort

The first cohort consists of 5,546 Finnish individuals in the 1966 Northern Finland Birth Cohort (NFBC66)[1] [Sabatti et al., 2008, Rantakallio, 1969].

Among the available phenotypes, we analyzed low-density lipoprotein, as it had the most genetic structure ($\lambda = 1.10$) among the phenotypes having genome-wide significant SNPs. We used a 2 megabase exclusion window, because genetic distances were not available. The relative performance of the different algorithms was similar to that for the WTCCC data. In particular, FaST-LMM-Select, which chose 300 SNPs, yielded a λ of 1.02. In contrast, using all available SNPs and correcting for proximal contamination gave $\lambda = 1.05$, showing inflation with respect to FaST-LMM-Select due to dilution. The traditional approach, which used all available SNPs but did not correct for proximal contamination, yielded a lower value ($\lambda = 1.00$), demonstrating the effect of deflation compared to the analysis that corrected for proximal contamination. As for power, using all SNPs (with or without correcting for proximal contamination) identified three loci

[1] For a description of the data see Section A.4

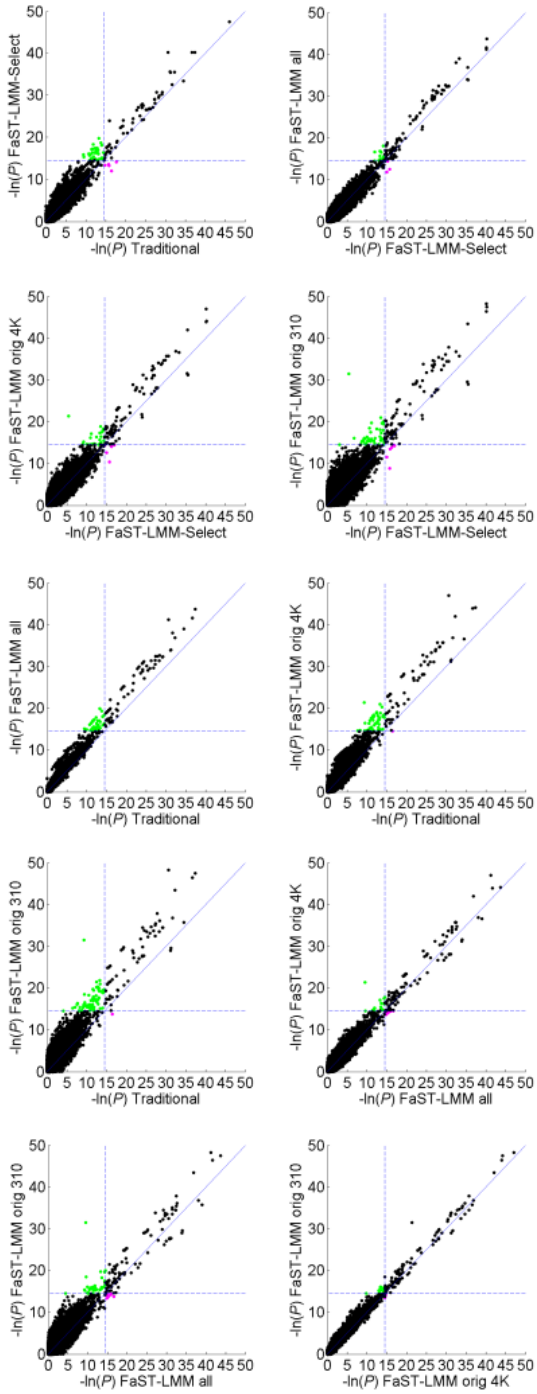


Figure 4.6. A comparison of P values for the algorithms described in Table 4.2. Each point in a plot shows the paired negative log P values of association for a particular SNP from two algorithms. Dashed lines show the genome-wide significance threshold (5×10^{-7}). Green points indicate SNPs called significant by the algorithm shown on the y axis but not the algorithm shown on the x axis, whereas magenta points indicate the opposite. The algorithm with the lower value for λ (see Table 4.2) is shown on the x axis.

as significant, ($p < 7.2 \times 10^{-8}$) as in Kang et al. [2010]. The first locus was near genes CELSR2, PSRC1, SORT1 on chromosome 1, the second was near APOB on chromosome 2, and the third was LDLR on chromosome 19. Associations with all three loci have been validated [Sabatti et al., 2008]. In contrast, FaST-LMM-Select identified these same three loci and one additional locus near genes FADS1 and FADS2 on chromosome 11, which also has been validated [Sabatti et al., 2008].

4.3.4. Genome-wide association study of smoking

Data for this cohort was obtained from the Genetic Analysis Workshop (GAW) 14 [Edenberg et al., 2005]. It consisted of autosomal SNP data from an Affymetrix SNP panel and a phenotype indicating whether an individual smoked a pack of cigarettes a day or more for six months or more (see Section A.2). The cohort included over eight ethnicities and numerous close family members—1,034 individuals in the dataset had parents, children, or siblings also in the dataset. As in the main paper, we used a 2 centimorgan exclusion window.

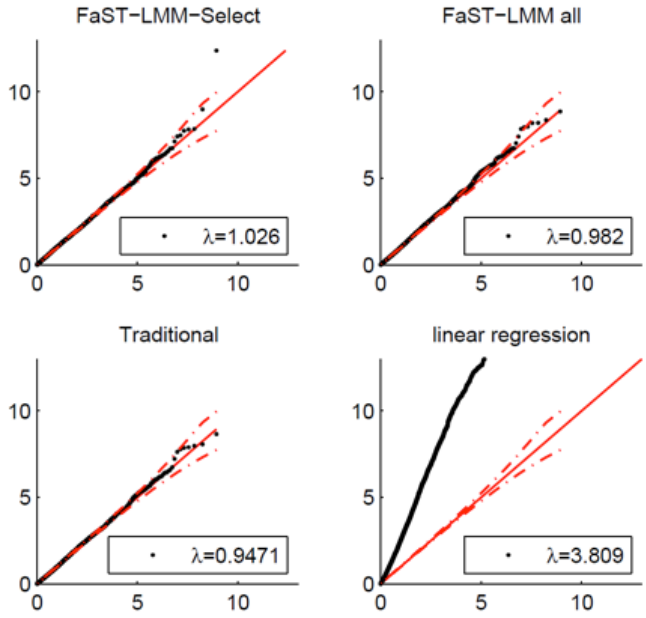
On this data, linear regression yielded $\lambda = 3.8$, significantly higher than 1.0 ($p < 0.001$), reflecting the large amount of genetic structure. Despite this substantial structure, FaST-LMM-Select chose only 650 SNPs and was well calibrated, yielding λ not significantly different from 1.0 ($p = 0.19$; Figure 4.7(a)). Interestingly, FaST-LMM-Select identified a single SNP, rs1950284, as significant ($p = 1.7 \times 10^{-8}$). While this association has not been validated, the SNP lies in the GPHN gene, for which a prior association with other forms of addiction has been reported [Enoch et al., 2012]. Use of all available SNPs in the similarity matrix while accounting for proximal contamination also yielded no significant deviation from $\lambda = 1.0$ ($p = 0.24$), but did not identify this SNP as significant. The traditional approach (use of all SNPs and not accounting for proximal contamination) yielded λ significantly lower than 1.0 ($p = 0.02$). This deflation presumably resulted from not accounting for proximal contamination. Statistical significance of deviation of λ from 1.0 was estimated using a Monte Carlo simulation of the null distribution (uniform on $[0, 1]$) with 1000 sampled distributions.

To demonstrate the robustness of FaST-LMM-Select to extremely strong genetic structure, we filtered the data to include only sib pairs ($N = 920$). Again, FaST-LMM-Select was well calibrated, yielding λ not significantly different from 1.0 ($p = 0.31$; Figure 4.7(b)). Here, the approach used 630 SNPs in the genetic similarity matrix. Possibly due to the reduced sample size, the SNP rs1950284 no longer reached genome-wide significance. Use of all available SNPs in the similarity matrix, either accounting or not accounting for proximal contamination, also yielded no significant deviation from $\lambda = 1.0$ ($p = 0.41$, $p = 0.16$).

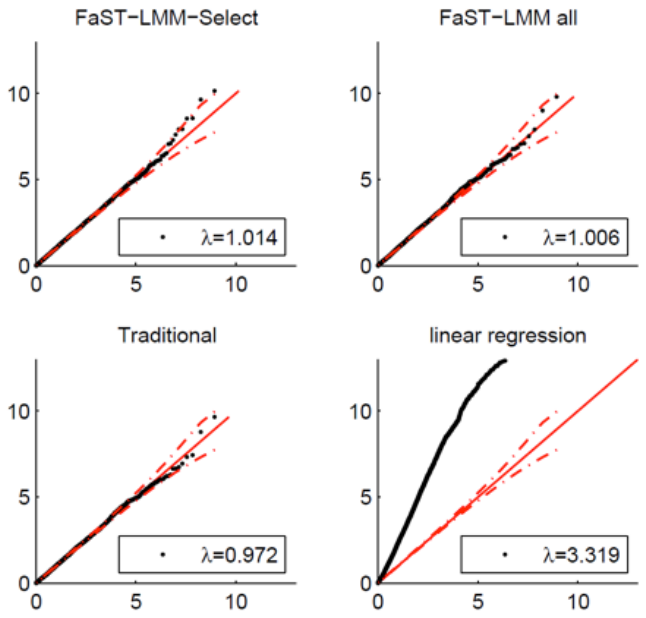
4.3.5. Genome-wide association study of flowering time in *A. thaliana*

The data was taken from a GWAS of 107 phenotypes on 199 *Arabidopsis thaliana* inbred lines [Atwell et al., 2010] (see Section A.5). *Arabidopsis thaliana* exhibits continuous isolation by distance at every geographic scale with the result that population genetic models assuming discrete populations work poorly on this species [Platt et al., 2010a].

FaST-LMM-Select chose 800 SNPs for the genetic similarity matrix. Note that values



(a) QQ plots for GAW14 using all data



(b) QQ plots for GAW14 using sib pairs only

Figure 4.7. Comparison of calibration obtained by FaST-LMM-Select for the analysis of GAW14 data. Quantile-quantile plots of negative log P values for FaST-LMM-Select, FaST-LMM all (using all available SNPs to estimate genetic similarity and accounting for proximal contamination), Traditional (using all SNPs to estimate genetic similarity but not accounting for proximal contamination), and linear regression, on a GWAS of (a) the GAW14 data and (b) a subset including only sib pairs. Dashed lines show 0.05 confidence intervals.

of λ were somewhat noisy due to the small sample size of this cohort. Consequently, we identified the first minimum using a grid search smoothed by a polynomial fit, rather than golden section search.

There were many strong associations in this cohort, making it difficult to evaluate calibration [Atwell et al., 2010]. Consequently, as in Atwell et al. [2010], we compared methods by their ability to identify SNPs that were likely a priori to be associated with a given phenotype. Following the main example used in Atwell et al. [2010], we analyzed the phenotype of flowering time at 10° Centigrade. For each method, we sorted SNPs by their P value of association, identifying the most strongly associated k SNPs for k ranging from 1 to 2000 (Atwell et al. [2010] selected approximately 2000 SNPs using an uncorrected approach, and approximately 250 SNPs using a LMM—see their Figure 3). Then, for each method and value of k, we determined how many of the k associations coincided with candidate SNPs, those that were within 20 kilobases (as in Atwell et al. [2010]) of a gene likely to be associated with flowering (Figure 4.8). The list of such genes was provided by Atwell et al. [2010] and was an updated version from the one used in their paper.

Over the range of k, FaST-LMM-Select generally identified the most candidate SNPs (i.e., true positives) among the top-ranked k SNPs, followed by FaST-LMM all (where all available SNPs were used in the genetic similarity matrix), the traditional LMM approach (which used all available SNPs and did not account for proximal contamination), and finally linear regression. At $k = 2000$, these methods (in order) identified 176, 148, 147, and 110 true positives. Only FaST-LMM-Select identified more SNPs than what would have been expected by chance (P values reported in Figure 4.8).

4.3.6. Experimental details

A likelihood ratio test was used to compute P values (see Section 2.2.5). The calibration of P values was assessed using the λ statistic, also known as the inflation factor from genomic control (see Section 2.3.1) Devlin and Roeder [1999], Balding [2006]. The value λ is defined as the ratio of the median observed to median theoretical test statistic. Values of λ substantially greater than (less than) 1.0 are indicative of inflation (deflation) (see Section 2.3.1). Missing SNP data was mean imputed. Runtimes were measured on a 40-core Dell PowerEdge R910 machine with a 2.0 GHz clock and 256 GB of RAM. All algorithms used the MKL for linear algebra computations.

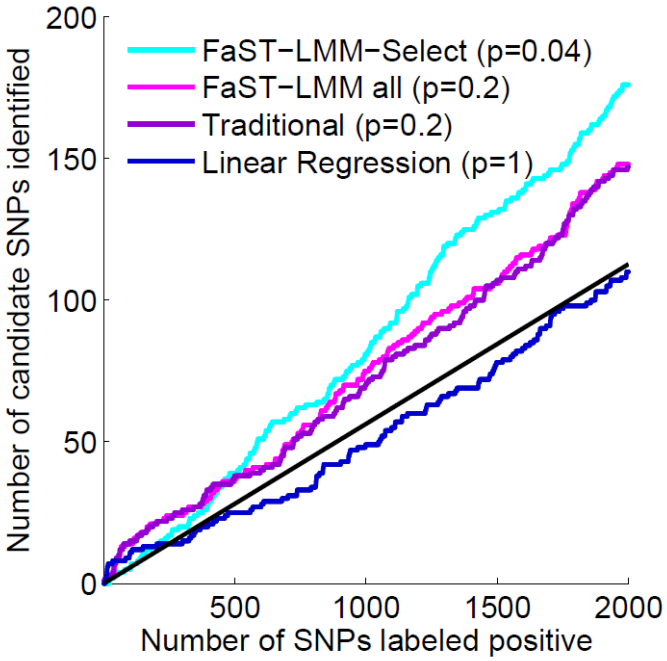


Figure 4.8. Enrichment of likely SNP associations for the trait of flowering time at 10° Centigrade for results obtained from FaST-LMM-Select and comparison methods. Number of candidate SNPs (i.e., true positives) identified versus the number of SNPs labeled positive (k) are plotted for each method. The methods include Fast-LMM-Select, Fast-LMM all, the traditional LMM approach which ignores proximal contamination, and linear regression. The solid black line shows what would be expected by chance. Two-sided P values for whether candidate SNPs were more enriched than by chance are shown adjacent to each curve. These P values were determined using the permutation method described in Supplementary Information 3.3 of Atwell et al. [2010], which preserves the linkage-disequilibrium structure in the SNPs.

4.4. Chapter summary and discussion

Here we pointed out two possible pitfalls when performing GWAS using LMMs, the effects of proximal contamination and dilution.

Proximal contamination refers to a loss of power due markers linked to the marker being tested entering the genetic similarity matrix. We demonstrated this effect on data from the WTCCC [Burton et al., 2007] as well as on GAW14 [Edenberg et al., 2005] and provided an efficient and accurate way co correct for this effect. To this end we developed a window-based method that avoids re-computing the matrix of genetic similarities by performing low rank updates.

Dilution refers to the effect of markers unrelated to the phenotype entering the matrix of genetic similarities, leading to both a loss in power as well as reduced ability to correct for confounding. To avoid dilution, we proposed a simple method for selecting markers to be included in computation of genetic similarities by association to the phenotype.

In synthetic experiments as well as in GWAS from four cohorts with substantial genetic structure we demonstrated that FaST-LMM-Select yields an improvement in power as well as calibration over the traditional approach of using genome-wide markers to compute genetic similarities.

Here, markers where selected by association to the phenotype. Another feature selection criterion that shows great promise is out of sample prediction [Lippert et al., 2013b].

So far we considered tests of single SNPs for association with a phenotype. In Section 5.1, based on FaST-LMM-Select we develop a test for association between sets of SNPs and a phenotype.

5. Aggregating multiple effects in linear mixed models

Approaches for testing sets of variants, such as a set of rare or common variants within a gene or pathway, for association with complex traits are a promising way to detect causal variants with smaller effect sizes or effects of rarer variants. In particular, set tests allow for aggregation of weak signal within a set, enable interplay among variants to be captured, and reduce the burden of multiple hypothesis testing. Unfortunately, until now, these approaches did not address confounding by family relatedness and population structure, a problem that is becoming more important as larger data sets are used to increase power.

In traditional GWAS one single variant at a time is tested for association with a heritable trait, overlooking interplay between SNPs, missing weak signal that aggregates in sets of related SNPs, and incurring a severe penalty for multiple testing.

More recently, sets of SNPs have been tested jointly in a gene-set enrichment style approach [Holden et al., 2008], and also in seeking association between rare variants within a gene and disease [Wu et al., 2011, Bansal et al., 2010]. As next generation sequencing rapidly becomes the norm, these set-based tests, complementary to single SNP tests, will become increasingly important.

Several types of approaches have been used to jointly test sets of SNPs: post-hoc, gene-set enrichment style approaches in which univariate P values are aggregated [Holden et al., 2008]; operator-based aggregation such as "collapsing" of SNP values [Braun and Buetow, 2011, Li and Leal, 2008]; and kernel based approaches such as a linear mixed models [Wu et al., 2010, 2011, Quon et al., 2013]. The latter methods can be interpreted as tests for significant local heritability in a genomic region [Quon et al., 2013].

However, until recently none of the existing methods for testing sets of SNPs handle confounders arising when related individuals or those of diverse ethnic backgrounds are included in the study. Such confounders, when not accounted for, result in spurious associations and loss in power [Balding, 2006, Price et al., 2010b]. Yet it is precisely these richly structured cohorts which yield the most power for discovery of the genetic underpinnings of complex traits. Moreover, such structure typically presents itself as data cohorts become larger and larger to enable the discovery of weak signals.

For cases, where it is not clear how to define such sets, sparse predictors of all genome-wide SNPs, use of shrinkage priors or employing stepwise forward selection has been successful [Yang et al., 2012, Schwender et al., 2011, Malo et al., 2008]. Applying a Laplacian prior leads to the Lasso [Li et al., 2011]. Other related priors have also been considered [Hoggart et al., 2008].

For the former application of testing pre-defined sets of markers for association to a phenotype we introduce a new approach called *FaST-LMM-Set*, a variance component based test that handles confounders (see Section 5.1). The model uses two random

effects—one to capture the set association signal and one to capture confounders. Based on the the FaST-LMM algorithm (see Section 3.4) we also introduce a computational speedup for two-random-effects models that makes this approach feasible even for extremely large cohorts, whereas it otherwise would not be. Experiments based on synthetic data demonstrate control of type I error and better power of the likelihood ratio test over a more traditional score test. Application of FaST-LMM-Set to the richly structured GAW14 data demonstrates that our method successfully corrects for population structure, while application of our method to WTCCC Crohn's disease demonstrates that our method additionally recovers genes not recoverable by univariate analysis.

For the latter case where such sets are not available, we propose *LMM-Lasso*, a model for multi-locus mapping using shrinkage estimators while accounting for relatedness in the mixed model framework. We show practical use in GWAS through retrospective analyses. In data from *Arabidopsis thaliana* and mouse, the benefits in modeling are demonstrated by significant improvements in prediction of phenotype from genotype in 91% of the phenotypes considered. At the same time the results are interpretable as the model dissects these predictions into components due to individual SNP effects and general population structure. In addition to improved prediction, enrichment of known candidate genes suggests that the associations retrieved by LMM-Lasso are more likely to be genuine.

5.1. A powerful and efficient set test for GWAS

In this section we introduce a new approach for set tests of genetic variants that handles confounders. As mentioned, our approach is based on the linear mixed model (LMM), which has an equivalence to linear regression. As we have argued in Chapters 2 and 4 this equivalence states that use of a LMM with a particular genetic similarity matrix is the same as regressing those SNPs used to estimate genetic similarity on the phenotype [Hayes et al., 2009, Listgarten et al., 2012]. One may choose to regress on SNPs for a number of reasons, including correction for confounders in GWAS [Yu et al., 2006, Kang et al., 2010, Listgarten et al., 2012], testing them for association with a phenotype [Wu et al., 2010, 2011], and conditioning on other causal SNPs to increase power [Kang et al., 2010, Atwell et al., 2010, Segura et al., 2012].

Independent of the use of LMMs for matters of population structure correction, the use of LMMs to jointly test sets of rare variants has become prevalent [Wu et al., 2010, 2011]. In our new approach, we marry the aforementioned uses of LMMs to perform set tests in the presence of confounders within a single, robust, and well-defined statistical model.

The first of the two random effects in our model captures confounders in a manner similar to the common usage of a LMM for correction of confounding variables in GWAS. By including this rich covariance structure among the individuals, the individuals effectively become de-correlated, thereby avoiding spurious signal and loss of power otherwise caused by such structure. The second random effect captures signal from the set of SNPs of interest (in a manner similar to the SKAT algorithm used to test sets of rare variants when no confounders are present [Wu et al., 2011, Lee et al., 2012b]).

Because of the aforementioned equivalence, our approach can also be viewed as a

form of linear regression with two distinct sets of covariates. The first set of covariates consists of markers that correct for confounders, that is, those which predict race and relatedness, for example. Their inclusion makes the data for individuals independently and identically distributed (i.e., knowing the value of these markers induces a common distribution from which the individuals are drawn). The second set of covariates consists of SNPs for a given set of interest, such as those SNPs belonging to a gene. We call our approach FaST-LMM-Set.

Computing the likelihood for our model—a LMM with two random effects—is, naively, extremely expensive, as it scales cubically with the number of individuals [Yu et al., 2006, Listgarten et al., 2010]. For example, on the 15,000 individual WTCCC data set we analyse, currently available algorithms would need to compute and store in memory genetic similarity matrices of dimension $15,000 \times 15,000$ and repeatedly perform cubic operations on them to test just a single set of SNPs. However, extending the work presented in Chapter 3 that made LMMs with a single random effect linear in the number of individuals [Lippert et al., 2011] to the two-variance component model needed here, we bypass this computational bottleneck, yielding a new two-random-effects algorithm which is linear in the number of individuals. This advance enables us to analyse data sets which could not otherwise be practically analysed, such as the 15,000 individual WTCCC cohort [Burton et al., 2007]. For example, using the naive cubic approach to test the gene set IL23R (containing 14 SNPs) took 13 hours as compared to one minute for our new approach (all on a single processor), demonstrating a factor speed-up of 780.

5.1.1. Linear mixed models with two variance components

The log likelihood in the linear regression view is given by

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \sigma_{\mathrm{g}}^2\right) = \log \int \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{G}\boldsymbol{u}\,;\, \sigma^2 \boldsymbol{I}\right) \mathcal{N}\left(\boldsymbol{v} \,\middle|\, \boldsymbol{0}\,;\, \frac{\sigma_{\mathrm{g}}^2}{S}\boldsymbol{I}\right) \mathrm{d}\boldsymbol{u} \tag{5.1}$$

where $\boldsymbol{y}$ is a $1 \times N$ vector of phenotype values for N individuals; $\boldsymbol{\beta}$ is the set of the fixed effects of the covariates stored in the design matrix $\boldsymbol{X}$; σ^2 is the residual variance in the regression; $\boldsymbol{u}$ are the $S \times 1$ random effects for the SNPs, stored in the design matrix $\boldsymbol{G}$ (dimension $N \times S$), and $\mathcal{N}\left(\boldsymbol{u} \,\middle|\, \boldsymbol{0}\,;\, \sigma_{\mathrm{g}}^2 \frac{1}{S}\boldsymbol{I}\right)$ is the distribution for those weights. That is, the random regression weights, $\boldsymbol{u}$ are marginalized over independent Normal distributions with equal variance σ_{g}^2/S. Equivalently, the log likelihood is sometimes written with random effects marginalized out,

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \sigma_{\mathrm{g}}^2\right) = \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right). \tag{5.2}$$

where the kernel, $\boldsymbol{K}$, is given by $\boldsymbol{K} = \frac{1}{S}\boldsymbol{G}\boldsymbol{G}^T$ as is the case, for example, when $\boldsymbol{K}$ is given by the realized relationship matrix (RRM) [Hayes et al., 2009, Lippert et al., 2011]. Given this equivalence, the SNPs used to estimate genetic similarity (those in $\boldsymbol{G}$) can be interpreted as a set of random covariates in the regression. In our model, we partition the random effects into two sets: one set of random effects, $\boldsymbol{u}_C$ (with design matrix $\boldsymbol{G}_C$), are used to correct for confounders, while the other, $\boldsymbol{u}_S$, are used to test our set of SNPs

in the corresponding design matrix, $\boldsymbol{G}_S$. The log likelihood is then written

$$\begin{aligned}&\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \sigma_C^2, \sigma_S^2\right)\\ =&\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{G}_C\boldsymbol{u}_C + \boldsymbol{G}_S\boldsymbol{u}_S\,;\, \sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{u}_C \,\middle|\, \boldsymbol{0}\,;\, \frac{\sigma_C^2}{S_C}\boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{u}_S \,\middle|\, \boldsymbol{0}\,;\, \frac{\sigma_S^2}{S_S}\boldsymbol{I}\right)\mathrm{d}\boldsymbol{u}_C\mathrm{d}\boldsymbol{u}_S,\end{aligned} \tag{5.3}$$

where each set of random effects has a separate variance (σ_C^2/S_C and σ_S^2/S_S). Equivalently, we can write this in the marginalized form

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \sigma_C^2, \sigma_S^2\right) = \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \frac{\sigma_C^2}{S_C}\boldsymbol{G}_C\boldsymbol{G}_C^\top + \frac{\sigma_S^2}{S_S}\boldsymbol{G}_S\boldsymbol{G}_S^\top + \sigma^2\boldsymbol{I}\right) \tag{5.4}$$

For convenience, we re-parameterize this as

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \sigma_C^2, \sigma_S^2\right) = \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma_\mathrm{g}^2\left(\frac{1-\tau}{S_C}\boldsymbol{G}_C\boldsymbol{G}_C^\top + \frac{\tau}{S_S}\boldsymbol{G}_S\boldsymbol{G}_S^\top\right) + \sigma^2\boldsymbol{I}\right), \tag{5.5}$$

where now the covariance matrix, $\boldsymbol{K}$, has been partitioned into two variance components:

$$\boldsymbol{K} = (1-\tau)\boldsymbol{K}_C + \tau\boldsymbol{K}_S, \tag{5.6}$$

using $\boldsymbol{K}_C = \frac{1}{S_C}\boldsymbol{G}_C\boldsymbol{G}_C^\top$ (where $\boldsymbol{G}_C$ is of dimension $N \times S_c$) to account for confounders, and $\boldsymbol{K}_S = \frac{1}{S_S}\boldsymbol{G}_S\boldsymbol{G}_S^\top$ (where $\boldsymbol{G}_S$ is of dimension $N \times S_S$) to account for signal from a pre-defined set of SNPs such as those within a gene; and $\tau \in [0, 1]$ is a scalar parameter estimated from the data by, for example, restricted maximum likelihood (REML). The null model for our set test is given by $\tau = 0$, while the alternative model allows $\tau \geq 0$.

Until recently, estimating the parameters and computing the likelihood of a LMM was cubic in the number of individuals. However, as shown in Section 3.4, when the number of SNPs used to estimate genetic similarity, S, is less than the cohort size, N, and when genetic similarity matrix, $\boldsymbol{K}$, factors as $\boldsymbol{G}\boldsymbol{G}^\top$ ($\boldsymbol{G}$ of dimension $N \times S$), then the computations (and memory requirement) become linear in N [Lippert et al., 2011]. So far this result has been applied in the context of correcting for confounders in a univariate GWAS, with just a single variance component.

In principle, the S SNPs for inclusion in $\boldsymbol{G}$ could be obtained by sampling SNPs genome-wide and relying on linkage disequilibrium, as was done in Section 3.5, or using feature selection (see Chapter 4) as one would do in any statistical modelling problem [Listgarten et al., 2012]. Here we used the feature selection method described in Section 4.2. We performed an uncorrected, univariate scan of the SNPs to select those which should be used to correct for confounders.

Thus, in our approach, we select S_c SNPs for $\boldsymbol{K}_C$ by first sorting all available SNPs according to their univariate linear-regression P values (in increasing order), and then evaluate the use of more and more SNPs according to this ordering, until we find the first minimum in λ, using a grid search. This resulted in 650 and 310 SNPs for the GAW14 and Crohn's analyses, respectively. Additionally, any SNPs that were being tested (i.e., those in $\boldsymbol{G}_S$), and those within 2 centimorgans, were removed from $\boldsymbol{G}_C$ so as not to contaminate the null model (see Section 4.1) [Listgarten et al., 2012].

For estimation of variance parameters, and computation of the likelihood ratio test statistic, we use the restricted likelihood. While the restricted likelihood does not yield a valid nested likelihood ratio test for the case of testing fixed effects, it is a valid likelihood when testing random effects, and can be computed in the same time and memory complexity as the (unrestricted) likelihood.

When testing sets in an uncorrected manner, that is, without accounting for confounders (which we did for comparison purposes), we omitted the portion of the variance which corrected for confounders, $\boldsymbol{K}_C$. In particular, we set $\tau = 1$, and tested the significance of σ^2_{g} with the same likelihood ratio test described next.

5.1.2. Statistical testing of variance components

We have now fully specified our model for doing set tests when confounders are present.

Likelihood ratio test

To obtain a P value on the set of SNPs of interest, such as those belonging to a gene (i.e., those in $\boldsymbol{G}_S$), we use a likelihood ratio test. In particular, to test the significance of the set of SNPs of interest, we compare the maximum restricted likelihood of the data with and without the set of SNPs of interest, that is, the maximum restricted likelihood of the alternative and null models. More formally, our null hypothesis is given by $\mathcal{H}_0 : \tau = 0$, while our alternative hypothesis is given by $\mathcal{H}_1 : \tau > 0$.

To obtain calibrated P values, we require an accurate estimate of the distribution of statistics under the null hypothesis. However, obtaining a sufficiently accurate estimate of this distribution is not straight-forward. Standard software uses a parametric form for this distribution of

$$\mathrm{LRT} \sim 0.5\chi^2_0 + 0.5\chi^2_1, \tag{5.7}$$

a 50-50 mixture of two χ^2 distributions, one with one degree of freedom, and the other with zero degrees of freedom, the latter which accounts for the fact that the tested parameter is on the boundary of the allowed space in the null model [Self and Liang, 1987, Dominicus et al., 2006]—that is, to account for the fact that $\tau = 0$ in the null model, and $\tau > 0$ in the alternative model (because it is a variance). The necessary regularity conditions for this include that the outcome variable can be partitioned into a large number of identically and independently distributed sub-vectors [Greven et al., 2008]—conditions which are not generally met in our setting because individuals may be arbitrarily related to one-another. It has been shown that when the regularity conditions are not met, the $0.5\chi^2_0 + 0.5\chi^2_1$ distribution yields conservative P values [Greven et al., 2008] because the mixing weight on the χ^2_0 component is too low at fifty percent. We have also found this to be the case in our setting (Table 5.1).

Although one might consider use of a parametric bootstrap to estimate the null distribution (e.g., [Greven et al., 2008]), such an approach dramatically increases the running time over computation of the test statistics themselves. Yet another alternative is to use an empirical distribution based on permutations. Although such an approach might be feasible, one can use many fewer permutations by instead assuming a parametric form of the null distribution and then fitting the few required parameters to the test statistics generated from the permutations. It is such an approach that we take here. Note that

this approach assumes that the null distribution of test statistics is the same across all tests, an assumption that has also been made in the small sample correction in SKAT-O and elsewhere [Lee et al., 2012a, Greven et al., 2008, Greven, 2007].

The parametric form of the null distribution that we assume, and to which we fit null distribution test statistics to, is inspired by Greven [2007], Greven et al. [2008], who reported that a mixture of χ^2_0 and a scaled χ^2_1, where a is the scaling parameter for the scaled chi-square distribution, yielded good type I control when testing a variance component in a single-component LMM. We use the same parametric form of the null distribution, except, to gain additional flexibility for the two-component LMM, we allow the degrees of freedom on the second component, d, to be different from 1 (finding this to be useful). That is, we use the null distribution,

$$a\text{LRT} \sim \pi\chi^2_0 + (1-\pi)\chi^2_d, \tag{5.8}$$

with free parameters π, a, and d. Using this distributional form, we found that a fit to the full range of test statistics yielded P values that were too liberal in the tail (Table 5.1). Thus, we instead fit our parametric form to only the more significant tail of the distribution of test statistics. (Note that in our experiments with just a single variance component, P values were also liberal in the tailthose for which $p \ll 0.05$).

We now describe the details of our approach for estimating the parameters of this null distribution. To generate a null statistic for a set, we permuted the individuals for only the SNPs in that set. Because we do not permute the SNPs (rather, the individuals), the pattern of linkage disequilibrium between the SNPs within a single test remains intact. Although we permute the individuals, who are not (generally) identically and independently distributed, we do so only for the SNPs in the test set, leaving any confounding signal among the covariates, the confounding SNPs and the phenotype intact. We found that parameter estimates stabilized with the use of 10 permutations per test (for both WTCCC and GAW14). Thus, our procedure has a runtime roughly a factor of ten larger than if we had not needed permutations. We used the same 10 permutations across all tests (and all SNPs within a gene).

Given this permutation-generated sample of statistics from the null distribution, we fit the parameters a, d, and π as follows. The χ^2_0 distribution is a Dirac delta function at 0that is, this component of the null distribution yields only test statistics of 0, and correspondingly $p = 1$. Furthermore, the $\chi^2_{d>0}$ yields a test statistic of 0 with measure zero. Consequently, one can obtain good estimates of the parameters simply by assuming that precisely those tests with variance parameter $\tau = 0$ belong to the χ^2_0 component, and then estimating π as the proportion of tests belonging to this component. We estimate a and d directly from the non-zero test statistics (those likely to belong to the $a\chi^2_d$ component) using a regression in which these parameters are adjusted such that resulting P values for LRT have the least squared error compared to the theoretical P values. Specifically, we use the $\log P$ value squared error, and only use the smallest 10% of P values in the regression. This truncated regression approach consistently yielded calibrated quantile-quantile plots (Figure 5.1) and also controlled type I error (Table 5.1). Furthermore, it yielded better power than the score test (Table 5.2).

In summary, our overall approach is as follows: (1) for each set to be tested, permute the individuals of the SNP belonging to this set; (2) compute the restricted LRT statistic

for this permuted data to obtain a test statistic from the null distribution; (3) repeat step (1) ten times; (4) estimate the proportion of test statistics drawn from the χ^2_0 component, $\hat{\pi}$, as the proportion of tests in which the parameter $\tau = 0$; (5) use the largest 10% of test statistics to perform a regression to fit the $a\chi^2_1$ component–that is find $\hat{a}$ and $\hat{d}$ which minimize the squared error of the log_{10} P values with their theoretical values (uniform distribution on $[\pi, 1]$); (6) compute the test statistic for all sets (non-permuted data) and then compute the corresponding P values for these using the null distribution

$$\hat{a}\text{LRT} \sim \hat{\pi}\chi^2_0 : (1 - \hat{\pi})\chi^2_{\hat{d}}.$$

In application to our real data (described later), our procedure yielded $\pi = 0.641$, $a = 2.29$, $d = 0.961$, on the GAW14 data, and $\pi = 0.643$, $a = 1.41$, $d = 0.85$ on the WTCCC data.

5.1.3. Linear-Time Computations

What remains left to explain is how to achieve the linear time speed-up in the case of two random effectsthe present setting. The crux of the cubic to linear time speed-up in the single random effect model was to bypass construction of $\boldsymbol{K}$ and the required spectral decomposition of $\boldsymbol{K}$ by recognizing that one can instead use $\boldsymbol{G}$ and the spectral decomposition of $\boldsymbol{G}$ [Lippert et al., 2011]. Note that we can view the two-random effects model as a single random effect with covariance $\boldsymbol{K} = (1 - \tau)\boldsymbol{K}_C + \tau\boldsymbol{K}_S$. To use the algebraic speed-up just mentioned, we observed that $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{G}^\top$, where now

$$\boldsymbol{G} = \left[(1 - \tau)^{1/2}\boldsymbol{G}_C, \tau^{1/2}\boldsymbol{G}_S\right], \tag{5.9}$$

So long as $S_S + S_C < N$, which was true for all of the data sets examined here, we obtained the linear time computations and memory footprint just as in Section 3.4 [Lippert et al., 2011]. (One might also consider using low rank update equations of the type used in Section 4.1.3 to correct for the effects of proximal contamination, although we did not implement this.) Finally, to perform parameter estimation in this two-random effects model, we used an approach similar to that reported in Lippert et al. [2011]. That is, we used a one-dimensional Brent search optimization routine to find the value of τ which maximized the restricted likelihood. For each call to the restricted likelihood for a particular value of τ, efficiently computations were performed as in Lippert et al. [2011], except using the two random effects.

5.1.4. Experiments

Data Sets and Methods

We formed SNP sets by grouping all variants within a window around a single gene to a set. More generally, this approach of forming sets from windows of nearby SNPs along the genome could be used to map an entire genome into sets, even when the SNPs do not lie in genes. In any case, it is not our goal here to evaluate different ways in which one might group SNPs, but to demonstrate that we can test sets of SNPs in the presence of confounders.

All analyses assumed additive effects of a SNP on phenotype, using a 0/1/2 encoding for each SNP (indicating the number of minor alleles for an individual). Missing SNP data was mean imputed. Multiple testing was accounted for with a Bonferroni correction.

WTCCC: For the WTCCC data described in Section A.1, we grouped SNPs into gene sets using gene positions provided on the USCSC Genome Browser[1] [Kent et al., 2002, Dreszer et al., 2012] using build hg19 (we also converted the original WTCCC annotations to this build), which yielded 13,850 gene sets. We concentrated our evaluations on Crohn's disease, as inflation for this phenotype was greatest with an uncorrected univariate analysis. The set sizes ranged from 1 to 748, with a mean value of 11, and a standard deviation of 24.

In counting hits for Crohn's disease (Table 5.4), we omitted any genes found in the MHC region because it is complicated by very long range linkage disequilibrium. We used positions 29-34 Mb on chromosome 6 as the boundaries of the MHC, as suggested by the MHC sequencing consortium [Pereyra et al., 2010].

GAW14: Because the SNPs for the GAW14 data (see Section A.2 mapped to only 251 non-singleton gene sets with this strategy, we formed sets for this analysis by using overlapping 1 centimorgan windows, yielding 2,157 sets. The set sizes ranged from 2 to 38, with a mean value of 5, and a standard deviation of 4.

