

Bayesian Methods for Multivariate Phenotype Analysis in Genome-wide Association Studies



Valentina Iotchkova
Mansfield College

Supervisor: Jonathan Marchini

Department of Statistics, University of Oxford

A thesis submitted in a partial fulfillment for

the Degree of DPhil in Statistics

Trinity Term 2013

Bayesian Methods for Multivariate Phenotype Analysis in Genome-wide Association Studies

Valentina Iotchkova, Mansfield College
Department of Statistics, University of Oxford
D.Phil thesis, Trinity term 2013

Abstract

Most genome-wide association studies search for genetic variants associated to a single trait of interest, despite the main interest usually being the understanding of a complex genotype-phenotype network. Furthermore, many studies collect data on multiple phenotypes, each measuring a different aspect of the biological system under consideration, therefore it can often make sense to jointly analyze the phenotypes. However this is rarely the case and there is a lack of well developed methods for multiple phenotype analysis. Here we propose novel approaches for genome-wide association analysis, which scan the genome one SNP at a time for association with multivariate traits.

The first half of this thesis focuses on an analytic model averaging approach which bi-partitions traits into associated and unassociated, fits all such models and measures evidence of association using a Bayes factor. The discrete nature of the model allows very fine control of prior beliefs about which sets of traits are more likely to be jointly associated. Using simulated data we show that this method can have much greater power than simpler approaches that do not explicitly model residual correlation between traits. On real data of six hematological parameters in 3 population cohorts (KORA, UKNBS and TwinsUK) from the HaemGen consortium, this model allows us to

uncover an association at the RCL locus that was not identified in the original analysis but has been validated in a much larger study.

In the second half of the thesis we propose and explore the properties of models that use priors encouraging sparse solutions, in the sense that genetic effects of phenotypes are shrunk towards zero when there is little evidence of association. To do this we explore and use spike and slab (SAS) priors. All methods combine both hypothesis testing, via calculation of a Bayes factor, and model selection, which occurs implicitly via the sparsity priors. We have successfully implemented a Variational Bayesian approach to fit this model, which provides a tractable approximation to the posterior distribution, and allows us to approximate the very high-dimensional integral required for the Bayes factor calculation. This approach has a number of desirable properties. It can handle missing phenotype data, which is a real feature of most studies. It allows for both correlation due to relatedness between subjects or population structure and residual phenotype correlation. It can be viewed as a sparse Bayesian multivariate generalization of the mixed model approaches that have become popular recently in the GWAS literature. In addition, the method is computationally fast and can be applied to millions of SNPs for a large number of phenotypes. Furthermore we apply our method to 15 glycans from 3 isolated population cohorts (ORCADES, KORCULA and VIS), where we uncover association at a known locus, not identified in the original study but discovered later in a larger one. We conclude by discussing future directions.

Acknowledgements

Foremost, I would like to express my gratitude to my D.Phil supervisor Jonathan Marchini for his patience, continuous support and invaluable guidance throughout my D.Phil journey. I would also like to thank my examiners Simon Myers and Heather Cordell for taking the time to consider my thesis and the Department of Statistics that has provided the funding for my years of D.Phil, for which I am extremely grateful.

A big thank you also goes to my team members Androniki Menelaou, Claire Churchhouse, Elena Frangou, Marie Forest, Olivier Delaneau, Jared O’Connel and Jason Liu, who established a great work environment, shared many sleepless nights, provided helpful advice and created unforgettable moments throughout this journey. I am thankful also to all my non-academic friends and fellow Taekwon-do practitioners from OUTKD, ITF England and “back home”, especially Kostadin Dimitrov, Lyubina Spasova, Charlotte McDermott and Paul Knighton, for all the support and amazing time we have had together making England and Bulgaria proud!

I am deeply grateful to my family, for their unconditional love and support throughout these years, despite none of them actually knowing what I do. My auntie and uncle for making me laugh and always being there for me, my father whose memory and partially character are always with me, my mother and grandparents for the wisdom and advice, even such I might not follow.

Last, but not least, I would also like to thank my new team leaders, team members and friends, Nicole Soranzo, Ewan Birney, Ian Dunham, Max Cocca, Klaudia Walter, Louella Vasquez, Lu Chen, Jie Huang, Stephen Watt, Dirk Dolle, Helena Kilpinnen, Sander Timmer, Sandro Morganella, Nils Kolling, Ester Cuenca, Padhraig Gromley, Yang Luo and Lucy Crooks, for their immense patience and support throughout the most difficult moments of my D.Phil journey, for tolerating my quirky/odd little habits, for making me laugh and providing me with small getaways to recharge.

Contents

Contents	v
1 Introduction and Background	1
1.1 Motivation	1
1.2 Introduction	3
1.3 Genome-wide Association Studies	5
1.3.1 Introduction	5
1.3.2 History and Development of GWAS	7
1.3.3 Findings	9
1.4 The Concept of Pleiotropy	12
1.5 Current methods for independent samples	15
1.5.1 Single phenotype model	16
1.5.2 Multi-trait methods for unrelated samples	16
1.6 Missing Phenotype Data	21
1.7 Relatedness between samples and population structure	22
1.8 Methods modelling sample relatedness	23
1.8.1 Univariate mixed effects models	23
1.8.2 Multi-trait methods for related samples	25

1.9	Summary	27
2	A Bayesian model averaging approach	30
2.1	Introduction	31
2.2	Model	31
2.2.1	Model space	33
2.3	Bayesian model averaging	34
2.3.1	Priors	37
2.3.2	Marginal Likelihoods, Bayes Factors, Parameter posteriors	40
2.3.3	Model averaging across single phenotype models	42
2.4	Bayes Factor Threshold	44
2.5	Univariate approach	45
2.6	Multivariate extension of univariate approach	45
2.7	AVG Method	47
2.8	Combining results across cohorts	48
2.9	Summary	48
3	Simulation studies	50
3.1	Data simulation	50
3.2	Simulation scenarios	51
3.3	Power of Bayesian and Frequentist methods	53
3.3.1	Single trait association models	54
3.3.2	Pleiotropic Models	57
3.4	Bayes factor for pleiotropy	60
3.5	Conclusions	62

4	Real Data Analysis	63
4.1	Data overview	63
4.1.1	Pre-processing	64
4.1.2	Prior on models of association	66
4.2	Results	68
4.2.1	Results at established blood traits loci	69
4.2.2	Genome-Wide Scans	77
4.3	Summary	82
5	A sparse Bayesian multivariate mixed model	84
5.1	Model	85
5.1.1	Prior on precision or inverse error covariance	86
5.1.2	Spike and Slab (SAS) prior	87
5.1.3	Model parametrization	89
5.1.4	Bayes Factors	90
5.1.5	Variational Bayesian Inference	90
5.1.6	VB approximation	92
5.1.7	VB SAS updates	92
5.1.8	Method initialization	95
5.1.9	Model selection	96
5.2	Extension to allow related samples	97
5.2.1	Pre-whitening	99
5.2.2	VB Approximation and initialization	101
5.2.3	Choice of M	102
5.2.4	Modelling missing phenotype data	104

5.2.5	Algorithm	106
5.2.6	Discussion	107
6	Simulation studies	109
6.1	Unrelated samples	109
6.1.1	Data simulation	109
6.1.2	Simulation scenarios	110
6.1.3	Power, $P=2$	113
6.1.4	Power, $P=10$	117
6.1.5	Power, $P=100$	119
6.1.6	Missing data imputation	120
6.1.7	Effect size estimates	122
6.2	Related samples	126
6.2.1	Data simulation	126
6.2.2	Choice of M	127
6.2.3	Modelling related samples	130
6.3	Summary	132
7	Real Data Analysis	133
7.1	Glycans data	133
7.2	Univariate analysis	134
7.3	VB SAS	137
7.3.1	GWAS scans	138
7.3.2	Known associations	141
7.4	Summary	150

8 Summary and Discussion	152
8.1 Summary	152
8.2 Discussion	156
8.2.1 Alternative Statistics for Association	156
8.2.2 Method Extensions	159
8.3 Conclusions	161
Appendix A - Distributions	162
Appendix B - Model averaging (A) method priors	167
Appendix C - Marginal Likelihood, Full Model	171
Appendix D - Bayes Factor, non-Full Models	174
Appendix E - Variational Bayes methods	178
Appendix F - VB SAS for unrelated samples	181
Appendix G - VB SAS for related samples	189
References	201

Chapter 1

Introduction and Background

“We are drowning in information and starving for knowledge.”

– Roger D. Rutherford

1.1 Motivation

With genotyping and sequencing technologies rapidly evolving, there has been an explosion of genome-wide association studies (GWAS) [Hindorff et al., 2013], each analysing data on hundreds to thousands of individuals. They have shown great success in identifying novel genetic variants associated with a variety of complex traits and diseases [Hindorff et al., 2013, Manolio et al., 2009], despite the over-simplified approach they typically use. The majority of these loci tend to have moderate-to-small effects and explain only a small fraction of the total phenotypic heritability (proportion of phenotypic variance attributable to genetic differences) [Manolio et al., 2009, Eichler et al., 2010] and it is thought that with a few thousand samples GWAS may still be underpowered to uncover

1. Introduction and Background

the missing heritability [Gibson, 2010, Ku et al., 2010]. The term 'missing heritability' stands for the discrepancy between the heritability explained by the set of variants found to be associated to a trait of interest and the expected (or apparent) total heritability of that trait, where the latter is most commonly estimated from population data by assessing the phenotypic correlations between monozygotic and dizygotic twins (overview given in [Zuk et al., 2012]). From a modelling perspective, GWAS have been aimed at assessing the potential relationship between a single genetic variant and a single observed characteristic, even when multiple phenotypes with joint relevance to a disease of interest are often available on each individual - for example in studies of haematological traits [Soranzo et al., 2009], lipids [Demirkan et al., 2011, Demirkan et al., 2012] and metabolic traits [Illig et al., 2009]. Improvements in modelling can be the key to a better understanding of these complex genotype-phenotype networks and the increasing evidence that established loci can affect multiple (potentially correlated) traits simultaneously [Cotsapas et al., 2011] suggests that multivariate phenotype modelling can be advantageous. Furthermore, by utilizing information in multiple traits, in particular modelling their correlation, we can increase power to detect additional genetic variants with smaller effects which may not necessarily be found using traditional single-phenotype approaches [Stephens, 2013, O'Reilly et al., 2012].

In this Chapter we introduce the main concepts, projects and developments in the field and then discuss current multi-trait methods for GWAS, describing their modelling assumptions, advantages and disadvantages, and show the place of our proposed methods in comparison to them. The aim of this thesis is to

show the benefits of multivariate association analysis, understand when and why it would be most useful, present novel approaches for multivariate GWAS and describe their advantages over existing ones. In general, we attempt to move a step closer towards a more realistic way of addressing the biological problem in question, by developing statistical methods that we think are better suited than what is currently available.

1.2 Introduction

Every individual (and living organism) carries genetic (or heritable) information, which is encoded in specific parts of our DNA (or RNA for certain viruses). The complete set of such information is called genotype. Specifically, the human DNA consists of 23 pairs of chromosomes, made up from three billion base pair nucleotides - Adenine (A), Guanine (G), Thymine (T), and Cytosine (C). In addition, we also possess various phenotypic characteristics, or observable properties. These are thought to be determined, up to a certain degree, by our genetic information, but could also be influenced by environmental or other non-genetic factors (Figure 1.1).

Understanding the genotype-phenotype map is very important for future progress in fields such as biology and medicine. In particular, uncovering the mechanisms of diseases could help us develop better means of treatment or possibly even lead to

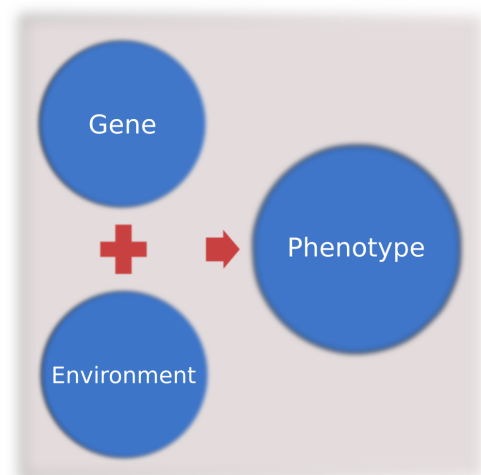


Figure 1.1: Phenotype, Genotype, Environment network

disease prevention [McCarthy et al., 2008].

The main objective thus is to learn about networks of the form of Figure 1.1. Clearly, if we have all possible genetic and non-genetic factors influencing a particular disease for example, then this problem can be modelled mathematically. As a result, an important question arises: how do we obtain environmental, phenotypic and genotypic information?

If we are interested in a particular (well-defined) disease, an obvious phenotype would be “an individual has/does not have the disease”. Such information is usually straightforward to obtain. In contrast, an interest in a particular disease symptom, such as “high blood pressure”, leads to a number of possible phenotypes. Two examples being the collections of either quantitative blood pressure measurements or binary “high”/“not high” blood pressure levels. On the other hand, even if we know exactly which trait we want to measure and how, this phenotype might not be easy (or even possible) to obtain.

The collection of environmental (and other non-genetic) factors also experiences the above-mentioned difficulties. However for both of these classes (environmental and phenotypic), there is another, much more general, issue beyond the definition of a particular phenotype and whether is it possible to measure it. Undoubtedly a more important question is: which traits and factors should we even consider? There is an infinite number of things that can be measured making it impossible to obtain full information about “everything”. Consequently, we rely on current knowledge, on diseases for example, to give us certain symptoms, characteristics and specific factors to attempt to record. Chances are, however, that there are

still many unknown related factors, which would make our task even harder.

In terms of genotypic information a number of genotyping techniques exist, for example DNA sequencing and DNA microarrays (Affymetrix, Illumina), which differ greatly in approach, cost, speed, accuracy and targeted genomic regions - from chips capturing specific positions of the genome to sequencing technologies that can read the genome base-by-base. With technologies rapidly evolving in this area, it is getting cheaper, faster and more accurate to extract the desired genotype information [Ragoussis, 2009]. It has to be noted here that genotyping errors can decrease the power for detecting associations between genetic variants and phenotypic measurements [Pompanon et al., 2005], however we will not discuss this any further within this work.

1.3 Genome-wide Association Studies

1.3.1 Introduction

Suppose we have a collection of genotypes, phenotypes and environmental factors on a number of individuals. The purpose of genetic association studies is to analyse this data in a statistical framework. They are aimed at revealing any signal between genetic variants and traits, and in their simplest form only consider one phenotype (most of the time binary) and a genotype at a single SNP for a number of individuals. A SNP or a single-nucleotide polymorphism is a single base pair mutation at a specific locus (usually consisting of two alleles), or in other words a position at which a single nucleotide differs between individuals.

1. Introduction and Background

Genetic association studies are used in order to investigate whether a genetic variant is associated with a particular trait. If we observe some correlation between them, it follows that a particular allele is observed more often in individuals with that trait, compared to the ones without it. If there is sufficient evidence to support this, it can be concluded that such a result could not have happened just by chance. Consequently, any individual carrying copies of this variant would be at an increased risk of having the associated trait.

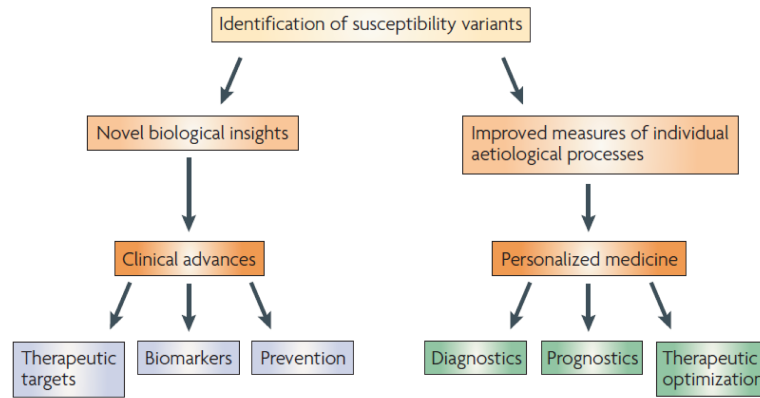


Figure 1.2: Clinical translations of findings from GWA studies [McCarthy et al., 2008]

What happens after genetic association studies and how can the acquired information be used? The objective is to translate our findings into improvements in medicine. Namely, having obtained results from genetic association studies there are two main directions in which this can be done [McCarthy et al., 2008] (Figure 1.2). In the first one (on the left), through novel biological insights into causal pathways we might be able to develop new treatments of patients and possibly even disease prevention. In the other one (on the right), through “genetic profiling”, we might be able to develop more personalized methods for disease

management.

1.3.2 History and Development of GWAS

Genome-Wide Association Studies (GWAS) are Genetic Association Studies examining the whole or most of the genome. In terms of number of polymorphisms analysed, 100 000 genotyped SNPs is accepted as a minimum in the NHGRI catalog [Hindorff et al., 2013], which contains all SNP-trait associations from published GWAS.

The objective of GWAS is to describe the susceptibility mechanism of diseases and traits of interest. Despite still being relatively new, they have already greatly improved our understanding of many traits and diseases, e.g. The Wellcome Trust Case-Control Consortium [WTCCC, 2007] conducted analysis on seven complex diseases and found many novel associations.

In Figure 1.3 a timeline of some of the major advances in GWAS is shown in which three of the most notable projects (the Human Genome Project, the International HapMap Project and the 1000 Genomes Project) have well-deserved places. The initiation of the Human Genome Project (HGP) in 1990 opens the doors to a new era. Determining the entire human DNA sequence and identifying the majority of the approximately 20000 – 25000 genes in the human DNA [International Human Genome Sequencing Consortium, 2004] lay down the foundations for GWAS.

Consequently, another contribution of great importance to the development in this field is made by the International HapMap Project

1. Introduction and Background

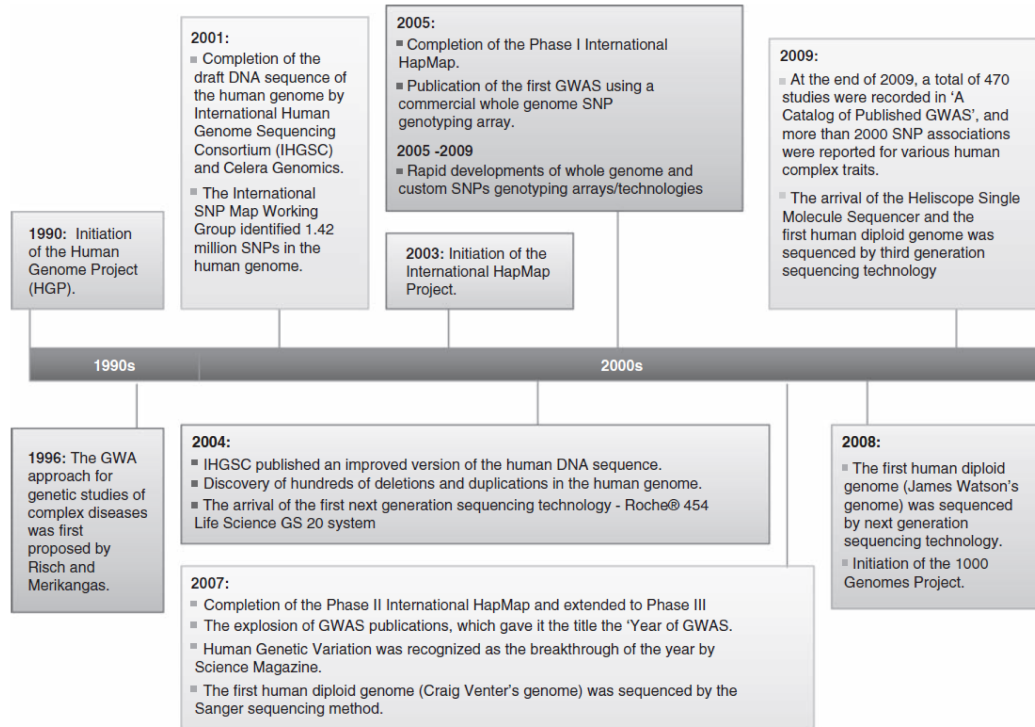


Figure 1.3: Major developments in GWAS and human genome sequencing [Ku et al., 2010]

[The International HapMap Consortium, 2007], the main aim of which was to determine the common patterns of DNA sequence variation and subsequently characterize the LD patterns between SNPs in a number of populations. LD or linkage disequilibrium is the non-random association among polymorphic sites. With its existence, LD significantly reduces the number of SNPs needed to be genotyped, which together with developments in genotyping techniques makes GWAS a feasible approach.

The 1000 Genomes Project [The 1000 Genomes Consortium, 2010] is the first project to sequence the genomes of a large number of individuals from different geographical backgrounds, using improving sequencing technologies. It aims to

find most genetic variants with minor allele frequencies greater than 1% in the populations studied and provide a detailed resource on human genetic variation. With its scale it brings us much closer to having a complete description of human DNA polymorphisms.

Publications of completed studies start appearing in 2002, however the first recorded GWAS [Klein et al., 2005] in the NHGRI catalog [Hindorff et al., 2013] is from 2005, due to the smaller number of markers of the earlier studies. In the following two years many more publications with significant findings start appearing and due to the “explosion of publications” in 2007 it is named the “Year of GWAS” [Ku et al., 2010] (Figure 1.3). One of the most notable ones being the WTCCC study [WTCCC, 2007], which examined 7 major diseases - bipolar disorder, coronary artery disease, Crohn’s disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes, and found 24 independent association signals. In recent years genome-wide association studies have become the study design of choice when searching for the genetic basis of diseases and other traits. The majority of studies have used genotyping chips to measure common sequence variation and have correlated this genetic variation with simple binary or quantitative phenotypes [Manolio, 2010].

1.3.3 Findings

The number of novel SNPs discovered through GWAS has been rapidly increasing over the last few years (Figure 1.4) and there are more than 1700 GWAS publications in [Hindorff et al., 2013] and over 11 000 SNPs identified so far. With this, GWAS has proven to perform well in identifying novel genetic loci and with

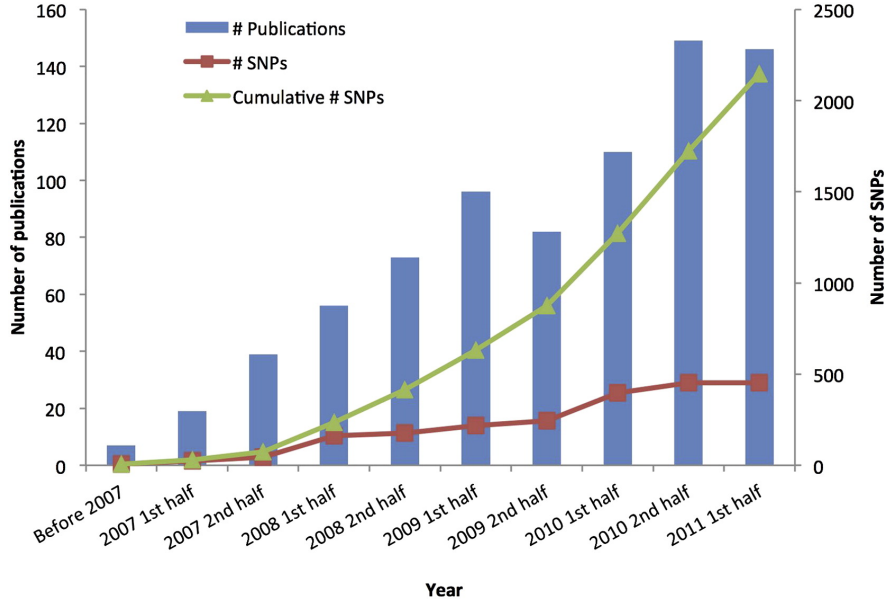


Figure 1.4: GWAS Discoveries over time. Data obtained from the Published GWAS Catalog. Only the top SNPs representing loci with association p values $< 5 \times 10^{-8}$ are included, and so that multiple counting is avoided, SNPs identified for the same traits with LD $r^2 > 0.8$ estimated from the entire HapMap samples are excluded [Visscher et al., 2012].

the cost reduction and the design of larger studies we can expect this trend to continue. An observation from the current results, however, is that most of the identified genetic variants have small effects and the risk alleles are mostly common, with allele frequency over 5% (bottom-right on Figure 1.5). This, of course, does not necessarily mean that rare alleles are not causal, since this can be a result from the fact that they are much harder to detect.

It is believed now that signals that are yet to be identified probably have smaller effect sizes and that GWAS still do not have enough power to detect them [Ku et al., 2010]. Thus, on a quest to find the missing heritability, various directions have been taken. For example, one way to increase the power for detecting causal variants is to increase the study sample sizes. For this

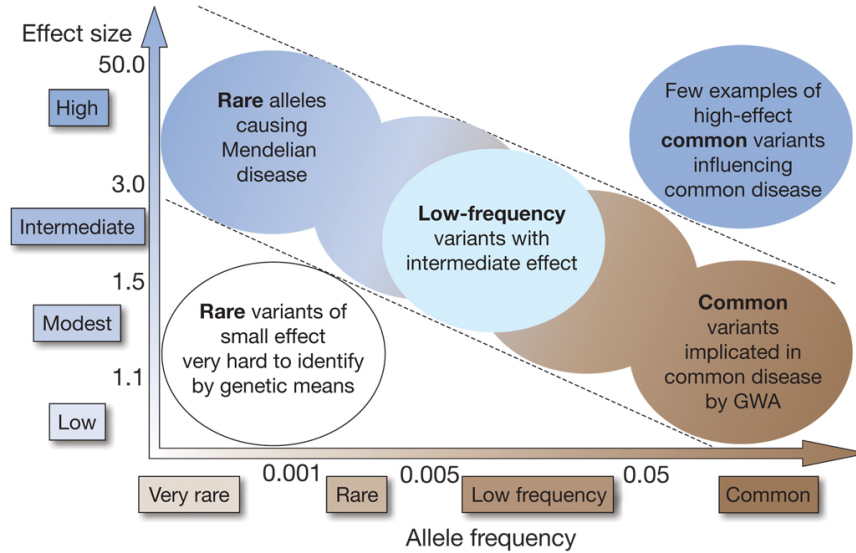


Figure 1.5: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio) [Manolio et al., 2009]

reason large GWAS meta-analyses have started being carried out, for example [Gieger et al., 2011] which aims to discover novel genetic determinants of megakaryopoiesis and platelet formation in over 66 000 samples, and this approach has proven to be fruitful for detection of new genetic variants. Therefore with the collection of larger samples or by combining datasets and analysing them together, we expect that it will be possible for many new hits to be identified.

Other answers to the missing heritability problem might be to target rare SNPs, which is still very expensive and requires state of the art sequencing technologies, however a number of projects have already undertaken to explore this task - for example the UK10K¹ projects, which aims to understand the rare variants predisposition to various diseases and complex traits by sequencing 10 000 individuals in the U.K.

¹<http://www.uk10k.org/>

On the other hand many cohorts record more than just a single trait, e.g. KORA [Wichmann et al., 2005], resulting in a number of, possibly correlated, phenotypes being measured. With the ability to obtain whole genome genotypes and its continuing cost reduction, in terms of data, cohorts would soon only be able to expand in sample sizes and number of different traits recorded. Most GWAS so far however, have been done using only one, typically binary, disease phenotype. Another possible way to increase power might then be to incorporate the correlation structure between traits by analysing them together in a single model rather than considering them one at a time.

1.4 The Concept of Pleiotropy

How can joint analysis help improve power in GWAS? For example if a gene is affecting more than one phenotype, with very small effects, then single trait analysis might not be able to pick up signal for any of them. However if these traits are highly correlated, some information about one should give us information about the others and in that case it might be possible to detect the joint signal.

Most diseases come with more than one symptom or risk factor and so for each disease a number of different characteristics/traits can be measured. For example among the risk factors for coronary heart disease (CHD) are blood pressure and different types of cholesterol levels (total, LDL and HDL cholesterol) [Wilson et al., 1998]. Higher blood pressure levels are typically associated with abnormal cholesterol levels (high total and LDL and low HDL cholesterol levels).

1. Introduction and Background

Therefore if a gene is influencing the cholesterol it is possible that it will also be affecting blood pressure, and through them putting you at high risk of coronary heart disease.

In a very broad sense a phenomenon in which one gene is affecting multiple phenotypes is usually referred to in the literature as pleiotropy (Figure 1.6). In the above example if a gene directly affects cholesterol levels but only indirectly blood pressure (through cholesterol levels), then it would not be considered pleiotropic. However if it directly affects both CHD risk factors, then pleiotropy would be observed.

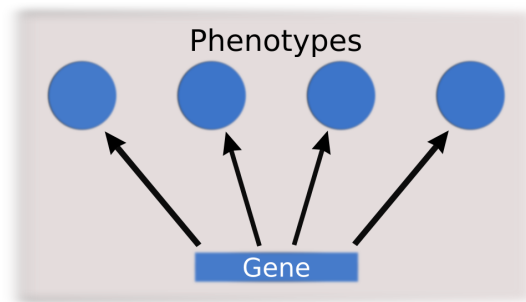


Figure 1.6: Pleiotropy

Unfortunately there are many different definitions and opinions on what pleiotropy exactly is, and the problem lies in the ambiguity of each of them (discussed in [Wagner and Zhang, 2011]). The first problem with the above definition is: how should the phenotypes be defined? If they are measurements of the same characteristic, should we really consider a gene associated to them to be pleiotropic? This would lead to a definition requiring unrelated phenotypes. However this is also not very practical as it is far too subjective. Another issue that arises together with this is measuring the degree of pleiotropy. This is usually taken to

1. Introduction and Background

be the number of associated traits, which also considering the above issues is not optimal, and in addition to that depends on the number of traits considered in a particular study. The other differences in the definitions of pleiotropy are concerning whether it is a single mutation affecting two traits or two closely linked mutations affecting one trait each. Furthermore, a case where an observed variant appears to affect two different phenotypes, could be a result of LD between two separate causal variants and the observed one, each of which affecting only one of the traits (Figure 1.7). This can typically occur if the actual causal variants are not typed and we only have information on a common tag. We should keep this in mind when interpreting results.

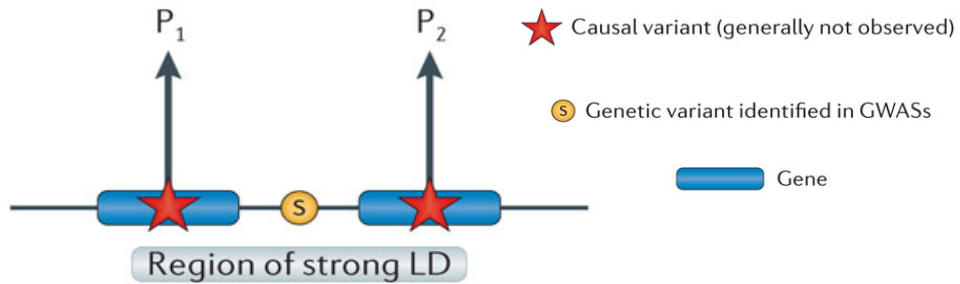


Figure 1.7: Spurious pleiotropy. Figure 1f from [Solovieff et al., 2013]

Sticking to our loose definition, there are a number of methods that test for pleiotropy, for example [Knott and Haley, 2000]. Another approach to measure the degree of pleiotropy in correlated traits would be to consider the “effective number” of associated traits instead of the total number, which can be done as in [Wagner, 2008]. There is already evidence for many pleiotropic loci and Solovieff et al. [Solovieff et al., 2013] argue that estimates of the levels of pleiotropy are likely to be conservative due to the current power limitations of GWAS. On

the other hand, current levels of pleiotropy could mean that pleiotropy is more likely to be “modular” than “universal”, i.e. that a gene can have a pleiotropic effect on phenotypes within a module rather than potentially affecting all traits [Stearns, 2010].

Another interest in this field lies in distinguishing between different types of pleiotropy at the molecular level - i.e. “genuine” and “spurious”, or types 1 and 2, respectively [Stearns, 2010, Wagner and Zhang, 2011]. Genuine referring to two distinct primary products arising from a single locus, and spurious - single molecular function with multiple consequences.

An example of pleiotropic findings is the 9p21 locus, which has shown to be associated with coronary artery disease, myocardial infarction and type 2 diabetes (T2D) [McPherson et al., 2007, Zeggini et al., 2007, Helgadottir et al., 2008]. The two highest scoring SNPs for CAD and T2D in this region were found to be very close to each other but independent and neither one showed association to the other disease [Helgadottir et al., 2007]. This means that these variants probably have independent functions, but further investigations are needed to understand/uncover the causal pathways underlying the diseases [Ku et al., 2010].

1.5 Current methods for independent samples

In this thesis we are only going to consider continuous phenotypes and additive genetic models, however generalisations of some of the methods discussed here as well as the ones we are about to propose might be possible.

1.5.1 Single phenotype model

Given a quantitative phenotype $\mathbf{Y}$ and a genotype $\mathbf{G}$ at a single locus for each of N individuals, the aim of GWAS is to find if there is any association between them. Each element of $\mathbf{G}$, G_i is typically coded 0,1 or 2 to represent the number of copies of the minor allele for individual i at that specific locus. When data is imputed this could either be inferred based on the best genotype or alternatively the expected number of minor alleles can be used. In term of statistical models, the most widely employed is the normal linear model with additive genetic effects

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \text{ , where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N) \quad (1.1)$$

where $\mathbf{X}$ is a matrix of covariates, $\boldsymbol{\beta}$ their effect sizes, $\boldsymbol{\alpha}$ the genotype effect size and σ^2 the error variance.

Typically, when a number P of phenotypes is available in a study, P individual genome-wide scans are performed for each of them separately [Soranzo et al., 2009, Illig et al., 2009, Demirkan et al., 2011, Demirkan et al., 2012].

1.5.2 Multi-trait methods for unrelated samples

Despite not much focus being placed on joint multiple phenotype analysis in comparison to univariate analysis, there are a number of approaches developed to analyse multiple (possibly correlated) traits, which we briefly review here. The focus in this section is on methods for unrelated samples and approaches that model relatedness between individuals will be introduces afterwards.

1. Introduction and Background

Possibly the simplest approach for modelling multiple phenotypes is to test each trait for association one at time and then use the smallest p-value across all traits as evidence of association. Some current methods focus on ways of combining univariate association tests [Yang et al., 2010], which are easy to implement and do not impose many assumptions on the phenotypes. Results from these approaches show an increase in power for detecting causal variants in some cases, however they still ignore a lot of the available information by not modelling the phenotype correlations explicitly.

Others use dimension reduction techniques, e.g. principal component analysis (PCA), and in [Klei et al., 2008] an improvement (in terms of power) to the PCA method is implemented, by using a principal component of heritability. These methods derive and analyse a single trait, which is a linear combination of all phenotypes. Results from these types of approach however might not be biologically interpretable.

[Ferreira and Purcell, 2009] propose the use of Canonical Correlation Analysis (CCA) for modelling multiple traits, which extracts the linear combination of traits that explain the largest possible amount of the covariation between the marker and all traits. CCA for a single SNP test reduces to multivariate linear regression with the SNP as the response variable and all traits as covariates. We will refer to this as the Reverse model in later Chapters. Although this approach is most appropriate for the analysis of normally distributed traits, it shows good performance even when considering non-normal or disease traits. O'Reilly et al. (2012) [O'Reilly et al., 2012] use a very similar approach in which ordinal

1. Introduction and Background

regression is used instead of linear regression and Wang (2007) [Wang, 2007] implement a Bayesian approach.

Further a class of Structural equation models (SEMs) is proposed in [Li et al., 2006]. SEM is a hierarchical system of regression relationships in which each variable may be a response or a predictor. Each model structure can be represented as a directed graph in which causal relationships between variables are denoted by directed edges between the variable nodes. SEMs, in contrast to the above methods aim to provide a precise description of the genetic architecture of a complex system. However quantitative trait loci (QTLs) in this approach are only included in the analysis if they are identified by univariate analysis, thus not aiming the method at improving power for discovery of novel loci.

An extension of the univariate least squares approach is also proposed in [Knott and Haley, 2000]. The multitrait least squares method is aimed at QTL detection, it is easy to implement and also allows for formal testing of whether there are two linked or one pleiotropic QTL affecting two traits. To test each hypothesis, approximate likelihood ratios are used. With increasing complexity, due to the number of parameters to be estimated, the increase in the degrees of freedom of the test statistic can restrain its practical use when the number of traits is large [Mangin et al., 1998].

In [Agakov et al., 2011] a multivariate Bayesian approach is described, which performs joint analysis of multiple genotypes, phenotypes and biomarkers. The sparse instrumental variable method models linear associations between the variables with the use of sparseness-inducing priors. It may be used for effective

screening of genotype-phenotype (and biomarker-phenotype) associations, however a limitation of it is the restriction on all model errors being uncorrelated (i.e. not allowing for non-diagonal noise covariance matrices).

Verzilli et al. (2005) [Verzilli et al., 2005] suggests an approach similar to our proposed model averaging (A) approach (Chapter 2), which in addition can model multiple SNPs at the same time, possibly all within the same candidate gene. The model allows for different sets of SNPs to be associated with each of the traits and inference using this model is carried out using MCMC. It is aimed at revealing the genetic variants (within a region) related to a set of traits, rather than characterizing the signal for a single genetic variant and it does not consider models with some zero and some non-zero effect sizes (it only fits what we refer to as the Full model).

Other authors [Rothman et al., 2010, Yin and Li, 2011] have studied the multivariate regression model (Eq. 1.2) and use a sparse prior on both the regression coefficients α and the precision matrix Ω using l_1 -norm penalties, which implies a graphical gaussian model (GGM) on the residual terms. The main focus of these methods was prediction and estimation of the GGM implied by Ω , whereas our focus is on hypothesis testing and subsequent model selection.

$$\mathbf{Y} = \mathbf{G}\alpha + \epsilon \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N, \Omega^{-1}) \quad (1.2)$$

The methods described in this section can perform simultaneous analysis on a number of phenotypes and some of them can also include a number of genetic

1. Introduction and Background

Study	Correlation	Sparsity	Extra
[Yang et al., 2010]	No	No	-
[Ferreira and Purcell, 2009]	Implicitly	No	CCA
[O'Reilly et al., 2012]	Implicitly	No	Ordinal G
[Wang, 2007]	Implicitly	No	Bayesian
[Klei et al., 2008]	Implicitly	No	Not biologically interpretable
[Li et al., 2006]	Implicitly	Yes	Cannot improve power
[Knott and Haley, 2000]	Yes	No	Not practical for large P
[Agakov et al., 2011]	No	Yes	-
[Verzilli et al., 2005]	Yes	No	MCMC
[Rothman et al., 2010]	Yes	Yes	aim: prediction and
[Yin and Li, 2011]	Yes	Yes	estimation of the GGM
A Method	Yes	Yes	Chapter 2
VB SAS Method	Yes	Yes	Chapter 5

Table 1.1: Methods for multiple phenotype analysis for unrelated samples. Correlation refers to whether the proposed method models between-phenotype correlation. Sparsity refers to whether the method considers cases in which only a subset of the phenotypes are associated to a particular genetic variant. The Extra column gives additional information about the methods.

variants. They all show nice properties in terms of outperforming univariate analysis (at least in some cases), some by an increased power for detecting causal variants and others for characterizing the genetic architecture of a causal system, however they all have their limitations. A brief summary with their key properties is given in Table 1.1 together with the A and VB SAS approaches we propose in Chapters 2 and 5.

The problem we consider is the assessment of association between multiple related continuous response variables and a single explanatory genotype variable. We do this in a Bayesian framework where we directly model the between-phenotype variances and using approaches that allow for both hypothesis testing and model selection. Our main focus is on improving power to detect true

associations between a given locus and at least one of the (possibly correlated) phenotypes and due to evidence of more “modular” (than “universal”) pleiotropy [Stearns, 2010] we consider and put more weight on sparse models of association. Our A approach is computationally tractable genome-wide for modest numbers of phenotypes (up to 10-15), whereas the VB SAS method can be applied to hundreds of traits. In addition, the VB SAS framework allows for and can perform imputation of sporadically missing phenotype data.

1.6 Missing Phenotype Data

Missing phenotype data is a real feature of many cohorts that have collected multiple phenotypes. When considering a single trait, analysis can be carried out using all individuals with non-missing data for that phenotype. However, when multiple traits are to be considered jointly (multivariately) then by using most basic models, only individuals with non-missing data for *any* of their phenotypes can be included in the analysis, which can reduce the sample size enormously. For example, if we have data on 10 phenotypes in 1000 samples and if we assume we have 1% missing data for each trait independently, then the resulting sample size for a multivariate study that cannot appropriately take into account phenotype missingness would be around 900 samples. Increasing the number of phenotypes to 100 with the same rate of missingness would certainly not leave enough samples to perform any analysis. So, this is a note to stress the importance of appropriate handling of missing data when considering multiple phenotypes in a single framework.

1.7 Relatedness between samples and population structure

Yet another real feature of many studies is population structure, which refers to the presence of systematic differences in allele frequencies between different subsamples of the population that can arise due to different ancestry. This mainly occurs due to study design, for example in isolated population cohorts (e.g. VIS, KORCULA, ORCADES), but can also be observed in supposedly unrelated samples, in which case it is referred to as "cryptic relatedness". Correcting for both of these types of population structure is very important in GWAS [Zhang and Deng, 2010, Voight and Pritchard, 2005], since it can lead to spurious associations [Campbell et al., 2005]. Many methods have been developed for correcting for stratification such as genomic control (GC) [Bacanu et al., 2000, Devlin and Roeder, 1999], principal component analysis (PCA) [Price et al., 2006] and mixed models [Kang et al., 2008], each of which has their advantages. With recent computational advances, mixed models have become possible to use in GWAS setting, although still being computationally very expensive, and are increasingly popular due their ability to address and correct confounding due to population stratification, family structure and cryptic relatedness [Price et al., 2010].

Linear mixed models extend basic linear models by the addition of a random effect, whose variance depends on a kinship matrix $\mathbf{K}$ that represents the similarity between individuals and through it the population structure. Kinship coefficients can either be inferred based on known pedigree information (which in GWAS is typically not known) or based on marker data. There are differences between

them and it is usually preferable to use the latter due to better accuracy, even if pedigree information is available [Astle and Balding, 2009].

1.8 Methods modelling sample relatedness

1.8.1 Univariate mixed effects models

We begin by introducing linear mixed models, widely used in single-phenotype analysis, such as [Yu et al., 2006, Aulchenko et al., 2007, Zhao et al., 2007, Kang et al., 2010], and then describe multivariate extensions that can handle multiple correlated traits.

Given a single phenotype $\mathbf{Y}$ and a single genetic variant $\mathbf{G}$, mixed effects models have an additional (random) term $\mathbf{u}$ to the basic association model in Equation 1.1, whose variance depends on the kinship matrix $\mathbf{K}$. The statistical model is the following:

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{u} + \boldsymbol{\epsilon} \quad (1.3)$$

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{2K}) \quad (1.4)$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N) \quad (1.5)$$

which written in matrix form is equivalent to

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(0, \sigma_g^2 \mathbf{2K} + \sigma_e^2 \mathbf{I}_N, \mathbf{I}_1) \quad (1.6)$$

and when no sample relatedness is observed, $2\mathbf{K} = \mathbf{I}_N$, it reduces to the basic association model in Equation 1.1. Here $\mathcal{N}(\cdot, \cdot, \cdot)$ refers to the matrix normal distribution, overview of which can be found in Appendix A.

Fitting such models is not computationally trivial and only in recent years due to computational advances has it become possible to employ them in GWAS. Similar models, on the other hand, have played a central role in linkage analysis and animal and plant breeding, where relationships between individuals are known explicitly [de los Campos et al., 2012].

Furthermore, this model is related to a Bayesian linear model in which a set of p SNPs, stored in the mean-centered and standardised $N \times p$ matrix $\mathbf{X}$, are included in the model

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1.7)$$

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_N) \quad (1.8)$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_N) \quad (1.9)$$

Equivalence between (1.3) and (1.7) occurs if

(a) we estimate $2\mathbf{K} = \mathbf{X}\mathbf{X}^T/p$, which is the usual covariance estimate between individuals based on the data at the p SNPs in $\mathbf{X}$.

(b) we let $\sigma_g^2 = p\sigma_\beta^2$

In this case $\mathbf{X}\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{X}\mathbf{X}^T\sigma_\beta^2) = N(\mathbf{0}, \sigma_g^2 2\mathbf{K})$ and $\mathbf{u} = \mathbf{X}\boldsymbol{\beta}$.

Recent work has involved fast implementations of fitting this model [Lippert et al., 2011, Zhou and Stephens, 2012], although this approach has been used within the package SOLAR¹ for many years. If the eigenvalue decomposition (SVD) of the kinship matrix is given by $\hat{\mathbf{K}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ then pre-multiplying Eqn. 1.6 by $\mathbf{U}^T$ results in the model

$$\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{2}\mathbf{\Lambda} + \sigma_e^2\mathbf{I}_N, \mathbf{I}_1) \quad (1.10)$$

which has diagonal covariance that simplifies and speeds up estimation of σ_g^2 and σ_e^2 .

1.8.2 Multi-trait methods for related samples

A number of multivariate phenotype extensions to the basic mixed effects models are becoming available. For example, [Korte et al., 2012] consider a mixed-model for $P = 2$ traits so that $\mathbf{Y}$ is a $N \times 2$ matrix. The model they use has the form

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{v} + \boldsymbol{\epsilon} \quad (1.11)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N, \mathbf{D}) \quad (1.12)$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, 2\mathbf{K}, \mathbf{B}) \quad (1.13)$$

where $\mathbf{B}$ and $\mathbf{D}$ are 2×2 matrices of covariances, where $\mathbf{B}$ models the correlation between the 2 phenotypes due to shared relationships between individuals and $\mathbf{D}$ models correlation due to non-genetic factors. Estimation is carried out using

¹<http://solar.txbiomedgenetics.org/>

maximum likelihood using the ASReml package [Gilmour et al., 2009].

[Rakitsch et al., 2012] consider another variant of a multivariate mixed model applied to measurements of alternatively spliced transcript isoforms estimated from RNA-seq. The model they use has the form

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{v} + \mathbf{u} + \boldsymbol{\epsilon} \quad (1.14)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N, \mathbf{D}) \quad (\mathbf{D} \text{ diagonal}) \quad (1.15)$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, 2\mathbf{K}, \mathbf{B}) \quad (1.16)$$

$$\mathbf{u}_n \sim N(0, \alpha^2, C_n), \quad n \in \{1, \dots, N\} \quad (1.17)$$

This model is similar to the model of [Korte et al., 2012] but has an additional term to model subject specific correlation between isoforms (traits) via the correlation matrices C_n , $n \in \{1, \dots, N\}$.

Additionally, [Zhou and Stephens, 2014] present a computationally efficient method for the model in [Korte et al., 2012] for a modest number of phenotypes (2-10) which is implemented in the software package GEMMA¹ [Zhou and Stephens, 2012]

The methods described in this section can perform simultaneous analysis on a number of (potentially correlated) phenotypes in samples of (potentially) related individuals. A brief summary with their key properties is given in Table 1.2 together with the VB SAS approach we propose in Chapter 5. The problem we consider is the assessment of association between multiple related continuous response variables and a single explanatory genotype variable. We do this in a

¹<http://stephenslab.uchicago.edu/software.html>

Study	Missingness	Sparsity	Extra
[Rakitsch et al., 2012]	No	No	-
[Korte et al., 2012]	No	No	Only 2 traits
[Zhou and Stephens, 2014]	Yes*	No	~ 2-10 traits
VB SAS Method	Yes	Yes	Many traits; Chapter 5

Table 1.2: Methods for multiple phenotype analysis for related samples. Missingness refers to whether the proposed method can accommodate missing phenotype data. Sparsity refers to whether the method considers cases in which only a subset of the phenotypes are associated to a particular genetic variant. The Extra column gives additional information about the methods. * means that a separate phenotype imputation algorithm is used.

Bayesian framework where we directly model the between-phenotype variances and using approaches that allow for both hypothesis testing and model selection. Our main focus is on improving power to detect true associations between a given locus and at least one of the (possibly correlated) phenotypes and due to evidence of more “modular” (than “universal”) pleiotropy [Stearns, 2010] we consider and put more weight on sparse models of association. Our VB SAS method uses Variational Bayes methods and “Spike and slab” shrinkage priors on the genetic effect sizes, to calculate approximate marginal likelihoods of competing models and can be applied genome-wide to tens-to-hundreds of traits genome-wide (depending on the amount of population structure). In addition, the VB SAS framework allows for and can perform imputation of sporadically missing phenotype data.

1.9 Summary

When modelling multiple traits in a joint model a straight-forward approach would be to fit a “full model” in which all of the traits are allowed to have some non-zero effect size. For continuous traits this can be achieved using multivari-

1. Introduction and Background

ate linear regression using either frequentist or Bayesian methods, that allow for correlated residual terms. Such an approach would sacrifice power when not all traits are associated, which may be the case, especially if the number of traits being analysed is large. For this reason we have developed several methods that allow for just a subset of the traits to be associated. Our first method is a model averaging approach that combines evidence across all possible models in which just a subset of traits are associated. An advantage of this approach is that it is very flexible in terms of the a priori information about which subsets of traits are more likely to be jointly associated. This approach becomes less computationally feasible as the number of traits increases beyond around 10 traits. Therefore we have also developed methods that use sparse continuous priors that shrink effect size parameters towards zero when possible. This allows much larger numbers of traits to be analysed jointly. We use Variational Bayesian methods to approximate the marginal likelihoods needed for inference [Jaakkola, 1997, Bishop and Tipping, 2000, Jaakkola and Jordan, 2000]. One advantage of these Bayesian approaches is that hypothesis testing and model selection are carried out at the same time. Our methods provide a single Bayes factor for each SNP that measures the evidence of association with the phenotypes which can be used to detect associated loci. Then once detected, the model fit can be examined to characterize the form of the association in terms of which subset of traits has the most evidence of joint association.

In addition our models allow for correlation between individuals, which can exist due to either close relationships between individuals (often by design in studies of families) or more distant relationships between individuals which may not be

explicitly known. We use multivariate random effects that allow us to capture this correlation and fixed effects to model the trait-SNP associations. Overall, our approach can be described as a multivariate mixed model that encourages sparsity of the fixed effects. A further advantage of this methods is that it can handle sporadic missing phenotype data which is a prevalent feature of real studies. In doing so we avoid the need to discard samples with at least one missing phenotype as this can substantially reduce power in the presence of many traits.

The main focus of the following chapters in this thesis is on the development (Chapters 2 and 5), assessment (Chapters 3 and 6) and application (Chapters 4 and 7) of such methods for joint analysis of multiple quantitative traits. We then conclude with a discussion and potential future in Chapter 8.

Chapter 2

A Bayesian model averaging approach

In this Chapter we outline and discuss a Bayesian model averaging (A) approach for simultaneous analysis of multivariate phenotypes in genome-wide association studies. It relies on the idea of bi-partitioning the traits into (disjoint) sets of associated and un-associated phenotypes (to each single genetic variant in question) and possesses the ability of assigning weights to each of these models that can reflect prior beliefs on the specific data we consider. The method is exact and computationally efficient for a modest number of traits ($\sim 2 - 10$) and it is designed for GWAS of unrelated individuals. In terms of power, it can outperform simpler approaches as shown in Chapter 3 for both pleiotropic and non-pleiotropic loci.

2.1 Introduction

We suppose we have $P > 1$ continuous traits $\mathbf{Y}_1, \dots, \mathbf{Y}_P$ and a genotype $\mathbf{G}$ at a single SNP for each of N unrelated individuals. $\mathbf{Y}_1, \dots, \mathbf{Y}_P$ and $\mathbf{G}$ are $N \times 1$ (column) vectors, in which each row contains the information for one individual. Each element of $\mathbf{G}$, G_i , is coded as 0, 1 or 2 and represents the number of copies of the minor allele for individual i at that specific locus, and each element Y_{ij} of $\mathbf{Y}_j$ is the j^{th} phenotype of the i^{th} individual.

Having data of this form, there are two natural questions that we would like to be able to answer:

Q1 Is there any association between the phenotypes and the genotype?

Q2 If there is, what kind of association is it, i.e. which phenotypes are associated to the genotype and which are not?