Experimental Set-Up to Assess Control of Type I Error and Power

We used synthetic data based on the real WTCCC data to assess the quality of our new method, as well as to compare it against a more traditional score test. In particular, to assess type I error, we used SNPs from the WTCCC data set, and then permuted the individuals for SNPs in each set tested so as to create null only test statistics. We permuted the data set a total of 72 times, yielding 997,200 null test statistics (because 13,850 sets were tested for each data set). We additionally permuted another 10 data sets in order to estimate the parameters of the null distribution (π,a,d).

For assessment of power, we again used the SNPs from the WTCCC data set, and then generated synthetic phenotypes using a linear mixed model. To do so, we first we fit the null model to the real data to obtain estimates of the parameters σ^2 and σ^2_{g}. Then we simulated the phenotype from a linear model

$$\boldsymbol{y} = \underbrace{\boldsymbol{G}_S\boldsymbol{\beta}_S}_{\text{signal}} + \underbrace{\boldsymbol{G}_C\boldsymbol{\beta}_C}_{\text{confounding}} \ ,$$

where the signal effects are drawn from a normal distribution such that their contribution to the phenotypic variance equals the genetic variance estimated on the real data ($\sigma^2_{\mathrm{g}} = 0.0125$)

$$\boldsymbol{\beta}_S \sim \mathcal{N}\left(\mathbf{0}\,;\, \frac{\sigma^2_{\mathrm{g}}}{S_S}\boldsymbol{I}\right). \tag{5.10}$$

[1] `http://genome.ucsc.edu/`

The confounding effects where drawn from a normal distribution, such that their contribution to the phenotypic variance equals the environmental variance estimated on the real data ($\sigma^2 = 0.094$).

$$\boldsymbol{\beta}_C \sim \mathcal{N}\left(\mathbf{0}\,;\, \frac{\sigma^2}{S_C}\boldsymbol{I}\right). \tag{5.11}$$

The same 310 confounding SNPs for $\boldsymbol{G}_C$ are as on the real data, and with all 321,839 SNPs further than 2 cM away from those in $\boldsymbol{G}_C$ for the causal SNPs used to form $\boldsymbol{G}_S$ (those contained in the true positive sets in our power experiments). We generated 5 phenotypes in this way. The resulting phenotypes behaved much like the real data in that, on average, we found 10 Bonferroni-corrected sets on each of 5 data sets, as compared to the 23 found on the real data (note that Table 5.4 does not include SNPs from the MHC region and therefore shows only 16). For both type I error and power experiments, we tested the same gene sets as on the real data.

When comparing our new LRT approach against a score-based test, we used the same score test as SKAT [Wu et al., 2011] which uses the Davies method to compute P values from the null distribution (still with our FaST-LMM-Set model).

Type I Error and Power on Synthetic Data

First we examined whether our new LRT approach controlled type I error. As described in the previous Section, we generated null-only test statistics by way of permutations on the WTCCC data, obtaining a total of roughly 1 million test statistics. The type I error was controlled (Table 5.1). Note that neither fitting the null distribution parameters with all test statistics, by way of maximum likelihood, nor use of a $0.5\chi^2_0 + 0.5\chi^2_1$ null distribution yielded calibrated P values. The first was liberal, while the latter was conservative (Table 5.1). Finally, Figure 5.1 additionally demonstrates good calibration of the entire range of P values from our method, for the same points as in Table 5.1.

Significance Level	$\alpha = 10^{-5}$	$\alpha = 10^{-4}$	$\alpha = 10^{-3}$
Fast-LMM-Set	1×10^{-5}	1.21×10^{-4}	1.01×10^{-3}
non-truncated ML	$2 \times 10^{-5}\star$	$1.83 \times 10^{-4}\star$	$1.26 \times 10^{-3}\star$
$0.5\chi^2_0 + 0.5\chi^2_1$	5×10^{-6}	$4 \times 10^{-5}\star$	$4.55 \times 10^{-4}\star$

Table 5.1. Type I error estimates for FaST-LMM-Set using one million tests across various levels of significance, α. The first row shows results for our new LRT-based method; the second row shows results when fitting the null distribution parameters using maximum likelihood with all test statistics (non-truncated ML); the third row shows results using a $0.5\chi^2_0 + 0.5\chi^2_1$ null distribution. Results significantly different from expected according to the binomial test ($p < 0.05$) are denoted with an asterisk. Next we compared the power of our LRT approach to a score test approach (using the same model) on synthetic data. Over five synthetic data sets and a range of significance levels, LRT found significantly more sets than the score test (Table 5.2). Furthermore, on the real WTCCC data, LRT again found significantly more sets passing the Bonferroni corrected significance threshold (Table 5.4).

α	LRT	score	P value
3.6×10^{-6}	44	26	0.03
10^{-5}	60	39	0.03
10^{-4}	172	138	0.047
10^{-3}	556	509	0.14
10^{-2}	2419	2195	0.0009

Table 5.2. Power experiments for FaST-LMM-Set. Number of tests with P values less than α. The last column shows the results of a binomial test comparing the number of tests found by LRT as compared to the score test. The first row denotes the Bonferroni threshold for the WTCCC data set.

Application to Real Data

We investigated our new approach on two data sets. The first was the Genetic Analysis Workshop (GAW) 14 [Edenberg et al., 2005], which included data from over eight ethnicities and numerous close family members—1,034 of the 1,261 individuals in the dataset had parents, children, or siblings also in the dataset. We used the smoking phenotype as it showed the most confounding. After filtering there were 7,579 SNPs available for analysis. The second data set was from the Wellcome Trust Case Control Consortium (WTCCC) disease with 14,925 individuals [Burton et al., 2007] and 356,441 SNPs after filtering. We used the Crohn's phenotypes because this was the one showing the most confounding in an uncorrected analysis. Unlike the WTCCC [Burton et al., 2007], we included non-white data for individuals and close family members to increase power and because the LMM can treat them properly [Price et al., 2010b, Kang et al., 2010, Astle and Balding, 2009].

To judge the degree of confounding due to genetic relatedness, and to ensure that our LMM approach could sufficiently correct for confounding, we ran both an uncorrected and corrected univariate analysis on each data set, because this is a well-understood test that has been reported on before. Here the extent of test statistic inflation due to unmodelled confounders was assessed using the λ statistic, also known as the inflation factor from genomic control [Devlin and Roeder, 1999]. The value λ is defined as the ratio of the median observed to median theoretical test statistic. Values of λ substantially greater than (less than) 1.0 are indicative of inflation (deflation) (see Section 2.3.1). As can be seen in Table 5.3, without correction, the test statistics are inflated. Although some might consider a λ of 1.08 (seen on the corrected analysis of WTCCC) as still moderately inflated, it has been shown that complex, highly polygenic traits lead to increases in λ [Yang et al., 2011b]. Moreover, the WTCCC themselves reported λ in the range of 1.08-1.11 upon removal of individuals from different races and also any related individuals (neither of which we removed), and upon adjustment with two principal components, suggesting that a λ of 1.08 is the result of polygenic influence [Burton et al., 2007].

Having established that both of our data sets required correction for confounders, and that the LMM with our chosen background kernel, $\boldsymbol{K}_S$, sufficiently corrected for confounders, we next applied FaST-LMM-Set, using the same LMM-correcting component

Method	GAW14	WTCCC
Uncorrected	3.80	1.30
FaST-LMM	1.01	1.08

Table 5.3. Genomic control λ of univariate tests for confounding-corrected and naive methods. FaST-LMM denotes a one-component (to correct for confounding) linear mixed model, testing one SNP fixed effect [Lippert et al., 2011]; Uncorrected refers to no correction for confounding (linear regression).

as in the univariate test. On GAW14, the uncorrected set analysis yielded 241 significant sets, whereas FaST-LMM-Set, which corrects for confounding, yielded none. It is thought that this data set contains little, if any signal (for example based on the univariate analysis). On WTCCC Crohn's disease, an uncorrected set analysis yielded 26 significant sets, whereas FaST-LMM-Set yielded 16 (Table 5.4). Next we investigate these sets in detail.

To validate the significant sets recovered on the WTCCC Crohn's phenotype we used a meta-analysis [Franke et al., 2010, Listgarten et al., 2012], using a 50 kilobase window of inclusion. Additionally, for the genes not found in the meta-analysis, we conducted a literature search[2]. Using our newly developed method, FaST-LMM-Set, we found 16 significant gene sets, of which all but one were found by the meta-analysis. The remaining gene, SLC24A4, performs a similar function to the validated gene SLC22A4both are cation transporters[3]–suggesting a promising candidate for follow-up.

Method	in meta-analysis	supported by literature	no support found
FaST-LMM-Set	15	1	0
FaST-LMM-Set-Score	7	0	0
FaST-LMM-Set (uncorrected)	17	3	6

Table 5.4. Validation of FaST-LMM-Set on WTCCC Crohn's disease. "FaST-LMM-Set" denotes our newly developed method which corrects for confounding using our new LRT approach; "FaST-LMM-Set (uncorrected)" is the same but does not correct for confounding with a second variance component; "FaST-LMM-Set-Score" refers to a score-based approximation to the LRT-based FaST-LMM-Set (and corrects for confounding), as described in Methods. Columns: "in meta-analysis" shows the number of significant sets validated by a meta-analysis [Franke et al., 2010]; "supported by literature" denotes the number of significant sets found by a literature search; "no support found" denotes the number of sets for which we found no support.

In the course of our analysis we noticed that some sets with small P values had almost no univariate signal in any of the SNPs. In particular, among the 16 sets in the WTCCC data supported by either meta-analysis or literature search, six (C1orf141, SAG, SLC24A4, SLC22A4, TCTA, and PTPN2) were missed by the univariate analysis (i.e., a SNP lying in any of the regions reported by Franke et al. [2010] was not found). One of

[2] Detailed validation results are provided in Supplemental Table 1 of Listgarten et al. [2013b].
[3] `www.genecards.org` [Rebhan et al., 1998]

the motivations for doing set analysis is to uncover signals for such cases. The intuition here is the same as in a univariate conditional GWAS analysis. That is, conditioning on variables can lead to an increase in power, revealing signal that would be hidden without the conditioning [Atwell et al., 2010, Segura et al., 2012]. Thus the set test acts not only to aggregate weak signal, but also to unmask signal hidden by covariates included by virtue of doing a set test. We decided to investigate one such case in detail. In particular, we computed the univariate P values for each of the 15 SNPs associated with the gene SLC22A4, marginally, as well as conditioned on all the other SNPs in this gene, using a LMM to correct for confounding. This gene was found to be associated with Crohn's disease using FaST-LMM-Set with $p = 7.6 \times 10^{-8}$. The smallest marginal univariate P value was 1.2×10^{-5}, but when we conditioned on the other SNPs in the set, the smallest conditional univariate P value obtained was 7×10^{-8}. This result demonstrates the increased power afforded by the set test owing to the interplay of SNPs within the gene that are missed by a univariate approach.

Next we computed the correlation between set size and set log P value, for each data set and algorithm, using Pearson correlation with those P values not equal to one (because of the one-sided nature of our test, these would clearly violate assumptions of Pearson correlation). We hypothesized that when confounders were not properly accounted for in the set analysis, that the more SNPs in a set, the more power the set would have to detect these confounders, and therefore the stronger the correlation between set size and P value would appear. Of course, we expect that among sets which are predictive of phenotype, that the larger the number of predictive SNPs in the set, the stronger the correlation between set size and P value will be. As such, on data with signal, we do expect to see some correlation between set size and P value; on data with no signal, we don't expect to see any. Indeed, this is what we observed, as summarized in Table 5.5. Note that when we permuted the Crohn's phenotype to remove signal, the correlation was further reduced to $\rho = 0.019$ ($p = 0.18$).

Dataset	FaST-LMM-Set (uncorrected)	FaST-LMM-Set
GAW14	**0.27** (2×10^{-34})	0.001 (0.98)
WTCCC	**0.051** (3×10^{-5})	0.025 (0.06)

Table 5.5. Pearson correlation of $\log_{10}(P)$ values with set size for tests using FaST-LMM-Set. P value is reported in parentheses next to the value for ρ. Significant entries are bolded. "FaST-LMM-Set" denotes our newly developed method; "FaST-LMM-Set (uncorrected)" is the same but does not correct for confounding using a second variance component.

5.1.5. Section summary and discussion

We have developed a novel, efficient approach for testing sets of genetic markers in the presence of confounding structure such as arises from ethnic diversity and family relatedness within a cohort. Application of this algorithm demonstrated that our method corrects for confounders and uncovers signal not recoverable by univariate analysis. Note, that a number of related approaches for confounder correction in variance component

tests have been proposed in parallel, but are based on the less powerful score test and do not consider linear time computations [Schifano et al., 2012, Oualkacha et al., 2013, Chen et al., 2013].

Although we did not analyze rare variant data, we have shown elsewhere that the underlying LMM methodology for correction of confounding works well to correct for confounding of rare variants in a univariate setting [Listgarten et al., 2013a]. Furthermore, others have already shown that LMM-based set tests work well for detection of sets of associated rare variants [Wu et al., 2011].

It follows that the hybrid approach that we presented here is likely to prove effective in the setting of testing sets of rare variants affected by confounding, although this remains to be investigated fully. We note, however, that we have found the use of a linear model on a case-control phenotype to yield inflated tests statistics when testing rare variants.

As in any regression/classification problem, too many effects relative to sample size can lead to problems of overfitting and/or loss of power. In the case of the mixed model, which integrates out its SNP random effects, one can see a loss of power if too many SNPs are used relative to the sample size [Lippert et al., 2013b]. We do not expect that this was a problem for analyses in the present work.

We have demonstrated that the LRT outperforms a score test when testing variance components in our setting (using the same underlying model). This is perhaps unsurprising given that the score test can be viewed as an approximation to the LRT by a second-order Taylor series expansion [Buse, 1982] in the neighbourhood of the null model. Furthermore, given its robust properties, the LRT is considered the benchmark for statistical testing [Dunson, 2008]. We note, however, that in some recent work [Lin and Tang, 2011], when testing for rare variants using a logistic fixed effects model, a score test was found to perform better than LRT, which was found to be liberal. Although the best test may depends on context, we note that Lin et al used a different model than we did and, in particular, did not use a variance component approach. Also, they used closed-form, asymptotic-based analytical LRT P values rather than making use of empirically-derived null distributions as we have done here.

For many cases of hidden structure in genetic data, the use of principal component-based covariates is sufficient for correction [Price et al., 2006], and thus these covariates could immediately be added to existing models such as SKAT [Wu et al., 2011] to achieve a set test that corrects for confounding. However, it is now widely accepted that there are various forms of confounding which cannot be corrected for by principal components, but for which a LMM adequately corrects [Yu et al., 2005a, Kang et al., 2010, Price et al., 2010b], and it is for these problems that we have developed our approach.

We here focused on testing SNPs in a manner similar to SKAT [Wu et al., 2011]. However, it would be straightforward to also adapt FaST-LMM-Set to the approach of SKAT-O, in which the original SKAT model is in effect combined with a collapsing-type approach [Lee et al., 2012b,a].

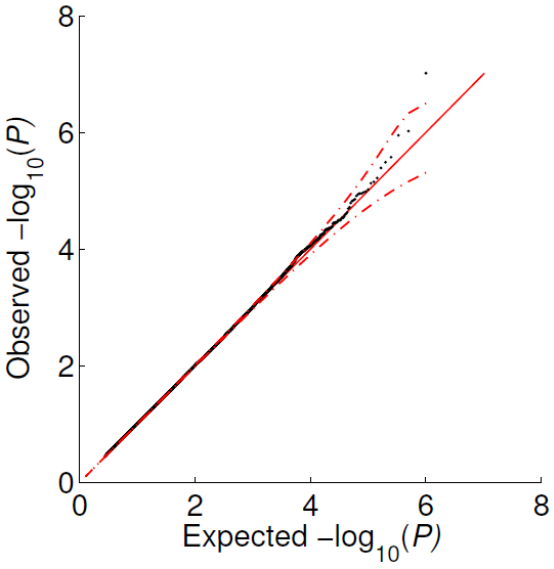


Figure 5.1. Quantile-quantile plot of observed and expected $-\log_{10} P$ values on the null-only WTCCC data sets (same data as used for Table 5.1) for FaST-LMM-Set. Dashed red error bars denote the 99% confidence interval around the solid red diagonal. Points shown are for null-only data (generated by permuting individuals in the SNPs to be testedsee Methods) and only for the non-unity P values (those assumed to belong to the non-zero degree of freedom component). The portion of the expected distribution of P values shown is uniform on the interval $[\pi, 1]$, where π is the mixing weight in the null distribution.

5.2. LMM-Lasso

Similar to the other models considered so far, the phenotype is the sum of individual genetic effects and random confounding effects. In brief, the phenotype of N samples $\boldsymbol{y} = [y_1, \ldots, y_N]$ is expressed as the sum of D fixed covariates, entailing SNPs and additional covariates $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, x_D]$

$$\boldsymbol{y} \sim \mathcal{N}\left(\underbrace{\boldsymbol{X}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\boldsymbol{G}\boldsymbol{u}}_{\text{random}} \; ; \sigma^2 \boldsymbol{I} \right). \tag{5.12}$$

The resulting mixed model is typically considered in the context of single candidate SNPs, i.e. restricting the sum in Equation (5.12) to a single SNP while ignoring all others [Yu et al., 2005a, Kang et al., 2008, 2010, Zhang et al., 2010, Lippert et al., 2011]. While computationally efficient and easy to interpret, this single SNP analysis is compromised by complex genetic architectures with some genetic factors masking others [Platt et al., 2010b]. Some improvement can be achieved by step-wise regression or forward selection, however this introduces side effects due to the ordering used [Yang et al., 2012]. Here, we consider joint inference in the model implied by Equation (5.12). Our approach assesses all SNPs at the same time while accounting for their interdependencies and without making any assumptions on their ordering. To allow for applications to genome-wide

SNP data, we place a Laplacian shrinkage prior over the fixed effects $\boldsymbol{\beta}_s$, assigning zero effect size to the majority of SNPs as done in the Lasso [Tibshirani, 1996].

Our new approach is called the LMM-Lasso since it combines the advantages of established linear mixed models with the Lasso.This allows for dissecting the explained variance in individual SNPs effects from the effects caused by population structure. The model complexity, i.e. the number of individual SNPs included in the model can either be selected through cross-validation, the Bayesian Information Criterion (BIC) or subsampling (for full details on parameter inference see Section 'Statistical model' and the supplementary material).

5.2.1. Linear mixed model Lasso model

Let $\boldsymbol{S}$ denote the $N \times D$ matrix of D SNPs that are modeled as fixed effects for N individuals, $\boldsymbol{s}_{:,d}$ is then the $N \times 1$ vector representing fixed effect d, while $\boldsymbol{x}_{:,d}$ is the $N \times 1$ vector representing the SNP d for all individuals.

We model the phenotype for N individuals, $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ as the sum of genetic effects β_d of the SNPs $\boldsymbol{x}_{:,d}$ and confounding influences $\boldsymbol{v}$

$$\boldsymbol{y} = \underbrace{\sum_{d=1}^{D} \boldsymbol{d}_{:,D}\beta_d}_{\text{genetic effects}} + \underbrace{\boldsymbol{v}}_{\text{confounding effects}} + \underbrace{\boldsymbol{e}}_{\text{noise}}. \tag{5.13}$$

The genetic effects are modeled as fixed effects, whereas the confounding influences are modeled as random effects. The sum is over genome-wide polymorphisms, where the great majority has zero effect size, i.e. $\beta_d = 0$, which is achieved by a Laplacian shrinkage prior on all weights. The random effect $\boldsymbol{v}$ is not observed itself. Instead, we assume that the distribution of $\boldsymbol{v}$ is Gaussian with covariance proportional to the genetic relationship matrix $\boldsymbol{K}$.

$$\boldsymbol{v} \sim \mathcal{N}\left(\mathbf{0}\,;\, \sigma_{\mathrm{g}}^2 \boldsymbol{K}\right).$$

Integrating out the random effects $\boldsymbol{v}$, we can write down the posterior distribution over the weight vector $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{K}, \gamma, \eta) \propto \underbrace{\mathcal{N}\left(\boldsymbol{y} \,|\, \sum_{d=1}^{D} \boldsymbol{x}_d\beta_d\,;\, \sigma_{\mathrm{g}}^2\boldsymbol{K} + \sigma^2\boldsymbol{I}\right)}_{\text{likelihood}} \underbrace{\prod_{d=1}^{D} \exp\left(-\eta|\beta_d|\right)}_{\text{prior}}, \tag{5.14}$$

where $\boldsymbol{X}$ is the genotype matrix, η is the hyperparameter for the Laplacian prior, σ^2 is the residual variance and σ_{g}^2 is the variance of the genetic random components.

As shown in Section 2.2.1, using the realized relationship matrix as the covariance matrix is equivalent to integrating over all SNPs while using an isotropic Gaussian prior [Goddard et al., 2009]. The choice of a Gaussian prior leads to a dense posterior distribution and thus reflects the *a priori* belief that a large fraction of SNPs may contribute a small fraction to the phenotype. This stands in sharp contrast to the generally accepted opinion that most SNPs are actually not associated with the phenotype. From our point of view, the covariance matrix $\boldsymbol{K}$ can be seen as modeling SNP effects

that are confounded due to population structure or are too small to be detected, while single SNPs that have a sufficiently large effect size are directly included in the model over $\boldsymbol{X}$.

Correction for population structure

Learning the hyperparameters $\boldsymbol{\Theta} = \{\eta, \sigma_{\mathrm{g}}^2, \sigma^2\}$ and the weights $\boldsymbol{\beta}$ jointly is a hard non-convex optimization problem. To obtain a practical and scalable algorithm, we first optimize $\sigma_{\mathrm{g}}^2, \sigma^2$ by Maximum Likelihood under the null model, ignoring the effect of individual SNPs (similar to the procedure introduced in Section 3.2.2 for univariate models [Kang et al., 2010, Zhang et al., 2010]).

Instead of working with $\sigma_{\mathrm{g}}^2, \sigma^2$ directly, we choose a different parametrization using $\gamma = \frac{\sigma_{\mathrm{g}}^2}{\sigma^2}$, whose estimator can be learnt more efficient by using the computational tricks proposed in Section 3.3:

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{S}, \boldsymbol{K}, \gamma, \eta) \propto \mathcal{N}\left(\boldsymbol{y} \,|\, \sum_{j=1}^{S} w_j \boldsymbol{s}_j \,;\, \sigma^2(\gamma \boldsymbol{K} + \boldsymbol{I})\right) \prod_{d=1}^{D} \exp\left(-\eta|\beta_d|\right). \tag{5.15}$$

In more detail, we compute the spectral decomposition of the covariance $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ to rotate the data such that the covariance matrix of the normal distribution is a diagonal matrix. We then employ a one-dimensional numerical optimizer to optimize γ.

Reduction to an ordinary Lasso problem

Having fixed γ, we use the spectral decomposition of $\boldsymbol{K}$ again to rotate our data such that the covariance matrix becomes isotropic:

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{K}, \gamma, \eta) \propto \mathcal{N}\left(\tilde{\boldsymbol{y}} \,|\, \sum_{d=1}^{D} \tilde{\boldsymbol{x}}_d \beta_d \,;\, \gamma\boldsymbol{\Lambda} + \boldsymbol{I}\right) \prod_{d=1}^{D} \exp\left(-\eta|\beta_d|\right). \tag{5.16}$$

Here, $\tilde{\boldsymbol{X}}$ denote the rotated and rescaled genotypes and $\tilde{\boldsymbol{y}}$ the respectively phenotypes:

$$\begin{aligned} \tilde{\boldsymbol{X}} &= (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{U}^\top \boldsymbol{X} \\ \tilde{\boldsymbol{y}} &= (\gamma\boldsymbol{\Lambda} + \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{U}^\top \boldsymbol{y}. \end{aligned}$$

For fixed γ and η, computing the most probable weights in Equation (5.16) is equivalent to the Lasso regression model, since maximizing the posterior with respect to $\boldsymbol{\beta}$ is the same as minimizing the negative log of ((5.16)):

$$\min_{\boldsymbol{\beta}} \frac{1}{\sigma^2} \sum_{n=1}^{N} (\tilde{y}_n - \tilde{\boldsymbol{x}}_{\boldsymbol{n},:}\boldsymbol{\beta})^2 + \eta\|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_1$ denotes the ℓ_1-norm of the vector $\boldsymbol{\beta}$.

In experiments, we choose η by cross-validation, minimizing the test set mean squared error. A different algorithm for solving the LMM-Lasso for general purposes is proposed in Schelldorfer et al. [2011], which includes generalized linear mixed models with ℓ_1-penalty.

5.2.2. Phenotype prediction

Using the BLUP predictor from Section 2.2.2 the phenotype $y_\star$ of a new test individual can be predicted by conditioning the joint distribution over all individuals on the training individuals:

$$y_\star \,|\, \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{x}_\star\boldsymbol{\beta} + \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) ; \sigma_\mathrm{g}^2 k_{\star,\star} + \sigma^2 - \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:} \left(\sigma_\mathrm{g}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} \sigma_\mathrm{g}^2 \boldsymbol{k}_{\star,:}^\top \right),$$

where $\boldsymbol{V}_{\boldsymbol{\theta}} = \left(\sigma_\mathrm{g}^2 \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)$.

5.2.3. Selecting the number of active SNPs

Model Selection in Lasso Methods can either be done by choosing the number of active SNPs or varying the hyperparameter η explicitly. For better interpretability, we chose to vary the number of active SNPs directly. The sample size provides a natural limit on how many SNPs one can select. If the sample size permits, we let the number of active SNPs vary in $\{0, 1, 2, \ldots, 10, 20, 30, \ldots, 100, 150, 200, 250\}$. For a fixed number of selected SNPs, we find the corresponding hyperparameter η by a combination of bracketing and bisection [Wu et al., 2009]. To speed up the computations we allowed inexact matches (± 10) if the number of selected SNPs exceeded 100.

For model selection, we employed different strategies:

1. 10-fold cross-validation: We split the data randomly into 10 folds. Each fold is once picked as test dataset, the other folds are used for training. The number of active SNPs is chosen such that the explained variance over all test sets is maximal.

2. Bayesian Information Criterion (BIC) is defined as

   $$BIC_{\beta,\Theta} := -2 \log \mathcal{L}\left(\boldsymbol{\beta}, \gamma\right) + d_{\boldsymbol{\beta},\delta} \log N,$$

   where d the degrees of freedom. Zou et al. [2007] show that the number of nonzero weights is an unbiased estimate for the degrees of freedom for the Lasso. When applying to the LMM-Lasso, we increment the degrees of freedom by one for fitting γ.

3. Stability Selection[Meinshausen and Bühlmann, 2010]: we fix the number of active SNPs to 20 and draw randomly 90% of the data 100 times. All SNPs that are selected in more than 50 repeats are in the active set.

In our experience, the BIC criterion is best suited for variable selection, cross-validation for prediction and stability selection for variable screening.

5.2.4. Experiments

Semi-empirical setting with known ground truth

We assessed the ability of LMM-Lasso to recover true genotype to phenotype associations in a semi-empirical simulated dataset. To ensure realistic characteristics of population structure, we simulated confounding such that it borrows key characteristics from the

Arabidopsis thaliana data [Atwell et al., 2010] as described in Section A.6. *Arabidopsis thaliana* is a strongly structured population that exhibits continuous isolation by distance at every geographic scale [Platt et al., 2010a].

To compare our method with existing techniques, we considered the standard Lasso, which models all SNPs jointly but without correcting for population structure, as well as univariate LMM, which effectively control for confounding, but consider each SNP in isolation. As a baseline, we also considered a standard univariate linear regression model (LM), which neither accounts for confounding nor considers joint effects because of complex genetic architectures. Both, the standard Lasso and LMM-Lasso were fit in identical ways. For LMM and the LMM-Lasso, we used the RRM as covariance matrix and fit γ on the null model. For univariate models, the ranking of individual SNPs was done according to their P values; for multivariate models, we considered the order of inclusion into the model. A fair comparison between the univariate and multivariate methods is difficult, as the univariate methods select blocks of linked markers, whereas the multivariate methods select only representative markers per block. For this reason we tried to account for these subtleties by excluding all additional markers within a region 10kb of the strongest associated marker within a region from the evaluation.

LMM-Lasso ranks causal SNPs higher than alternative methods

First, we compared the alternative methods in terms of their accuracy in recovering SNPs with a true simulated association (Figure 5.2(a)). Methods that account for population structure (LMM-Lasso, LMM) are most accurate, with LMM-Lasso performing best. Although the linear mixed model performs well at recovering strong associations, the independent statistical testing falls short in detecting weaker associations that are likely masked by stronger effects (see Figure 5.3(a)). Comparing methods that account for population structure and naive methods, we observe that accounting for this confounding effect avoids the selection of SNPs that merely reflect relatedness without a causal effect (see Figure 5.3(b)). An alternative evaluation, which considers the receiver operating characteristic curve are given in Figure 5.2(a), yields identical conclusions.

Next, we explored the impact of variable simulation settings. As common in the literature, we used the area under the precision-recall curve as a summary performance measure to compare different algorithms. Precision and recall both depend on the decision threshold, above which a marker is predicted to be positive. By varying this threshold, one obtains a precision-recall curve.

Figure 5.4(a) shows the area under the precision-recall curve as a function of an increasing ratio of population structure and independent environmental noise. When confounding population structure is weak, both the Lasso and the LMM-Lasso perform similar. As expected, the benefits of population structure correction in LMM-Lasso are most pronounced in the regime of strong confounding. We also examined the ability of each method to recover genetic effects for increasing complexities of the genetic model, varying the number of true causal SNPs while keeping the overall genetic heritability fixed (Figure 5.4(b)). LMM-Lasso performs better than alternative methods for the whole range of considered settings with the difference in accuracy being the largest for complex genetic architectures. In a nutshell these results show, that in the regime of a larger number of true weak associations, it is advantageous to include a genetic similarity

$\boldsymbol{K}$ that accounts for some of the weak effects [Yang et al., 2010].

The identical effect is observed when varying the ratio between true genetic signal versus confounding and noise (Figure 5.4(c)). Again, the performance of the LMM-Lasso is superior to all other methods, and the strengths are particularly visible for medium signal to noise ratios.

LMM-Lasso explains the genetic architecture of complex traits in model systems

Having shown the accuracy of LMM-Lasso in recovering causal SNPs in simulations, we now demonstrate that the LMM-Lasso better models the genotype-to-phenotype map in *A. thaliana* and mouse. Here, we focus on 20 flowering time phenotypes for *A. thaliana*, which are well characterized (see Section A.5), and 273 mouse phenotypes, which are relevant to human health (see Section A.7).

LMM-Lasso more accurately predicts phenotype from genotype and uncovers sparser genetic models We perform phenotype prediction to investigate the capability of alternative methods to explain the joint effect of groups of SNPs on phenotypes. To measure the predictive power, we assessed which fraction of the total phenotypic variation can be explained by the genotype using different methods [Ober et al., 2012]. Explained variance is defined as the fraction of the total variance of the phenotype that can be explained by the model and in our experiments equals one minus the mean squared error, as we preprocessed the data to have zero mean and unit variance. We avoided prediction on the training data, as for all methods, this leads to anti-conservative estimates of variance explained because of overfitting (see Figure 5.7 for a comparison).

Figures 5.5(a) and 5.5(a) show the explained variance of the two methods on the independent test dataset for each phenotype in the two datasets. For both model organisms, LMM-Lasso explained at least as much variation as the Lasso. We omitted the univariate methods, as their performance is generally lower because of the simplistic assumption of a single causal SNP (see Figure 5.7 for comparative predictions in *A. thaliana*). In a fraction of 85.00% of the *A. thaliana* and 91.58% of the mouse phenotypes, LMM-Lasso was more accurate in predicting the phenotype, and thus explained a greater fraction of the phenotype variability from genetic factors than the Lasso. In contrast, Lasso achieved better performance in only 15% of the *A. thaliana* and 8.42% of the mouse phenotypes. Beyond an assessment of the genetic component of phenotypes, LMM-Lasso dissects the phenotypic variability into the contributions of individual SNPs and of population structure. Figures 5.5(c) and 5.5(d) show the number of SNPs selected in the respective genetic models for prediction. With the exception of two phenotypes, LMM-Lasso selected substantially fewer SNPs than the Lasso, suggesting that the Lasso includes additional SNPs into the model to capture the effect of population structure through an additional set of individual SNPs. This observation is in line with the insights derived from the simulation setting where the majority of excess SNPs selected by Lasso are indeed driven by population effects. Although the genetic models fit by LMM-Lasso are substantially sparser, they nevertheless suggest complex genetic control by multiple loci. In 90.00% of *A. thaliana* and in 66.06% of the mouse phenotypes, LMM-Lasso selected more than one SNP, in 40.00% and 45.49% of the respective cases, the number of SNPs in the model was > 10.

LMM-Lasso allows for dissecting individual SNP effects from global genetic effects driven by population structure Next, we investigated the ability of LMM-Lasso to differentiate between individual genetic effects and effects caused by population structure. Figure 5.6 shows the explained variances for the phenotype flowering time (measured at 10°C) for *A. thaliana*. Again, these estimates were obtained using a cross-validation approach. It is known that flowering is strikingly confounded by with population structure [Zhao et al., 2007], which explains why the LMM-Lasso already captured a substantial fraction (45.17%) of the phenotypic variance, when using realized relationships alone (number of active SNPs=0). Because of the small sample size, cross-validation can underestimate the true explained variance [Hastie et al., 2001]. Nevertheless, cross-validation is fair for comparison and conservative, as it avoids possible overfitting.

For increasing number of SNPs included in the model, the explained variance of LMM-Lasso gradually shifted from the kernel to the effects of individual SNPs. In this example, the best performance (48.87%) was reached with 30 SNPs in the model, where the relative contribution of the random effect model was 33.10% and of the individual SNPs are 15.77%. In comparison, Lasso explained at most 46.53% of the total variance, when 125 SNPs were included in the model.

Associations found by LMM-Lasso are enriched for SNPs in proximity to known candidate genes Finally, we considered the associations retrieved by alternative methods in terms of their enrichment near candidate genes with known implications for flowering in *A. thaliana*. To avoid the negative effects of proximal contamination demonstrated in Section 4.1, we avoided inclusion of interest in the genetic similarity matrix $\boldsymbol{K}$. Consequently, we applied LMM-Lasso on a per-chromosome basis estimating the effect of population structure from all remaining chromosomes.

To obtain a comparable cut-off of significance, we applied stability selection for both the LMM-Lasso and the Lasso [Meinshausen and Bühlmann, 2010]. Table 5.6 shows that the LMM-Lasso found a greater number of SNPs linked to candidate genes for 12 phenotypes, whereas Lasso retrieved a greater number for only 6 phenotypes. In the remaining two phenotypes, both methods performed identically. We also investigated to what extent the solution is affected by different selection thresholds (see Figure 5.8). Reassuringly, the LMM-Lasso outperformed the standard Lasso over a large range of different values. It is difficult to compare the multivariate approaches with univariate techniques in a quantitative manner, as the univariate models tend to retrieve complete LD-Blocks. Thus, we revert to reporting the P values of the univariate methods for the SNPs detected by the LMM-Lasso. We also considered to what extent the findings provide evidence for allelic heterogeneity or the existence of an imperfectly tagged causal locus. Overall, 14.75% of the SNPs linked to candidate genes and selected by the LMM-Lasso appear as adjacent pairs (Table 5.7), that is, having a distance < 10 kb from each other, whereas 5.56% of the SNPs selected by the Lasso do. From all activated SNPs, 8.18% selected by LMM-Lasso and 18.96% selected by the Lasso have at least a second active SNP in close proximity.

Phenotype	LMM-Lasso	Lasso
LD	**5/54**	4/69
LDV	**5/63**	3/69
SD	**3/55**	2/61
SDV	**5/54**	2/60
FT10	1/48	**4/67**
FT16	3/51	**4/68**
FT22	**2/54**	1/64
2W	**3/53**	2/65
8W	2/51	**4/59**
FLC	**5/52**	3/53
FRI	3/43	3/46
8WGHFT	**4/59**	2/66
8WGHLN	1/48	**4/58**
0WGHFT	**4/58**	3/63
FTField	**4/61**	3/69
FTDiameterField	1/49	1/51
FTGH	1/49	**2/61**
LN10	**3/50**	2/67
LN16	2/58	**3/64**
LN22	**4/54**	2/65

Table 5.6. Associations close to flowering candidate genes in *A. thaliana* **detected by LMM-Lasso and Lasso.** We report true positives/positives (TP/P) for LMM-Lasso and Lasso for all phenotypes related to flowering time in *Arabidopsis thaliana*. P are all activated SNPs and TP are all activated SNPs that are close to candidate genes.

Phenotype	Chrom.	Position	GeneID	LM	LMM
LD	4	(466307,466800)	AT4G01060	(2.55,6.40)	(3.37,4.20)
2W	4	(454542,460246)	AT4G01060	(8.29,1.89)	(6.03,4.26)
FLC	4	(205170,210657)	AT4G00450	(6.88,5.40)	(5.01,4.78)
FRI	4	(268809,268990)	AT4G00650	(20.91,15.13)	(17.45,13.65)
FRI	4	(268990,276143)	AT4G00650	(15.13,17.36)	(13.65,14.37)

Table 5.7. List of flowering candidate genes in *A. thaliana* **containing multiple associations.** List of all candidate genes that have two activated SNPs in close proximity for all phenotype related to flowering time of *Arabidopsis thaliana*. The last two columns show the $-\log_{10}$ transformed P values for the linear and the linear mixed model.