2.2 Model

We employ a (multivariate) linear model of the following form:

$$\mathbf{Y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma}) \quad (2.1)$$

where $\mathbf{Y} = (\mathbf{Y}_1 \ \mathbf{Y}_2 \ \dots \ \mathbf{Y}_P)$ is an $N \times P$ matrix of phenotypes, each column of which represents a single trait, $\boldsymbol{\mu} = (\mu_1 \ \mu_2 \ \dots \ \mu_P)$ is a $1 \times P$ (row) vector of means, $\mathbf{1}$ is an $N \times 1$ matrix each element of which is one, $\mathbf{G}$ is an $N \times 1$ (column)

2. A Bayesian model averaging approach

vector of genotypes at a single SNP, $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_P)$ is a $1 \times P$ (row) vector of genotype effect sizes, one for each phenotype, and $\boldsymbol{\epsilon}$ is an $N \times P$ matrix of normally distributed errors. We assume that we have independent samples, but we allow for correlation between phenotypes that is not explained by $\mathbf{G}$. Here $\mathcal{N}(\mathbf{M}, \mathbf{V}, \boldsymbol{\Sigma})$ denotes the Matrix Normal Distribution with mean $\mathbf{M}$ and row and column covariance matrices $\mathbf{V}$ and $\boldsymbol{\Sigma}$. Thus by $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma})$ we mean that the errors are normally distributed such that rows are independent, each multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. More on the Matrix Normal Distribution can be found in Appendix A (page 162) and [Dawid, 1981].

The intercept of Model 2.1 is not of interest for answering questions **Q1** and **Q2**, which are concerned with the possible genetic effects on each of the traits. Without loss of generality then we can remove the mean term from our model by mean-centering $\mathbf{G}$ and all of the phenotypes $\mathbf{Y}_j$ (i.e. $\mathbf{G} := \mathbf{G} - \bar{\mathbf{G}}$ and $\mathbf{Y}_j := \mathbf{Y}_j - \bar{\mathbf{Y}}_j$ for all j). Additionally, for comparability between the traits, we also scale each of them to have a unit variance.

With the data now transformed and the intercept removed, we employ the following normal linear model:

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma}) \tag{2.2}$$

where $\mathbf{Y}$ is a matrix of mean-centered and scaled phenotypes, $\mathbf{G}$ a vector of mean-centered genotypes, and everything else is as in Model 2.1.

2. A Bayesian model averaging approach

The advantage of using Model 2.2 to modelling each phenotype separately is that we can use information across different traits by allowing for a non-diagonal error covariance matrix. Correlated errors can be interpreted as factor or factors (not related to the genetic variant) that influence some or all of the traits.

Model 2.2 can also easily be extended to include other covariates, e.g. age, sex of individuals in the study, etc. This can be done by regressing out all of these non-genetic factors from each of our traits and then using the residuals as phenotypes in Model 2.1, or after scaling and mean centering in Model 2.2.

2.2.1 Model space

In total 2^P possible models of the form of Model 2.2 exist based on P phenotypes, where each is either associated to the genetic variant $\mathbf{G}$ or not (i.e. each $\alpha_i = 0$ or $\alpha_i \in \mathbb{R}$).

We distinguish between three classes of models - “Null” in which none of the traits depends on the genotype, “Full” in which all traits are considered associated, “Other” in which there is at least one associated and one unassociated trait.

Full Model

$$\begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \dots & \mathbf{Y}_P \end{pmatrix} = \mathbf{G} \begin{pmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_P \end{pmatrix} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma}) \quad (2.3)$$

$\alpha_i \in \mathbb{R}$ for all i

Null Model

$$\begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \dots & \mathbf{Y}_P \end{pmatrix} = \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma}) \quad (2.4)$$

Other - Non-null, Non-full Models

$$\begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \dots & \mathbf{Y}_P \end{pmatrix} = \mathbf{G} \begin{pmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_P \end{pmatrix} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma}) \quad (2.5)$$

$\alpha_i \in \mathbb{R}$ or $\alpha_i = 0$ for each i , (at least one of each)

We denote all possible models by M_i , $i = 0, 1, \dots, 2^P - 1$, where throughout we assume that M_0 is the null model, M_j for $j = 1, 2, \dots, P$ are the models with trait j considered associated and all other traits unassociated to the particular variant. All other models $P + 1, \dots, 2^P - 1$ being pleiotropic (with at least two phenotypes being associated). An example on $P = 2$ phenotypes is given in Figure 2.1.

Under our model specification, answering question **Q1** is equivalent to distinguishing between $\boldsymbol{\alpha} = 0$ and $\boldsymbol{\alpha} \neq 0$. In the two phenotype example in Figure 2.1 this translates to a comparison between the null model and the group of association models (the 3 models in gray in Figure 2.1). Similarly, assuming we find association, question **Q2** is equivalent to uncovering exactly which $\alpha_i = 0$, which translates to distinguishing between the 3 association models (coloured in gray).

2.3 Bayesian model averaging

In order to address questions **Q1** and **Q2** for the models described in Section 2.2 and 2.2.1, we use a Bayesian approach. Given a prior distribution $\mathbb{P}(\boldsymbol{\alpha}, \boldsymbol{\Sigma} | \text{Model})$

2. A Bayesian model averaging approach

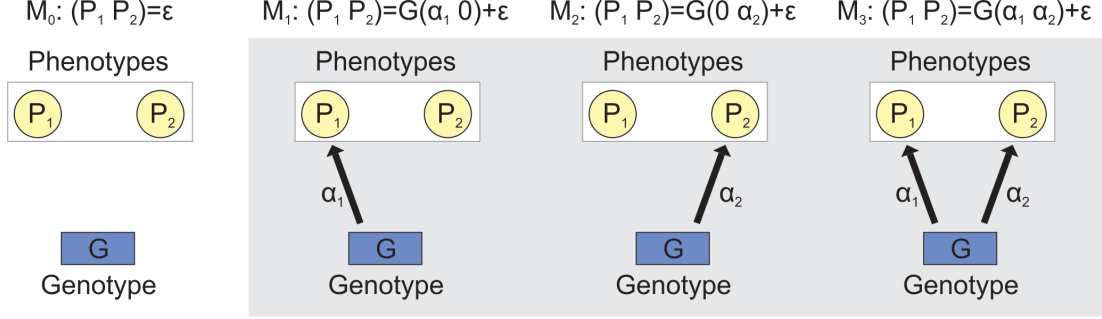


Figure 2.1: Graphical representation of all models on 2 phenotypes.

for the parameters α and Σ for each model M_i , we can find the marginal likelihood $\mathbb{P}(\text{Data}|\text{Model}) = \mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{Model})$. We do this by conditioning on the prior parameter distributions and integrating out the parameters to get:

$$\mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{Model}) = \int \int \mathbb{P}(\mathbf{Y}|\mathbf{G}, \alpha, \Sigma, \text{Model}) \mathbb{P}(\alpha, \Sigma|\text{Model}) d\alpha d\Sigma \quad (2.6)$$

Then for every two models under consideration, we can calculate a Bayes factor, as the ratio of their marginal likelihoods. For example a comparison between the Full (Model 2.3) and the Null (Model 2.4) models on p phenotypes, results in $BF_{\text{Full}} = \mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{Full}) / \mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{Null})$.

Additionally for each model, its posterior probability can be found given a pre-specified model prior as

$$\mathbb{P}(\text{Model}|\mathbf{Y}, \mathbf{G}) \propto \mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{Model}) \mathbb{P}(\text{Model}|\mathbf{G})$$

Returning to the first problem we are trying to solve, i.e. identifying associated and potentially causal genetic variants, we need a test statistic which is to com-

2. A Bayesian model averaging approach

pare the set of models of association to the null model. In the Bayesian framework we can define a Bayes factor of association as the ratio of $\mathbb{P}(\text{Data}|\text{Model of association})$ and the marginal likelihood under the null model. This would then give us the usual posterior-prior odds between the models of interest.

Now $\mathbb{P}(\text{Data}|\text{Model of association}) = \sum_{i=1}^{2^P-1} \mathbb{P}(\text{Data}|M_i)\mathbb{P}(M_i|\text{Model of association})$, which yields a definition of the Bayes factor of association

$$BF_A = \sum_{i=1}^{2^P-1} \frac{\mathbb{P}(\text{Data}|M_i)}{\mathbb{P}(\text{Data}|M_0)} \mathbb{P}(M_i|\text{Model of association}) \quad (2.7)$$

Denote by BF_i the Bayes factor comparing model M_i to the null and by p_i the prior on model M_i to be a model of association. Equation 2.7 then becomes

$$BF_A = \sum_{i=1}^{2^P-1} BF_i p_i \quad (2.8)$$

The resulting single Bayes factor statistic has an interpretation of a model averaging approach in which the p_i 's are the weights of the corresponding model Bayes factors, which can be chosen to reflect our prior belief. From now on by a model (or an individual model) Bayes factor we refer to a Bayes factor comparing model M_i (for some $i \neq 0$) and M_0 .

In terms of determining if a variant is associated the model averaging Bayes factor BF_A can be used. As for finding the exact type of association, the models can be compared, either using the individual model Bayes factors BF_i or their posterior probabilities.

2.3.1 Priors

In the Bayesian framework we need to specify priors for the parameters for each of the models under consideration. We start by describing the full model priors and then extend them to give comparable priors for the rest of the models.

We use Normal Inverse Wishart priors on $(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$, which is the matrix-variate extension of the Normal Inverse Gamma prior, commonly used in current Bayesian GWAS analyses. This means that we take

- $\boldsymbol{\Sigma} \sim IW(c, \boldsymbol{Q})$ (Notation given in Appendix A)
- $\boldsymbol{\alpha}|\boldsymbol{\Sigma} \sim N(\boldsymbol{m}, v\boldsymbol{\Sigma})$

The Inverse Wishart distribution is considered suitable for modelling covariance matrices (just as inverse-gamma is for variances). The Normal distribution is the most widely used one for effect sizes in univariate analysis and its extension, the multivariate normal, is the natural choice for multivariate analysis. In addition, the Normal Inverse Wishart prior has a conjugate structure, which makes analytical computation of the marginal likelihood integrals possible. Its potential disadvantage however is that, it might not truly reflect our prior belief, in particular with the prior on the effect sizes depending on the model covariance matrix. In our case however, this seems to be a reasonable assumption, based on the idea that if two traits are highly correlated then we might expect either both or neither of them to be associated to a genetic variant. On the other hand even if they are both associated they can still have somewhat different effects and it is not clear how well would this be accommodated. Taking all this into account

2. A Bayesian model averaging approach

and keeping in mind the potential problems, we still choose to use this prior for our analysis.

Here we discuss our choice of hyperparameters for the prior distributions. Genetic variants can have either positive or negative effects on each trait in question, which motivates zero mean of the effect size distribution apriori. In addition, the expectation of very small effect sizes suggests that the marginal expectation of the model variance would be close to the phenotype variance itself (which is 1 after scaling the trait). A suitable prior on the error variance in univariate analysis is $\Gamma(2, 3)$ (default prior in SNPTEST¹). In our multivariate case, we can choose the hyperparameters of the prior distribution on the error covariance matrix, so that they are equivalent to the single-trait model ones, when we have only one trait in our model. In addition, both for computational simplicity and because error correlations need not be the same (or even in the same direction) as phenotype correlations, we choose to use priors that result in uncorrelated error variance expectation. More details can be found in Appendix B, page 167. The remaining hyperparameter v we set to be equal to 0.02 as in [Marchini, 2010].

Specifically, the priors on the error covariance matrix Σ and the genetic effect sizes α that we use are the following:

- $\Sigma \sim IW(5 + P, 4\mathbf{I})$ and
- $\alpha|\Sigma \sim N(\mathbf{0}, 0.02\Sigma)$

and corresponds to setting $\mathbf{m} = \mathbf{0}$, $v = 0.02$, $c = 5 + P$, $\mathbf{Q} = 4\mathbf{I}$.

¹https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

2. A Bayesian model averaging approach

For all other non-null models we use the same prior on the error covariance, however we force the relevant entries of α for each model to be zero. For a given non-null model we have a partition of the traits into an associated and unassociated set. Given this partition we split α into α_+ and α_- , which denote the effect sizes of the associated and unassociated traits respectively. This means that $\alpha_- = 0$.

We use the prior $\alpha|\Sigma \sim N(\mathbf{0}, v\Sigma)$ on the full vector, and force the α_- coefficients to zero. This results in a prior on the associated trait effect sizes of α_+ take a prior $\alpha_+|\Sigma \sim N(\mathbf{0}, v\Sigma^*)$ (obtained by conditioning on $\alpha_- = \mathbf{0}$), where $\Sigma^* = \Sigma_{++} - \Sigma_{+-}\Sigma_{--}^{-1}\Sigma_{-+}$ (definitions of the Σ matrices can be found in Appendix A under the Matrix Normal or Inverse Wishart distributions, where $+$:= 2 and $-$:= 1).

Prior weights on each of the models (M_i) can be chosen to represent any prior belief we might have, or in the case of no such insight available either a uniform prior or a sparsity inducing prior (e.g. using Binomial distribution on the number of associated traits and a uniform prior on all models of different class) can be used. In our simulation study (Chapter 3) we use a uniform prior, whereas in the real data analysis in Chapter 4 we use a very specific biology motivated non-uniform prior.

2.3.2 Marginal Likelihoods, Bayes Factors, Parameter posteriors

In our framework we assume that the data given the parameters is normally distributed in each of our models, i.e. the **likelihood** is

$$\mathbb{P}(\mathbf{Y}|\mathbf{G}, M_i, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{G}\boldsymbol{\alpha}, \mathbf{I}, \boldsymbol{\Sigma}) \quad (2.9)$$

2.3.2.1 Full model

The **priors** in our full model are

$$\boldsymbol{\Sigma} \sim IW(c, \mathbf{Q}) \text{ and } \boldsymbol{\alpha}|\boldsymbol{\Sigma} \sim N(\mathbf{m}, v\boldsymbol{\Sigma}) \quad (2.10)$$

with hyperparameters as in Subsection 2.3.1.

By the conjugacy of the priors, analytical expression of the **marginal likelihood** can be obtained and is of the following form (derivation of which is given in Appendix C):

$$\mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{ Full model}) = \frac{|\mathbf{Q}|^{c/2}}{|\mathbf{Q}_F^*|^{c^*/2}} \frac{v^{*P/2}}{v^{P/2}} \frac{\Gamma_P[c^*/2]}{\Gamma_P[c/2]} \frac{1}{\pi^{(NP/2)}} \quad (2.11)$$

where $\Gamma_P(\cdot)$ is the multivariate gamma function, and

- $v^* = (\mathbf{G}^T \mathbf{G} + v^{-1})^{-1}$
- $\mathbf{m}^* = (\mathbf{G}^T \mathbf{G} + v^{-1})^{-1}(v^{-1}\mathbf{m} + \mathbf{G}^T \mathbf{Y})$
- $c^* = c + N$

2. A Bayesian model averaging approach

- $\mathbf{Q}_F^* = \mathbf{Q} + \mathbf{Y}^T \mathbf{Y} + \mathbf{m}^T \mathbf{v}^{-1} \mathbf{m} - \mathbf{m}^{*T} \mathbf{v}^{*-1} \mathbf{m}^*$

The **posteriors of the parameters** stay in the same families of distributions (again by the conjugacy of the priors) and are given by:

$$\Sigma | Data \sim IW(c^*, \mathbf{Q}_F^*) \text{ and } \boldsymbol{\alpha} | \Sigma, Data \sim N(\mathbf{m}^*, v^* \Sigma) \quad (2.12)$$

2.3.2.2 Null model

Analogously for the null model very similar expressions can be derived. And the **Bayes factor** comparing the full and null models takes the form:

$$BF_{Full} = \frac{\mathbb{P}(\mathbf{Y} | \mathbf{G}, \text{Full model})}{\mathbb{P}(\mathbf{Y} | \mathbf{G}, \text{Null model})} = \frac{|\mathbf{Q}_0^*|^{c^*/2} v^{*P/2}}{|\mathbf{Q}_F^*|^{c^*/2} v^{P/2}} \quad (2.13)$$

where $\mathbf{Q}_0^* = \mathbf{Q} + \mathbf{Y}^T \mathbf{Y}$.

2.3.2.3 Other models

The **priors** in all non-null/ non-full models are

$$\Sigma \sim IW(c, \mathbf{Q}) \text{ and } \boldsymbol{\alpha} | \Sigma \sim N(\mathbf{m}, v \Sigma^*) \quad (2.14)$$

with hyperparameters as in Subsection 2.3.1. For simplicity of the notation used, here $\boldsymbol{\alpha}$ denotes only the non-zero elements of the effects. These priors do not allow us to analytically derive the marginal likelihood integrals, however by splitting the phenotypes in groups according to them being related to the variant or not, we can derive an analytical expression directly for the Bayes factors (derivation

2. A Bayesian model averaging approach

given in Appendix D).

$$BF_i = \frac{\mathbb{P}(\mathbf{Y}|\mathbf{G}, M_i)}{\mathbb{P}(\mathbf{Y}|\mathbf{G}, \text{Null})} = \frac{|\mathbf{Q}_0^*|^{c^*/2} v^{*P_+/2}}{|\mathbf{Q}_i^*|^{c^*/2} v^{P_+/2}} \quad (2.15)$$

where P_+ is the number of associated traits for model M_i , and

- $\mathbf{S} = \mathbf{I} - \mathbf{Y}_-(\mathbf{Q}_{--} + \mathbf{Y}_-^T \mathbf{Y}_-)^{-1} \mathbf{Y}_-^T$
- $v^* = (\mathbf{G}^T \mathbf{S} \mathbf{G} + v^{-1})^{-1}$
- $\mathbf{m}^* = v^* \mathbf{G}^T \mathbf{S} \mathbf{Y}_+$
- $c^* = c + N$
- $\mathbf{Q}_0^* = \mathbf{Q}^* + \mathbf{Y}_+^T \mathbf{S} \mathbf{Y}_+$
- $\mathbf{Q}_i^* = \mathbf{Q}^* + \mathbf{Y}_+^T \mathbf{S} \mathbf{Y}_+ - \mathbf{m}^{*T} v^{*-1} \mathbf{m}^*$

Here $\mathbf{Y}_+$ and $\mathbf{Y}_-$ are the associated and unassociated phenotypes for model M_i , $\mathbf{Q}_{--}$ and $\mathbf{Q}^*$ are derived from the $\mathbf{Q}$ matrix just as the relevant Σ entries (definitions in Appendix A with $- := 1$ and $+ := 2$). Note, $\mathbf{S}$ is an $N \times N$ matrix and for computational speed in an algorithm implementation, direct calculation can be avoided by appropriately re-writing the $\mathbf{Q}^*$ equations.

2.3.3 Model averaging across single phenotype models

Following the idea of Equations 2.7 and 2.8 in expressing a Bayes factor of association BF_A , another statistic can be defined to represent the Bayes factor of association on models with only a single phenotype considered related to the

2. A Bayesian model averaging approach

genetic variant:

$$BF_S := \sum_{i=1}^P BF_i p_i^s \quad (2.16)$$

where BF_S is a model average between the single trait association models only and p_i^s is the prior on model M_i given that it is a single trait association model. It is important to make it clear that these are *joint* models of all P phenotypes assuming just one of the traits is associated and $P - 1$ traits are not associated. These joint models are different from modelling each trait independently, since the joint error structure of the phenotypes is modelled via the joint approach.

The BF_S statistics is a special case of BF_A where we put no prior belief in the pleiotropic models i.e. models with more than one trait associated. The reason for introducing this statistic is that a comparison between BF_A and BF_S can then tell us something about the degree of pleiotropy of a variant. This however can also be done by directly comparing the set of pleiotropic and the set of non-pleiotropic models of association and measuring their difference by a Bayes factor of pleiotropy, which can be defined as follows:

$$BF_P := \frac{\sum_{i=P+1}^{2^P-1} \mathbb{P}(\text{Data}|M_i) p_i^m}{\sum_{i=1}^P \mathbb{P}(\text{Data}|M_i) p_i^s} = \frac{\sum_{i=P+1}^{2^P-1} BF_i p_i^m}{\sum_{i=1}^P BF_i p_i^s} \quad (2.17)$$

where p_i^s are defined as in Equation 2.16 and p_i^m is the prior on model M_i given that it is a multiple trait association model.

2.4 Bayes Factor Threshold

A question which arises with our proposed method is what values of our test statistic (BF_A) should be considered to give us sufficient evidence for association. In [WTCCC, 2007] a prior probability of association of 10^{-5} is suggested for univariate analysis Bayes factors ([Marchini, 2009]), based on an expectation of 10 detectable genes involved in a disease and 10^6 independent regions in the genome. If C_a , C_b are the relative costs of false discovery and false non-discovery, then it is suggested to declare a variant associated if the posterior probability of association is greater than $C_a/(C_a + C_b)$, or equivalently if $BF \geq C_a/C_b \times 1/\text{Prior Odds}$. For a discovery step $C_a = 1$ and $C_b = 100$ are proposed in [Marchini, 2009] giving a Bayes factor threshold of 10^3 . A similar threshold for our multivariate approach statistic would need to be investigated, however for now we employ the same threshold.

In frequentist testing 5×10^{-8} is accepted as the genome-wide significance P-value threshold. Due to the different rationale for setting the thresholds in Bayesian and frequentist approaches, exact comparison is not possible, however under general assumptions [Sellke et al., 2001], $BF < -1/(e p \log(p))$ for $p < 1/e$ [Stephens and Balding, 2009]. This translates $p = 5 \times 10^{-8}$ into $\log_{10} BF < 5.64$ (upper bound slightly above the proposed 10^5 BF threshold by [WTCCC, 2007]) (10^3 BF upper limit translates to $p = 10^{-4.5}$ however actual correspondence may be for a much smaller P-value). In addition, BF and P-value produce approximately the same ranking if the variance of the normal prior on the effect sizes in the Bayesian case is proportional to the sample size and asymptotically proportional to $1/(\text{MAF}(1 - \text{MAF}))$ [Stephens and Balding, 2009].

2.5 Univariate approach

The usual approach of analysing multiple traits in GWAS is by performing separate analyses for each phenotype of interest (e.g. [Soranzo et al., 2009]). This in turn (in the Bayesian framework) gives P Bayes factors BF_i^u , $i = 1, 2, \dots, P$, where each one is obtained by ignoring any information about the others. To compare our multivariate approaches to such univariate approach use a Bayesian model average over all P univariate models as

$$BF_U := \sum_{i=1}^P BF_i^u p_i^u \quad (2.18)$$

where p_i^u is the prior on univariate model i . This statistic (although not necessarily the best way to combine the univariate Bayes factors) should still give us an idea of the differences between our approach and univariate GWAS analysis.

2.6 Multivariate extension of univariate approach

Another interpretation of the univariate models and model Bayes factors is that of a Full versus Null model comparisons (on a single trait). A multivariate extension of this is also to consider the Full and Null models on all subsets of the traits. An example with the models on two traits is given in Figure 2.2.

Then again model averaging over the resulting model Bayes factors gives

$$BF_J = \sum_{i=1}^{2^P-1} BF_i^u p_i^j \quad (2.19)$$

This is another multivariate modelling approach and a comparison between

2. A Bayesian model averaging approach

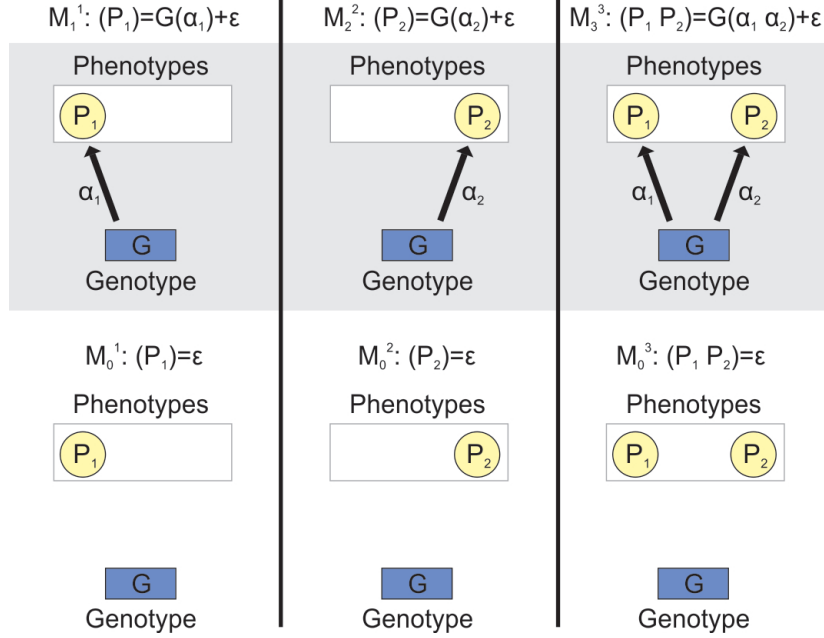


Figure 2.2: Graphical representation of the Full (top) and corresponding Null (bottom) models on subsets of 2 phenotypes.

it and the univariate method can also be informative about the potential advantages/disadvantages of performing joint analysis. Additionally the univariate model Bayes factors are part of BF_J , which should make it easier to see whether we would gain anything by using the more complex approach. Note however that this method would not give us comparable results between the model Bayes factors as more than one model could be true here (e.g. in a 2 phenotype example if the variant is causal to both traits, all 3 models should show association). A similar measure of pleiotropy here as BF_P hence would not be sensible and results from the individual model comparisons might not be easily interpreted. We then only include this approach for comparison purposes and use the same priors on the model parameters. Note, however, that results obtained from both multivariate methods should be the same if our data comes from a model with

uncorrelated errors.

2.7 AVG Method

In addition to our proposed multivariate approach in Section 2.1, we also consider model averaging between all of its multivariate models and all of the univariate ones from Section 2.5. Here we define

$$BF_{AVG} := \frac{BF_A + BF_U}{2} \quad (2.20)$$

The reason for introducing this approach is that in the case where at least one of BF_A and BF_U pick up some association signal, then so should the BF_{AVG} . Moreover if neither of them do, it should also show no association.

Strictly speaking such a definition (average of Bayes factors) does not necessarily yield an actual Bayes factor, however in our setting this is the case under the following conditions. Each of the univariate Bayes factors part of BF_U can be written as follows

$$BF_i^u = \frac{\mathbb{P}(\mathbf{Y}_i | \mathbf{G}, \text{Full})}{\mathbb{P}(\mathbf{Y}_i | \text{Null})} = \frac{\mathbb{P}(\mathbf{Y}_i | \mathbf{G}, \text{Full}) \mathbb{P}(\mathbf{Y}_{-i} | \mathbf{Y}_i, \mathbf{G}, \text{Full})}{\mathbb{P}(\mathbf{Y}_i | \text{Null}) \mathbb{P}(\mathbf{Y}_{-i} | \mathbf{Y}_i, \text{Null})}$$

if and only if the set of all phenotypes with the exception of phenotype $\mathbf{Y}_i$ (denoted as $\mathbf{Y}_{-i}$) are conditionally independent of $\mathbf{G}$ given phenotype $\mathbf{Y}_i$. With this representation we now have the same denominator as for all models part of BF_A and thus the BF_{AVG} statistic can be thought of as a model averaging Bayes factor, where each model is a multivariate phenotype one and the set of

models coming from BF_U represent special models in which some phenotypes are conditionally independent of the genotype given the others. The final averaging between BF_A and BF_U is part of the prior on each of the models.

2.8 Combining results across cohorts

In Chapter 4 we apply our methods to 3 distinct cohorts with the same set of multiple phenotypes. In order to combine results across these 3 cohorts we combine Bayes factors by multiplying them together at the sub-model level, which can be thought of as an assumption of independence of the three datasets. Meta-analysis of multiple phenotype GWAS is an open problem in the field which we have not focussed on here. Even though this might not be the best way to combine data from different cohorts, it has still allowed us to do a preliminary comparison between univariate and multivariate methods. Future work would include extending the approaches to include random effects to model different cohorts.

2.9 Summary

In this Chapter, we considered the problem of modelling multivariate phenotypes in GWAS and proposed Bayesian model averaging approaches to assess the association between a single genetic variant and some or all of the traits of interest. Our framework allows for simultaneous testing and description of the type of association and provides an alternative to the commonly used univariate GWAS models. In the next two chapters, we illustrate the properties of our methods on

2. A Bayesian model averaging approach

both simulated data (Chapter 3) and GWAS data on 6 haematological parameters from [Soranzo et al., 2009].

We have to note that recently a study has been published on a method similar to our proposed approach, which instead of bi-partitioning the phenotypes, splits them into three types of classes of un-associated, directly unassociated and indirectly associated (independent of the genetic variant given the directly associated traits) [Stephens, 2013]. This method was not available/published at the time this analysis has been carried out.

Chapter 3

Simulation studies

In this Chapter we use simulated data in order to investigate the performance of the methods from Chapter 2 and of particular interest to us is understanding the differences between the multivariate and univariate phenotype methods.

3.1 Data simulation

We start by explaining how we generate datasets.

- First we simulate a vector of N genotypes $\mathbf{G}$ at a single locus, each element of which is thought to come from a Binomial $Bin(2, p_{\mathbf{G}})$ distribution, where $p_{\mathbf{G}}$ is the minor allele frequency at that locus.
- Next we simulate N multivariate (P -variate) normal samples with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ for the model errors $\mathbf{\epsilon}$
- Finally, according to some vectors of means $\mathbf{\mu}$ and effect sizes $\mathbf{\alpha}$, we set $\mathbf{Y} = \mathbf{1}\mathbf{\mu} + \mathbf{G}\mathbf{\alpha} + \mathbf{\epsilon}$, where $\mathbf{1}$ denotes an $N \times 1$ (column) vector of ones.

This way we obtain simulated genotypes and multivariate normal samples of phenotypes, with pre-specified genotype minor allele frequencies, error covariance structure, vectors of phenotype means and genotype effect sizes. Knowing the "truth", i.e. the model and parameters under which data is generated, would enable us to check the performance of the different methods from Chapter 2 and compare multivariate and univariate trait analyses.

3.2 Simulation scenarios

We simulate data under a number of different scenarios based on the parameters in Table 3.2. In each case we generate 100 000 samples, which we analyse using all different approaches from Chapter 2 (also briefly summarised in Table 3.1). We then find and compare the power for detecting associations between the methods.

We only generate data with non-negative effect sizes, as due to the shape of the multivariate normal distribution (likelihood of the data) in terms of power, data coming from a model with $\{ \boldsymbol{\alpha} = (\alpha_1 \ \alpha_2) \text{ and error correlation } \rho \}$ would give the same results as data with $\{ \boldsymbol{\alpha} = (-\alpha_1 \ -\alpha_2) \text{ and error correlation } \rho \}$ and data with $\{ \boldsymbol{\alpha} = (\alpha_1 \ -\alpha_2) \text{ and error correlation } -\rho \}$.

3. Simulation studies

BF	Expression	Further information
BF_A	$\sum_{i=1}^{2^P-1} \frac{\mathbb{P}(\mathbf{Y} \mathbf{G}, M_i)}{\mathbb{P}(\mathbf{Y} \text{Null})} p_i$	Each model M_i is a bi-partition of the set of phenotypes into a set of associated and unassociated to the genotype
BF_S	$\sum_{i=1}^P \frac{\mathbb{P}(\mathbf{Y} \mathbf{G}, M_i)}{\mathbb{P}(\mathbf{Y} \text{Null})} p_i^s$	Each model M_i has a single phenotype as associated to the genotype and all others as unassociated
BF_U	$\sum_{i=1}^P \frac{\mathbb{P}(\mathbf{Y}_i \mathbf{G}, \text{Full})}{\mathbb{P}(\mathbf{Y}_i \text{Null})} p_i^u$	Each model considers only a single trait and ignores any possible association of all other traits
BF_J	$\sum_{i=1}^{2^P-1} \frac{\mathbb{P}(\mathbf{Y}' \mathbf{G}, \text{Full})}{\mathbb{P}(\mathbf{Y}' \text{Null})} p_i^j$	Each model is a Full (or Null) model on a subset $\mathbf{Y}'$ of the traits $\mathbf{Y}$ and is ignoring the remaining traits not in $\mathbf{Y}'$
BF_{AVG}	$\frac{1}{2} (BF_A + BF_U)$	
BF_P	$\frac{\sum_{i=P+1}^{2^P-1} \mathbb{P}(\mathbf{Y} \mathbf{G}, M_i) p_i^m}{\sum_{i=1}^P \mathbb{P}(\mathbf{Y} \mathbf{G}, M_i) p_i^s}$	Each model in the denominator is a model with exactly one associated to the genotype phenotype (with all others considered unassociated) and each model in the numerator is a model with more than one associated phenotype (with all others considered unassociated)

Table 3.1: Bayes factor summary for the statistics introduced in Chapter 2.

Sample size	$N = 2000$ individuals
Number of phenotypes	$P = 2$
Minor allele frequency	$p_G = 0.1$
Vector of intercepts	$\mathbf{m} = (10 \ 0)$
Error variances	$\sigma_1^2 = \sigma_2^2 = 1$
Effect sizes	$\boldsymbol{\alpha} = \begin{cases} (0 \ 0) \\ (0 \ 0.1) \\ (0.2 \ 0.1) \end{cases}$
Error correlations	$\rho = \begin{cases} 0 \\ \pm 0.2 \\ \pm 0.4 \\ \pm 0.9 \end{cases}$

Table 3.2: Values of the parameters used for data simulation. All 21 possible combinations are considered

3.3 Power of Bayesian and Frequentist methods

It is usually the case that Bayesian test statistics do not have an analytic distributional form when data arise from a particular model, such as a null model of no association. This complicates the direct assessment of the power of a Bayesian method and comparison to other (possibly non-Bayesian) approaches.

We use simulations under null and non-null models of association to calculate empirical distributions of all the test statistics we consider, and use these distributions to assess power. The power of a given statistic at a specific false positive

rate α is given by the percentage of test statistics from the non-null distribution which occur above a threshold of q_α , where q_α is the $100(1 - \alpha)$ -quantile of null distribution for the statistic. By varying α we can assess power at different levels of significance and construct Receiver operating characteristic (ROC) curves.

3.3.1 Single trait association models

Figure 3.1 shows ROC curves for all of the methods in Chapter 2 under a model where just one of the two traits is associated ($\alpha = (0 \ 0.1)$).

The first observation is that in the absence of a correlation in the noise ($\rho = 0$) all methods perform the same. Under this assumption, this can be explained by the fact that fitting any bivariate model in this case would be equivalent to fitting both of its univariate submodels separately. Hence the Bayes factor BF_1 comparing model $(\mathbf{Y}_1 \ \mathbf{Y}_2) = \mathbf{G}(\alpha_1 \ 0) + \boldsymbol{\epsilon}$ to the bivariate null and the univariate Bayes factor BF_1^u comparing $(\mathbf{Y}_1) = \mathbf{G}(\alpha_1) + \boldsymbol{\epsilon}$ to its corresponding null should be the same.

As the error correlation increases, irrespective of its sign, the power of all multivariate approaches also increases. Figure 3.2 gives an illustration of why this occurs (using much larger effect sizes for clarity) by plotting the joint distribution of two phenotypes under a model where just one of the two traits is associated, for 3 different values of the error correlation (ρ). The phenotypes are coloured according to genotype which illustrates how the separation between genotype classes increases with ρ . This increasing separation between genotype groups can be detected by the multivariate approaches only.

3. Simulation studies

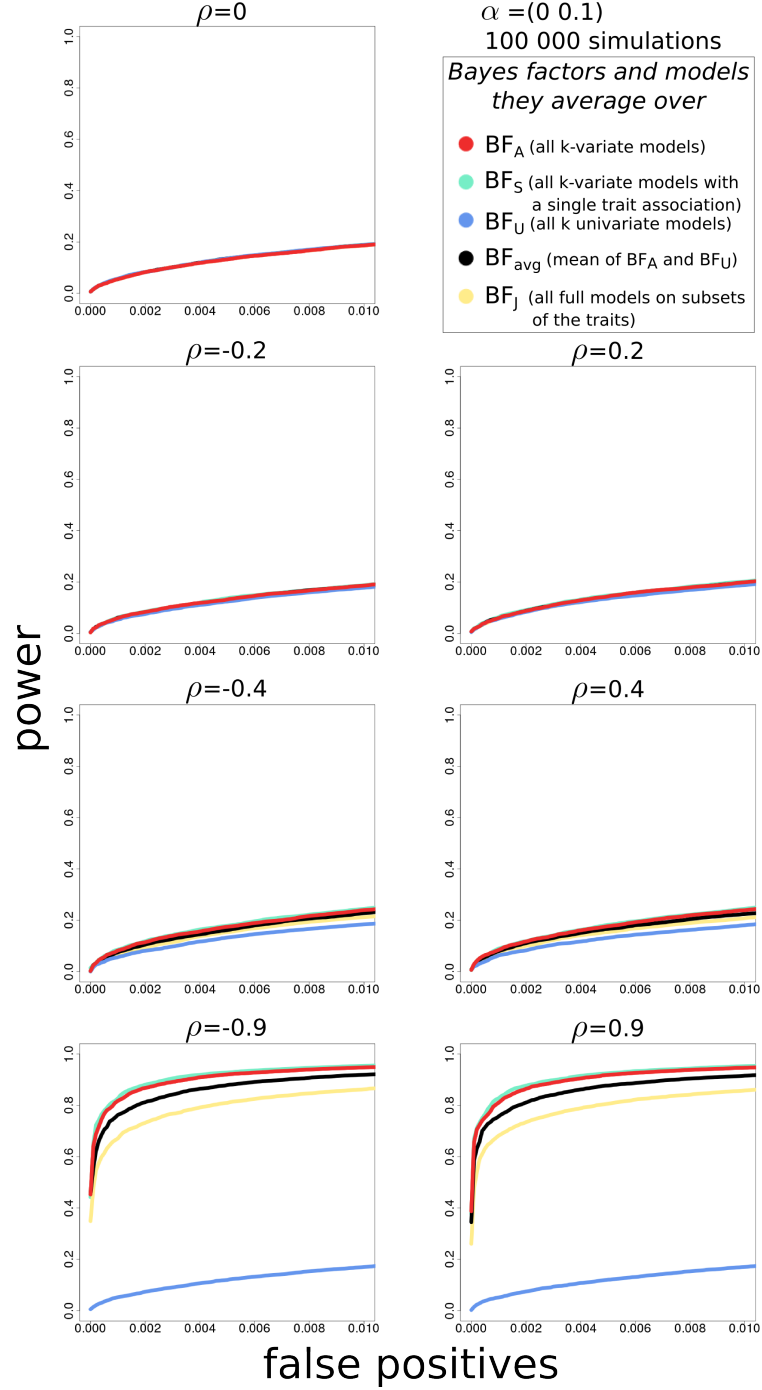


Figure 3.1: ROC curves for the different Bayes factors from Chapter 2 obtained from 100 000 datasets generated for 2 traits, with only one of them associated to the genetic variant (effect size 0.1), for $n = 2\,000$ individuals, for each of the 7 different error correlation structures from Table 3.2.

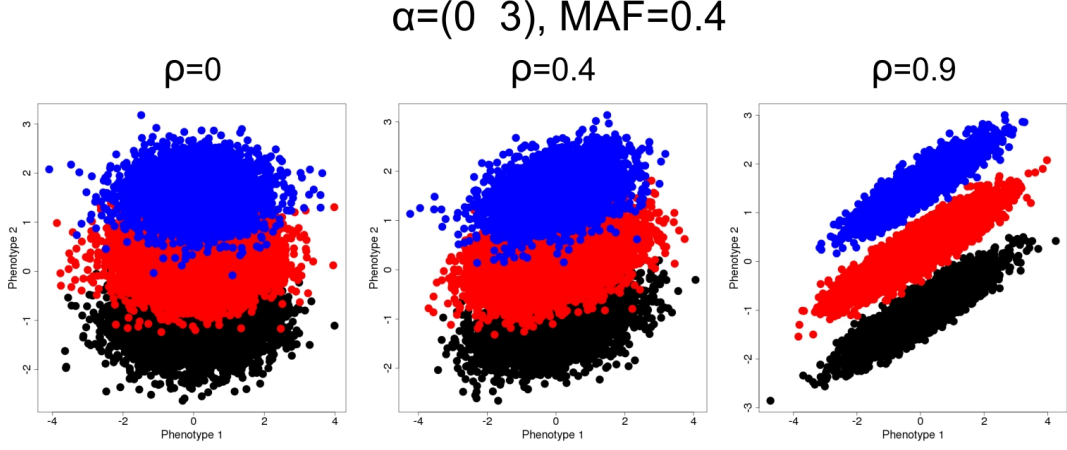


Figure 3.2: Data on 2 traits, one associated (effect size 3), $MAF = 0.4$, $n = 20\,000$, model error correlation ρ . Colours represent $G = 0$ in black, $G = 1$ in red and $G = 2$ in blue.

To quantify this effect we can calculate the squared Mahalanobis distance [Maesschalck et al., 2000], between the vector of effect sizes under the null and alternative models, which takes into account the correlation structure of the errors. This measure is a weighted sum of the inner product of α with the eigenvectors of covariance matrix, with eigenvalues as weights. Since effect sizes are zero under the null this reduces to $M_{\Sigma}^2 = \alpha \Sigma^{-1} \alpha^T$. For the three examples in Figure 3.2 where $\alpha = (0 \ 0.1)$ we get $M_{\Sigma}^2 = 9, 10.7, 47.4$ for $\rho = 0, 0.4, 0.9$ respectively. Table 3.3 shows this metric for the scenarios in Figure 3.1. The pattern of this metric is correlated with the differences in power between the multivariate and univariate approaches. It may be possible to derive the power of our multivariate approaches as a function of M_{Σ}^2 but we have not done this here.

Figure 3.1 also shows that the order of the different approaches remains the same in all scenarios with $BF_S > BF_A > BF_{AVG} > BF_J$. These differences only become apparent when the correlation increases. The true model is fit by

$\alpha = (0 \ 0.1)$							
ρ	0	0.2	0.4	0.9	-0.2	-0.4	-0.9
M_{Σ}^2	0.01	0.0104	0.0119	0.0526	0.0104	0.0119	0.0526

Table 3.3: Squared Mahalanobis distances from α to $\mathbf{0}$, with error correlation ρ and unit error variances (calculated with the parameters used for the data simulations).

both BF_S and BF_A with BF_S putting more prior weight on the true model, and explains why BF_S has slightly more power in this case. BF_{AVG} model averages between all models fit by BF_A and the less powered here BF_U , hence the ordering $BF_A > BF_{AVG}$. Finally, BF_J appears to have the least power out of the multivariate approaches and this is because it does not explicitly fit the true model (as opposed to all other multivariate methods that do).

3.3.2 Pleiotropic Models

Figure 3.3 shows ROC curves for all of the methods in Chapter 2 under a model where both of the two traits are associated ($\alpha = (0.2 \ 0.1)$). Again we observe very similar power for all methods when $\rho = 0$. As ρ becomes more negative, we observe that the multivariate approaches are more powerful than the univariate approach BF_U . When ρ is positive, we first observe a drop in power in all multivariate models for $\rho = 0.2$ and then $\rho = 0.4$, followed by an increase in power for $\rho = 0.9$. Consistent results are shown by the squared Mahalanobis distances for these scenarios (Table 3.4).

Figure 3.4 provides an illustration of the phenotype-by-genotype group separation as ρ varies. Table 3.5 details the values of M_{Σ}^2 for each of the examples. Here

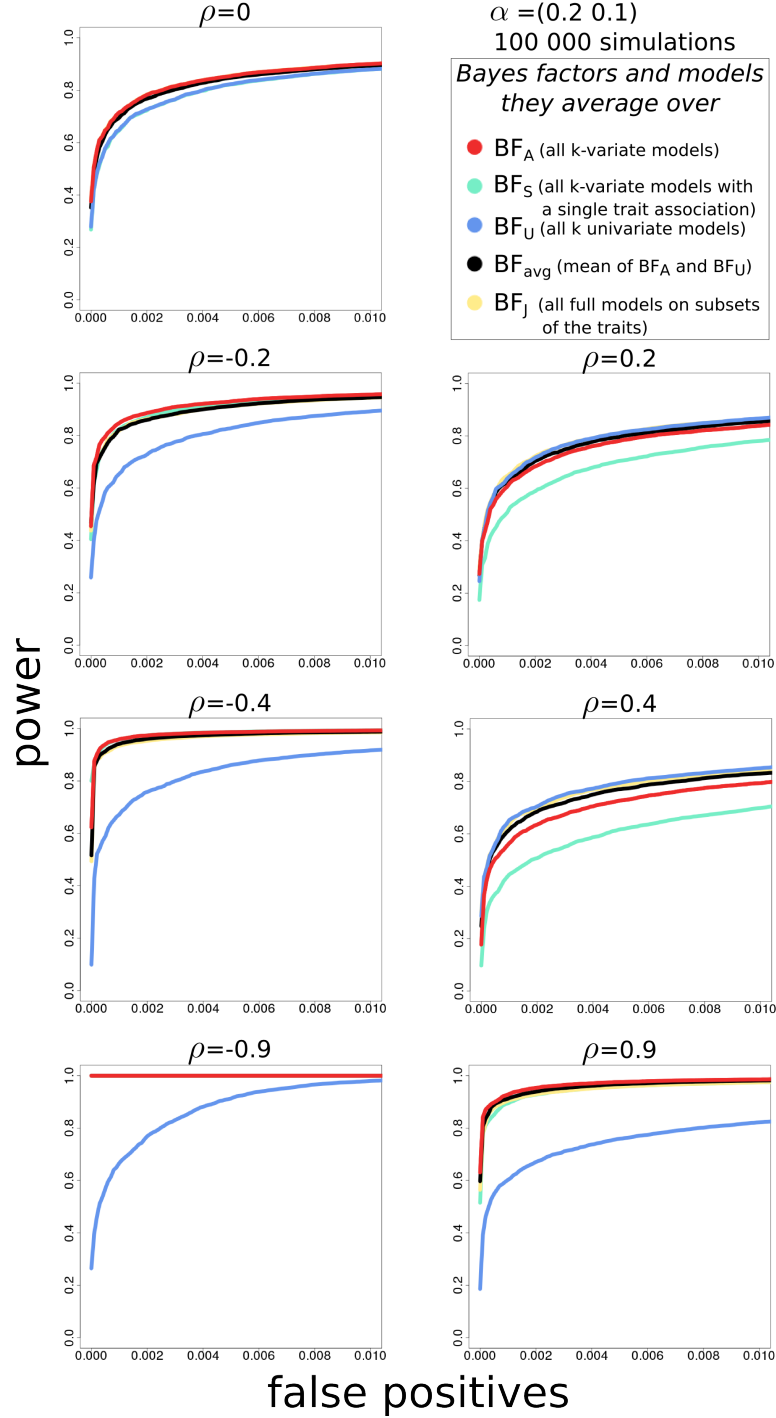


Figure 3.3: ROC curves for the 5 approaches. Data generated for 2 traits, both associated (effect sizes 0.2 and 0.1), 100 000 datasets generated, $n = 2\,000$ individuals, for the 7 different error correlation structures from Table 3.2.

3. Simulation studies

$\alpha = (0.2 \ 0.1)$							
ρ	0	0.2	0.4	0.9	-0.2	-0.4	-0.9
M_{Σ}^2	0.05	0.0438	0.0405	0.0737	0.0604	0.0786	0.4526

Table 3.4: Squared Mahalanobis distances from α to $\mathbf{0}$, with error correlation ρ and unit error variances (calculated with the parameters used for the data simulations).

under negative noise correlations we observe clearer separation of the clusters of points, compared to the no correlation case ($\rho = 0$) and under positive correlation we see the opposite, with the best case being $\rho = -0.9$ and the worst $\rho = 0.9$, which is again consistent with the Mahalanobis distance results.

In Figure 3.3 we observe power ordering of $BF_J > BF_A$ whenever $BF_U > BF_S$ (and vice versa). This is because the first two methods fit all models in the corresponding second two with additionally both fitting the Full model. By construction BF_{AVG} achieves power between the one for BF_A and BF_U . Overall, best power for all scenarios is achieved by BF_J , however as described in Chapter 2, results from this approach are not easily interpretable, with BF_{AVG} being the next best method and followed by BF_A .

ρ	0	0.4	0.9	-0.4	-0.9
M_{Σ}^2	25	18.3	17.9	41.2	245.3

Table 3.5: Squared Mahalanobis distances for data from Figure 3.4.

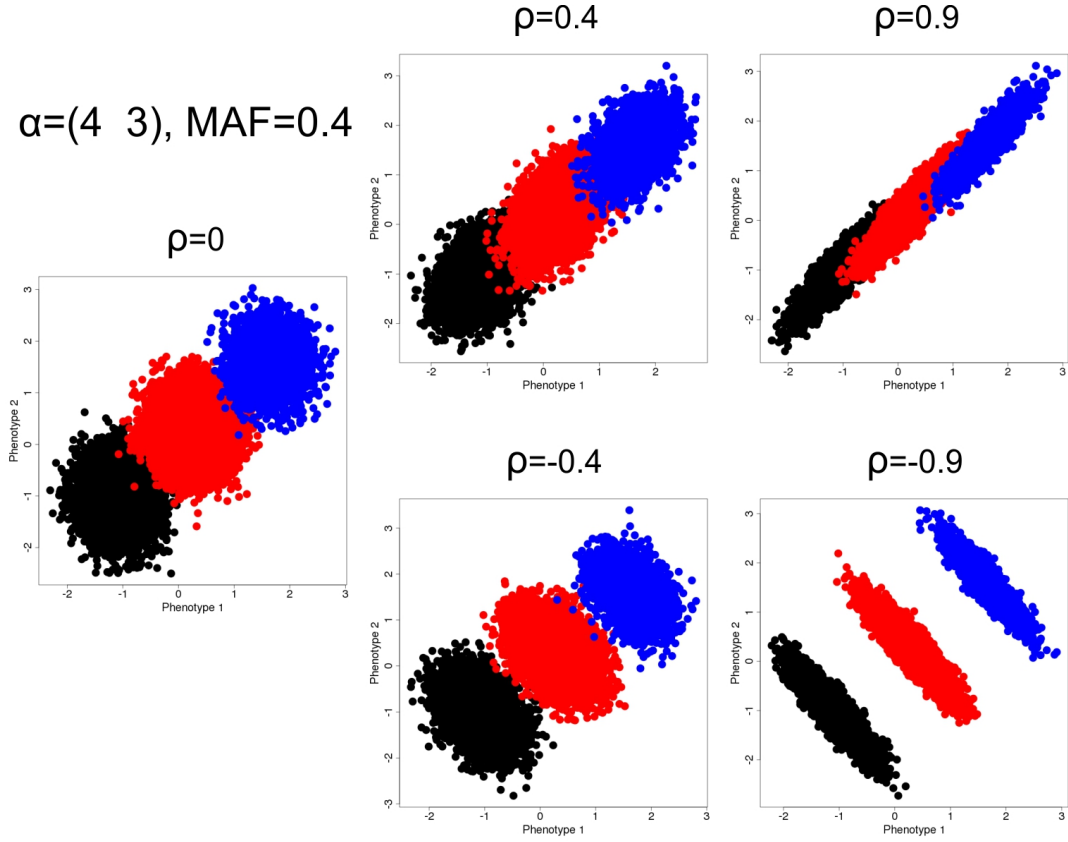


Figure 3.4: Data on 2 traits, both associated (effect size 4 and 3), $MAF = 0.4$, $n = 20\,000$, model error correlation ρ . Colours represent $G = 0$ in black, $G = 1$ in red and $G = 2$ in blue.

3.4 Bayes factor for pleiotropy

Figure 3.5 shows the power of the pleiotropy statistic BF_P for the models from Table 3.2 for different values of ρ , where the null distributions are obtained from the single trait association models and the alternatives - from their corresponding pleiotropic ones. Note that we define the null for BF_P as association with exactly one associated trait (which is different to the model of no association which we use as a null for the other BF statistics).

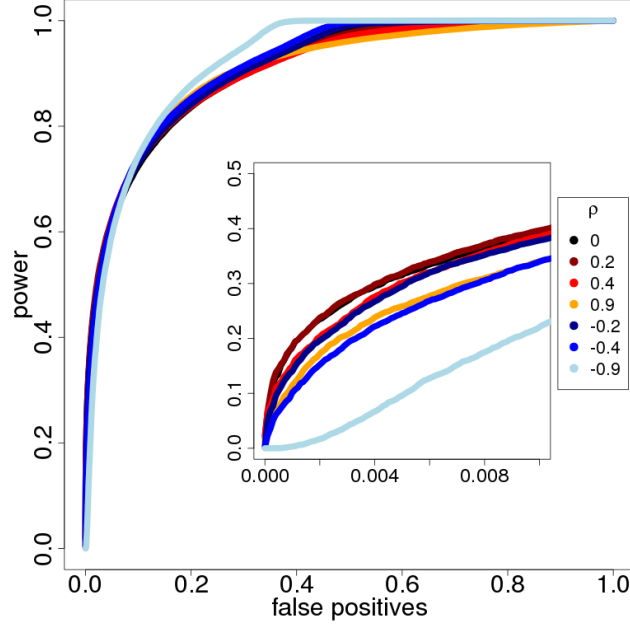


Figure 3.5: ROC curves for BF_P for different model error correlations.

When we allow for $< 10\%$ false positives, BF_P seems to perform best when there is no or small positive error correlation between the traits, with the worst possible case being large negative correlation. An interesting observation is that the most powerful case for finding associations is the least powerful for actually detecting pleiotropy.

The power for detecting pleiotropy varies between 20% and 40% when allowing for 1% false positives. This shows that we have more power for detecting associations in data coming from a pleiotropic model than detecting pleiotropy itself.

3.5 Conclusions

Our proposed multivariate approach shows an increase in power over univariate analysis in some scenarios with pleiotropic data and all single trait association cases. It also shows an increase in the posterior probability of association for traits with weak effects when other phenotypes are also associated to the genetic variant (with stronger effects). The Bayes factor of pleiotropy allows us to distinguish between single trait and pleiotropic association models.

Chapter 4

Real Data Analysis

In this Chapter we apply our methods from Chapter 2 to a real study of multiple haematological phenotypes that has been previously analysed with univariate GWAS methods.

4.1 Data overview

We analyze the HaemGen consortium dataset [Soranzo et al., 2009] which comprises of measurements on eight hematological parameters (Table 4.1) on over 4600 individuals from 3 population cohorts - KORA [Wichmann et al., 2005], TwinsUK ¹ and UKBS-CC1 from WTCCC ², together with a number of covariates and imputed genome-wide genotypes. Full specifications of the data can be found in the supplementary material of [Soranzo et al., 2009].

¹<http://www.twinsuk.ac.uk/>

²<http://www.wtccc.org.uk/>

Red Blood Cell	HB	Hemoglobin
	MCH	Mean corpuscular hemoglobin
	MCHC	Mean corpuscular hemoglobin concentration
	MCV	Mean corpuscular volume
	RBC	Cell Count
Platelet Cell	PLT	Platelet count
	MPV	Mean platelet volume
Red Blood Cell	WBC	Cell count

Table 4.1: Hematological traits

After the replication stage in [Soranzo et al., 2009], 22 loci (23 SNPs) were found to be associated with these traits by performing 8 single phenotype analyses. In addition to that pleiotropy was *observed* at some loci. As such the HaemGen dataset is a particularly interesting dataset to work with in order to check how our method performs.

4.1.1 Pre-processing

The phenotypes were transformed to allow us to use models that assume multivariate normality. We followed [Soranzo et al., 2009] and take the log of RBC, MPV and WBC and leave all other traits untransformed. Histograms, pairs plots and correlations for UKBS are given in Figure 4.1. From the histograms it can be seen the normal approximation seem plausible, so the variable transformations seem appropriate. We can also observe that some traits are highly positively/negatively correlated (MCH, MCV - 0.94 ; logMPV, PLT - -0.48), whereas others don't seem to be related at all. Plots for the other two datasets are similar and are not shown here. Since MCH and MCHC [Soranzo et al., 2009] are de-

4. Real Data Analysis

rived from other traits (RBC, HB and MCV) we remove these from the analysis. From the full dataset, we remove individuals with missing data for any of the

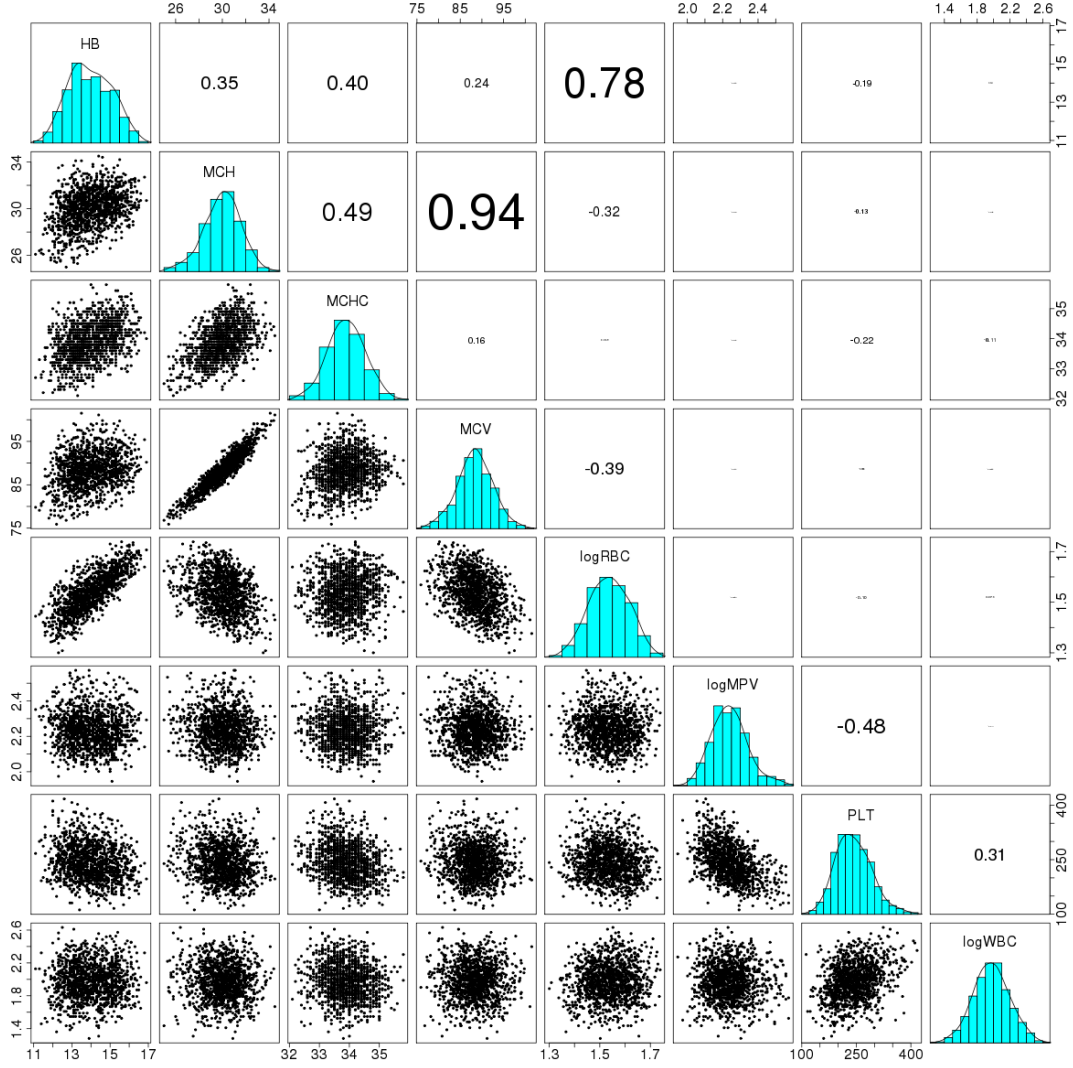


Figure 4.1: Pairs plots, histograms and correlations - UKBS

phenotypes or the covariates. In addition for TwinsUK we only include unrelated individuals in our analysis. At the end we are left with a total of 4583 samples after removing a number of potential outliers that did not seem to satisfy the

(multivariate) normality assumption of the data (after transformation).

We included all SNPs with data for *all* three cohorts, leaving a total of over 2 million variants for our analysis. We use imputed data and for each variant, we consider individual i to have a missing genotype if the sum of the probabilities of each of the 3 possible genotype values is less than 0.9, and do not include samples with missing genotype for the analysis of that particular SNP.

We used covariates Sex and Age for all three cohorts, together with Instrument (factor coding for instrument used to make measurements) for TwinsUK and Geographical region for UKBS.

4.1.2 Prior on models of association

An advantage of our Bayesian approach is that we can fully specify our prior beliefs on the models that we are fitting. The blood traits measured in this study represent properties of three classes of blood cells - red and white blood cells and platelets.

Each of these groups has different biological functions and it makes sense to think that traits within the same class are more likely to be more similar than traits from different classes. Each of these cell types are derived from stem cells via a process of differentiation known as hematopoiesis. Figure 4.2 from [Metcalf, 2007] gives an overview of what is currently known about this process in terms of the possible trajectories of stem cells to blood cell types via intermediate stages. The process, as represented in the figure, has a tree like structure,

with some blood cell types (on the right) sharing intermediate cell types. For example, we can see that the red blood cells (Erythrocyte) and the platelets are much closer to each other than to any other pair of the three classes. If we consider a single mutation affecting a particular cell (or the process of forming that cell) in the tree, then we can look at the different possible lists of affected cells. Thus we can distinguish a case in which only one cell type is affected (i.e. red, white or platelets). Another, in which platelets and erythrocytes are both affected and a third in which all three groups are affected. Additionally, we can consider the case where a single cell of a particular type is affected.

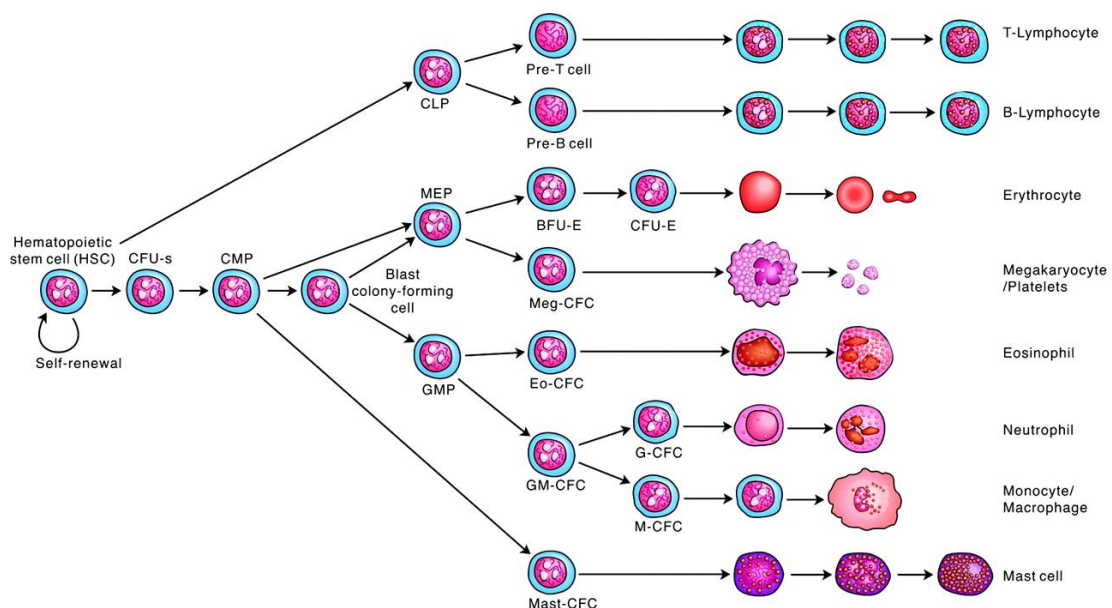


Figure 4.2: Hematopoiesis [Metcalf, 2007]

Having discussed this process with our collaborator Dr Nicole Soranzo, we decided to use a prior with more weight on the models described above, as well as the model with red and white blood cell traits associated, and this with white blood cells and platelets associated. The final two we include for completeness. Then we put a weight of 0.8 (and uniform within the group) on models in Table

4.2 and a weight of 0.2 (and again uniform within) to the rest.

	Red blood cell			Platelet		White blood cell
	HB	MCV	RBC	PLT	MPV	WBC
Single trait	HB	-	-	-	-	-
	-	MCV	-	-	-	-
	-	-	RBC	-	-	-
	-	-	-	MPV	-	-
	-	-	-	-	PLT	-
	-	-	-	-	-	WBC
RBC	HB	MCV	RBC	-	-	-
PLT	-	-	-	MPV	PLT	-
RBC & PLT	HB	MCV	RBC	MPV	PLT	-
RBC & WBC	HB	MCV	RBC	-	-	WBC
PLT & WBC	-	-	-	MPV	PLT	WBC
All	HB	MCV	RBC	MPV	PLT	WBC

Table 4.2: Models with a larger prior belief, each row representing a single model

4.2 Results

We first look at the 22 loci identified in [Soranzo et al., 2009] as they have been discovered using the same data described in this Chapter and subsequently successfully replicated. Thus we can use them to compare results from the univariate and our multivariate approach using real data. After that we perform genome-wide analysis and comment on our findings.

4.2.1 Results at established blood traits loci

We employ all approaches from Chapter 2 on the 22 loci discovered in [Soranzo et al., 2009], results from which are reported in Table 4.3. Our main focus is on comparison of our approach (BF_A) to the univariate one (BF_U). In the BF_A column of the table of results there are a number of variants coloured in yellow and blue. The ones in blue are SNPs for which $BF_U > 10^3$ but $BF_A < 10^3$ and in yellow the ones for which $BF_U < 10^3$ but $BF_A > 10^3$. In other words, based on a threshold of 10^3 on both statistics the ones in blue would be missed by our approach and the ones in yellow would be missed by the univariate method but identified in ours. Based on the same threshold of 10^3 however the extension of the multivariate approach BF_{AVG} would not miss any of the loci that the univariate approach identifies and in addition would also identify all but the last SNP in yellow.

Some loci show a larger signal for BF_U than for our multivariate approaches, for example SNP rs9609565 on chromosome 22. In Figure 4.3 we provide a fine-scale view of the Bayes Factors at this SNP for each sub-model for each cohort, as well as for the combined analysis across all 3 cohorts. We can see that the signal here is entirely driven by the KORA dataset. Since the best models in KORA seem to be pleiotropic (on at least 3 traits) but not part of the favoured sets of models apriori (identified by the process of haematopoiesis), it follows that the multivariate statistics would be down-weighted compared to the univariate ones.

chr	SNP	position	HB	MCV	RBC	MPV	PLT	WBC	BF_U	BF_J	BF_S	BF_A	BF_P	BF_{AVG}	max BF_i
22	rs5756506	35797338	0.75	2.27	-1.44	-1.20	-0.67	0.78	1.52	4.89	3.22	4.58	1.58	4.28	6.44
6	rs11970772	42033268	-1.24	4.16	1.46	-1.01	-0.46	-0.85	3.38	3.02	-0.83	1.16	2.22	3.09	3.03
6	rs1800562	26201120	0.44	8.73	0.48	-0.01	-0.43	-0.13	7.95	7.88	0.60	7.09	6.71	7.71	7.83
22	rs9609565	31197528	-0.54	2.92	-1.24	-1.12	-1.21	-1.42	2.14	1.76	-0.29	-0.09	0.29	1.84	1.72
6	rs9402686	135469510	-1.48	12.16	8.31	-0.52	1.17	0.45	11.38	17.14	4.77	17.94	13.39	17.64	19.95
7	rs7385804	100073906	-0.96	1.68	3.31	-0.54	-0.54	-1.59	2.54	2.19	-0.18	0.87	1.26	2.25	2.83
1	rs10914144	170216373	-0.50	-1.08	-1.11	4.30	1.16	-1.29	3.52	3.62	3.32	4.35	1.24	4.11	5.47
15	rs11071720	61129049	-1.59	-0.89	-1.19	2.09	-0.71	-1.00	1.32	0.94	-0.05	-0.19	-0.27	1.03	0.70
11	rs11602954	192856	-1.32	-0.49	-0.96	2.28	0.08	-0.63	1.51	1.26	0.32	1.26	1.14	1.40	2.31
3	rs12485738	56840816	-1.24	-0.88	-1.51	8.33	1.08	-0.55	7.55	7.22	4.65	5.76	1.32	7.26	6.81
1	rs1668873	203502613	-1.59	-0.52	-1.00	6.16	0.24	-0.98	5.38	5.01	3.31	3.56	0.35	5.08	4.51
17	rs2138852	24727475	-1.78	-1.37	-0.81	9.44	0.30	0.64	8.66	8.54	5.30	7.55	2.48	8.39	8.93
10	rs2393967	64803162	-1.20	0.01	-1.14	4.51	0.11	-1.38	3.73	3.38	1.45	1.79	0.47	3.43	2.83
7	rs342293	106159455	-1.30	-1.43	-1.71	8.61	2.50	-1.03	7.83	7.47	4.39	5.85	1.68	7.53	6.98
20	rs6136489	1871734	-1.38	-0.53	-1.06	3.29	0.45	0.29	2.51	6.01	5.23	5.06	-0.35	4.76	6.01
2	rs647316	31318333	-1.44	-1.10	-1.28	3.29	0.19	-1.11	2.52	2.15	1.13	1.18	0.08	2.23	2.12
12	rs7961894	120849966	-0.62	-0.53	-0.41	16.13	1.01	-0.44	15.35	15.92	14.76	15.48	0.91	15.42	16.46
18	rs1790576	65669823	-1.23	-1.28	-1.42	4.12	-0.57	1.19	3.34	3.90	4.10	4.35	0.36	4.09	5.80
12	rs11065987	110556807	-0.69	-1.54	-0.25	-1.08	2.68	-1.34	1.90	2.83	4.44	4.08	-1.22	3.78	5.22
12	rs11066301	111355755	-0.76	-1.35	-0.28	-1.10	3.14	-1.08	2.37	3.52	4.96	4.60	-1.20	4.30	5.74
6	rs210135	33648670	-0.95	-0.70	-0.59	-1.20	6.45	0.71	5.67	5.34	4.18	4.05	-0.23	5.38	5.50
9	rs385893	4753176	-1.62	1.27	-0.12	-0.89	5.55	-0.61	4.78	4.84	5.02	4.97	-0.09	4.89	6.84
17	rs1007654	35364880	-0.74	-1.37	-1.08	-1.33	-1.54	3.44	2.66	2.29	3.47	3.09	-1.66	2.93	4.25

Table 4.3: Results from our analysis on the 22 loci (23 SNPs) identified in [Soranzo et al., 2009]. Columns 4-16 show Bayes factors on a $\log_{10}$ scale, the first 6 labeled by the phenotype name show univariate Bayes factors for these traits. The next 6 columns show statistics obtained by the five approaches described in Chapter 2 and a Bayes factor of pleiotropy. The last column gives the maximum $\log_{10}$ Bayes factor over all multivariate models, which is also shown in the table by the colouring of the univariate statistics, where a trait is coloured in red if it is considered associated in the model with the largest Bayes factor.

4. Real Data Analysis

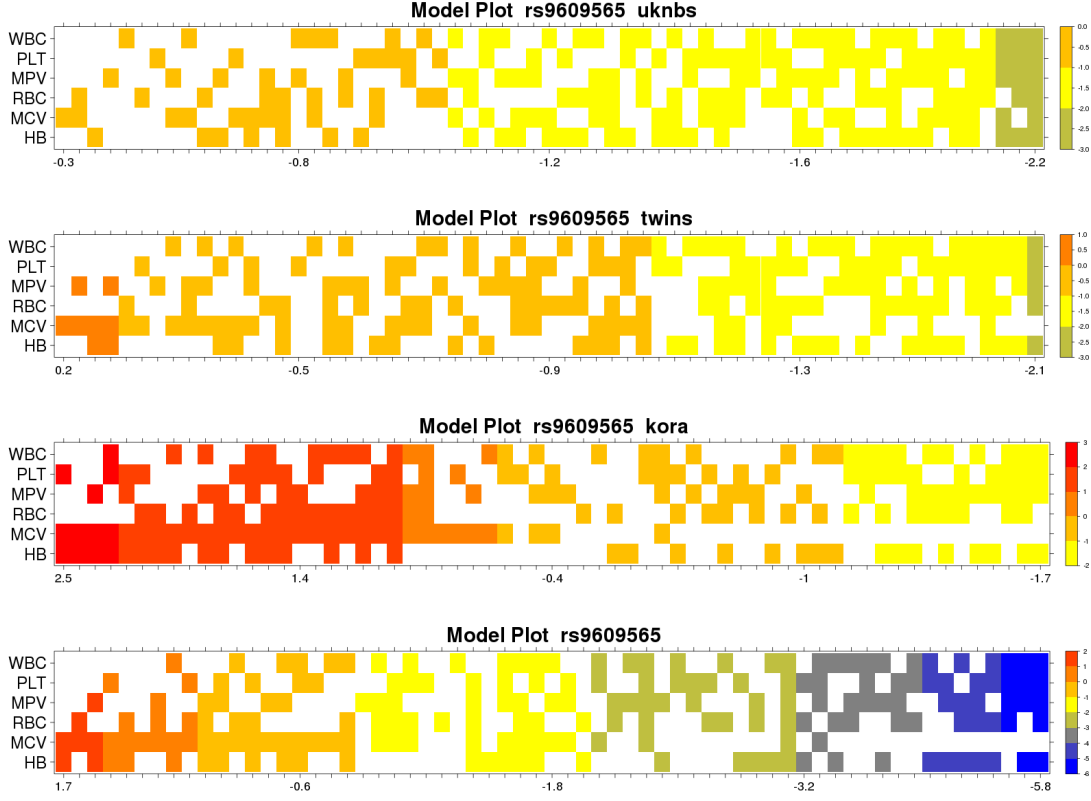


Figure 4.3: Model plots for rs9609565 for the 3 datasets and a combined one. A model plot shows a representation of the individual model Bayes factors (columns in plot) in which models are ordered by the size of their statistic, starting from the largest on the left to the lowest on the right. Each of the models is shown as a combination of white and coloured boxes corresponding to each trait, where the colour stands for association.

Going back to Table 4.3 the Bayes factor of pleiotropy column shows consistent results with the univariate statistics, i.e. when more than one univariate $\log_{10}$ Bayes factor is positive BF_P tends to also be positive and vice versa. Additionally the model with the largest Bayes factor (coloured in red) also to some extent reflects the results from univariate analysis, where no large univariate Bayes factor is excluded from the best multivariate model. Furthermore a number of traits

with negative univariate $\log_{10}$ Bayes factors are also included in the best scoring joint model.

All loci in Table 4.3 have been identified in [Soranzo et al., 2009] using the same three datasets. However, now some of these SNPs do not show any significant univariate statistics. The explanation for this is that our analysis is performed in a different way - by removing missing data for all traits (they only remove the missing data for the particular trait they analyse) and combining results from the three cohorts by multiplying each of the resulting Bayes factors. Additionally, the samples of unrelated individuals which we obtain in the data filtering stage might not be the same as the ones used in [Soranzo et al., 2009]. However, as our purpose is to compare the multivariate and univariate trait approaches, we leave this aside.

The best multivariate model associations found in practice tend to be ones with only one or two traits associated (models in red in Table 4.3). This is most likely the case due to the actual biology behind those associations or the studies still being underpowered to detect all associations (due to smaller effect sizes). The general intuition behind using multivariate models is the aim to get more power to detect associations if there are multiple phenotypes associated to a genetic variant. Despite that, even in models with a single true association, we expect the multivariate approach to be advantageous due to the better noise measurement of the collection of phenotypes and as such potentially better "denoising" of real SNP associations.

In Figure 4.4 we plot BF_A versus BF_S and BF_{AVG} versus BF_U at the 22 loci to highlight differences and similarities between multivariate and univariate approaches. The multivariate BFs (x-axes) are always almost at least as large as the univariate BFs (y-axes). In both plots the Bayes factors of the methods on the y-axes are included in the model averages of the methods on the x-axes, therefore this slight shift is expected due to model averaging over a larger number of models. We can then argue that for the SNPs around the diagonal we do not gain anything by using the more complex approach, but we also do not seem to lose much either. From the plot on the left we can tell whether a variant is pleiotropic or not and from the one on the right we can infer if our multivariate approach (BF_A) outperforms the univariate one (BF_U) (since BF_{AVG} is averaging over the models in BF_A and BF_U). Note that by using BF_{AVG} here instead of directly BF_A , we would never lose much to the univariate approach.

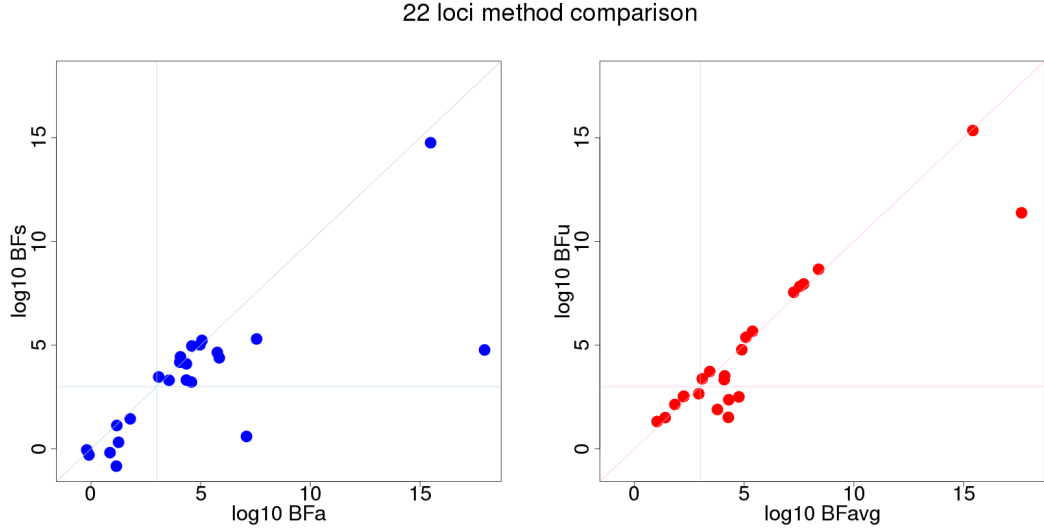


Figure 4.4: Left: Comparison between the multivariate (BF_A) and multivariate on only single trait association (BF_S) methods. Right: Comparison between the multivariate method extension (BF_{AVG}) and the univariate approach (BF_U).

We observe two loci with large pleiotropic effects, corresponding to two SNPs on chromosome 6 (rs1800562 and rs9402686 from Table 4.3). From the plot on the right we find 4 loci for which $BF_{AVG} > 10^3 > BF_U$ (corresponding to 4 of the SNPs in yellow from Table 4.3) and the rs9402686 variant which shows a boost in signal.

Figure 4.5 shows a more detailed view of the performance of different approaches at the HBS1L-MYB1 locus (rs9402686, chr6). For this region we can see a boost in signal in BF_{AVG} compared to univariate analysis, Figure 4.5 (bottom plots) and as well as that a boost from BF_S to BF_A which would suggest that this might be a pleiotropic region. This is expected, as can be seen from Table 4.3, because this SNP shows (univariate) association to more than one of the traits (MCV, RBC, PLT, WBC) and in particular has very strong signal in MCV and RBC. Bayes factor of pleiotropy and model comparison plots are given in Figures 4.6 and 4.7, respectively. For this genetic variant we get $\log_{10}BF_P = 13.39$, which provides strong evidence of it being pleiotropic.

On Figure 4.7, we can see that the model with the highest $\log_{10}BF$ is the model with exactly (MCV, RBC, PLT, WBC) associated to the the genetic variant, the same as we would have expected from univariate phenotype analyses. In addition to the best model, a large number of the following models also identify MCV and RBC as associated (and more than a half of them consider RBC and PLT as well).

4. Real Data Analysis

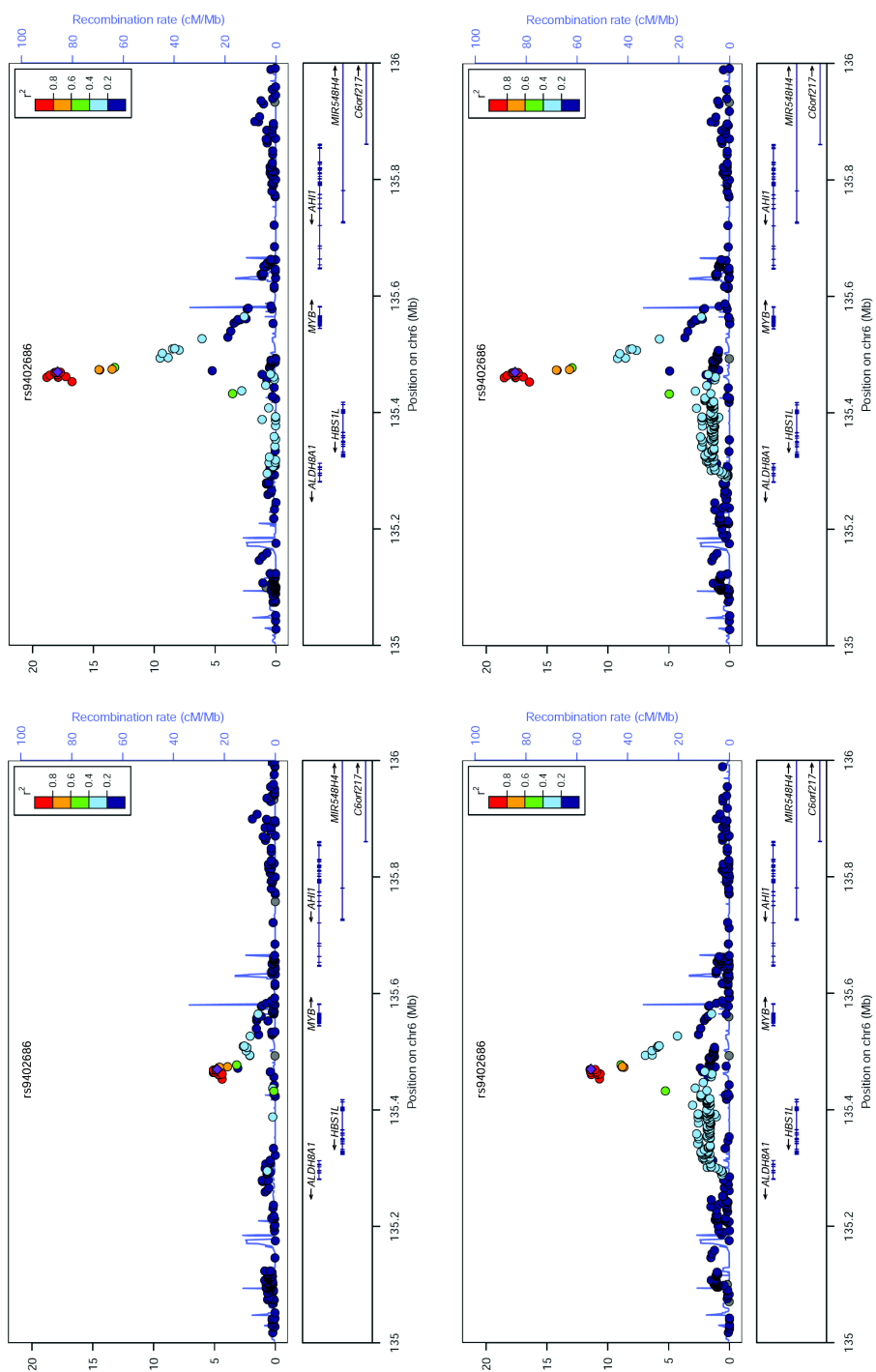


Figure 4.5: Hit plots for the HBS1L-MYB region on chromosome 6, from the analysis of the non-pleiotropic BF_S (top left) and BF_A (top right) multivariate approaches as well as from univariate BF_U (bottom left) and BF_{AVG} (bottom right) approaches. Annotated SNP is rs9402686 and plots done using [Pruim et al., 2010].

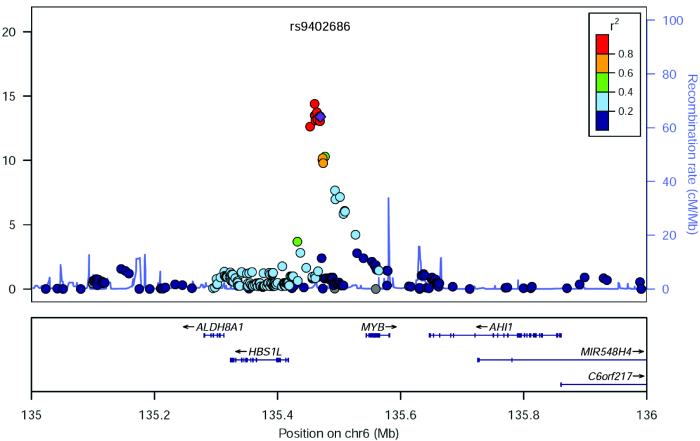


Figure 4.6: HBS1L-MYB log10BF pleiotropy

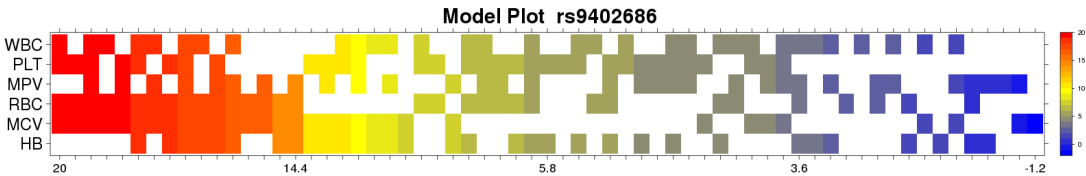


Figure 4.7: Model plot chr6, HBS1L-MYB, rs9402686. Models from top scoring to lowest one (left to right).

4.2.2 Genome-Wide Scans

Method comparison plots from Genome-wide analysis are given in Figure 4.8, from which it can be seen that a number of SNPs exhibit an increase in signal of association in the method allowing for pleiotropic models BF_A compared to the one on single association models (plot on the left). On the right plot a number of genetic variants show a larger BF_{AVG} statistic compared to the univariate BF_U , which are regions for which our multivariate approach (BF_A) would have an advantage over univariate analysis. After removing all SNPs in regions close to ones discovered in [Soranzo et al., 2009] we find a locus of interest.

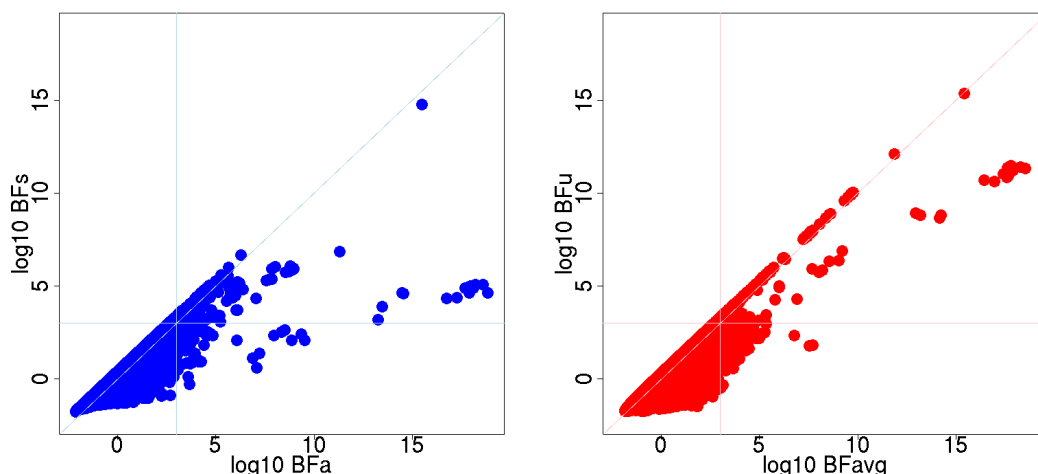


Figure 4.8: Genome-wide log10 BF comparisons.

A SNP in the RCL1 locus on chromosome 9 (rs10974815), which has not been reported in [Soranzo et al., 2009] shows a boost in signal over univariate analysis using our multivariate approach. RCL1 has been replicated in a larger study and found to be independent from another gene nearby AK3, which is identified by one of the 22 loci in [Soranzo et al., 2009].

Signal plots for AK3 and RCL1 are given in Figure 4.9. For AK3 they show similar values of the Bayes factors from our multivariate approach BF_A and the single trait association method BF_S , which suggests that this locus is not pleiotropic. Additionally from the bottom plots again similar values for both BF_U and BF_{AVG} are observed. In contrast, the Bayes factors for RCL1 seem to increase in the latter cases (right plots compared to left), suggesting possible pleiotropy at RCL1. This gives us a reason to believe that these two regions, even though not so far apart might have different functions. It also suggests that our approach can be used for dissecting signals and picking up/classifying differences between them. A similar interpretation can be drawn from the Bayes factors of pleiotropy in Figure 4.10 since AK3 does not appear to be a pleiotropic region as opposed to RCL1.

From the model plots in Figure 4.11 we can indeed see that our method picks up rs10974815 (RCL1) as a pleiotropic locus, (most likely) associated to MCV, RBC and PLT. On the other hand the best model for rs385893 (near AK3) is the same as the one for rs10974815, which is thus again pleiotropic. BF_P however suggests that there is no pleiotropy for AK3. If we look closer at the model plots (Figure 4.11) for both regions we can notice different patterns in the top scoring models. RCL1 favours all models with MCV and RBC associated, and shows weaker evidence for association with PLT. AK3, on the other hand, shows a strong association signal with PLT, with MCV and RBC possibly quite weakly associated.

4. Real Data Analysis

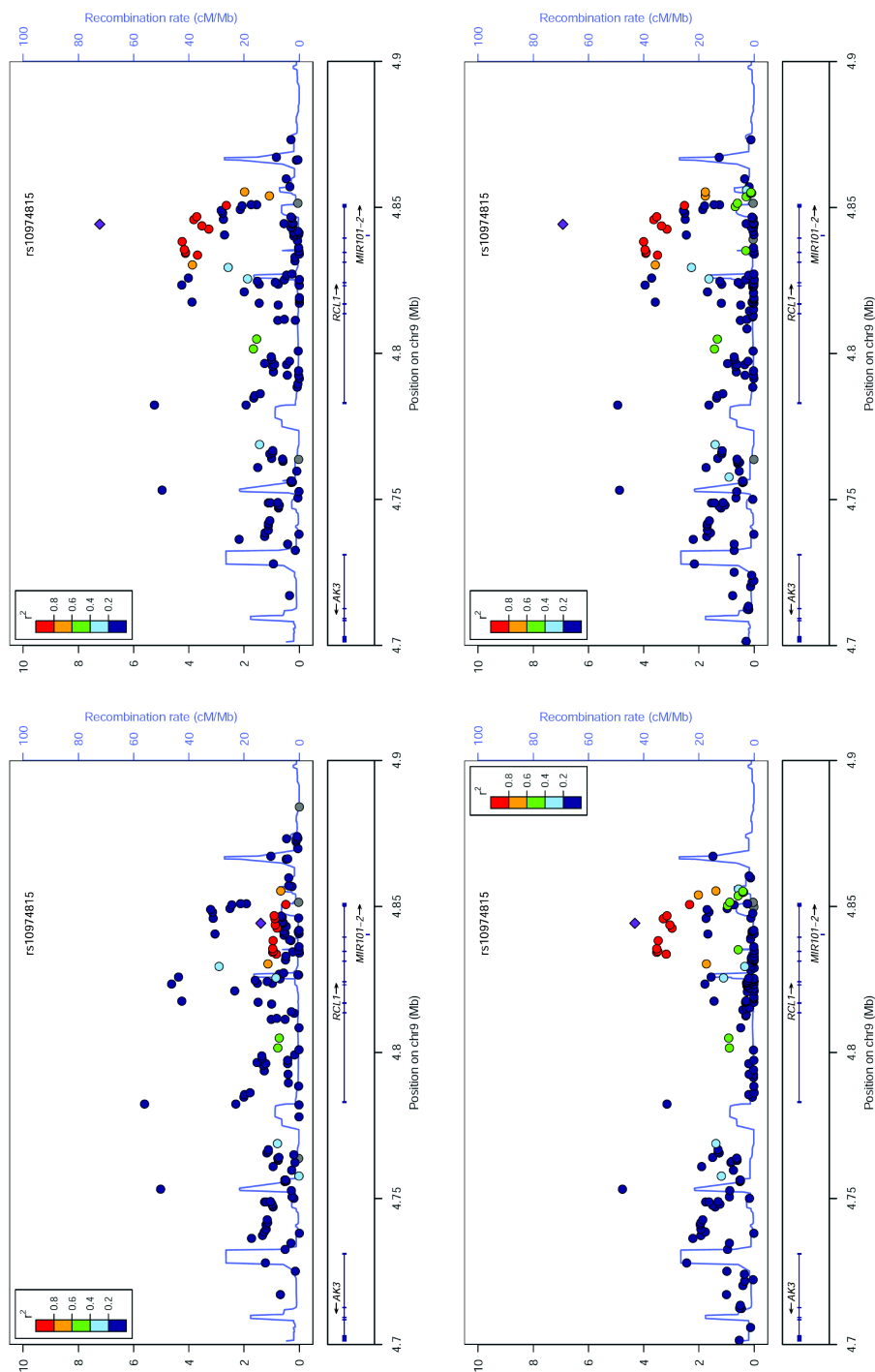


Figure 4.9: Hit plots for the region around RCL1 (annotated) and AK3 (top SNP around 4.75 Mb) : for BF_S (top left), BF_A (top right), BF_U (bottom left) and BF_{AVG} (bottom right) (on a $\log_{10}$ scale)

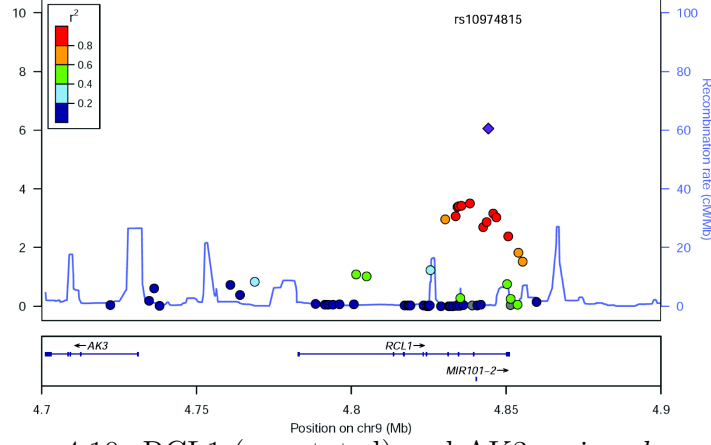


Figure 4.10: RCL1 (annotated) and AK3 region, $\log_{10}BF_P$

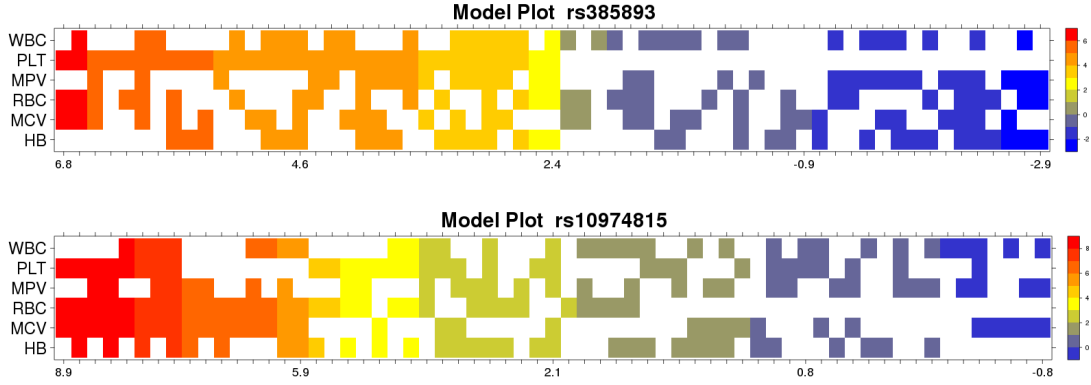


Figure 4.11: Model plots chr9, AK3 (top), RCL1 (bottom). Models from top scoring to lowest one (left to right)

After performing the genome-wide scans, a common variant on chromosome 20 (MAF 0.12–0.15 for the 3 datasets), not previously declared in [Soranzo et al., 2009] showed a boost in signal in the multivariate approach compared to univariate analysis, resulting in $\log_{10}BF_{AVG} = 2.90$ and $\log_{10}BF_U = 0.003$, with univariate $\log_{10}$ Bayes factors 0.52, -0.15 , 0.17, -0.44 , -1.11 and -1.16 . The SNP does not show any evidence of pleiotropy with $\log_{10}BF_P = -0.28$ ($\log_{10}BF_A = 3.20$ and $\log_{10}BF_S = 3.35$). Manhattan plots of the 4 approaches are given in Figure 4.12 and a model in Figure 4.13.

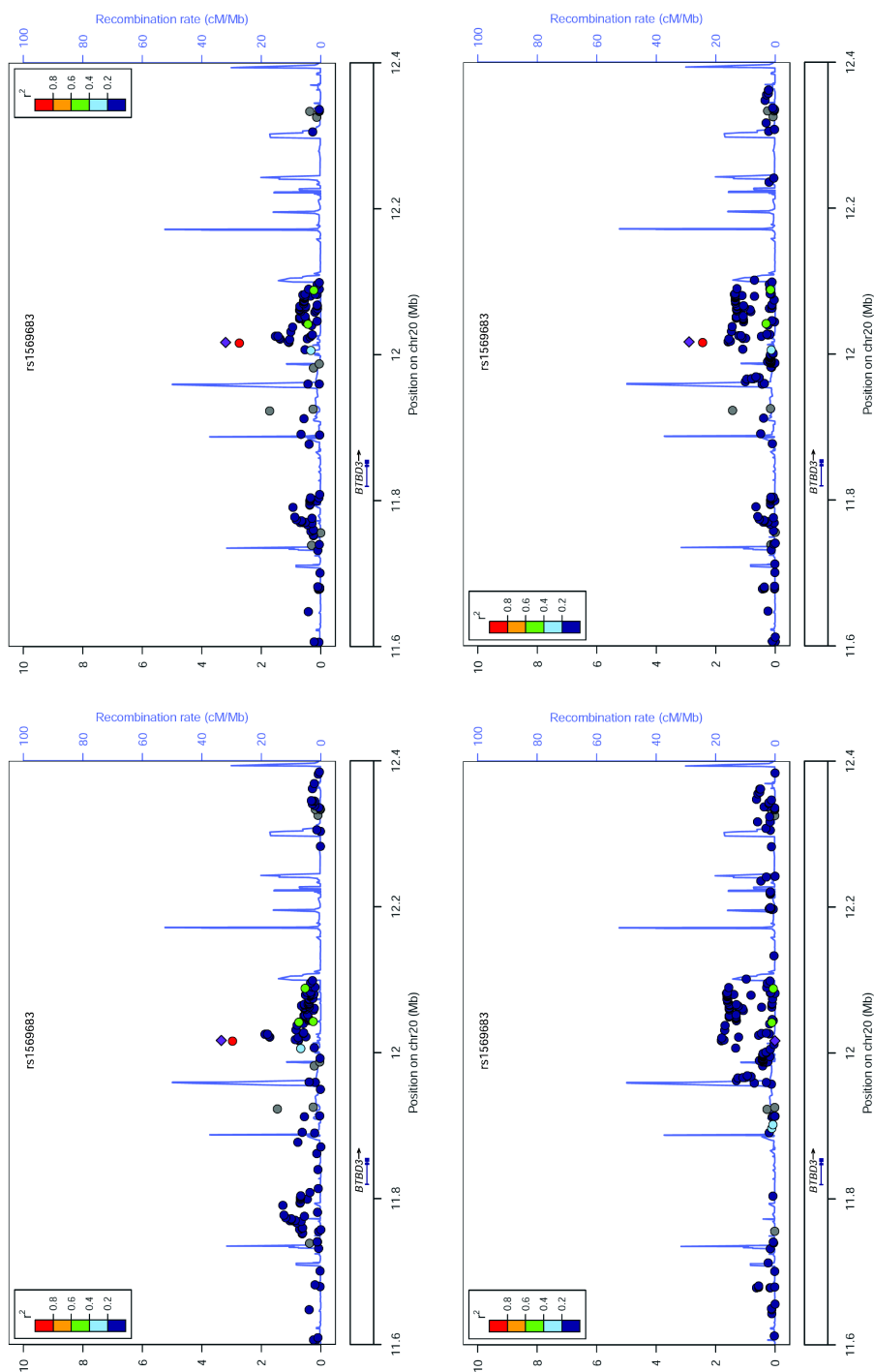


Figure 4.12: Model plot and $\log_{10}BF$ hit plots for rs1569683 on chr20, BF_S (top-left), BF_A (top-right), univariate BF_U (bottom-left) and BF_{AVG} (bottom-right).

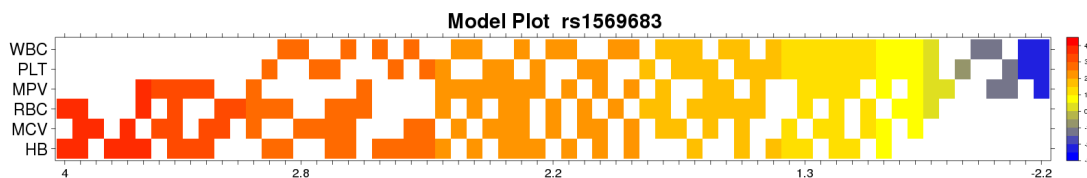


Figure 4.13: Model plot and for rs1569683 on chr20

From the hit plots we can clearly see an advantage of the multivariate approaches to the univariate one and again no signs of pleiotropy appear. The latter agrees with the model plot results, which show the top model (in fact the top 2 models) as multiple trait association models, closely followed by single trait models. We conclude from this analysis that this locus is potentially interesting for further analysis.

4.3 Summary

After applying our multivariate approach to the dataset on blood trait measurements, we found genetic variants that show larger signal of association to one or more of the phenotypes in comparison to the usual univariate approach, e.g. RCL1, which has been replicated in a larger study. Another marker on chr20, which shows no (or very weak) association to any of the traits by analysing them one at a time, reaches a Bayes factor value of $10^{3.2}$ in our multivariate approach. Further investigation of this variant would need to be carried out.

In addition to the ability to detect association of generic markers in our approach, the Bayes factor of pleiotropy and in particular the individual model Bayes factors (with the use of the models plots) provide a way to distinguish

between signals coming from genetic variants with (possibly) different functions (e.g. RCL1 and AK3).

This method does have some limitations. The method can not easily handle missing phenotypic data or relatedness between samples. In addition, the way in which we have combined results across cohorts is rather simplistic. These limitations motivated our work in subsequent chapters of this thesis.

Chapter 5

A sparse Bayesian multivariate mixed model

A limitation of the model averaging (MA) approach proposed in Chapter 2 is that it does not scale well with the number of traits. In this Chapter we propose a method, that aims to address this and other limitations. Our new model allows for correlations between individuals due to related samples, imputes missing phenotypes and models sparse signals of association where just a subset of traits are associated. We use Variational Bayes methods to get an approximate measure of association and employ spike and slab priors [Mitchell and Beauchamp, 1988] on the effect sizes, which allows us to discretize our model space and shrink irrelevant effect size coefficients to zero. This approach is well suited to the situation where a relatively small proportion of large number of traits are associated.

5.1 Model

We assume P phenotypes and a genotype at a single SNP are available for N unrelated individuals, which are stored in an $N \times P$ matrix $\mathbf{Y}$ and an $N \times 1$ matrix $\mathbf{G}$, respectively. Further we assume that the phenotypes $\mathbf{Y}$ and the SNP genotypes $\mathbf{G}$ have been mean centred and that the phenotypes have been scaled to have unit variance. We model the association between the SNP and the phenotypes using a multivariate linear regression model and we use a Spike and Slab (SAS) prior on the effect sizes to allow for a sparse solution. Here we give the details of the model of association, and the details of the prior structure on the precision matrix and SAS prior in following sub-sections.