5.2.5. Section summary and discussion

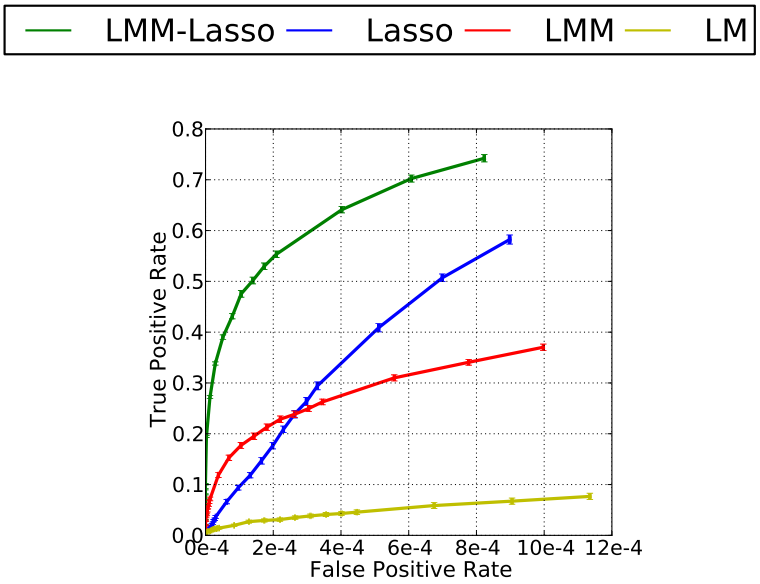
Here, we have presented a Lasso multi-marker mixed model (LMM-Lasso) for detecting genetic associations in the presence of confounding influences such as population structure. The approach combines the attractive properties of mixed models that allow for elegant correction for confounding effects and those of multi-marker models that con-

sider the joint effects of sets of genetic markers rather than one single locus. Thus, LMM-Lasso leads to improved recovery of true genetic effects, even in challenging settings with complex genetic architectures, weak effects of individual markers or presence of strong confounding effects.

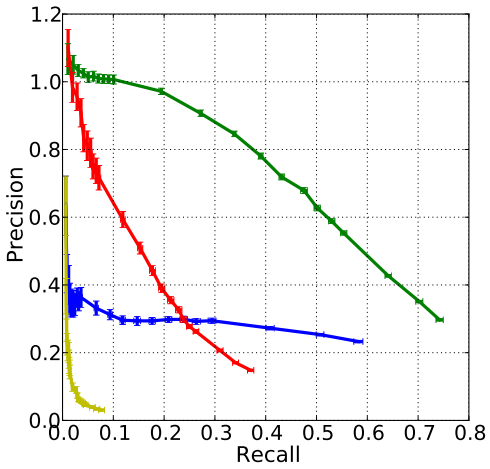
LMM-Lasso is relevant for genome-wide association studies on complex phenotypes, particularly the large number of phenotypes whose genetic basis is conjectured to be multifactorial [Flint and Mackay, 2009]. We show the practical utility of LMM-Lasso to real settings in different retrospective analyses in *Arabidopsis thaliana* and mouse. First, LMM-Lasso is better able to predict phenotype from genotype, suggesting that the underlying model accounting for both, population structure effects and multi-locus effects, is a better fit to real genetic architectures. It is widely accepted that the missing heritability in single-locus genome wide association mapping can often be explained by a large number of loci that have a joint effect on the phenotype [Yang et al., 2010] while leading only to weak signals of association if considered independently. In addition to recovering greater fractions of the heritable component of quantitative traits, LMM-Lasso allows for differentiating between variation that is broad-scale genetic and hence likely caused by population structure and individual genetic effects. In *Arabidopsis* and mouse, this approach revealed substantially sparser genetic models than naive Lasso approaches. Second, LMM-Lasso retrieves genetic associations that are enriched for known candidate genes. In line with the findings in Yang et al. [2012], we retrieved an increased rate of physically adjacent SNPs selected in proximity to candidate genes.

Neither the concept to account for population structure nor multivariate modeling of the genetic data are novel *per se*. An approach for distinct populations based on multi-task learning is presented in Puniyani et al. [2010]. However, with the notable exception of Schelldorfer et al. [2011], these approaches do not include random effects to control for confounding. Our approach to combining mixed models with Lasso is much more scalable and efficient, enabling the application to genome wide settings. In Foster et al. [2007], a combination of linear mixed models and the Lasso is also proposed, but the markers are modeled as random Lasso effects. Among prior work on mixed models, few considered joint effects of multiple loci. Perhaps closest related are variance component models [Yang et al., 2010]. The strength of the approach presented here is the combination of the regime of variance component modeling and multivariate models for several individual effects which is instrumental for the increase in genetic variation our model can explain.

In summary, we believe that LMM-Lasso is a useful addition to the current toolbox of computational models for unraveling genotype-phenotype relationships. As sample sizes increase, the power of detecting multifactorial effects will quickly rise. However, multi-marker mapping is inherently linked to the challenge of some markers being picked up by the model due to their correlation with a confounding variable, such as population structure. In a pure Lasso regression model, it is unclear which markers merely reflect these hidden confounders. LMM-Lasso, in contrast, explains confounding explicitly as random effect, and thus, helps to resolve the ambiguity between individual genetic effects and phenotype variability due to population structure.

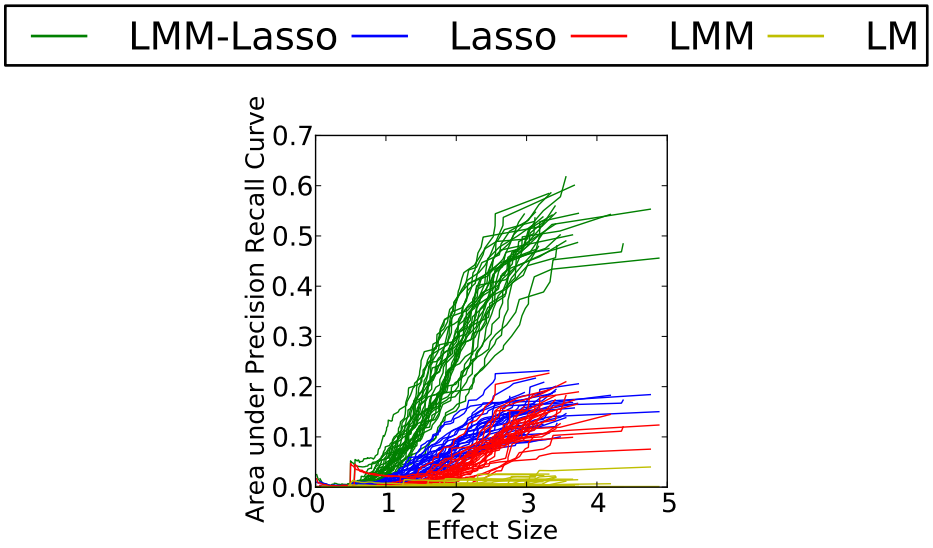


(a) ROC

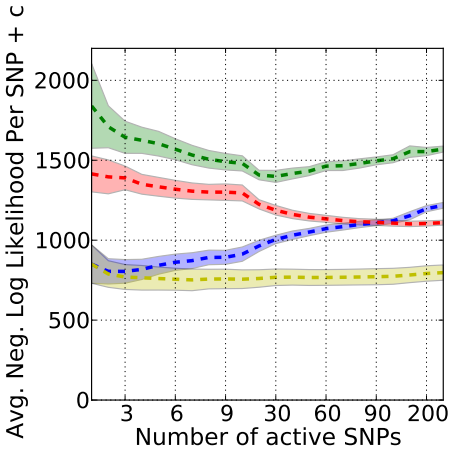


(b) Precision/Recall

Figure 5.2. Evaluation of LMM-Lasso and alternative methods on a semi-empirical GWAS dataset mimicking population structure as found in *Arabidopsis thaliana*. (a) Receiver operating characteristics (ROC) for recovering simulated causal SNPs using alternative methods. Shown is the True Positive Rate (TPR) as a function of the False Positive Rate (FPR). (b) Alternative evaluation of each method on the identical dataset using Precision-Recall. Shown is the precision as a function of the recall.

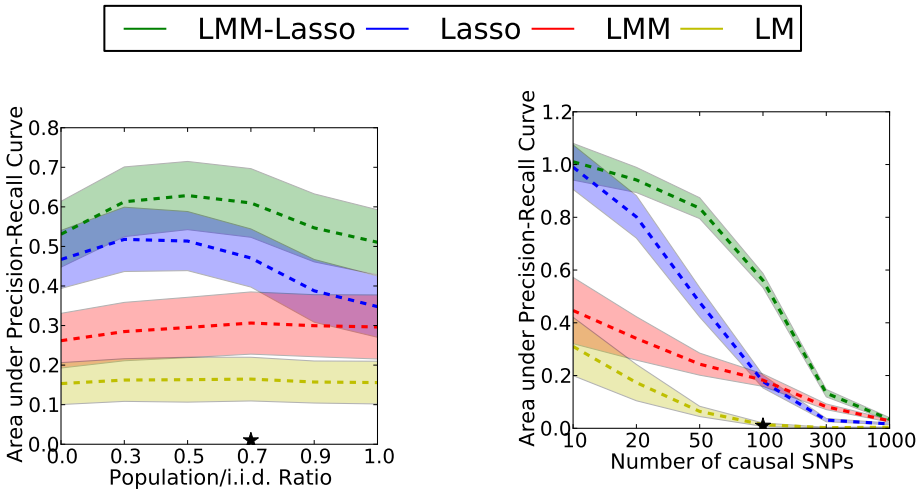


(a) Effect size vs. area under precision-recall curve



(b) Averaged neg. log likelihood vs. number of active SNPs

Figure 5.3. Evaluation of LMM-Lasso and alternative methods on semi-empirical GWAS dataset. (a) Area under the precision-recall curve as a function of the total effect size of all causal SNPs. (b) Averaged negative log-likelihood of the selected SNPs under the multivariate normal distribution $\mathcal{N}(\mathbf{0}; \boldsymbol{K})$ as a function of the number of SNPs that are active in the model. The smaller the negative log likelihood is, the more the SNPs are correlated with the population structure. For the LMM-Lasso and the Lasso active SNPs have been selected by following the regularization path. For linear mixed model (LMM) and linear model (LM), the set of active SNPs have been obtained in ascending order of the P value obtained. In the beginning, Lasso and the linear model choose SNPs that heavily reflect the population structure, while the mixed model approaches don't. In both figures the number of causal SNPs was 100.

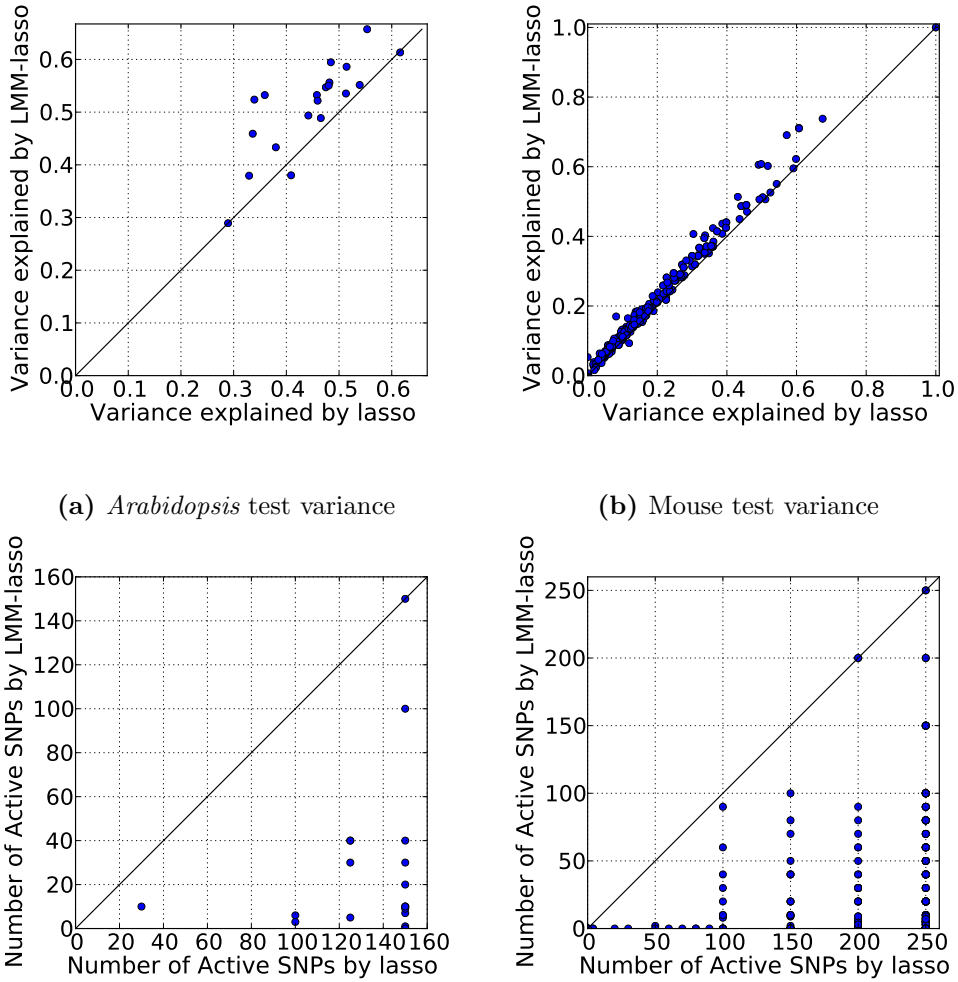


(a) Population structure strength

(b) Trait complexity: Varying Number of Causal SNPs

(c) Trait complexity: Varying signal strength

Figure 5.4. Evaluation of LMM-Lasso and alternative methods on semi-empirical GWAS dataset for different simulation settings. Area under precision-recall curve for finding the true simulated associations. Alternative simulation parameters have been varied in a chosen range. (a) Evaluation for different relative strength of population structure. (b) Evaluation for true simulated genetic models with increasing complexity (more causal SNPs). (c) Evaluation for variable signal to noise ratio.



(a) *Arabidopsis* test variance

(b) Mouse test variance

(c) *Arabidopsis* number of SNPs

(d) Mouse number of SNPs

Figure 5.5. Predictive power and sparsity of the fitted genetic models for Lasso and LMM-Lasso applied to quantitative traits in model systems. Considered were flowering phenotypes in *Arabidopsis thaliana* and bio-chemical and physiological phenotypes with relevance for human healthy profiled in mouse. Comparative evaluations include the fraction of phenotype variance predicted and the complexity of the fitted genetic model (number of active SNPs). (a) Explained variance in *Arabidopsis thaliana*. (b) Explained variance in mouse. (c) Complexity of fitted models in *Arabidopsis thaliana*. (d) Complexity of fitted models in mouse.

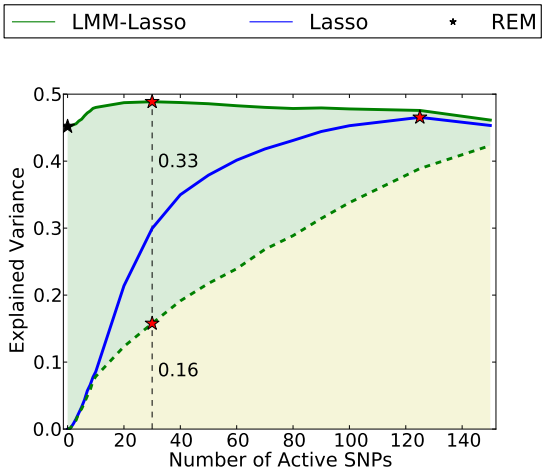
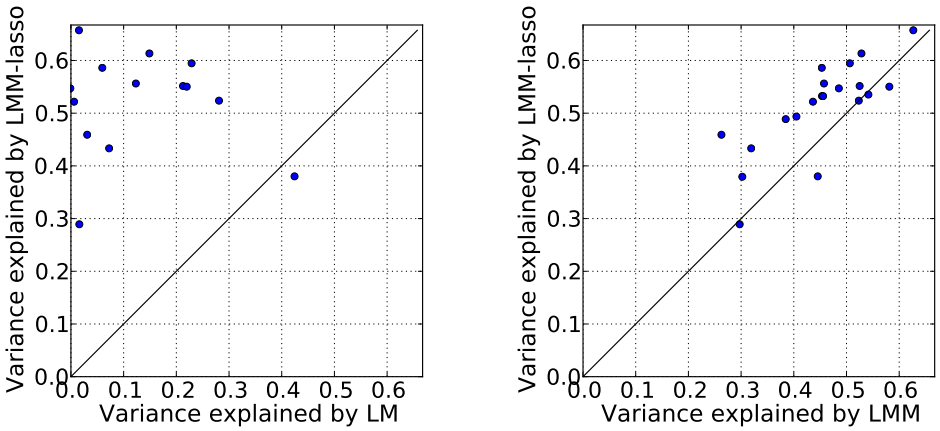


Figure 5.6. Variance dissection of *A. thaliana* flowering time into individual SNP effects and global genetic background driven by population structure using LMM-Lasso. Shown is the explained variance on an independent test set as a function of the number of active SNPs for the flowering phenotype (10° C) in *Arabidopsis thaliana*. In blue, the predictive test set variance of the Lasso as a function of the number of SNPs in the model. In green, the total predictive variance of LMM-Lasso for different sparsity levels. The shaded area indicates the fraction of variance LMM-Lasso explains by means of population structure (yellow) an population structure (green). LMM-Lasso without additional SNPs in the model corresponds to a genetic random effect model as in common usage (black star).



(a) LMM-Lasso vs. linear model.

(b) LMM-Lasso vs. linear mixed model.

Figure 5.7. Comparison of predictive power and sparsity obtained by LMM-Lasso and alternative methods on quantitative traits in *Arabidopsis thaliana*. Maximal explained variance on an independent test set (a) LMM-Lasso vs. linear model including the top associated SNP. (b) LMM-Lasso vs. linear mixed model including the top associated SNP.

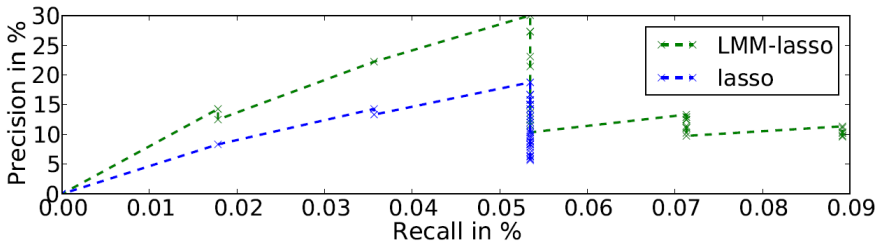


Figure 5.8. Precision-Recall curves for recovery of proximal SNPs for LMM-Lasso and Lasso on FLC gene expression in Arabidopsis thaliana. Precision-Recall Curve for recovering SNPs in proximity to known candidate genes using alternative methods. Shown is precision (TP/(TP+FP)) as a a function of the recall (TP/(TP+FN)). Each point in the plot corresponds to a specific selection threshold.

6. Linear mixed models for multiple related traits

Aiming at a holistic view of how genetics shapes the phenotype of an organism, it is already common practice to measure high dimensional phenotypes for a number of samples. Phenotype data spanning multiple omics can for example entail disease phenotypes, capture physiological measurements of each sample, molecular phenotypes, or expression levels of thousands of genes [Schadt et al., 2005, Bennett et al., 2010, Smith and Kruglyak, 2008, Psaty et al., 2009]. In this setting, the approach of testing for effects of single genetic variants on a single trait measurement will eventually reach its limits. To fully utilize such data, methods are needed, that take dependencies among phenotypic variables into account and allow to formulate and test complex composite hypotheses involving many phenotypes.

For example we may be interested in the effect of genetic regulators, that cause differences of the measured value for a multivariate phenotype or groups of related phenotypes, or to test for a *cis*-effect of a SNP while conditioning on broad effects of *trans*-regulators. Because of the size of genomes, the number of such hypotheses to be considered can be enormous and hence well powered and calibrated statistical tests are needed to make sense of these rich data [Spencer et al., 2009]. As in the case of univariate phenotypes considered so far in this thesis, samples are usually not independent, but rather get confounded by genetic structure or by hidden external influences.

To this end, we discuss multivariate approaches to data modeling and hypothesis testing and propose an accurate and efficient multivariate modeling framework that scales to thousands of traits.

The matrix-variate normal distribution, which underlies most of the multivariate modeling approaches discussed in this chapter, is introduced in Section 6.1.2. It is a distribution over matrices of real values, where the covariance between two entries of a matrix is given as the product between a covariance between their rows and a covariance between their columns. If the data matrix in the matrix-variate normal distribution is fully observed [1], it is possible to derive an efficient inference scheme called the "flip-flop algorithm". In this iterative approach, which we present in Section 6.1.2, estimation of the covariances on rows and columns of the matrix is decoupled efficiently by exploiting the fact, that the covariance factorizes into a Kronecker product[2] of the two respective covariances [Dutilleul, 1999, Zhang and Schneider, 2010]. Matrix-variate normal models have important applications in various fields. These models have been used as regularizer for multi-output prediction, jointly modeling the similarity between tasks and samples [Zhang and Schneider, 2010]. In related work in Gaussian processes, generaliza-

[1] This case is also called a *balanced design*. In contrast to the case, where some entries in the data matrix may be missing, which is called an *unbalanced design*

[2] See the definition in Equation (6.1).

tions of matrix-variate normal distributions have been used for inference of vector-valued functions [Bonilla et al., 2008, Alvarez and Lawrence, 2011]. Such models with Kronecker factored covariance have applications in geostatistics [Wackernagel, 2003], in collaborative filtering [Yu et al., 2009], multi-task prediction [Yu et al., 2005b, Bonilla et al., 2008], statistical testing on matrix-variate data [Allen and Tibshirani, 2010] and statistical genetics [Lynch and Walsh, 1998].

Attempts to make linear fixed effects models multivariate are reviewed in Section 6.1.3. These models, which have been developed for hypothesis testing in high-dimensional data, are are based on the matrix-variate normal distribution, efficient maximum-likelihood inference can be derived either in closed form or as variants of the "flip-flop algorithm".

When treating these multivariate fixed effects as random, the covariance of the marginalized likelihood becomes a sum of a random effects covariance as well as a noise covariance matrix and for this reason does not factor into a Kronecker product, even if the data matrix is fully observed. As efficient inference techniques similar to the "flip-flop algorithm" are inappropriate, inference scales cubic in the number of samples times cubic in the number of phenotypes. This severe runtime bottleneck already makes application to a handful of phenotypes hardly practical.

In theory it would be possible to make the simplifying modeling assumption that the data is directly given by a noise-free matrix-variate distribution. This has been used with some success, but there are clear motivations for using a model that includes additional noise. For example in a closely related multi-task regression setting, a noise-free matrix-variate Gaussian process leads to a cancelation of information sharing between the various prediction tasks [Bonilla et al., 2008]. This effect, also known from the geostatistics literature [Wackernagel, 2003], eliminates any benefit from multivariate modeling compared to naïve approaches.

In Section 6.2.1, we address these shortcomings and propose a general framework to scale random-effects models with *i.i.d.* observation noise to high-dimensional phenotype data, involving hundreds to thousands of phenotypes [Stegle et al., 2011, Yan et al., 2011]. Although in this model the covariance matrix does not factorize into a Kronecker product, we show how efficient parameter inference can still be done.

To this end, we provide derivations of both the log-likelihood and gradients with respect to hyperparameters that can be computed in the same asymptotic runtime as iterations of the "flip-flop algorithm" on a noise-free model. This allows for parameter learning of covariance matrices of size $10^5 \times 10^5$, or even bigger, which would not be possible if done naïvely.

We show how for any combination of covariances, evaluation of model likelihood and gradients with respect to individual covariance parameters is tractable.

Then, in Section 6.2.2 we apply this framework to structure learning in Gaussian graphical models, while accounting for a confounding non-*i.i.d.* sample structure. This generalization of the *Graphical Lasso* [Banerjee et al., 2008, Friedman et al., 2008] allows to jointly learn and account for a sparse inverse covariance matrix between features and a structured (non-diagonal) sample covariance. The low rank component of the sample covariance is used to account for confounding effects, as is done in other models for genomics [Leek and Storey, 2007, Stegle et al., 2010].

We illustrate this generalization called "Kronecker GLASSO" on synthetic datasets and heterogeneous protein signaling and gene expression data, where the aim is to recover the

hidden network structures. We show that our approach is able to recover the confounding structure, when it is known, and reveals sparse biological networks that are in better agreement with known components of the latent network structure.

After having introduced multivariate variants of both fixed-effects models that allow for hypothesis testing in multivariate scenarios as well as random-effects models that allow for modeling latent dependencies in the data, in Section 6.3.2 we combine these to achieve multivariate linear mixed models as a full generalization of univariate linear mixed models.

Bivariate mixed models have already been applied in GWAS to increase power to detect pleiotropic effects on correlated traits [Korte et al., 2012] as well as for the analysis of traits measured in different environments [Korte et al., 2012, Yang et al., 2011a].

Existing models have been proposed for the unbalanced case and for this reason come at a tremendous computational cost. We show in Section 6.3.2 for the balanced design case, how inference in these models can be done efficiently.

Finally, we give an outlook in Section 6.3.3 on how these matrix-variate linear mixed models will allow to phrase and test complex hypotheses involving a large number of target variables, by using appropriate row and column design matrices.

6.1. Simple multivariate identities and models

Here, we review some important Kronecker identities that are used throughout this Chapter and the matrix-variate normal distribution. We also introduce a number of multivariate models that rely on fixed effects only, that serve as a stepping stone for the development of the multivariate random and mixed models in this chapter.

6.1.1. Kronecker product identities

First we introduce some notation. For any $R \times C$ matrix $\boldsymbol{A}$, we define $\mathrm{vec}(\boldsymbol{A})$ to be the vector obtained by concatenating the columns of $\boldsymbol{A}$; further, let $\boldsymbol{A} \otimes \boldsymbol{B}$ denote the *Kronecker product*[3] between matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\mathrm{vec}(\boldsymbol{A}) = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{LM} \end{pmatrix}; \qquad \boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1C}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2C}\boldsymbol{B} \\ \dots & \dots & \dots & \vdots \\ a_{R1}\boldsymbol{B} & a_{R2}\boldsymbol{B} & \dots & a_{RC}\boldsymbol{B} \end{pmatrix}. \tag{6.1}$$

The first identity allows to write a product of a Kronecker product matrix with a vectorized matrix in terms of ordinary matrix products:

$$(\boldsymbol{C} \otimes \boldsymbol{B}^\top)\mathrm{vec}\,(\boldsymbol{A}) = \mathrm{vec}(\boldsymbol{BAC}). \tag{6.2}$$

Also the determinant of the Kronecker product of two full-rank matrices $\boldsymbol{C} \in \mathbb{R}^{M \times M}$ and $\boldsymbol{R} \in \mathbb{R}^{N \times N}$ can be written as the product of the determinants of the individual matrices times the rank

$$|\boldsymbol{C} \otimes \boldsymbol{R}| = |\boldsymbol{C}|^N \cdot |\boldsymbol{R}|^M. \tag{6.3}$$

[3] or *tensor product*

$$(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}) = (\boldsymbol{U_C} \otimes \boldsymbol{U_R})(\boldsymbol{\Lambda_C} \otimes \boldsymbol{\Lambda_R} + \sigma^2 \boldsymbol{I})(\boldsymbol{U_C}^\top \otimes \boldsymbol{U_R}^\top), \tag{6.4}$$

where $\boldsymbol{C} = \boldsymbol{U_C} \boldsymbol{\Lambda_C} \boldsymbol{U_C}^\top$ is the spectral decomposition of $\boldsymbol{C}$, and similarly for $\boldsymbol{R}$.

On a related note, singular value decompositions and Kronecker product identities were also used for efficient covariance computation in graph kernel research [Vishwanathan et al., 2010].

6.1.2. The matrix-variate normal distribution

The matrix-variate normal distribution[4] is a distribution over matrices of real values, where the covariance between two entries of a matrix is given as the product between a covariance between their rows and a covariance between their columns.

We say, that an N-by-G matrix $\boldsymbol{Y}$ with N rows and G columns follows a matrix-variate normal distribution with mean $\boldsymbol{M}$, row covariance $\boldsymbol{R}$ and column covariance $\boldsymbol{C}$,

$$\boldsymbol{Y} \sim \mathcal{N}_{NM}\left(\boldsymbol{M}\,;\, \boldsymbol{R}\,,\, \boldsymbol{C}\right), \tag{6.5}$$

iff the vectorized matrix $\mathrm{vec}\,(Y)$ follows a multivariate normal distribution with a covariance matrix, that consists of the Kronecker product between $\boldsymbol{C}$ and $\boldsymbol{R}$.

$$\mathrm{vec}\,(\boldsymbol{Y}) \sim \mathcal{N}\left(\mathrm{vec}\,(\boldsymbol{M})\,;\, \boldsymbol{C} \otimes \boldsymbol{R}\right). \tag{6.6}$$

In these expressions the row covariance $\boldsymbol{R}$ is a symmetric positive semi-definite matrix of size N-by-N, the column covariance $\boldsymbol{C}$ is a symmetric positive semi-definite matrix of size G-by-G, and the mean $\boldsymbol{M}$ is of size N-by-G.

In order to find a suitable form for the matrix-variate normal distribution, we write out the expression for the multivariate normal density (6.6) as

$$(2\pi)^{-\frac{N \cdot D}{2}} \cdot |\boldsymbol{C} \otimes \boldsymbol{R}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}\mathrm{vec}\,(\boldsymbol{Y}_\mathrm{r})^\top \,(\boldsymbol{C} \otimes \boldsymbol{R})^{-1}\,\mathrm{vec}\,(\boldsymbol{Y}_\mathrm{r})^\top\right),$$

where $\boldsymbol{Y}_\mathrm{r}$ equals the residuals of the data $\boldsymbol{Y}$ after subtracting the mean $\boldsymbol{M}$.

$$\boldsymbol{Y}_\mathrm{r} = (\boldsymbol{Y} - \boldsymbol{M})$$

To achieve a form of the matrix-variate normal distribution, that allows for efficient evaluation, one can exploit identity (6.2) for the squared form and identity (6.3) for the determinant to get

$$(2\pi)^{-\frac{N \cdot M}{2}} \cdot |\boldsymbol{C}|^{-\frac{N}{2}} \cdot |\boldsymbol{R}|^{-\frac{M}{2}} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{Y}_\mathrm{r}^\top \boldsymbol{R}^{-1}\boldsymbol{Y}_\mathrm{r}\right)\right).$$

Also the logarithm of the matrix-variate normal density can be evaluated without explicitly computing the Kronecker product as

$$-\frac{N \cdot M}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{C}| - \frac{M}{2}\log|\boldsymbol{R}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{Y}_\mathrm{r}^\top \boldsymbol{R}^{-1}\boldsymbol{Y}_\mathrm{r}\right). \tag{6.7}$$

[4] For a comprehensive review of this family of distributions see for example [Dawid, 1981].

Maximum likelihood estimation

We assume in the following, that the mean $\boldsymbol{M}$ of the matrix-variate normal distribution is known and given in advance. A sensible estimate of the mean in the matrix-variate normal distribution requires additional modeling assumptions, like in the examples discussed in Section 6.1.3. First, free-form estimation of the covariance matrix is reviewed, which can be performed using closed form updates. Second, gradients are provided for the case, where the covariance matrices are given as the function of a set of hyperparameters.

Maximum likelihood estimation using the "flip-flop algorithm"

In case of free-form estimation of the covariance matrix, the "flip-flop algorithm" can be used [Dutilleul, 1999]. The central insight that leads to the algorithm is, that instead of setting the gradient of the logarithm of the matrix-variate normal density with respect to all entries of the covariance matrix to be equal to zero, we can get an equivalent system of equations by setting the gradients with respect to entries of its inverse to zero. The equation we have to solve with respect to the (i, j)-th entry of the inverse of $\boldsymbol{R}$ equals

$$0 = \frac{M}{2}\mathrm{tr}\left(\boldsymbol{R}\frac{\boldsymbol{R}^{-1}}{\partial[\boldsymbol{R}^{-1}]_{i,j}}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{Y}_{\mathrm{r}}^{\top}\frac{\boldsymbol{R}^{-1}}{\partial[\boldsymbol{R}^{-1}]_{i,j}}\boldsymbol{Y}_{\mathrm{r}}\right).$$

As the derivative of the inverse with respect to the (i, j)-th entry equals a matrix of only zeros, except for the (i, j)-th entry, which equals one, this expression simplifies to

$$0 = \frac{M}{2}[\boldsymbol{R}]_{i,j} - \frac{1}{2}[\boldsymbol{Y}_{\mathrm{r}}]_{j,:}\boldsymbol{C}^{-1}[\boldsymbol{Y}_{\mathrm{r}}^{\top}]_{:,i}.$$

This can easily be solved for all i and j to obtain the maximum likelihood estimate $\boldsymbol{R}$ as

$$\hat{\boldsymbol{R}} = \frac{1}{M}\boldsymbol{Y}_{\mathrm{r}}\boldsymbol{C}^{-1}\boldsymbol{Y}_{\mathrm{r}}^{\top}.$$

An analogous estimator follows for the column covariance as

$$\hat{\boldsymbol{C}} = \frac{1}{N}\boldsymbol{Y}_{\mathrm{r}}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Y}_{\mathrm{r}}. \tag{6.8}$$

The "flip-flop algorithm" now iterates between computing these two estimators, where the next estimate of $\boldsymbol{R}$ is computed from data that is transformed using the current estimate of $\boldsymbol{C}$ and vice versa, until convergence is achieved[5]. One round of these updates can be evaluated in $O(N^3 + M^3)$.

The algorithm can be modified to include several types of regularization terms. For example for the case, where the entries of the inverse of the covariance matrices are penalized by their absolute values, the algorithm involves a similar iterative procedure, where closed form updates are replaced by solving graphical Lasso problems on similarly transformed data [Zhang and Schneider, 2010].

[5] Note, that the algorithm as stated above assumes that both matrices are positive definite. In practice, it is possible to constrain the matrices to be positive definite, by adding a small positive constant to the diagonal when performing an update.

Maximum-likelihood estimation by gradient descent

If the matrices $\boldsymbol{C}$ and $\boldsymbol{R}$ are given as the function of hyperparameters $\boldsymbol{\Theta}_{\boldsymbol{C}}$ and $\boldsymbol{\Theta}_{\boldsymbol{R}}$ respectively, we can revert to gradient based optimizers. For this alternative procedure we need to take the derivative with respect to each parameter $\theta_{\boldsymbol{R}} \in \boldsymbol{\Theta}_{\boldsymbol{R}}$ and $\theta_{\boldsymbol{C}} \in \boldsymbol{\Theta}_{\boldsymbol{C}}$.

$$\frac{\partial \log \mathcal{N}_{NM}\left(\boldsymbol{Y} \,\middle|\, \boldsymbol{M}\,;\, \boldsymbol{R}\,,\, \boldsymbol{C}\right)}{\partial \theta_{\boldsymbol{R}}} = -\frac{M}{2}\mathrm{tr}\left(\boldsymbol{R}^{-1}\frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{Y}_{\mathrm{r}}^{\top}\boldsymbol{R}^{-1}\frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}}\boldsymbol{R}^{-1}\boldsymbol{Y}_{\mathrm{r}}\right).$$

The derivative with respect to a parameter $\theta_{\boldsymbol{C}} \in \boldsymbol{\Theta}_{\boldsymbol{C}}$ is analogous:

$$\frac{\partial \log \mathcal{N}_{NM}\left(\boldsymbol{Y} \,\middle|\, \boldsymbol{M}\,;\, \boldsymbol{R}\,,\, \boldsymbol{C}\right)}{\partial \theta_{\boldsymbol{C}}} = -\frac{N}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_{\boldsymbol{C}}}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_{\boldsymbol{C}}}\boldsymbol{C}^{-1}\boldsymbol{Y}_{\mathrm{r}}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Y}_{\mathrm{r}}\right).$$

Both of these derivatives, as well as the logarithm of the matrix-variate likelihood can be evaluated in $O(N^3 + G^3)$.

6.1.3. Existing multivariate linear fixed effects models

Here, we give an overview over multivariate linear fixed effects models. All the models presented here, are generalizations of the linear regression model presented in Section 2.1 to the case, where more than one target variable is considered.

In the following, let the N-by-G matrix of observations $\boldsymbol{Y}$ be defined as

$$\boldsymbol{Y} = \begin{bmatrix} | & | & & | \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_M \\ | & | & & | \end{bmatrix}, \tag{6.9}$$

where each of the vectors $\boldsymbol{y}_g$ contains the set of N samples for the g-th target variable or dimension. For example in the case of a study of gene expression, we might have measured the expression values of G genes (variables, dimensions) for N individuals (samples).

In linear regression, as introduced in Section 2.1, each $\boldsymbol{y}_g$ would be modeled by an independent linear model

$$\boldsymbol{y}_g = \boldsymbol{X}\boldsymbol{\beta}_g + \boldsymbol{e}_g, \tag{6.10}$$

where the D-by-1 vector $\boldsymbol{\beta}_g$ is the vector of unknown fixed effects for the g-th regression and the N-by-D matrix $\boldsymbol{X}$ is a fixed design matrix, that is a shared over all target dimensions, and $\boldsymbol{e}_g$ is normally distributed observation noise with unknown variance $\sigma^2_{\boldsymbol{e}_g}$.

In principle, it is possible to write down a joint likelihood over the individual likelihoods of G linear regressions.

$$\mathcal{L}\left(\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G\}, \{\sigma^2_{\boldsymbol{e}_1}, \ldots, \sigma^2_{\boldsymbol{e}_G}\}\right) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_g \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}_g\,;\, \sigma^2_{\boldsymbol{e}_g}\right)$$

In order to find a more intuitive expression of the joint likelihood, we introduce the D-by-G matrix $\boldsymbol{B}$, where the g-th column contains the fixed effects $\boldsymbol{\beta}_g$ of the g-th linear model.

$$\boldsymbol{B} = \begin{bmatrix} | & | & & | \\ \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_G \\ | & | & & | \end{bmatrix}$$

Then, we can jointly write down the linear regression models in terms of a matrix-variate normal distribution on $\boldsymbol{Y}$ with mean $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{B}$.

$$\boldsymbol{Y} \sim \mathcal{N}_{NM}\left(\boldsymbol{X}\boldsymbol{B}\,;\, \boldsymbol{I}_N\,,\, \operatorname{diag}\left(\begin{array}{cccc} \sigma^2_{e_1}, & \sigma^2_{e_2}, & \ldots, & \sigma^2_{e_G} \end{array}\right)\right). \tag{6.11}$$

As in many applications the values obtained for different dimensions are dependent, the approach of performing G independent linear regressions seems wasteful. Below, we give a review of generalizations of the linear regression model that allow for joint modeling of fixed effects in the multivariate case.

The MANOVA model

Given a sufficient number of samples with measurements for each of the target dimensions, it is possible to estimate covariances between the observations. This is used in the multivariate analysis of variance MANOVA model [Roy, 1957] to jointly estimate the linear regression coefficients $\boldsymbol{\beta}_g$ for each of the target dimensions together with a full G-by-G covariance matrix $\boldsymbol{C}$ between the target dimensions.

Using the definition of the matrix-variate normal distribution, we can write down the MANOVA model as

$$\boldsymbol{Y} \sim \mathcal{N}_{NM}\left(\boldsymbol{X}\boldsymbol{B}\,;\, \boldsymbol{I}_N\,,\, \boldsymbol{C}\right). \tag{6.12}$$

While the covariance estimates get coupled by $\boldsymbol{C}$, each dimension is still modeled by a separate univariate regression model $\boldsymbol{X}\boldsymbol{\beta}_g$.

Maximum likelihood estimation Inference in the MANOVA model can be performed by maximizing the likelihood with respect to the parameters $\boldsymbol{B}$ and $\boldsymbol{C}$. Taking the derivative of the likelihood with respect to each entry $[\boldsymbol{B}]_{d,g}$ of the weight matrix and jointly setting these to zero, we get the system of equations

$$\boldsymbol{0} = \boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{C}^{-1} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\boldsymbol{C}^{-1}. \tag{6.13}$$

Assuming that $\boldsymbol{C}$ is full rank, this can be right-multiplied by $\boldsymbol{C}$ and solved for an expression for $\boldsymbol{B}_\mathrm{M}$ independent of $\boldsymbol{C}$ and equals the maximum likelihood estimators of G independent linear regression models:

$$\boldsymbol{B}_\mathrm{M} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}. \tag{6.14}$$

Plugging this estimator back into the likelihood, we get the residual $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$, which we can use in Equation (6.8) to obtain $\boldsymbol{C}_\mathrm{M}$ as

$$\boldsymbol{C}_\mathrm{M} = \frac{1}{N}\boldsymbol{Y}^\top\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{Y} \tag{6.15}$$

Seemingly unrelated regressions

A slightly more general model than MANOVA, is the seemingly unrelated regressions model [Zellner, 1962, 1963]. In the case of seemingly unrelated regressions, each target dimension g is allowed to have a distinct design matrix $\boldsymbol{X}_g$ of size N-by-D_g. As the term

in the mean of the likelihood can not be written as a Kronecker product, the likelihood is written in terms of a specially structured general linear model on the vectorized matrix of target values $\mathrm{vec}(\boldsymbol{Y})$. Define

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{X}_G \end{bmatrix} \tag{6.16}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_G \end{bmatrix} \tag{6.17}$$

Then the likelihood of the seemingly unrelated regressions model becomes

$$\mathrm{vec}(\boldsymbol{Y}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{C} \otimes \boldsymbol{I}_N). \tag{6.18}$$

Estimation in seemingly unrelated regressions Even though the seemingly unrelated regressions model is very similar to MANOVA, estimation of the fixed effects is not independent for each dimension. To overcome the problem, that dimensions get coupled by the unknown covariance matrix $\boldsymbol{C}$, a two-step procedure is applied for estimation in seemingly unrelated regressions [Zellner, 1962]. First, an estimate $\hat{\boldsymbol{C}}$ of the covariance matrix $\boldsymbol{C}$ is estimated based on the model with the maximum likelihood estimators of independent regressions on the target dimensions.

$$\hat{\boldsymbol{y}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{y}_g. \tag{6.19}$$

Using this procedure, the estimate $\hat{\boldsymbol{C}}$ is

$$\hat{\boldsymbol{C}} = \frac{1}{N} \begin{bmatrix} | & | & & | \\ \hat{\boldsymbol{y}}_1 & \hat{\boldsymbol{y}}_2 & \dots & \hat{\boldsymbol{y}}_G \\ | & | & & | \end{bmatrix}^\top \begin{bmatrix} | & | & & | \\ \hat{\boldsymbol{y}}_1 & \hat{\boldsymbol{y}}_2 & \dots & \hat{\boldsymbol{y}}_G \\ | & | & & | \end{bmatrix}. \tag{6.20}$$

The growth curve model

In MANOVA, each target dimension is modeled by a separate univariate vector of fixed effects $\boldsymbol{\beta}_g$. Even though the dimensions get coupled during estimation by ways of the covariance matrix $\boldsymbol{C}$, we have observed, that the maximum likelihood estimators of the fixed effects is identical to the G maximum likelihood estimators in the univariate regression models. The growth curve model [Potthoff and Roy, 1964] achieves multivariate fixed effects by replacing the linear model in MANOVA by a bilinear model

$$\boldsymbol{Y} \sim \mathcal{N}_{NM}\left(\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top; \boldsymbol{I}_N, \boldsymbol{C}\right), \tag{6.21}$$

where $\boldsymbol{C}$ is an unknown G-by-G covariance matrix between target dimensions, $\boldsymbol{B}$ is a D-by-M matrix of unknown fixed effect weights, and the matrices $\boldsymbol{X}$ and $\boldsymbol{A}$ are design

matrices. To avoid ambiguity, the N-by-D matrix $\boldsymbol{X}$ is typically referred to as the *between-individuals design matrix* and the M-by-G matrix $\boldsymbol{A}$ is typically referred to as the *within-individuals design matrix* [Kollo and von Rosen, 2005].

Intuitively, the between-individuals design matrix contains a set of features, say for example a set of SNPs or a treatment condition, that differs between individuals. The features defined in the within-individuals design matrix contains are the same over all individuals, but differ between the various target dimensions, a feature might for example indicate, which genes are regulated by a common transcription factor. Note, that the growth-curve model generalizes the MANOVA model, as can be seen by choosing the within-individuals design matrix to be the G-dimensional identity matrix.

Maximum likelihood estimation By a conditional argument on correlated least squares estimators with unknown covariance [Rao, 1967], the maximum likelihood estimator $\boldsymbol{B}_{\mathrm{M}}$ of $\boldsymbol{B}$ and $\boldsymbol{C}_{\mathrm{M}}$ of $\boldsymbol{C}$ in the growth curve model can be derived [Rao, 1965]. $\boldsymbol{B}$ follows as

$$\boldsymbol{B}_{\mathrm{M}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \hat{\boldsymbol{C}}^{-1} \boldsymbol{A} \left(\boldsymbol{A}^\top \hat{\boldsymbol{C}}^{-1} \boldsymbol{A}\right)^{-1}, \tag{6.22}$$

where

$$\hat{\boldsymbol{C}} = \frac{1}{N} \boldsymbol{Y}^\top \left(\boldsymbol{I} - \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right) \boldsymbol{Y}. \tag{6.23}$$

Using $\boldsymbol{B}_{\mathrm{M}}$, the maximum-likelihood estimator $\boldsymbol{C}_{\mathrm{M}}$ of $\boldsymbol{C}$ is determined as

$$\boldsymbol{C}_{\mathrm{M}} = \hat{\boldsymbol{C}} + \frac{1}{N} \boldsymbol{P}^\top \boldsymbol{Y}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \boldsymbol{P}, \tag{6.24}$$

where

$$\boldsymbol{P} = \boldsymbol{I} - \hat{\boldsymbol{C}}^{-1} \boldsymbol{A} \left(\boldsymbol{A}^\top \hat{\boldsymbol{C}}^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top. \tag{6.25}$$

The sum of profiles model

Even though the growth curve model is already quite general, it still has a strong limitation in fixed-effects modeling. This limitation is, that there is an interaction weight $[\boldsymbol{B}]_{d,m}$ for all combinations of features d in the between-individuals design matrix and all features m in the within-individuals design matrix. In order to incorporate prior knowledge on features that are known to exert no interaction, one would have to constrain the respective entries in the fixed effects matrix $\boldsymbol{B}$ to zero, by introducing additional linear constraints. The sum of profiles model [Verbyla and Venables, 1988] allows to explicitly phrase such knowledge about what features interact (and more importantly on what features do not interact) by replacing the single bilinear term in the growth curve model by the sum over J bilinear terms, where each term might involve fixed effects of different dimensionality.

$$\boldsymbol{Y} \sim \mathcal{N}_{NM} \left(\sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j \,;\, \boldsymbol{I}_N \,,\, \boldsymbol{C} \right), \tag{6.26}$$

with each of J fixed unknown effects matrices $\boldsymbol{B}_j$ of sizes D_j-by-M_j and between-individuals design matrices $\boldsymbol{X}_j$ of sizes N-by-D_j and the within-individuals design matrices of sizes M_j-by-G. Again, the G-by-G covariance matrix between target dimensions is considered unknown.

Estimation in the sum of profiles model While for special cases of the model closed form maximum likelihood estimation is possible [Von Rosen, 1989], the standard way to perform maximum likelihood estimation in the general sum of profiles model is to perform an algebraic reduction to an equivalent seemingly unrelated regressions model [Verbyla and Venables, 1988, Kollo and von Rosen, 2005].

Multivariate linear models with Kronecker product covariance structure

In all the multivariate fixed-effects models considered so far, the samples have been assumed to be independent. The multivariate linear model with Kronecker product covariance structure [Srivastava et al., 2008, 2009] is a generalization of the growth curve model to the case, where not only the target dimensions are dependent, but also the samples may covary.

$$\boldsymbol{Y} \sim \mathcal{N}_{NM}\left(\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{\top}; \boldsymbol{R}, \boldsymbol{C}\right), \tag{6.27}$$

where as in the growth curve model in Equation (6.21) the between-individuals design matrix $\boldsymbol{X}$ and the within-individuals design matrix $\boldsymbol{A}$ are given. The matrices to be estimated are the fixed effects weights $\boldsymbol{B}$, the N-by-N sample covariance matrix $\boldsymbol{R}$, and the covariance between target dimensions $\boldsymbol{C}$.

Maximum likelihood estimation Maximum-likelihood estimators for the multivariate linear model with Kronecker product structure can be found by a variant of the "flip-flop-algorithm" [Srivastava et al., 2009]. The maximum-likelihood estimate $\boldsymbol{B}_{\mathrm{M}}$ of the fixed-effects is given based on the current estimate $\boldsymbol{C}_{\mathrm{M}}$ of the covariance between target dimensions.

$$\boldsymbol{B}_{\mathrm{M}} = \left(\boldsymbol{X}^{\top}\boldsymbol{S}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{S}^{-1}\boldsymbol{Y}{\boldsymbol{C}_{\mathrm{M}}}^{-1}\boldsymbol{A}\left(\boldsymbol{A}^{\top}{\boldsymbol{C}_{\mathrm{M}}}^{-1}\boldsymbol{A}\right)^{-1}, \tag{6.28}$$

where

$$\boldsymbol{S} = \boldsymbol{Y}\left({\boldsymbol{C}_{\mathrm{M}}}^{-1} - {\boldsymbol{C}_{\mathrm{M}}}^{-1}\boldsymbol{A}\left(\boldsymbol{A}^{\top}{\boldsymbol{C}_{\mathrm{M}}}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\top}{\boldsymbol{C}_{\mathrm{M}}}^{-1}\right)\boldsymbol{Y}^{\top}, \tag{6.29}$$

The maximum-likelihood estimator of the covariance between dimensions is given as a function of the current estimate $\boldsymbol{R}_{\mathrm{M}}$ of the samples covariance.

$$\boldsymbol{C}_{\mathrm{M}} = \frac{1}{N}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_{\mathrm{M}}\boldsymbol{A}^{\top}\right)^{\top}{\boldsymbol{R}_{\mathrm{M}}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_{\mathrm{M}}\boldsymbol{A}^{\top}\right) \tag{6.30}$$

Finally, an estimate $\boldsymbol{R}_{\mathrm{M}}$ of the sample-covariance is given as a function of the current estimates of the fixed effects and the covariance between target dimensions.

$$\boldsymbol{R}_{\mathrm{M}} = \frac{1}{G}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_{\mathrm{M}}\boldsymbol{A}^{\top}\right){\boldsymbol{C}_{\mathrm{M}}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_{\mathrm{M}}\boldsymbol{A}^{\top}\right)^{\top} \tag{6.31}$$

6.2. Efficient multivariate random effects models

Assume we are given a data matrix $\boldsymbol{Y}$ with N rows and G columns, where N is the number of samples with G features each. As an example, think of N as a number of samples in a micro-array experiment, where in each sample the expression levels of the

same G genes are measured; here, $y_{r,c}$ would be the expression level of gene $c \in [1, \ldots, G]$ in experiment $r \in [1, \ldots, N]$.

For modeling $\boldsymbol{Y}$, we first introduce N-by-G additional random effects $\boldsymbol{Z}$, which can be thought of as the noise-free observations. The data $\boldsymbol{Y}$ is then given by $\boldsymbol{Z}$ plus *i.i.d.* Gaussian observation noise. The likelihood function follows as

$$\mathcal{L}\left(\boldsymbol{Z}, \sigma^2\right) = \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \operatorname{vec}(\boldsymbol{Z})\,;\, \sigma^2 \boldsymbol{I}_{N \cdot G}\right).$$

In order to model the dependence structure between the observed values, we assume that the matrix of random-effects $\boldsymbol{Z}$ follows a matrix-variate normal distribution given by Equation (6.6) with zero mean.

$$\int \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \operatorname{vec}(\boldsymbol{Z})\,;\, \sigma^2 \boldsymbol{I}_{N \cdot D}\right) \mathcal{N}_{NM}\left(\boldsymbol{Z} \,\middle|\, \boldsymbol{0}_{N \cdot G}\,;\, \boldsymbol{R}, \boldsymbol{C}\right) \mathrm{d}\boldsymbol{Z}.$$

Marginalizing over the noise-free observations $\boldsymbol{Z}$ results in the marginal likelihood of the observed data $\boldsymbol{Y}$

$$\mathcal{L}\left(\boldsymbol{C}, \boldsymbol{R}, \sigma^2\right) = \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \boldsymbol{0}_{N \cdot G}\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}_{N \cdot G}\right). \tag{6.32}$$

We observe, that random observation noise causes addition of a constant diagonal term to the covariance matrix. For this reason the covariance no longer factorizes into a Kronecker product[6], rendering the algorithms presented in Section 6.1.2 inapplicable. Naive inference on the other would require computation and storage of a large $(N \cdot G)$-by-$(N \cdot G)$ covariance matrix and perform operations that have a runtime in the order of $O((N \cdot G)^3)$. In practice this already prohibits application of the model to moderate sizes of data.

6.2.1. Efficient parameter estimation in multivariate random-effects models

We achieve efficient evaluation and parameter estimation of the multivariate random effects model given in Equation (6.32) by exploiting the compatibility of a Kronecker product plus a constant diagonal term with the spectral decomposition (see Equation (6.4)).

Likelihood evaluation

As shown in Section E.2.1, using this identity together with the identities in Equations (6.2) and (6.2), the logarithm of the likelihood in Equation (6.32) follows as

$$\begin{aligned}\log \mathcal{L}\left(\boldsymbol{C}, \boldsymbol{R}, \sigma^2\right) = &- \frac{N \cdot G}{2} \ln(2\pi) - \frac{1}{2} \ln \left| \boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I} \right| \\ &- \frac{1}{2} \operatorname{vec}(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}})^\top (\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1} \operatorname{vec}(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}}).\end{aligned} \tag{6.33}$$

Analogous to the FaST-LMM algorithm presented in Section 3.3, this term can be interpreted as a multivariate normal distribution with a diagonal covariance matrix on rotated data

$$\mathcal{L}\left(\boldsymbol{C}, \boldsymbol{R}, \sigma^2\right) = \mathcal{N}\Big(\operatorname{vec}(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}}) \,\Big|\, \boldsymbol{0}\,;\, (\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})\Big). \tag{6.34}$$

[6] Note that for $\sigma^2 = 0$, the likelihood model in Equation (6.32) reduces to the matrix-variate normal distribution in Equation (6.6).

Gradient evaluation

Derivatives of the log marginal likelihood with respect to a particular covariance parameter $\theta_{\boldsymbol{R}} \in \boldsymbol{\Theta}_{\boldsymbol{R}}$ can be expressed as

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\boldsymbol{C}, \boldsymbol{R}, \sigma^2\right)}{\partial \theta_{\boldsymbol{R}}} = &-\frac{1}{2}\operatorname{diag}\Big(\left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right)^{-1}\Big)^{\top} \operatorname{diag}\Big(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes (\boldsymbol{U}_{\boldsymbol{R}}^{\top} \frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}} \boldsymbol{U}_{\boldsymbol{R}})\Big) \\ &+ \frac{1}{2}\operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot (\boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}})\right)^{\top} \operatorname{vec}\Big(\boldsymbol{U}_{\boldsymbol{R}}^{\top} \frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}} \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot (\boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}})\right) \boldsymbol{\Lambda}_{\boldsymbol{C}}\Big),\end{aligned} \tag{6.35}$$

where the entries of the N-times-G matrix $\boldsymbol{D}_{\boldsymbol{\Lambda}}$ are defined as

$$[\boldsymbol{D}_{\boldsymbol{\Lambda}}]_{r,j} = \frac{1}{[\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c}[\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} + \sigma^2}.$$

Analogous expressions follow for partial derivatives with respect to $\theta_{\boldsymbol{C}} \in \boldsymbol{\Theta}_{\boldsymbol{C}}$ and the noise level σ^2. Full details of all derivations can be found in Section E.2.

Runtime and memory complexity

A naïve implementation for optimizing the likelihood (6.32) with respect to the hyperparameters would have runtime complexity $\mathcal{O}(N^3 \cdot G^3)$ and memory complexity $\mathcal{O}(N^2 \cdot G^2)$. Using the likelihood and derivative as expressed in Equations (6.33) and (6.35), each evaluation with new kernel parameters involves solving the symmetric eigenvalue problems of both $\boldsymbol{R}$ and $\boldsymbol{C}$, together having a runtime complexity of $\mathcal{O}(N^3 + G^3)$. Explicit evaluation of any matrix Kronecker products is not necessary, resulting in a low memory complexity of $\mathcal{O}(N^2 + G^2)$.

6.2.2. Graphical Lasso in the presence of confounders

Estimation of sparse inverse covariance matrices is widely used to identify undirected network structures from observational data. However, non-*i.i.d.* observations due to hidden confounding variables may hinder accurate recovery of the true network structure. If not accounted for, confounders may lead to a large number of false positive edges. This is of particular relevance in biological applications, where observational data are often heterogeneous, combining measurements from different labs, data obtained under various perturbations or from a range of measurement platforms.

As an application of the random-effects model described in Section 6.2.1, in Section 6.2.2 we propose an approach to learning sparse inverse covariance matrices between features, while accounting for covariation between samples due to confounders. First, we briefly review the "orthogonal" approaches that account for the corresponding types of sample and feature covariance we set out to model.

Explaining feature dependencies using the Graphical Lasso

A common approach to model relationships between variables in a graphical model is the graphical Lasso. It has been used in the context of biological studies to recover the hidden network structure of gene-gene interrelationships, for instance [Menéndez et al.,

2010]. The graphical Lasso assumes a multivariate Gaussian distribution on features with a sparse precision (inverse covariance) matrix. The sparsity is induced by an L_1 penalty on the entries of $\boldsymbol{C}^{-1}$, the inverse of the feature covariance matrix.

Under the simplifying assumption of *i.i.d.* samples, the posterior distribution of $\boldsymbol{C}$ under this model is proportional to the joint distribution of $\boldsymbol{C}$ and the data $\boldsymbol{Y}$

$$p\left(\boldsymbol{Y}, \boldsymbol{C}\right) = p\left(\boldsymbol{C}\right) \prod_{r=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{r,:} \,\middle|\, \boldsymbol{0}_D \,;\, \boldsymbol{C}\right). \tag{6.36}$$

Here, the prior is defined in terms of the precision matrix $\boldsymbol{C}^{-1}$.

$$p\left(\boldsymbol{C}\right) \propto \exp\left(-\eta \left\|\boldsymbol{C}^{-1}\right\|_1\right) \;\; [\boldsymbol{C}^{-1} \succ \boldsymbol{0}], \tag{6.37}$$

with $\|\boldsymbol{A}\|_1$ defined as the sum over all absolute values of the matrix entries. Note that this prior is only nonzero for positive-definite matrices $\boldsymbol{C}^{-1}$.

Modeling confounders using the Gaussian process latent variable model

Confounders are unobserved variables that can lead to spurious associations between observed variables and to covariation between samples. A possible approach to identify such confounders is dimensionality reduction. Here we briefly review two dimensionality reduction methods, (dual) probabilistic principal components analysis and its generalization, the Gaussian process latent variable model [Lawrence, 2004, 2005]. In the context of applications, these methods have previously been applied to identify regulatory processes [Yeung and Ruzzo, 2001], and to recover confounding factors with broad effects on many features [Leek and Storey, 2007, Stegle et al., 2010].

In dual probabilistic principal components analysis [Lawrence, 2005], the observed data $\boldsymbol{Y}$ is explained as a linear combination of k latent variables ("factors"), plus independent observation noise. The model is as follows:

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{W} + \boldsymbol{E},$$

where $\boldsymbol{G} \in \mathbb{R}^{N \times k}$ contains the values of k latent variables ("factors"), $\boldsymbol{W} \in \mathbb{R}^{k \times M}$ contains independent standard-normally distributed weights that specify the mapping between latent and observed variables. Finally, $\boldsymbol{E} \in \mathbb{R}^{N \times G}$ contains *i.i.d.* Gaussian noise with $E_{rc} \sim \mathcal{N}\left(N \,;\, 0\right)\sigma^2$. Marginalizing over the weights $\boldsymbol{W}$ yields the likelihood as a function of $\boldsymbol{G}$:

$$\mathcal{L}\left(\boldsymbol{G}\right) = \prod_{c=1}^{G} \mathcal{N}\left(\boldsymbol{Y}_{:,c} \,;\, \boldsymbol{0}_N, \boldsymbol{G}\boldsymbol{G}^\top + \sigma^2 \boldsymbol{I}_N\right). \tag{6.38}$$

Learning the latent factors $\boldsymbol{G}$ and the observation noise variance σ^2 can be done by maximum likelihood. The more general gaussian process latent variable model [Lawrence, 2005] is obtained by replacing $\boldsymbol{G}\boldsymbol{G}^\top$ in (6.38) with a more general Gram matrix $\boldsymbol{R}$, with $R_{r,s} = \kappa\big((g_{r,1}, \ldots, g_{r,k}), (g_{s,1}, \ldots, g_{s,k})\big)$ for some covariance function $\kappa : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$.

Combining the two models

We propose to combine these two different explanations of the data into one coherent model. Instead of treating either the samples or the features as being (conditionally) independent, we aim to learn a joint covariance for the observed data matrix $\boldsymbol{Y}$. This model, called Kronecker GLASSO, is a special instance of the multivariate random effects model introduced in Section 6.2.1, as the data likelihood can be written as:

$$\mathcal{L}\left(\boldsymbol{R}, \boldsymbol{C}^{-1}, \sigma^2\right) = \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \mathbf{0}_{N \cdot G}\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}_{N \cdot G}\right). \tag{6.39}$$

Here, we build on the model components introduced in Section 6.2.2 and Section 6.2.2. We use the sparse L_1 penalty (6.37) for the feature inverse covariance $\boldsymbol{C}^{-1}$ and use a linear kernel for the covariance on rows $\boldsymbol{R} = \boldsymbol{G}\boldsymbol{G}^\top + \rho^2 \boldsymbol{I}_N$. Learning the model parameters proceeds via maximum a-priori inference, optimizing the log likelihood implied by Equation (6.39) with respect to $\boldsymbol{G}$ and $\boldsymbol{C}^{-1}$, and the hyperparameters σ^2, ρ^2. By combining the graphical Lasso and Gaussian process latent variable model in this way, we can recover a network structure in the presence of confounders.

An equivalent generative model can be obtained in a similar way as in dual probabilistic principal components analysis. The main difference is that now, the rows of the weight matrix $\boldsymbol{W}$ are sampled from a $\mathcal{N}(\mathbf{0}_G, \boldsymbol{C})$ distribution instead of a $\mathcal{N}(\mathbf{0}_G, \boldsymbol{I}_G)$ distribution. This generative model for $\boldsymbol{Y}$ given latent variables $\boldsymbol{G} \in \mathbb{R}^{N \times k}$ and feature covariance $\boldsymbol{C} \in \mathbb{R}^{G \times G}$ is of the form $\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{W} + \rho\boldsymbol{V} + \boldsymbol{E}$, where $\boldsymbol{W} \in \mathbb{R}^{k \times G}$, $\boldsymbol{V} \in \mathbb{R}^{N \times G}$ and $\boldsymbol{E} \in \mathbb{R}^{N \times G}$ are jointly independent with distributions $\operatorname{vec}(\boldsymbol{W}) \sim \mathcal{N}(\mathbf{0}_{k \cdot G}, \boldsymbol{C} \otimes \boldsymbol{I}_K)$, $\operatorname{vec}(\boldsymbol{V}) \sim \mathcal{N}(\mathbf{0}_{NG}, \boldsymbol{C} \otimes \boldsymbol{I}_N)$ and $\operatorname{vec}(\boldsymbol{E}) \sim \mathcal{N}(\mathbf{0}_{NG}, \sigma^2 \boldsymbol{I}_{NG})$.

Inference in the joint model

As already mentioned in Section 6.2.1, parameter inference in the Kronecker GLASSO model implied by Equation (6.39), when done naïvely, is intractable for all but very low dimensional data matrices $\boldsymbol{Y}$. Even using the tricks discussed in Section 6.2.1, free-form sparse inverse covariance updates for $\boldsymbol{C}^{-1}$ are intractable under the L_1 penalty when depending on gradient updates.

Similar as in Section 6.2.1, the first step towards efficient inference is to introduce $N \times G$ additional latent variables $\boldsymbol{Z}$, which can be thought of as the noise-free observations:

$$p(\boldsymbol{Y} | \boldsymbol{Z}, \sigma^2) = \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y})\,;\, \operatorname{vec}(\boldsymbol{Z}), \sigma^2 \boldsymbol{I}_{N \cdot G}\right) \tag{6.40}$$

$$p(\boldsymbol{Z} | \boldsymbol{R}, \boldsymbol{C}) = \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Z})\,;\, \mathbf{0}_{N \cdot G}, \boldsymbol{C} \otimes \boldsymbol{R}\right). \tag{6.41}$$

We consider the latent variables $\boldsymbol{Z}$ as additional model parameters. We now optimize the distribution $p(\boldsymbol{Y}, \boldsymbol{C}^{-1} \,|\, \boldsymbol{Z}, \boldsymbol{R}, \sigma^2) = p(\boldsymbol{Y} \,|\, \boldsymbol{Z}, \sigma^2) p(\boldsymbol{Z} \,|\, \boldsymbol{R}, \boldsymbol{C}) p(\boldsymbol{C}^{-1})$ with respect to the unknown parameters $\boldsymbol{Z}$, $\boldsymbol{C}^{-1}$, σ^2, and $\boldsymbol{R}$ (which depends on $\boldsymbol{G}$ and kernel parameters $\boldsymbol{\Theta}_{\boldsymbol{R}}$) by iterating through the following steps:

1. Optimize for $\sigma^2, \boldsymbol{R}$ after integrating out $\boldsymbol{Z}$, for fixed $\boldsymbol{C}$:

$$\begin{aligned} &\operatorname*{argmax}_{\sigma^2, \boldsymbol{\Theta}_{\boldsymbol{R}}, \boldsymbol{G}} p(\boldsymbol{Y} \,|\, \boldsymbol{C}, \boldsymbol{R}(\boldsymbol{\Theta}_{\boldsymbol{R}}, \boldsymbol{G}), \sigma^2) = \\ &\operatorname*{argmax}_{\sigma^2, \boldsymbol{\Theta}_{\boldsymbol{R}}, \boldsymbol{G}} \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y})\,;\, \mathbf{0}_{N \cdot G}, \boldsymbol{C} \otimes \boldsymbol{R}(\boldsymbol{\Theta}_{\boldsymbol{R}}, \boldsymbol{G}) + \sigma^2 \boldsymbol{I}_{N \cdot G}\right) \end{aligned} \tag{6.42}$$

2. Calculate the expectation of $\boldsymbol{Z}$ for fixed $\boldsymbol{R}$, $\boldsymbol{C}$, and σ^2 :

$$\mathrm{vec}(\hat{\boldsymbol{Z}}) = (\boldsymbol{C} \otimes \boldsymbol{R})(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}_{N \cdot G})^{-1} \mathrm{vec}(\boldsymbol{Y})$$

3. Optimize $\hat{\boldsymbol{C}}^{-1}$ for fixed $\boldsymbol{R}$ and $\hat{\boldsymbol{Z}}$:

$$\underset{\hat{\boldsymbol{C}}^{-1}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{C}}^{-1} \mid \hat{\boldsymbol{Z}}, \boldsymbol{R}) = \underset{\hat{\boldsymbol{C}}^{-1}}{\operatorname{argmax}} \mathcal{N}\left(\mathrm{vec}(\hat{\boldsymbol{Z}})\,;\, \boldsymbol{0}, \hat{\boldsymbol{C}} \otimes \boldsymbol{R}\right) p(\hat{\boldsymbol{C}}^{-1})$$

and set $\boldsymbol{C} = \hat{\boldsymbol{C}}$.

As a stopping criterion we consider the relative reduction of the negative log-marginal likelihood (Equation (6.39)) plus the regularizer on $\boldsymbol{C}^{-1}$. The choice to optimize $\hat{\boldsymbol{C}}^{-1}$ for fixed $\hat{\boldsymbol{Z}}$ is motivated by computational considerations, as this subproblem then reduces to conventional graphical Lasso; a full expectation-maximization approach with latent variables $\boldsymbol{Z}$ does not seem feasible. Step 1 can be done using the efficient likelihood evaluations and gradients described in Section 6.2.1. We will now discuss step 3 in more detail.

Optimizing for $\hat{\boldsymbol{C}}^{-1}$ The third step, optimizing with respect to $\hat{\boldsymbol{C}}^{-1}$, can be done efficiently, using similar ideas as in Section 6.2.1. First consider:

$$\ln \mathcal{N}\left(\mathrm{vec}(\hat{\boldsymbol{Z}})\,;\, \boldsymbol{0}_{N \cdot G}, \hat{\boldsymbol{C}} \otimes \boldsymbol{R}\right) = -\frac{N \cdot G}{2} \ln(2\pi) - \frac{1}{2} \ln \left|\hat{\boldsymbol{C}} \otimes \boldsymbol{R}\right| - \frac{1}{2} \mathrm{vec}(\hat{\boldsymbol{Z}})^\top (\hat{\boldsymbol{C}} \otimes \boldsymbol{R})^{-1} \mathrm{vec}(\hat{\boldsymbol{Z}}).$$

Now, using the Kronecker identity (6.2) and

$$\ln |\boldsymbol{A} \otimes \boldsymbol{B}| \quad = \mathrm{rank}(\boldsymbol{B}) \ln |\boldsymbol{A}| + \mathrm{rank}(\boldsymbol{A}) \ln |\boldsymbol{B}|\,,$$

we can rewrite the log likelihood as:

$$\begin{aligned} &\ln \mathcal{N}\left(\mathrm{vec}(\hat{\boldsymbol{Z}})\,;\, \boldsymbol{0}, \hat{\boldsymbol{C}} \otimes \boldsymbol{R}\right) p(\hat{\boldsymbol{C}}^{-1}) \\ &= -\tfrac{N \cdot G}{2} \ln(2\pi) - \tfrac{1}{2} G \ln |\boldsymbol{R}| + \tfrac{1}{2} N \ln \left|\hat{\boldsymbol{C}}^{-1}\right| - \tfrac{1}{2} \mathrm{tr}(\hat{\boldsymbol{Z}}^\top \boldsymbol{R}^{-1} \hat{\boldsymbol{Z}} \hat{\boldsymbol{C}}^{-1}). \end{aligned}$$

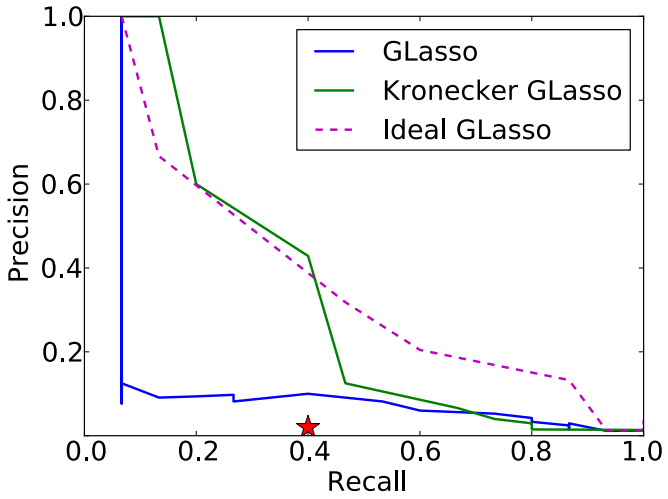
Thus we obtain a standard graphical Lasso problem with covariance matrix $\hat{\boldsymbol{Z}}^\top \boldsymbol{R}^{-1} \hat{\boldsymbol{Z}}$:

$$\underset{\hat{\boldsymbol{C}}^{-1}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{C}}^{-1} \mid \hat{\boldsymbol{Z}}, \boldsymbol{R}) = \underset{\hat{\boldsymbol{C}}^{-1} \succ \boldsymbol{0}}{\operatorname{argmax}} \left(-\frac{1}{2} \mathrm{tr}(\hat{\boldsymbol{Z}}^\top \boldsymbol{R}^{-1} \hat{\boldsymbol{Z}} \hat{\boldsymbol{C}}^{-1}) + \frac{1}{2} N \ln \left|\hat{\boldsymbol{C}}^{-1}\right| - \eta \left\|\hat{\boldsymbol{C}}^{-1}\right\|_1\right). \tag{6.43}$$

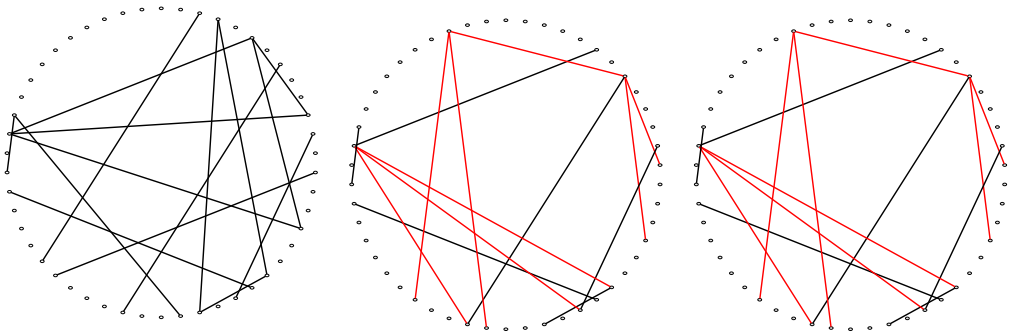
The inverse sample covariance $\boldsymbol{R}^{-1}$ in Equation (6.43) rotates the data covariance, similar as in the established "flip-flop algorithm" for inference in matrix-variate normal distributions [Dutilleul, 1999, Zhang and Schneider, 2010].

6.2.3. Experiments

In this Section, we describe three experiments with the generalized graphical Lasso.



(a) Precision-recall curve.



(b) Ground truth.

(c) Kron. GLASSO

(d) Ideal GLASSO

Figure 6.1. Network reconstruction by Kronecker GLASSO and comparison methods. (a) Precision-recall curve, when varying the sparsity penalty η. Compared are the standard graphical Lasso (GLASSO), our algorithm with Kronecker structure (Kronecker GLASSO) and as a reference an idealized setting, applying standard graphical Lasso to a similar dataset without confounding influences (Ideal GLASSO). The model that accounts for confounders approaches the performance of an idealized model, while standard graphical Lasso finds a large fraction of false positive edges. (b) Ground truth network. Recovered networks for (c) Kronecker GLASSO and (d) Ideal GLASSO at 40% recall (star in (a)). False positive predicted edges are colored in red. Because of the effect of confounders, standard GLASSO predicted an excess of edges to 4 of the nodes.

Simulation study

First, we considered an artificial dataset to illustrate the effect of confounding factors on the solution quality of sparse inverse covariance estimation. We created synthetic data, with $N = 100$ samples and $G = 50$ features according to the generative model described in Section 6.2.2. We generated the sparse inverse column covariance $\boldsymbol{C}^{-1}$ choosing edges at random with a sparsity level of 1%. Non-zero entries of the inverse covariance were drawn from a Gaussian with mean 1 and variance 2. The row covariance matrix $\boldsymbol{R}$ was

created from $k = 3$ random factors $\boldsymbol{g}_i$, each drawn from unit variance *i.i.d.* Gaussian variables. The weighting between the confounders and the *i.i.d.* component ρ^2 was set such that the factors explained equal variance, which corresponds to moderate extent of confounding influences. Finally, we added independent Gaussian observation noise, choosing a signal-to-noise ratio of 10%.

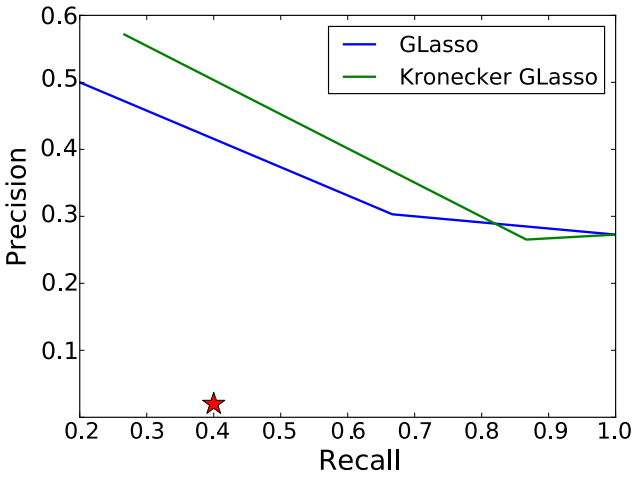
Next, we applied different methods to reconstruct the true simulated network. We considered standard graphical Lasso and our Kronecker model that accounts for the confounding influence (Kronecker GLASSO). For reference, we also considered an idealized setting, applying graphical Lasso to a similar dataset without the confounding effects (Ideal GLASSO), obtained by setting $\boldsymbol{G} = \boldsymbol{0}_{N \cdot k}$ in the generative model. To determine an appropriate latent dimensionality of Kronecker GLASSO, we used the Bayesian information criterion criterion on multiple restarts with $k = 1$ to $k = 5$ latent factors. For all models we varied the sparsity parameter of the graphical lasso, setting $\eta = 5^x$, with x linearly interpolated between -8 and 3. The solution set of lasso-based algorithms is typically unstable and depends on slight variation of the data. To improve the stability of all methods, we employed stability selection [Meinshausen and Bühlmann, 2010], applying each algorithm for all regularization parameters 100 times to randomly drawn subsets containing 90% of the data. We then considered edges that were found in at least 50% of all 100 restarts.