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{G}\boldsymbol{\alpha}, \mathbf{I}_N, \boldsymbol{\Omega}^{-1}) \quad (5.1)$$

$$\alpha_q \sim \pi N(\alpha_q | 0, \sigma_g^2) + (1 - \pi)\delta(\alpha_q) \quad q \in 1, \dots, P \quad (5.2)$$

$$\sigma_g^2 \sim \Gamma^{-1}(d_1, d_2) \quad (5.3)$$

$$\pi \sim \text{Beta}(b_1, b_2) \quad (5.4)$$

$$\boldsymbol{\Omega} \sim \text{Wi}(\nu, \mathbf{V}) \quad (5.5)$$

where $\boldsymbol{\alpha}$ is $1 \times P$ vector of effect size parameters and $\mathcal{N}(\mathbf{G}\boldsymbol{\alpha}, \mathbf{I}_N, \boldsymbol{\Omega}^{-1})$ denotes the Matrix Normal Distribution [Dawid, 1981] with row and column covariance matrices specified by the $N \times N$ identity matrix $\mathbf{I}_N$ and the $P \times P$ precision matrix $\boldsymbol{\Omega}$, and mean given by the $N \times P$ matrix of zeros. This distribution implies that the errors are normally distributed such that rows are independent, each multivariate normal with mean 0 and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$. In general the Matrix Normal (MN) distribution can model correlation between rows (i.e.

5. A sparse Bayesian multivariate mixed model

individuals) and between columns (i.e. phenotypes). We use the MN model here as it provides the framework for extending the method to handle related individuals, by allowing correlations between rows of $\mathbf{Y}$. Appendix A includes more details of the MN distribution as well as other distributional forms and relevant results.

5.1.1 Prior on precision or inverse error covariance

The model for the precision needs to capture correlation between phenotypes that is not due to shared association via the term $\mathbf{G}\boldsymbol{\alpha}$ in Eqn. 5.1. As in our previous MA method we use the standard inverse Wishart distribution for the prior on the error covariance

$$\boldsymbol{\Sigma} \sim IW_P(\nu, \mathbf{Q}) \tag{5.6}$$

or equivalently, a Wishart distribution for the precision matrix,

$$\boldsymbol{\Omega} \sim Wi_P(\nu, \mathbf{V}) \tag{5.7}$$

where $\mathbf{V} = \mathbf{Q}^{-1}$. In most GWAS the size of the genetic effects are small so that the residual variance of each phenotype will often be quite close to the phenotypic variance. Since we have assumed that all phenotypes have been scaled to unit variance it makes sense to use a prior on the residual variances that is centred on 1.

Setting $\nu = 5 + P$ and $\mathbf{Q} = 4\mathbf{I}_P$ ($\mathbf{V} = 0.25\mathbf{I}_P$) gives $\mathbb{E}(\boldsymbol{\Sigma}) = \mathbf{I}_p$. When $P = 1$ the $\boldsymbol{\Sigma} \sim IW_1(6, 4I_1) = IG(3, 2)$ which is the same prior used in SNPTEST for sin-

5. A sparse Bayesian multivariate mixed model

gle trait Bayesian analysis (see Supplementary Material of [Marchini and Howie, 2010]).

Other prior distributions that assume sparsity of the precision matrix could also be used but we have not explored them in this work. Such priors might be useful if we were interested in inferring conditional independency relationships between the error terms of each phenotype, which is not our main interest here.

5.1.2 Spike and Slab (SAS) prior

The so called spike and slab (SAS) prior that we used has been quite widely adopted in sparse regression modelling literature [Mitchell and Beauchamp, 1988] and allows us to circumvent enumerating over all possible sub-models, as we did using the MA approach in Chapter 2. Our prior models the effect size parameter for the q^{th} trait (α_q) as a mixture of a point mass at 0 (the “spike”) and a $N(0, \sigma_g^2)$ distribution (the “slab”). An illustration is given in Figure 5.1. We use a conjugate Beta prior for the parameter π which controls the probability of a non-zero effect size, and an Inverse Gamma prior on the effect size variance parameter σ_g^2 . This type of prior has been used for single phenotype analysis before [Logsdon et al., 2010, Carbonetto and Stephens, 2012] within a Variational Bayes framework.

Through application of our model on real data we have found that it is important to use a prior on the variance of the slab (σ_g^2) that discourages small values. We have found that if small values of σ_g^2 are given appreciable weight then real effect sizes can sometimes be shrunk excessively towards zero. We have found that setting $d_1 = d_2 = 4.5$ performs well in both simulations and real data. We

5. A sparse Bayesian multivariate mixed model

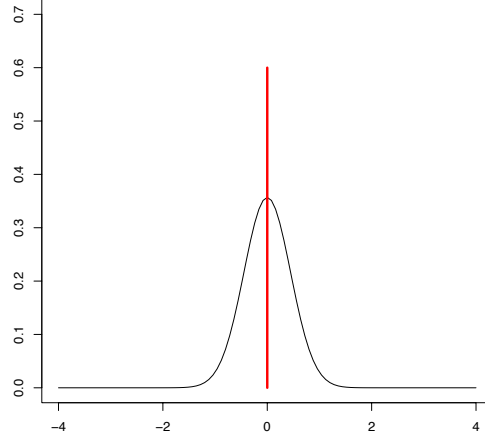


Figure 5.1: Illustration of the Spike and Slab (SAS) prior with a "spike" (technically a point mass) at zero and a normal distribution.

have briefly investigated a couple of other sets of hyper-parameters and results seem to be fairly robust with priors restricting the use of small effect size variance values.

The parameter π in our model represents the proportion of the traits associated to a genetic variant. The hyperparameters (b_1 and b_2) of prior on π should be chosen according to the particular phenotypes being analyzed. The results can be sensitive to choice of b_1 and b_2 and some thought needs to be given on determining our belief on the proportion of traits being associated to a given genetic variant for the data in question.

For computational simplicity a factorisation of the effect size parameters α_q is typically performed for this type of prior. This can be done in multiple ways (which are mathematically equivalent) and following [Titsias and Lazaro-Gredilla, 2011] we use the following reparameterization of the model

5. A sparse Bayesian multivariate mixed model

$$\alpha_q = \tilde{\alpha}_q s_q \quad q \in 1, \dots, P \quad (5.8)$$

$$\tilde{\alpha}_q \sim N(\alpha_q | 0, \sigma_g^2) \quad q \in 1, \dots, P \quad (5.9)$$

$$s_q \sim \text{Bernoulli}(\pi) \quad q \in 1, \dots, P \quad (5.10)$$

which gives the desired spike and slab prior distribution on each α_q . Here s_q is a binary variable that codes for association/no association for trait q and $\tilde{\alpha}_q$ is the effect size of that trait given association. A parameterisation of this form is desirable for practical use in approximate inference methods.

5.1.3 Model parametrization

With the selected priors, we can describe our model with the following log-densities (up to a constant)

$$\begin{aligned} \log P(\mathbf{Y} | \boldsymbol{\alpha}, \boldsymbol{\Omega}) &= \frac{N}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T + \text{const} \\ \log P(\sigma_g^2) &= -(d_1 + 1) \log \sigma_g^2 - d_2 \sigma_g^{2-1} + \text{const} \\ \log P(\mathbf{s}) &= \sum_{q=1}^P s_q \log \pi + \sum_{q=1}^P (1 - s_q) \log(1 - \pi) + \text{const} \\ \log P(\pi) &= (b_1 - 1) \log \pi + (b_2 - 1) \log(1 - \pi) + \text{const} \\ \log P(\tilde{\boldsymbol{\alpha}}) &= -\frac{1}{2} \sum_{q=1}^P \log \sigma_g^2 - \frac{1}{2} \sum_{q=1}^P \tilde{\alpha}_q^2 \sigma_g^{2-1} + \text{const} \\ \log P(\boldsymbol{\Omega}) &= \frac{\nu - P - 1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{V}^{-1} \boldsymbol{\Omega}) + \text{const} \end{aligned}$$

and full set of parameters is given by $\theta = (\boldsymbol{\Omega}, \sigma_g^2, \{\tilde{\alpha}_q\}_{\forall q}, \{s_q\}_{\forall q}, \pi)$

5.1.4 Bayes Factors

We carry out hypothesis testing via calculation of a Bayes factor between two competing models for a set of data $\mathbf{D}$. We use M_0 to denote the Null model of no association ($\boldsymbol{\alpha} = 0$ in model (5.1)) between the SNP and any of the phenotypes and M_1 to denote the model which assumes that there is association present to at least one of the traits. The Bayes factor that compares these two models is the ratio of their marginal likelihoods

$$BF = \frac{P(\mathbf{D}|M_1)}{P(\mathbf{D}|M_0)} \quad (5.11)$$

where $P(\mathbf{D}|M) = \int P(\mathbf{D}|\theta, M)P(\theta|M)d\theta$, θ are the parameters of the model and $P(\theta|M)$ is their prior for that specific model.

5.1.5 Variational Bayesian Inference

The integral required to calculate the marginal likelihood of our model ($P(\mathbf{D}|M_1)$) is not tractable, so either an approximation or sampling methods are needed. As we wish to apply this method at millions of SNPs across the genome the need for a computationally efficient and numerically stable method is important. To do this we employ Variational Bayesian (VB) methods [Jaakkola, 1997, Bishop and Tipping, 2000, Jaakkola and Jordan, 2000]. The idea of VB is to find an approximation to a given posterior distribution $P(\theta|\mathbf{D}, M) \approx Q(\theta)$ where θ is the vector of parameters of the model M . The true log marginal likelihood can be written as

$$\log P(\mathbf{D}|M) = F(Q(\theta)) + D_{KL}(Q(\theta)||P(\theta|\mathbf{D}, M)) \quad (5.12)$$

5. A sparse Bayesian multivariate mixed model

where $F(Q(\theta)) = \int \log \left[\frac{P(\theta, \mathbf{D}|M)}{Q(\theta)} \right] Q(\theta) d\theta$ and $D_{KL}(Q(\theta) || P(\theta | \mathbf{D}, M))$ is the Kullback-Liebler (KL) divergence (see Appendix E). Minimising the KL divergence (which is bounded below by 0) is equivalent to maximising $F(Q(\theta))$ with respect to $Q(\theta)$ and by doing so $F(Q(\theta))$ is a lower-bound approximation to $\log P(\mathbf{D}|M)$.

The VB method requires us to specify a form of the approximate posterior $Q(\theta)$ in order to find its optimal distributions and parameters. The most widely used approximation is known as the mean field approximation and assumes only that the approximate posterior has a factorized form $Q(\theta) = \prod_i Q(\theta_i)$. The VB algorithm maximises $F(Q(\theta))$, by iteratively updating each $Q(\theta_i)$ in a coordinate ascent fashion. The update for $Q(\theta_i)$ that maximises the functional $F(Q(\theta))$ (derived in Appendix F) is given by

$$\log Q(\theta_i) = \mathbb{E}_{Q(\theta_{-i})} \left(\log P(\mathbf{D}, \theta | M) \right) + \text{const} \quad (5.13)$$

which involves taking the expectation of the log-likelihoods with respect to variational posterior distributions of all parameters apart from $Q(\theta_i)$. The precise form of the distribution $Q(\theta_i)$ will depend upon the form of the likelihood and the priors. The applicability of VB often depends upon whether $Q(\theta_i)$ reduces to a tractable parametric form and also whether the various expectations needed when calculating $Q(\theta_i)$ are tractable. The VB algorithm cycles through updates for each $Q(\theta_i)$, at each step increasing $F(Q(\theta))$ until convergence is reached. One key feature of this mean field approximation is that it is not necessary to specify the precise form of the approximate posterior in advance and this is determined by the factorisation assumption.

5.1.6 VB approximation

In our approach, we use a slightly relaxed form of the mean-field approximation, where we assume that the posterior distributions Q of the model parameters Θ factorize in the following way

$$Q(\theta) = Q(\Omega)Q(\pi)Q(\sigma_g^2) \prod_{q=1}^P Q(\tilde{\alpha}_q | s_q)Q(s_q) \quad (5.14)$$

The relaxation, in terms of posterior dependence of the effect sizes on the state of association (or not), allows for discreteness in our model and more meaningful inference on the distributions of effect sizes a posteriori.

The marginal likelihood for the null model ($P(\mathbf{D}|M_0)$) using VB reduces to the same analytic expression as in the model averaging method from Chapter 2. To fit the alternative model M_1 we initialize the parameters (Section 5.1.8) and iterate the VB updates (Section 5.1.7) until the variational bound $F(Q(\theta))$ converges within a tolerance of 10^{-3} . We then use the bound $F(Q(\theta))$ as the approximation of $\log P(\mathbf{D}|M_1)$ and use this together with the marginal likelihood from the null model to calculate an approximate Bayes Factor. Since our model uses VB to fit our model that involves a SAS prior structure, we refer to our approach as VB SAS.

5.1.7 VB SAS updates

With our version of the mean field approximation, all parameter posterior distributions remain in the same family of distributions, which makes this VB approach feasible. We have derived the VB updates for all of the parameters,

5. A sparse Bayesian multivariate mixed model

all the required moment calculations and the expression for the bound on the marginal likelihood and these are given below. Full specification and derivation of the updates and the approximate log-marginal likelihood of the association model M_1 are given in Appendix F.

Approximate posterior distribution for Ω :

$$\begin{aligned} Q(\Omega) &= Wi(\nu^*, \mathbf{V}^*) \\ \nu^* &= \nu + N \\ \mathbf{V}^* &= [\mathbf{V}^{-1} + \mathbb{E}[(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})]]^{-1} \end{aligned} \tag{5.15}$$

Approximate posterior distribution for σ_g^2

$$\begin{aligned} Q(\sigma_g^2) &= \Gamma^{-1}(d_1^*, d_2^*) \\ d_1^* &= d_1 + \frac{P}{2} \\ d_2^* &= d_2 + \frac{1}{2} \sum_{q=1}^P \mathbb{E}[\tilde{\alpha}_q^2] \end{aligned} \tag{5.16}$$

Approximate posterior distribution for π

$$\begin{aligned} Q(\pi) &= B(b_1^*, b_2^*) \\ b_1^* &= b_1 + \sum_{q=1}^P \mathbb{E}(s_q) \\ b_2^* &= b_2 + \sum_{q=1}^P (1 - \mathbb{E}(s_q)) \end{aligned} \tag{5.17}$$

5. A sparse Bayesian multivariate mixed model

Approximate posterior distribution for $\{\tilde{\alpha}_q, s_q\}$

$$Q(s_q) = \text{Bernoulli}(\pi_q^*) \quad (5.18)$$

$$Q(\tilde{\alpha}_q | s_q) = N\left(s_q m_q, s_q v_q + (1 - s_q) \mathbb{E}(\sigma_g^{2^{-1}})^{-1}\right) \quad (5.19)$$

$$\pi_q^* = \frac{1}{1 + e^{-\rho_q}}$$

$$\rho_q = \mathbb{E}\left(\log \frac{\pi}{1 - \pi}\right) + \frac{m_q^2}{2v_q} + \frac{1}{2} \log v_q + \frac{1}{2} \log \left[\mathbb{E}(\sigma_g^{2^{-1}})\right]$$

$$v_q = \left[\mathbf{G}^T \mathbf{G} \mathbb{E}(\boldsymbol{\Omega})_{q,q} + \mathbb{E}(\sigma_g^{2^{-1}})\right]^{-1}$$

$$m_q = v_q \mathbf{G}^T \left[\mathbb{E}(\boldsymbol{\Omega})_{q,q} \mathbf{Y}_q + \sum_{i \neq q} \mathbb{E}(\boldsymbol{\Omega})_{q,i} (\mathbf{Y}_i - \mathbf{G} \mathbb{E}(\alpha_i)) \right]$$

Required moments

$$\mathbb{E}(\tilde{\alpha}_q^2) = \pi_q^*(v_q + m_q^2) + (1 - \pi_q^*) \frac{d_2^*}{d_1^*}$$

$$\mathbb{E}(s_q) = \pi_q^*$$

$$\mathbb{E}(\sigma_g^{2^{-1}}) = \frac{d_1^*}{d_2^*}$$

$$\mathbb{E}(\boldsymbol{\Omega}) = \boldsymbol{\nu}^* \mathbf{V}^*$$

$$\mathbb{E}(\alpha_q) = m_q \pi_q^*$$

$$\mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha})_{qr} = \mathbb{E}(\alpha_q \alpha_r)$$

$$\mathbb{E}(\alpha_q \alpha_q) = \pi_q^*(v_q + m_q^2)$$

$$\mathbb{E}(\alpha_q \alpha_r) = \pi_q^* \pi_r m_q m_r \quad q \neq r$$

$$\mathbb{E}(\tilde{\alpha}_q^2) = \pi_q^*(v_q + m_q^2) + (1 - \pi_q^*) \frac{d_2^*}{d_1^*}$$

$$\mathbb{E}(\log(\pi)) = \psi(b_1^*) - \psi(b_1^* + b_2^*)$$

$$\mathbb{E}(\log(1 - \pi)) = \psi(b_2^*) - \psi(b_1^* + b_2^*)$$

where ψ denotes the digamma function

Approximate marginal likelihood

$$\begin{aligned}
F(Q(\theta)) &= \int \log \left[\frac{P(\theta, \mathbf{D}|M)}{Q(\theta)} \right] Q(\theta) d\theta \\
&= \mathbb{E}[\log P(\mathbf{D}, \theta) - \log Q(\theta)] \\
&= \frac{1}{2} \text{tr} \left[(\mathbf{V}^{*-1} - \mathbf{V}^{-1} - \mathbf{Y}^T \mathbf{Y} + \mathbf{Y}^T \mathbf{G} \boldsymbol{\gamma} + \boldsymbol{\gamma}^T \mathbf{G}^T \mathbf{Y} - \mathbf{G}^T \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha})) \boldsymbol{\nu}^* \mathbf{V}^* \right] \\
&\quad + \left(d_2^* - d_2 - \frac{1}{2} \sum_{q=1}^P \mathbb{E}(\tilde{\alpha}_q^2) \right) \frac{d_1^*}{d_2^*} + d_1 \log d_2 - d_1^* \log d_2^* - \log \Gamma(d_1) \\
&\quad + \log \Gamma(d_1^*) + \frac{P}{2} - \frac{\nu}{2} \log |\mathbf{V}| + \frac{\nu^*}{2} \log |\mathbf{V}^*| - \sum_{q=1}^P (1 - \pi_q^*) \log(1 - \pi_q^*) \\
&\quad - \log \Gamma_P\left(\frac{\nu}{2}\right) + \log \Gamma_P\left(\frac{\nu^*}{2}\right) - \log B(b_1, b_2) + \log B(b_1^*, b_2^*) - \sum_{q=1}^P \pi_q^* \log \pi_q^* \\
&\quad + \frac{1}{2} \sum_{q=1}^P \left(\pi_q^* \log v_q - (1 - \pi_q^*) \frac{d_1^*}{d_2^*} \right) + (b_1 - b_1^* + \sum_{q=1}^P \pi_q^*) \mathbb{E}(\log \pi) \\
&\quad + (b_2 - b_2^* + \sum_{q=1}^P (1 - \pi_q^*)) \mathbb{E}(\log(1 - \pi)) - \frac{NP}{2} \log(\pi)
\end{aligned}$$

where $\gamma = \mathbb{E}(\alpha_q)$

5.1.8 Method initialization

VB approaches iterate over all model parameters and update their approximate posterior distributions. At each step this results in an increase of the log-marginal likelihood. However before they can be updated, these parameters need to be initialized. Two main points need to be considered here. First, the number of required iterations until convergence would depend on the VB starting points.

5. A sparse Bayesian multivariate mixed model

Therefore initializing the algorithm as close as possible to the true posterior would be desirable. Second, and perhaps more important, the (approximate) marginal likelihood need not be unimodal and VB is only guaranteed to converge to a local maximum (and not a global one), stressing even further the importance of the parameter initialization.

We initialize the model by setting $\pi_q^* = 1 \forall q$. We then set m_q and v_q in (5.18) and (5.19) equal to the marginal posterior mean and variance from fitting a “full” model to all traits with the form

$$\begin{aligned} \mathbf{Y} &= \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \mathbf{I}, \boldsymbol{\Sigma}) \\ \boldsymbol{\alpha} &\sim N(0, v\boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} &\sim IW(c, \mathbf{Q}) \end{aligned} \tag{5.20}$$

This model is a multivariate linear regression model with fully conjugate Normal Inverse Wishart prior structure so parameter estimates are analytic and fast to compute. The remaining parameters $\{\nu, \mathbf{V}^*, d_1^*, d_2^*, b_1^*, b_2^*\}$ can all be initialized via first application of their VB updates.

5.1.9 Model selection

Once a signal of association has been detected inference shifts to inference of which phenotypes are associated. The implementation of the VB-SAS method involves specification of a latent variable s_q for each trait that represents the probability of association (see Appendix F). Model selection can occur by examining

the components of the VB approximation $\pi_q^* = Q(s_q = 1)$ and inferring that the q th trait is associated if π_q^* is above a threshold.

5.2 Extension to allow related samples

We have extended our model to handle multiple correlated traits and correlated individuals in a way that allows VB inference. To do this we introduce an unobserved random effect matrix $\mathbf{S}$ of dimension $N \times M$ such that

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (5.21)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_N, \boldsymbol{\Omega}^{-1}) \quad (5.22)$$

$$\mathbf{S} \sim N(0, 2\mathbf{K}, \mathbf{I}_M) \quad (5.23)$$

Here $\boldsymbol{\beta}$ is a $M \times P$ matrix of mixing coefficients that induce correlations between columns of the term $\mathbf{S}\boldsymbol{\beta}$ and $\mathbf{D} = \boldsymbol{\Omega}^{-1}$ is an error covariance matrix. Under this model, conditional on $\boldsymbol{\beta}$ we have $\mathbf{S}\boldsymbol{\beta} \sim \mathcal{N}(0, 2\mathbf{K}, \mathbf{B})$, where $\mathbf{K}$ is the kinship matrix and $\mathbf{B} = \boldsymbol{\beta}^T \boldsymbol{\beta}$ and the trait covariances are given by

$$\text{Cov}(Y_{ik}, Y_{jl}) = 2K_{ij}B_{kl} + D_{kl}\delta_{ij} \quad (5.24)$$

In this model the covariance within an individual between traits k and l , ($k \neq l$),

$$\text{Cov}(Y_{ik}, Y_{il}) = B_{kl} + D_{kl} \quad (5.25)$$

When all individuals can be assumed to be unrelated, so that $2\mathbf{K} = \mathbf{I}_N$, then the model reduces to model (Eqn. 5.1) with $\boldsymbol{\Sigma} = \mathbf{B} + \mathbf{D}$.

5. A sparse Bayesian multivariate mixed model

With regard to standard linear mixed models Eqn. 5.21 aims to model

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{G}_u\mathbf{u} + \boldsymbol{\epsilon} \quad (5.26)$$

where $\mathbf{G}_u$ is an N by R matrix of R (unknown) causal loci and $\mathbf{u}$ is the matrix of their loadings for each of the phenotypes. Moreover, for univariate response linear mixed models (P=1), this translates to

$$\mathbf{Y}_q = \mathbf{G}\boldsymbol{\alpha}_q + \mathbf{v} + \boldsymbol{\epsilon}_q \quad (5.27)$$

$$\mathbf{v} = N(0, \sigma_g^2 \mathbf{K}) \quad (5.28)$$

$$\boldsymbol{\epsilon}_q = N(0, \sigma_e^2 \mathbf{I}) \quad (5.29)$$

where the kinship matrix $\mathbf{K}$ is an estimate of the unknown $\frac{1}{R}\mathbf{G}_u\mathbf{G}_u^T$ and σ_g^2 and σ_e^2 are the genetic and non-genetic components of the trait variance, respectively. Phenotypic narrow sense heritability h^2 can be derived from those as $\frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. When generalizing this model to P phenotypes, we allow for a causal variant to be unique to a trait or to be shared between traits, in the latter case possibly with different effects. This in turn can create various correlations between the phenotypes attributable to the genetic effects which is what the $\mathbf{B}$ covariance matrix is there to model. Generally $\mathbf{B}$ and $\mathbf{D}$ correspond to σ_g^2 and σ_e^2 in the multivariate linear mixed effects model.

In our proposed model, as before, we use a Wishart prior on the trait non-genetic precision parameters

$$\boldsymbol{\Omega} = \mathbf{D}^{-1} \sim Wi(e, \mathbf{E}) \quad (5.30)$$

5. A sparse Bayesian multivariate mixed model

using $e = 5 + P$, $\mathbf{E} = 0.25 \mathbf{I}_P$ by default so that $\mathbb{E}(\sigma^2) = V(\sigma^2) = 1$ for $P = 1$. We use a Normal prior on the elements of the mixing matrix

$$\beta_{mq} \sim N(0, \lambda^2) \quad (5.31)$$

which implies $\mathbf{B} \sim W_P(M, \lambda^2 \mathbf{I}_P)$. We discuss the choice of M in a following subsection. Conditional on $\mathbf{B}$ and $\mathbf{D}$ the heritability of each trait is defined as $h_q^2 = B_{qq}/(B_{qq} + D_{qq})$. Also, $B_{qq} \sim \lambda^2 \chi_M^2$ and $D_{qq} \sim IG(3, 2)$. Setting $\lambda^2 = 1/M$ results in marginal expected heritability of 0.5. Alternatively $\lambda^2 = 1/(2M)$ gives a prior on h^2 that has good coverage of plausible values of the heritability and seems rather insensitive to the choice of M (Figure 5.2).

We have developed a VB algorithm for this model and full detail of the updates and marginal likelihood calculation are available in Appendix G. In the case of correlated individuals the marginal likelihood calculation for the null model also needs to be calculated using the VB method, as VB does not reduce to an analytic expression as it did when we considered unrelated samples.

5.2.1 Pre-whitening

An important computational step prior to employing the VB framework is data transformation or pre-whitening, just as in univariate mixed model approaches (e.g. [Lippert et al., 2011]). It requires a spectral decomposition of twice the kinship matrix $2\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, after which the phenotype data and each of the genotypes need to be premultiplied by $\mathbf{U}^T$. We refer to this as the pre-whitening step and what it creates is a much simpler model to work with where now the new

5. A sparse Bayesian multivariate mixed model

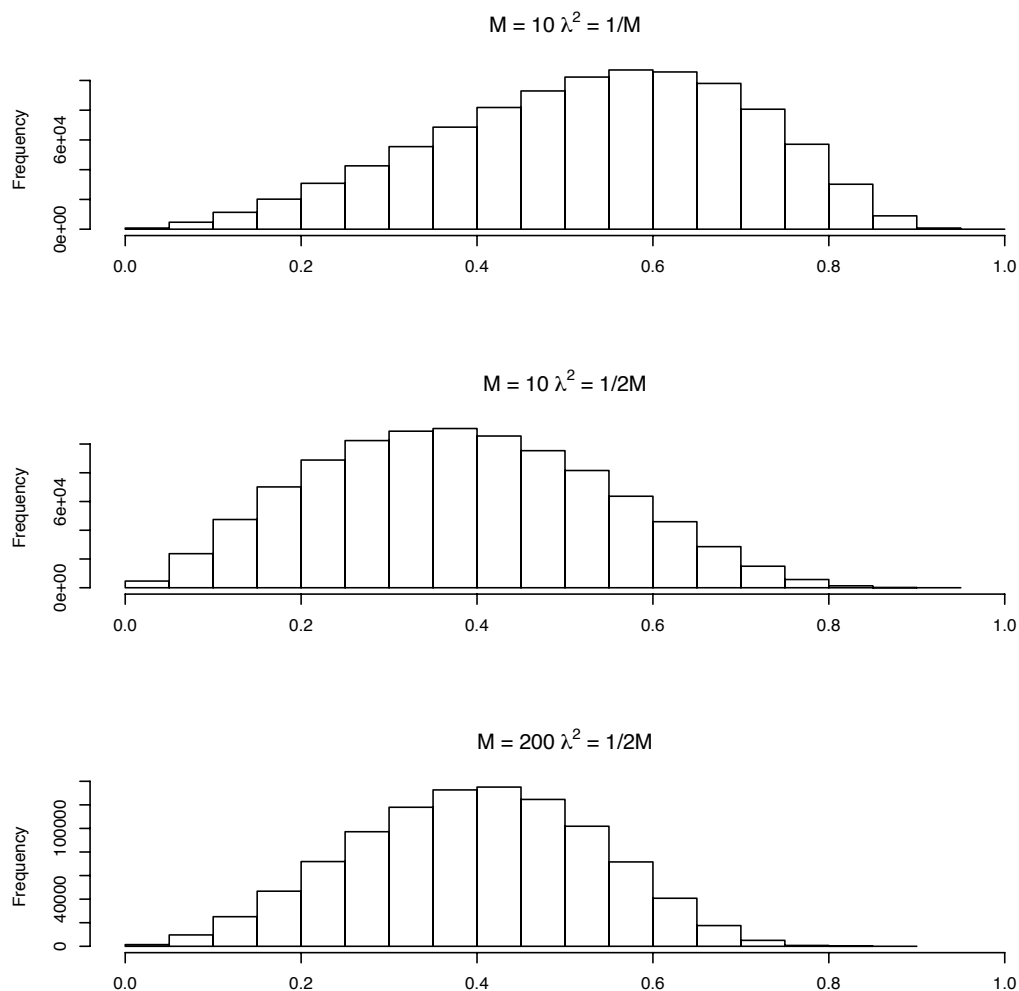


Figure 5.2: The implied prior on heritability h^2 by assuming $d_q^2 \sim \Gamma^{-1}(3, 2)$ and $\beta_{mq} \sim N(0, \lambda^2)$. We investigated three combinations (top) $M = 10$ and $\lambda^2 = 1/M$, (middle) $M = 10$ and $\lambda^2 = 1/(2M)$ and (bottom) $M = 1 = 200$ and $\lambda^2 = 1/(2M)$.

transformed data follows the following model, where without loss of generality

5. A sparse Bayesian multivariate mixed model

assume $\mathbf{Y}$, $\mathbf{G}$ and $\mathbf{S}$ refer to their pre-whitened values from now on.

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (5.32)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_N, \boldsymbol{\Omega}^{-1}) \quad (5.33)$$

$$\mathbf{S} \sim \mathcal{N}(0, \boldsymbol{\Lambda}, \mathbf{I}_M) \quad (5.34)$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix.

5.2.2 VB Approximation and initialization

Now with the additional parameters in the model, we assume the following factorisation of the posterior distributions of the parameters $\boldsymbol{\theta}$

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\Omega})Q(\boldsymbol{\pi})Q(\sigma_g^2) \prod_{q=1}^P Q(\tilde{\alpha}_q | s_q)Q(s_q) \prod_{m=1}^M Q(\mathbf{S}_m) \prod_{m=1}^M Q(\boldsymbol{\beta}_m) \quad (5.35)$$

where $\mathbf{S}_m$ refers to the m 'th column of the $\mathbf{S}$ matrix and $\boldsymbol{\beta}_m$ to the m 'th row.

Initialization of the additional two parameters in this approach is done as follows. We initialise the posterior mean of the $\mathbf{S}_m$ variables at zero and variance $\boldsymbol{\Lambda}$. Then the posterior means of the diagonal elements of $\boldsymbol{\beta}$ are set to 0.5 and all other entries are set to zero.

Under the null model, all other parameters can be initialised using the VB updates. For the alternative model, it is computationally faster if we start the approach for $\mathbf{S}$ and $\boldsymbol{\beta}$ from posterior parameter estimates from the null model, and set all other parameters as for the case with unrelated samples.

5.2.3 Choice of M

The parameter M represents the number of random effects in our model. Our belief about it can range from zero, when we have unrelated data, to P , which represents the case where we would need a distinct random effect for each trait. Within the Variational Bayes framework we need to update each parameter in the model at each iteration until convergence, which will be computationally faster with smaller number of parameters to loop over. This motivates our desire to choose M to be as small as possible, but at the same time large enough to capture correlation between traits and samples.

In our Bayesian formulation of the model, our interest lies in discovering association between the SNP of interest and the set (or subset) of traits under study. Therefore we want to use the same M for both the null and alternative models in order to test for a SNP effect. We have investigated a strategy in which we choose M under the null model, which would only require us to do this selection once for each dataset. We use the Deviance information criterion, or DIC, to choose M . DIC is a Bayesian measure of fit (or adequacy), penalized by a term for model complexity [Spiegelhalter et al., 2002]. DIC has been used in genetics for hierarchical models of gene-environment interactions [Fikse et al., 2003, Rekaya et al., 2003], genetic evaluations in animal breeding [Legarra A. et al., 2005] and multiple QTL mapping [Shriner and Yi, 2009]. Other ways of choosing the number of random effects required might be possible, but will not be considered throughout this work.

5. A sparse Bayesian multivariate mixed model

DIC has the following form:

$$DIC = \bar{D}(\Theta) + p_D \quad (5.36)$$

where

$$D(\Theta) = -2 \log \mathbb{P}(Y|\Theta) + \text{const} \quad (5.37)$$

is the deviance statistic. The posterior mean deviance $\bar{D}(\Theta)$ is suggested as a Bayesian measure of fit or adequacy [Spiegelhalter et al., 2002] and p_D is a penalty for model complexity of the form

$$p_D = \bar{D}(\Theta) - D(\bar{\Theta}) \quad (5.38)$$

where $\bar{D}(\Theta) = \mathbb{E}_{\mathbb{P}(\Theta|Y)} D(\Theta)$ and $D(\bar{\Theta}) = D(\mathbb{E}_{\mathbb{P}(\Theta|Y)} \Theta)$.

In our method, the log-likelihood under the null is

$$\log \mathbb{P}(\mathbf{Y}|\Theta) = -\frac{1}{2} \text{tr}(\mathbf{\Omega}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta})) + \frac{N}{2} \log |\mathbf{\Omega}| - \frac{NP}{2} \log(2\pi) \quad (5.39)$$

and under the factorization of the posterior distributions that we are assuming in VB, we can calculate the complexity term as

$$p_D = \text{tr}(\mathbb{E} \mathbf{\Omega}(\mathbb{E}(\boldsymbol{\beta}^T \mathbf{S}^T \mathbf{S} \boldsymbol{\beta}) - \mathbb{E}(\boldsymbol{\beta})^T \mathbb{E}(\mathbf{S})^T \mathbb{E}(\mathbf{S}) \mathbb{E}(\boldsymbol{\beta}))) - N \mathbb{E} \log |\mathbf{\Omega}| + N \log |\mathbb{E}(\mathbf{\Omega})| \quad (5.40)$$

where $\mathbb{E}(\log |\mathbf{\Omega}|) = \sum_{i=1}^P \psi(\frac{e^* + 1 - i}{2}) + P \log 2 + \log |\mathbf{E}^*|$ and ψ is the digamma function.

5.2.4 Modelling missing phenotype data

In GWAS that analyse one phenotype at a time, any samples with missing phenotype data are usually excluded from the analysis. When jointly analysing multiple traits it will usually be the case that the pattern of missingness differs between traits. The majority of available approaches require that there is no missing phenotype data. The simplest approach of analysing only samples with no missing data across all phenotypes can significantly decrease sample size compared to approaches that analyse one phenotype at a time. Our VB methods are able to use all samples irrespective of the pattern of missing data. Missing phenotype values are modelled at each iteration of the VB algorithm under the null model as random variables and their posterior distribution is in turn updated.

We have considered simpler approaches to imputing missing data before the analysis. These include k-nearest neighbour (knn) imputation using Euclidean distance (kkn package in R) and a version of knn using Mahalanobis distance. The first method uses Euclidean distance to measure the distance between points and therefore ignores any possible correlation between traits. Hence we think that Mahalanobis distance would make more use of the structure of the data in general and in addition would reduce to Euclidean distance if the traits are uncorrelated. Our imputation method, however uses our null model in order to impute the missing data, which takes into account (and models) phenotype correlation, but in addition it also takes into account (and models) relatedness between samples, which is definitely more advantageous.

5. A sparse Bayesian multivariate mixed model

Missing data can be handled within the Bayesian framework. We assume that the missing phenotypes are parameters of the model and obtain a VB approximation to the posterior distribution on these parameters in the same way as other parameters. We can view the matrix normal likelihood of our model (if there were no missing phenotype values) as a likelihood for the non-missing data plus a prior on the missing data. Suppose that we have already pre-whitened our data and $\mathbf{G} := \mathbf{U}^T \mathbf{G}^{(0)}$, $\mathbf{Y} := \mathbf{U}^T \mathbf{Y}^{(0)}$, where $\mathbf{Y}^{(0)}$ and $\mathbf{G}^{(0)}$ are our original data. Then we can find the approximate posterior distribution of the missing phenotypes via the following VB update equation, when assuming that the posterior of the missing data is independent of the posteriors of all other model parameters

$$\begin{aligned} \log Q(\text{missing } \mathbf{Y}_i^{(0)}) &= \mathbb{E}_{Q(\theta \setminus \mathbf{Y})} (\log P(\mathbf{D}, \theta)) + \text{const} \\ &= -\frac{1}{2} \sum_{i=1}^N \mathbb{E} (\mathbf{Y}_i^{(0)} - \mathbf{U}_i \mathbf{S} \boldsymbol{\beta} - \mathbf{U}_i \mathbf{G} \boldsymbol{\alpha}) \boldsymbol{\Omega} (\mathbf{Y}_i^{(0)} - \mathbf{U}_i \mathbf{S} \boldsymbol{\beta} - \mathbf{U}_i \mathbf{G} \boldsymbol{\alpha})^T \\ &\quad + \text{const} \end{aligned} \tag{5.41}$$

Then each row of the phenotype data would follow a normal distribution with mean $H = \mathbf{U}_i \mathbb{E}(\mathbf{S} \boldsymbol{\beta} - \mathbf{G} \boldsymbol{\alpha})$ and variance $F = \mathbb{E}(\boldsymbol{\Omega})$. However here we are interested in only updating the missing values, therefore we require the conditional distributions of the missing values, given the non-missing ones.

We can use the properties about conditional distributions of the multivariate normal distribution (as in Appendix A) to derive a multivariate normal distribution for each row of the missing phenotypes. Then all $\mathbf{Y}^T \mathbf{Y}$ terms would need to become $\mathbb{E}(\mathbf{Y}^T \mathbf{Y})$ for the data imputation run of the model.

5.2.5 Algorithm

The full set of steps needed to fit our VB SAS model to related samples consists of the following 6 steps.

1. **Pre-whiten data** required for computational purposes only. The complexity of this step is $O(N^3)$ and only needs to be done once, after which the cost of the remaining operations is reduced to $O(N)$ in terms of number of samples N (just as in univariate mixed models like [Lippert et al., 2011]), $O(M)$ for the number of random effects and $O(P^3)$ for the number of phenotypes. The last of which due to determinants and inverses that need to be calculated at each step.
2. **Choose M** ensures we use enough random effects in our model to account for the genetic correlation in the data, but not to use more than enough for computational speed.
3. **Impute missing data** deals with the missing phenotype data problem. The missing data is then fixed at the imputed values.
4. **Fit model M_0** fits the null model once, having already fixed the number of random effects needed and having imputed the data. This is used also in order to initialise the parameters of the alternative model.
5. **Fit model M_1** fits the alternative model for each genetic variant in our analysis.
6. **Refit model M_0** re-estimates the non-genetic precision parameters under the null, with distribution of S and β keeping their posterior distribution

5. A sparse Bayesian multivariate mixed model

from Step 5. We have found that this step can make a big difference at some SNPs that show strong signal of association. By fixing the VB approximation of S and β from the model M_1 fit, we stop these terms soaking up any of the variance associated with the SNP effect.

5.2.6 Discussion

In the simplest possible case with no missing data and no sample relatedness, the null model marginal likelihood requires no updates and is analytical to compute. The alternative marginal likelihood on the other hand, approximated by the VB mean field approximation, is a lower bound to the true marginal likelihood therefore making the VB Bayes factor a lower bound to the true Bayes factor. As such, the independence assumptions made for the posterior distributions may lead to a loss of power to detect true positives but cannot result in an increased false positive rate.

On the other hand, when modelling sample relatedness the null model marginal likelihood is also approximated and the Bayes factor no longer needs to be a lower bound of the true Bayes factor. So suppose that the true BF is equal to $\frac{a}{b}$ and the VB SAS BF is equal to $\frac{a-x}{b-y}$, where a and b are the true marginal likelihoods under the two models and x and y are the differences between their true and approximated values due to the VB mean field approximations. Two main points then need to be made: (a) in order for VB SAS to introduce a false positive that a true calculation doesn't, we would need to have $\frac{a}{b} < 10^5 \leq \frac{a-x}{b-y}$, which would be the case if $x < 10^5 y$; (b) we would expect that x would be at least as large as (or similar in magnitude to) y because VB SAS makes more

5. A sparse Bayesian multivariate mixed model

independence assumptions under the alternative in addition to those for the null, however this has not been investigated further in this thesis. Based on these reasons, we still think that the variational Bayes modelling and in particular the assumptions made about the form of the posterior distributions should not lead to an increased false positive rate.

Chapter 6

Simulation studies

In this Chapter we use simulated data in order to investigate the performance of the proposed Variational Bayes method in Chapter 5 for both unrelated and related samples.

6.1 Unrelated samples

6.1.1 Data simulation

We start by explaining how we generate datasets. Since we only consider unrelated individuals here, this can be done from Eqn 5.1, or equivalently as in Chapter 3:

- First we simulate a vector of N genotypes $\mathbf{G}$ at a single locus, each element of which is thought to come from a Binomial $Bin(2, p_{\mathbf{G}})$ distribution, where $p_{\mathbf{G}}$ is the minor allele frequency at that locus. We use $p_{\mathbf{G}}$ in $\{0.05 - 0.5\}$.
- Next we simulate N multivariate (P -variate) normal samples with mean $\mathbf{0}$

and covariance matrix Σ for the model errors ϵ .

- Finally, according to some vectors of means μ and effect sizes α , we set $Y = \mathbf{1}\mu + G\alpha + \epsilon$, where $\mathbf{1}$ denotes an $N \times 1$ (column) vector of ones.

This way we obtain simulated genotypes and multivariate normal samples of phenotypes, with pre-specified genotype minor allele frequencies, error covariance structure, vectors of phenotype means and genotype effect sizes. Knowing the “truth”, i.e. the model and parameters under which data is generated, would enable us to check the performance of our method from Chapter 5 in comparison to various other approaches.

6.1.2 Simulation scenarios

We simulate datasets consisting of $N = 2000$ unrelated individuals with P traits where a subset of p traits are associated with a SNP and the remaining $P - p$ traits are unassociated. We consider 8 different combinations of P and p and these are shown in Table 6.1 together with a description of which subset of traits were associated and with which effect sizes (α).

We also vary the amount of correlation between the residual errors of the phenotypes. As the number of traits increases it becomes impossible to fully evaluate all possible combinations of trait effects and error covariance structures. We have chosen to use a simple autoregressive process AR(1), for different strengths of correlation ρ , such that $\Sigma_{ij} = \rho^{|i-j|}$. When ρ is positive this induces positive correlation between all traits but when ρ is negative there will be a mixture of positive and negative correlations between the residual errors of the traits. In

6. Simulation Studies

Scenario	P	p	Effect sizes
1	2	2	$\alpha_1 = \alpha_2 = 0.125$
2	2	2	$\alpha_1 = 0.125, \alpha_2 = 0.05$
3	2	1	$\alpha_1 = 0.125, \alpha_2 = 0$
4	10	1	$\alpha_1 = 0.125, \alpha_i = 0 \ \forall i \neq 1$
5	10	3	$\alpha = \{0.125 \ 0.05 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0.08\}$
6	10	6	$\alpha = \{0.125 \ 0.05 \ 0.08 \ 0 \ 0 \ 0.05 \ 0.03 \ 0 \ 0 \ 0.1\}$
7	100	1	$\alpha_1 = 0.125, \alpha_i = 0 \ \forall i \neq 1$
8	100	6	$\alpha_{1,2,3} = \{0.125, 0.05, 0.08\}, \alpha_{54} = 0.1, \alpha_{99,100} = \{0.05, 0.08\}$

Table 6.1: This table lists the parameters used to simulate the datasets of unrelated samples of P traits where a subset of p traits are associated. The effect sizes of the p traits are given in the 4th column. Effect sizes of 0.125 and 0.05 refer to 0.5% and 0.1% of the traits variance being explained by a genetic variant.

both cases the magnitude of the correlations decreases with increasing difference between the index number of the traits. Under this model it then matters which subsets of traits we chose to be associated as this dictates the residual correlations between these associated traits.

We carry out 10 000 simulations under the models in Table 6.1 for values of ρ ranging in $[-0.9, 0.9]$, by steps of 0.1. In addition we also simulate sets of 10 000 datasets under the null model for each value of ρ and use the empirical null distribution to calculate the 1% significance level of each approach. This allows us to compare the power of Bayesian (and Frequentist) approaches. We apply and compare the following methods to the simulated data:

- P-value Min - minimum p-value method - where we take and analyse each trait separately in a frequentist setting and then take the minimum of the p-values obtained
- BF U from Chapter 3 - where we analyse each trait separately in a Bayesian

setting and average over univariate Bayes factors.

- BF True - where we fit the true model, i.e. fixing all effect sizes of traits we simulate as unassociated to zero - in practice this will not be known, but we include it here for comparison purposes as it should produce an upper bound for power calculations.
- BF Full - where we fit the full model with all effect sizes having normal priors.
- P-value Reverse, where we consider the genotype as response and the phenotypes as predictors in a normal linear model
- P-value MP, which is the MultiPhen approach [O'Reilly et al., 2012], which treats the genotype as an ordinal response and the traits as predictors
- BF A - which is our model averaging approach from Chapter 2. Here we consider putting a binomial (P, π) prior on the number of associated traits in the models, where $\pi = 1/2, 1/3, 1/10$ for $P = 2, 10, 100$ respectively.
- BF VB SAS from Chapter 5 which uses a Spike and Slab prior on the effect sizes. The prior on the proportion of traits associated π here we take to be $Beta(b_1, b_2)$, where $b_1 = 1$ and $b_2 = 1, 2, 10$ for $P = 2, 10, 100$ respectively.

Next we present power estimates for the different approaches, calculated at one percent false positive rate from 10 000 simulations for each scenario considered.

6.1.3 Power, $P=2$

For the 2-trait cases from Table 6.1, power calculations at 1% false positive rate are given in Figure 6.1 for all methods discussed previously. The first thing to note is that both univariate statistics - the minimum p-value approach and the univariate Bayes factor average remain pretty much constant with respect to ρ , which is expected as these methods do not take into account residual trait correlation. They also show very similar power estimates to each other, which is expected as they are effectively modelling the same thing.

On the other hand, in comparison to them, all other approaches demonstrate an increase in power for most levels of residual correlation, which can be more than 30% at times. As discussed before in Chapter 3 when analysing the (BF) A approach in depth, there are also cases for which we observe a drop in power in particular in A approach compared to the univariate approaches. The same can be observed here and applies to all multi-trait approaches that model the residual correlation, but they seem to be affected more in certain cases than others. For example the (MultiPhen) P-value MP has the lowest power when only one of the traits is associated to a genetic variant (top plot) and the A and VB SAS Bayes factors show the highest power (apart from the true model, which of course in theory is not known). On the other hand when we consider 2 traits with the same effect sizes (bottom plot) and high positive residual correlation, A and VB SAS seem to achieve lower power than MP. Here the Reverse and Full models also show an advantage over A, VB SAS and MP, even though for $\rho > 0.3$ they are all less powerful than the Min p-value approach. For the middle plot, where we consider different effect sizes of the two traits, A, VB SAS and MP show very

similar power, apart from perhaps a slight disadvantage of MP for values of ρ in $\{-0.1; -0.4\}$. The Reverse and Full model for this scenario however would be preferred instead of these for ρ in $\{0.2; 0.6\}$. The slight disadvantage of MP in terms of power when compared to Canonical Correlation analysis - or the Reverse here, has been noted in [O'Reilly et al., 2012].

The Reverse and MP methods both consider the genotype as a response and the phenotypes are predictors for all scenarios on two traits the Reverse shows a slight advantage. The VB SAS method and the A method we think of as similar approaches in terms of the effects they attempt to model. Note we have chosen the priors on each of the possible models in the A approach to correspond to the prior on the proportion of associated traits in the VB SAS method, and here they show very similar power for all scenarios.

The problem with the multivariate approaches having a drop in power for certain correlation structures when multiple traits are associated to a single SNP is discussed further about MP in [O'Reilly et al., 2012], where it is noted that its power is lowest when the association of one of the traits with the genotype is entirely explained by the other trait, i.e., when one trait is conditionally independent of the genotype given the other trait. Furthermore they show a correspondence of this to the ratio of the effect sizes $\alpha_i/\alpha_j = r$ (the correlation between the phenotypes), that is, when the genetic effects are in the same direction as the correlation.

Despite VB SAS not achieving the best power in this case, as a reminder, we want to mention that the initial reason for developing this approach was in order

to address the limitation of A of possible genome-wide analysis of only up to a modest number of traits (about 10). Thus we want to focus more on the scenarios that are about to follow.

6. Simulation Studies

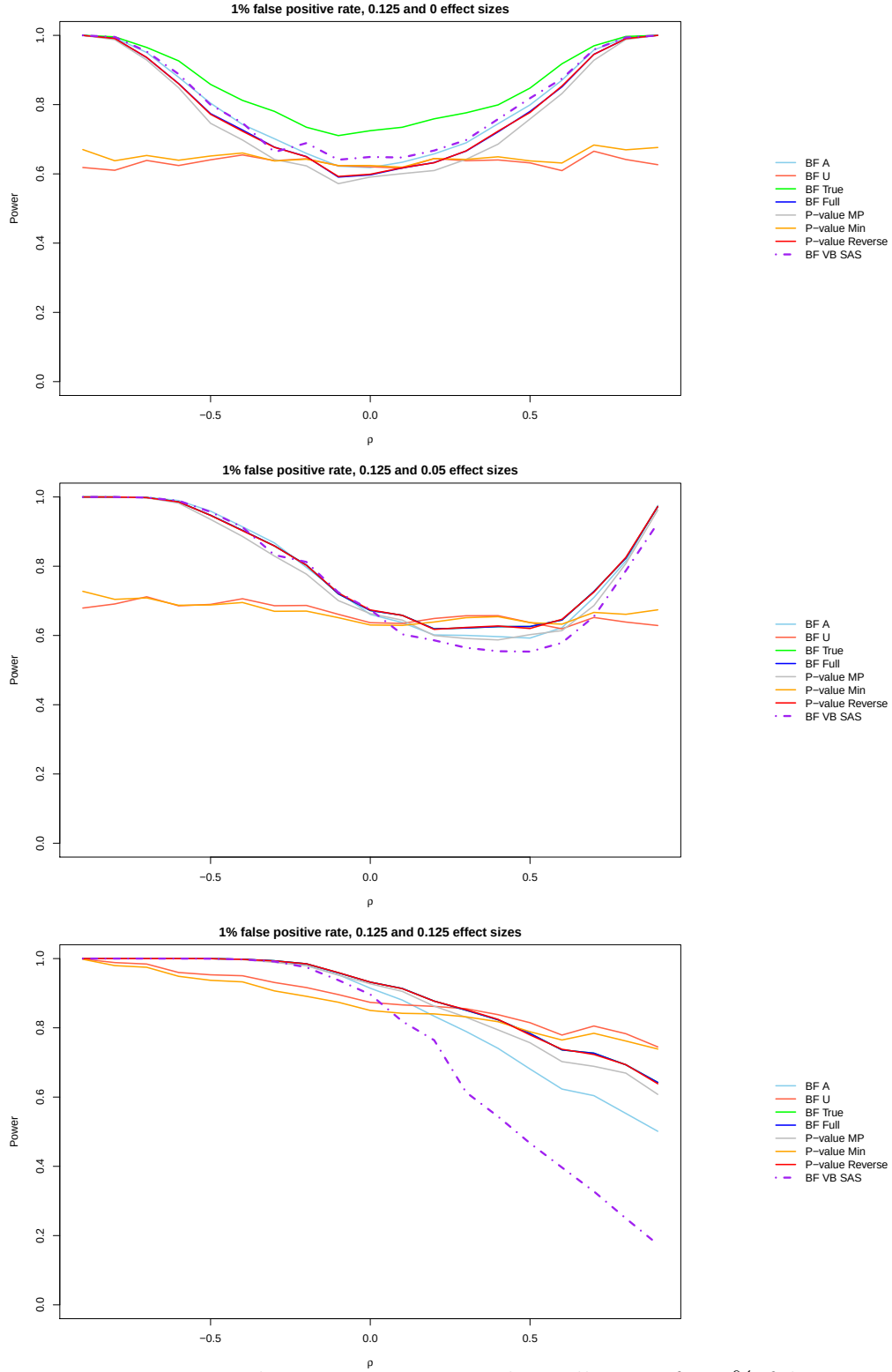


Figure 6.1: Power to detect association, when allowing for 1% false positives, estimated using 10 000 simulations with $P = 2$ traits for simulation scenarios 1 (top), 2 (middle) and 3 (bottom) from Table 6.1 for different levels of error correlations ρ .

6.1.4 Power, $P=10$

For 10 traits, power at 1% false positive rate is shown on Figure 6.2. We are now considering scenarios with 1 (top), 3 (middle) and 6 (bottom) traits associated, respectively, out of a total of 10. As expected again the minimum p-value and univariate Bayes factor average methods are fairly constant with respect to the residual correlation in the model, and show very similar power. The Full and the Reverse also always have very similar power and slightly outperform MP.

As previously discussed here the Full model is not the "true" one but it still outperforms all other methods for 6 associated traits, however as expected we see a large drop of power in the sparser models (fewer traits associated) of the Full, Reverse and MP models compared to the A and VB SAS.

Here unlike in the 2-trait case, perhaps the most advantageous methods to use are VB SAS and A, as they almost always outperform the univariate methods, and when they do not, there is only a negligible loss in power, similar to the one when they are outperformed by any of the other methods (in particular for 6 traits associated).

6. Simulation Studies

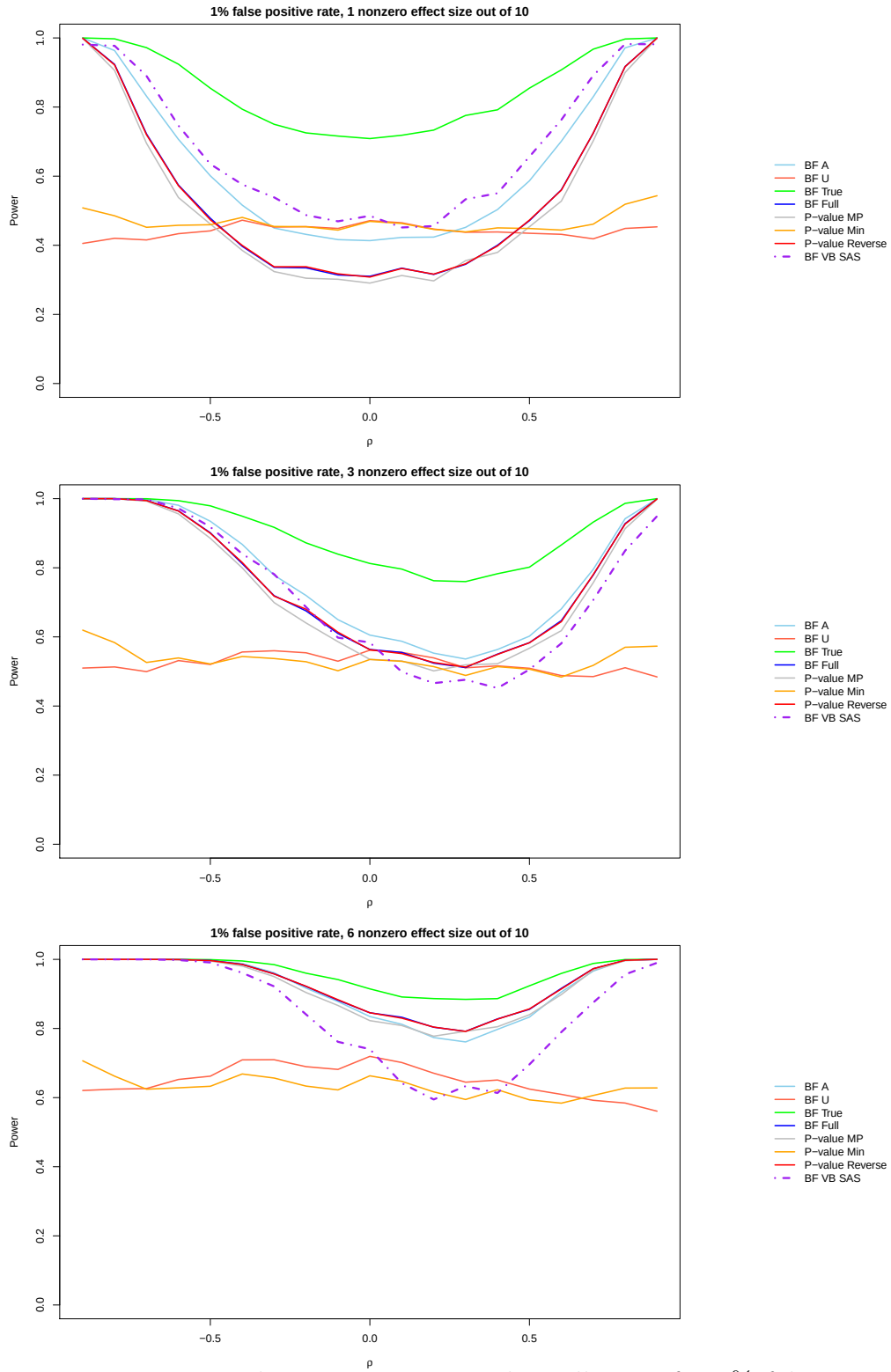


Figure 6.2: Power to detect association, when allowing for 1% false positives, estimated using 10 000 simulations with $P = 10$ traits for simulation scenarios 4 (top), 5 (middle) and 6 (bottom) from Table 6.1 for different levels of error correlations ρ .

6.1.5 Power, $P=100$

For 100 traits, power at 1% false positive rate is shown on Figure 6.3. Here, the number of traits considered is too large to be able to use our model averaging approach A from Chapter 2 and MP, and we have also excluded results on the average of univariate Bayes factors as power estimates are always very similar to the minimum p-value approach.

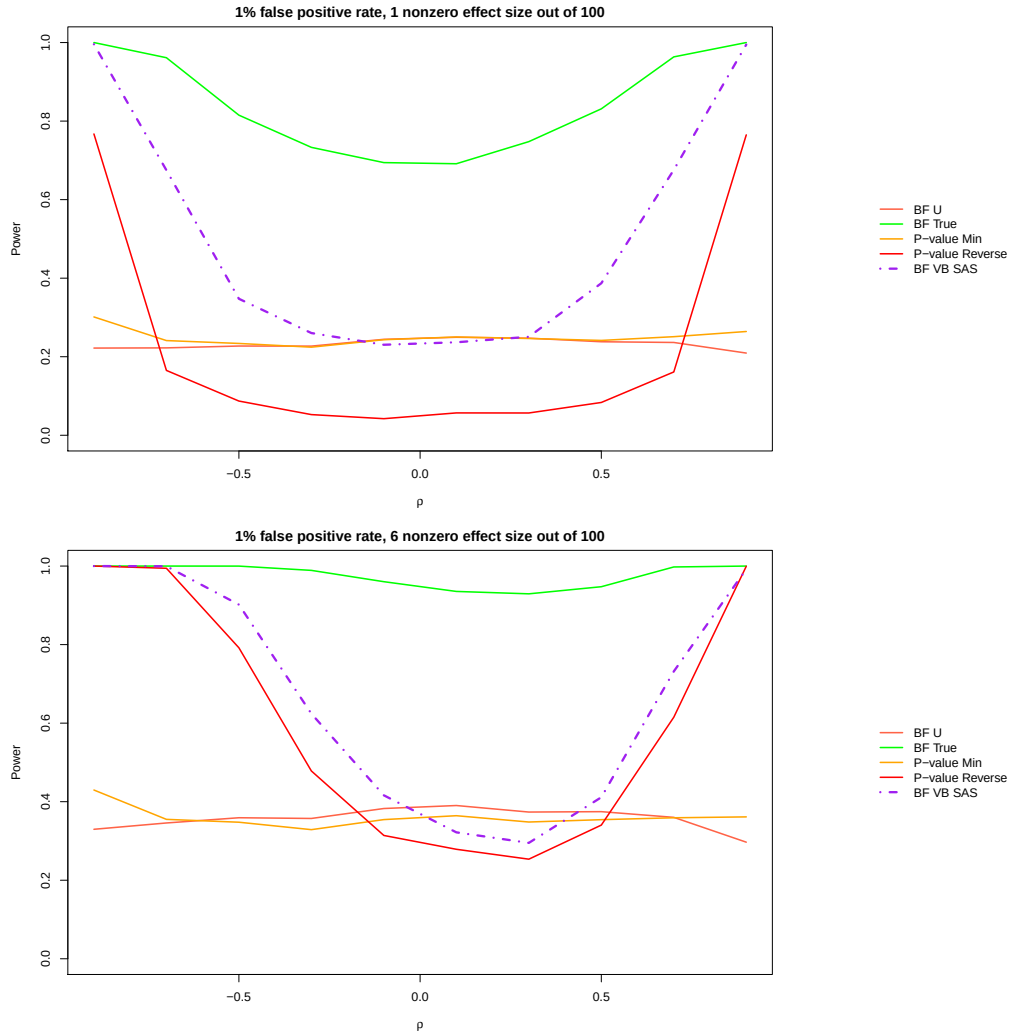


Figure 6.3: Power to detect association, when allowing for 1% false positives, estimated using 10 000 simulations with $P = 100$ traits for simulation scenarios 7 (top) and 8 (bottom) from Table 6.1 for different levels of error correlations ρ .

For such a large number of traits the advantage of VB SAS over all other methods is much more pronounced. In particular the Reverse model (and consequently the Full via similarity) completely lose power to find a single phenotype-genotype association in the presence of 99 unassociated traits. The advantage of the Spike and Slab prior here lies in its properties to shrink the irrelevant coefficients to zero (or almost zero), which for sparse models (like 1 out of 100 traits being associated) would be a clear advantage over the assumption of the Full model that all traits can be associated to the genetic variant in question.

6.1.6 Missing data imputation

We carried out a simulation to assess the utility of our method for imputing missing phenotypes. We compare our method to an approach that omits all samples with any missing data. We use 1000 simulated datasets for a number of scenarios from Table 6.1 and calculate ($\log_{10}$) Bayes factors from the VB SAS approaches without any missing data. We then set 1% of the data for each trait as missing (independently) and re-analyse the data using (1) our built in imputation method (2) the VB SAS method after removing all samples with any missing phenotype data. Figure 6.4 shows comparisons between these methods for $\rho = -0.6$ and scenarios in which 6 or 1 of the traits are considered associated to a genetic variant out of a total of 10 phenotypes.

In the case with 1 out of 10 traits associated (top plot on Figure 6.4), we can see that $\log_{10}$ Bayes factors obtained from our imputation approach correlate well with those obtained from the full data (before removing 1% data) (left plot). Note that by removing 1% of the data for each trait (at random) this results in

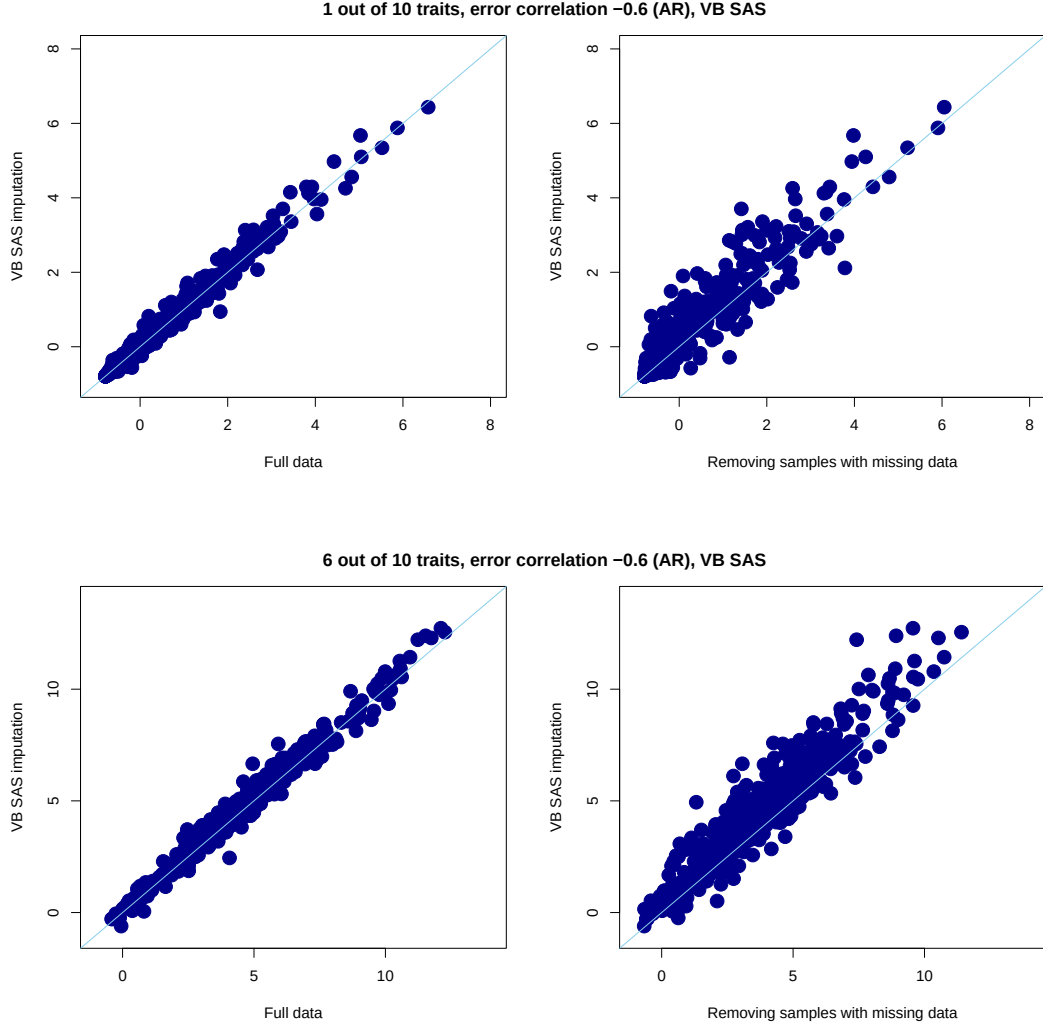


Figure 6.4: VB SAS data imputation comparison - 1 percent per phenotype missing data, $\rho = -0.6$, sample size 2000, 1000 datasets, 1 trait associated out of 10 (top) and 6 traits associated out of 10 (bottom). Left plot compares full data $\log_{10}$ BF to $\log_{10}$ BF from data imputation approach. Right plot compares data imputation approach to $\log_{10}$ BF obtained by analysing the data in which each sample with any missing data is removed (nearly 10% data removed).

1% of the data to be imputed, but if we were to remove all samples with any missing data this results in nearly 10% reduction in sample size. Unsurprisingly this would lead to reduction in $\log_{10}$ Bayes factors compared to using the full

data, and thus compared to our imputation approach (right plot). In fact the same can be observed for all other scenarios with the difference in the magnitude of the Bayes factors and plots for those are not shown here.

In general, we have also looked at extending our approach so it does not explicitly fix or impute the missing phenotype data, but to leave it and model it in the approach as another random variable. This in turn is more computationally intensive as missing data parameters need to be updated at every iteration even in the alternative model. We observed negligible differences in the resulting Bayes Factors on a $\log_{10}$ scale and have not investigated any further (data not shown here).

6.1.7 Effect size estimates

We further investigate the posterior effect size distributions produced by the VB approximation and in particular address possible issues with suitability and appropriate coverage of the obtained intervals. This is relevant for the following Chapter on real data analysis.

We revisit the simulations on six associated out of ten phenotypes and the BF Full, BF VB SAS and univariate analysis P-value approaches, for which we look at the posterior effect size estimates and their corresponding posterior standard deviations (or standard errors for the latter). Figures 6.5 and 6.6 show corresponding boxplots from all simulations for ρ equal to 0 and 0.4 (data for other values of ρ considered but not shown here). Tables 6.2 and 6.3 provide per phenotype coverage for the three methods and two error covariance structures.

6. Simulation Studies

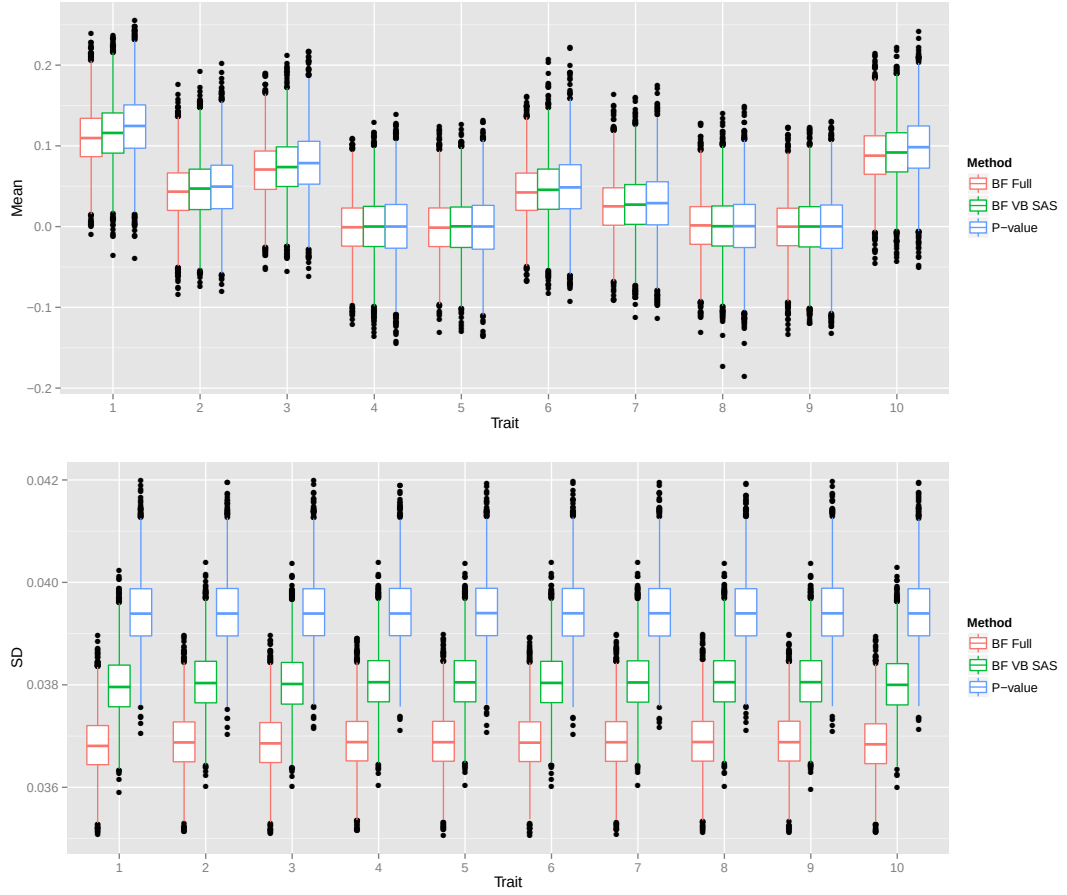


Figure 6.5: Posterior mean (top) and standard deviation estimates (bottom) (or standard errors for the frequentist approach) of the effect sizes for 6 out of 10 traits associated with $\rho = 0$ for the BF Full, BF VB SAS and univariate least squares regression P-value approaches.

Method	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
P-value	0.946	0.953	0.951	0.946	0.944	0.951	0.952	0.947	0.946	0.947
Full	0.945	0.955	0.953	0.962	0.963	0.964	0.962	0.965	0.962	0.947
VB SAS	0.952	0.958	0.956	0.956	0.954	0.959	0.959	0.957	0.955	0.952

Table 6.2: Coverage of credible sets/confidence intervals from the simulations in Figure 6.5 for $\rho = 0$ at 95% level.

6. Simulation Studies

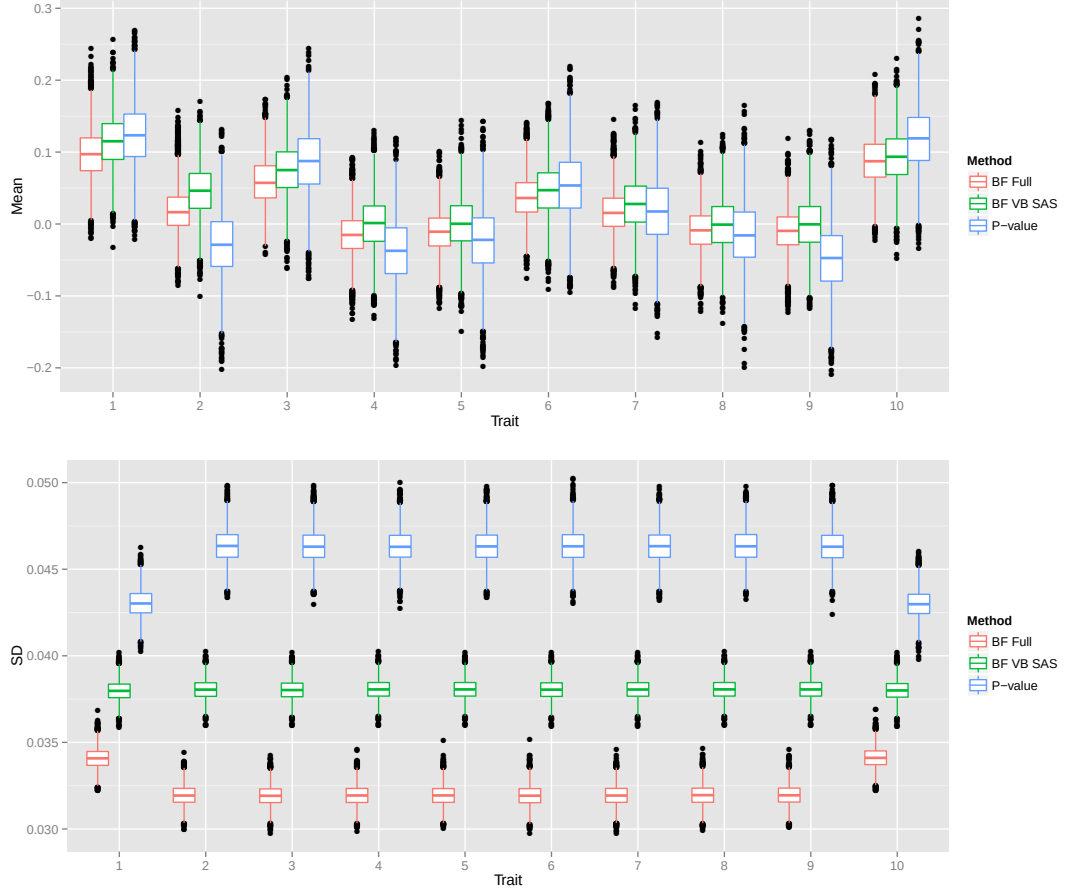


Figure 6.6: Posterior mean (top) and standard deviation estimates (bottom) (or standard errors for the frequentist approach) of the effect sizes for 6 out of 10 traits associated with $\rho = 0.4$ for the BF Full, BF VB SAS and univariate least squares regression P-value approaches.

Method	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
P-value	0.947	0.606	0.951	0.868	0.924	0.945	0.945	0.938	0.820	0.928
Full	0.869	0.844	0.903	0.938	0.956	0.945	0.943	0.956	0.950	0.947
VB SAS	0.952	0.958	0.958	0.963	0.955	0.952	0.963	0.964	0.960	0.952

Table 6.3: Coverage of credible sets/confidence intervals from the simulations in Figure 6.6 for $\rho = 0.4$ at 95% level.

We observe similar mean effect size estimates between approaches, with frequentist ones slightly larger than the Bayesian (due to the priors used for the latter), and with much smaller difference between methods when no residual correlation is present. The same direction of trend is present for the posterior standard deviations/errors, resulting in much tighter credible sets over the CI for the frequentist approach. This effect becomes much stronger when the error correlation between the trait increases - Figure 6.6 for $\rho = 0.4$ versus Figure 6.5 for $\rho = 0$ (data for other values of ρ not shown here). Explanation for this could be the ability of multivariate approaches to model the error correlation of the traits and as such potentially better de-noise the effect of the variant, giving more confidence in the exact SNP effect.

VB SAS uses an approximation to the posterior distribution of parameters and it is possible that the narrow credible sets for the effect sizes are due to the approximation itself. To investigate this further we calculated their coverage from the full set of simulations as the proportion of credible sets that contain the true effect size values. We found that for most traits all methods achieved appropriate coverage and additionally observed VB SAS (and the Full model) to have overall better coverage than the frequentist method despite the more narrow credible sets. To create a full picture of the accuracy of the VB SAS approximation to the posterior computationally intensive approaches (e.g. MCMC) can be employed, however we have not investigated this further in this thesis.

6.2 Related samples

6.2.1 Data simulation

To investigate the ability of our VB approach to control for relatedness between samples we simulate data on 400 sets of 4 siblings with $P = 10$ traits. We first simulate genotypes at a set of “causal” SNPs and then simulate phenotypes conditional upon those genotypes. By allowing overlapping sets of SNPs to affect each trait we induce correlation between the traits.

First we generate genotypes at L loci for 4 sibs in 400 families, which we store in the $1600 \times L$ matrix $\mathbf{G}$. We do this by generating each individual’s maternal and paternal genotypes at each locus, each with MAF $p_l \sim U(0.05, 0.5)$. Then we generate each child genotype by conditioning on the genotypes of their mother and father. In addition to causal SNPs we also simulate a number of non-causal SNPs.

We then simulate traits conditional upon genotype using the model

$$\mathbf{Y} = \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{6.1}$$

where we draw $\boldsymbol{\epsilon}$ from a $N(\mathbf{0}, \mathbf{I}, \boldsymbol{\Sigma})$ distribution. In the simulations that follow we set the effect size matrix $\boldsymbol{\alpha}$ to be a block band matrix structure, in which the sparsity and magnitude of effects control the level of relatedness between samples. Note that all similarity between the phenotypes of siblings is assumed to be due to genetic effects and no environmental relatedness is introduced, which if present may produce stronger phenotypic similarity and can lead to confounding when

using our proposed approach just like in all other mixed effects models that do not model environmental relatedness.

6.2.2 Choice of M

We carried out simulations to assess our method's performance in estimating M using the DIC measure by varying the parameter L in the simulations. We simulated error covariance matrices from a $IW(21, 10 \mathbf{I}_{10})$ distribution, which has mean $\mathbf{I}_{10}$. We generate 1000 datasets for each of the following 3 scenarios and calculate our DIC measures for each M between 1 and 10 (we consider the maximum M to be the number of traits as this allows a separate random effect for each trait).

Scenario 1 We assume only two genetic variants to have effects on the 10 phenotypes ($L = 2$), and generate effect sizes in the following way.

$$\alpha = \begin{pmatrix} \alpha_{1\,1} & \alpha_{1\,2} & \alpha_{1\,3} & \alpha_{1\,4} & \alpha_{1\,5} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha_{2\,6} & \alpha_{2\,7} & \alpha_{2\,8} & \alpha_{2\,9} & \alpha_{2\,10} \end{pmatrix}$$

where each nonzero α_{ij} is generated from a normal distribution with mean $\lambda = 1$ and standard deviation $\sigma = 0.1$ and is assigned a plus or minus sign at random.

Table 6.4 shows the proportion of times we would select each value of M based on a minimum DIC score. According to the results 93% of the time we would select the correct model and if we are wrong then we would most likely choose a model with 3 random effects, rather than the two we simulated the data under.

6. Simulation Studies

M	2	3	4	5
# min DIC(M)	0.929	0.067	0.003	0.001

Table 6.4: Data simulated with $M = 2$. Table shows the proportion of times we would select the number of random effects M based on DIC. 1000 simulations on 10 traits on 4 sibs for 400 families.

Scenario 2 We assume five genetic variants to have effects on the 10 phenotypes ($L = 5$), and generate effect sizes in the following way.

$$\alpha = \begin{pmatrix} \alpha_{1\,1} & \alpha_{1\,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_{2\,3} & \alpha_{2\,4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha_{3\,5} & \alpha_{3\,6} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{4\,7} & \alpha_{4\,8} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{5\,9} & \alpha_{5\,10} \end{pmatrix}$$

where each nonzero α_{ij} is generated from a normal distribution with mean $\lambda = 1$ and standard deviation $\sigma = 0.1$ and is assigned a plus or minus sign at random.

Table 6.5 shows the proportion of times we would select each value of M based on a minimum DIC score. According to the results 99% of the time we would select the correct model and if we are wrong we would choose a model with 4 or 6 random effects, rather than the five we simulated the data under.

M	4	5	6
# min DIC(M)	0.005	0.992	0.003

Table 6.5: Data simulated with $M = 5$. Table shows the proportion of times we would select the number of random effects M based on DIC. 1000 simulations on 10 traits on 4 sibs for 400 families.

Scenario 3 We assume all ten genetic variants to have effects on the 10 phenotypes ($L = 10$), and generate effect sizes in the following way.

$$\alpha = \begin{pmatrix} \alpha_{1\ 1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_{2\ 2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_{3\ 3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha_{4\ 4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha_{5\ 5} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha_{6\ 6} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{7\ 7} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{8\ 8} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{9\ 9} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{10\ 10} \end{pmatrix}$$

where each nonzero α_{ij} is generated from a normal distribution with mean $\lambda = 1$ and standard deviation $\sigma = 0.1$ and is assigned a plus or minus sign at random.

Table 6.6 shows the proportion of times we would select each value of M based on a minimum DIC score. According to the results we would almost always choose a model with less random effects than we should, the most likely values being around 8.

M	6	7	8	9	10
# min DIC(M)	0.002	0.128	0.637	0.230	0.003

Table 6.6: Data simulated with $M = 10$. Table shows the proportion of times we would select the number of random effects M based on DIC. 1000 simulations on 10 traits on 4 sibs for 400 families.

DIC seems to perform well for data coming from small to moderate M (2 and 5), but seems to underestimate the number of unobserved random effects, when $M = P = 10$ here. The effect sizes that we have used for the simulations here are large (around 1) and we are using common variants with MAF between 0.2 and 0.5, resulting in large portion of the phenotypic variance due to the genetic part of the model. Overall, these results indicate that the DIC measure can select a suitable number of random effects.

6.2.3 Modelling related samples

We carried out simulations to test how the VB SAS approach deals with and corrects for population structure. We used an effect size matrix of the form

$$\begin{aligned}\alpha_{lj} &\sim N((-1)^j \lambda, \sigma^2) \quad j \in \{\max(1, \lceil lP/L \rceil - 1), \min(\lceil lP/L \rceil + 1, P)\} \quad \forall l \\ \alpha_{lj} &= 0 \text{ otherwise}\end{aligned}$$

The values of λ and σ^2 will dictate the level of correlation between traits. Using a small σ^2 will generate two sets of traits with very high positive correlation within each set but negative correlation between sets. Using a large λ increases the heritability. We used the parameters $P = 10$, $L = 100$, $\lambda = 0.5$ and $\sigma = 0.1$.

We fit the VB SAS method at all the non-causal SNPs. We fit the model using $\mathbf{K}$ equal to the expected kinship matrix for the 400 families of 4 sibs. Specifically, for pairs of sibs $K_{ij} = 1/4$ and is equal to zero for unrelated individuals. We also fit the model using $\mathbf{K} = 1/2\mathbf{I}_N$ which assumes unrelated samples as this allows us to assess how well the method does in correcting for relatedness. We used 1000

simulations.

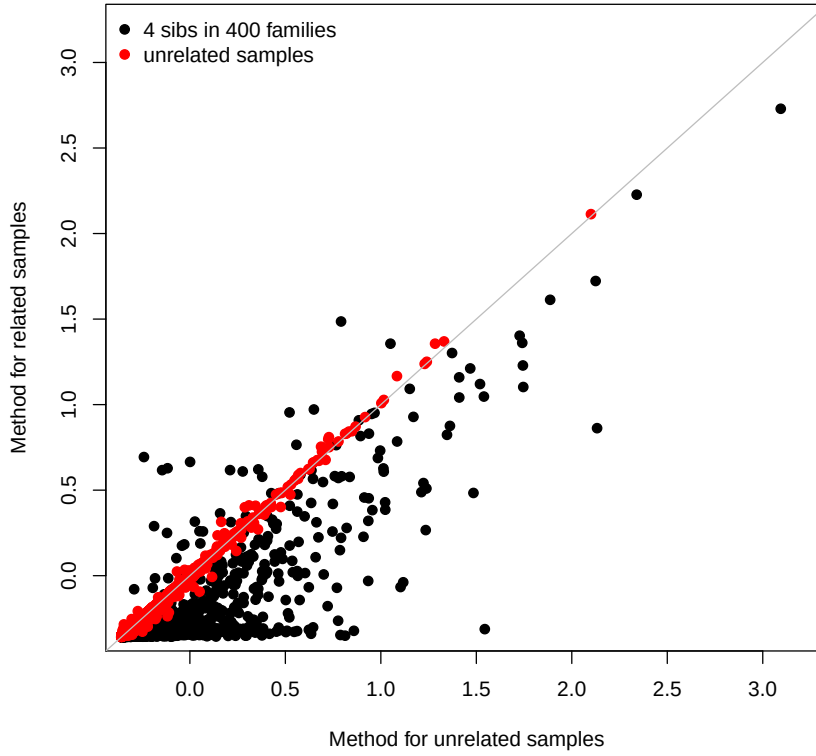


Figure 6.7: Comparison between $\log_{10}$ Bayes Factor from methods for related and unrelated samples for two types of data under the null - one for unrelated data and one for data on 4 siblings in 400 families with $\lambda = 0.5$ and $\sigma = 0.1$.

Figure 6.7 shows $\log_{10}BF$ s from VB SAS when accounting for population structure (y-axis) and when fitting the model for unrelated samples (x-axis). It shows very similar results for data on unrelated individuals (in red) where results from the two approaches correlate very well. On the other hand, we observe much larger Bayes factors in the method that ignores population structure, when the data is in fact generated to be related, when compared to the method that mod-

els sample relatedness. Perhaps more is to be desired as there are still a few points in the top-left part of the plot, however if you note the axes, none of these points would be considered significant. Hence a much larger number of simulations would be necessary to conclude the power of the method to correct for population structure, however at this point the observations suggest that we can better deal with underlying population structure under the null by our mixed effects model.

6.3 Summary

To summarise the work in this Chapter, we performed simulation studies under various scenarios for unrelated individuals and observed overall best performance of our proposed VB SAS approach in cases where the underlying genotype-phenotype structure is sparse. In general, we expect that SNPs would only contribute to the variability of a small number of traits, which is exactly the strength of the Variational method and we would suggest applying it to data on more than say 10 phenotypes.

For our model that allows for related samples we demonstrated the ability of our method to accommodate and model missing phenotype data and show its clear advantage over removing samples with missing phenotype data. We demonstrated that the method has good performance at estimating the latent dimensionality of the random effects term. Finally, we showed that the method does not exhibit the same levels of test statistic inflation when compared to a method that assumes unrelated individuals.

Chapter 7

Real Data Analysis

In this Chapter we analyse data on 15 glycans phenotypes from 3 isolated population cohorts, which have been previously analysed one trait at a time [Lauc et al., 2010]. Here it is not feasible to use the model averaging approach A due to (i) the number of phenotypes and (ii) the population structure of the data, therefore we only employ our Variational sparsity inducing VB SAS method.

7.1 Glycans data

The sample sizes of the three cohorts are : ORCADES 842, VIS 907 and KORCULA 786. These are isolated populations, from Scotland and Croatia, with individuals showing high levels of relatedness. A full description of the cohorts and genotyping can be found in [Lauc et al., 2010]. For our analysis we use imputed genotype data on ~ 9 million SNPs, filtered on $MAF > 0.01$ and imputation score $INFO > 0.4$.

A total of 15 phenotypes have been collected on each of the samples. Phenotypes labelled DG1 - DG13 refer to different structurally related groups of glycans. The traits FUC-C and FUC-A are derived from the original variables to calculate the percentage of glycan structures containing core (FUC-C) or antennary (FUC-A) fucose, yielding a total of 15 glycan traits for analysis. All cohorts have sporadically missing phenotypes, overall around 5% of individuals have some missing phenotypes.

The logarithm for each of the 15 Glycan phenotypes were regressed on age, sex and fibrinogen levels, as done in [Lauc et al., 2010]. Residuals that were greater than three times the interquartile range were removed as outliers and set as missing phenotypes. The linear regressions were then refitted with these outliers removed. The residuals from this final set of regressions are the variables used in the subsequent analysis. The kinship matrix $\mathbf{K}$ used in all the univariate and mixed model analysis was calculated from the genotype data using the approach of [Hayes et al., 2009].

Correlation between phenotypes can be seen in Figure 7.1 for ORCADES and is similar for the other two cohorts (figures not included here).

7.2 Univariate analysis

The original analysis of this dataset carried out separate analysis of each phenotype, combined the GWAS scans via meta-analysis and this discovered 3 associated loci at loci 12q24.31, 14q23.3 and 19p13.3. A subsequent larger meta-analysis [Huffman et al., 2011] added in an additional cohort of 656 samples from

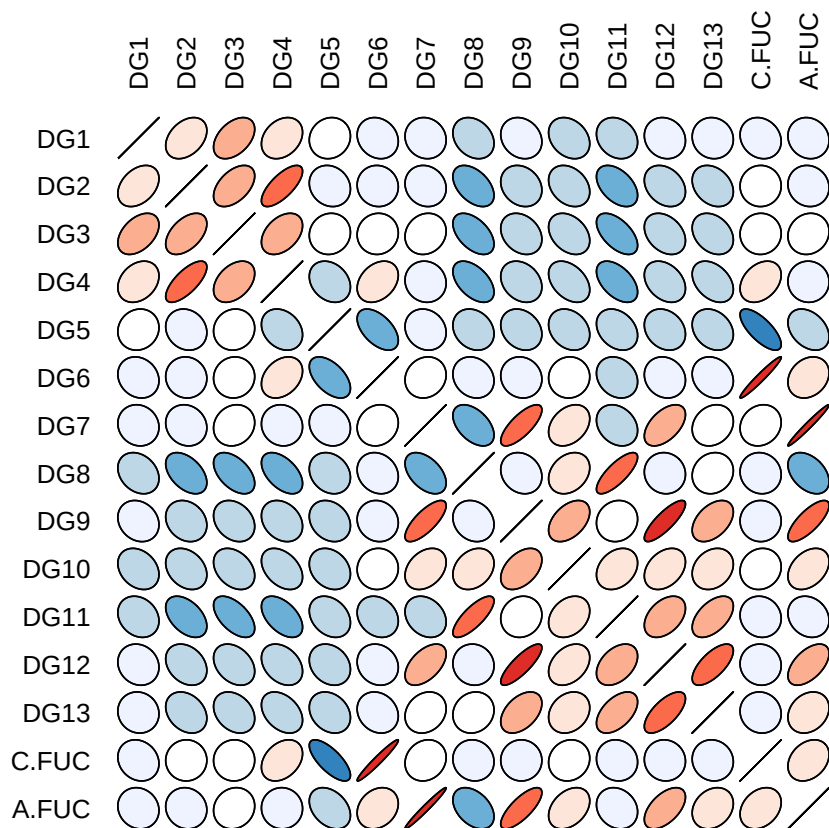


Figure 7.1: Glycans sample correlation plots - ORCADES.

the Northern Swedish Population Health Study (NSPHS) and found additional associations at 2q21 and 11q25.

We applied univariate analysis by fitting a mixed model genome-wide at each SNP and for each phenotype in turn using FaST-LMM [Lippert et al., 2011]. For each cohort we obtained the minimum P-value (Min) for all traits and corrected this statistic for multiple-comparisons using a Bonferroni correction. Genome-wide scans for the three cohort from this approach are shown in Figure 7.2. These plots show that associations at 19p13.3 and 14q23.3 are clearly visible in all 3

7. Real Data Analysis

cohorts, but the association at 12q24.31 can only be seen in VIS. The associations at 2q21 and 11q25 are not visible in this analysis.

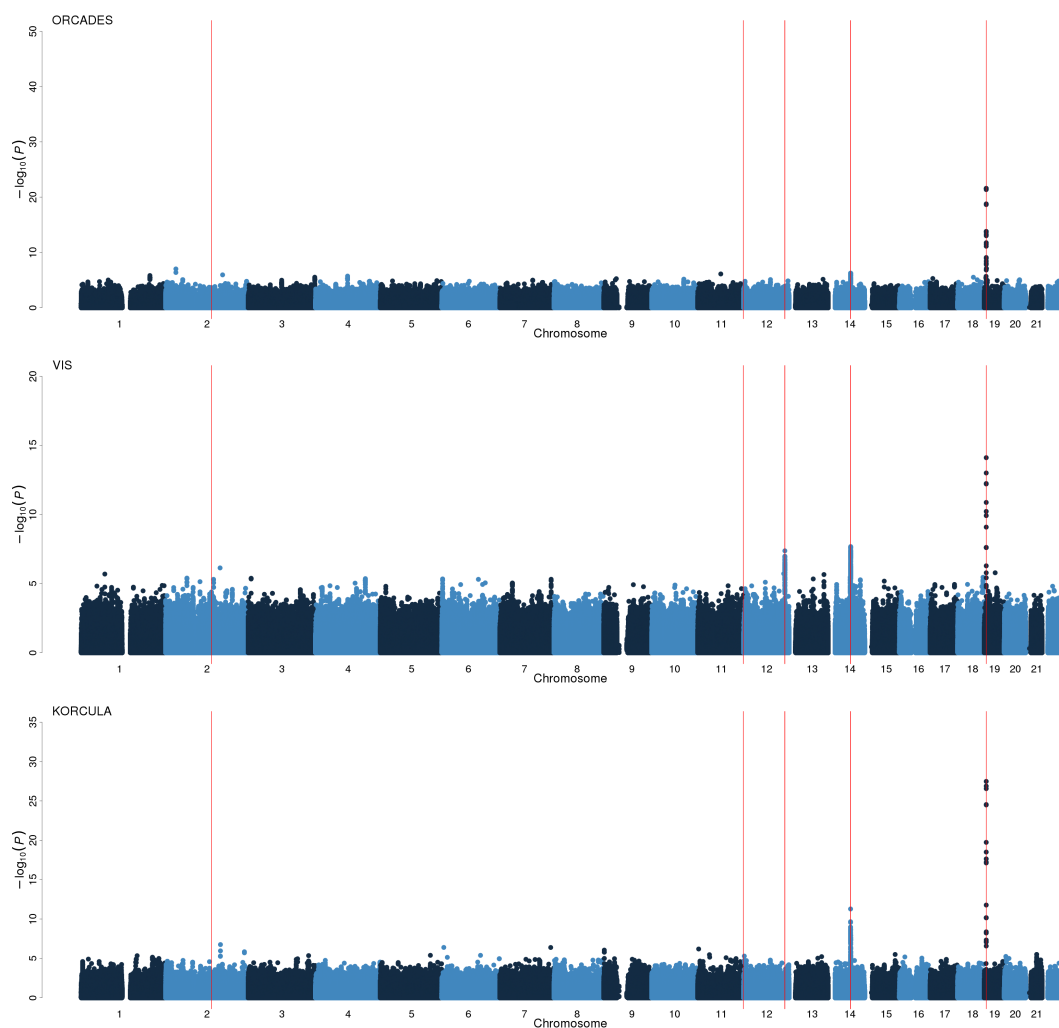


Figure 7.2: Glycans P-value Min from FaST-LMM genome-wide manhattan plots for the three cohorts. Vertical red lines show the positions of the known loci for the analysed traits.

7.3 VB SAS

We then applied the VB SAS method to each of the 3 cohorts. For this analysis we set the prior on the proportion of traits associated to be $\pi \sim \text{Beta}(1, 3)$, which favours a quarter of the 15 traits to be associated apriori. All other priors are set as in Chapter 5.

The first step of our approach is choosing the number of random effects to be used in the modelling. As discussed in the methodology, we do this under the null model according to deviance information criterion (DIC). Tables 7.1, 7.2 and 7.3 show the DIC values for the three datasets and point to optimal number of random effects of 6 for ORCADES and 5 for VIS and KORCULA.

ORCADES

Trait	DIC	Trait	DIC	Trait	DIC
$M = 1$	-28421.1	$M = 6$	-29069.0	$M = 11$	-28781.7
$M = 2$	-28722.3	$M = 7$	-29031.2	$M = 12$	-28711.7
$M = 3$	-28933.3	$M = 8$	-28981.3	$M = 13$	-28642.1
$M = 4$	-29004.0	$M = 9$	-28919.1	$M = 14$	-28575.6
$M = 5$	-29055.7	$M = 10$	-28853.7	$M = 15$	-28505.0

Table 7.1: DIC values for M ranging from 1 to 15, ORCADES dataset. $M = 6$ shows the smallest DIC value and is therefore chosen for the analysis.

VIS

Trait	DIC	Trait	DIC	Trait	DIC
$M = 1$	-31861.2	$M = 6$	-32347.2	$M = 11$	-32882.3
$M = 2$	-32038.0	$M = 7$	-32323.6	$M = 12$	-31797.0
$M = 3$	-32327.4	$M = 8$	-32194.1	$M = 13$	-31708.0
$M = 4$	-32421.8	$M = 9$	-32079.3	$M = 14$	-31612.2
$M = 5$	-32477.2	$M = 10$	-31981.3	$M = 15$	-31532.2

Table 7.2: DIC values for M ranging from 1 to 15, VIS dataset. $M = 5$ shows the smallest DIC value and is therefore chosen for the analysis.

KORCULA

Trait	DIC	Trait	DIC	Trait	DIC
$M = 1$	-26789.0	$M = 6$	-27547.3	$M = 11$	-27068.2
$M = 2$	-27028.3	$M = 7$	-27504.5	$M = 12$	-26977.2
$M = 3$	-27346.0	$M = 8$	-27440.7	$M = 13$	-26909.4
$M = 4$	-27577.2	$M = 9$	-27331.2	$M = 14$	-26805.3
$M = 5$	-27593.4	$M = 10$	-27211.8	$M = 15$	-26714.9

Table 7.3: DIC values for M ranging from 1 to 15, KORCULA dataset. $M = 5$ shows the smallest DIC value and is therefore chosen for the analysis.

7.3.1 GWAS scans

Genome-wide manhattan plots from our VB SAS approach in ORCADES, VIS and KORCULA, together with the corresponding FaST-LMM scan are shown in Figures 7.3 7.4 and 7.5, respectively. For all three cohorts VB SAS uncovers all known loci that its corresponding univariate analysis does. In addition, the 2q21 locus from [Huffman et al., 2011] can be found with our approach in VIS and KORCULA, but cannot be seen in the univariate approach.

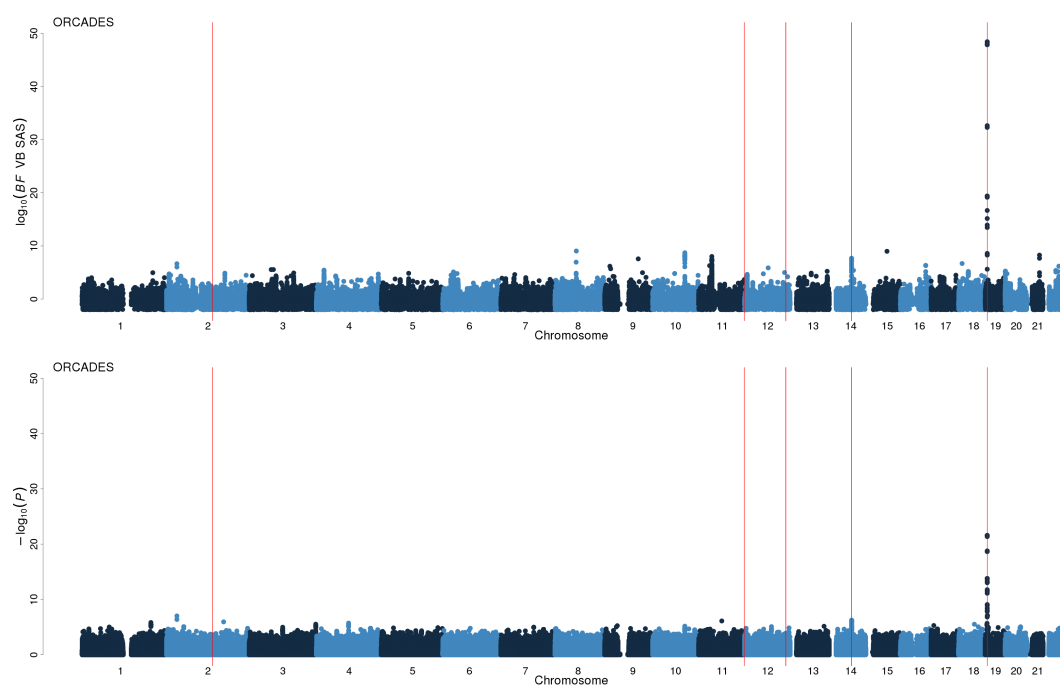


Figure 7.3: Genome-wide manhattan plot for ORCADES from VB SAS (top) and FaST-LMM (bottom). Vertical red lines show the positions of the known loci for the analysed traits.

7. Real Data Analysis

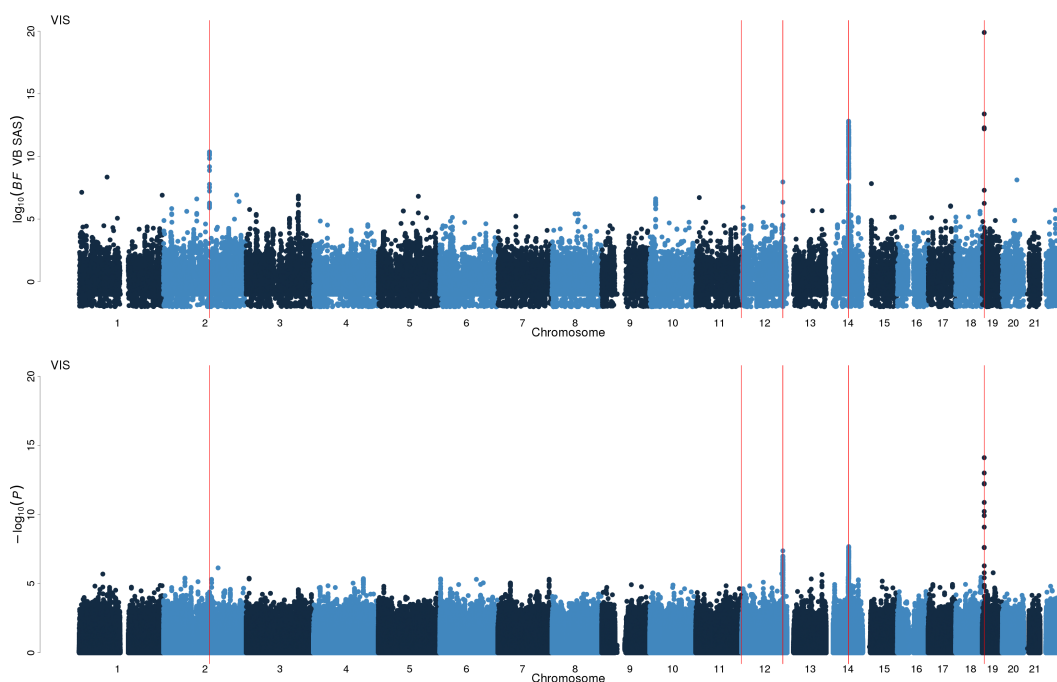


Figure 7.4: Genome-wide manhattan plot for VIS from VB SAS (top) and FaST-LMM (bottom). Vertical red lines show the positions of the known loci for the analysed traits.