Figure 6.1(a) shows the precision-recall curve for each algorithm. Kronecker GLASSO performed considerably better than standard graphical Lasso, approaching the performance of the ideal model without confounders. Figures 6.1b-d show the reconstructed networks at 40% recall. While Kronecker GLASSO reconstructed the same network as the ideal model, standard graphical Lasso found an excess of false positive edges.

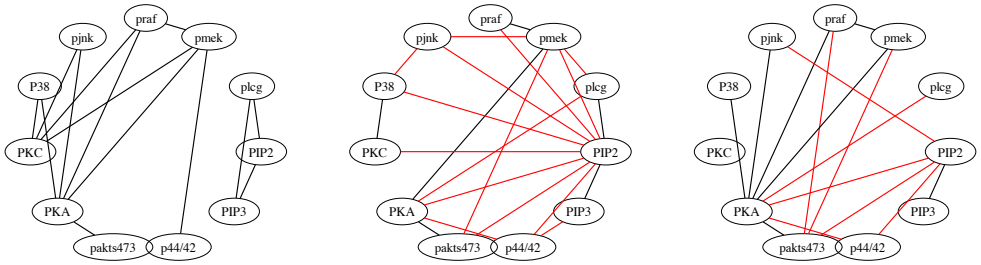
Network reconstruction of protein-signaling networks

Important practical applications of the graphical Lasso include the reconstruction of gene and protein networks. Here, we revisit the extensively studied protein signaling data from[7] Sachs et al. [2005]. The dataset provides observational data of the activations of 11 proteins under various external stimuli. We combined measurements from the first 3 experiments, yielding a heterogeneous mix of 2,666 samples that are not expected to be an *i.i.d.* sample set. To make the inference more difficult, we selected a random fraction of 10% of the samples, yielding a final data matrix of size 266 times 11. We used the directed ground truth network and moralized the graph structure to obtain an undirected ground truth network. Parameter choice and stability selection were done as in the simulation study.

Figure 6.2 shows the results. Analogous to the simulation setting, the Kronecker GLASSO model found true network links with greater accuracy than standard graphical Lasso. This results suggest that our model is suitable to account for confounding variation as it occurs in real settings.



(a) Precision-recall curve.



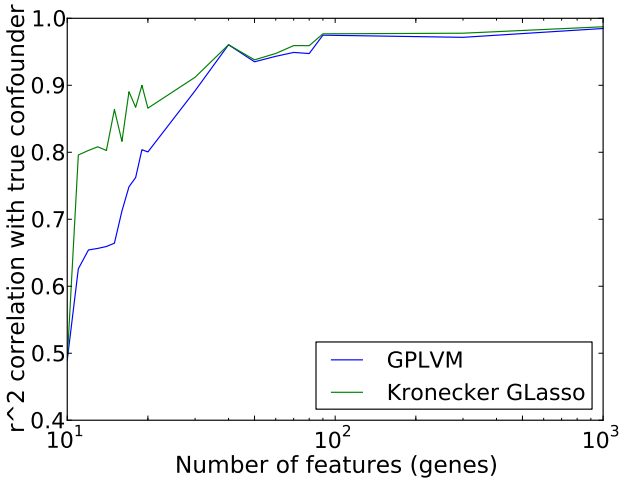
(b) Ground truth. **(c)** GLASSO **(d)** Kron. GLASSO

Figure 6.2. Network reconstruction of a protein signaling network from Sachs et al. [2005] (a) Precision-recall curve, when varying the sparsity penalty η. Compared are the standard graphical Lasso, and our algorithm with Kronecker structure (Kronecker GLASSO). Standard graphical Lasso (GLASSO), not accounting for confounders, found more false positive edges for a wide range of recall rates. (b) Ground truth network. Recovered networks for (c) the graphical Lasso (GLASSO) and (d) Kronecker GLASSO at 40% recall (star in (a)). False positive edge predictions are colored in red.

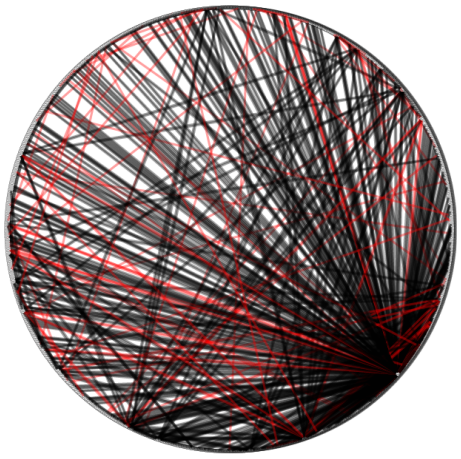
Large-scale application to yeast gene expression data

Next, we considered an application to large-scale gene expression profiling data from yeast. We revisited the dataset from Smith and Kruglyak [2008], consisting of 109 genetically diverse yeast strains, each of which has been expression profiled in two environmental conditions (glucose and ethanol)[8]. Because the confounder in this dataset is known explicitly, we tested the ability of Kronecker GLASSO to recover it from observational data. Because of missing complete ground truth information, we could not evaluate the network reconstruction quality directly. An appropriate regularization parameter was

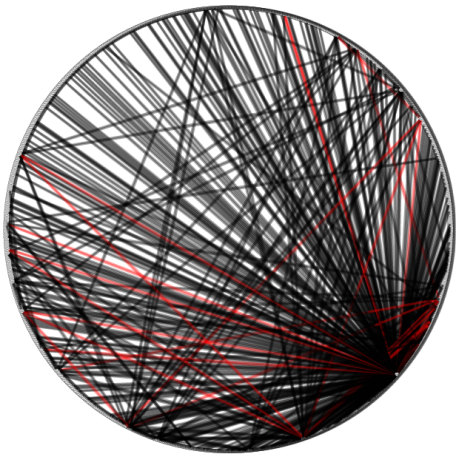
[7] See Section A.8 for more details.
[8] See Section A.9 for additional information.



(a) Confounder reconstruction



(b) GLASSO consistency (68%).



(c) Kron. GLASSO consistency (74%)

Figure 6.3. Comparison of Kronecker GLASSO and GLASSO on an eQTL study in yeast. (a) Correlation coefficient between learned confounding factor and true environmental condition for different subsets of all features (genes). Compared are the standard Gaussian process latent variable model with a linear covariance and our proposed model that accounts for low rank confounders and sparse gene-gene relationships (Kronecker GLASSO). Kronecker GLASSO is able to better recover the hidden confounder by accounting for the covariance structure between genes. Consistency of edges on the largest network with 1,000 nodes learnt using (b) the GLASSO or using (c) Kronecker GLASSO on the joint dataset, comparing the results when combining both conditions with those for a single condition (glucose).

selected by means of cross validation, evaluating the marginal likelihood on a test set (analogous to the procedure described in Friedman et al. [2008]). To simplify the comparison to the known confounding factor, we chose a fixed number of confounders that we set to $k = 1$.

Recovery of the known confounder Figure 6.3(c) shows the r^2 correlation coefficient between the inferred factor and the true environmental condition for increasing number of features (genes) that were used for learning. In particular for small numbers of genes, accounting for the network structure between genes improved the ability to recover the true confounding effect.

Consistency of obtained networks Next, we tested the consistency when applying graphical Lasso and Kronecker GLASSO to data that combines both conditions, glucose and ethanol, comparing to the recovered network from a single condition alone (glucose). The respective networks are shown in Figures 6.3(b) and 6.3(c). The Kronecker GLASSO model identifies more consistent edges, which shows the susceptibility of standard graphical Lasso to the confounder, here the environmental influence.

6.3. Efficient multivariate linear mixed models

Various forms of multivariate linear mixed models have been proposed in the past. These models, which are reviewed in Section 6.3.1, were mostly motivated by applications in animal breeding, where several traits are evaluated for culling of animals to optimize the genetic composition of farm animals towards a large yield in these traits. While the models proposed in this field are all rather expressive, in the sense that some of these allow for unbalanced designs, or varying noise levels between different traits, they share some deficits that make them hardly applicable to genome-wide association testing in a large number of traits. In contrast to applications of mixed models in genetic association studies, where one usually is interested in testing the effects of fixed effects, while the random effects are treated as nuisance parameters and integrated out, in classical animal breeding the situation is the exact opposite. Here, the genetic contribution to a trait, which is selected for, is modeled as a random-effect, while non-genetic influences that differ between samples are usually explained away in the form of univariate fixed effects that only involve a between-individuals design matrix. Second, expressiveness of the noise model as well as allowing for unbalanced designs come at the cost that inference in these models is expensive and in practice does not scale to more than a couple of traits.

To this end we propose a class of large-scale multivariate linear mixed models, that allows to efficiently analyze the sum of fixed terms composed from between-individuals designs and within-individuals designs. We derive a scalable inference scheme that, at the cost of requiring a balanced setting (i. e. each individual has to be observed for all traits), allows to jointly analyze and perform statistical testing of the effects of genetic markers.

6.3.1. Previous multivariate mixed models

A number of related models have been proposed before. These are reviewed in the following.

A multivariate mixed model for balanced designs

A model for selection of animals based on their random genetic effect on a set of correlated traits with a balanced design was proposed by Thompson [1973].

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, (\boldsymbol{I}_G \otimes \boldsymbol{X}) \operatorname{vec}(\boldsymbol{B});\ \boldsymbol{R} \otimes \boldsymbol{C} + \boldsymbol{I} \otimes \boldsymbol{D}\right), \tag{6.44}$$

where $\boldsymbol{D}$ is a diagonal matrix of unknown noise covariances between target variables, and $\boldsymbol{R}$ is a known sample covariance matrix, $\boldsymbol{C}$ is a G-by-G matrix of random-effects covariances between dimensions and $\boldsymbol{R}$ is an N-by-N matrix of random-effects covariances between samples.

Maximum likelihood inference for the model has been considered in Meyer [1985].

A multivariate mixed model for unbalanced designs

As the model above applies only to the balanced case, where the data matrix $\boldsymbol{Y}$ is observed for all individuals and all traits. A similar model was proposed for the unbalanced case, where not all individuals are observed for each trait [Schaeffer et al., 1978].

For each of G traits, let $\boldsymbol{y}_c$ be a vector of N_c observations for the trait c, with $c \in [1, \ldots, G]$. Each $\boldsymbol{y}_c$ is modeled as the sum of fixed effects $\boldsymbol{\beta}_c$, random effects $\boldsymbol{v}_c$ and noise $\boldsymbol{\epsilon}_c$.

$$\boldsymbol{y}_c = \boldsymbol{Z}_c \boldsymbol{X} \boldsymbol{\beta}_c + \boldsymbol{Z}_c \boldsymbol{G} \boldsymbol{v}_c + \boldsymbol{\epsilon}_c,$$

where $\boldsymbol{X}$ is a design matrix for the fixed effects of covariates and $\boldsymbol{G}$ is a design matrix of random genetic effects. and $\boldsymbol{Z}_c$ is an indicator matrix of individuals for which the trait c is observed.

The vector of all phenotypic observations $\boldsymbol{y}$ is given by the vertical concatenation of all $\boldsymbol{y}_c$.

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \ldots \\ \boldsymbol{y}_G \end{bmatrix}. \tag{6.45}$$

The fixed effect design matrix $\boldsymbol{X}_{\text{full}}$ for the full model is given by the direct sum [Searle et al., 1966] of all $\boldsymbol{Z}_c \boldsymbol{X}$.

$$\boldsymbol{X}_{\text{full}} = \begin{bmatrix} \boldsymbol{Z}_1 \boldsymbol{X} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 \boldsymbol{X} & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{Z}_G \boldsymbol{X} \end{bmatrix} \tag{6.46}$$

The noise is assumed independent between individuals and traits, with a different noise level per target trait.

$$\boldsymbol{V} = \begin{bmatrix} \sigma_1^2 \boldsymbol{I}_{N_1} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_2^2 \boldsymbol{I}_{N_2} & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \sigma_G^2 \boldsymbol{I}_{N_G} \end{bmatrix}. \tag{6.47}$$

Integrating over the random effects $\boldsymbol{v}_c$ for all traits $\boldsymbol{c}$, then the total contribution of the random effects $\boldsymbol{K}$ to variance for all observations is a multiplicative between the covariance between individuals $\boldsymbol{R} = \boldsymbol{G}\boldsymbol{G}^\top$ and the covariance between traits i and j $c_{i,j}$, for all $i \in [1, \ldots, G]$ and $j \in [1, \ldots, G]$.

$$\boldsymbol{K} = \begin{bmatrix} c_{1,1}\boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^\top & c_{1,2}\boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_2^\top & \ldots & c_{1,G}\boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_G^\top \\ c_{2,1}\boldsymbol{Z}_2\boldsymbol{R}\boldsymbol{Z}_1^\top & c_{2,2}\boldsymbol{Z}_2\boldsymbol{R}\boldsymbol{Z}_2^\top & \ldots & c_{2,G}\boldsymbol{Z}_2\boldsymbol{R}\boldsymbol{Z}_G^\top \\ \ldots & \ldots & \ldots & \ldots \\ c_{1,G}\boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_G^\top & c_{2,G}\boldsymbol{Z}_2\boldsymbol{R}\boldsymbol{Z}_G^\top & \ldots & c_{G,G}\boldsymbol{Z}_G\boldsymbol{R}\boldsymbol{Z}_G^\top \end{bmatrix}. \tag{6.48}$$

It follows that the distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}\,;\, \boldsymbol{K} + \boldsymbol{V}\right). \tag{6.49}$$

If all individuals are observed for all traits, (all $\boldsymbol{Z}_c = \boldsymbol{I}_N$), then $\boldsymbol{K}$ and $\boldsymbol{V}$ have Kronecker structure and the model is identical to the model considered in Section 6.3.1.

6.3.2. Large-scale multivariate linear mixed models for balanced designs

We propose an efficient algorithmic framework for inference and statistical testing in multivariate linear mixed models. Our approach generalizes linear mixed models on univariate phenotypes to the setting, where matrix-variate values depend on both row-wise as well as column-wise features and their interactions. As a multivariate mixed model, our approach allows for conditioning on random effects with a multivariate Gaussian prior defined by row and column covariance structures.

By concentrating on the balanced design case, we derive efficient algorithms, that should allow genome-wide testing for association between genetic markers and a large number of phenotypes. Similar to our efficient algorithm for parameter estimation in multivariate random-effects models (see Section 6.2.1) use of the spectral decomposition of row and column covariance matrices allows us to apply Kronecker product identities despite modeling *i.i.d.* noise.

The multivariate extensions to linear mixed models aims at explaining the variation of a matrix-variate data matrix $\boldsymbol{Y}$ with N rows and G columns using a matrix-variate normal model:

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, (\boldsymbol{A} \otimes \boldsymbol{X}) \underbrace{\operatorname{vec}(\boldsymbol{B})}_{\text{fixed effects}} + \underbrace{\boldsymbol{Z}}_{\text{random effects}} ; \sigma^2\boldsymbol{I}\right). \tag{6.50}$$

The model in Equation (6.50) generalizes the univariate mixed model in Equation (2.15). Again, $\boldsymbol{B}$ denotes the fixed-effects. The matrix $\boldsymbol{A}$ is the design matrix of the column effect and $\boldsymbol{X}$ of the row effect. For example in a test for gene-environment interaction, the matrix $\boldsymbol{X}$ could be a genetic marker and the matrix $\boldsymbol{A}$ could be a binary indicator of an environment-specific intervention. In this case the matrix $\boldsymbol{B}$ would contain interaction effects between the genetic marker and the intervention. When testing for a joint constant effect of a marker on all target variables, the matrix $\boldsymbol{X}$ would also be a genetic marker and the matrix $\boldsymbol{A}$ would be a 1-by-G row-vector of ones, thereby replicating the effects $\boldsymbol{B}$ across all phenotypes. A column design matrix equal to the G-by-G identity matrix on the

other hand would allow for (seemingly) unrelated effects, yielding a direct generalization of the seemingly unrelated regressions model to cases where both rows and columns covary (see Section 6.1.3).

Analogous as in the linear mixed models introduced in Section 2.2, we assume that the random effects $\boldsymbol{Z}$ follow a normal distribution, but as in Section 6.2, we assume it to be matrix-variate:

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Z}) \,\middle|\, \boldsymbol{0}\,;\, \boldsymbol{C} \otimes \boldsymbol{R}\right). \tag{6.51}$$

Integrating over the random effects distributions, we obtain the marginal likelihood model

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, (\boldsymbol{A} \otimes \boldsymbol{X})\operatorname{vec}(\boldsymbol{B})\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right). \tag{6.52}$$

Using the Kronecker vec operation from Equation (E.2), we can rewrite the mean term in Equation (6.52) as a product of row-wise fixed effects $\boldsymbol{X}$, the weight matrix $\boldsymbol{B}$, and column-wise fixed effects $\boldsymbol{A}$:

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \operatorname{vec}\left(\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right)\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right). \tag{6.53}$$

We argue that this is a natural extension to linear fixed effects, as marginalization of these fixed effects over a matrix-variate prior distribution, would also result in a multivariate random effect. In general, we may assume a sum of fixed effects, which results in the model we consider in the remainder of this chapter:

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \operatorname{vec}\left(\sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top\right)\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right), \tag{6.54}$$

where each $\boldsymbol{X}_j$ is an N-by-D_j matrix, with $\sum_{j=1}^{J} D_j = D$, and each $\boldsymbol{A}_j$ is a G-by-M_j matrix, with $\sum_{j=1}^{J} M_j = M$ The mean term is a linear function in the stacked Kronecker products of the design matrix for each of the J terms in the sum. We defining the complete $(N \cdot G)$-by-$(D \cdot M)$ design matrix $\boldsymbol{\Phi}$ as follows

$$\Phi = \begin{bmatrix} \boldsymbol{A}_1 \otimes \boldsymbol{X}_1, & \ldots, & \boldsymbol{A}_J \otimes \boldsymbol{X}_J \end{bmatrix}. \tag{6.55}$$

The $(D \cdot M)$-by-1 vector $\boldsymbol{\beta}$ of concatenated fixed effects is defined as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \operatorname{vec}(\boldsymbol{B}_1) \\ \vdots \\ \operatorname{vec}(\boldsymbol{B}_J) \end{bmatrix}. \tag{6.56}$$

Using the definition of Φ and $\boldsymbol{\beta}$ the matrix-variate mixed model in Equation (6.54) can be written as

$$\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \Phi\boldsymbol{\beta}\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right). \tag{6.57}$$

Further generalizations to sums of Kronecker products in the covariance are also possible, however omitted for brevity.

Efficient parameter estimation

Having defined the multivariate mixed model in Equation (6.54), we now discuss how to perform efficient parameter inference. Evaluation of the marginal likelihood can be done efficiently using linear algebra identities previously proposed in Section E.2 for the multivariate random effects model.

Again using the spectral decomposition of the Kronecker product plus a constant diagonal term given in Equation (6.4), the log of the likelihood model (Equation (6.54)) can then be written as

$$\begin{aligned}
&\ln \mathcal{L}\left(\underbrace{\theta_{\boldsymbol{C}}, \theta_{\boldsymbol{R}}, \sigma^2, \{\boldsymbol{B}_j\}}_{\boldsymbol{\Theta}}\right) = \\
&-\frac{R \cdot C}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right| \\
&-\frac{1}{2}\mathrm{vec}(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_{\mathrm{r}} \boldsymbol{U}_{\boldsymbol{C}})^\top (\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1} \mathrm{vec}(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_{\mathrm{r}} \boldsymbol{U}_{\boldsymbol{C}}),
\end{aligned} \tag{6.58}$$

allowing for efficient evaluation. Here, we have defined the $\boldsymbol{Y}_{\mathrm{r}}$ to be the residuals after the fixed effects have been subtracted from the data.

$$\boldsymbol{Y}_{\mathrm{r}} = \left(\boldsymbol{Y} - \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top\right). \tag{6.59}$$

The variable $\boldsymbol{\Theta}$ denotes all the parameters to be fit explicitly where $\theta_{\boldsymbol{C}}$ and $\theta_{\boldsymbol{R}}$ parameterize the column and row covariances respectively.

Gradient-based optimization of model parameters

Starting from the rotated representation in Equation (6.58), efficient evaluation of parameter derivatives are feasible. Derivatives with respect to the covariance parameters $\theta_{\boldsymbol{C}}$, $\theta_{\boldsymbol{R}}$ and σ^2 have previously been considered in Section 6.3.2.

$$\frac{\nabla \log \mathcal{L}\left(\boldsymbol{R}, \boldsymbol{C}, \sigma^2, \boldsymbol{\beta}\right)}{\nabla \boldsymbol{\beta}} = \begin{bmatrix} \mathrm{vec}\left(\frac{\nabla \log \mathcal{L}\left(\boldsymbol{R}, \boldsymbol{C}, \sigma^2, \boldsymbol{\beta}\right)}{\nabla \boldsymbol{B}_1}\right) \\ \vdots \\ \mathrm{vec}\left(\frac{\nabla \log \mathcal{L}\left(\boldsymbol{R}, \boldsymbol{C}, \sigma^2, \boldsymbol{\beta}\right)}{\nabla \boldsymbol{B}_J}\right) \end{bmatrix}. \tag{6.60}$$

Each D-by-M matrix holding gradients with respect to fixed effects $\boldsymbol{B}_k$ can be evaluated efficiently as derived in Equation (E.29):

$$\frac{\nabla \log \mathcal{L}\left(\boldsymbol{R}, \boldsymbol{C}, \sigma^2, \boldsymbol{\beta}\right)}{\nabla \boldsymbol{B}_k} = -\boldsymbol{X}_k^\top \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot (\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_{\mathrm{r}} \boldsymbol{U}_{\boldsymbol{C}})\right) \boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_k,$$

where the entries of the N-times-G matrix $\boldsymbol{D}_{\boldsymbol{\Lambda}}$ are defined as

$$[\boldsymbol{D}_{\boldsymbol{\Lambda}}]_{r,c} = \frac{1}{[\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c}[\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} + \sigma^2}.$$

Choosing the right order to perform multiplications, each of these gradient matrices takes $O(N^2G + NG^2 + N^2D_k + G^2M_k) = O(N^2G + NG^2)$ time to be evaluated, where multiplication of the residuals with the eigenvectors is the dominating operation. As this operation can be cached, evaluating the full gradient for all fixed effects takes $O(N^2D + G^2M + N^2G + NG^2)$ time. Space requirement is $O(NG)$.

Joint optimization of the variance components and the fixed effect parameters

Ideally all model parameters would be re-estimated for each test. As the problem involves a large number of parameters, computation of the Hessian matrix for optimization can be prohibitive. The easiest way to circumvent computation of the Hessian is by use of a quasi-Newton method like L-BFGS or its bounded version L-BFGS-B [Liu and Nocedal, 1989, Byrd et al., 1995, Zhu et al., 1997].

Note, that the approach suffers from local optima and requires some care. For example could the alternative model erroneously achieve a lower likelihood than the null model, only due to local optima. This problem can be circumvented using multiple restarts at different initialization, including the null model as an initialization.

Fixing the variance components

When performing a large number of tests on a genome-wide scale, re-estimation of the parameters of the covariance matrices $\boldsymbol{C}$ and $\boldsymbol{R}$ on every alternative model is likely going to be prohibitive, despite the computational methods presented here. For this case, we propose to estimate these on the null model only and keep them fixed throughout the testing procedure. A similar approximation is widely applied in the context of univariate linear mixed models [Zhang et al., 2010, Kang et al., 2010].

Given the covariance matrices, the set of maximum likelihood fixed effects can be evaluated in closed form using the generalized least squares estimator.

$$\boldsymbol{\beta}_{\mathrm{M}} = \left(\boldsymbol{\Phi}^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \mathrm{vec}\left(\boldsymbol{Y}\right), \tag{6.61}$$

Using the efficient evaluation derived in Section E.3.4, the fixed effects (Equation (6.61)) can be evaluated efficiently in $O(ND^2 + NGM^2 + D^3M^3)$ or $O(GM^2 + NGD^2 + D^3M^3)$ time and $O(NG)$ memory.

6.3.3. Phrasing hypotheses in matrix variate linear mixed models

$\boldsymbol{A}_j \in \mathbb{R}^{G \times M_j}$ is a matrix, that replicates the j-th fixed effects matrix $\boldsymbol{X}_j \in \mathbb{R}^{N \times D_j}$. $\boldsymbol{A}_j^\top$ typically would be a binary matrix, but could also have real valued features of the target dimensions. Using different versions of $\boldsymbol{A}_j^\top$ corresponds to choosing a testing strategy. For example, when $\boldsymbol{A}_j$ is the $G \times G$ Identity matrix, then one would fit an independent weight to every column of $\boldsymbol{Y}$, when $\boldsymbol{A}_j$ is a column-vector of ones, then one would fit a single joint weight to all columns of $\boldsymbol{Y}$.

$$\mathcal{N}\left(\mathrm{vec}\left(\boldsymbol{Y}\right) \middle| \sum_j \boldsymbol{A}_j \otimes \boldsymbol{X}_j \mathrm{vec}\left(\boldsymbol{B}_j\right); \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)$$

As long as $D_j \leq N$, $M_j \leq G$, the rank of $\boldsymbol{X}_j$ is D_j and the rank of $\boldsymbol{A}_j$ is M_j, and the rank of $[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J]$ is $\sum_{j=1}^{J} M_j \leq N$ the number of degrees of freedom of a single $\boldsymbol{B}_j$ is $D_j \cdot M_j$.

Efficient statistical testing

For the likelihood ratio test the null distribution of the likelihood ratio statistic can be obtained in closed form. The number of degrees of freedom of the test is the difference in the fixed effects between the null model and the alternative model. For example, when testing all fixed effects of a single Kronecker fixed effects term $\boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j$, then the number of degrees of freedom is $D_j \cdot M_j$, the number of entries in the matrix $\boldsymbol{B}_j$.

6.4. Chapter summary and discussion

We have shown an efficient scheme for parameter learning in matrix-variate normal distributions with *i.i.d.* observation noise. By exploiting linear algebra tricks, we have shown how hyperparameter optimization for the row and column covariances can be carried out without evaluating the prohibitive full covariance, thereby greatly reducing computational and memory complexity.

As an application of our framework, we have proposed a method that accounts for confounding influences while estimating a sparse inverse covariance structure. Our approach extends the Graphical Lasso, generalizing the rigid assumption of *i.i.d.* samples to more general sample covariances. For this purpose, we employ a Kronecker product covariance structure and learn a low-rank covariance between samples, thereby accounting for potential confounding influences. We provided synthetic and real world examples where our method is of practical use, reducing the number of false positive edges learned.

Additionally, we have presented an efficient parameter inference scheme for multivariate mixed models and statistical testing within the multivariate framework. This approach shows great promise for applications in a number of fields, one of which is genetics.

A natural application of the proposed approach would be in genetic association studies involving multiple phenotypes, where it enables matrix-variate association tests between individual genotype markers and groups of phenotypes while accounting for confounding variation. Our approach is particularly well suited for such application domain, because the number of hypothesis to be considered requires efficient computations and accounting for confounding variation is critical. Relatedness between individuals together with other unmeasured confounding factors can severely break the *i.i.d.* assumption, when not taken into account [Leek and Storey, 2007]. There are numerous possible designs and modeling choices how matrix variate mixed models can be used in genetics to test entire pathways [Wang et al., 2011] or phenotypic groups determined by means of clustering. In future work we plan to explore such applications in full detail.

7. Conclusions

Up to today a large number of findings have been made due to genome-wide association studies, with numbers growing rapidly. Still, GWAS are often criticised for the results falling behind the initial expectations. In the mid 2000's, a number of early GWAS discoveries of a considerable number of common variants with moderate effect sizes quickly led to the hope, that soon a large part of the causal mutations and causal processes for many common diseases and complex phenotypes would be uncovered. Despite great efforts and spending, the insights gained from GWAS still greatly lack almost any clinically relevant findings, leading to a prolonged phase of disappointment and criticism.

Today it seems clear, that expectations where overly high, relying on assumptions that seldom hold in practice. Even though a number of common causal variants with moderate to large effect sizes, which can easily detected in the GWAS framework, have been determined, these still make up only for a small fraction of the total heritable portion of a phenotype. The hypothesis that common diseases are caused by only a handful of common variants, that can easily be detected by GWAS, seems to apply rarely at most. On the contrary, many phenotypes have convincingly been shown to be much more polygenic and effect sizes much smaller then what initially has been assumed [Yang et al., 2010, 2011b, Stahl et al., 2012]. Also the influence of uncommon and rare variants seems to be larger than anticipated.

Instead of just creating more and more data within the traditional GWAS design, researchers are now thinking about ways to adapt their design according to the lessons learned. On the one side, huge cohorts involving rich genotype and phenotype data are being assembled, to achieve the power to detect weaker effects tagged by common variants. As costs for genotyping common variants have been dropping drastically, large public consortia and health organizations like Kaiser Permanente are performing large-scale genotyping on cohorts involving hundreds of thousands of individuals, for which over the course of decades they have collected a wealth of phenotypic information and clinical data. In many European countries like the UK, where public health records of high quality are available for a large fraction of the population, these are being complemented by genotype data. When assembling such gigantic and high-powered cohorts, hidden relatedness or hidden environmental influences, would almost inevitably lead to biases in the results of the analysis, when not taken into account. It is also due to the methodological contributions like the ones presented in this thesis, that robust off-the-shelve analysis tools for such huge cohorts are are now readily accessible even to non-experts. Not only does FaST-LMM automatically correct for many problems in a rapid manner, without requiring explicit knowledge of the problem or the cause, other advances implemented in FaST-LMM, like variable selection, or correcting for the effects of proximal contamination, also yield an increase in power over traditional analyses (see Chapters 3 and 4) [Lippert et al., 2011, Listgarten et al., 2012, 2013a, Lippert et al., 2013a,b].

Investigation of the full spectrum of variation present in the genome in a depth that goes beyond genotyping of pre-defined panels od common variants requires the use of sequencing technology. While it is viable to genotype large cohorts using microarrays, the cohort sizes for which deep sequencing data are being produced are considerably smaller. In contrast to typical genotype arrays that tag a large fraction of the common variation at a low price, the cost for obtaining the depth in sequencing required for a complete picture of the genotypic variation, including uncommon and rare variants, is still considerably higher. Depending on the depth of sequencing, the cost for obtaining a complete human genome is still in the order of several thousand of US Dollars and lately has been stagnating.

In many studies the cost, which still prohibits study of the whole genome in large cohorts, is reduced by sequencing only certain parts of the genome, for example the complete exonic region, a targeted region that is associated to a phenotype, or the whole genome at lower depth. While such data is likely to many additional contain causal variants that are not covered by standard SNP arrays, typical sample sizes would still be too small to obtain the power to significantly associate most individual uncommon and rare variants.

Set tests are a method of choice to utilize sequencing data to detect genetic regions involving rare variants that explain a significant fraction of variation [Price et al., 2010a, Wu et al., 2011]. As we have found the result of set tests to be especially vulnerable to genetic relatedness [Quon et al., 2013], methods that correct for genetic structure as proposed in Chapter 5 should prove useful to overcome this problem [Listgarten et al., 2013b, Oualkacha et al., 2013]. Additionally, shallow sequencing makes it hard to reliably call and detect rare variants across all individuals. Ideally such variability in calling quality should be accounted for in a GWAS analysis. As long as sequencing data is still too expensive to obtain at a sufficiently high depth, ways to adjust variance components to address these biases are a worthy target for further investigation.

Another avenue that will increase in importance is the study the phenotypic variation of each individual. By collecting a wealth of phenotypic data on the organismal as well as on the molecular level, traditional GWAS methods that build on the simplistic model that look at isolated pairs of mutations and phenotype are going to be insufficient to fully utilize such rich data and gain a complete picture of the complex mechanisms involved in phenotypic regulation. Multivariate methods for the joined analysis of multiple phenotypes as considered in Chapter 6 are going to prove useful in a broad range of applications. For example in the context of RNA sequencing experiments, multivariate mixed models can be used for testing both joined, as well as differential genetic regulation of alternative transcript isoforms [Rakitsch et al., 2012]. Accurate models that set these data in relation and allow to dissect associations to multiple phenotypes, to a phenotype over time or to molecular phenotypes for different tissues, would allow to answer questions that go beyond simple association between variants and phenotypes. For example conditional independence tests can be used to perform inference over causal mechanisms and regulatory pathways [Pearl, 2000, Lawlor et al., 2008].

With most of the data being made available on storage platforms like dbgap, methods that enable to accurately combine the results of multiple existing studies of the same or related phenotypes are also likely to become important. Such meta analyses have to cope with population differences in the study cohorts, as well as with external batch effects

due to different experimental protocols. The mixed model framework is a promising means to develop methods, that allow joint analysis while taking these subtle differences into account [Han and Eskin, 2011, Furlotte et al., 2012, Lippert et al., 2013a].

A large number of initial hurdles in studies of complex phenotypes have now been taken, promising a new wave of important findings from genome-wide association studies. The next step is to now start thinking about how to turn these findings into practical value. These for example include using GWAS hits as disease markers in genetic tests for elevated risk for common and late-onset diseases. Similarly, prediction of drug-genotype interactions may help avoid serious side effects by to serious side effects. So far such genetic tests mostly are based on a small number of rare markers with large effect sizes.

As annotation of GWAS hits is still scarce and most effect sizes are small, utilization of such variants might not seem to be a trivial task. But similar to the prediction of breeding values of animals and plants [Hayes et al., 2009], multifactorial models, including the ones considered in Chapter 5, could use a large number of weakly tagging variants to predict elevated risk for complex diseases. As these models aggregate the effects of many variants they should not large individual effect sizes of the variants in order to achieve a reliable prediction of heritable traits or elevated disease risk [Daetwyler et al., 2008, de los Campos et al., 2010, Zhou et al., 2013, Rakitsch et al., 2013, Wray et al., 2013]. For the same reasons even statistical significance of the individual variants should not be a necessity to warrant use in prediction.

In order to turn such predictors into a resource for clinicians, we have to also address a number of problems that go beyond pure basic research. Building robust and easily accessible database solutions that provide secure access to and reliable storage for anonymized data at scale will be crucial. The information has to be presented in a way that is in concordance with privacy concerns, while still providing enough insights to be useful. Another large hurdle is the long and expensive approval processes that has to be overcome before such resources would be able to hit the market.

Despite these complications, I think that the results of GWAS will prove their value in foreseeable future, making it an exciting time to be contributing to this field and following the discoveries made.

A. Datasets

We have used a number of datasets from different organisms.

All analyses additive effect of a SNP on the phenotype, using a 0/1/2 minor-allele count encoding for each SNP. Missing SNP data were mean imputed.

A.1. Wellcome Trust Case Control Consortium 1

The Wellcome Trust Case Control Consortium (WTCCC) 1 data consists of the SNP and phenotype data for seven common diseases: bipolar disorder (BP), coronary artery disease (CAD), hypertension (HT), Chron's disease (CD), rheumatoid arthritis (RA), type I diabetes (T1D), and type II diabetes (T2D). Each phenotype group contains about 1,500 individuals. In addition, a set of approximately 2,000 controls from the UK Blood Service Control Group (NBS) was included. A second control group from the 1958 British Birth Cohort (58C) was not included, as permissions for the data precluded use by a commercial organization.

The difference between values of λ from an uncorrected analysis (Armitage Trend test) and those from Kang et al. [2010] averaged 0.01 across the phenotypes with a standard deviation of 0.01, indicating that the absence of the 58C data in our analysis had little effect on inflation/deflation. In these initial analyses, we found a substantial over-representation of P values equal to one, and traced this to the existence of more than two thousand non-varying SNPs or *single-nucleotide constants* (SNCs). In addition, we found (not surprisingly) that SNPs with very low minor-allele frequencies led to skewed P value distributions. Consequently, we employed a more conservative SNP filter, also described by the WTCCC in Burton et al. [2007], wherein a SNP was excluded if either its minor-allele frequency less than 0.01, it was missing in greater than one percent of the individuals, or it was in the extended MHC region. After filtering, 368,584 SNPs remained for the data used for Chapter 3 and Sections 4.1.1 and 4.1.2. The remaining studies in Chapter 4 and in Chapter 5 made use of an additional list of poor quality SNPs obtained from the WTCCC such that the number of SNPs was reduced to 356,441 SNPs.

In our initial analysis, we excluded individuals and SNPs as described in both Kang et al. [2010] and the primary analysis [Burton et al., 2007]. In total, there were 14,925 individuals across the seven phenotypes and control.

Our analyses for a given disease phenotype used data from the NBS group and the remaining six phenotypes as controls. In theory such an approach might introduce errors, as for example disease causing alleles in other diseases might appear protective as well as lead to a loss in power to detect causal variants that have pleiotropic effects. In practice though, this approach has been shown to in increase power because of the larger sample size without introducing false positives, when confounding factors are accounted for using an LMM [Lippert et al., 2013a].

A.2. Genetic Analysis Workshop 14

The Genetic Analysis Workshop (GAW) 14 data [Edenberg et al., 2005] consisted of autosomal SNP data from an Affymetrix SNP panel and a phenotype indicating whether an individual smoked a pack of cigarettes a day or more for six months or more. The cohort included over eight ethnicities and numerous close family members—1,034 individuals in the dataset had parents, children, or siblings also in the dataset. In addition to the curation provided by GAW, we excluded a SNP when either (i) its minor allele frequency was less than 0.05, (ii) its values were missing in more than 10% of the population, or (iii) its allele frequencies were not in Hardy-Weinberg equilibrium ($p < 0.001$). In addition, we excluded an individual when more than 10% of SNP values were missing. After filtering, there were 7,579 SNPs across 1,261 individuals.

A.3. Large-scale synthetic dataset based on GAW14

The GAW14 data (See Section A.2) was used as the basis for generating large synthetic datasets. Datasets GAW14.x, $x = 1, 5, 10, 20, 50$, and 100 were generated. Roughly, the synthetic GAW14.x dataset was constructed by "copying" the original dataset x times. For each white, black, and Hispanic individual in the original data (1,238 individuals), x individuals were created in the copy. The family relationships among these individuals were similarly copied from the original pedigree. For each individual with no parents, data for each SNP was sampled using the race-based marginal frequency of that SNP in the original dataset. A phenotype for each individual was then sampled from a generalized linear mixed model (GLMM) with a logistic link function whose parameters were adjusted to mimic that of the real data. In particular, the offset and genetic-variance parameters of the GLMM were adjusted so that (i) the phenotype frequency in the real and synthetic data were almost the same, and (ii) the genetic variance parameter of a LMM fit to the real and synthetic data were comparable. It was assumed that there were no fixed effects. GAW14 and GAW14.1 had almost identical runtimes.