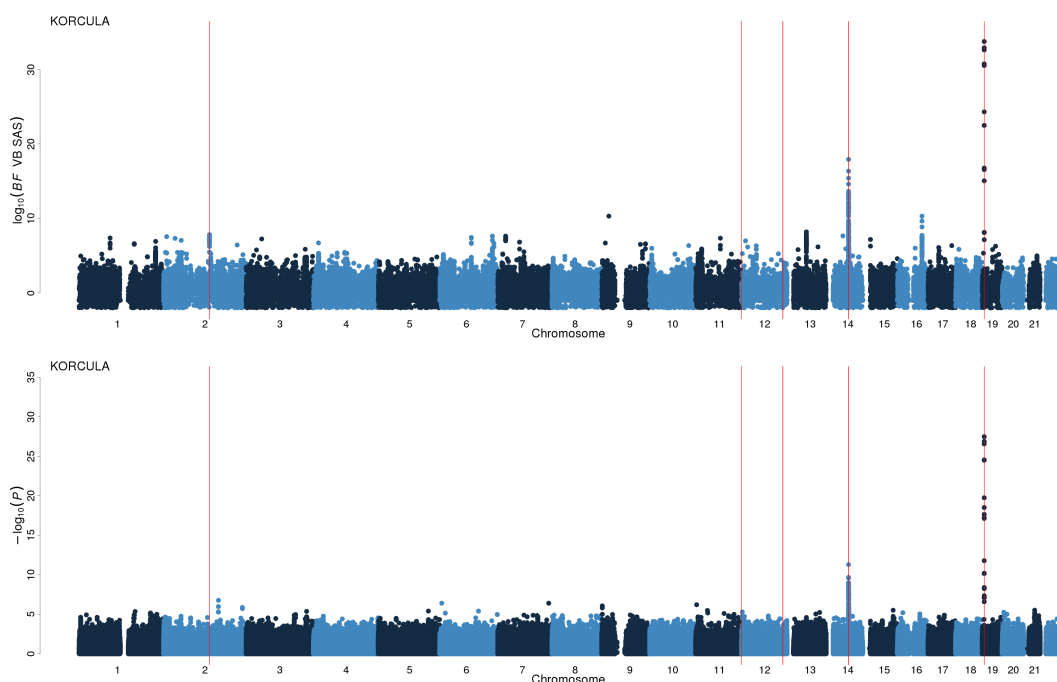


Figure 7.5: Genome-wide manhattan plot for KORCULA from VB SAS (top) and FaST-LMM (bottom). Vertical red lines show the positions of the known loci for the analysed traits.

7.3.2 Known associations

FUT6/FUT3 Manhattan plots for the 19p13.3 locus for the three cohorts using both approaches is shown in Figure 7.6. In all three cohorts we observe larger BF statistics than the corresponding p-values as well as a lower background noise in VB SAS than univariate approach. Effect size plots for this locus are shown in Figure 7.7, where for Fast-LMM for each phenotype the estimated mean effect size is plotted along with bars for 2 standard errors away from it. Then for each trait the magnitude of the p-value is reported via the colouring. For VB SAS we plot the posterior mean of each phenotype effect size along with two standard deviation distance bars. Here the colouring refers to the magnitude of the posterior probability of association. [Lauc et al., 2010] only report association to the traits with the largest effect sizes - DG7, DG9, DG12 and A.FUC, however our VB SAS method additionally classifies DG8, DG10 and DG11 as associated in all datasets.

Note, in Chapter 6 we saw much narrower intervals from VB SAS when compared to univariate analysis for traits with highly correlated errors. As expected, for example here in ORCADES the traits with extremely high correlations to other traits from Figure 7.1 have tighter posterior intervals as opposed to the ones that are less correlated to the rest like DG13, DG10, DG1.

7. Real Data Analysis

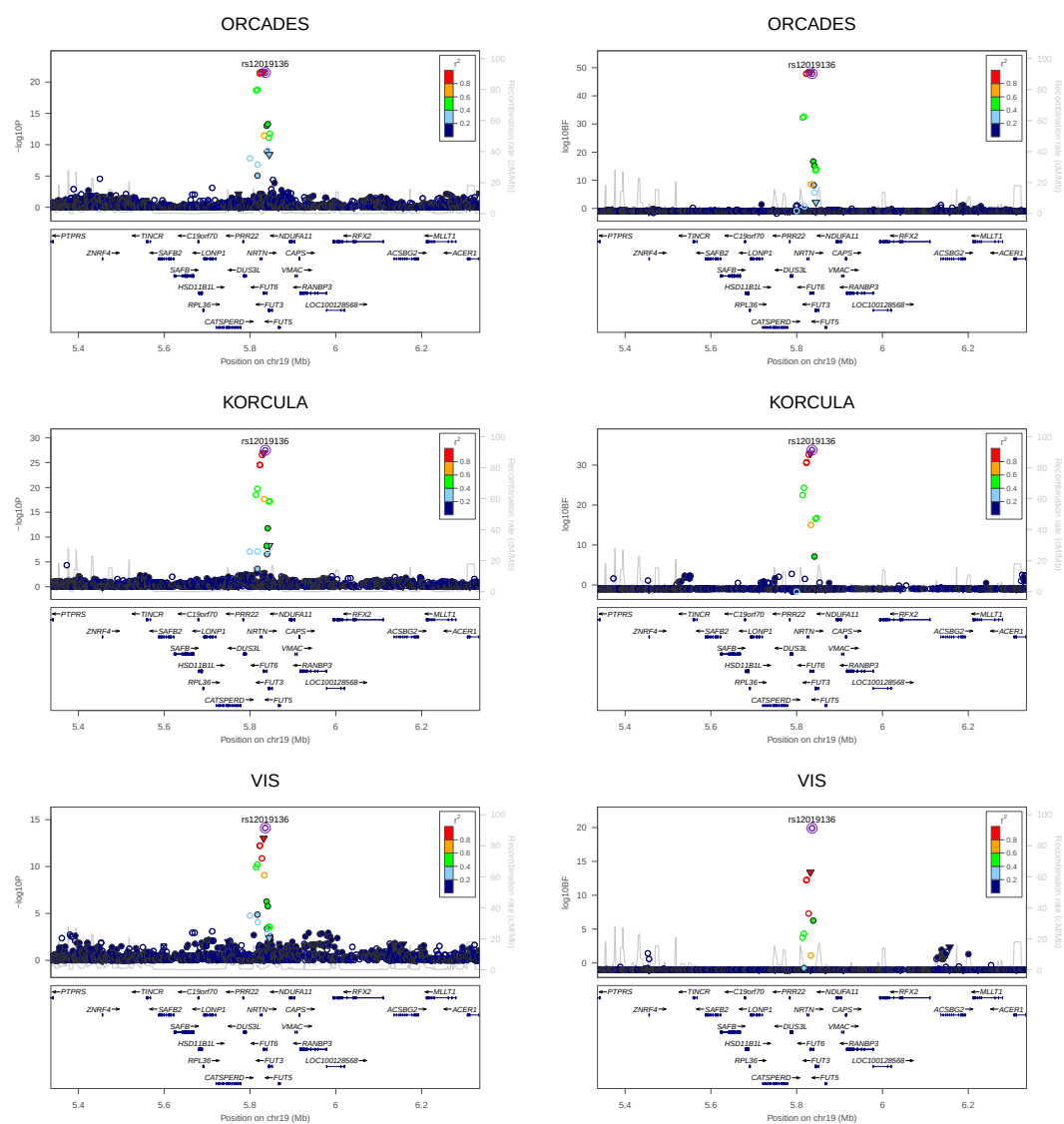


Figure 7.6: Manhattan plots, Min p-value approach (left) and VB SAS (right) - three datasets, chr19 FUT6/FUT3 locus. VB SAS figures are truncated at $\log_{10}BF$ of -2 .

7. Real Data Analysis

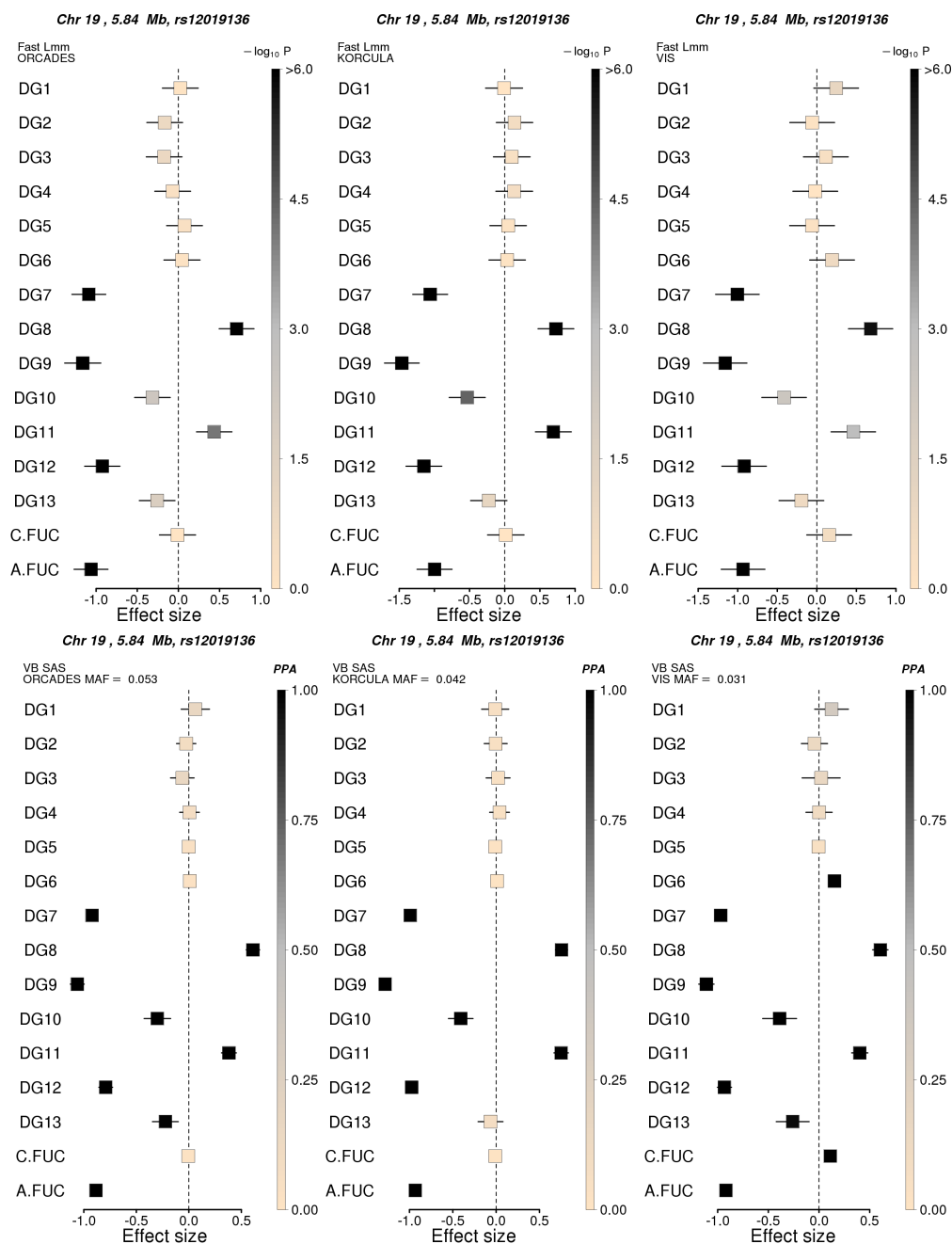


Figure 7.7: Effect size plots, Min p-value approach (top) and VB SAS(bottom).

7. Real Data Analysis

FUT8 Manhattan plots for the 14q23.3 locus for the three cohorts using both approaches is shown in Figure 7.8. Again we observe larger BF statistics than the corresponding p-values for the top associated variants as well as a lower background noise in VB SAS than univariate approach. Effect size plots for this locus are shown in Figures 7.9, which has been reported to have associations to DG6. With VB SAS we can identify this and in addition point to association to C.FUC and DG1 in all three cohorts.

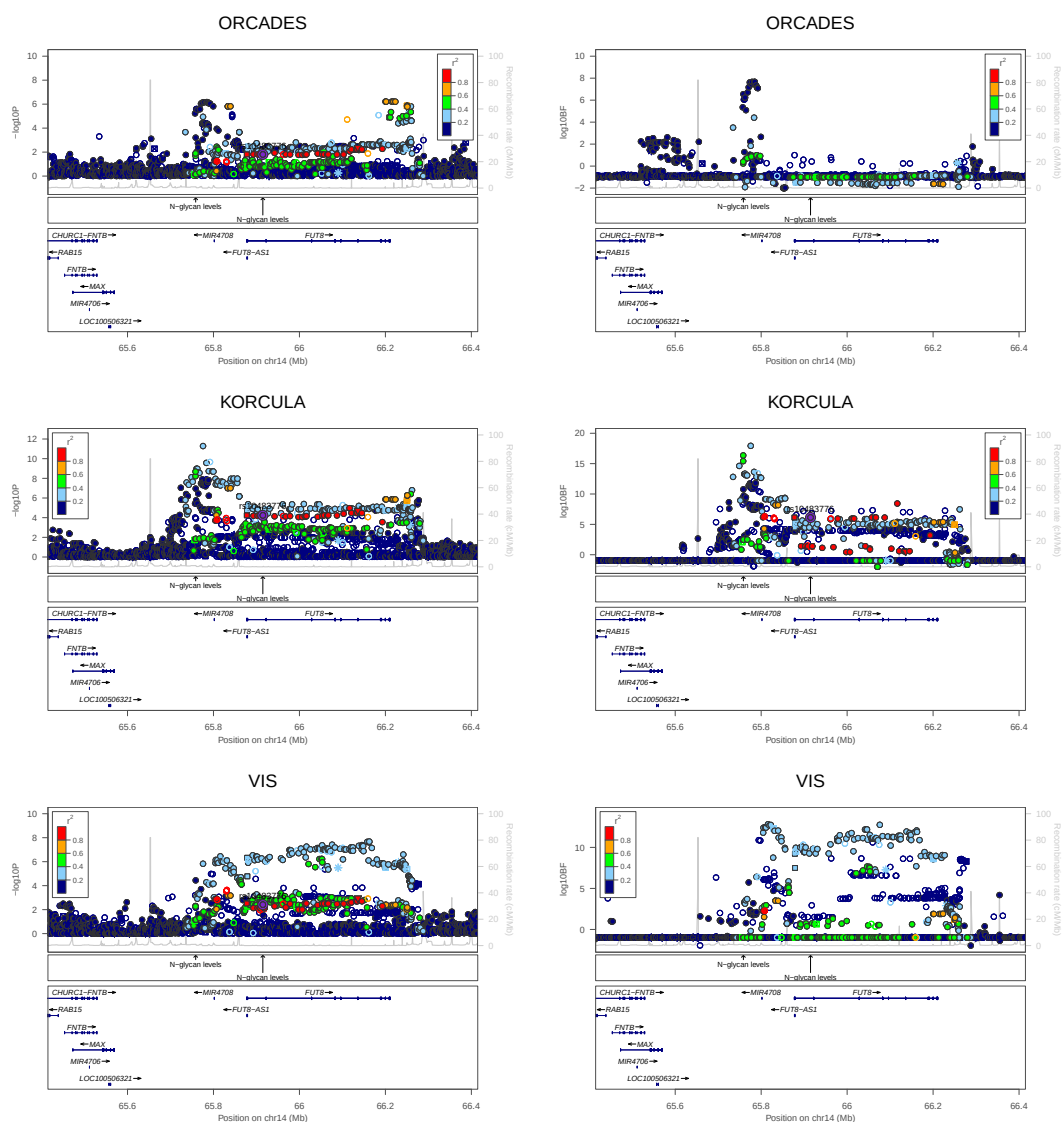


Figure 7.8: Manhattan plots, Min p-value approach (left) and VB SAS (right) - three datasets, FUT8.

7. Real Data Analysis

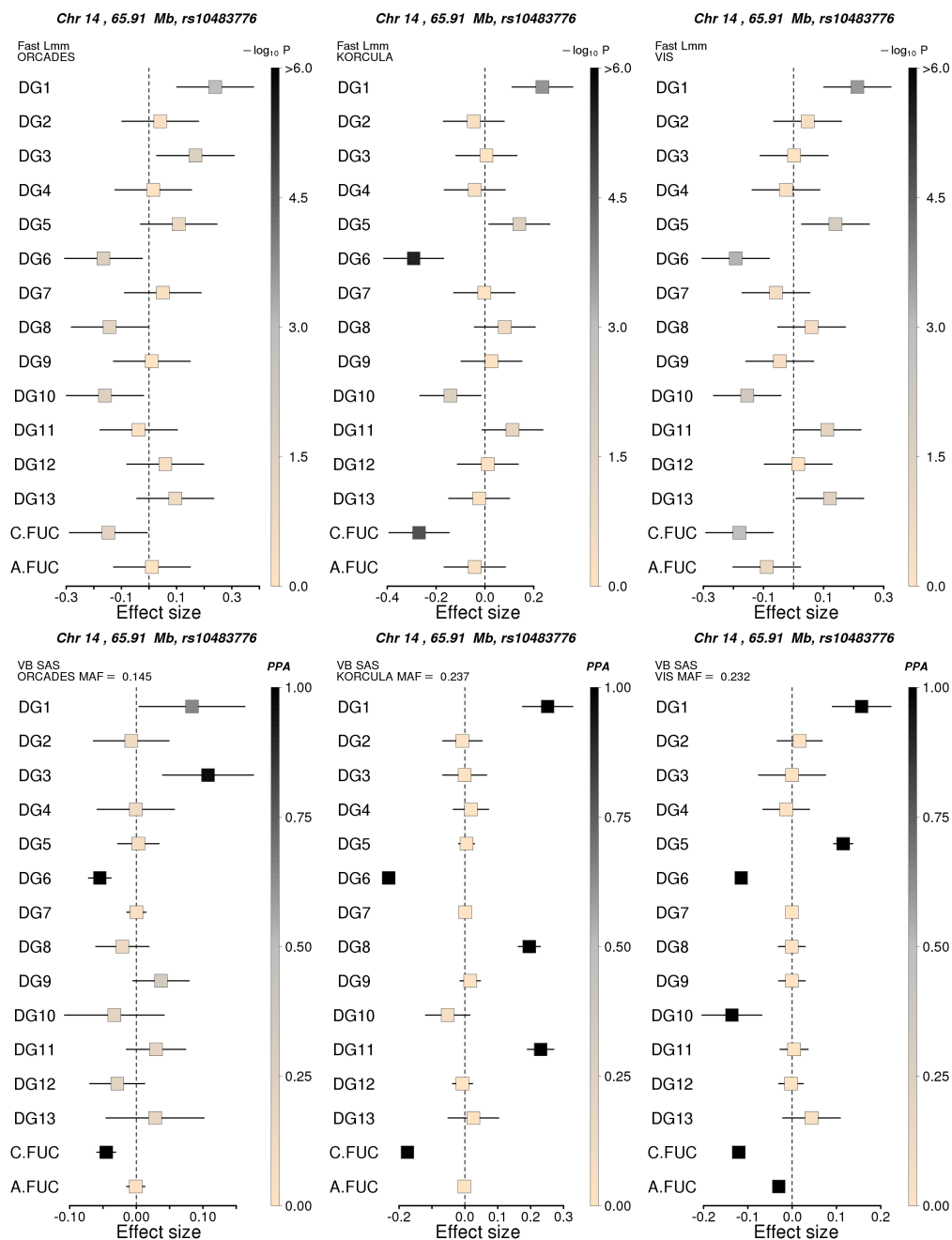


Figure 7.9: Effect size plots, Min p-value approach (top) and VB SAS (bottom) - three datasets, rs10483776, FUT8.

7. Real Data Analysis

HNF1A Manhattan plot for the 12q24.31 locus are given in Figure 7.10. Association signal is only present in VIS. Effect size plots for a variant in that locus are shown in Figures 7.11 for which association to only DG7 and DG11 is previously reported. Our VB SAS approach suggests additional association to DG8, DG9, DG12, A.FUC, DG6, DG10 and C.FUC in VIS.

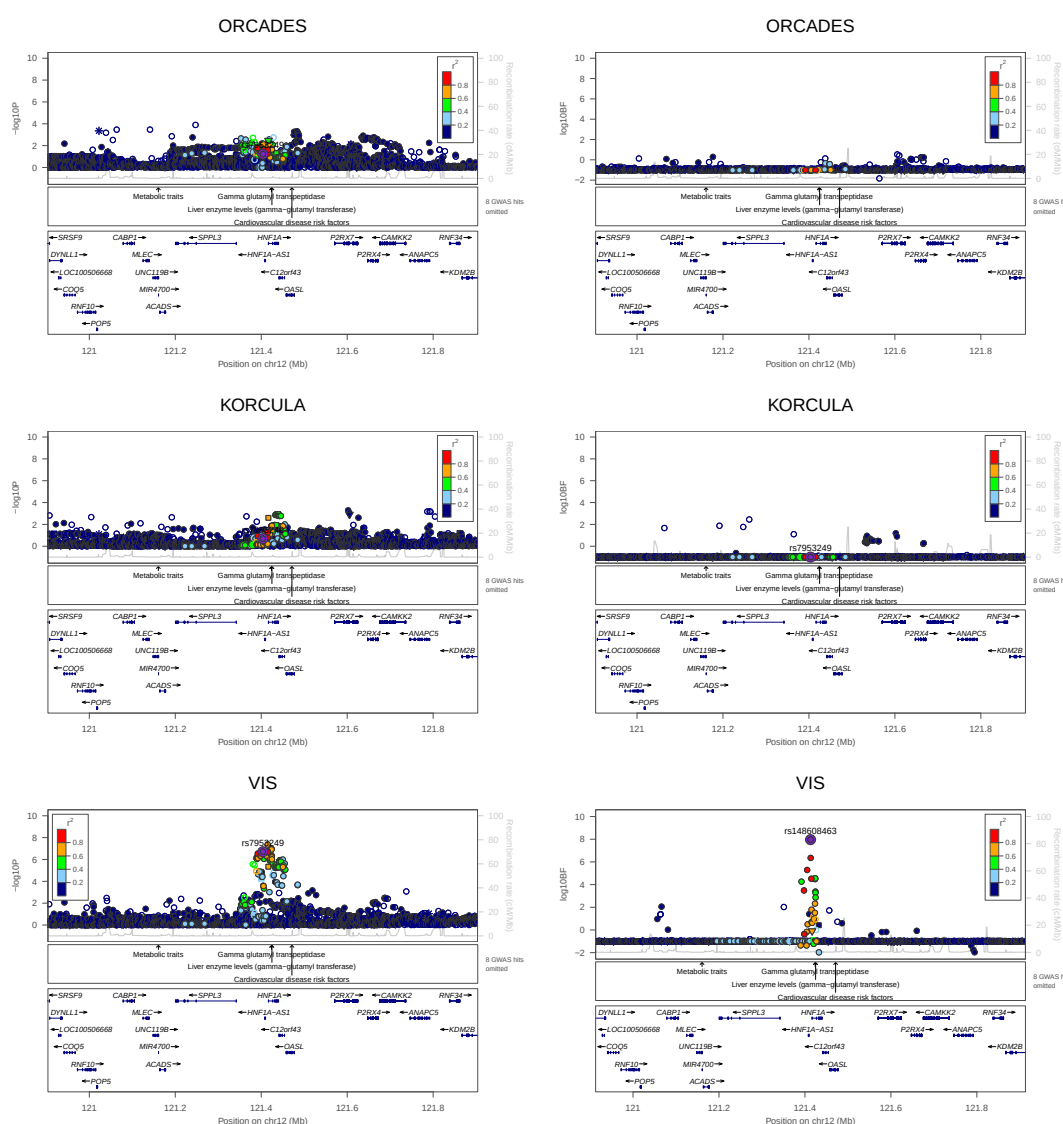


Figure 7.10: Manhattan plots, Min p-value approach (left) and VB SAS (right) - three datasets, HNF1A.

7. Real Data Analysis

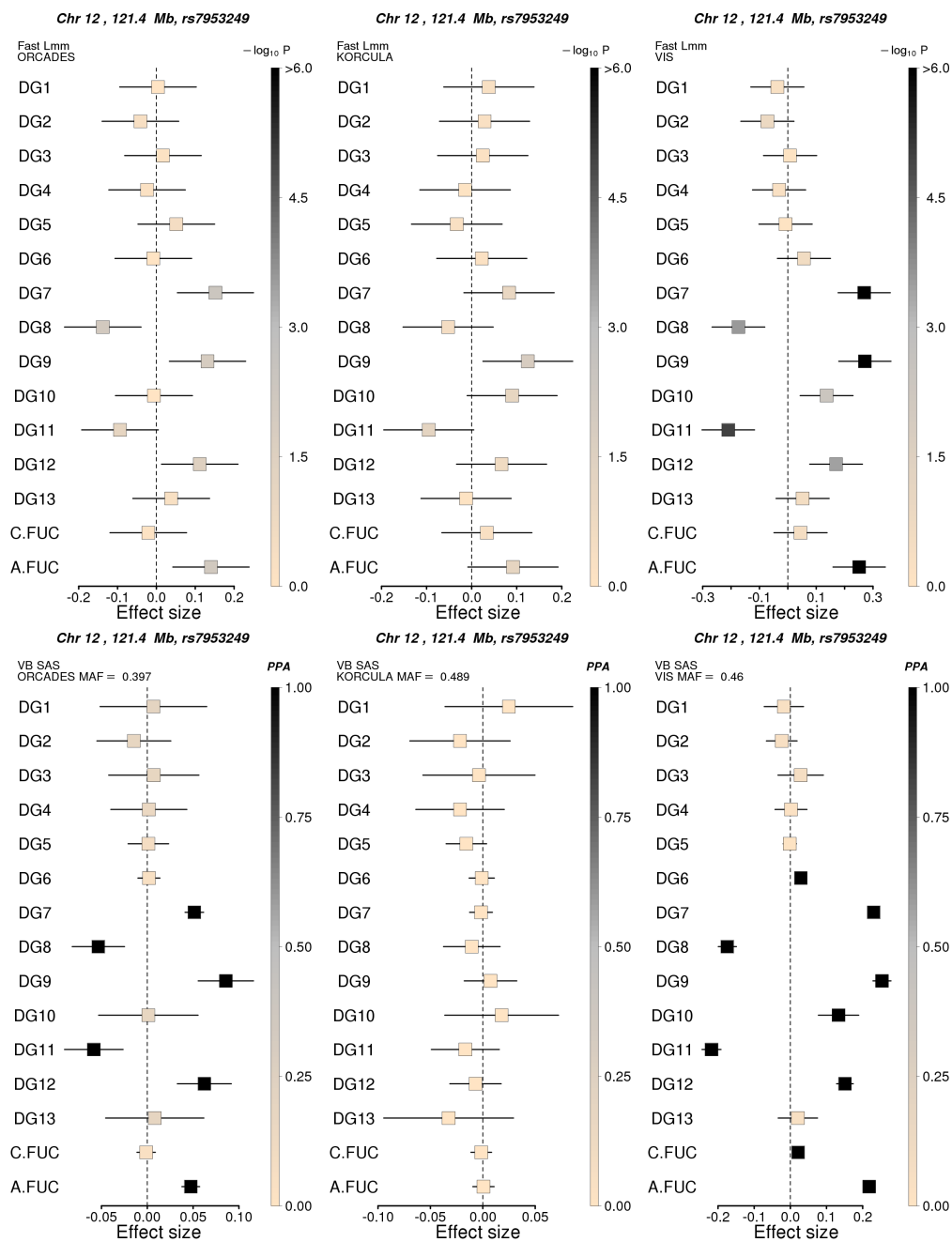


Figure 7.11: Effect size plots, Min p-value approach (top) and VB (bottom) - three datasets, chr12, position 121403724, rs7953249.

7. Real Data Analysis

MGAT5 After the pilot study [Lauc et al., 2010], [Huffman et al., 2011] discovered association to DG11 in locus 2q21, which VB SAS detects in VIS and KORCULA. Manhattan plots of this region are shown in Figure 7.12. Effect size estimates are given in Figure 7.13 and show association of DG11 in all three cohort from our method. Additionally, suggested possible associations to DG12 in VIS and DG9 in KORCULA are observed.

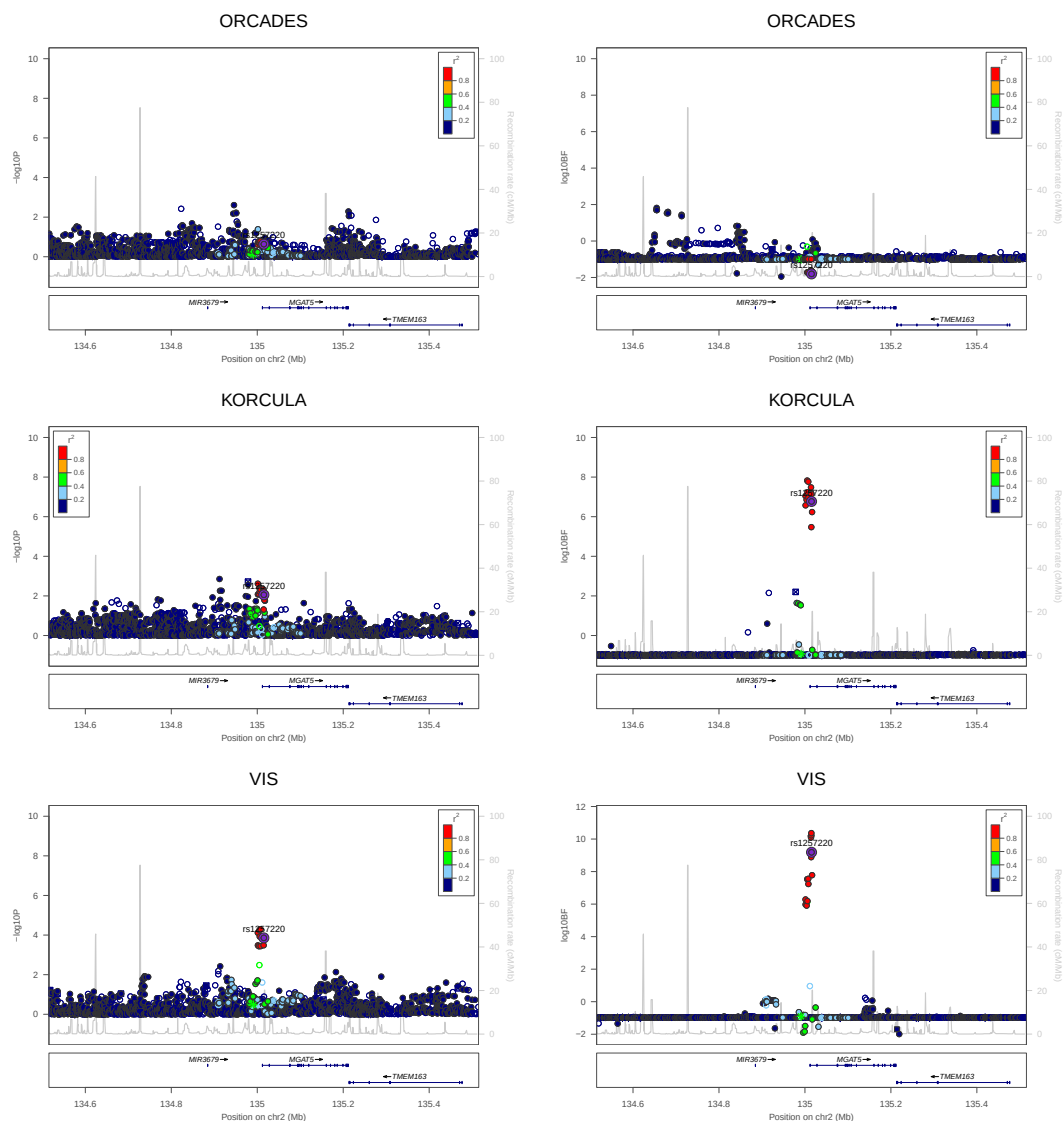


Figure 7.12: Manhattan plots, Min p-value approach (left) and VB (right) - three datasets, chr2.

7. Real Data Analysis

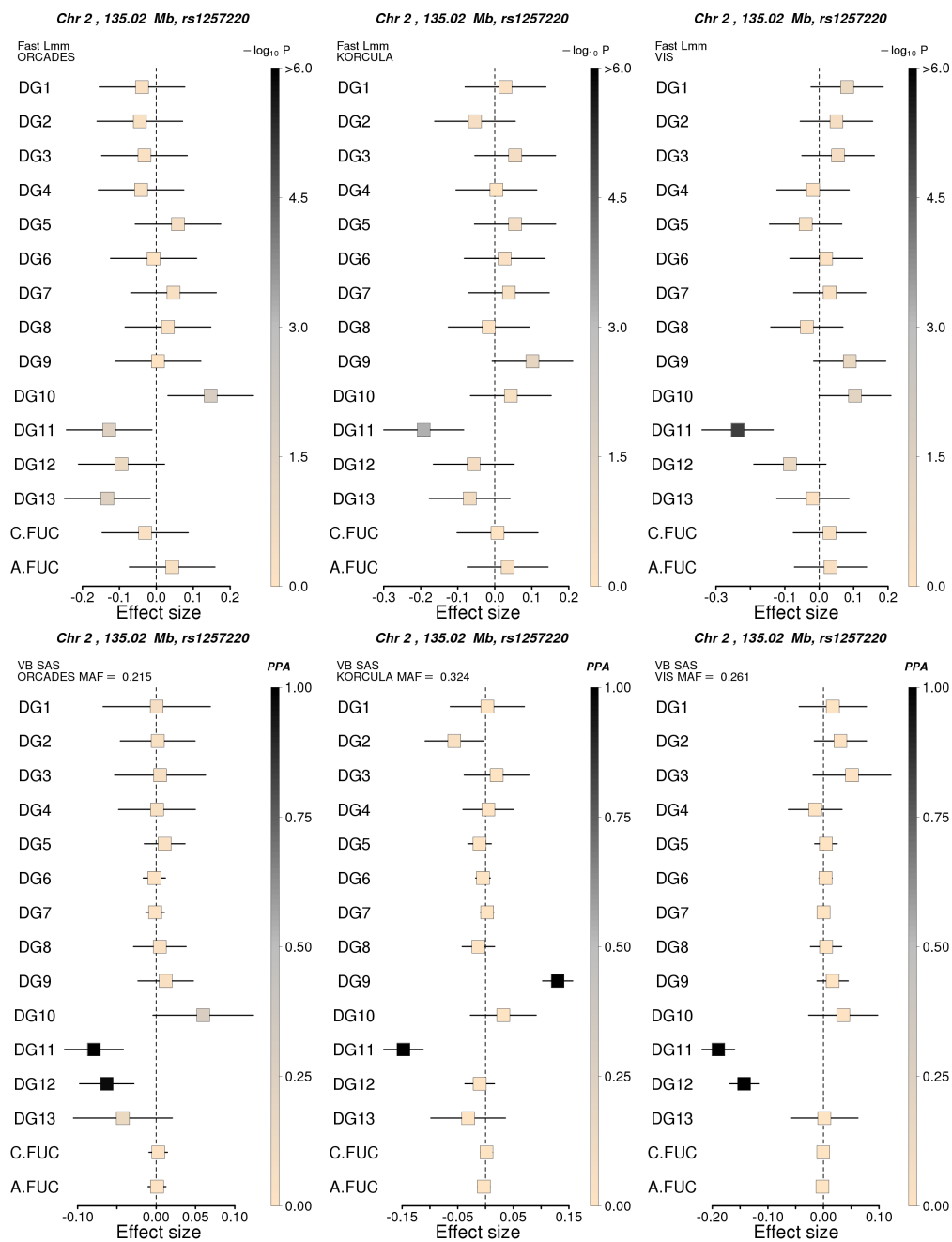


Figure 7.13: Effect size plots, Min p-value approach (top) and VB (bottom) - three datasets, chr2, position 135014685, rs1257218.

7.4 Summary

In this Chapter, we have shown that our VB SAS method can be applied successfully to real data, where it identifies known associations. For the 2q21 locus - with smaller sample size than needed for the original finding. In addition, based on the GWAS manhattan plots, we can observe a number of additional potentially novel loci, however due to time limitations for thesis submission, this analysis is yet to be completed.

Briefly - for the majority of those, there is only signal in one of the three cohorts, which may be due to insufficient power to detect those signals as all of the remaining top variants from VB SAS suggest much smaller effect sizes than for known loci. Furthermore, a large proportion of these variants are common and manhattan plot association signals contain sets of variants in LD with large BF. Often some signal is observed, even though lower (insignificant), for the same cohort from univariate analysis. These signals could potentially be 'real' novel discoveries, however larger sample sizes would be needed for replication.

On the other hand there is also a large set of variants with low MAF (1-5%) which look much less convincing and perhaps some or all of those could reflect false positives. Further and more stringent filtering may be of benefit for our analysis here. For 'real' associations of this type even larger sample sizes than the ones for common variants will most likely be needed for replication.

Finally, a comment on the suitability of the model for the data in general has to be made. VB SAS assumes multivariate normality of the phenotypes and even

though data has been cleaned as much as possible and approximate normality of each trait has been achieved (marginally), multivariate normality of the set of traits is a much stronger assumption in general. Real data is almost always going to deviate to some extent from this assumption and this should be kept in mind when interpreting the analysis results.

Chapter 8

Summary and Discussion

8.1 Summary

Genome-wide association studies have become a key step for targeting genetic variants that can explain variability in disease states and complex traits. They have already provided thousands of such loci to various binary and quantitative phenotypes. Their simplistic modelling in analysing the potential association of a single phenotype on a single SNP, however leave room for improvement. Here we have proposed novel methods that can analyse multiple traits together in a multivariate regression model - an analytic Bayesian model averaging (A) method in Chapter 2 and a Variational Bayesian sparsity inducing (VB SAS) method in Chapter 5. Both of these are aimed at improving power for detection of causal variants, which they can achieve through modelling of phenotype correlations as well as allowing for sparsity in the phenotype-genotype associations.

In Chapter 2 we introduced the model averaging A method framework and described its usage for data analysis of unrelated individuals. The most desirable properties of the method are its analytical tractability and computational effi-

ciency for modest number of phenotypes ($\sim 2 - 10$), as well as its flexibility in terms of prior model specification. The approach possesses the ability to both test for association and distinguish its exact type. These properties are then assessed in Chapter 3 for two phenotypes, where simulation study shows which are the most powerful scenarios (combinations of effect sizes and error correlations) and when in general are multivariate approaches more advantageous in comparison to univariate ones. In particular, in the presence of residual correlation between phenotypes, power can be gained over simpler approaches when (i) only one of a group of correlated traits is associated to a SNP or (ii) if the direction of effects of the associated traits is not the same as the direction of the noise correlation. In addition, the larger this difference, the bigger boost in power is observed. On the contrary, for cases in which the genotypic association to one phenotype is entirely explained through the other one, the univariate approaches would be most powerful.

Next, in Chapter 4 we applied our A method to data on 6 haematological parameters in individuals from three population cohorts, where we discovered a locus (RCL1) not identified in the original study [Soranzo et al., 2009] with the same data, but was later discovered in a study with a much larger sample size [Gieger et al., 2011]. In addition, through our discrete model structure we can observe the evidence for each phenotype bi-partition and can potentially distinguish independent signals, e.g. AK3 and RCL1 - where they both show association to a SNP, but the signal appears to be different. Related to this, our approach can also distinguish and test for evidence of pleiotropy.

Despite the desirable properties of the A method, it comes with its limitations: (i) it does not scale well with number of traits (it requires 2^P models to be fitted at every SNP for P phenotypes); (ii) it is developed only for analysis of unrelated samples and analytically tractable extension is not possible (to the best of my knowledge); (iii) it cannot accommodate missing phenotype data, which is a real feature of most cohort data.

To address all above-mentioned limitations, we have developed and introduced in Chapter 5 a Variational Bayesian method using sparsity inducing spike and slab (SAS) priors on the effect size coefficients. It is a multivariate extension to currently used mixed model approaches and can accommodate (i) large number of traits (tens to hundreds, depending on levels of relatedness); (ii) both unrelated and related samples; (iii) it can internally model and impute missing phenotype data; (iv) through the SAS prior it still keeps the discrete structure of the models and can be viewed as model averaging over individual association models; (v) it provides the ability to both perform inference through the computation of approximate Bayes factors, as well as distinguish its exact type through posterior probabilities of association of each trait (as part of the model).

In addition, the SAS prior can be used to introduce sparsity in the model to control for the level of pleiotropy in the data, i.e. we can put larger prior on models with small number of traits associated, which agrees with currently most commonly observed “modular” pleiotropy (as opposed to “universal” one).

In Chapter 5 we introduce the computational algorithm, VB updates, as well as the computational complexity of the method. Using techniques as in [Lippert et al., 2011],

we reduce the $O(N^3)$ complexity to $O(N)$ after an initial spectral decomposition step of the kinship matrix. With regard to the number of phenotypes, the most expensive steps are the computations of inverses and determinants at the cost of $O(P^3)$ operations each. When analysing data on related individuals we introduce ($M \leq P$) random effects to model the population structure and so the computational speed of the algorithm depends on the underlying sample relatedness structure. Thus, data with few different underlying genetic processes would be faster to model than data in which every trait needs a separate random effect as the number of parameters in our model space would be smaller.

Next, in Chapter 6 we conduct another simulation study, this time on small (2), modest (10) and large (100) number of traits and compute power for detecting association for a wide range of methods. Results for the VB SAS approach agree with those from the A one (as analysed before in Chapter 3) and in general both methods have similar power for modest number of traits. The best improvement in power for the VB SAS method comes when dealing with larger number of correlated phenotypes in a sparse phenotype-genotype network, which we believe to reflect biology better.

Afterwards, we illustrate the advantage of missing phenotype data imputation - a real feature of most studies. Then we conclude the Chapter with discussion on choice of number of random effects and accounting for population structure. We choose M , the number of random effects, based on DIC, a Bayesian method for model comparison, which appears to perform well in simple simulated data.

Finally, in the previous Chapter, we apply the VB SAS method to data on 15 glycans phenotypes in three cohorts of isolated populations. These datasets contain both correlated phenotypes and correlated individuals and in addition have around 5% of samples missing at least one phenotype value. We recover known association signals from the original conducted study [Lauc et al., 2010] and in addition uncover a locus discovered in a larger study [Huffman et al., 2011]. Unfortunately due to time restrictions in submitting this thesis, we did not carry out as much testing of the method on real data as we would have liked to. In particular further analysis of the glycans data has not yet been performed, however a number of potentially novel loci from our VB SAS approach appear in our genome-wide manhattan plots and would need to be followed up.

8.2 Discussion

8.2.1 Alternative Statistics for Association

The obvious way to estimate the Bayes Factor when using the VB SAS approach is the ratio of the lower bounds on the marginal likelihoods of the VB SAS model and the null model M_0 respectively as

$$BF_0 = \frac{\exp(F(Q(\hat{\theta}|SAS)))}{\exp(F(Q(\hat{\theta}|M_0)))} \quad (8.1)$$

However, the lower bound in the numerator for the VB SAS model is meant to approximate $P(D|M_1)$ where M_1 is the model that assumes at least one of the traits is associated. Use of the SAS prior means that there is a point mass of prior probability $P(M_0)$ at the null model where all traits are unassociated. In

8. Summary and Discussion

other words, the numerator is approximating a weighted mixture of $P(D|M_1)$ and $P(D|M_0)$ as

$$BF_0 \approx \frac{P(D|M_0)P(M_0) + P(D|M_1)(1 - P(M_0))}{P(D|M_0)} \quad (8.2)$$

$$= P(M_0) + \frac{P(D|M_1)}{P(D|M_0)}(1 - P(M_0)) \quad (8.3)$$

Adjusted Bayes Factor A suitable alternative to adjust for this difference is to calculate an adjusted Bayes Factor as

$$\frac{P(D|M_1)}{P(D|M_0)} \approx \frac{BF_0 - P(M_0)}{1 - P(M_0)} = BF_1 \quad (8.4)$$

where under our prior the probability of no traits being associated, $P(M_0)$, is

$$\begin{aligned} P(M_0) = \mathbb{E}[(1 - \pi)^P] &= \int_0^1 (1 - \pi)^P \frac{\Gamma(b_1 + b_2)}{\Gamma(b_1)\Gamma(b_2)} \pi^{b_1-1} (1 - \pi)^{b_2-1} d\pi \\ &= \frac{\Gamma(b_1 + b_2)\Gamma(P + b_2)}{\Gamma(b_2)\Gamma(P + b_1 + b_2)} \end{aligned} \quad (8.5)$$

Using Posterior Probability of Association We can also use an alternative method of calculating a Bayes factor from the VB SAS models. Using the VB model fit we can calculate the posterior probability of at least one trait being associated (PPA) as

$$PPA = 1 - \prod_{q=1}^P (1 - \pi_q^*) \quad (8.6)$$

8. Summary and Discussion

and the posterior log odds ($pLOD$) of association are given by

$$pLOD = \log_{10} \frac{PPA}{1 - PPA} = \log_{10} \left[\frac{1}{\prod_{q=1}^P (1 - \pi_q^*)} - 1 \right] \quad (8.7)$$

where π_q^* is the VB posterior probability that trait q is associated and $\rho_q = \log_{10} \frac{\pi_q^*}{1 - \pi_q^*}$ (see Appendices G and H). When at least one of the π_q^* 's tends to 1 then $pLOD \rightarrow \infty$. In this case (say when at least one $\rho_q > 3$) we can approximate the $pLOD$ using

$$pLOD = \log_{10} \left[\frac{1}{\prod_{q=1}^P (1 - \pi_q^*)} \right] \quad (8.8)$$

$$= \sum_{q=1}^P \rho_q + \log_{10}(1 + 10^{-\rho_q}) \quad (8.9)$$

Since the log posterior odds is related to the Bayes Factor via

$$pLOD = \log_{10} BF + \log_{10} PO \quad (8.10)$$

$$PO = \frac{1 - P(M_0)}{P(M_0)} \quad (8.11)$$

where $P(M_0)$ is given by Eq. 8.5, we can derive an alternative estimate of the $\log_{10}$ Bayes Factor (BF_2) using

$$\log_{10} BF_2 = pLOD - \log_{10} PO \quad (8.12)$$

This calculation avoids explicitly evaluating the PPA which may become rounded up to 1.

Using MS statistics [Zhou and Stephens, 2014] discuss the use and derivation of the Wald test statistic

$$\sqrt{\hat{\alpha}V(\hat{\alpha})^{-1}\hat{\alpha}^T} \quad (8.13)$$

in a multivariate mixed effects model. A version of this can be employed in our approach, by replacing the maximum likelihood estimates of the model parameters with their posterior estimates from the VB SAS approach. No asymptotic properties would be available in this case, however permutation testing can be used to infer the significance of the resulting test statistics.

8.2.2 Method Extensions

Missing data Currently VB SAS models missing phenotype data under the null model, after which it is imputed and fixed for association testing. The VB SAS framework allows for missing phenotype data to be treated as a random variable in our alternative model as well, in which case imputation uncertainty can be taken into account by not explicitly fixing the data. We have briefly investigated this approach and found very small differences in the resulting Bayes Factors on a $\log_{10}$ scale, at a much greater computational cost.

Spike and Slab Our method uses a spike and slab prior with a zero mean effect size. This can be generalized and extended further to

$$\alpha_q \sim \frac{\pi}{2}N(\alpha_q|\mu, \sigma_g^2) + \frac{\pi}{2}N(\alpha_q|-\mu, \sigma_g^2) + (1-\pi)\delta(\alpha_q) \quad q \in 1, \dots, P \quad (8.14)$$

for a non-zero μ . This would put a larger prior on a non-zero mean effect size (for the same π) and moreover move the 'slab' away from zero so it targets the effect

sizes (given association) and not compete with the 'spike' when no association is present. We think this could lead to better power for detecting association.

Multiple SNPs Considering multiple genetic variants jointly is of interest as it may be useful in getting a better understanding of complex biological systems, since for complex traits there will be many SNPs that contribute to the disease risk. It has been shown that jointly modelling multiple SNPs can lead to increased power to detect new variants [Listgarten et al., 2012, Segura et al., 2012] and recent work in this area includes application of Lasso-type estimators [Hoggart et al., 2008] as well as MCMC approaches [Guan and Stephens, 2011] and VB approaches [Logsdon et al., 2010]. An extension of the VB SAS to analyse multiple SNPs simultaneously in addition to the multiple phenotypes is also possible and using the VB framework it is very easy to implement.

Multiple cohorts In study designs, data is often collected on multiple populations (e.g. KORCULA, VIS, ORCADES). To analyse such data appropriately the VB SAS approach can readily be extended in a hierarchical fashion to accommodate and model such data.

Study of polygenic effects Another interesting aspect in terms of our current model specification would be to investigate the number of distinct polygenic bases of all traits under consideration. Inferring the rank of the polygenic effects $S\beta$ would be interesting in order to better understand the common genetic bases of different traits. In addition it would be interesting to investigate and characterise the differences between the $S\beta$ and Ω terms when there is a SNP in the model, compared to the null, as this contains information on how a single SNP (or

multiple ones if present in the model) affects the ratio between the genetic and non-genetic phenotypic variances.

Mixtures of types of phenotypes So far we have focussed exclusively on analysis of quantitative traits, whereas many studies will have measured phenotypes that are binary, counts, survival measures and ordered and unordered factors. Future work could involve extending our approach to handle mixtures of phenotypes of different types. Some work on analysis of a single binary and single continuous phenotypes exists in the literature via the use of probit models [Liu et al., 2009, Hahn et al., 2010]. Others have shown that treating binary phenotypes as if they were continuous phenotypes works well [Kang et al., 2010].

8.3 Conclusions

In this thesis, we presented two methods for multivariate phenotype analysis in genome-wide associations studies and highlighted their advantages over simpler uni-trait approaches as well as simpler competing methods for this type of problem. We tested them in terms of power assessment via simulations and applicability to and performance on real data. The exploration of different approaches to this problem is important and it is an ongoing research area, since multivariate approaches are still under-appreciated and not commonly used when at the same time they could greatly further current understanding of phenotype-genotype interactions. Lots of challenges lie ahead, but we look forward to meet them and try to understand the hidden information in the data.

Appendix A - Distributions

Multivariate Normal

Let $\mathbf{x}$ be a $1 \times k$ real-valued (random) vector such that $\mathbf{x} \sim N(\mathbf{m}, \Sigma)$, where $\mathbf{m}$ is a $1 \times k$ vector of means ($\mathbb{E}(\mathbf{x}) = \mathbf{m}$) and Σ is a $k \times k$ (positive-definite) covariance matrix. Then $\mathbf{x}$ has the following density function:

Multivariate Normal Density

$$\mathbb{P}(\mathbf{x}) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{m})\Sigma^{-1}(\mathbf{x} - \mathbf{m})^T\right\}$$

Conditional distributions

Let $\mathbf{x} \sim N(\mathbf{m}, \Sigma)$, $\mathbf{x} = (\mathbf{x}_1 \ \mathbf{x}_2)$ and $\mathbf{m} = (\mathbf{m}_1 \ \mathbf{m}_2)$, where $\mathbf{x}_i$ and $\mathbf{m}_i$ are $1 \times k_i$ vectors and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ with Σ_{ij} - $k_i \times k_j$ matrices. Then we have

- $\mathbf{x}_1 \sim N(\mathbf{m}_1, \Sigma_{11})$
- $\mathbf{x}_2 | \mathbf{x}_1 = \mathbf{a} \sim N(\mathbf{m}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{a} - \mathbf{m}_1), \Sigma^*)$,
where $\Sigma^* = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Matrix Normal

Let $\mathbf{Y}$ be an $n \times k$ real-valued (random) matrix such that $\mathbf{Y} \sim \mathcal{N}(\mathbf{M}, \mathbf{V}, \mathbf{\Sigma})$, where $\mathbf{M}$ is an $n \times k$ matrix of means ($\mathbb{E}(\mathbf{Y}) = \mathbf{M}$), $\mathbf{V}$ is an $n \times n$ (row) covariance matrix and $\mathbf{\Sigma}$ is a $k \times k$ (column) covariance matrix. Then $\mathbf{Y}$ has the following density function:

Matrix Normal Density

$$\mathbb{P}(\mathbf{Y}) = (2\pi)^{-\frac{nk}{2}} |\mathbf{V}|^{-\frac{k}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \text{tr}[\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{M})^T]\right\}$$

Conditional distributions

Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{M}, \mathbf{I}, \mathbf{\Sigma})$, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ and $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2)$, where $\mathbf{Y}_i$ and $\mathbf{M}_i$ are $n \times k_i$ matrices and $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$ with $\mathbf{\Sigma}_{ij}$ - $k_i \times k_j$ matrices. Then we have

- $\mathbf{Y}_1 \sim \mathcal{N}(\mathbf{M}_1, \mathbf{I}, \mathbf{\Sigma}_{11})$
- $\mathbf{Y}_2 | \mathbf{Y}_1 = \mathbf{A} \sim \mathcal{N}(\mathbf{M}_2 + (\mathbf{A} - \mathbf{M}_1)\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}, \mathbf{I}, \mathbf{\Sigma}^*),$
where $\mathbf{\Sigma}^* = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$

Inverse Wishart

Let Σ be a $k \times k$ real-valued (symmetric) positive-definite (random) matrix such that $\Sigma \sim IW_k(c, \mathbf{Q})$. Then Σ has the following density function:

Inverse Wishart Density

$$\mathbb{P}(\Sigma) = \frac{2^{-kc/2}}{\Gamma_k[c/2]} |\mathbf{Q}|^{c/2} |\Sigma|^{-(c+k+1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\Sigma^{-1} \mathbf{Q})\right\}$$

Expectation

$$\mathbb{E}(\Sigma) = \frac{\mathbf{Q}}{c - k - 1}$$

Marginal and conditional distributions

Let $\Sigma \sim IW_k(c, \mathbf{Q})$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ and $\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}$, where Σ_{ij} and $\mathbf{Q}_{ij}$ are $k_i \times k_j$ matrices. Then we have

- Σ_{11} is independent of $\Sigma' = \Sigma_{11}^{-1} \Sigma_{12}$ and $\Sigma^* = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$
- $\Sigma_{11} \sim IW_{k_1}(c - 2k_2, \mathbf{Q}_{11})$
- $\Sigma^* \sim IW_{k_2}(c - k_1, \mathbf{Q}^*)$, where $\mathbf{Q}^* = \mathbf{Q}_{22} - \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}$
- $\Sigma' | \Sigma^* \sim N(\mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}, \mathbf{Q}_{11}^{-1}, \Sigma^*)$

Relationship to Inverse Gamma distribution

The $IW_k(c, \mathbf{Q})$ distribution is an extension of the $IG(a, b)$ distribution to a matrix valued variables, so when $\mathbf{Q}$ is a scalar (i.e. when $k = 1$) IW reduces to IG for

some values of the parameters. If $\sigma^2 \sim IG(a, b)$ then

$$P(\sigma^2) = \frac{1}{\Gamma[a]} b^a (\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}, \quad \mathbb{E}(\sigma^2) = \frac{b}{a-1} \quad (15)$$

$$V(\sigma^2) = \frac{b^2}{(a-1)^2(a-2)}, \quad \mathbb{E}(\sigma^{-2}) = \frac{a}{b} \quad (16)$$

then $\sigma^2 \sim IW_1(2a, 2b)$. In other words $IW_1(2a, 2b) = IG(a, b)$.