A.4. 1966 Northern Finland Birth Cohort

The 1966 Northern Finland Birth Cohort (NFBC66) [Sabatti et al., 2008, Rantakallio, 1969] was analyzed in Section 4.3.3. Genotype data were available for 5,546 Finnish individuals, all with genotyping completeness > 95%. We prepared the data for analysis exactly as in Kang et al. [2010]. In particular, we excluded individuals from further analysis because they had withdrawn consent (15), had discrepancies between reported sex and sex determined from the X chromosome (14), were sample duplications (2), were too related to another subject (77), had more than 5% missing genotypes (1) or had no phenotype data (111), leaving 5,326 individuals for analysis. In addition, we excluded SNPs from the original set of 368,177 when there were more than two discordant genotype calls between different methods (4,711), when the allele frequencies were not in Hardy-Weinberg equilibrium ($p < 10^{-4}$; 5,260), when more than 5% of the individuals had missing values (2,535), or when the minor allele frequency was less than 1% (27,002), leaving 331,475 SNPs for analysis. We adjusted the nine phenotypes used in the original

data for sex, pregnancy status, and use of oral contraceptives.

A.5. Meta-analysis of 107 phenotypes in A. thaliana

The data was taken from a GWAS of 107 phenotypes on 199 *Arabidopsis thaliana* inbred lines [Atwell et al., 2010]. The lines were genotyped using a 250K Affymetrix SNP-tiling array containing 248,584 SNPs [Kim et al., 2007].

We study the group of phenotypes related to the flowering time of the plants. We exclude phenotypes that were measured for less than 150 accessions to avoid possible small sample size effects. 20 out of 23 flowering phenotypes pass this sample size threshold.

For the experiments in Section 4.3.5, we excluded a SNP when its minor allele frequency was less than 0.05, in addition to the data preparation provided by Atwell et al. [2010]. We did not filter SNPs based on deviation from Hardy-Weinberg equilibrium, as such a filter would have excluded all SNPs (using a threshold $p < 0.001$). After filtering, there were 206,612 SNPs.

After quality filtering each genotype comprises 216,130 single nucleotide polymorphisms per accession for the experiments in Section 4.3.5.

A.6. Semi-empirical data

We used as basis for our simulation real genomic data from *Arabidopsis thaliana*. Genotype data for 1,196 plants is available from Horton et al. [2012]. For simulating the population driven effects, we used the real phenotype leaf number at flowering time (LN, 16°C, 16 hrs daylight) which is available for 176 plants. Univariate analyses as done in Atwell et al. [2010] have shown that the phenotype has an excess of associations when we do not correct for population structure while after correction the p-values are approximately uniformly distributed. First, we fit a random effects model to LN to determine the fraction of genetic and residual variance which we subsequently used to predict the population structure for the remaining 1,120 plants. We then simulated the phenotypes as follows:

$$\boldsymbol{y} = \sigma_{\text{sig}}\boldsymbol{y}_{sig} + (1 - \sigma_{\text{sig}})[\sigma_{\text{pop}}\boldsymbol{y}_{\text{pop}} + (1 - \sigma_{\text{pop}})\boldsymbol{e}_i],$$

where $\boldsymbol{y}_{\text{sig}} = \boldsymbol{X}^{(k)}\boldsymbol{\beta}$, $\boldsymbol{X}^{(k)}$ is the SNP data for the k causal SNPs, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{I})$ and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{I})$. The first two causal SNPs are drawn such that they are in close linkage disequilibrium (distance between 1kb and 10kb), the remaining causal SNPs are randomly drawn from the complete genome.

The default settings used for the simulation experiments were $\sigma_{\text{sig}} = 0.7$, $\sigma_{\text{pop}} = 0.5$ and $k = 100$. To determine the influence of the population strength, we considered $\sigma_{\text{sig}} = 0.5$, $k = 20$ and varied $\sigma_{\text{pop}} \in \{0.0, 0.3, 0.5, 0.7, 0.9, 1.0\}$. In experiments to assess the impact of the overall noise, we fixed $k = 100, \sigma_{\text{pop}} = 0.5$, and let σ_{sig} vary in $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. Finally, we considered different numbers of causal SNPs $k \in \{10, 20, 500, 100, 300, 1000\}$ and fixed $\sigma_{\text{sig}} = 0.7, \sigma_{\text{pop}} = 0.5$. For the LD experiments, we used the $\sigma_{\text{sig}} = 0.7$, $\sigma_{\text{pop}} = 0.5$ and $k = 10$. We simulated 30 phenotypes for all settings.

A.7. Mouse data

We also obtained genotype and phenotype data for 1940 mice from a study of Valdar et al. [2006]. Each genotype comprises 12,226 single nucleotide polymorphisms. All mice were derived from eight inbred strains and were crossed to produce a heterogenous stock. The phenotypes span a large variety of different measurements ranging from biochemistry to behavioral traits. Here, we focused on 273 phenotypes which have numeric or binary values.

Preprocessing We standardized the SNP data which has the effect that SNPs with a smaller MAF have a larger effect size. On the phenotypes, we performed a Box-Cox transformation [Box and Cox, 1964] and subsequently standardized the data.

A.8. Sachs signaling

In Section 6.2.3 we analyze the extensively studied protein signaling data from Sachs et al. [2005]. The dataset provides observational data of the activations of 11 proteins under various external stimuli. We combined measurements from the first 3 experiments, yielding a heterogeneous mix of 2,666 samples that are not expected to be an *i.i.d.* sample set.

A.9. Smith and Kruglyak data

In Section 6.2.3 we analyzed the dataset from Smith and Kruglyak [2008], consisting of 109 genetically diverse yeast strains, each of which has been expression profiled in two environmental conditions (glucose and ethanol).

B. Score and information for linear mixed models

Here we provide a summary of a few useful quantities for computations of linear mixed models.

The log likelihood of the linear mixed model equals

$$\log \mathcal{L}\left(\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\beta}\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\log|\boldsymbol{H}_{\boldsymbol{\gamma}}| - \frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right).$$

B.1. Score and observed information for a model parameter

$$\frac{\nabla \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\nabla \boldsymbol{\beta}} = \frac{1}{\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top {\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\boldsymbol{X} \tag{B.1}$$

$$\frac{\nabla^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\nabla^2 \boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{X}^\top {\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\boldsymbol{X} \tag{B.2}$$

$$\frac{\nabla \partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\nabla \boldsymbol{\beta} \partial \gamma_i} = -\frac{1}{\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}{\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\boldsymbol{X} \tag{B.3}$$

$$\frac{\nabla \partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\nabla \boldsymbol{\beta} \partial \sigma^2} = -\frac{1}{\sigma^4}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top {\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\boldsymbol{X} \tag{B.4}$$

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \gamma_i} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}\right) + \frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top {\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}{\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right). \tag{B.5}$$

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \gamma_i \partial \gamma_j} =& \frac{1}{2}\mathrm{tr}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_j} - \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}\frac{\partial^2 \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i \partial \gamma_j}\right) \\ &+ \frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top {\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\left(\frac{\partial^2 \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i \partial \gamma_j} - 2\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}{\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_j}\right){\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right).\end{aligned} \tag{B.6}$$

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^\top {\boldsymbol{H}_{\boldsymbol{\gamma}}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right). \tag{B.7}$$

$$\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial^2 \sigma^2} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{B.8}$$

$$\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \sigma^2 \partial \gamma_i} = -\frac{1}{2\sigma^4}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i} \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{B.9}$$

B.2. Fisher Information

This assumes that the expectation is taken over the likelihood, meaning that the parameters are the correct ones.

$$-E\left[\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \gamma_i \partial \gamma_j}\right] = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i} \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_j}\right). \tag{B.10}$$

$$-E\left[\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial^2 \sigma^2}\right] = \frac{N}{2\sigma^4}. \tag{B.11}$$

$$-E\left[\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \sigma^2 \partial \gamma_i}\right] = \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}\right). \tag{B.12}$$

B.3. Average Information

$$\begin{aligned}
&\frac{1}{2}\left(-E\left[\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \gamma_i \partial \gamma_j}\right] - \frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \gamma_i \partial \gamma_j}\right) = \\
&\frac{1}{4}\mathrm{tr}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1} \frac{\partial^2 \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i \partial \gamma_j}\right) \\
&+ \frac{1}{4\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1}\left(\frac{\partial^2 \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i \partial \gamma_j} - 2\frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i} \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_j}\right) \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right).
\end{aligned} \tag{B.13}$$

$$\frac{1}{2}\left(-E\left[\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial^2 \sigma^2}\right] - \frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial^2 \sigma^2}\right) = \frac{1}{2\sigma^6}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{B.14}$$

$$\begin{aligned}
&\frac{1}{2}\left(-E\left[\frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \sigma^2 \partial \gamma_i}\right] - \frac{\partial^2 \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}\right)}{\partial \sigma^2 \partial \gamma_i}\right) = \\
&- \frac{1}{4\sigma^4}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i} \boldsymbol{H}_{\boldsymbol{\gamma}}{}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) - \frac{1}{4\sigma^2}\mathrm{tr}\left(\boldsymbol{H}_{\boldsymbol{\gamma}}^{-1} \frac{\partial \boldsymbol{H}_{\boldsymbol{\gamma}}}{\partial \gamma_i}\right).
\end{aligned} \tag{B.15}$$

C. Linear mixed model derivations

The matrices $\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}$ and $\boldsymbol{Q} = \left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)\boldsymbol{X}^\top\boldsymbol{H}^{-1}$ meet all the requirements needed for restricted maximum likelihood from Section 2.2.3.

Proposition C.1. $\operatorname{rank}(\boldsymbol{S}) = N - D$.

Proof. This is shown in Proposition C.16. □

Proposition C.2. $\operatorname{rank}(\boldsymbol{Q}) = D$.

$$\boldsymbol{Q} = \left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}$$

Proof. As both $\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}$ is a D-by-D matrix of full rank and $\boldsymbol{X}^\top$ has rank D as long as all the columns of $\boldsymbol{X}$ are linearly independent, their product $\boldsymbol{Q}$ is also of rank D.

Proposition C.3. *The two projections are statistically independent under the model.*
$\Leftrightarrow \mathbb{C}\mathrm{ov}(\boldsymbol{S}\boldsymbol{y}, \boldsymbol{Q}\boldsymbol{y}) = \boldsymbol{0}$.
$\Leftrightarrow \boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{Q}^\top = \boldsymbol{0}$.

Proof.

$$\begin{aligned}
\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{Q}^\top &= \boldsymbol{S}\underbrace{\boldsymbol{H}\boldsymbol{H}^{-1}}_{\boldsymbol{I}}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \underbrace{\boldsymbol{S}\boldsymbol{X}}_{\boldsymbol{0}}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{0}. \quad \square
\end{aligned}$$

Proposition C.4. *The expected value of $\boldsymbol{S}\boldsymbol{y}$ under the model is zero.*
$\Leftrightarrow \mathbb{E}(\boldsymbol{S}\boldsymbol{y}) = \boldsymbol{0}$.
$\Leftrightarrow \boldsymbol{S}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$.
$\Leftarrow \boldsymbol{S}\boldsymbol{X} = \boldsymbol{0}$.

Proof.

$$\underbrace{\boldsymbol{S}\boldsymbol{X}}_{\boldsymbol{0}}\boldsymbol{\beta} = \boldsymbol{0}. \quad \square$$

Proposition C.5. $\operatorname{rank}(\boldsymbol{Q}\boldsymbol{X}) = D$.

Proof.

$$\begin{aligned}
\boldsymbol{Q}\boldsymbol{X} &= \left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X} \\
&= \boldsymbol{I}. \quad \square
\end{aligned} \tag{C.1}$$

C.1. The restricted likelihood

Let in the following Z and R be defined as follows:

$$\mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta}\,;\, \sigma^2\boldsymbol{H}\right) = \underbrace{\left(2\pi\sigma^2\right)^{-\frac{N}{2}} |\boldsymbol{H}|^{-\frac{1}{2}}}_{Z} \cdot \exp\left(-\frac{1}{2\sigma^2}\underbrace{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}_{R}\right).$$

Let in the following $Z_{\boldsymbol{S}}$ and $R_{\boldsymbol{S}}$ be defined as follows:

$$\mathcal{N}\left(\boldsymbol{U}\boldsymbol{y} \,\middle|\, \boldsymbol{0}\,;\, \sigma^2\boldsymbol{\Lambda}\right) = \underbrace{\left(2\pi\sigma^2\right)^{-\frac{N-D}{2}} |\boldsymbol{\Lambda}|^{-\frac{1}{2}}}_{Z_{\boldsymbol{S}}} \cdot \exp\left(-\frac{1}{2\sigma^2}\underbrace{\boldsymbol{y}^\top\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{y}}_{R_{\boldsymbol{S}}}\right).$$

Let in the following $Z_{\boldsymbol{Q}}$,and $R_{\boldsymbol{Q}}$ be defined as follows:

$$\mathcal{N}\left(\boldsymbol{Q}\boldsymbol{y} \,\middle|\, \boldsymbol{w}\,;\, \sigma^2\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right) = \underbrace{\left(2\pi\sigma^2\right)^{-\frac{D}{2}} |\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}|^{-\frac{1}{2}}}_{Z_{\boldsymbol{Q}}} \cdot \exp\left(-\frac{1}{2\sigma^2}R_{\boldsymbol{Q}}\right),$$

with

$$R_{\boldsymbol{Q}} = (\boldsymbol{Q}\boldsymbol{y}-\boldsymbol{\beta})^\top \left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}(\boldsymbol{Q}\boldsymbol{y}-\boldsymbol{\beta}).$$

Proposition C.6.

$$R = R_{\boldsymbol{S}} + R_{\boldsymbol{Q}}.$$

Proof.

$$\begin{aligned}
R = & & (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
= & (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \underbrace{\begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}^{-\top}}_{\boldsymbol{I}} \boldsymbol{H}^{-1} \underbrace{\begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}}_{\boldsymbol{I}} (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
= & (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}^\top \left(\begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}\boldsymbol{H}\begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix}^\top\right)^{-1} \begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{Q}\end{bmatrix} (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
= & \begin{bmatrix}\boldsymbol{U}^\top\boldsymbol{y}\\ \boldsymbol{Q}\boldsymbol{y}-\boldsymbol{\beta}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{U}^\top\boldsymbol{y}\\ \boldsymbol{Q}\boldsymbol{y}-\boldsymbol{\beta}\end{bmatrix} \\
= & \underbrace{\boldsymbol{y}^\top\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{y}}_{R_{\boldsymbol{S}}} + \underbrace{(\boldsymbol{Q}\boldsymbol{y}-\boldsymbol{\beta})^\top\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}(\boldsymbol{Q}\boldsymbol{y}-\boldsymbol{\beta})}_{R_{\boldsymbol{Q}}} \\
= & & R_{\boldsymbol{S}} + R_{\boldsymbol{Q}} \quad \square
\end{aligned}$$

The following Proposition was stated in Harville [1974] without a complete proof:

Proposition C.7.

$$Z = \left|\boldsymbol{X}^\top\boldsymbol{X}\right|^{-\frac{1}{2}} \cdot Z_{\boldsymbol{S}} \cdot Z_{\boldsymbol{Q}}.$$

Proof.

$$
\begin{aligned}
Z_{\boldsymbol{S}} \cdot Z_{\boldsymbol{Q}} = & \left(2\pi\sigma^2\right)^{-\frac{N-D}{2}} |\boldsymbol{\Lambda}|^{-\frac{1}{2}} \cdot \left(2\pi\sigma^2\right)^{-\frac{D}{2}} |\boldsymbol{X}^\top \boldsymbol{H}^{-1} \boldsymbol{X}|^{-\frac{1}{2}} \\
= & \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \left| \begin{bmatrix} \boldsymbol{\Lambda} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}^\top \boldsymbol{H}^{-1} \boldsymbol{X} \end{bmatrix} \right|^{-\frac{1}{2}} \\
= & \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \left| \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{Q} \end{bmatrix} \boldsymbol{H} \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{Q} \end{bmatrix}^\top \right|^{-\frac{1}{2}} \\
= & \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \left| \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{Q} \end{bmatrix} \right|^{-\frac{1}{2}} \cdot |\boldsymbol{H}|^{-\frac{1}{2}} \cdot \left| \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{Q} \end{bmatrix}^\top \right|^{-\frac{1}{2}} \\
= & \underbrace{\left(2\pi\sigma^2\right)^{-\frac{N}{2}} |\boldsymbol{H}|^{-\frac{1}{2}}}_{Z} \cdot \left| \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{Q} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{Q} \end{bmatrix} \right|^{-\frac{1}{2}} \\
= & Z \cdot \left| \underbrace{\boldsymbol{U}^\top \boldsymbol{U}}_{\boldsymbol{I}} \right|^{-\frac{1}{2}} \cdot \left| \boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{Q} \underbrace{\boldsymbol{U} \left(\boldsymbol{U}^\top \boldsymbol{U}\right)^{-1} \boldsymbol{U}^\top}_{\boldsymbol{S}} \boldsymbol{Q}^\top \right|^{-\frac{1}{2}} \\
= & Z \cdot \left| \underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top}_{\boldsymbol{0}} + \underbrace{\boldsymbol{Q}\boldsymbol{X}}_{\boldsymbol{I}} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \underbrace{\boldsymbol{X}^\top \boldsymbol{Q}^\top}_{\boldsymbol{I}} \right|^{-\frac{1}{2}} \\
= & Z \cdot \left| \boldsymbol{X}^\top \boldsymbol{X} \right|^{\frac{1}{2}}. \quad \square
\end{aligned}
$$

The following Proposition was stated in Harville [1974] without a proof:

Lemma C.8. *Given* $\boldsymbol{S} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)\boldsymbol{X}^\top\right) \in \mathbb{R}^{(N\times N}$, $\boldsymbol{Q} = \left(\boldsymbol{X}^\top \boldsymbol{H}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{H}^{-1} \in \mathbb{R}^{D\times N}$, $\boldsymbol{U} \in \mathbb{R}^{N\times(N-D)}$, *with the economy spectral decomposition of* $\boldsymbol{SHS}$ *given as* $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$

$$
\mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta};\, \sigma^2 \boldsymbol{H}\right) = |\boldsymbol{X}^\top \boldsymbol{X}|^{-\frac{1}{2}} \cdot \mathcal{N}\left(\boldsymbol{U}\boldsymbol{y} \,\middle|\, \boldsymbol{0};\, \sigma^2 \boldsymbol{\Lambda}\right) \cdot \mathcal{N}\left(\boldsymbol{Q}\boldsymbol{y} \,\middle|\, \boldsymbol{w};\, \sigma^2 \boldsymbol{X}^\top \boldsymbol{H}^{-1} \boldsymbol{X}\right).
$$

Proof.

$$
\begin{aligned}
\mathcal{N}\left(\boldsymbol{U}\boldsymbol{y} \,\middle|\, \boldsymbol{0};\, \sigma^2 \boldsymbol{\Lambda}\right) \cdot \mathcal{N}\left(\boldsymbol{Q}\boldsymbol{y} \,\middle|\, \boldsymbol{w};\, \sigma^2 \boldsymbol{X}^\top \boldsymbol{H}^{-1} \boldsymbol{X}\right) = & \; Z_{\boldsymbol{S}} \cdot \exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{S}}\right) \cdot Z_{\boldsymbol{Q}} \cdot \exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{Q}}\right) \\
= & \underbrace{Z_{\boldsymbol{S}} \cdot Z_{\boldsymbol{Q}}}_{Z \cdot \left|\boldsymbol{X}^\top \boldsymbol{X}\right|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2\sigma^2} \underbrace{\left(R_{\boldsymbol{S}} + R_{\boldsymbol{Q}}\right)}_{R}\right) \\
= & \left|\boldsymbol{X}^\top \boldsymbol{X}\right|^{\frac{1}{2}} \cdot Z \cdot \exp\left(-\frac{1}{2\sigma^2} R\right) \\
= & \left|\boldsymbol{X}^\top \boldsymbol{X}\right|^{\frac{1}{2}} \cdot \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\boldsymbol{\beta};\, \sigma^2 \boldsymbol{H}\right). \quad \square
\end{aligned}
$$

Lemma C.9. *Let $\boldsymbol{H}_\gamma$ be defined as $\gamma\boldsymbol{K}+\boldsymbol{I}$, where $\boldsymbol{K}$ is a positive semi-definite matrix and $\gamma \geq 0$. Further, let the spectral decomposition of $\boldsymbol{K}$ be $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. Then, the spectral decomposition of $\boldsymbol{H}_\gamma$ is given by $\boldsymbol{U}(\gamma\boldsymbol{\Lambda}+\boldsymbol{I}_N)\boldsymbol{U}^\top$, where $(\gamma\boldsymbol{\Lambda}+\boldsymbol{I}_N)$ is a diagonal matrix holding the N eigenvalues of $\boldsymbol{H}_\gamma$ and the eigenvectors are unchanged.*

Proof.

$$\begin{aligned}\boldsymbol{H}_\gamma &= \gamma\boldsymbol{K}+\boldsymbol{I}\\ &= \gamma\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top+\boldsymbol{I}\\ &= \boldsymbol{U}(\gamma\boldsymbol{\Lambda})\boldsymbol{U}^\top+\boldsymbol{U}\boldsymbol{U}^\top\\ &= \boldsymbol{U}(\gamma\boldsymbol{\Lambda}+\boldsymbol{I})\boldsymbol{U}^\top. \quad \square\end{aligned}$$

The following Lemma was stated and used in Patterson and Thompson [1971]:

Lemma C.10. *Let $\boldsymbol{K}$ be a positive semi-definite matrix, and the spectral decomoposition of $\boldsymbol{K}$ be $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. Then the Moore-Penrose pseudo inverse of $\boldsymbol{K}$ is given by $\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda}^\dagger$ is obtained by inverting the non-zero diagonal entries, contained in the upper diagonal part $\boldsymbol{\Lambda}_1$ of the diagonal matrix $\boldsymbol{\Lambda}$.*

Proof. This is proven, by Propositions C.11 to C.14, the four properties of the Moore-Penrose pseudo inverse. $\square$

Proposition C.11. $\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)$ *is symmetric.*

Proof.

$$\begin{aligned}\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right) &= \boldsymbol{U}\boldsymbol{\Lambda}\underbrace{\boldsymbol{U}^\top\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\\ &= \boldsymbol{U}\begin{bmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_1^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{U}^\top\\ &= \boldsymbol{U}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{U}^\top\\ &= \boldsymbol{U}_1\boldsymbol{U}_1^\top. \quad \square\end{aligned}$$

Proposition C.12. $\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)$ *is symmetric.*

Proof.

$$\begin{aligned}\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right) &= \boldsymbol{U}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{U}^\top\\ &= \boldsymbol{U}_1\boldsymbol{U}_1^\top. \quad \square\end{aligned}$$

Proposition C.13. $\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)$ *is a weak inverse of* $\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)$.

Proof.

$$\begin{aligned}\underbrace{\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)}_{\boldsymbol{U}_1\boldsymbol{U}_1^\top}\underbrace{\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)}_{\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top} &= \boldsymbol{U}_1\underbrace{\boldsymbol{U}_1^\top\boldsymbol{U}_1}_{\boldsymbol{I}}\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top \\ &= \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top \\ &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top. \quad \square\end{aligned}$$

Proposition C.14. $\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)$ *is a weak inverse of* $\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)$.

Proof.

$$\begin{aligned}\underbrace{\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)}_{\boldsymbol{U}_1\boldsymbol{U}_1^\top}\underbrace{\left(\boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top\right)}_{\boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top} &= \boldsymbol{U}_1\underbrace{\boldsymbol{U}_1^\top\boldsymbol{U}_1}_{\boldsymbol{I}}\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top \\ &= \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top \\ &= \boldsymbol{U}\boldsymbol{\Lambda}^\dagger\boldsymbol{U}^\top. \quad \square\end{aligned}$$

The determinant of the genetic similarity matrix, $\left|\boldsymbol{U}\left(\gamma\boldsymbol{\Lambda}+\boldsymbol{I}\right)\boldsymbol{U}^\top\right|$ can be written using the property that $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$, the fact that $|\boldsymbol{U}| = \left|\boldsymbol{U}^\top\right| = 1$, and that the determinant of a diagonal matrix is the product of the diagonal entries. The inverse of the genetic similarity matrix can be rewritten using the property that $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, the fact that $\boldsymbol{U}^{-1} = \boldsymbol{U}^\top$ and $\boldsymbol{U}^{-\top} = \boldsymbol{U}$

The following lemma was stated and used in Patterson and Thompson [1971]:

Lemma C.15. *Let* $\boldsymbol{H}_1$ *be defined as* $\boldsymbol{K}+\boldsymbol{I}$ *and* $\boldsymbol{H}_\gamma$ *be defined as* $\gamma\boldsymbol{K}+\boldsymbol{I}$*, where* $\boldsymbol{K}$ *is a positive semi-definite matrix and* $\gamma \geq 0$*. Further, let the economy spectral decomposition of* $\boldsymbol{S}\boldsymbol{H}_1\boldsymbol{S}$ *be* $\boldsymbol{U}_{\boldsymbol{S}}\left(\boldsymbol{\Sigma}+\boldsymbol{I}_{N-D}\right)\boldsymbol{U}_{\boldsymbol{S}}^\top$*. Then, the economy spectral decomposition of* $\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}$ *is given by* $\boldsymbol{U}\left(\gamma\boldsymbol{\Sigma}+\boldsymbol{I}_{N-D}\right)\boldsymbol{U}^\top$*, where* $\left(\gamma\boldsymbol{\Sigma}+\boldsymbol{I}_{N-D}\right)$ *is a diagonal matrix holding the* $N-D$ *non-zero eigenvalues of* $\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S}$ *and the first* $N-D$ *eigenvectors given as columns of* $\boldsymbol{U}_{\boldsymbol{S}}$ *are unchanged.*

Proof.

$$\begin{aligned}\boldsymbol{S}\boldsymbol{H}_\gamma\boldsymbol{S} &= \boldsymbol{S}\left(\gamma\boldsymbol{K}+\boldsymbol{I}\right)\boldsymbol{S} \qquad \text{(C.2)} \\ &= \boldsymbol{S}\left(\gamma\underbrace{\left(\boldsymbol{K}+\boldsymbol{I}\right)}_{\boldsymbol{H}_1}+\left(1-\gamma\right)\boldsymbol{I}\right)\boldsymbol{S} \\ &= \gamma\boldsymbol{S}\boldsymbol{H}_1\boldsymbol{S}+\left(1-\gamma\right)\boldsymbol{S} \\ &= \gamma\boldsymbol{U}_{\boldsymbol{S}}\left(\boldsymbol{\Sigma}+\boldsymbol{I}\right)\boldsymbol{U}_{\boldsymbol{S}}^\top+\left(1-\gamma\right)\boldsymbol{U}_{\boldsymbol{S}}\boldsymbol{U}_{\boldsymbol{S}}^\top \\ &= \boldsymbol{U}_{\boldsymbol{S}}\left(\gamma\boldsymbol{\Sigma}+\boldsymbol{I}_{N-D}\right)\boldsymbol{U}_{\boldsymbol{S}}^\top,\end{aligned}$$

where we used idempotency of $\boldsymbol{S}$ and Proposition C.32 to replace $\boldsymbol{S}$ by $\boldsymbol{U}_{\boldsymbol{S}}\boldsymbol{U}_{\boldsymbol{S}}^\top$.[1] $\square$

[1] The proof relies on $\boldsymbol{H}_\gamma$ to be full rank, which is always true by definition.

C.1.1. Orthogonal projection matrices

Given an N-by-D matrix $\boldsymbol{X}$ of full column rank D, the N-by-N orthogonal projection matrix $\boldsymbol{S}$ is defined as

$$\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top. \tag{C.3}$$

Proposition C.16. *$\boldsymbol{S}$ is singular of rank $N-D$ has $N-D$ eigenvalues that equal one. and D eigenvalues that equal zero.*

Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{X}$, where $\boldsymbol{U}$ is an N-by-N orthogonal matrix, holding the left singular vectors as columns, $\boldsymbol{V}$ is an D-by-D orthogonal matrix, holding the right singular vectors as columns, and $\boldsymbol{\Sigma}$ is an N-by-D diagonal matrix, holding the D singular values on the diagonal.

Proof.

$$\begin{aligned}
\boldsymbol{S} &= \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \\
&= \underbrace{\boldsymbol{I}}_{\boldsymbol{U}\boldsymbol{U}^\top} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\left(\boldsymbol{V}\boldsymbol{\Sigma}\underbrace{\boldsymbol{U}^\top\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Sigma}^\top\boldsymbol{V}^\top\right)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \\
&= \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}\boldsymbol{\Sigma}\underbrace{\boldsymbol{V}^\top\boldsymbol{V}}_{\boldsymbol{I}}{}^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\underbrace{\boldsymbol{V}^\top\boldsymbol{V}}_{\boldsymbol{I}}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \\
&= \boldsymbol{U}\left(\boldsymbol{I} - \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\boldsymbol{\Sigma}^\top\right)\boldsymbol{U}^\top \\
&= \boldsymbol{U}\left(\boldsymbol{I} - \begin{bmatrix}\boldsymbol{I}_D & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\right)\boldsymbol{U}^\top \\
&= \boldsymbol{U}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{N-D}\end{bmatrix}\boldsymbol{U}^\top
\end{aligned}$$

It follows that the rank of $\boldsymbol{S}$ is $N-D$ and that all $N-D$ non-zero eigenvalues equal one. □

Proposition C.17. *$\boldsymbol{S}$ is orthogonal to $\boldsymbol{X}$*

Proof.

$$\begin{aligned}
\boldsymbol{S}\boldsymbol{X} &= \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{X} \\
&= \boldsymbol{X} - \boldsymbol{X}\underbrace{\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}}_{\boldsymbol{I}} \\
&= \boldsymbol{0}. \quad \square
\end{aligned}$$

Proposition C.18. *For any vector $\boldsymbol{a}$, the projection $\boldsymbol{S}\boldsymbol{a}$ is orthogonal to $\boldsymbol{X}$.*

Proof.

$$\begin{aligned}(\boldsymbol{S}\boldsymbol{a})^\top \boldsymbol{X} &= \boldsymbol{a}^\top \underbrace{\boldsymbol{S}\boldsymbol{X}}_{\mathbf{0}} \\ &= \mathbf{0}. \quad \square\end{aligned}$$

Proposition C.19. *$\boldsymbol{S}$ is idempotent.*

Proof.

$$\begin{aligned}\boldsymbol{S}\boldsymbol{S} &= \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right) \\ &= \boldsymbol{I} - 2\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top + \boldsymbol{X}\underbrace{\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}}_{\boldsymbol{I}}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \\ &= \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \\ &= \boldsymbol{S}. \quad \square\end{aligned}$$

Proposition C.20. *Any matrix $\boldsymbol{A}$ that is orthogonal to $\boldsymbol{X}$, stays constant multiplication with $\boldsymbol{S}$.*

Proof.

$$\begin{aligned}\boldsymbol{S}\boldsymbol{A} &= \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{A} \\ &= \boldsymbol{A} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\underbrace{\boldsymbol{X}^\top\boldsymbol{A}}_{\mathbf{0}}s \\ &= \boldsymbol{A}. \quad \square\end{aligned}$$

C.1.2. Conjugate projection matrices

Given an N-by-D matrix $\boldsymbol{X}$ of full column rank D, a positive definite N-by$N$ matrix $\boldsymbol{H}^{-1}$ of full rank, the N-by-N matrix $\boldsymbol{P}$ is defined as

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}.$$

Proposition C.21. *$\boldsymbol{P}^\top$ is orthogonal to $\boldsymbol{X}$.*

Proof.

$$\begin{aligned}\boldsymbol{P}\boldsymbol{X} &= \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\right)\boldsymbol{X} \\ &= \boldsymbol{X} - \boldsymbol{X}\underbrace{\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}}_{\boldsymbol{I}} \\ &= \mathbf{0}. \quad \square\end{aligned} \tag{C.4}$$

Proposition C.22. $\boldsymbol{P}$ *is idempotent.*

Proof.

$$\begin{aligned}
\boldsymbol{PP} = & \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\right)\boldsymbol{P} \\
= & \boldsymbol{P} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{P} \\
= & \boldsymbol{P} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1} \\
& + \boldsymbol{X}\underbrace{\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}}_{\boldsymbol{I}}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1} \\
= & \boldsymbol{P}. \quad \square
\end{aligned} \tag{C.5}$$

Proposition C.23. $\boldsymbol{SP} = \boldsymbol{S}$, *where* $\boldsymbol{S} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)$
$\Leftrightarrow \boldsymbol{P}^\top\boldsymbol{S} = \boldsymbol{S}$

Proof.

$$\begin{aligned}
\boldsymbol{SP} = & \boldsymbol{S}\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\right) \\
= & \boldsymbol{S} - \underbrace{\boldsymbol{SX}}_{\boldsymbol{0}}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1} \\
= & \boldsymbol{S}. \quad \square
\end{aligned} \tag{C.6}$$

Proposition C.24. $\boldsymbol{PS} = \boldsymbol{P}$, *where* $\boldsymbol{S} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)$
$\Leftrightarrow \boldsymbol{SP}^\top = \boldsymbol{P}^\top$

Proof.

$$\begin{aligned}
\boldsymbol{PS} = & \boldsymbol{P}\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right) \\
= & \boldsymbol{P} - \underbrace{\boldsymbol{PX}}_{\boldsymbol{0}}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \\
= & \boldsymbol{P}. \quad \square
\end{aligned} \tag{C.7}$$

Proposition C.25. $\boldsymbol{H}^{-1}\boldsymbol{P}$ *is symmetric.*
$\Leftrightarrow \boldsymbol{H}^{-1}\boldsymbol{P} = \boldsymbol{P}^\top\boldsymbol{H}^{-1}$.

Proof.

$$\begin{aligned}
\boldsymbol{H}^{-1}\boldsymbol{P} &= \boldsymbol{H}^{-1}\left(\boldsymbol{I}-\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\right)\\
&= \boldsymbol{H}^{-1}-\boldsymbol{H}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\\
&= \left(\boldsymbol{I}-\boldsymbol{H}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{H}^{-1}\\
&= \boldsymbol{P}^\top\boldsymbol{H}^{-1}. \quad \square
\end{aligned} \tag{C.8}$$

Proposition C.26. $\boldsymbol{PH}$ *is symmetric.*
$\Leftrightarrow \boldsymbol{PH} = \boldsymbol{HP}^\top$.

Proof.

$$\begin{aligned}
\boldsymbol{PH} &= \left(\boldsymbol{I}-\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{H}^{-1}\right)\boldsymbol{H}\\
&= \boldsymbol{H}-\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\\
&= \boldsymbol{H}-\underbrace{\boldsymbol{H}\boldsymbol{H}^{-1}}_{\boldsymbol{I}}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\\
&= \boldsymbol{HP}^\top. \quad \square
\end{aligned} \tag{C.9}$$

Lemma C.27. $\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}$ *is the Moore-Penrose pseudoinverse of* $\boldsymbol{SHS}$*, where* $\boldsymbol{S} = \left(\boldsymbol{I}-\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)$ *[Kang et al., 2008].*

Proof. The four properties of the Moore-Penrose pseudoinverse are proven below in Propositions C.28 to C.31. $\square$

Proposition C.28. $(\boldsymbol{SHS})\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)$ *is symmetric.*

Proof.

$$\begin{aligned}
(\boldsymbol{SHS})\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right) &= \boldsymbol{SH}\underbrace{\boldsymbol{P}^\top\boldsymbol{H}^{-1}}_{\boldsymbol{H}^{-1}\boldsymbol{P}}\boldsymbol{P}\\
&= \boldsymbol{S}\underbrace{\boldsymbol{H}\boldsymbol{H}^{-1}}_{\boldsymbol{I}}\underbrace{\boldsymbol{P}\boldsymbol{P}}_{\boldsymbol{P}}\\
&= \boldsymbol{SP}\\
&= \boldsymbol{S}.
\end{aligned}$$

As $\boldsymbol{S}$ is symmetric, $(\boldsymbol{SHS})\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)$ is also symmetric. $\square$

Proposition C.29. $\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)(\boldsymbol{SHS})$ *is symmetric.*

Proof.

$$\begin{aligned}\left(\boldsymbol{P}^\top \boldsymbol{H}^{-1}\boldsymbol{P}\right)(\boldsymbol{SHS}) &= \boldsymbol{P}^\top \boldsymbol{H}^{-1}\underbrace{\boldsymbol{PS}}_{\boldsymbol{P}}\boldsymbol{HS}\\ &= \boldsymbol{P}^\top \underbrace{\boldsymbol{H}^{-1}\boldsymbol{P}}_{\boldsymbol{P}^\top\boldsymbol{H}^{-1}}\boldsymbol{HS}\\ &= \underbrace{\boldsymbol{P}^\top\boldsymbol{P}^\top}_{\boldsymbol{P}^\top}\underbrace{\boldsymbol{H}^{-1}\boldsymbol{H}}_{\boldsymbol{I}}\boldsymbol{S}\\ &= \boldsymbol{P}^\top\boldsymbol{S}\\ &= \boldsymbol{S}.\end{aligned}$$

As $\boldsymbol{S}$ is symmetric, $\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)(\boldsymbol{SHS})$ is also symmetric. □

Proposition C.30. *$\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}$ is a weak inverse of $\boldsymbol{SHS}$.*

Proof.

$$\begin{aligned}\underbrace{\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)(\boldsymbol{SHS})}_{\boldsymbol{S}}\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right) &= \underbrace{\boldsymbol{SP}^\top}_{\boldsymbol{P}^\top}\boldsymbol{H}^{-1}\boldsymbol{P}\\ &= \boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}.\end{aligned}$$ □

Proposition C.31. *$\boldsymbol{SHS}$ is a weak inverse of $\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}$.*

Proof.

$$\begin{aligned}\underbrace{(\boldsymbol{SHS})\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)}_{\boldsymbol{S}}(\boldsymbol{SHS}) &= \underbrace{\boldsymbol{SS}}_{\boldsymbol{S}}\boldsymbol{HS}\\ &= \boldsymbol{SHS}.\end{aligned}$$ □

Lemma C.32. *Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the economy spectral decomposition of $(\boldsymbol{SHS})$, where $\boldsymbol{\Lambda}$ is an $(N-D)$-by-$(N-D)$ matrix, holding the non-zero eigenvalues of $(\boldsymbol{SHS})$, and $\boldsymbol{U}$ is the N-by-$(N-D)$ matrix, holding the corresponding $N-D$ eigenvectors of $(\boldsymbol{SHS})$ as columns. Then $\boldsymbol{S}=\boldsymbol{U}\boldsymbol{U}^\top$ [Patterson and Thompson, 1971].*

Proof.

$$\begin{aligned}\boldsymbol{S} &= (\boldsymbol{SHS})\left(\boldsymbol{P}^\top\boldsymbol{H}^{-1}\boldsymbol{P}\right)\\ &= (\boldsymbol{SHS})(\boldsymbol{SHS})^\dagger\\ &= \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\right)\\ &= \boldsymbol{U}\boldsymbol{\Lambda}\underbrace{\boldsymbol{U}^\top\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\\ &= \boldsymbol{U}\underbrace{\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}}_{\boldsymbol{I}}\boldsymbol{U}^\top\\ &= \boldsymbol{U}\boldsymbol{U}^\top.\end{aligned}$$ □

D. Derivations for FaST linear mixed models

Here we provide additional derivations that are used in Chapter 3.

D.1. Derivation of the low-rank quadratic form

Let $\boldsymbol{K}$ be a rank k genetic similarity matrix whose spectral decomposition can be written

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2^\top = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top + \boldsymbol{U}_2\left[\boldsymbol{0}\right]\boldsymbol{U}_2^\top = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top, \tag{D.1}$$

where

$$\boldsymbol{U} = \left[\boldsymbol{U}_1, \boldsymbol{U}_2\right], \tag{D.2}$$

The N-times-k matrix $\boldsymbol{U}_1$ contains the eigenvectors corresponding to non-zero eigenvalues, and the N-times-$N-k$ matrix $\boldsymbol{U}_2$. Using the fact that $\boldsymbol{U}$ is a normal N-times-N matrix, that is, $\boldsymbol{U}^{-1} = \boldsymbol{U}^\top$, we have

$$\boldsymbol{I}_N = \boldsymbol{U}\boldsymbol{U}^\top = \left[\boldsymbol{U}_1, \boldsymbol{U}_2\right]\left[\boldsymbol{U}_1, \boldsymbol{U}_2\right]^\top = \boldsymbol{U}_1\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{U}_2^\top. \tag{D.3}$$

Solving Equation (D.3) for $\boldsymbol{U}_2\boldsymbol{U}_2^\top$, we get

$$\boldsymbol{U}_2\boldsymbol{U}_2^\top = \boldsymbol{I}_N - \boldsymbol{U}_1\boldsymbol{U}_1^\top. \tag{D.4}$$

Further, because the columns of $\boldsymbol{U}$ are orthonormal, it follows that

$$\boldsymbol{I}_N = \boldsymbol{U}^\top\boldsymbol{U}, \tag{D.5}$$

$$\boldsymbol{I}_k = \boldsymbol{U}_1^\top\boldsymbol{U}_1, \tag{D.6}$$

$$\boldsymbol{I}_{N-k} = \boldsymbol{U}_2^\top\boldsymbol{U}_2. \tag{D.7}$$

Let $\boldsymbol{a} \equiv (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Our goal is to efficiently evaluate $\boldsymbol{a}^\top\left(\boldsymbol{K} + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{a}$. Substituting the spectral decomposition for $\boldsymbol{K}$ into this expression, we have

$$\boldsymbol{a}^\top\left(\boldsymbol{K} + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{a} = \left(\boldsymbol{U}^\top\boldsymbol{a}\right)^\top\left(\gamma\boldsymbol{\Lambda} + \boldsymbol{I}\right)^{-1}\left(\boldsymbol{U}^\top\boldsymbol{a}\right). \tag{D.8}$$

Using Equation (D.2), we can stack the matrix product in blocks involving $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to re-write this expression as

$$\begin{bmatrix}\boldsymbol{U}_1^\top\boldsymbol{a} & \boldsymbol{U}_2^\top\boldsymbol{a}\end{bmatrix}^\top\begin{bmatrix}\left(\gamma\boldsymbol{\Lambda}_1 + \boldsymbol{I}_k\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N-k}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_1^\top\boldsymbol{a} & \boldsymbol{U}_2^\top\boldsymbol{a}\end{bmatrix}. \tag{D.9}$$

As the off-diagonal blocks of the central matrix are equal to zero, the quadratic form reduces to the sum of two terms, namely

$$\left(\boldsymbol{U}_1^\top \boldsymbol{a}\right)^\top (\gamma \boldsymbol{\Lambda}_1 + \boldsymbol{I}_k)^{-1} \left(\boldsymbol{U}_1^\top \boldsymbol{a}\right) + \left(\boldsymbol{U}_2^\top \boldsymbol{a}\right)^\top \boldsymbol{I}_{N-k} \left(\boldsymbol{U}_2^\top \boldsymbol{a}\right). \tag{D.10}$$

Substituting $\boldsymbol{U}_2^\top \boldsymbol{U}_2$ for $\boldsymbol{I}_{N-k}$ (using Equation (D.7)), the second term becomes

$$\begin{aligned}\left(\boldsymbol{U}_2^\top \boldsymbol{a}\right)^\top \boldsymbol{I}_{N-k} \left(\boldsymbol{U}_2^\top \boldsymbol{a}\right) =& \boldsymbol{a}^\top \boldsymbol{U}_2 \boldsymbol{I}_{N-k} \boldsymbol{U}_2^\top \boldsymbol{a} \\ =& \boldsymbol{a}^\top \boldsymbol{U}_2 \left(\boldsymbol{U}_2^\top \boldsymbol{U}_2\right) \boldsymbol{U}_2^\top \boldsymbol{a}.\end{aligned} \tag{D.11}$$

Finally, using Equation (D.4), we can eliminate $\boldsymbol{U}_2$ to obtain

$$\left(\boldsymbol{U}_2 \boldsymbol{U}_2^\top \boldsymbol{a}\right)^\top \left(\boldsymbol{U}_2 \boldsymbol{U}_2^\top \boldsymbol{a}\right) = \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1 \boldsymbol{U}_1^\top\right) \boldsymbol{a}\right)^\top \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1 \boldsymbol{U}_1^\top\right) \boldsymbol{a}\right). \tag{D.12}$$

Substituting the expression (D.12) into (D.10), we obtain $\boldsymbol{a}^\top (\gamma \boldsymbol{K} + \boldsymbol{I})^{-1} \boldsymbol{a}$ equals

$$\left(\boldsymbol{U}_1^\top \boldsymbol{a}\right)^\top (\gamma \boldsymbol{\Lambda}_1 + \boldsymbol{I}_k)^{-1} \left(\boldsymbol{U}_1^\top \boldsymbol{a}\right) + \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1 \boldsymbol{U}_1^\top\right) \boldsymbol{a}\right)^\top \left(\left(\boldsymbol{I}_N - \boldsymbol{U}_1 \boldsymbol{U}_1^\top\right) \boldsymbol{a}\right). \tag{D.13}$$

Substituting $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ for $\boldsymbol{a}$, we obtain Equation (3.24).

D.2. FaST compressed linear mixed models

Let $\boldsymbol{Z}$ be the $n \times g$ group indicator matrix that assigns each of n individuals to exactly one group. Let $\boldsymbol{K}$ be the g-by-g group similarity matrix. Then the genetic similarity matrix in compressed mixed models is $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$. In order to combine the ideas of compression and FaST-LMM, we need to compute the spectral decomposition of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$.

If $\boldsymbol{K}$ factors into $\boldsymbol{G}\boldsymbol{G}^\top$, where $\boldsymbol{G}$ is a $g \times s_\mathrm{c}$ matrix of SNP data, Then the genetic similarity matrix in compressed mixed models is $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{Z}^\top$. In this case the spectral decomposition of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$ can be found as described in Section D.2.1. If $\boldsymbol{K}$ does not factor, then the spectral decomposition of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$ can be found as described in Section D.2.2.

D.2.1. Spectral decomposition of the compressed similarity matrix, when the group similarity matrix factors

For our argument, we use the fact that, given a matrix $\boldsymbol{A}$, both $\boldsymbol{A}\boldsymbol{A}^\top$ as well as $\boldsymbol{A}^\top\boldsymbol{A}$ share the same eigenvalues, and that these eigenvalues are given by the square of the singular values of $\boldsymbol{A}$. The eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$ are given by the left, and the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ are given by the right singular vectors of $\boldsymbol{A}$ respectively.

So the eigenvalues of $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{Z}^\top$ are the same as the eigenvalues of

$$\boldsymbol{G}^\top\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{G} = \boldsymbol{G}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}\boldsymbol{G}.$$

Using the same argument, the latter matrix has the same eigenvalues as

$$(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}\boldsymbol{G}\boldsymbol{G}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}.$$

These eigenvalues are given by the square of the singular values of $(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}\boldsymbol{G}$, where $(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}$ is a $g \times g$ diagonal matrix holding the square root of the number of members of each group on the diagonal. Because $(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}$ is diagonal, multiplication can be done in $O(gs_c)$ time.

Let $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^\top$ be the SVD of $(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}\boldsymbol{G}$. Then the following holds:

$$\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^\top \boldsymbol{Z}^\top = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}\boldsymbol{G}\boldsymbol{G}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\boldsymbol{Z}^\top, \tag{D.14}$$

where $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}$ is a $g \times g$ diagonal matrix, holding one over the square root of the number of members of each group on its diagonal. Substituting $(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}\boldsymbol{G}$ by its SVD, we get

$$\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^\top \boldsymbol{V}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\boldsymbol{Z}^\top. \tag{D.15}$$

Finally, by orthonormality of $\boldsymbol{V}$, this expression simplifies to

$$\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}}\boldsymbol{\Lambda}\tilde{\boldsymbol{U}}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\boldsymbol{Z}^\top, \tag{D.16}$$

where $\boldsymbol{\Lambda} = \tilde{\boldsymbol{\Lambda}}^2$ is a diagonal matrix, holding the non-zero eigenvalues of $\boldsymbol{G}\boldsymbol{G}^\top$ on its diagonal. The columns of $\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}}$ are orthonormal, as can be seen by

$$\tilde{\boldsymbol{U}}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\boldsymbol{Z}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{U}}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{U}}^\top \tilde{\boldsymbol{U}} = \boldsymbol{I}_g. \tag{D.17}$$

where we have once again used the fact that $(\boldsymbol{Z}^\top \boldsymbol{Z})$ is diagonal. It follows that the eigenvectors of $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^\top \boldsymbol{Z}^\top$ are given by

$$\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}},$$

completing the spectral decomposition of $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{G}^\top \boldsymbol{Z}$.

Note that the rotation of the data by $(\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}})^\top$ can be done efficiently by multiplying the data by the transpose of the rows of $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}}$ belonging to the respective cluster.

D.2.2. Spectral decomposition of the compressed similarity matrix, when the group similarity matrix does not factor

Here we extend the arguments in Section D.2 to any positive semi-definite $g \times g$ group genetic similarity matrix $\boldsymbol{K}$. In this case, the spectral decomposition of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ can also be computed efficiently, namely from the spectral decomposition of

$$(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}\boldsymbol{K}(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2} = \tilde{\boldsymbol{U}}\boldsymbol{\Lambda}\tilde{\boldsymbol{U}}^\top,$$

which can be computed in $O(g^3)$ runtime. As $\boldsymbol{K}$ is positive semi-definite, there always exists some square root $\boldsymbol{G}$ of $\boldsymbol{K}$, such that $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{G}^\top$. In Section D.2, we have shown, that $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$ and $(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}\boldsymbol{K}(\boldsymbol{Z}^\top \boldsymbol{Z})^{1/2}$ have the same eigenvalues. Consequently, we can compute the eigenvalues $\boldsymbol{\Lambda}$ of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$ from the spectral decomposition $\tilde{\boldsymbol{U}}\boldsymbol{\Lambda}\tilde{\boldsymbol{U}}^\top$. Analogous to the derivation in Equations (D.14)–(D.16), it follows that the eigenvectors of $\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}^\top$ are $\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}\tilde{\boldsymbol{U}}$, where by Equation (D.17), the columns are orthonormal.

E. Kronecker Product Derivations

Here we provide derivations used for high-dimensional linear mixed models described in Chapter 6.

E.1. Kronecker product identities

In the following, we repeatedly use two well-known facts about Kronecker products [Bernstein, 2009, Petersen and Pedersen, 2006].

E.1.1. Vectorization of Kronecker products

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}, \tag{E.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times M}$, $\boldsymbol{B} \in \mathbb{R}^{P \times Q}$, $\boldsymbol{C} \in \mathbb{R}^{M \times R}$ and $\boldsymbol{D} \in \mathbb{R}^{Q \times S}$ are matrices.

Let vec be an operation that concatenates the columns of an $N \times M$ matrix into a vector of length $N \operatorname{cot} M$. Then the following three statements hold:

$$\begin{aligned} \operatorname{vec}(\boldsymbol{ABC}) &= (\boldsymbol{C}^\top \otimes \boldsymbol{A})\operatorname{vec}(\boldsymbol{B}) \\ \operatorname{vec}(\boldsymbol{ABC})^\top &= \operatorname{vec}(\boldsymbol{B})^\top (\boldsymbol{C} \otimes \boldsymbol{A}^\top) \\ (\boldsymbol{A} \otimes \boldsymbol{C})\operatorname{vec}(\boldsymbol{B}) &= \operatorname{vec}\left(\boldsymbol{CBA}^\top\right) \end{aligned} \tag{E.2}$$

E.1.2. Singular value decomposition of a Kronecker product

In the following, we make heavy use of the eigenvalue decomposition of $\boldsymbol{C} \otimes \boldsymbol{R} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is an $N \cdot G$-by-$N \cdot G$ orthonormal matrix, holding the eigenvectors of $\boldsymbol{C}$ and $\boldsymbol{\Lambda} = \boldsymbol{\Lambda_C} \otimes \boldsymbol{\Lambda_R}$ is an $N \cdot G$-by-$N \cdot G$ diagonal matrix, holding the corresponding eigenvalues on the diagonal. This decomposition can be efficiently obtained from the composition of the individual Kronecker terms (after some reordering), i.e. $\boldsymbol{C} \otimes \boldsymbol{R} = (\boldsymbol{U_C} \otimes \boldsymbol{U_R})(\boldsymbol{\Lambda_C} \otimes \boldsymbol{\Lambda_R})(\boldsymbol{U_C}^\top \otimes \boldsymbol{U_R}^\top)$. Then simply substituting in the definition of the singular value decomposition leads to

$$\boldsymbol{C} \otimes \boldsymbol{R} = (\boldsymbol{U_C} \otimes \boldsymbol{U_R})(\boldsymbol{\Lambda_C} \otimes \boldsymbol{\Lambda_R})(\boldsymbol{U_C}^\top \otimes \boldsymbol{U_R}^\top). \tag{E.3}$$

E.1.3. Efficient evaluation of covariance term inverse times a vector

The first expression exploiting the eigenvalue decomposition that is going to be used frequently is the product of some vectorized N-by-G matrix $\boldsymbol{A}$ and the inverse of a

Kronecker covariance with independent noise matrix [Stegle et al., 2011].

$$\begin{aligned}\left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \operatorname{vec}(\boldsymbol{A}) &= \left(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}}\right) \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right)^{-1} \left(\boldsymbol{U}_{\boldsymbol{C}}^\top \otimes \boldsymbol{U}_{\boldsymbol{R}}^\top\right) \operatorname{vec}(\boldsymbol{A}) \\ &= \left(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}}\right) \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right)^{-1} \operatorname{vec}\left(\underbrace{\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{A} \boldsymbol{U}_{\boldsymbol{C}}}_{\tilde{\boldsymbol{A}}}\right) \\ &= \left(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}}\right) \operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{A}}\right) \qquad \text{(E.4)} \\ &= \operatorname{vec}\left(\boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{A}}\right) \boldsymbol{U}_{\boldsymbol{C}}^\top\right), \qquad \text{(E.5)}\end{aligned}$$

with the $N \times G$ matrix $\boldsymbol{D}_{\boldsymbol{\Lambda}}$ defined as having entries

$$[\boldsymbol{D}_{\boldsymbol{\Lambda}}]_{r,c} = \frac{1}{[\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c} \cdot [\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} + \sigma^2},$$

for all $r \in \{1, \ldots, N\}$ and $c \in \{1, \ldots, G\}$.

E.2. Covariance estimation in matrix-variate random effects models

Here, we give details of an efficient implementation of the tensor random effects model derived in Section 6.2.

The basic model

We start with a general form a random effects model where the covariance has a Kronecker structure:

$$\mathcal{L}\left(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{R}, \boldsymbol{C}, \sigma^2\right) = \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}); \operatorname{vec}(\boldsymbol{M}), \boldsymbol{C}(\boldsymbol{\Theta}_{\boldsymbol{C}}) \otimes \boldsymbol{R}(\boldsymbol{\Theta}_{\boldsymbol{R}}) + \sigma^2 \boldsymbol{I}\right). \qquad \text{(E.6)}$$

Here, $\boldsymbol{Y}$ is the data matrix with N rows (samples) and G columns (features). We defined $\boldsymbol{R}(\boldsymbol{\Theta}_{\boldsymbol{R}})$ as the row "row covariance" of the data matrix and $\boldsymbol{C}(\boldsymbol{\Theta}_{\boldsymbol{C}})$ corresponds to the "column covariance".

For notational convenience we will drop the dependence of the covariance matrices on additional hyperparameters $\boldsymbol{\Theta}_{\boldsymbol{R}}$ and $\boldsymbol{\Theta}_{\boldsymbol{C}}$, respectively. Furthermore, we will make the simplifying assumption that $\boldsymbol{C}$ is kept constant, i.e. has no parameters that need to be adapted during learning. Importantly, this is no restriction for general solutions as all calculations can be performed with respect to other covariance as well, for example iteratively optimizing hyperparameters of $\boldsymbol{R}$ and $\boldsymbol{C}$ in turn.

To implement parameter optimization of the covariance parameters of the model in Equation (E.6), we require efficient evaluation of the marginal likelihood and the gradients with respect to hyperparameters.

E.2.1. Efficient evaluation of the log likelihood

The term we want to evaluate is the log-likelihood, given by the log of the multivariate Normal density

$$\begin{aligned}\log \mathcal{L}\left(\boldsymbol{M}, \boldsymbol{\Theta}_{\boldsymbol{R}}, \boldsymbol{\Theta}_{\boldsymbol{C}}, \sigma^2\right) = & \quad \log \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \operatorname{vec}(\boldsymbol{M}) ; \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right) \\ = & \quad -\frac{N \cdot D}{2} \log 2\pi - \frac{1}{2} \underbrace{\log |\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}_{N \cdot D}|}_{\text{log-det}} \\ & -\frac{1}{2} \underbrace{\operatorname{vec}(\boldsymbol{Y}-\boldsymbol{M})^{\top} (\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}_{N \cdot D})^{-1} \operatorname{vec}(\boldsymbol{Y}-\boldsymbol{M})}_{\text{squared form}} .\end{aligned} \tag{E.7}$$

We derive efficient solutions for the logarithm of the determinant of $\boldsymbol{C} \otimes \boldsymbol{R}$ and the squared form separately.

Efficient evaluation of the log-det

Assuming that we have the eigenvalue decompositions for $\boldsymbol{R}$ and $\boldsymbol{C}$, the logarithm of the determinant can be written as

$$\begin{aligned}\log |\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}_{N \cdot G}| &= \log |(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}})(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}})(\boldsymbol{U}_{\boldsymbol{C}}^{\top} \otimes \boldsymbol{U}_{\boldsymbol{R}}^{\top}) + \sigma^2 \boldsymbol{I}_{N \cdot G}| \\ &= \log |(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}})| + \log |(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}_{N \cdot G})| + \log |\boldsymbol{U}_{\boldsymbol{C}}^{\top} \otimes \boldsymbol{U}_{\boldsymbol{R}}^{\top}| \\ &= \log |(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}_{N \cdot D})| \\ &= \sum_{r=1}^{N} \sum_{c=1}^{G} -\log([\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} \cdot [\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c} + \sigma^2).\end{aligned} \tag{E.8}$$

This term can be evaluated in $O(N \cdot G)$.

Efficient evaluation of the squared form

The squared form in the log marginal likelihood can be evaluated efficiently using the expression in Equation (E.4). Writing the residuals $\boldsymbol{Y}_{\mathrm{r}} = \boldsymbol{Y} - \boldsymbol{M}$, the squared form can be evaluated as

$$\begin{aligned}\operatorname{vec}(\boldsymbol{Y}_{\mathrm{r}})^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \operatorname{vec}(\boldsymbol{Y}_{\mathrm{r}}) &= \operatorname{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y}_{\mathrm{r}} \boldsymbol{U}_{\boldsymbol{C}}\right)^{\top} (\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}}) \operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right) \\ &= \operatorname{vec}\left(\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^{\top} \operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right) \\ &= \sum_{r=1}^{N} \sum_{c=1}^{G} \frac{\left[\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right]_{r,c}^2}{[\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} \cdot [\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c} + \sigma^2}\end{aligned} \tag{E.9}$$

As this term involves the multiplication of the data matrix $\boldsymbol{Y}_{\mathrm{r}}$ with $\boldsymbol{U}_{\boldsymbol{C}}$ and $\boldsymbol{U}_{\boldsymbol{R}}^{\top}$, it can be evaluated in $O(N^2 G + N G^2)$.

E.2.2. Efficient evaluation of the gradients of covariance parameters

Here, the aim is to evaluate the gradient of Equation (E.6) w.r.t. a particular covariance parameter θ.

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{M}, \boldsymbol{C}(\theta_{\boldsymbol{C}}), \boldsymbol{R}(\theta_{\boldsymbol{R}}), \sigma^2\right)}{\partial \theta} = -\frac{1}{2}\frac{\partial \log |\boldsymbol{C} + \sigma^2 \boldsymbol{I}|}{\partial \theta} - \frac{1}{2}\frac{\partial \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top (\boldsymbol{C} + \sigma^2 \boldsymbol{I})^{-1} \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})}{\partial} \tag{E.10}$$

The derivative consists of the derivative of the determinant term and the derivative of the squared form. Each of these is given separately for the cases where θ is either the noise parameter σ^2, or a row covariance parameter $\theta_{\boldsymbol{R}} \in \boldsymbol{\Theta}_{\boldsymbol{R}}$. The case of a column covariance parameter $\theta_{\boldsymbol{C}} \in \boldsymbol{\Theta}_{\boldsymbol{C}}$ is analogous to the case of a row covariance parameter and therefore is omitted for brevity.

E.2.3. Derivatiaves w.r.t. noise variance σ^2

Here, we provide the gradient of Equation (E.6) w.r.t. the noise parameter σ^2.

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{M}, \boldsymbol{C}, \boldsymbol{R}, \sigma^2\right)}{\partial \sigma^2} = -\frac{1}{2}\frac{\partial \log |\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}|}{\partial \sigma^2} - \frac{1}{2}\frac{\partial \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top (\boldsymbol{C} + \sigma^2 \boldsymbol{I})^{-1} \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})}{\partial \sigma^2}$$

Derivatiaves of the log-det term w.r.t. noise variance σ^2

$$\begin{aligned}
\frac{\partial \log |\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}|}{\partial \sigma^2} &= \mathrm{tr}\left[(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I})^{-1} \frac{\partial (\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I})}{\partial \sigma^2}\right] \\
&= \mathrm{tr}\left[(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I})^{-1} \frac{\partial (\sigma^2 \boldsymbol{I})}{\partial \sigma^2}\right] \\
&= \mathrm{tr}\left[\boldsymbol{U}(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{U}^\top \boldsymbol{I}\right] \\
&= \mathrm{tr}\left[(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1}\right] \\
&= \sum_{r=1}^{N} \sum_{c=1}^{G} \frac{1}{[\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} \cdot [\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c} + \sigma^2}.
\end{aligned} \tag{E.11}$$

Squared form derivatives w.r.t. noise variance σ^2

$$\frac{\partial \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})}{\partial \sigma^2}$$

The derivative only affects the covariance term

$$\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top \frac{\partial \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1}}{\partial \sigma^2} \mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}}).$$

Using $\frac{\partial \boldsymbol{A}^{-1}}{\partial \sigma^2} = -\boldsymbol{A}^{-1}\left[\frac{\partial \boldsymbol{A}}{\partial \sigma^2}\right]\boldsymbol{A}^{-1}$, this becomes

$$-\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^{\top}\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\left(\frac{\partial\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)}{\partial\sigma^2}\right)\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}}).$$

The remaining matrix derivative equals the identity matrix and vanishes, leaving

$$-\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^{\top}\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}}).$$

In this term we use the Kronecker expression (E.4) on both sides, giving

$$-\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^{\top}\left(\boldsymbol{U}_{\boldsymbol{C}}^{\top}\otimes\boldsymbol{U}_{\boldsymbol{R}}^{\top}\right)\left(\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}}\right)\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right).$$

Using orthogonality of the eigenvectors, the middle part vanishes, leaving the vector product

$$-\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^{\top}\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right).$$

This can be evaluated as the sum

$$\sum_{r=1}^{N}\sum_{c=1}^{G}\frac{\left[\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right]_{r,c}^{2}}{\left(\left[\boldsymbol{\Lambda}_{\boldsymbol{R}}\right]_{r,r}\cdot\left[\boldsymbol{\Lambda}_{\boldsymbol{C}}\right]_{c,c}+\sigma^2\right)^{2}}. \tag{E.12}$$

E.2.4. Derivatives w.r.t a row covariance parameter

Here, we provide the gradient of Equation (E.6) w.r.t. a particular row covariance parameter $\theta_{\boldsymbol{R}} \in \boldsymbol{\Theta}_{\boldsymbol{R}}$.

$$\frac{\partial \log\mathcal{L}\left(\boldsymbol{M},\boldsymbol{C},\boldsymbol{R},\sigma^2\right)}{\partial\theta_{\boldsymbol{R}}} = -\frac{1}{2}\frac{\partial\log\left|\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right|}{\partial\theta_{\boldsymbol{R}}} - \frac{1}{2}\frac{\partial\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^{\top}(\boldsymbol{C}+\sigma^2\boldsymbol{I})^{-1}\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})}{\partial\theta_{\boldsymbol{R}}} \tag{E.13}$$

Derivatives of the determinant w.r.t. a row covariance parameter $\theta_{\boldsymbol{R}}$

$$\begin{aligned}
\frac{\partial \log|\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}|}{\partial\theta_{\boldsymbol{R}}} =&\mathrm{tr}\left((\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I})^{-1}\frac{\partial(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I})}{\partial\theta_{\boldsymbol{R}}}\right)\\
=&\mathrm{tr}\left[(\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}})(\boldsymbol{\Lambda}+\sigma^2)^{-1}(\boldsymbol{U}_{\boldsymbol{C}}^\top\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top)(\boldsymbol{C}\otimes\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}})\right].\\
&\text{Using the identity } (\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D})=\boldsymbol{AC}\otimes\boldsymbol{BD} \text{ this equals}\\
&\mathrm{tr}\left[(\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}})(\boldsymbol{\Lambda}+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{U}_{\boldsymbol{C}}^\top\boldsymbol{C}\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}})\right]\\
&\text{Using } \mathrm{tr}(\boldsymbol{AB})=\mathrm{tr}(\boldsymbol{BA}), \text{ this becomes}\\
&\mathrm{tr}\left[(\boldsymbol{\Lambda}+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{U}_{\boldsymbol{C}}^\top\boldsymbol{C}\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}})(\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}})\right]\\
=&\mathrm{tr}\left[(\boldsymbol{\Lambda}+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{U}_{\boldsymbol{C}}^\top\boldsymbol{C}\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}})\right]\\
=&\mathrm{tr}\left[(\boldsymbol{\Lambda}_{\boldsymbol{C}}\otimes\boldsymbol{\Lambda}_{\boldsymbol{R}}+\sigma^2\boldsymbol{I})^{-1}\Big(\boldsymbol{\Lambda}_{\boldsymbol{C}}\otimes(\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}})\Big)\right]\\
=&\mathrm{diag}\left((\boldsymbol{\Lambda}_{\boldsymbol{C}}\otimes\boldsymbol{\Lambda}_{\boldsymbol{R}}+\sigma^2\boldsymbol{I})^{-1}\right)^\top\Big(\mathrm{diag}(\boldsymbol{\Lambda}_{\boldsymbol{C}})\otimes\mathrm{diag}(\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}})\Big)
\end{aligned} \tag{E.14}$$

As this derivation only involves the trace, we just need the diagonal of the Kronecker product, which only involves the diagonal of $(\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}})$.

Derivatives of the squared form w.r.t. a row covariance parameter $\theta_{\boldsymbol{R}}$

$$\begin{aligned}
&\frac{\partial\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})}{\partial\theta_{\boldsymbol{R}}}\\
=&\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top\frac{\partial\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}}{\partial\theta_{\boldsymbol{R}}}\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})\\
&\text{using } \frac{\partial\boldsymbol{A}^{-1}}{\partial\theta_{\boldsymbol{R}}}=-\boldsymbol{A}^{-1}\frac{\partial\boldsymbol{A}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{A}^{-1} \text{ this becomes}\\
=&-\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})^\top\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\left(\frac{\partial\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)}{\partial\theta_{\boldsymbol{R}}}\right)\left(\boldsymbol{C}\otimes\boldsymbol{R}+\sigma^2\boldsymbol{I}\right)^{-1}\mathrm{vec}(\boldsymbol{Y}_{\mathrm{r}})\\
&\text{using the efficient Kronecker expression (E.4) on both sides, this becomes}\\
=&-\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top\left(\boldsymbol{U}_{\boldsymbol{C}}^\top\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top\right)\left(\boldsymbol{C}\otimes\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\right)(\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}})\,\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right).\\
=&-\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top\left(\boldsymbol{U}_{\boldsymbol{C}}^\top\boldsymbol{C}\boldsymbol{U}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}}\right)\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right).\\
=&-\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top\left(\boldsymbol{\Lambda}_{\boldsymbol{C}}\otimes\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}}\right)\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right).\\
=&-\mathrm{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top\mathrm{vec}\left(\left(\boldsymbol{U}_{\boldsymbol{R}}^\top\frac{\partial\boldsymbol{R}}{\partial\theta_{\boldsymbol{R}}}\boldsymbol{U}_{\boldsymbol{R}}\right)\left(\boldsymbol{D}_{\boldsymbol{\Lambda}}\odot\tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)\boldsymbol{\Lambda}_{\boldsymbol{C}}\right).
\end{aligned} \tag{E.15}$$

Derivatives for the Kronecker GPLVM

In latent variable models, the covariance parameters represent the values of a latent variables of a given row (or column) of $\boldsymbol{Y}$. These latent variables serve as inputs to a kernel function a possibly non-linear positive semi-definite kernel function $\mathrm{k}(.,.)$ that describe the covariance between any two rows (or columns). Models of this kind are an extension of the Gaussian process latent variable model [Lawrence, 2004, 2005].

Let the N-by-K matrix holding the K latent dimensions of all of N rows be $\boldsymbol{G_R}$ ($\boldsymbol{G_C}$ for columns). For example for the case, where the row-covariance is modeled in this way, $\boldsymbol{R}$ is defined as follows.

$$\boldsymbol{R} = \begin{bmatrix} \mathrm{k}\left([\boldsymbol{G_R}]_{1,:}, [\boldsymbol{G_R}]_{1,:}\right) & \dots & \mathrm{k}\left([\boldsymbol{G_R}]_{1,:}, [\boldsymbol{G_R}]_{N,:}\right) \\ \vdots & \ddots & \vdots \\ \mathrm{k}\left([\boldsymbol{G_R}]_{N,:}, [\boldsymbol{G_R}]_{1,:}\right) & \dots & \mathrm{k}\left([\boldsymbol{G_R}]_{N,:}, [\boldsymbol{G_R}]_{N,:}\right) \end{bmatrix} \tag{E.16}$$

In this case, each parameter $\theta_{\boldsymbol{R}}$ represents a value of a single entry $[\boldsymbol{G}]_{r,k}$ of $\boldsymbol{G_R}$ for $r \in [1, \dots, N]$ and $k \in [1, \dots, K]$. The derivative of a kernel value with respect to $\theta_{\boldsymbol{R}}$ follows from applying the chain rule to each argument of k.

$$\frac{\partial \mathrm{k}\left([\boldsymbol{G_R}]_{i,:}, [\boldsymbol{G_R}]_{j,:}\right)}{\partial \theta_{\boldsymbol{R}}} = \frac{\partial \mathrm{k}\left([\boldsymbol{G_R}]_{i,:}, [\boldsymbol{G_R}]_{j,:}\right)}{\partial [\boldsymbol{G_R}]_{i,k}} \frac{\partial [\boldsymbol{G_R}]_{i,k}}{\partial \theta_{\boldsymbol{R}}} + \frac{\partial \mathrm{k}\left([\boldsymbol{G_R}]_{i,:}, [\boldsymbol{G_R}]_{j,:}\right)}{\partial [\boldsymbol{G_R}]_{j,k}} \frac{\partial [\boldsymbol{G_R}]_{i,k}}{\partial \theta_{\boldsymbol{R}}},$$

where the derivatives of the entries of $\boldsymbol{G_R}$ are zero except for $[\boldsymbol{G_R}]_{r,k}$.

$$\begin{aligned} \tfrac{\partial [\boldsymbol{G_R}]_{i,k}}{\partial \theta_{\boldsymbol{R}}} &= 1 \quad \text{for } i = r. & \qquad \tfrac{\partial [\boldsymbol{G_R}]_{j,k}}{\partial \theta_{\boldsymbol{R}}} &= 1 \quad \text{for } j = r. \\ \tfrac{\partial [\boldsymbol{G_R}]_{i,k}}{\partial \theta_{\boldsymbol{R}}} &= 0 \quad \text{for } i \neq r. & \qquad \tfrac{\partial [\boldsymbol{G_R}]_{j,k}}{\partial \theta_{\boldsymbol{R}}} &= 0 \quad \text{for } j \neq r. \end{aligned}$$

We observe, that in this case the derivative of $\theta_{\boldsymbol{R}}$ only has an effect on the r^{th} row and the r^{th} column of $\boldsymbol{R}$.