Wishart Density

Let Σ be a $k \times k$ real-valued (symmetric) positive-definite (random) matrix such that $\Sigma \sim W_k(\nu, Q)$ then Σ has the following density function:

$$P(\Sigma) = \frac{2^{-k\nu/2}}{\Gamma_k[\nu/2]} |\mathbf{Q}|^{-\nu/2} |\Sigma|^{(\nu-k-1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\mathbf{Q}^{-1}\Sigma)\right\} \quad (17)$$

Student t distribution

The generalised t -distribution is given by

$$t_v(x|\mu, \sigma^2) = c \left[1 + \frac{1}{v} \left(\frac{x - \mu}{\sigma} \right)^2 \right]^{-\left(\frac{v+1}{2}\right)} \quad (18)$$

$$c = \frac{\Gamma(v/2 + 1/2)}{\Gamma(v/2)} \frac{1}{\sqrt{v\pi\sigma^2}} \quad (19)$$

with properties

$$\mathbb{E}(X) = \mu \quad (20)$$

$$V(X) = \frac{v\sigma^2}{(v-2)} \quad (21)$$

Multivariate t distribution

The multivariate T distribution in k dimensions is given by

$$T_v(x|\mu, \Sigma) = c \left[1 + \frac{1}{v} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]^{-\left(\frac{v+1}{2}\right)} \quad (22)$$

$$c = \frac{\Gamma(v/2 + k/2) |\Sigma|^{-1/2}}{\Gamma(v/2) v^{k/2} \pi^{k/2}} \quad (23)$$

with properties

$$\mathbb{E}(X) = \mu \quad \text{if } v > 1 \quad (24)$$

$$V(X) = \frac{v}{(v-2)} \Sigma \quad \text{if } v > 2 \quad (25)$$

Appendix B - Model averaging

(A) method priors

Inverse Gamma $IG(a, b)$ and Inverse Wishart $IW(c, \mathbf{Q})$ distributions, where a, b, c are scalars and $\mathbf{Q}$ is a $k \times k$ positive definite matrix. The IW distribution is an extension of the IG distribution to a matrix valued variables, so when $\mathbf{Q}$ is a scalar (i.e. when $k = 1$) IW reduces to IG for some values of the parameters.

$IW(c, \mathbf{Q})$ to $IG(a, b)$ when $k = 1$

Suppose $\Sigma \sim IW(c, \mathbf{Q})$ and $\sigma^2 \sim IG(a, b)$.

When $k = 1$, $\Sigma = \sigma^2$ for some parameters c and $\mathbf{Q}$.

$$\mathbb{P}(\sigma^2) = \frac{1}{\Gamma[a]} b^a (\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

$$\text{Appendix } \mathbb{P}(\Sigma) = \frac{2^{-kc/2}}{\Gamma_k[c/2]} |\mathbf{Q}|^{c/2} |\Sigma|^{-(c+k+1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\Sigma^{-1} \mathbf{Q})\right\}$$

$$\text{When } k = 1 \text{ the latter becomes } \mathbb{P}(\Sigma) = \frac{2^{-c/2}}{\Gamma_1[c/2]} \mathbf{Q}^{c/2} |\Sigma|^{-(c+2)/2} \exp\left\{-\frac{1}{2} \Sigma^{-1} \mathbf{Q}\right\}$$

Equating the parameters, we get $\boxed{Q = 2b}$ and $\boxed{c = 2a}$. So when $k = 1$
 $IW(c, Q)$ reduces to $IG(\frac{c}{2}, \frac{Q}{2})$. [Or $IG(a, b) = IW(2a, 2b)$.]

One phenotype model

Suppose we have a single phenotype $\mathbf{P}$, mean-centered and scaled so that it has mean 0 and variance 1, and a mean-centered genotype $\mathbf{G}$ at a single SNP.

Model $\mathbf{P} = \gamma \mathbf{G} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$.

Priors $\gamma | \sigma^2 \sim N(\mu, v\sigma^2)$, $\sigma^2 \sim IG(a, b)$

We'd expect σ^2 to be around 1 ($= \text{Var}(\mathbf{P})$) as we expect to see small or no genotype effects. So we want the prior distribution of σ^2 to have mean 1, i.e. $\frac{b}{a-1} = 1$, which implies $\boxed{b = a - 1}$.

The variance of a $IG(a, b)$ random variable is $\frac{b^2}{(a-1)^2(a-2)}$.
(And we want $a > 2$ in order for the variance to be positive.)

It is natural to take $\text{Var}(\sigma^2) = 1$, i.e.

$$1 = \frac{b^2}{(a-1)^2(a-2)} = \frac{(a-1)^2}{(a-1)^2(a-2)} = \frac{1}{(a-2)}, \text{ which gives } \boxed{a = 3} \text{ and } \boxed{b = 2}.$$

For μ and v take $\boxed{\mu = 0}$ and $\boxed{v = 0.02}$ as in page 7 from [Marchini, 2010]
{<http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.v2.pdf>}

Multiple phenotypes model

Phenotypes P ($P_1, P_2, \dots, P_k$) mean-centered and scaled (separately) so that each one of them has mean zero and variance 1. G - genotype at a single SNP.

Model $\mathbf{P} = \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma})$.

Priors $\boldsymbol{\alpha}|\boldsymbol{\Sigma} \sim N(\mathbf{M}, v\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \sim IW(c, \mathbf{Q})$

When $k = 1$, $IW(c, \mathbf{Q}) = IG(\frac{c}{2}, \frac{\mathbf{Q}}{2})$, so fix $\frac{c}{2} = a$ (by Section 1), i.e. fix $\boxed{c = 2a}$.

We would expect the diagonal entries of $\boldsymbol{\Sigma}$ - Σ_{ii} to be around 1 ($= \text{var}(\mathbf{P}_i)$) as we expect to see small or no effects. So we want the prior distribution of $\boldsymbol{\Sigma}$ to have mean with diagonal entries all equal to 1, i.e. $\frac{\mathbf{Q}_{ii}}{c-2} = \frac{\mathbf{Q}_{ii}}{2a-2} = 1$, which implies $\boxed{\mathbf{Q}_{ii} = 2(a-1) = 2b}$ for all i .

All non-diagonal elements of $\mathbf{Q}$ take to be zero as we would like to allow for both positive and negative correlations in the model.

Hence we get $\boxed{c = 2a}$ and $\boxed{\mathbf{Q} = 2b\mathbf{I}}$, which by the previous section for the single phenotype case gives $\boxed{c = 6 = 5 + k}$ and $\boxed{\mathbf{Q} = 4\mathbf{I}}$.

For $\mathbf{M}$ and v take $\boxed{M_i = 0}$ for all i and $\boxed{v = 0.02}$ as in the one phenotype case.

- One phenotype model priors: $\boxed{\gamma|\sigma^2 \sim N(0, 0.02\sigma^2), \sigma^2 \sim IG(3, 2)}$
- Multiple phenotypes model priors: $\boxed{\boldsymbol{\alpha}|\boldsymbol{\Sigma} \sim N(\mathbf{0}, 0.02\boldsymbol{\Sigma}), \boldsymbol{\Sigma} \sim IW(5 + k, 4\mathbf{I})}$

Appendix C - Marginal Likelihood, Full Model

Without loss of generality in this Appendix assume that $c := c - P + 1$

The marginal likelihood is $\mathbb{P}(\text{Data}|\text{Full}) = \int \int \mathbb{P}(\text{Data}|\boldsymbol{\alpha}, \boldsymbol{\Sigma}) \mathbb{P}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) d\boldsymbol{\alpha} d\boldsymbol{\Sigma}$

with priors for $\boldsymbol{\alpha}|\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ taken to be $N(\mathbf{m}, v\boldsymbol{\Sigma})$ and $IW(c, \mathbf{Q})$, giving

$$\mathbb{P}(\boldsymbol{\alpha}|\boldsymbol{\Sigma}) = (2\pi)^{-\frac{P}{2}} v^{-\frac{P}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \text{tr}[v^{-1}(\boldsymbol{\alpha} - \mathbf{m})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha} - \mathbf{m})^T]\right\} \text{ and}$$

$$\mathbb{P}(\boldsymbol{\Sigma}) = \frac{2^{-P(c+P-1)/2}}{\Gamma_P[(c+P-1)/2]} |\mathbf{Q}|^{(c+P-1)/2} |\boldsymbol{\Sigma}|^{-(c+2P)/2} \exp\left\{-\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q})\right\}.$$

Under the normal linear model

$$\mathbb{P}(\text{Data}|\boldsymbol{\alpha}, \boldsymbol{\Sigma}) = (2\pi)^{-NP/2} |\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{1}{2} \text{tr}((\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T)\right)$$

$$\begin{aligned} \text{So then } \mathbb{P}(\text{Data}|\text{Full}) &= \frac{2^{-P(c+P-1)/2} |\mathbf{Q}|^{(c+P-1)/2}}{(2\pi)^{(NP/2)+(pP/2)} v^{P/2} \Gamma_P[(c+P-1)/2]} \times \\ &\int \int |\boldsymbol{\Sigma}|^{-((c+2P)/2)-(1/2)-(N/2)} \exp\left\{-\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q})\right\} \times \\ &\exp\left\{-\frac{1}{2} \text{tr}[v^{-1}(\boldsymbol{\alpha} - \mathbf{m})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha} - \mathbf{m})^T]\right\} \times \\ &\exp\left(-\frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T\right) d\boldsymbol{\alpha} d\boldsymbol{\Sigma} \end{aligned}$$

By the properties of traces, we have

$$\text{tr}[v^{-1}(\boldsymbol{\alpha} - \mathbf{m})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha} - \mathbf{m})^T] = \text{tr}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha} - \mathbf{m})^T v^{-1}(\boldsymbol{\alpha} - \mathbf{m})] \quad (1) \text{ and}$$

$$tr((\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T) = tr(\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})) \quad (2)$$

and the sum of traces 1 and 2 above is equal to

$$tr \left[\boldsymbol{\Sigma}^{-1} \left[(\boldsymbol{\alpha} - \mathbf{m})^T v^{-1} (\boldsymbol{\alpha} - \mathbf{m}) + (\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha}) \right] \right]$$

We want this of the form $tr \left[\boldsymbol{\Sigma}^{-1} \left[(\boldsymbol{\alpha} - \mathbf{m}^*)^T (v^*)^{-1} (\boldsymbol{\alpha} - \mathbf{m}^*) + C_{\boldsymbol{\alpha}} \right] \right]$, where $C_{\boldsymbol{\alpha}}$ is constant with respect to $\boldsymbol{\alpha}$. We know this can be done since by the properties of conjugate priors the posterior and prior parameters belong to the same family of distributions.

$$\text{Now } (\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \mathbf{G}^T \mathbf{G} \boldsymbol{\alpha} + \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{G} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{G}^T \mathbf{Y} \quad (3)$$

$$\text{and } (\boldsymbol{\alpha} - \mathbf{m})^T v^{-1} (\boldsymbol{\alpha} - \mathbf{m}) = \boldsymbol{\alpha}^T v^{-1} \boldsymbol{\alpha} + \mathbf{m}^T v^{-1} \mathbf{m} - \boldsymbol{\alpha}^T v^{-1} \mathbf{m} - \mathbf{m}^T v^{-1} \boldsymbol{\alpha} \quad (4)$$

$$\text{Let } \mathbf{m}^+ := v^{-1} \mathbf{m} + \mathbf{G}^T \mathbf{Y},$$

$$\text{then } (3)+(4) = \boldsymbol{\alpha}^T (\mathbf{G}^T \mathbf{G} + v^{-1}) \boldsymbol{\alpha} + \mathbf{Y}^T \mathbf{Y} + \mathbf{m}^T v^{-1} \mathbf{m} - \boldsymbol{\alpha}^T \mathbf{m}^+ - (\mathbf{m}^+)^T \quad (5)$$

$$\text{Then take } v^* := (\mathbf{G}^T \mathbf{G} + v^{-1})^{-1},$$

$$\text{giving } (5) = \boldsymbol{\alpha}^T v^{*-1} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{m}^+ - (\mathbf{m}^+)^T \boldsymbol{\alpha} + \mathbf{Y}^T \mathbf{Y} + \mathbf{m}^T v^{-1} \mathbf{m}$$

$$\text{Want to get } \boldsymbol{\alpha}^T v^{*-1} \boldsymbol{\alpha} - (\mathbf{m}^*)^T (v^*)^{-1} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T v^{*-1} \mathbf{m}^* + (\mathbf{m}^*)^T v^{*-1} \mathbf{m}^* + C_{\boldsymbol{\alpha}}$$

$$\text{So if we set } \mathbf{m}^* := v^* \mathbf{m}^+ = (\mathbf{G}^T \mathbf{G} + v^{-1})^{-1} (v^{-1} \mathbf{m} + \mathbf{G}^T \mathbf{Y})$$

$$\text{then } (5) = (\boldsymbol{\alpha} - \mathbf{m}^*)^T v^{*-1} (\boldsymbol{\alpha} - \mathbf{m}^*) + C_{\boldsymbol{\alpha}},$$

$$\text{where } C_{\boldsymbol{\alpha}} = \mathbf{Y}^T \mathbf{Y} + \mathbf{m}^T v^{-1} \mathbf{m} - (\mathbf{m}^*)^T v^{*-1} \mathbf{m}^*$$

Our marginal likelihood then becomes:

$$\begin{aligned} \mathbb{P}(\text{Data}|\text{Full}) &= \frac{2^{-P(c+P-1)/2} |\mathbf{Q}|^{(c+P-1)/2} v^{*P/2}}{(2\pi)^{(NP/2)} v^{P/2} \Gamma_P[(c+P-1)/2]} \int |\boldsymbol{\Sigma}|^{-((c+2P)/2)-(N/2)} \times \\ &\exp\{-\frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{Q})\} \exp\{-\frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}C_{\boldsymbol{\alpha}})\} \times \end{aligned}$$

$$\begin{aligned} & \int \frac{|\Sigma|^{-(1/2)}}{(2\pi)^{(1P/2)} v^{*P/2}} \exp\left\{-\frac{1}{2}\text{tr}[v^{*-1}(\alpha - \mathbf{m}^*)\Sigma^{-1}(\alpha - \mathbf{m}^*)^T]\right\} d\alpha d\Sigma \\ &= \frac{2^{-P(c+P-1)/2} |\mathbf{Q}|^{(c+P-1)/2} v^{*P/2}}{(2\pi)^{(NP/2)} v^{P/2} \Gamma_P[(c+P-1)/2]} \int |\Sigma|^{-((c+2P)/2)-(N/2)} \exp\left\{-\frac{1}{2}\text{tr}(\Sigma^{-1}(\mathbf{Q}+C_\alpha))\right\} d\Sigma, \end{aligned}$$

since the multivariate normal density intergates to 1

And if we let $\mathbf{Q}^* := \mathbf{Q} + C_\alpha$, and $c^* := c + N$ then

$$\begin{aligned} \mathbb{P}(\text{Data}|\text{Full}) &= \frac{|\mathbf{Q}|^{(c+P-1)/2} v^{*P/2} 2^{PN/2} \Gamma_P[(c^* + P - 1)/2]}{|\mathbf{Q}^*|^{(c^*+P-1)/2} (2\pi)^{(NP/2)} v^{P/2} \Gamma_P[(c + P - 1)/2]} \times \\ & \int \frac{2^{-P(c^*+P-1)/2}}{\Gamma_P[(c^* + P - 1)/2]} |\mathbf{Q}^*|^{(c^*+P-1)/2} |\Sigma|^{-((c^*+2P)/2)} \exp\left\{-\frac{1}{2}\text{tr}(\Sigma^{-1}(\mathbf{Q}^*))\right\} d\Sigma \\ &= \frac{|\mathbf{Q}|^{(c+P-1)/2} v^{*P/2} 2^{PN/2} \Gamma_P[(c^* + P - 1)/2]}{|\mathbf{Q}^*|^{(c^*+P-1)/2} (2\pi)^{(NP/2)} v^{P/2} \Gamma_P[(c + P - 1)/2]}, \end{aligned}$$

since the density of $IW(c^*, \mathbf{Q}^*)$ integrates to 1.

$$\text{So } \mathbb{P}(\text{Data}|\text{Full}) = \frac{|\mathbf{Q}|^{(c+P-1)/2} v^{*P/2} \Gamma_P[(c^* + P - 1)/2]}{|\mathbf{Q}^*|^{(c^*+P-1)/2} v^{P/2} \Gamma_P[(c + P - 1)/2]} \frac{1}{\pi^{(NP/2)}}, \text{ where}$$

- $c^* = c + N$
- $v^* = (\mathbf{G}^T \mathbf{G} + v^{-1})^{-1}$
- $\mathbf{m}^* = (\mathbf{G}^T \mathbf{G} + v^{-1})^{-1}(v^{-1}\mathbf{m} + \mathbf{G}^T \mathbf{Y})$
- $\mathbf{Q}^* = \mathbf{Q} + \mathbf{Y}^T \mathbf{Y} + \mathbf{m}^T v^{-1} \mathbf{m} - \mathbf{m}^{*T} v^{*-1} \mathbf{m}^*$

Similarly for a null model we can derive

$$\mathbb{P}(\text{Data}|\text{Null}) = \frac{|\mathbf{Q}|^{(c+P-1)/2} \Gamma_P[(c^* + P - 1)/2]}{|\mathbf{Q}^*|^{(c^*+P-1)/2} \Gamma_P[(c + P - 1)/2]} \frac{1}{\pi^{(NP/2)}}, \text{ where}$$

- $c^* = c + N$
- $\mathbf{Q}^* = \mathbf{Q} + \mathbf{Y}^T \mathbf{Y}$

Appendix D - Bayes Factor, non-Full Models

Without loss of generality in this Appendix assume that $c := c - P + 1$

Suppose that Appendix for a particular model M_i , $\mathbf{Y} = (\mathbf{Y}_+ \mathbf{Y}_-)$, where $\mathbf{Y}_+$ is the set of associated and $\mathbf{Y}_-$ the unassociated phenotypes. Then $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_+ \boldsymbol{\alpha}_-) = (\boldsymbol{\alpha}_+ \mathbf{0})$. Suppose also that $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_+ & \boldsymbol{\Sigma}_{+-} \\ \boldsymbol{\Sigma}_{-+} & \boldsymbol{\Sigma}_- \end{pmatrix}$, $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_+ - \boldsymbol{\Sigma}_{+-}\boldsymbol{\Sigma}_-^{-1}\boldsymbol{\Sigma}_{-+}$ and $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_-^{-1}\boldsymbol{\Sigma}_{-+}$

We can then split $\mathbb{P}(\mathbf{Y}|\mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i)$ into $\mathbb{P}(\mathbf{Y}_-|\mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i) \times \mathbb{P}(\mathbf{Y}_+|\mathbf{Y}_-, \mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i)$, where $\mathbb{P}(\mathbf{Y}_-|\mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i) = \mathbb{P}(\mathbf{Y}_-|\boldsymbol{\Sigma}_-, M_i)$. $\mathbb{P}(\boldsymbol{\alpha}|\boldsymbol{\Sigma}, M_i) = \mathbb{P}(\boldsymbol{\alpha}_-|\boldsymbol{\Sigma}, M_i) \times \mathbb{P}(\boldsymbol{\alpha}_+|(\boldsymbol{\alpha}_- = 0), \boldsymbol{\Sigma}, M_i) = 1 \times \mathbb{P}(\boldsymbol{\alpha}_+|\boldsymbol{\alpha}_-, \boldsymbol{\Sigma}, M_i)$ and $\mathbb{P}(\boldsymbol{\Sigma}) = \mathbb{P}(\boldsymbol{\Sigma}_-, \boldsymbol{\Sigma}_+, \boldsymbol{\Sigma}_{+-})$

Then we can make change of variables from $(\boldsymbol{\Sigma}_-, \boldsymbol{\Sigma}_{+-}, \boldsymbol{\Sigma}_+)$ to $(\boldsymbol{\Sigma}_-, \boldsymbol{\Sigma}', \boldsymbol{\Sigma}^*)$ in the marginal likelihood integral introducing a Jacobian $\mathbf{J}$ for the transformation. Due to the definitions of the new variables, this $\mathbf{J}$ will be only a function of $\boldsymbol{\Sigma}_-$.

By the properties of the distributions (as in Appendix A),

$$\mathbb{P}(\boldsymbol{\alpha}_+|(\boldsymbol{\alpha}_- = 0), \boldsymbol{\Sigma}, M_i) = \mathbb{P}(\boldsymbol{\alpha}_+|(\boldsymbol{\alpha}_- = 0), \boldsymbol{\Sigma}^*, M_i) \text{ and}$$

$$\mathbb{P}(\mathbf{Y}_+|\mathbf{Y}_-, \mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i) = \mathbb{P}(\mathbf{Y}_+|\mathbf{Y}_-, \mathbf{G}, \boldsymbol{\alpha}_+, \boldsymbol{\Sigma}^*, \boldsymbol{\Sigma}', M_i), \text{ giving the marginal likelihood}$$

$$\mathbb{P}(\mathbf{Y}|\mathbf{G}, M_i) = \int \int \mathbb{P}(\mathbf{Y}|\mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i) \mathbb{P}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}, M_i) d\boldsymbol{\alpha} d\boldsymbol{\Sigma}$$

$$= \int \mathbb{P}(\mathbf{Y}_-|\boldsymbol{\Sigma}_-, M_i) \mathbb{P}(\boldsymbol{\Sigma}_-) |\mathbf{J}| d\boldsymbol{\Sigma}_- \times$$

$$\int \int \int \mathbb{P}(\mathbf{Y}_+|\mathbf{Y}_-, \mathbf{G}, \boldsymbol{\alpha}_+, \boldsymbol{\Sigma}^*, \boldsymbol{\Sigma}', M_i) \mathbb{P}(\boldsymbol{\alpha}_+|(\boldsymbol{\alpha}_- = 0), \boldsymbol{\Sigma}^*, M_i) \mathbb{P}(\boldsymbol{\Sigma}^*) \mathbb{P}(\boldsymbol{\Sigma}') d\boldsymbol{\alpha}_+ d\boldsymbol{\Sigma}^* d\boldsymbol{\Sigma}'$$

For a Bayes factor calculation, comparing this model to the null, the first term will cancel out,

so we are only interested in the second one.

Now we find $L := \int \int \mathbb{P}(Y_+ | Y_-, G, \alpha_+, \Sigma^*, \Sigma') \mathbb{P}(\alpha_+ | (\alpha_- = 0), \Sigma^*) \mathbb{P}(\Sigma^*) \mathbb{P}(\Sigma') d\alpha_+ d\Sigma^* d\Sigma'$,

where

- $\mathbb{P}(Y_+ | Y_-, \alpha_+, \Sigma^*, \Sigma') \sim N(M, I, \Sigma')$, where $M = G\alpha + Y_- \Sigma'$
- $\alpha_+ \sim N(0, v\Sigma^*)$
- $\Sigma^* \sim IW(c + (p - P), Q^*)$, let $c^* = c + p - P$
- $\Sigma' | \Sigma^* \sim N(Q', Q_{--}^{-1}, \Sigma^*)$

so that

- $\mathbb{P}(Y_+ | Y_-, \alpha, \Sigma^*, \Sigma') = (2\pi)^{-PN/2} |\Sigma^*|^{-N/2} e^{-\frac{1}{2} \text{tr}((Y_+ - M)\Sigma^{*-1}(Y_+ - M)^T)}$, where $M = G\alpha + Y_- \Sigma'$
- $\mathbb{P}(\alpha) = (2\pi)^{-P/2} v^{-P/2} |\Sigma^*|^{-1/2} e^{-\frac{1}{2} v^{-1} \alpha \Sigma^{*-1} \alpha^T}$
- $\mathbb{P}(\Sigma^*) = \frac{2^{-P(c^* + P - 1)/2}}{\Gamma((c^* + P - 1)/2)} |Q^*|^{(c^* + P - 1)/2} |\Sigma^*|^{-(c^* + 2P)/2} e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} Q^*)}$
- $\mathbb{P}(\Sigma' | \Sigma^*) = (2\pi)^{-P(p - P)/2} |Q_{--}^{-1}|^{-P/2} |\Sigma^*|^{-(p - P)/2} e^{-\frac{1}{2} \text{tr}(Q_{--}(\Sigma' - Q')\Sigma^{*-1}(\Sigma' - Q')^T)}$

First integrate out Σ' :

Collect terms involving Σ' and complete the square to a matrix normal distribution.

Terms containing Σ' : $e^{-\frac{1}{2} \text{tr}((Y_+ - M)\Sigma^{*-1}(Y_+ - M)^T)}$, where $M = G\alpha + Y_- \Sigma'$ and $e^{-\frac{1}{2} \text{tr}(Q_{--}(\Sigma' - Q')\Sigma^{*-1}(\Sigma' - Q')^T)}$

Let $Y = Y_+ - G\alpha$, then $e^{-\frac{1}{2} \text{tr}((Y_+ - M)\Sigma^{*-1}(Y_+ - M)^T)} = e^{-\frac{1}{2} \text{tr}((Y - Y_- \Sigma')\Sigma^{*-1}(Y - Y_- \Sigma')^T)}$

Then $e^{-\frac{1}{2} \text{tr}((Y - Y_- \Sigma')\Sigma^{*-1}(Y - Y_- \Sigma')^T)} e^{-\frac{1}{2} \text{tr}(Q_{--}(\Sigma' - Q')\Sigma^{*-1}(\Sigma' - Q')^T)}$
 $= e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1}[(Y - Y_- \Sigma')^T(Y - Y_- \Sigma') + (\Sigma' - Q')^T Q_{--}(\Sigma' - Q')])}$

Now $(Y - Y_- \Sigma')^T(Y - Y_- \Sigma') + (\Sigma' - Q')^T Q_{--}(\Sigma' - Q')$
 $= \Sigma'^T(Q_{--} + Y_-^T Y_-)\Sigma' - \Sigma'^T(Y_-^T Y + Q_{--} Q') - (Y_-^T Y + Q_{--} Q')^T \Sigma' + Y^T Y + Q'^T Q_{--} Q'$

Complete the square to $(\Sigma' - M_{\Sigma'})^T V_{\Sigma'}^{-1}(\Sigma' - M_{\Sigma'}) + c_{\Sigma'}$

$= \Sigma'^T V_{\Sigma'}^{-1} \Sigma' - \Sigma'^T V_{\Sigma'}^{-1} M_{\Sigma'} - M_{\Sigma'}^T V_{\Sigma'}^{-1} \Sigma' + M_{\Sigma'}^T V_{\Sigma'}^{-1} M_{\Sigma'} + c_{\Sigma'}$

$$\Rightarrow V_{\Sigma'}^{-1} = Q_{--} + Y_-^T Y_-, M_{\Sigma'} = V_{\Sigma'}(Y_-^T Y + Q_{--} Q')$$

$$c_{\Sigma'} = Y^T Y + Q'^T Q_{--} Q' - M_{\Sigma'}^T V_{\Sigma'}^{-1} M_{\Sigma'}$$

$$\Rightarrow \mathbb{P}(Y_+ | Y_-, \alpha, \Sigma^*, \Sigma') \mathbb{P}(\Sigma' | \Sigma^*)$$

$$= (2\pi)^{-PN/2} |\Sigma^*|^{-N/2} (2\pi)^{-P(p-P)/2} |Q_{--}^{-1}|^{-P/2} |\Sigma^*|^{-(p-P)/2} \times$$

$$e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} c_{\Sigma'})} e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} (\Sigma' - M_{\Sigma'})^T V_{\Sigma'}^{-1} (\Sigma' - M_{\Sigma'}))}$$
 and the other terms in the integral are constant in Σ'

$$\text{So that } \int \mathbb{P}(Y_+ | Y_-, \alpha, \Sigma^*, \Sigma') \mathbb{P}(\Sigma' | \Sigma^*) d\Sigma'$$

$$= (2\pi)^{-PN/2} |\Sigma^*|^{-N/2} e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} c_{\Sigma'})} |Q_{--}^{-1}|^{-P/2} |V_{\Sigma'}|^{P/2} \times$$

$$\int (2\pi)^{-P(p-P)/2} |V_{\Sigma'}|^{-P/2} |\Sigma^*|^{-(p-P)/2} e^{-\frac{1}{2} \text{tr}(V_{\Sigma'}^{-1} (\Sigma' - M_{\Sigma'})^T \Sigma^{*-1} (\Sigma' - M_{\Sigma'})^T)} d\Sigma'$$

The integral is equal to 1 since this is the matrix normal density.

$$\Rightarrow L = \frac{|V_{\Sigma'}|^{P/2}}{|Q_{--}^{-1}|^{P/2}} \int (2\pi)^{-PN/2} |\Sigma^*|^{-N/2} e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} c_{\Sigma'})} \mathbb{P}(\alpha | \Sigma^*) \mathbb{P}(\Sigma^*) d\alpha d\Sigma^*$$

Now let us look at $e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} c_{\Sigma'})}$:

$$c_{\Sigma'} = Y^T Y + Q'^T Q_{--} Q' - M_{\Sigma'}^T V_{\Sigma'}^{-1} M_{\Sigma'}, \text{ where } M_{\Sigma'} = V_{\Sigma'}(Y_-^T Y + Q_{--} Q') \text{ and } V_{\Sigma'}^{-1} = Q_{--} + Y_-^T Y_-, \text{ and } Y = Y_+ - G\alpha$$

$$\begin{aligned} \text{So } c_{\Sigma'} &= (Y_+ - G\alpha)^T (Y_+ - G\alpha) + Q'^T Q_{--} Q' - \\ & (Y_-^T (Y_+ - G\alpha) + Q_{--} Q')^T V_{\Sigma'} (Y_-^T (Y_+ - G\alpha) + Q_{--} Q') \\ &= (Y_+ - G\alpha)^T (Y_+ - G\alpha) + Q'^T Q_{--} Q' - (Y_+ - G\alpha)^T Y_- V_{\Sigma'} Y_-^T (Y_+ - G\alpha) - \\ & (Y_+ - G\alpha)^T Y_- V_{\Sigma'} Q_{--} Q' - (Q_{--} Q')^T V_{\Sigma'} Y_-^T (Y_+ - G\alpha) - (Q_{--} Q')^T V_{\Sigma'} Q_{--} Q' \end{aligned}$$

$$\mathbf{1. \text{ if } Q \text{ diagonal: } } c_{\Sigma'} = (Y_+ - G\alpha)^T (I - Y_- V_{\Sigma'} Y_-^T) (Y_+ - G\alpha)$$

And if we let $V = (I - Y_- V_{\Sigma'} Y_-^T)^{-1}$ and decompose $V^{-1} = U^T U$ (cholesky, U-upper triangular), then $c_{\Sigma'} = (Y_+ - G\alpha)^T U^T U (Y_+ - G\alpha)$

$$= (UY_+ - UG\alpha)^T (UY_+ - UG\alpha) = (Y_{new} - G_{new}\alpha)^T (Y_{new} - G_{new}\alpha)$$

Then $e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} c_{\Sigma'})} =$

$$e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1} (Y_{new} - G_{new}\alpha)^T (Y_{new} - G_{new}\alpha))} = e^{-\frac{1}{2} \text{tr}((Y_{new} - G_{new}\alpha) \Sigma^{*-1} (Y_{new} - G_{new}\alpha)^T)}$$

$$\text{And } L = \frac{|V_{\Sigma'}|^{P/2}}{|Q_{--}^{-1}|^{P/2}} \int (2\pi)^{-PN/2} |\Sigma^*|^{-N/2} e^{-\frac{1}{2} \text{tr}((Y_{new} - G_{new}\alpha)\Sigma^{*-1}(Y_{new} - G_{new}\alpha)^T)} \mathbb{P}(\alpha|\Sigma^*) \mathbb{P}(\Sigma^*) d\alpha d\Sigma^*$$

$$\text{Or } L = \frac{|V_{\Sigma'}|^{P/2}}{|Q_{--}^{-1}|^{P/2}} \int \mathbb{P}(Y_{new}|G_{new}, \alpha, \Sigma^*) \mathbb{P}(\alpha|\Sigma^*) \mathbb{P}(\Sigma^*) d\alpha d\Sigma^*$$

Now the factor in front of the integral will cancel out when calculating Bayes factors (assuming we use the same priors on Q_{--} for all models). Then the problem would be equivalent to fitting a full model to the new data (only associated phenotypes), when using the same priors on the effect sizes and the covariance matrix. (Priors are the same since we are using Q diagonal, with the difference that c now would be c+p-P rather than just c.)

$$\mathbf{2. \text{ if } Q \text{ is not diagonal: } } c_{\Sigma'} = (Y_+ - G\alpha - M)^T V^{-1} (Y_+ - G\alpha - M) + C,$$

where $V = (I - Y_- V_{\Sigma'} Y_-^T)^{-1}$ (as before), $M = V Y_- V_{\Sigma'} Q_{--} Q'$ and

$$\begin{aligned} C &= -M^T V^{-1} M + Q'^T Q_{--} Q' - Q'^T Q_{--} V_{\Sigma'} Q_{--} Q' \\ &= Q'^T Q_{--} (Q_{--}^{-1} - V_{\Sigma'} - V_{\Sigma'} Y_-^T V Y_- V_{\Sigma'}) Q_{--} Q' \end{aligned}$$

$$Q' = Q_{--}^{-1} Q_{-+} \Rightarrow C = Q_{-+} (Q_{--}^{-1} - V_{\Sigma'} (V_{\Sigma'}^{-1} - Y_-^T V Y_-) V_{\Sigma'}) Q_{-+}$$

Then again decompose $V^{-1} = U^T U$ and define $Y_{new} = U(Y_+ - M)$ and $G_{new} = UG$ in order to get

$$L = \frac{|V_{\Sigma'}|^{P/2}}{|Q_{--}^{-1}|^{P/2}} \times \int (2\pi)^{-PN/2} |\Sigma^*|^{-N/2} e^{-\frac{1}{2} \text{tr}((Y_{new} - G_{new}\alpha)\Sigma^{*-1}(Y_{new} - G_{new}\alpha)^T)} e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1}C)} \mathbb{P}(\alpha|\Sigma^*) \mathbb{P}(\Sigma^*) d\alpha d\Sigma^*$$

$$\text{Or } L = \frac{|V_{\Sigma'}|^{P/2}}{|Q_{--}^{-1}|^{P/2}} \int e^{-\frac{1}{2} \text{tr}(\Sigma^{*-1}C)} \mathbb{P}(Y_{new}|G_{new}, \alpha, \Sigma^*) \mathbb{P}(\alpha|\Sigma^*) \mathbb{P}(\Sigma^*) d\alpha d\Sigma^*$$

$$= \frac{|V_{\Sigma'}|^{P/2}}{|Q_{--}^{-1}|^{P/2}} \frac{|Q^*|^{(c^*+P-1)/2}}{|Q^* + C|^{(c^*+P-1)/2}} \int \mathbb{P}(Y_{new}|G_{new}, \alpha, \Sigma^*) \mathbb{P}(\alpha|\Sigma^*) \mathbb{P}_{\neq}(\Sigma^*) d\alpha d\Sigma^*,$$

where $\mathbb{P}_{\neq}(\Sigma^*) \sim IW(c+p-P, Q^* + C)$. Here again the term in front of the integral will cancel out when calculating Bayes factors (again assuming we use the same priors for all models).

Again the problem is equivalent to fitting the full model to the new data (only associated phenotypes), when using the same prior on the effect sizes and with a prior on the covariance matrix changed from $IW(c, Q_{++})$ to $IW(c+p-P, Q^* + C)$.

Appendix E - Variational Bayes methods

Here we give some general details of the method of Variational Bayes and in the next two appendices we give full details of the derivations of the VB implementations for sparse Bayesian models that uses the SAS prior on the effect sizes. We consider the cases of unrelated and related samples separately.

Theory of Variational Bayes

KL divergence and marginal likelihoods

The KL divergence between a posterior distribution $P(\theta|\mathbf{D}, M)$ and an approximate posterior $Q(\theta)$ is given by

$$D_{KL}(Q(\theta)||P(\theta|\mathbf{D}, M)) = \int \log \left[\frac{Q(\theta)}{P(\theta|\mathbf{D}, M)} \right] Q(\theta) d\theta \quad (26)$$

$$= \int \log \left[\frac{Q(\theta)P(\mathbf{D}|M)}{P(\theta, \mathbf{D}|M)} \right] Q(\theta) d\theta \quad (27)$$

$$= \log P(\mathbf{D}|M) - \int \log \left[\frac{p(\theta, \mathbf{D}|M)}{Q(\theta)} \right] Q(\theta) d\theta \quad (28)$$

The log marginal likelihood can then be written as

$$\log P(\mathbf{D}|M) = F(Q(\theta)) + D_{KL}(Q(\theta)||P(\theta|\mathbf{D}, M)) \quad (29)$$

where $F(Q(\theta)) = \int \log \left[\frac{P(\theta, \mathbf{D}|M)}{Q(\theta)} \right] Q(\theta) d\theta$. Note that since $D_{KL}(Q(\theta)||P(\theta|\mathbf{D}, M)) \geq 0$ we have that $F(Q(\theta)) \leq \log P(\mathbf{D}|M)$ so that $F(Q(\theta))$ is a lower bound on the log-marginal likelihood.

VB updates

Writing $Q(\theta) = Q(\theta_i)Q(\theta_{-i})$ where $\theta_{-i} = \theta \setminus \theta_i$ the lower-bound can be re-written as

$$F(Q(\theta)) = \int \log \left[\frac{P(\theta, \mathbf{D}|M)}{Q(\theta_i)Q(\theta_{-i})} \right] q(\theta_i)Q(\theta_{-i})d\theta_i d\theta_{-i} \quad (30)$$

$$= \int Q(\theta_i) \left[\int \log P(\mathbf{D}, \theta|M) Q(\theta_{-i}) d\theta_{-i} \right] d\theta_i \quad (31)$$

$$- \int Q(\theta_i) \log Q(\theta_i) d\theta_i - \sum_{j \neq i} \int Q(\theta_j) \log Q(\theta_j) d\theta_j \quad (32)$$

If we let $Q^*(\theta_i) = \frac{1}{Z} \exp \left[\int \log P(\mathbf{D}, \theta|M) Q(\theta_{-i}) d\theta_{-i} \right]$ where Z is a normalising constant and write $H(Q(\theta_j)) = - \int Q(\theta_j) \log Q(\theta_j) d\theta_j$ as the entropy of $Q(\theta_j)$ then

$$F(Q(\theta)) = \int Q(\theta_i) \log \frac{Q^*(\theta_i)}{Q(\theta_i)} d\theta_i + \sum_{j \neq i} H(Q(\theta_j)) + \log Z \quad (33)$$

$$= -D_{KL}(Q(\theta_i)||Q^*(\theta_i)) + \sum_{j \neq i} H(Q(\theta_j)) + \log Z \quad (34)$$

We then see that $F(Q(\theta))$ is maximised when $Q(\theta_i) = Q^*(\theta_i)$ as this choice minimises the Kullback-Liebler divergence term. Thus the update for $Q(\theta_i)$ is given by

$$Q(\theta_i) \propto \exp \left[\int \log P(\mathbf{D}, \theta | M) Q(\theta_{-i}) d\theta_{-i} \right] \quad (35)$$

or

$$\log Q(\theta_i) = \mathbb{E}_{Q(\theta_{-i})} \left(\log P(\mathbf{D}, \theta | M) \right) + \text{const} \quad (36)$$

Appendix F - VB SAS for unrelated samples

The model can be described as

$$\begin{aligned}
 \text{Appendix } \mathbf{Y} &\sim \mathcal{N}(\mathbf{Y} | \mathbf{G}\boldsymbol{\alpha}, \mathbf{I}, \boldsymbol{\Omega}) \\
 \alpha_q &\sim \pi N(\alpha_q | 0, \sigma_g^2) + (1 - \pi)\delta(\alpha_q) \quad q \in 1, \dots, P \\
 \sigma_g^2 &\sim \Gamma^{-1}(d_1, d_2) \\
 \pi &\sim \text{Beta}(b_1, b_2) \\
 \boldsymbol{\Omega} &\sim \text{Wi}(\nu, \mathbf{V})
 \end{aligned}$$

Following [Titsias and Lazaro-Gredilla, 2011] we use a reparameterization of the model that is more amenable to VB inference such that α_q is represented as the product of a *Bernoulli*(π) random variable s_q and a $N(0, \sigma_g^2)$ random variable $\tilde{\alpha}_q$

so that $\alpha_q = \tilde{\alpha}_q s_q, q \in 1, \dots, P$. The relevant log densities are then given by

$$\begin{aligned}
\log P(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\Omega}) &= \frac{N}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T + \text{const} \\
\log P(\sigma_g^2) &= -(d_1 + 1) \log \sigma_g^2 - d_2 \sigma_g^{2-1} + \text{const} \\
\log P(\mathbf{s}) &= \sum_{q=1}^P s_q \log \pi + \sum_{q=1}^P (1 - s_q) \log(1 - \pi) + \text{const} \\
\log P(\pi) &= (b_1 - 1) \log \pi + (b_2 - 1) \log(1 - \pi) + \text{const} \\
\log P(\tilde{\boldsymbol{\alpha}}) &= -\frac{1}{2} \sum_{q=1}^P \log \sigma_g^2 - \frac{1}{2} \sum_{q=1}^P \tilde{\alpha}_q^2 \sigma_g^{2-1} + \text{const} \\
\log P(\boldsymbol{\Omega}) &= \frac{\nu - P - 1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{V}^{-1}\boldsymbol{\Omega}) + \text{const}
\end{aligned}$$

The full set of parameters is $\theta = (\boldsymbol{\Omega}, \{\sigma_q^2\}_{\forall q}, \tilde{\boldsymbol{\alpha}}, \mathbf{s}, \pi)$ and

$$\begin{aligned}
\log P(\mathbf{D}, \theta|M) &= \log P(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\Omega}) + \log P(\tilde{\boldsymbol{\alpha}}) + \log P(\mathbf{s}) \\
&\quad + \log P(\pi) + \log P(\sigma_g^2) + \log P(\boldsymbol{\Omega}) + \text{const}
\end{aligned}$$

We use a factorization for Q of the following form

$$Q(\theta) = Q(\boldsymbol{\Omega}) \prod_{q=1}^P Q(\tilde{\alpha}_q, s_q) Q(\sigma_g^2) Q(\pi)$$

Approximate posterior distribution for $\boldsymbol{\Omega}$

$$\begin{aligned}
\log Q(\boldsymbol{\Omega}) &= \mathbb{E}_{Q(\theta|\mathbf{D})} (\log P(\mathbf{D}, \theta)) + \text{const} \\
&= \frac{\nu - P - 1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{V}^{-1}\boldsymbol{\Omega}) \\
&\quad + \frac{N}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}[\mathbb{E}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha}) \boldsymbol{\Omega}] + \text{const}
\end{aligned}$$

Hence $Q(\boldsymbol{\Omega}) = Wi(\nu^*, \mathbf{V}^*)$, where $\nu^* = \nu + N$ and

$$\mathbf{V}^* = [\mathbf{V}^{-1} + \mathbb{E}[(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})]]^{-1}$$

Approximate posterior distribution for σ_g^2

$$\begin{aligned} \log Q(\sigma_g^2) &= \mathbb{E}_{Q(\theta|\sigma_g^2)}(\log P(\mathbf{D}, \theta)) + \text{const} \\ &= -(d_1 + 1) \log \sigma_g^2 - d_2 \sigma_g^{2-1} - \frac{1}{2} \sum_{q=1}^P \log \sigma_g^2 - \frac{1}{2} \sum_{q=1}^P \mathbb{E}(\tilde{\alpha}_q^2) \sigma_g^{2-1} + \text{const} \end{aligned}$$

Hence $Q(\sigma_g^2) = \Gamma^{-1}(d_1^*, d_2^*)$, where $d_1^* = d_1 + \frac{P}{2}$ and $d_2^* = d_2 + \frac{1}{2} \sum_{q=1}^P \mathbb{E}[\tilde{\alpha}_q^2]$

Approximate posterior distribution for π

$$\begin{aligned} \log Q(\pi) &= \mathbb{E}_{Q(\theta|\pi)}(\log P(\mathbf{D}, \theta)) + \text{const} \\ &= (b_1 - 1) \log(\pi) + (b_2 - 1) \log(1 - \pi) \\ &\quad + \sum_{q=1}^P \mathbb{E}(s_q) \log \pi + \sum_{q=1}^P \mathbb{E}(1 - s_q) \log(1 - \pi) + \text{const} \end{aligned}$$

Therefore $Q(\pi) \sim B(b_1^*, b_2^*)$, where $b_1^* = b_1 + \sum_{q=1}^P \mathbb{E}(s_q)$ and $b_2^* = b_2 + \sum_{q=1}^P (1 - \mathbb{E}(s_q))$

Approximate posterior distribution for $\{\tilde{\alpha}_q, s_q\}$

$$\begin{aligned}\log Q(\tilde{\alpha}_q, s_q) &= \mathbb{E}_{Q(\theta \setminus \{\tilde{\alpha}_q, s_q\})}(\log P(\mathbf{D}, \theta)) + \text{const} \\ &= \mathbb{E} \left[-\frac{1}{2} \text{tr}((\mathbf{Y} - \mathbf{G}\tilde{\alpha}\mathbf{s})\mathbf{\Omega}(\mathbf{Y} - \mathbf{G}\tilde{\alpha}\mathbf{s})^T) \right. \\ &\quad \left. + s_q \log \pi + (1 - s_q) \log(1 - \pi) - \frac{1}{2} \tilde{\alpha}_q^2 \sigma_g^{2-1} \right] + \text{const}\end{aligned}$$

Decompose $Q(\tilde{\alpha}_q, s_q) = Q(\tilde{\alpha}_q | s_q)Q(s_q)$. Then it follows that

$$\log Q(\tilde{\alpha}_q | s_q) = \mathbb{E} \left[-\frac{1}{2} \text{tr}((\mathbf{Y} - \mathbf{G}\tilde{\alpha}\mathbf{s})\mathbf{\Omega}(\mathbf{Y} - \mathbf{G}\tilde{\alpha}\mathbf{s})^T) - \frac{1}{2} \tilde{\alpha}_q^2 \sigma_g^{2-1} \right] + \text{const}$$

Therefore

$$\begin{aligned}Q(\tilde{\alpha}_q | s_q = 1) &\sim N(m_q, v_q) \\ Q(\tilde{\alpha}_q | s_q = 0) &\sim N(0, \mathbb{E}(\sigma_g^{2-1})^{-1})\end{aligned}$$

where $v_q = \left[\mathbf{G}^T \mathbf{G} \mathbb{E}(\mathbf{\Omega})_{q,q} + \mathbb{E}(\sigma_g^{2-1}) \right]^{-1}$
 $m_q = v_q \mathbf{G}^T \left[\mathbb{E}(\mathbf{\Omega})_{q,q} \mathbf{Y}_q + \sum_{i \neq q} \mathbb{E}(\mathbf{\Omega})_{q,i} (\mathbf{Y}_i - \mathbf{G} \mathbb{E}(\alpha_i)) \right]$ and $\mathbf{Y}_q$ denotes the q th column of $\mathbf{Y}$. Then

$$\begin{aligned}\log Q(s_q = 1) &= \log Q(\tilde{\alpha}_q, s_q = 1) - \log Q(\tilde{\alpha}_q | s_q = 1) \\ &= \mathbb{E} \left[-\frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{G}\tilde{\alpha})\mathbf{\Omega}(\mathbf{Y} - \mathbf{G}\tilde{\alpha})^T + \log \pi - \frac{1}{2} \tilde{\alpha}_q^2 \sigma_g^{2-1} \right] \\ &\quad + \frac{1}{2} v_q^{-1} (\tilde{\alpha}_q - m_q)^2 + \frac{1}{2} \log v_q + \text{const} \\ &= \frac{1}{2} v_q^{-1} m_q^2 + \frac{1}{2} \log v_q + \mathbb{E}(\log \pi) + \text{const}\end{aligned}$$

and

$$\begin{aligned}
\log Q(s_q = 0) &= \log Q(\tilde{\alpha}_q, s_q = 0) - \log Q(\tilde{\alpha}_q | s_q = 0) \\
&= \mathbb{E} \left[-\frac{1}{2} \text{tr}(\mathbf{Y} \mathbf{\Omega} \mathbf{Y}^T) + \log(1 - \pi) - \frac{1}{2} \tilde{\alpha}_q^2 \sigma_g^{2-1} \right] \\
&\quad + \frac{1}{2} \mathbb{E}(\sigma_g^{2-1}) (\tilde{\alpha}_q)^2 + \frac{1}{2} \log \mathbb{E}(\sigma_g^{2-1})^{-1} + \text{const} \\
&= \frac{1}{2} \log \mathbb{E}(\sigma_g^{2-1})^{-1} + \mathbb{E}(\log(1 - \pi)) + \text{const}
\end{aligned}$$

If we write $\pi_q^* = Q(s_q = 1)$ and note that the constants in Eqns. (37) and (37) are the same then we have

$$\rho_q = \log \frac{\pi_q^*}{1 - \pi_q^*} = \mathbb{E} \left(\log \frac{\pi}{1 - \pi} \right) + \frac{m_q^2}{2v_q} + \frac{1}{2} \log v_q + \frac{1}{2} \log \left[\mathbb{E}(\sigma_g^{2-1}) \right]$$

and $\pi_q^* = \frac{1}{1 + e^{-\rho_q}}$. Then the joint approximate posterior distribution of $\tilde{\alpha}_q$ and s_q can be written in the following form

$$Q(\tilde{\alpha}_q, s_q) = Q(\tilde{\alpha}_q | s_q) Q(s_q) = \mathcal{N} \left(s_q m_q, s_q v_q + (1 - s_q) \mathbb{E}(\sigma_g^{2-1})^{-1} \right) (\pi_q^*)^{s_q} (1 - \pi_q^*)^{1-s_q}$$

and this is of the same form as the prior. Also the expectation of $\alpha_q = \tilde{\alpha}_q s_q$ is $\pi_q^* m_q$. Additionally let us denote $m_q^* = s_q m_q$ and $v_q^* = s_q v_q + (1 - s_q) \mathbb{E}(\sigma_g^{2-1})^{-1}$.