As the kernel is symmetric in both arguments, the derivative of the log-likelihood can be computed using only using a row vector $\boldsymbol{d}_{r,k}$ holding the non-zero derivatives with respect to the first input of the kernel w.r.t. $[\boldsymbol{G_R}]_{r,k}$.

$$\boldsymbol{d}_{r,k} = \begin{bmatrix} \frac{\partial \mathrm{k}\left([\boldsymbol{G_R}]_{r,:}, [\boldsymbol{G_R}]_{1,:}\right)}{\partial [\boldsymbol{G_R}]_{r,k}}, & \dots, & \frac{\partial \mathrm{k}\left([\boldsymbol{G_R}]_{r,:}, [\boldsymbol{G_R}]_{N,:}\right)}{\partial [\boldsymbol{G_R}]_{r,k}} \end{bmatrix}. \tag{E.17}$$

In the efficient derivative evaluations provided above, the derivative of $\boldsymbol{R}$ always appears rotated by $\boldsymbol{U_R}$ from both sides.

$$\boldsymbol{U_R}^\top \frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}} \boldsymbol{U_R} = [\boldsymbol{U_R}]_{r,:}^\top \boldsymbol{d}_{r,k} \boldsymbol{U_R} + \boldsymbol{U_R}^\top \boldsymbol{d}_{r,k}^\top [\boldsymbol{U_R}]_{r,:} \tag{E.18}$$

Derivative of the log-determinant Let $\theta_{\boldsymbol{R}}$ be the value of $[\boldsymbol{G_R}]_{r,k}$. Using the derivatives in (E.17), we can evaluate the derivative of the log-determinant.

$$
\begin{aligned}
&\frac{\partial \log |\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}|}{\partial \theta_{\boldsymbol{R}}} \\
=&\operatorname{diag}\left((\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1}\right)^\top \left(\operatorname{diag}(\boldsymbol{\Lambda}_{\boldsymbol{C}}) \otimes \operatorname{diag}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}} \boldsymbol{U}_{\boldsymbol{R}}\right)\right) \\
=&\operatorname{diag}\left((\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1}\right)^\top \left(\operatorname{diag}(\boldsymbol{\Lambda}_{\boldsymbol{C}}) \otimes \operatorname{diag}\left([\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}^\top \boldsymbol{d}_{r,k} \boldsymbol{U}_{\boldsymbol{R}} + \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{d}_{r,k}^\top [\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}\right)\right) \\
=&2\operatorname{diag}\left((\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I})^{-1}\right)^\top \left(\operatorname{diag}(\boldsymbol{\Lambda}_{\boldsymbol{C}}) \otimes \left([\boldsymbol{U}_{\boldsymbol{R}}]_{r,:} \odot (\boldsymbol{d}_{r,k} \boldsymbol{U}_{\boldsymbol{R}})\right)^\top\right). \qquad \text{(E.19)}
\end{aligned}
$$

Derivative of the squared form Let $\theta_{\boldsymbol{R}}$ be the value of $[\boldsymbol{G}_{\boldsymbol{R}}]_{r,k}$. Using the derivatives in (E.17), we can evaluate the derivative of the log-determinant.

$$
\begin{aligned}
&-\operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top \operatorname{vec}\left(\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \frac{\partial \boldsymbol{R}}{\partial \theta_{\boldsymbol{R}}} \boldsymbol{U}_{\boldsymbol{R}}\right)\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right) \boldsymbol{\Lambda}_{\boldsymbol{C}}\right) \\
=&-\operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top \operatorname{vec}\left(\left([\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}^\top \boldsymbol{d}_{r,k} \boldsymbol{U}_{\boldsymbol{R}} + \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{d}_{r,k} [\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}\right)\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right) \boldsymbol{\Lambda}_{\boldsymbol{C}}\right) \\
=&-\operatorname{tr}\left(\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top \left([\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}^\top \boldsymbol{d}_{r,k} \boldsymbol{U}_{\boldsymbol{R}} + \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{d}_{r,k}^\top [\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}\right)\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right) \boldsymbol{\Lambda}_{\boldsymbol{C}}\right) \\
=&-2\boldsymbol{d}_{r,k} \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right) \boldsymbol{\Lambda}_{\boldsymbol{C}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_{\mathrm{r}}\right)^\top [\boldsymbol{U}_{\boldsymbol{R}}]_{r,:}^\top. \qquad \text{(E.20)}
\end{aligned}
$$

E.3. Efficient computations for matrix-variate linear mixed models

In the matrix-variate linear mixed model the mean $\operatorname{vec}(\boldsymbol{M})$ in the basic model (E.6) is modeled as the sum of J terms.

$$
\begin{aligned}
&\log \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \operatorname{vec}\left(\sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top\right);\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right) \\
&\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \sum_j \boldsymbol{A}_j \otimes \boldsymbol{X}_j \operatorname{vec}(\boldsymbol{B}_j);\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right) \qquad \text{(E.21)}
\end{aligned}
$$

We define the complete design matrix Φ as the horizontal concatenation of all the Kronecker design terms

$$
\Phi = [\boldsymbol{A}_1 \otimes \boldsymbol{X}_1, \ldots, \boldsymbol{A}_J \otimes \boldsymbol{X}_J] \qquad \text{(E.22)}
$$

and the column vector of all fixed effects $\boldsymbol{\beta}$ as the vertical concatenation of all vectorized weight matrices

$$
\boldsymbol{\beta} = \begin{bmatrix} \operatorname{vec}(\boldsymbol{B}_1) \\ \vdots \\ \operatorname{vec}(\boldsymbol{B}_J) \end{bmatrix}.
$$

$$
\mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \Phi\boldsymbol{\beta};\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)
$$

In the matrix-variate mixed model the mean $\mathrm{vec}\,(\boldsymbol{M})$ is given by a linear term in the basic model (E.6)

E.3.1. Efficient evaluation of the matrix-variate mixed model likelihood

$$\log \mathcal{N}\left(\mathrm{vec}\,(\boldsymbol{Y}) \,\middle|\, \sum_{j=1}^{J} \boldsymbol{A}_j \otimes \boldsymbol{X}_j \mathrm{vec}\,(\boldsymbol{B}_j)\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)$$

Apply the vec-trick to the mean term:

$$\log \mathcal{N}\left(\mathrm{vec}\,(\boldsymbol{Y}) \,\middle|\, \mathrm{vec}\left(\sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}^\top\right)\,;\, \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)$$

Writing out the log normal distribution this equals

$$-\frac{C \cdot R}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}| - \frac{1}{2}\mathrm{vec}\,(\boldsymbol{Y}_\mathrm{r})^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \mathrm{vec}\,(\boldsymbol{Y}_\mathrm{r})\,, \quad \text{(E.23)}$$

where the matrix of residuals of $\boldsymbol{Y}$ in this context is defined as

$$\boldsymbol{Y}_\mathrm{r} = \boldsymbol{Y} - \underbrace{\sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top}_{\boldsymbol{M}}\,. \quad \text{(E.24)}$$

Using the spectral decomposition of the covariance matrix, the log likelihood (Equation (E.23)) and be written using residuals rotated by the eigenvectors of the covariance matrix.

$$-\frac{C \cdot R}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Lambda} + \sigma^2 \boldsymbol{I}| - \frac{1}{2}\mathrm{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_\mathrm{r} \boldsymbol{U}_{\boldsymbol{C}}\right)^\top \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right)^{-1} \mathrm{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_\mathrm{r} \boldsymbol{U}_{\boldsymbol{C}}\right), \quad \text{(E.25)}$$

E.3.2. Derivative of the rotated residual term

$$\begin{aligned}
\frac{\partial}{\partial [\boldsymbol{B}_k]_{a,b}}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_\mathrm{r} \boldsymbol{U}_{\boldsymbol{C}}\right) &= \frac{\partial}{\partial [\boldsymbol{B}_k]_{a,b}}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} - \boldsymbol{U}_{\boldsymbol{R}}^\top \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top \boldsymbol{U}_{\boldsymbol{C}}\right) \\
&= \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_k \frac{\partial \boldsymbol{B}_k}{\partial [\boldsymbol{B}_k]_{a,b}} \boldsymbol{A}_k^\top \boldsymbol{U}_{\boldsymbol{C}} \\
&= \boldsymbol{U}_{\boldsymbol{R}}^\top [\boldsymbol{X}_k]_{:,a} [\boldsymbol{A}_k]_{:,b}^\top \boldsymbol{U}_{\boldsymbol{C}} \quad \text{(E.26)}
\end{aligned}$$

E.3.3. Estimation of the fixed effects

Score vector for fixed effects

From the definition of $\boldsymbol{\beta}$ in Equation (6.56), a single fixed effect can be identified as a single entry $[\boldsymbol{B}_k]_{a,b}$ of the fixed effect matrix $\boldsymbol{B}_j$.

$$\begin{aligned}
&\frac{\nabla}{\nabla \boldsymbol{\beta}} \log \mathcal{N}\left(\operatorname{vec}(\boldsymbol{Y}) \,\middle|\, \Phi\boldsymbol{\beta}; \boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right) \\
=&\frac{\nabla}{\nabla \boldsymbol{\beta}} - \frac{1}{2} \left(\operatorname{vec}(\boldsymbol{Y}) - \Phi\boldsymbol{\beta}\right)^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \left(\operatorname{vec}(\boldsymbol{Y}) - \Phi\boldsymbol{\beta}\right) \\
=&\Phi^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \underbrace{\left(\operatorname{vec}(\boldsymbol{Y}) - \Phi\boldsymbol{\beta}\right)}_{\operatorname{vec}\left(\boldsymbol{Y} - \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top\right)}
\end{aligned}$$

The right term of this expression for the gradient can easily be identified as the vectorized matrix of residuals $\boldsymbol{Y}_\mathrm{r}$.

Efficient evaluation of the gradient with respect to a single fixed effect

$$\begin{aligned}
&\frac{\partial}{\partial [\boldsymbol{B}_k]_{a,b}} \left(-\frac{1}{2} \operatorname{vec}(\boldsymbol{Y}_\mathrm{r})^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \operatorname{vec}(\boldsymbol{Y}_\mathrm{r}) \right) \\
&\quad - \operatorname{vec}(\boldsymbol{Y}_\mathrm{r})^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \operatorname{vec}\left(\frac{\partial \boldsymbol{Y}_\mathrm{r}}{\partial [\boldsymbol{B}_k]_{a,b}}\right)
\end{aligned}$$

Using the efficient Kronecker inverse formula (E.4), we get

$$- \operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_\mathrm{r}\right)^\top \left(\boldsymbol{U}_{\boldsymbol{C}}^\top \otimes \boldsymbol{U}_{\boldsymbol{R}}^\top\right) \operatorname{vec}\left(\frac{\partial \boldsymbol{Y}_\mathrm{r}}{\partial [\boldsymbol{B}_k]_{a,b} r}\right).$$

Resolving the derivative of the data term on the right yields

$$\operatorname{vec}\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_\mathrm{r}\right)^\top \operatorname{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top [\boldsymbol{X}_k]_{:,a} [\boldsymbol{A}_k]_{:,b}^\top \boldsymbol{U}_{\boldsymbol{C}}\right).$$

This scalar value can be identified as the trace of the product of the two matrices.

$$- \operatorname{tr}\left(\left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_\mathrm{r}\right)^\top \boldsymbol{U}_{\boldsymbol{R}}^\top [\boldsymbol{X}_k]_{:,a} [\boldsymbol{A}_k]_{:,b}^\top \boldsymbol{U}_{\boldsymbol{C}}\right).$$

Using the properties $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A})$ and $\operatorname{tr}(\boldsymbol{A}^\top) = \operatorname{tr}(\boldsymbol{A})$, this equals the scalar

$$- [\boldsymbol{X}_k]_{:,a}^\top \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_\mathrm{r}\right) \boldsymbol{U}_{\boldsymbol{C}}^\top [\boldsymbol{A}_k]_{:,b}. \tag{E.27}$$

When stacking together the derivatives for all a and b the gradient with respect to all entries of $\boldsymbol{B}_k$ follows as

$$\begin{aligned}
&\frac{\nabla}{\nabla \boldsymbol{B}_k} \left(-\frac{1}{2} \operatorname{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_\mathrm{r} \boldsymbol{U}_{\boldsymbol{C}}\right)^\top \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right)^{-1} \operatorname{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{Y}_\mathrm{r} \boldsymbol{U}_{\boldsymbol{C}}\right) \right) \\
=& - \boldsymbol{X}_k^\top \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}_\mathrm{r}\right) \boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_k \qquad \text{(E.28)} \\
=& \boldsymbol{X}_k^\top \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{C} \times \boldsymbol{R}} \odot \left(\sum_{j=1}^{J} \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j \boldsymbol{B}_j \boldsymbol{A}_j^\top \boldsymbol{U}_{\boldsymbol{C}} \right) \right) \boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_k^\top - \boldsymbol{X}_k^\top \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D}_{\boldsymbol{\Lambda}} \odot \tilde{\boldsymbol{Y}}\right) \boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_k^\top \qquad \text{(E.29)}
\end{aligned}$$

E.3.4. Closed form maximum likelihood estimate of the fixed effects

By setting the gradient of the log-likelihood with respect to the weight vector to zero, we can solve for the maximum-likelihood weight estimate $\boldsymbol{\beta}_{\mathrm{M}}$:

$$\frac{\nabla}{\nabla \boldsymbol{\beta}_{\mathrm{M}}} \left(-\frac{1}{2} \left(\mathrm{vec}\left(\boldsymbol{Y}\right) - \Phi\boldsymbol{\beta} \right)^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \left(\mathrm{vec}\left(\boldsymbol{Y}\right) - \Phi\boldsymbol{\beta} \right) \right) = \mathbf{0} \tag{E.30}$$

$$\Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \Phi \boldsymbol{\beta}_{\mathrm{M}} - \Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \mathrm{vec}\left(\boldsymbol{Y}\right) = \mathbf{0} \tag{E.31}$$

$$\Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \Phi \boldsymbol{\beta} = \Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \mathrm{vec}\left(\boldsymbol{Y}\right) \tag{E.32}$$

$$\boldsymbol{\beta}_{\mathrm{M}} = \Big(\underbrace{\Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \Phi}_{\text{left term}} \Big)^{-1} \underbrace{\Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \mathrm{vec}\left(\boldsymbol{Y}\right)}_{\text{right term}} \tag{E.33}$$

Kronecker structures are used to to simplify this expression for both terms separately.

Right term

The column vector can be written as the vertical concatenation of

$$\Phi^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \mathrm{vec}\left(\boldsymbol{Y}\right) = \begin{bmatrix} \mathrm{vec}\left(\boldsymbol{X}_1^T \boldsymbol{U}_{\boldsymbol{R}} (\boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} \odot \boldsymbol{D}) \boldsymbol{U}_{\boldsymbol{C}}^{\top} \boldsymbol{A}_1 \right) \\ \vdots \\ \mathrm{vec}\left(\boldsymbol{X}_J^T \boldsymbol{U}_{\boldsymbol{R}} (\boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} \odot \boldsymbol{D}) \boldsymbol{U}_{\boldsymbol{C}}^{\top} \boldsymbol{A}_J \right) \end{bmatrix} \tag{E.34}$$

For the j^{th} term, this looks as follows:

$$\left(\boldsymbol{A}_j \otimes \boldsymbol{X}_j \right)^{\top} \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I} \right)^{-1} \mathrm{vec}\left(\boldsymbol{Y}\right) \tag{E.35}$$

$$\left(\boldsymbol{A}_j^{\top} \otimes \boldsymbol{X}_j^{\top} \right) \left(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}} \right) \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I} \right)^{-1} \left(\boldsymbol{U}_{\boldsymbol{C}}^{\top} \otimes \boldsymbol{U}_{\boldsymbol{R}}^{\top} \right) \mathrm{vec}\left(\boldsymbol{Y}\right) \tag{E.36}$$

$$\left(\boldsymbol{A}_j^{\top} \otimes \boldsymbol{X}_j^{\top} \right) \left(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}} \right) \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I} \right)^{-1} \mathrm{vec}\left(\boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} \right) \tag{E.37}$$

$$\left(\boldsymbol{A}_j^{\top} \otimes \boldsymbol{X}_j^{\top} \right) \left(\boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{U}_{\boldsymbol{R}} \right) \mathrm{vec}\left(\boldsymbol{D} \odot \boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} \right) \tag{E.38}$$

$$\left(\boldsymbol{A}_j^{\top} \otimes \boldsymbol{X}_j^{\top} \right) \mathrm{vec}\left(\boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D} \odot \boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} \right) \boldsymbol{U}_{\boldsymbol{C}}^{\top} \right) \tag{E.39}$$

$$\mathrm{vec}\left(\boldsymbol{X}_j^{\top} \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{D} \odot \boldsymbol{U}_{\boldsymbol{R}}^{\top} \boldsymbol{Y} \boldsymbol{U}_{\boldsymbol{C}} \right) \boldsymbol{U}_{\boldsymbol{C}}^{\top} \boldsymbol{A}_j \right) \tag{E.40}$$

Left term

$$\Phi^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} \Phi \tag{E.41}$$

This matrix to be inverted has a block structure, where the $(i,j)^{\text{th}}$ block involves the i^{th} and the j^{th} fixed effects:

$$(\boldsymbol{A}_i \otimes \boldsymbol{X}_i)^\top \left(\boldsymbol{C} \otimes \boldsymbol{R} + \sigma^2 \boldsymbol{I}\right)^{-1} (\boldsymbol{A}_j \otimes \boldsymbol{X}_j)$$

$$\left(\boldsymbol{A}_i^\top \boldsymbol{U}_{\boldsymbol{C}} \otimes \boldsymbol{X}_i^\top \boldsymbol{U}_{\boldsymbol{R}}\right) \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{R}} + \sigma^2 \boldsymbol{I}\right)^{-1} \left(\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_j \otimes \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j\right)$$

$$\sum_{r=1}^{N} \sum_{c}^{G} \frac{1}{[\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c}[\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} + \sigma^2} \left([\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_i]_{c,:}^\top \otimes [\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_i]_{r,:}^\top\right) \left([\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_j]_{c,:} \otimes [\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j]_{r,:}\right)$$

$$\sum_{r=1}^{N} \sum_{c}^{G} \frac{1}{[\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c}[\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} + \sigma^2} \left([\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_i]_{c,:} [\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_j]_{c,:}\right) \otimes \left([\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j]_{r,:}^\top [\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j]_{r,:}\right)$$

This can be simplified in two different ways, depending on which one is cheaper to compute. If the number of target variables G is smaller than the number os samples N ($G < N$), we use

$$\sum_{c=1}^{G} \left([\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_i]_{c,:}^\top [\boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}_j]_{c,:}\right) \otimes \left(\boldsymbol{X}_i^\top \boldsymbol{U}_{\boldsymbol{R}} \left(\boldsymbol{\Lambda}_{\boldsymbol{R}} [\boldsymbol{\Lambda}_{\boldsymbol{C}}]_{c,c} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j\right) \tag{E.42}$$

else, we use

$$\sum_{r=1}^{N} \left(\boldsymbol{A}_i^\top \boldsymbol{U}_{\boldsymbol{C}} \left(\boldsymbol{\Lambda}_{\boldsymbol{C}} [\boldsymbol{\Lambda}_{\boldsymbol{R}}]_{r,r} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{U}_{\boldsymbol{C}}^\top \boldsymbol{A}\right) \otimes \left([\boldsymbol{X}_i^\top \boldsymbol{U}_{\boldsymbol{R}}]_{:,r} [\boldsymbol{U}_{\boldsymbol{R}}^\top \boldsymbol{X}_j]_{r,:}\right) \tag{E.43}$$

Let $D = \sum_{j=1}^{J} D_j$ be the total number of row effects and $M = \sum_{j=1}^{J} M_j$ be the total number of column effects. Then Equation (E.42) takes $O(ND^2 + NGM^2)$ time to evaluate and Equation (E.43)takes $O(GM^2 + NGD^2)$ time to evaluate. Inversion of the whole term takes $O(D^3 M^3)$.

List of Tables

List of Figures

Bibliography

G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97–101, 2001.

A. Agresti. *Categorical data analysis*, volume 359. Wiley-interscience, 2002.

G. Allen and R. Tibshirani. Inference with transposable data: Modeling the effects of row and column correlations. *Arxiv preprint arXiv:1004.0209*, 2010.

H. L. Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467 (7317):832–838, 2010.

D. M. Altshuler, E. S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, et al. A map of human genome variation from population scale sequencing. *Nature*, (467):1061–1073, 2010.

M. A. Alvarez and N. D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12(5):1459–1500, 2011.

N. Amin, C. M. van Duijn, and Y. S. Aulchenko. A genomic background based method for association analysis in related individuals. *PloS one*, 2(12):e1274, 2007.

W. Astle and D. J. Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24:451–471, 2009.

W. J. Astle. *Population Structure abd Cryptic Relatedness in Genetic Association Studies*. PhD thesis, Imperial College, London, 2009.

S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

Y. S. Aulchenko and H. C. de Koning, Dirk-Jan. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177:577–585, September 2007.

Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

X. Autosomes Chromosome. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:1, 2012.

D. J. Balding. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, 7:781–791, Oct 2006.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

V. Bansal, O. Libiger, A. Torkamani, and N. J. Schork. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785, 2010.

B. J. Bennett, C. R. Farber, L. Orozco, H. M. Kang, A. Ghazalpour, N. Siemers, M. Neubauer, I. Neuhaus, R. Yordanova, B. Guan, et al. A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome research*, 20 (2):281–290, 2010.

D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

D. Berrar, W. Dubitzky, and M. Granzow, editors. *A Practical Approach to Microarray Data Analysis*. Kluwer, Norwell, MA, 2003.

J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237, 2013.

S. Bolormaa, B. Hayes, K. Savin, R. Hawken, W. Barendse, P. Arthur, R. Herd, and M. Goddard. Genome-wide association studies for feedlot and growth traits in cattle. *Journal of animal science*, 89(6):1684–1697, 2011.

E. Bonilla, K. Chai, and C. Williams. Multi-task Gaussian process prediction. *Advances in Neural Information Processing Systems*, 20:153–160, 2008.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

R. Braun and K. Buetow. Pathways of distinction analysis: a new technique for multi–SNP analysis of GWAS data. *PLoS genetics*, 7(6):e1002101, 2011.

M. Bulmer. *Francis Galton: pioneer of heredity and biometry*. Johns Hopkins University Press, 2003.

D. Burbridge. Francis galton on twins, heredity and social class. *The British Journal for the History of Science*, 34(03):323–340, 2001.

P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

A. Buse. The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, 36(3a):153–157, 1982.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

C. D. Campbell, E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman, L. C. Groop, D. Altshuler, K. G. Ardlie, and J. N. Hirschhorn. Demonstrating stratification in a european american population. *Nature genetics*, 37(8):868–872, 2005.

J. Cao, K. Schneeberger, S. Ossowski, T. Günther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, et al. Whole-genome sequencing of multiple arabidopsis thaliana populations. *Nature genetics*, 43(10):956–963, 2011.

H. Chen, J. B. Meigs, and J. Dupuis. Sequence kernel association test for quantitative traits in family samples. *Genetic epidemiology*, 37(2):196–204, 2013.

H. Chial. Rare genetic disorders: Learning about genetic disease through gene mapping, SNPs, and microarray data. *Nature Education*, 1(1), 2008.

H. D. Daetwyler, B. Villanueva, and J. A. Woolliams. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, 3(10):e3395, 2008.

A. P. Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265274, 1981.

P. I. de Bakker, M. A. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics*, 17(R2):R122–R128, 2008.

G. de los Campos, D. Gianola, and D. B. Allison. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12): 880–886, 2010.

E. Demidenko. *Mixed Models: Theory and Applications*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2004.

A. P. Dempster, C. M. Patel, M. R. Selwyn, and A. J. Roth. Statistical and computational aspects of mixed model analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(2):203–214, 1984.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55: 997–1004, Dec 1999.

C. B. Do, J. Y. Tung, E. Dorfman, A. K. Kiefer, E. M. Drabant, U. Francke, J. L. Mountain, S. M. Goldman, C. M. Tanner, J. W. Langston, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. *PLoS genetics*, 7(6):e1002141, 2011.

A. Dominicus, A. Skrondal, H. K. Gjessing, N. L. Pedersen, and J. Palmgren. Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior genetics*, 36(2): 331–340, 2006.

T. R. Dreszer, D. Karolchik, A. S. Zweig, A. S. Hinrichs, B. J. Raney, R. M. Kuhn, L. R. Meyer, M. Wong, C. A. Sloan, K. R. Rosenbloom, et al. The ucsc genome browser database: extensions and updates 2011. *Nucleic acids research*, 40(D1):D918–D923, 2012.

D. B. Dunson. *Random effect and latent variable model selection*, volume 192. Springer, 2008.

P. Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

H. Edenberg, L. Bierut, P. Boyce, M. Cao, S. Cawley, R. Chiles, K. Doheny, M. Hansen, T. Hinrichs, K. Jones, et al. Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC genetics*, 6(Suppl 1):S2, 2005.

G. B. Ehret, P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S.-J. Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478 (7367):103–9, 2011.

M.-A. Enoch, Z. Zhou, M. Kimura, D. C. Mash, Q. Yuan, and D. Goldman. Gabaergic gene expression in postmortem hippocampus from alcoholics and cocaine addicts; corresponding findings in alcohol-naive p and np rats. *PloS one*, 7(1):e29369, 2012.

M. P. Epstein, A. S. Allen, and G. A. Satten. A simple and improved correction for population stratification in case-control studies. *American journal of human genetics*, 80(5):921, 2007.

W. J. Ewens and R. S. Spielman. The transmission/disequilibrium test: history, subdivision, and admixture. *American journal of human genetics*, 57(2):455, 1995.

R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

J. Flint and T. F. Mackay. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome research*, 19(5):723–733, 2009.

S. D. Foster, A. P. Verbyla, and W. S. Pitchford. Incorporating lasso effects into a mixed model for quantitative trait loci detection. *Journal of agricultural, biological, and environmental statistics*, 12(2):300–314, 2007.

A. Franke, D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125, 2010.

M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato, et al. Assessing the impact of population stratification on genetic association studies. *Nature genetics*, 36(4):388–393, 2004.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

N. A. Furlotte, E. Y. Kang, A. Van Nas, C. R. Farber, A. J. Lusis, and E. Eskin. Increasing association mapping power and resolution in mouse genetic studies through the use of meta-analysis for structured populations. *Genetics*, 191(3):959–967, 2012.

N. Fusi, O. Stegle, and N. D. Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS computational biology*, 8(1):e1002330, 2012.

N. Fusi, C. Lippert, K. Borgwardt, N. D. Lawrence, and O. Stegle. Detecting regulatory gene-environment interactions with unmeasured environmental factors. *Bioinformatics*, 2013.

F. Galton. The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society of London*, 61(369-377):401–413, 1897.

F. Galton. A diagram of heredity. *Nature*, 57:293, 1898.

S. F. Galton. *Hereditary genius*. Macmillan and Company, 1869.

M. E. Goddard, N. R. Wray, K. Verbyla, and P. M. Visscher. Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, 24(4):517–529, 2009.

S. Greven. *Non-standard problems in inference for additive and linear mixed models*. Cuvillier Verlag, 2007.

S. Greven, C. M. Crainiceanu, H. Küchenhoff, and A. Peters. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4), 2008.

A. J. Griffiths, S. R. Wessler, R. C. Lewontin, W. M. Gelbart, D. T. Suzuki, and J. H. Miller. *An introduction to genetic analysis*. wH Freeman, 2004.

B. Han and E. Eskin. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, 88(5):586–598, 2011.

O. J. Hardy and X. Vekemans. Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, 2 (4):618–620, 2002.

H. O. Hartley and J. N. K. Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1 and 2):93–108, June 1967.

D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biokatrika*, 61(2):383–385, August 1974.

T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

B. J. Hayes, P. Visscher, and M. Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(01):47–60, 2009.

C. Henderson. *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, 1984.

C. R. Henderson. Estimation of genetic parameters. *The Annals of Mathematical Statistics*, (2):309–310, June 1950.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*, 4 (7):e1000130, 2008.

M. Holden, S. Deng, L. Wojnowski, and B. Kulle. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, 24 (23):2784–2785, 2008.

M. W. Horton, A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone, B. J. Vilhjálmsson, et al. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the regmap panel. *Nature genetics*, 44(2):212–216, 2012.

X. Huang, X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics*, 42(11):961–967, 2010.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 107, 2008.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, 42:348–354, Apr 2010.

W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

F. R. Kidd, Judith R.and Friedlaender, W. C. Speed, A. J. Pakstis, F. M. De La Vega, and K. K. Kidd. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics*, 2(1): 1–13, 2011.

H. S. Kim, J. D. Minna, and M. A. White. GWAS meets tcga to illuminate mechanisms of cancer predisposition. *Cell*, 152(3):387–389, 2013.

S. Kim, V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature genetics*, 39(9):1151–1155, 2007.

G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, pages 495–502, 1970.

W. C. Knowler, R. Williams, D. Pettitt, and A. Steinberg. Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in american indians with genetic admixture. *American journal of human genetics*, 43(4):520, 1988.

T. Kollo and D. von Rosen. *Advanced Multivariate Statistics with Matrices*. Springer, 2005.

A. Korte, B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9):1066–1071, 2012.

C. G. Lambert and L. J. Black. Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics*, 13(2):195–203, 2012.

E. S. Lander et al. The new genomics: global views of biology. *Science*, 274(5287): 536–539, 1996.

D. A. Lawlor, R. M. Harbord, J. A. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16(329-336):3, 2004.

N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

S. Lee, F. A. Wright, and F. Zou. Control of population stratification by correlation-selected principal components. *Biometrics*, 67(3):967–974, 2011.

S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 2012a.

S. Lee, M. C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012b.

S. H. Lee, J. H. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS genetics*, 4(10):e1000231, 2008.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

J. Li, K. Das, G. Fu, R. Li, and R. Wu. The Bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523, 2011.

D.-Y. Lin and Z.-Z. Tang. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3): 354–367, 2011.

C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8 (10):833–835, 2011.

C. Lippert, J. Listgarten, R. I. Davidson, S. Baxter, H. Poong, C. M. Kadie, and D. Heckerman. An exhaustive epistatic snp association analysis on expanded wellcome trust data. *Scientific reports*, 3, 2013a.

C. Lippert, G. Quon, J. Listgarten, and D. Heckerman. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports*, (In press), 2013b.

J. Listgarten, C. Kadie, E. E. Schadt, and D. Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 107: 16465–16470, Sep 2010.

J. Listgarten, C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):3–4, 2012.

J. Listgarten, C. Lippert, and D. Heckerman. Fast-lmm select for addressing confounding from spatial structure and rare variants. *Nature Genetics*, (In press), 2013a.

J. Listgarten, C. Lippert, E. Y. Kang, X. Jing, C. M. Kadie, and D. Heckerman. A powerful and efficient set test for genetic markers that handles confounding. *Bioinformatics*, 2013b.

D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

I. Lobo. Multifactorial inheritance and genetic disease. *Nature Education*, 1(1), 2008.

M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA, 1998.

B. Maher. The case of the missing heritability. *Nature*, 456(7218):18–21, 2008.

G. Malécot. *Les mathématiques de l'hérédité*. Masson, 1948.

N. Malo, O. Libiger, and N. J. Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375–385, 2008.

M. Malosetti, C. G. van der Linden, B. Vosman, and F. A. van Eeuwijk. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. *Genetics*, 175(2):879–889, 2007.

T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

G. Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

I. Mathieson and G. McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 2012.

J. McClellan and M.-C. King. Genetic heterogeneity in human disease. *Cell*, 141(2): 210–217, 2010.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

G. Mendel. Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, 44, 1866.

P. Menéndez, Y. Kourmpetis, C. Ter Braak, and F. van Eeuwijk. Gene regulatory networks from multifactorial perturbations using graphical lasso: Application to the dream4 challenge. *PLoS One*, 5(12):e14147, 2010.

T. Meuwissen, B. Hayes, and M. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

K. Meyer. Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics*, pages 153–165, 1985.

T. H. Morgan, A. H. Sturtevant, H. J. Muller, and C. B. Bridges. *The mechanism of Mendelian heredity*. Holt, 1922.

N. E. Morton. Sequential tests for the detection of linkage. *American journal of human genetics*, 7(3):277, 1955.

R. Nassir, R. Kosoy, C. Tian, P. White, L. Butler, G. Silva, R. Kittles, M. Alarcon-Riquelme, P. Gregersen, J. Belmont, et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC genetics*, 10(1):39, 2009.

A. Nejati-Javaremi, C. Smith, and J. Gibson. Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of Animal Science*, 75(7): 1738–1745, 1997.

J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

U. Ober, J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. Mackay, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. *PLoS genetics*, 8 (5):e1002685, 2012.

K. Oualkacha, Z. Dastani, R. Li, P. E. Cingolani, T. D. Spector, C. J. Hammond, J. B. Richards, A. Ciampi, and C. M. Greenwood. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology*, 2013.

M. Palma, E. Ristori, E. Ricevuto, G. Giannini, A. Gulino, et al. Brca1 and brca2: the genetic testing and the current management options for mutation carriers. *Critical reviews in oncology/hematology*, 57(1):1, 2006.

J.-H. Park, S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*, 42(7):570–575, 2010.

J.-H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni Jr, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, December 1971.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190+, December 2006.

J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

F. Pereyra, X. Jia, P. J. McLaren, A. Telenti, P. I. de Bakker, B. D. Walker, S. Ripke, C. J. Brumme, S. L. Pulit, M. Carrington, et al. The major genetic determinants of hiv-1 control affect hla class i peptide presentation. *Science (New York, NY)*, 330 (6010):1551, 2010.

K. B. Petersen and M. S. Pedersen. The matrix cookbook. Technical report, Technical University of Denmark, 2006.

M. Pirinen, P. Donnelly, and C. C. Spencer. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *arXiv preprint arXiv:1207.4886*, 2012.

A. Platt, M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio, N. W. Mulyati, J. Ågren, O. Bossdorf, D. Byers, K. Donohue, et al. The scale of population structure in Arabidopsis thaliana. *PLoS genetics*, 6(2):e1000843, 2010a.

A. Platt, B. J. Vilhjálmsson, and M. Nordborg. Conditions under which genome-wide association studies will be positively misleading. *Genetics*, 186(3):1045–1052, 2010b.

R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3):313–326, 1964.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, August 2006.

A. L. Price, J. Butler, N. Patterson, C. Capelli, V. L. Pascali, F. Scarnicci, A. Ruiz-Linares, L. Groop, A. A. Saetta, P. Korkolopoulou, et al. Discerning the ancestry of european americans in genetic association studies. *PLoS genetics*, 4(1):e236, 2008.

A. L. Price, G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L.-J. Wei, and S. R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics*, 86(6):832, 2010a.

A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, 11:459–463, Jun 2010b.

J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease–common variant or not? *Human molecular genetics*, 11(20):2417–2423, 2002.

J. K. Pritchard and N. A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal of Human Genetics*, 65(1):220–228, 1999.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000a.

J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American journal of human genetics*, 67(1):170, 2000b.

B. M. Psaty, C. J. O'Donnell, V. Gudnason, K. L. Lunetta, A. R. Folsom, J. I. Rotter, A. G. Uitterlinden, T. B. Harris, J. C. Witteman, E. Boerwinkle, et al. Cohorts for heart and aging research in genomic epidemiology (charge) consortium design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics*, 2(1):73–80, 2009.

K. Puniyani, S. Kim, and E. P. Xing. Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics*, 26(12):i208–i216, 2010.

G. Quon, C. Lippert, D. Heckerman, and J. Listgarten. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic acids research*, 41(4): 2095–2104, 2013.

B. Rakitsch, C. Lippert, H. Topa, K. Borgwardt, A. Honkela, and O. Stegle. A mixed model approach for joint genetic analysis of alternatively spliced transcript isoforms using rna-seq data. *arXiv preprint arXiv:1210.2850*, 2012.

B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2): 206–214, 2013.

P. Rantakallio. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scandinavica*, 193:Suppl–193, 1969.

C. Rao, Radhakrishna. Least squares theory using an estiamted dispersion matrix and its application to measurement of signals. In L. M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability*, pages 355–372. University of California Press, 1967.

C. R. Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4):447–458, 1965.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, December 2005. ISBN 026218253X.

M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. Genecards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664, 1998.

D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510, 2001.

N. Risch, K. Merikangas, et al. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

G. K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

S. N. Roy. *Some Aspects of Multivariate Analysis*. Wiley, 1957.

C. Sabatti et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46, 2008.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523, 2005.

E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

L. Schaeffer, J. Wilton, and R. Thompson. Simultaneous estimation of variance and covariance components from multitrait mixed model equations. *Biometrics*, pages 199–208, 1978.

J. Schelldorfer, P. Bühlmann, G. DE, and S. VAN. Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

E. D. Schifano, M. P. Epstein, L. F. Bielak, M. A. Jhun, S. L. Kardia, P. A. Peyser, and X. Lin. Snp set association analysis for familial data. *Genetic Epidemiology*, 36(8): 797–810, 2012.

B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001.

H. Schwender, I. Ruczinski, and K. Ickstadt. Testing SNPs and sets of SNPs for importance in association studies. *Biostatistics*, 12(1):18–32, 2011.

S. R. Searle et al. Matrix algebra for the biological sciences (including applications in statistics). *Matrix algebra for the biological sciences (including applications in statistics).*, 1966.

V. Segura, B. J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, and M. Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 2012.

S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

E. Setakis, H. Stirnadel, and D. J. Balding. Logistic regression protects against population structure in genetic association studies. *Genome research*, 16(2):290–296, 2006.

E. Smith and L. Kruglyak. Gene–environment interaction in yeast gene expression. *PLoS Biology*, 6(4):e83, 2008.

E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. L. Allen, C. M. Lindgren, J. Luan, R. Mägi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42 (11):937–948, 2010.

C. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5):e1000477, 2009.

M. S. Srivastava, T. von Rosen, and D. von Rosen. Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17 (4):357–370, 2008.

M. S. Srivastava, T. von Rosen, and D. von Rosen. Estimation and testing in general multivariate linear models with kronecker product covariance structure. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, pages 137–163, 2009.

E. A. Stahl, D. Wegmann, G. Trynka, J. Gutierrez-Achury, R. Do, B. F. Voight, P. Kraft, R. Chen, H. J. Kallberg, F. A. Kurreeman, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics*, 2012.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology*, 6(5):e1000770, 2010.

O. Stegle, C. Lippert, J. M. Mooij, N. D. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate Gaussian models with iid observation noise. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 630–638, 2011.

O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

G. R. Svishcheva, T. I. Axenovich, N. M. Belonogova, C. M. van Duijn, and Y. S. Aulchenko. Rapid variance components-based method for whole-genome association analysis. *Nature genetics*, 2012.

T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307): 707–713, 2010.

R. Thompson. The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics*, pages 527–550, 1973.

C. Tian, P. K. Gregersen, and M. F. Seldin. Accounting for ancestry: population substructure and genome-wide association studies. *Human molecular genetics*, 17(R2): R143–R150, 2008.

F. Tian, P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*, 43(2):159–162, 2011.

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

M. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J.R. Statistical Society, Series B*, 61:6111–622, 1999.

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38(8):879–887, 2006.

A. P. Verbyla and W. N. Venables. An extension of the growth curve model. *Biometrika*, 75(1):129–138, 1988.

B. Villanueva, R. Pong-Wong, J. Fernandez, and M. Toro. Benefits from marker-assisted selection under an additive polygenic genetic model. *Journal of animal science*, 83(8): 1747–1752, 2005.

S. V. N. Vishwanathan, N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

P. M. Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489–490, 2008.

P. M. Visscher, B. McEVOY, and J. Yang. From galton to GWAS: quantitative genetics of human height. *Genetics Research*, 92(5):371, 2011.

P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *American journal of human genetics*, 90(1):7, 2012.

D. Von Rosen. Maximum likelihood estimators in multivariate linear normal models. *Journal of multivariate analysis*, 31(2):187–200, 1989.

H. Wackernagel. *Multivariate geostatistics*. Springer Verlag, 2003.

N. Wade. A decade later, genetic map yields few new cures. *New York Times*, 12, 2010.

G. Wahba. *Spline models for observational data*, volume 59. Society for Industrial Mathematics, 1990.

K. Wang, H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, et al. Common genetic variants on 5p14. 1 associate with autism spectrum disorders. *Nature*, 459(7246):528–533, 2009.

L. Wang, P. Jia, R. D. Wolfinger, X. Chen, B. L. Grayson, T. M. Aune, and Z. Zhao. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics*, 27(5):686–692, 2011.

Y. Wang, R. Localio, and T. R. Rebbeck. Bias correction with a single null marker for population stratification in candidate gene association studies. *Human Heredity*, 59 (3):165–175, 2005.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, pages 60–62, 1938.

N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507–515, 2013.

S. Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645): 330–338, 1922.

S. Wright. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 19(6):506, 1934a.

S. Wright. The results of crosses between inbred strains of guinea pigs, differing in number of digits. *Genetics*, 19(6):537, 1934b.

M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

F. Yan, Z. Xu, and Y. A. Qi. Sparse matrix-variate Gaussian process blockmodels for network modeling. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 745–752, Corvallis, Oregon, 2011. AUAI Press.

J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76, 2011a.

J. Yang, M. N. Weedon, S. Purcell, G. Lettre, K. Estrada, C. J. Willer, A. V. Smith, E. Ingelsson, J. R. O'Connell, M. Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011b.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

K. Yeung and W. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763, 2001.

J. Yu, G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2005a.

J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, 38:203–208, Feb 2006.

K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd International Coference on Machine Learning*, 2005b.

K. Yu, J. Lafferty, S. Zhu, and Y. Hong. Large-scale collaborative prediction using a nonparametric random effects model. In *Proceedings of the 26th International Coference on Machine Learning*, 2009.

A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368, June 1962.

A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 58(298):977–992, December 1963.

S. Zhang, X. Zhu, and H. Zhao. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic epidemiology*, 24(1):44–56, 2002.

Y. Zhang and J. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. *Advances in Neural Information Processing Systems*, 23:2550–2558, 2010.

Z. Zhang, E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, 42:355–360, Apr 2010.

K. Zhao, M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, et al. An Arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1):e4, 2007.

X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.