Approximate marginal likelihood

$$\begin{aligned}
F(Q(\theta)) &= \int \log \left[\frac{P(\theta, \mathbf{D} | M)}{Q(\theta)} \right] Q(\theta) d\theta \\
&= \mathbb{E}[\log P(\mathbf{D}, \theta) - \log Q(\theta)] \\
&= \mathbb{E} \left[-\frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{G}\boldsymbol{\alpha})^T + \frac{N}{2} \log |\boldsymbol{\Omega}| - \frac{NP}{2} \log(2\pi) \right. \\
&\quad - (d_1 + 1) \log \sigma_g^2 - d_2 \sigma_g^{2-1} + d_1 \log d_2 - \log \Gamma(d_1) \\
&\quad + \sum_{q=1}^P s_q \log \pi + \sum_{q=1}^P (1 - s_q) \log(1 - \pi) - \sum_{q=1}^P s_q \log \pi_q^* - \sum_{q=1}^P (1 - s_q) \log(1 - \pi_q^*) \\
&\quad - \frac{1}{2} \sum_{q=1}^P \log \sigma_g^2 - \frac{1}{2} \sum_{q=1}^P \tilde{\alpha}_q^2 \sigma_g^{2-1} - \cancel{\frac{P}{2} \log(2\pi)} \\
&\quad + (b_1 - 1) \log(\pi) + (b_2 - 1) \log(1 - \pi) - \log B(b_1, b_2) \\
&\quad + (d_1^* + 1) \log \sigma_g^2 + d_2^* \sigma_g^{2-1} - d_1^* \log d_2^* + \log \Gamma(d_1^*) \\
&\quad + \frac{1}{2} \sum_{q=1}^P \log v_q^* + \frac{1}{2} \sum_{q=1}^P (\tilde{\alpha}_q - m_q^*)^2 v_q^{*-1} + \cancel{\frac{P}{2} \log(2\pi)} \\
&\quad - (b_1^* - 1) \log(\pi) - (b_2^* - 1) \log(1 - \pi) + \log B(b_1^*, b_2^*) \\
&\quad + \frac{\nu - P - 1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{V}^{-1} \boldsymbol{\Omega}) - \frac{\nu P}{2} \log 2 - \frac{\nu}{2} \log |\mathbf{V}| - \log \Gamma_P\left(\frac{\nu}{2}\right) \\
&\quad \left. - \frac{\nu^* - P - 1}{2} \log |\boldsymbol{\Omega}| + \frac{1}{2} \text{tr}(\mathbf{V}^{*-1} \boldsymbol{\Omega}) + \frac{\nu^* P}{2} \log 2 + \frac{\nu^*}{2} \log |\mathbf{V}^*| + \log \Gamma_P\left(\frac{\nu^*}{2}\right) \right] \\
&= \frac{1}{2} \text{tr} \left[\left(\mathbf{V}^{*-1} - \mathbf{V}^{-1} - \mathbf{Y}^T \mathbf{Y} + \mathbf{Y}^T \mathbf{G} \boldsymbol{\gamma} + \boldsymbol{\gamma}^T \mathbf{G}^T \mathbf{Y} - \mathbf{G}^T \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha}) \right) \nu^* \mathbf{V}^* \right] \\
&\quad + \left(d_2^* - d_2 - \frac{1}{2} \sum_{q=1}^P \mathbb{E}(\tilde{\alpha}_q^2) \right) \frac{d_1^*}{d_2^*} + d_1 \log d_2 - d_1^* \log d_2^* - \log \Gamma(d_1) + \log \Gamma(d_1^*) + \frac{P}{2} \\
&\quad - \frac{\nu}{2} \log |\mathbf{V}| + \frac{\nu^*}{2} \log |\mathbf{V}^*| - \log \Gamma_P\left(\frac{\nu}{2}\right) + \log \Gamma_P\left(\frac{\nu^*}{2}\right) - \log B(b_1, b_2) + \log B(b_1^*, b_2^*) \\
&\quad - \sum_{q=1}^P \pi_q^* \log \pi_q^* - \sum_{q=1}^P (1 - \pi_q^*) \log(1 - \pi_q^*) + \frac{1}{2} \sum_{q=1}^P \left(\pi_q^* \log v_q - (1 - \pi_q^*) \frac{d_1^*}{d_2^*} \right) \\
&\quad + (b_1 - b_1^* + \sum_{q=1}^P \pi_q^*) \mathbb{E}(\log \pi) + (b_2 - b_2^* + \sum_{q=1}^P (1 - \pi_q^*)) \mathbb{E}(\log(1 - \pi)) - \frac{NP}{2} \log(\pi)
\end{aligned}$$

Parameter updates At each iteration of the algorithm the parameters of VB approximation are updated and various moments need to be calculated for the marginal likelihood approximation. The details are as follows

$$\begin{aligned}
d_1^* &= d_1 + \frac{P}{2} \\
d_2^* &= d_2 + \frac{1}{2} \sum_{q=1}^P \left[\pi_q^* (v_q + m_q^2) + (1 - \pi_q^*) \frac{d_2^*}{d_1^*} \right] \\
v_q &= \left[\mathbf{G}^T \mathbf{G} \nu^* \mathbf{V}_{q,q}^* + \frac{d_1^*}{d_2^*} \right]^{-1} \\
m_q &= v_q \nu^* \mathbf{G}^T \left[\mathbf{V}_{q,q}^* \mathbf{Y}_q + \sum_{i \neq q} \mathbf{V}_{q,i}^* (\mathbf{Y}_i - \mathbf{G} \pi_i^* m_i) \right] \\
\rho_q &= \psi(b_1^*) - \psi(b_2^*) + \frac{1}{2} \log \frac{d_1^*}{d_2^*} + \frac{m_q^2}{2v_q} + \frac{1}{2} \log v_q \\
\pi_q^* &= \frac{1}{1 + e^{-\rho_q}} \\
b_1^* &= b_1 + \sum_{q=1}^P \pi_q^* \\
b_2^* &= b_2 + P - \sum_{q=1}^P \pi_q^* \\
\nu^* &= \nu + N \\
\mathbf{V}^* &= [\mathbf{V}^{-1} + \mathbf{Y}^T \mathbf{Y} - \boldsymbol{\gamma}^T \mathbf{G}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{G} \boldsymbol{\gamma} + \mathbf{G}^T \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha})]^{-1} \\
\boldsymbol{\gamma}_q &= m_q * \pi_q \\
\mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha})_{qr} &= \mathbb{E}(\alpha_q \alpha_r) \\
\mathbb{E}(\alpha_q \alpha_q) &= \pi_q (v_q + m_q^2) \\
\mathbb{E}(\alpha_q \alpha_r) &= \pi_q \pi_r m_q m_r \quad q \neq r \\
\mathbb{E}(\tilde{\alpha}_q^2) &= \pi_q^* (v_q + m_q^2) + (1 - \pi_q^*) \frac{d_2^*}{d_1^*}
\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\log(\pi)) &= \psi(b_1^*) - \psi(b_1^* + b_2^*) \\ \mathbb{E}(\log(1 - \pi)) &= \psi(b_2^*) - \psi(b_1^* + b_2^*)\end{aligned}$$

where ψ denotes the digamma function.

Method initialization First fix d_1^* and ν^* as their values do not need to get updated at each iteration. Then initialize

$$\begin{aligned}m_q &= (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{Y}_q \\ \mathbf{V}^* &= ((\nu^* - P - 1)(\mathbf{Y} - \mathbf{G}\mathbf{M}_\alpha)^T (\mathbf{Y} - \mathbf{G}\mathbf{M}_\alpha) / (N - 1))^{-1} \\ \pi_q^* &= 1 \quad \forall q \\ v_q &= \left[(\mathbf{G}^T \mathbf{G})^{-1} \left(\left(1 - \frac{N - 1}{\nu^* - P - 1}\right) \mathbf{V}^{*-1} - \mathbf{V}^{-1} \right) \right]_{qq} \\ d_2^* &= d_2 + 0.5 \sum_{q=1}^P (v_q + m_q^2)\end{aligned}$$

All parameters are then updated in the following order - b_1^* , b_2^* , $\pi_q^* \forall q$, d_2^* , $\mathbf{V}^*$, $v_q \forall q$, $m_q \forall q$.

Default priors $d_1 = 4.5$, $d_2 = 4.5$, $\nu = 5 + P$, $\mathbf{V} = 0.25\mathbf{I}$, $b_1 = 1$, $b_2 = 3$.

Appendix G - VB SAS for related samples

Appendix

The model can be described as

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{G}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\beta}, \mathbf{I}_N, \boldsymbol{\Omega}^{-1}) \quad (37)$$

$$\mathbf{S} \sim \mathcal{N}(0, 2\mathbf{K}, I_M) \quad (38)$$

$$\beta_{mq} \sim \mathcal{N}(0, \lambda^2) \quad q \in 1, \dots, P, m \in 1, \dots, M \quad (39)$$

$$\alpha_q \sim \pi \mathcal{N}(\alpha_q | 0, \sigma_g^2) + (1 - \pi) \delta(\alpha_q) \quad q \in 1, \dots, P \quad (40)$$

$$\sigma_g^2 \sim \Gamma^{-1}(d_1, d_2) \quad (41)$$

$$\pi \sim \text{Beta}(b_1, b_2) \quad (42)$$

$$\boldsymbol{\Omega} \sim \text{Wi}(e, \mathbf{E}) \quad (43)$$

Following [Titsias and Lazaro-Gredilla, 2011] we use a reparameterization of the model that is more amenable to VB inference such that α_q is represented as the product of a *Bernoulli*(π) random variable s_q and a $\mathcal{N}(0, \sigma_g^2)$ random variable $\tilde{\alpha}_q$

so that $\alpha_q = \tilde{\alpha}_q s_q, q \in 1, \dots, P$. The relevant log densities are then given by

$$\begin{aligned}
\log P(\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{S}, \boldsymbol{\beta}, \mathbf{D}) &= -\frac{1}{2} \text{tr}((\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T) \\
&\quad + \frac{N}{2} \log |\boldsymbol{\Omega}| + \text{const} \\
\log P(\sigma_g^2) &= -(d_1 + 1) \log \sigma_g^2 - d_2 \sigma_g^{2^{-1}} + \text{const} \\
\log P(\mathbf{s}) &= \sum_{q=1}^P s_q \log \pi + \sum_{q=1}^P (1 - s_q) \log(1 - \pi) + \text{const} \\
\log P(\pi) &= (b_1 - 1) \log \pi + (b_2 - 1) \log(1 - \pi) + \text{const} \\
\log P(\tilde{\boldsymbol{\alpha}}) &= -\frac{1}{2} \sum_{q=1}^P \log \sigma_g^2 - \frac{1}{2} \sum_{q=1}^P \tilde{\alpha}_q^2 \sigma_g^{2^{-1}} + \text{const} \\
\log P(\boldsymbol{\beta}) &= -\frac{1}{2} \sum_{q=1}^P \sum_{m=1}^M \log \lambda^2 - \frac{1}{2} \sum_{q=1}^P \sum_{m=1}^M \beta_{mq}^2 \lambda^{-2} + \text{const} \\
\log P(\mathbf{S}) &= -\frac{1}{2} \sum_{m=1}^M \log |2\mathbf{K}| - \frac{1}{2} \sum_{m=1}^M \mathbf{S}_m^T (2\mathbf{K})^{-1} \mathbf{S}_m + \text{const} \\
\log P(\boldsymbol{\Omega}) &= \frac{e - P - 1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \text{tr}(\mathbf{E}^{-1} \boldsymbol{\Omega}) + \text{const}
\end{aligned}$$

The full set of parameters is $\theta = (\tilde{\boldsymbol{\alpha}}, \mathbf{s}, \pi, \boldsymbol{\beta}, \mathbf{S}, \boldsymbol{\Omega})$ and

$$\begin{aligned}
\log P(\mathbf{D}, \theta | M) &= \log P(\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{S}, \boldsymbol{\beta}, \mathbf{D}) + \log P(\tilde{\boldsymbol{\alpha}}) + \log P(\mathbf{s}) + \log P(\pi) \\
&\quad + \log P(\mathbf{S}) + \log P(\boldsymbol{\beta}) + \log P(\boldsymbol{\Omega}) + \log P(\sigma_g^2) + \text{const}
\end{aligned}$$

We use a factorization for Q of the following form

$$Q(\theta) = Q(\sigma_g^2) Q(\pi) Q(\boldsymbol{\Omega}) \prod_{m=1}^M Q(\boldsymbol{\beta}_m) \prod_{m=1}^M Q(\mathbf{S}_m) \prod_{q=1}^P Q(\tilde{\alpha}_q, s_q)$$

Approximate posterior distribution for Ω

$$\begin{aligned}
 \log Q(\Omega) &= \mathbb{E}_{Q(\theta \setminus \Omega)} (\log P(\mathbf{D}, \theta)) + \text{const} \\
 &= \frac{e - P - 1}{2} \log |\Omega| - \frac{1}{2} \text{tr}(\mathbf{E}^{-1} \Omega) \\
 &\quad - \frac{1}{2} \mathbb{E} \text{tr}((\mathbf{Y} - \mathbf{S}\beta - \mathbf{G}\alpha) \Omega (\mathbf{Y} - \mathbf{S}\beta - \mathbf{G}\alpha)^T) \\
 &\quad + \frac{N}{2} \log |\Omega| + \text{const}
 \end{aligned}$$

Hence $Q(\Omega) = Wi(e^*, \mathbf{E}^*)$, where $e^* = e + N$ and

$$\mathbf{E}^* = (\mathbf{E}^{-1} + \mathbb{E}[(\mathbf{Y} - \mathbf{S}\beta - \mathbf{G}\alpha)^T(\mathbf{Y} - \mathbf{S}\beta - \mathbf{G}\alpha)])^{-1}$$

Approximate posterior distribution for σ_g^2

$$\begin{aligned}
 \log Q(\sigma_g^2) &= \mathbb{E}_{Q(\theta \setminus \sigma_g^2)} (\log P(\mathbf{D}, \theta)) + \text{const} \\
 &= -(d_1 + 1) \log \sigma_g^2 - d_2 \sigma_g^{-2} - \frac{1}{2} \sum_{q=1}^P \log \sigma_g^2 - \frac{1}{2} \sum_{q=1}^P \mathbb{E}(\tilde{\alpha}_q^2) \sigma_g^{-2} + \text{const}
 \end{aligned}$$

Hence $Q(\sigma_g^2) = \Gamma^{-1}(d_1^*, d_2^*)$, where $d_1^* = d_1 + \frac{P}{2}$ and $d_2^* = d_2 + \frac{1}{2} \sum_{q=1}^P \mathbb{E}[\tilde{\alpha}_q^2]$

Approximate posterior distribution for β_m .

$$\begin{aligned}
 \log Q(\beta_{m.}) &= \mathbb{E}_{Q(\theta \setminus \beta_{m.})} (\log P(\mathbf{D}, \theta)) + \text{const} \\
 &= -\frac{1}{2} \mathbb{E} \text{tr} \left[(\mathbf{Y} - \sum_m \mathbf{S}_m \beta_{m.} - \mathbf{G}\alpha) \Omega (\mathbf{Y} - \sum_m \mathbf{S}_m \beta_{m.} - \mathbf{G}\alpha)^T \right] \\
 &\quad - \frac{1}{2} \sum_{q=1}^P \beta_{mq}^2 \lambda^{-2} + \text{const}
 \end{aligned}$$

Hence $Q(\beta_{m.}) \sim \mathcal{N}(\mathbf{M}_m, I_1, \mathbf{V}_m)$, where $\mathbf{V}_m = [\mathbb{E}(\mathbf{S}_m^T \mathbf{S}_m) \mathbb{E}(\Omega) + \lambda^{-2} I_P]^{-1}$ and

$$\mathbf{M}_m = \mathbb{E}(\mathbf{S}_m)^T \mathbf{X}_m \mathbb{E}(\Omega) \mathbf{V}_m = \mathbb{E}(\mathbf{S}_m)^T (\mathbf{Y} - \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}) - \sum_{k \neq m} \mathbb{E}(\mathbf{S}_k) \mathbb{E}(\boldsymbol{\beta}_{k.})) \mathbb{E}(\Omega) \mathbf{V}_m$$

Approximate posterior distribution for $\mathbf{S}_m$

$$\begin{aligned} \log Q(\mathbf{S}_m) &= \mathbb{E}_{Q(\theta \setminus \mathbf{S}_m)} (\log P(\mathbf{D}, \theta)) + \text{const} \\ &= -\frac{1}{2} \mathbf{S}_m^T (2\mathbf{K})^{-1} \mathbf{S}_m \\ &\quad - \frac{1}{2} \mathbb{E} \operatorname{tr} \left[(\mathbf{Y} - \mathbf{G} \boldsymbol{\alpha} - \sum_{m=1}^P \mathbf{S}_m \boldsymbol{\beta}_{m.}) \Omega (\mathbf{Y} - \mathbf{G} \boldsymbol{\alpha} - \sum_{m=1}^P \mathbf{S}_m \boldsymbol{\beta}_{m.})^T \right] + \text{const} \\ &= -\frac{1}{2} \mathbf{S}_m^T (2\mathbf{K})^{-1} \mathbf{S}_m \\ &\quad - \frac{1}{2} \operatorname{tr}(\mathbf{S}_m \mathbb{E}(\boldsymbol{\beta}_{m.} \Omega \boldsymbol{\beta}_{m.}^T) \mathbf{S}_m^T) \\ &\quad + \frac{1}{2} \operatorname{tr}(\mathbf{S}_m \mathbb{E}(\boldsymbol{\beta}_{m.} \Omega) \mathbf{X}_m^T) + \frac{1}{2} \operatorname{tr}(\mathbf{X}_m \mathbb{E}(\Omega \boldsymbol{\beta}_{m.}^T) \mathbf{S}_m^T) + \text{const} \\ &\quad \text{where } \mathbf{X}_m = \mathbf{Y} - \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}) - \sum_{k \neq m} \mathbb{E}(\mathbf{S}_k \boldsymbol{\beta}_{k.}) \end{aligned}$$

Hence $Q(\mathbf{S}_m) \sim \mathcal{N}(\mathbf{v}_m^*, \boldsymbol{\Psi}_m^*, I_1)$, where $\boldsymbol{\Psi}_m^* = [(2\mathbf{K})^{-1} + \mathbb{E}(\boldsymbol{\beta}_{m.} \Omega \boldsymbol{\beta}_{m.}^T) \mathbf{I}_N]^{-1}$ and $\mathbf{v}_m^* = \boldsymbol{\Psi}_m^* \mathbf{X}_m \mathbb{E}(\Omega \boldsymbol{\beta}_{m.}^T) = \boldsymbol{\Psi}_m^* (\mathbf{Y} - \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}) - \sum_{k \neq m} \mathbb{E}(\mathbf{S}_k \boldsymbol{\beta}_{k.})) \mathbb{E}(\Omega) \mathbb{E}(\boldsymbol{\beta}_{m.})^T$

Approximate posterior distribution for π

$$\begin{aligned} \log Q(\pi) &= \mathbb{E}_{Q(\theta \setminus \pi)} (\log P(\mathbf{D}, \theta)) + \text{const} \\ &= \mathbb{E} [(b_1 - 1) \log(\pi) + (b_2 - 1) \log(1 - \pi) \\ &\quad + \sum_{q=1}^P s_q \log \pi + \sum_{q=1}^P (1 - s_q) \log(1 - \pi)] + \text{const} \end{aligned}$$

Therefore $Q(\pi) \sim B(b_1^*, b_2^*)$, where $b_1^* = b_1 + \sum_{q=1}^P \mathbb{E}(s_q)$ and $b_2^* = b_2 + \sum_{q=1}^P (1 - \mathbb{E}(s_q))$

Approximate posterior distribution for $\{\tilde{\alpha}_q, s_q\}$

$$\begin{aligned}
\log Q(\tilde{\alpha}_q, s_q) &= \mathbb{E}_{Q(\theta \setminus \{\tilde{\alpha}_q, s_q\})} (\log P(\mathbf{D}, \theta)) + \text{const} \\
&= \mathbb{E} \left[-\frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T \right. \\
&\quad \left. + s_q \log \pi + (1 - s_q) \log(1 - \pi) - \frac{1}{2} \tilde{\alpha}_q^2 \sigma_g^{2-1} \right] + \text{const} \\
&\text{where } \alpha_q = \tilde{\alpha}_q s_q \quad \forall q
\end{aligned} \tag{44}$$

Let $\mathbf{Z} = \mathbf{Y} - \mathbf{S}\boldsymbol{\beta}$. Then

$$\begin{aligned}
\text{tr}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T &= \text{tr}(\mathbf{Z} - \mathbf{G}\boldsymbol{\alpha})\boldsymbol{\Omega}(\mathbf{Z} - \mathbf{G}\boldsymbol{\alpha})^T \\
&= \boldsymbol{\alpha}\mathbf{G}^T\mathbf{G}\boldsymbol{\Omega}\boldsymbol{\alpha}^T - \boldsymbol{\alpha}\boldsymbol{\Omega}\mathbf{Z}^T\mathbf{G} - \mathbf{G}^T\mathbf{Z}\boldsymbol{\Omega}\boldsymbol{\alpha}^T + \text{const}
\end{aligned}$$

Therefore

$$\begin{aligned}
\log Q(\tilde{\alpha}_q, s_q) &= -\frac{1}{2} \mathbb{E} \left[\sum_{i,j} \tilde{\alpha}_i \tilde{\alpha}_j s_i s_j \Omega_{i,j} \mathbf{G}^T \mathbf{G} - 2 \sum_{i,j} \tilde{\alpha}_i s_i \Omega_{i,j} (Z^T \mathbf{G})_{j,1} \right. \\
&\quad \left. + s_q \log \pi + (1 - s_q) \log(1 - \pi) - \frac{1}{2} \tilde{\alpha}_q^2 \sigma_g^{2-1} \right] + \text{const}
\end{aligned}$$

Decompose $Q(\tilde{\alpha}_q, s_q) = Q(\tilde{\alpha}_q | s_q) Q(s_q)$. Then it follows that

$$\begin{aligned}
\log Q(\tilde{\alpha}_q, s_q = 1) &= \log \pi - \frac{1}{2} \tilde{\alpha}_q^2 \mathbb{E}(\sigma_g^{2-1}) - \frac{1}{2} \tilde{\alpha}_q^2 \mathbb{E}(\Omega_{q,q}) \mathbf{G}^T \mathbf{G} \\
&\quad + \sum_j \tilde{\alpha}_q \mathbb{E}(\Omega_{q,j}) (\mathbb{E}(\mathbf{Z})^T \mathbf{G})_{j,1} - \frac{1}{2} \sum_{q \neq j} \tilde{\alpha}_q \mathbb{E}(\tilde{\alpha}_j) \mathbb{E}(s_j) \mathbb{E}(\Omega_{q,j}) \mathbf{G}^T \mathbf{G} + \text{const}
\end{aligned}$$

and

$$\log Q(\tilde{\alpha}_q, s_q = 0) = \log(1 - \pi) - \frac{1}{2} \tilde{\alpha}_q^2 \mathbb{E}(\sigma_g^{2-1}) + \text{const}$$

Therefore

$$\begin{aligned} Q(\tilde{\alpha}_q | s_q = 1) &\sim \mathcal{N}(m_q, v_q) \\ Q(\tilde{\alpha}_q | s_q = 0) &\sim \mathcal{N}(0, \mathbb{E}(\sigma_g^{2^{-1}})^{-1}) \end{aligned}$$

where $v_q = \left[\mathbf{G}^T \mathbf{G} \mathbb{E}(\Omega_{q,q}) + \mathbb{E}(\sigma_g^{2^{-1}}) \right]^{-1}$ and

$$m_q = v_q \left[\sum_j \mathbb{E}(\Omega_{q,j}) (\mathbb{E}(Z)^T \mathbf{G})_{j,1} - \sum_{q \neq j} \mathbb{E}(\tilde{\alpha}_j) \mathbb{E}(s_j) \mathbb{E}(\Omega_{q,j}) \mathbf{G}^T \mathbf{G} \right]$$

Then

$$\begin{aligned} \log Q(s_q = 1) &= \log Q(\tilde{\alpha}_q, s_q = 1) - \log Q(\tilde{\alpha}_q | s_q = 1) \\ &= \frac{1}{2} v_q^{-1} m_q^2 + \frac{1}{2} \log v_q + \mathbb{E}(\log \pi) + \text{const} \end{aligned}$$

and

$$\begin{aligned} \log Q(s_q = 0) &= \log Q(\tilde{\alpha}_q, s_q = 0) - \log Q(\tilde{\alpha}_q | s_q = 0) \\ &= \frac{1}{2} \log \mathbb{E}(\sigma_g^{2^{-1}})^{-1} + \mathbb{E}(\log(1 - \pi)) + \text{const} \end{aligned}$$

If we write $\pi_q^* = Q(s_q = 1)$ and note that the constants in the above equations are the same then we have

$$\rho_q = \log \frac{\pi_q^*}{1 - \pi_q^*} = \mathbb{E} \left(\log \frac{\pi}{1 - \pi} \right) + \frac{m_q^2}{2v_q} + \frac{1}{2} \log v_q + \frac{1}{2} \log \left[\mathbb{E}(\sigma_g^{2^{-1}}) \right]$$

and $\pi_q^* = \frac{1}{1 + e^{-\rho_q}}$. Then the joint approximate posterior distribution of $\tilde{\alpha}_q$ and s_q can be written in the following form

$$Q(\tilde{\alpha}_q, s_q) = Q(\tilde{\alpha}_q | s_q) Q(s_q) = \mathcal{N} \left(s_q m_q, s_q v_q + (1 - s_q) \mathbb{E}(\sigma_g^{2^{-1}})^{-1} \right) (\pi_q^*)^{s_q} (1 - \pi_q^*)^{1-s_q}$$

and this is of the same form as the prior. Also the expectation of $\alpha_q = \tilde{\alpha}_q s_q$ is $\pi_q^* m_q$. Let $m_q^* = s_q m_q$ and $v_q^* = s_q v_q + (1 - s_q) \mathbb{E}(\sigma_g^{2-1})^{-1}$. Then $Q(\tilde{\alpha}_q | s_q) = \mathcal{N}(m_q^*, v_q^*)$.

Approximate marginal likelihood The approximation to the marginal likelihood is given by

$$\begin{aligned}
F(Q(\theta)) &= \int \log \left[\frac{P(\theta, \mathbf{D} | M)}{Q(\theta)} \right] Q(\theta) d\theta \\
&= \mathbb{E}[\log P(\mathbf{D}, \theta) - \log Q(\theta)] \\
&= \mathbb{E} \left[-\frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{S}\beta - \mathbf{G}\alpha)^T \mathbf{\Omega} (\mathbf{Y} - \mathbf{S}\beta - \mathbf{G}\alpha) + \cancel{\frac{n}{2} \log |\mathbf{\Omega}|} - \frac{NP}{2} \log(2\pi) \right. \\
&\quad \left. - \frac{PM}{2} \log \lambda^2 - \frac{1}{2} \sum_{q,m} \beta_{mq}^2 \lambda^{2-1} - \cancel{\frac{MP}{2} \log(2\pi)} \right. \\
&\quad \left. - \frac{1}{2} \sum_{m=1}^M \log |2\mathbf{K}| - \cancel{\frac{MN}{2} \log(2\pi)} - \frac{1}{2} \sum_{m=1}^M \mathbf{S}_m^T (2\mathbf{K})^{-1} \mathbf{S}_m \right. \\
&\quad \left. + \cancel{\frac{e-P-1}{2} \log |\mathbf{\Omega}|} - \frac{1}{2} \text{tr}(\mathbf{E}^{-1} \mathbf{\Omega}) - \frac{eP}{2} \log(2) - \frac{e}{2} \log |\mathbf{E}| - \log \Gamma_P \left(\frac{e}{2} \right) \right. \\
&\quad \left. - \cancel{(d_1+1) \log \sigma_g^2} - d_2 \sigma_g^{2-1} + d_1 \log d_2 - \log \Gamma(d_1) \right. \\
&\quad \left. + \sum_{q=1}^P s_q \log \pi + \sum_{q=1}^P (1 - s_q) \log(1 - \pi) - \sum_{q=1}^P s_q \log \pi_q^* - \sum_{q=1}^P (1 - s_q) \log(1 - \pi_q^*) \right. \\
&\quad \left. - \cancel{\frac{1}{2} \sum_{q=1}^P \log \sigma_g^2} - \frac{1}{2} \sum_{q=1}^P \tilde{\alpha}_q^2 \sigma_g^{2-1} - \cancel{\frac{P}{2} \log(2\pi)} \right. \\
&\quad \left. + (b_1 - 1) \log(\pi) + (b_2 - 1) \log(1 - \pi) - \log B(b_1, b_2) \right. \\
&\quad \left. + \frac{1}{2} \sum_{m=1}^M \log |\mathbf{V}_m| + \frac{1}{2} \sum_{m=1}^M (\beta_{m.} - \mathbf{M}_m)^T \mathbf{V}_m^{-1} (\beta_{m.} - \mathbf{M}_m) + \cancel{\frac{MP}{2} \log(2\pi)} \right. \\
&\quad \left. + \frac{1}{2} \sum_{m=1}^M \log |\mathbf{\Psi}^*| + \cancel{\frac{MN}{2} \log(2\pi)} + \frac{1}{2} \sum_{m=1}^M (\mathbf{S}_m - \mathbf{v}_m^*)^T (\mathbf{\Psi}_m^*)^{-1} (\mathbf{S}_m - \mathbf{v}_m^*) \right. \\
&\quad \left. - \cancel{\frac{e^* - P - 1}{2} \log |\mathbf{\Omega}|} + \frac{1}{2} \text{tr}(\mathbf{E}^{*-1} \mathbf{\Omega}) + \frac{e^* P}{2} \log(2) + \frac{e^*}{2} \log |\mathbf{E}^*| + \log \Gamma_P \left(\frac{e^*}{2} \right) \right. \\
&\quad \left. + \cancel{(d_1^* + 1) \log \sigma_g^2} + d_2^* \sigma_g^{2-1} - d_1^* \log d_2^* + \log \Gamma(d_1^*) \right. \\
&\quad \left. + \frac{1}{2} \sum_{q=1}^P \log v_q^* + \frac{1}{2} \sum_{q=1}^P (\tilde{\alpha}_q - m_q^*)^2 v_q^{*-1} + \cancel{\frac{P}{2} \log(2\pi)} \right. \\
&\quad \left. - (b_1^* - 1) \log(\pi) - (b_2^* - 1) \log(1 - \pi) + \log B(b_1^*, b_2^*) \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \text{tr} \left(e^* \mathbf{E}^* (\mathbf{E}^{*-1} - \mathbf{E}^{-1} - \mathbb{E}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})) \right) \\
&\quad - \frac{NP}{2} \log(\pi) + \frac{e^*}{2} \log |\mathbf{E}^*| + \log \Gamma_P \left(\frac{e^*}{2} \right) - \frac{e}{2} \log |\mathbf{E}| - \log \Gamma_P \left(\frac{e}{2} \right) \\
&\quad + \frac{PM}{2} (1 - \log \lambda^2) + \frac{NM}{2} - \frac{M}{2} \log |2\mathbf{K}| + \frac{1}{2} \sum_{m=1}^M |\boldsymbol{\Psi}_m^*| + \frac{1}{2} \sum_{m=1}^M \log |\mathbf{V}_m| \\
&\quad - \frac{1}{2} \sum_{m=1}^M \lambda^{2-1} \left[\mathbf{M}_m^T \mathbf{M}_m + \text{diag}(\mathbf{V}_q) \right] - \frac{1}{2} \sum_{m=1}^M \left[\mathbf{v}_m^{*T} (2\mathbf{K})^{-1} \mathbf{v}_m^* + \text{tr}((2\mathbf{K})^{-1} \boldsymbol{\Psi}_m^*) \right] \\
&\quad + \frac{1}{2} \sum_{q=1}^P \left[\pi_q^* \log v_q - (1 - \pi_q^*) \log \frac{d_1^*}{d_2^*} \right] + \frac{P}{2} - \frac{1}{2} \frac{d_1^*}{d_2^*} \sum_{q=1}^P \left[\pi_q^* v_q + (1 - \pi_q^*) \frac{d_2^*}{d_1^*} + \pi_q^* m_q^2 \right] \\
&\quad - \sum_{q=1}^P \pi_q^* \log \pi_q^* - \sum_{q=1}^P (1 - \pi_q^*) \log (1 - \pi_q^*) - \log B(b_1, b_2) + \log B(b_1^*, b_2^*) \\
&\quad + (d_2^* - d_2) \frac{d_1^*}{d_2^*} + d_1 \log d_2 - d_1^* \log d_2^* + \log \frac{\Gamma(d_1^*)}{\Gamma(d_1)} \\
&\quad - (b_1^* - b_1 - \sum_{q=1}^P \pi_q^*) [\psi(b_1^*) - \psi(b_1^* + b_2^*)] - (b_2^* - b_2 + \sum_{q=1}^P (\pi_q^* - 1)) [\psi(b_2^*) - \psi(b_1^* + b_2^*)]
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha}) = \\
&\quad \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T [\boldsymbol{\Lambda} \mathbf{M} + \mathbf{G} \mathbb{E}(\boldsymbol{\alpha})] - [\boldsymbol{\Lambda} \mathbf{M} + \mathbf{G} \mathbb{E}(\boldsymbol{\alpha})]^T \mathbf{Y} \\
&\quad + \mathbf{M}^T \boldsymbol{\Lambda}^T \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}) + \mathbb{E}(\boldsymbol{\alpha})^T \mathbf{G}^T \boldsymbol{\Lambda} \mathbf{M} \\
&\quad + \mathbf{G}^T \mathbf{G} \mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha}) + \mathbb{E}(\boldsymbol{\beta}^T \mathbf{Z} \boldsymbol{\beta})
\end{aligned}$$

where $\mathbf{Z}$ is given by

$$\begin{aligned}
\mathbf{Z}_{mm} &= \text{tr}(\boldsymbol{\Psi}_m^*) + (\mathbf{v}_m^*)^T (\mathbf{v}_m^*) \quad m \in \{1, \dots, M\} \\
\mathbf{Z}_{ms} &= (\mathbf{v}_m^*)^T (\mathbf{v}_s^*) \quad m \neq s
\end{aligned}$$

$\mathbb{E}(\boldsymbol{\alpha})$ by

$$\mathbb{E}(\alpha_q) = m_q \pi_q^*$$

$\mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{\alpha})$ by

$$\begin{aligned}\mathbb{E}(\alpha_q \alpha_r) &= m_q \pi_q^* m_r \pi_r^* & q \neq r \\ \mathbb{E}(\alpha_q^2) &= (m_q^2 + v_q) \pi_q^*\end{aligned}$$

$\mathbf{M}$ by

$$\mathbf{M} = (\mathbf{M}_1^T \dots \mathbf{M}_M^T)^T$$

$\boldsymbol{\Lambda}$ by

$$\boldsymbol{\Lambda} = (\mathbf{v}_1^* \dots \mathbf{v}_M^*)$$

and $\mathbb{E}(\boldsymbol{\beta}^T \mathbf{Z} \boldsymbol{\beta})$ by

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}_{m.}^T \mathbf{Z} \boldsymbol{\beta}_{l.}) &= \mathbf{M}_{m.}^T \mathbf{Z} \mathbf{M}_{l.} & m \neq l \\ \mathbb{E}(\boldsymbol{\beta}_{m.}^T \mathbf{Z} \boldsymbol{\beta}_{m.}) &= \mathbf{M}_{m.}^T \mathbf{Z} \mathbf{M}_{m.} + \text{tr}(\mathbf{Z} \mathbf{V}_m)\end{aligned}$$

Parameter updates At each iteration of the algorithm the parameters of VB approximation are updated and various moments need to be calculated for the marginal likelihood approximation. The details are as follows

$$\begin{aligned}
e^* &= e + N \\
E^* &= (E^{-1} + \mathbb{E}(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha}))^{-1} \\
d_1^* &= d_1 + \frac{P}{2} \\
d_2^* &= d_2 + \frac{1}{2} \sum_{q=1}^P \left[\pi_q^*(v_q + m_q^2) + (1 - \pi_q^*) \frac{d_2^*}{d_1^*} \right] \\
v_q &= \left[e^* E_{q,q}^* \mathbf{G}\mathbf{G} + \frac{d_1^*}{d_2^*} \right]^{-1} \\
m_q &= v_q e^* \left[\sum_j \mathbf{E}_{qj}^* (\mathbf{Y}_j - \boldsymbol{\Lambda} \mathbf{M}_{\cdot j})^T \mathbf{G} - \sum_{j \neq q} \pi_q^* m_q \mathbf{E}_{qj}^* \mathbf{G}\mathbf{G} \right] \\
\mathbf{V}_m &= [Z_{m,m} e^* \mathbf{E}^* + \lambda^{-2} I_P]^{-1} \\
\mathbf{M}_m &= e^* \mathbf{v}_m^*{}^T (\mathbf{Y} - \mathbf{G}\mathbb{E}(\boldsymbol{\alpha}) - \sum_{k \neq m} \mathbf{v}_k^* \mathbf{M}_k) \mathbf{E}^* \mathbf{V}^m \\
\boldsymbol{\Lambda} &= (\mathbf{v}_1^* \dots \mathbf{v}_M^*) \\
\boldsymbol{\Psi}_m^* &= [(2K)^{-1} + \text{tr}(e^* \mathbf{E}^* (\mathbf{M}_m^T \mathbf{M}_m + \mathbf{V}_m)) \mathbf{I}_N]^{-1} \\
\mathbf{v}_m^* &= \boldsymbol{\Psi}_m^* \left[e^* \left(\mathbf{Y} - \mathbf{G}\mathbb{E}(\boldsymbol{\alpha}) - \sum_{l \neq m} \mathbf{v}_l^* \mathbf{M}_l \right) \mathbf{E}^* \mathbf{M}_m^T \right] \\
\rho_q &= \psi(b_1^*) - \psi(b_2^*) + \frac{1}{2} \log \frac{d_1^*}{d_2^*} + \frac{m_q^2}{2v_q} + \frac{1}{2} \log v_q \\
\pi_q^* &= \frac{1}{1 + e^{-\rho_q}} \\
b_1^* &= b_1 + \sum_{q=1}^P \pi_q^* \\
b_2^* &= b_2 + P - \sum_{q=1}^P \pi_q^*
\end{aligned}$$

Computational step

For related data, quicker updates can be obtained by factorizing $2\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and then prewhitening by $\mathbf{U}^T$ (data then is $\mathbf{Y}_U$). This would result in diagonal covariance matrices in the updates of $\mathbf{S}$. This can be adapted for inclusion of missing data (see below), but for computational speed when no missing data is present we use a prewhitening step

- decompose $2\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$
- prewhiten $\mathbf{G}^u = \mathbf{U}^T\mathbf{G}$ and $\mathbf{Y}^u = \mathbf{U}^T\mathbf{Y}$
- run method for $\mathbf{Y}^u$ and $\mathbf{G}^u$, where the equivalent prior for $\mathbf{S}$ is now $\mathcal{N}(0, \mathbf{\Lambda}, I_M)$, where $\mathbf{\Lambda}$ is diagonal, and everything else remains the same.
- then the updates for the now diagonal $\mathbf{\Psi}_m^*$ become faster (as well as the updates for $\mathbf{v}_m^*$)
- finally expressions involving traces and determinants of $\mathbf{\Psi}^*$ can be evaluated quickly since it is diagonal.

Suppose that we have prewhitened our data and $\mathbf{G} := \mathbf{U}^T\mathbf{G}^{(0)}$, $\mathbf{Y} := \mathbf{U}^T\mathbf{Y}^{(0)}$, where $\mathbf{Y}^{(0)}$ and $\mathbf{G}^{(0)}$ are our original data.

$$\begin{aligned}
 \log Q(\text{missing } \mathbf{Y}_i^{(0)}) &= \mathbb{E}_{Q(\theta|\mathbf{Y})} (\log P(\mathbf{D}, \theta)) + \text{const} \\
 &= -\frac{1}{2} \sum_{i=1}^N \mathbb{E} (\mathbf{Y}_i^{(0)} - \mathbf{U}_i \mathbf{S} \boldsymbol{\beta} - \mathbf{U}_i \mathbf{G} \boldsymbol{\alpha}) \Omega (\mathbf{Y}_i^{(0)} - \mathbf{U}_i \mathbf{S} \boldsymbol{\beta} - \mathbf{U}_i \mathbf{G} \boldsymbol{\alpha})^T \\
 &\quad + \text{const}
 \end{aligned} \tag{45}$$

Then each row of the phenotype data would follow a normal distribution with mean $H = \mathbf{U}_i \mathbb{E}(\mathbf{S}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})$ and variance $F = \mathbb{E}(\boldsymbol{\Omega})$. However here we are interested in only updating the missing values, therefore we require the conditional distributions of the missing values, given the non-missing ones.

We can use the properties about conditional distributions of the multivariate normal distribution (as in Appendix) to derive a multivariate normal distribution for each row of the missing phenotypes. Then all $\mathbf{Y}^T \mathbf{Y}$ terms will need to become $\mathbb{E}(\mathbf{Y}^T \mathbf{Y})$

References

- [Agakov et al., 2011] Agakov, F., McKeigue, P., Krohn, J., and Storkey, A. (2011). Sparse instrumental variables (SPIV) for genome-wide studies. *Neural Information Processing Systems (NIPS 2010)*, 23:28–36. [18](#), [20](#)
- [Astle and Balding, 2009] Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471. [23](#)
- [Aulchenko et al., 2007] Aulchenko, Y. S., de Koning, D. J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585. [23](#)
- [Bacanu et al., 2000] Bacanu, S.-A., Devlin, B., and Roeder, K. (2000). The power of genomic control. *American Journal of Human Genetics*, 66(6):1933–1944. [22](#)
- [Bishop and Tipping, 2000] Bishop, C. M. and Tipping, M. E. (2000). Variational relevance vector machines. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53. [28](#), [90](#)

REFERENCES

- [Campbell et al., 2005] Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G., and Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, 37(8):868–872. [22](#)
- [Carbonetto and Stephens, 2012] Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108. [87](#)
- [Cotsapas et al., 2011] Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., Abecasis, G. R., Barrett, J. C., Behrens, T., Cho, J., De Jager, P. L., Elder, J. T., Graham, R. R., Gregersen, P., Klareskog, L., Siminovitch, K. A., Van Heel, D. A., Wijmenga, C., Worthington, J., Todd, J. A., Hafler, D. A., Rich, S. S., Daly, M. J., and FOCiS Network of Consortia (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8):e1002254. [2](#)
- [Dawid, 1981] Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274. [10.1093/biomet/68.1.265](#). [32](#), [85](#)
- [de los Campos et al., 2012] de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2012). Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics*. [24](#)
- [Demirkan et al., 2011] Demirkan, A., Amin, N., Isaacs, A., Jarvelin, M.-R., Whitfield, J. B., Wichmann, H.-E., Kyvik, K. O., Rudan, I., Gieger, C., Hicks, A. A., Johansson, Å., Hottenga, J. J., Smith, J. J., Wild, S. H., Pedersen,

REFERENCES

- N. L., Willemsen, G., Mangino, M., Hayward, C., Uitterlinden, A. G., Hofman, A., Witteman, J., Montgomery, G. W., Pietiläinen, K. H., Rantanen, T., Kaprio, J., Döring, A., Pramstaller, P. P., Gyllensten, U., de Geus, E. J., Penninx, B. W., Wilson, J. F., Rivadeneria, F., Magnusson, P. K., Boomsma, D. I., Spector, T., Campbell, H., Hoehne, B., Martin, N. G., Oostra, B. A., McCarthy, M., Peltonen-Palotie, L., Aulchenko, Y., Visscher, P. M., Ripatti, S., Janssens, A. C. J., and van Duijn, C. M. (2011). Genetic architecture of circulating lipid levels. *European Journal of Human Genetics*, 19(7):813–819. [2](#), [16](#)
- [Demirkan et al., 2012] Demirkan, A., van Duijn, C. M., Ugocsai, P., Isaacs, A., Pramstaller, P. P., Liebisch, G., Wilson, J. F., Johansson, Å., Rudan, I., Aulchenko, Y. S., Kirichenko, A. V., Janssens, A. C. J. W., Jansen, R. C., Gnewuch, C., Domingues, F. S., Pattaro, C., Wild, S. H., Jonasson, I., Polasek, O., Zorkoltseva, I. V., Hofman, A., Karssen, L. C., Struchalin, M., Floyd, J., Igl, W., Biloglav, Z., Broer, L., Pfeufer, A., Pichler, I., Campbell, S., Zaboli, G., Kolcic, I., Rivadeneira, F., Huffman, J., Hastie, N. D., Uitterlinden, A., Franke, L., Franklin, C. S., Vitart, V., DIAGRAM Consortium, Nelson, C. P., Preuss, M., CARDIoGRAM Consortium, Bis, J. C., O'Donnell, C. J., Franceschini, N., CHARGE Consortium, Witteman, J. C. M., Axenovich, T., Oostra, B. A., Meitinger, T., Hicks, A. A., Hayward, C., Wright, A. F., Gyllensten, U., Campbell, H., Schmitz, G., and on behalf of the EUROSPAN consortium (2012). Genome-wide association study identifies novel loci associated with circulating Phospho- and Sphingolipid concentrations. *PLoS genetics*, 8(2):e1002490. [2](#), [16](#)

REFERENCES

- [Devlin and Roeder, 1999] Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004. [22](#)
- [Eichler et al., 2010] Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450. [1](#)
- [Ferreira and Purcell, 2009] Ferreira, M. A. R. and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics*, 25(1):132–133. [17](#), [20](#)
- [Fikse et al., 2003] Fikse, W., Rekaya, R., and Weigel, K. (2003). Genotype by environment interaction for milk production in guernsey cattle. *Journal of Dairy Science*, 86(5):1821–1827. [102](#)
- [Gibson, 2010] Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*, 42(7):558–560. [2](#)
- [Gieger et al., 2011] Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A. H., Labrune, Y., Lopez, L. M., Mägi, R., Meacham, S., Okada, Y., Pirastu, N., Sorice, R., Teumer, A., Voss, K., Zhang, W., Ramirez-Solis, R., Bis, J. C., Ellinghaus, D., Gögele, M., Hottenga, J. J., Langenberg, C., Kovacs, P., O’Reilly, P. F., Shin, S.-Y., Esko, T., Hartiala, J., Kanoni, S., Murgia, F., Parsa, A., Stephens, J., van der Harst, P., Ellen van der Schoot, C., Allayee, H., Attwood, A., Balkau, B., Bastardot, F., Basu, S., Baumeister, S. E., Biino, G., Bomba, L., Bonnefond, A., Cambien, F., Chambers, J. C., Cucca, F., D’Adamo, P., Davies, G., de Boer, R. A., de Geus, E. J. C., Döring, A., Elliott, P., Erdmann, J., Evans,

REFERENCES

- D. M., Falchi, M., Feng, W., Folsom, A. R., Frazer, I. H., Gibson, Q. D., Glazer, N. L., Hammond, C., Hartikainen, A.-L., Heckbert, S. R., Hengstenberg, C., Hersch, M., Illig, T., Loos, R. J. F., Jolley, J., Tee Khaw, K., Kühnel, B., Kyrtsonis, M.-C., Lagou, V., Lloyd-Jones, H., Lumley, T., Mangino, M., Maschio, A., Mateo Leach, I., McKnight, B., Memari, Y., Mitchell, B. D., Montgomery, G. W., Nakamura, Y., Nauck, M., Navis, G., Nöthlings, U., Nolte, I. M., Porteous, D. J., Pouta, A., Pramstaller, P. P., Pullat, J., Ring, S. M., Rotter, J. I., Ruggiero, D., Ruukonen, A., Sala, C., Samani, N. J., Sambrook, J., Schlessinger, D., Schreiber, S., Schunkert, H., Scott, J., Smith, N. L., Snieder, H., Starr, J. M., Stumvoll, M., Takahashi, A., Wilson Tang, W. H., Taylor, K., Tenesa, A., Lay Thein, S., Tönjes, A., Uda, M., Ulivi, S., van Veldhuisen, D. J., Visscher, P. M., Völker, U., Wichmann, H.-E., Wiggins, K. L., Willemsen, G., Yang, T.-P., Hua Zhao, J., Zitting, P., Bradley, J. R., Dedoussis, G. V., Gasparini, P., Hazen, S. L., Metspalu, A., Pirastu, M., Shuldiner, A. R., Joost van Pelt, L., Zwaginga, J.-J., Boomsma, D. I., Deary, I. J., Franke, A., Froguel, P., Ganesh, S. K., Jarvelin, M.-R., Martin, N. G., Meisinger, C., Psaty, B. M., Spector, T. D., Wareham, N. J., Akkerman, J.-W. N., Ciullo, M., Deloukas, P., Greinacher, A., Jupe, S., Kamatani, N., Khadake, J., Kooner, J. S., Penninger, J., Prokopenko, I., Stemple, D., Toniolo, D., Wernisch, L., Sanna, S., Hicks, A. A., Rendon, A., Ferreira, M. A., Ouwehand, W. H., and Soranzo, N. (2011). New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208. [11](#), [153](#)
- [Gilmour et al., 2009] Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. (2009). ASReml user guide release 3.0. *VSN International Ltd.* [26](#)

REFERENCES

- [Guan and Stephens, 2011] Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815. [160](#)
- [Hahn et al., 2010] Hahn, P. R., Carvalho, C. M., Carlos, M., and Scott, J. G. (2010). A sparse factor-analytic probit model for congressional voting patterns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):619–635. [161](#)
- [Hayes et al., 2009] Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(01):47–60. [134](#)
- [Helgadottir et al., 2008] Helgadottir, A., Thorleifsson, G., Magnusson, K. P., Gretarsdottir, S., Steinthorsdottir, V., and Manolescu, A. (2008). The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genetics*, 40:217–224. [10.1038/ng.72](#). [15](#)
- [Helgadottir et al., 2007] Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., and Jonasdottir, A. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, 316:1491–1493. [10.1126/science.1142842](#). [15](#)
- [Hindorff et al., 2013] Hindorff, L., MacArthur, J., Morales, J., Junkins, H., Hall, P., Klemm, A., and Manolio, T. (2013). A catalog of published genome-wide association studies. <http://www.genome.gov/gwastudies/>. [1](#), [7](#), [9](#)

REFERENCES

- [Hoggart et al., 2008] Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130–e1000130. [160](#)
- [Huffman et al., 2011] Huffman, J. E., Knežević, A., Vitart, V., Kattla, J., Adamczyk, B., Novokmet, M., Igl, W., Pučić, M., Zgaga, L., Johannson, Å., Redžić, I., Gornik, O., Zemunik, T., Polasek, O., Kolcic, I., Pehlic, M., Koeleman, C. A. M., Campbell, S., Wild, S. H., Hastie, N. D., Campbell, H., Gyllenstein, U., Wuhler, M., Wilson, J. F., Hayward, C., Rudan, I., Rudd, P. M., Wright, A. F., and Lauc, G. (2011). Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. *Human Molecular Genetics*, 20(24):5000–5011. [134](#), [138](#), [148](#), [156](#)
- [Illig et al., 2009] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B. S., Mewes, H.-W., Meitinger, T., Hrabé De Angelis, M., Kronenberg, F., Soranzo, N., Wichmann, H.-E., Spector, T. D., Adamski, J., and Suhre, K. (2009). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*, 42(2):137–141. [2](#), [16](#)
- [International Human Genome Sequencing Consortium, 2004] International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945. [7](#)
- [Jaakkola, 1997] Jaakkola, T. S. (1997). Variational methods for inference and estimation in graphical models. *PhD thesis, Department of Brain and Cognitive*

REFERENCES

- Sciences, Massachusetts Institute of Technology.* [28](#), [90](#)
- [Jaakkola and Jordan, 2000] Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian logistic regression: a variational approach. *Statistics and Computing*, 10(1):25–37. [28](#), [90](#)
- [Kang et al., 2010] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354. [23](#), [161](#)
- [Kang et al., 2008] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723. [22](#)
- [Klei et al., 2008] Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, 32(1):9–19. [17](#), [20](#)
- [Klein et al., 2005] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., and Haynes, C. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308:385–389. 10.1126/science.1109557. [9](#)
- [Knott and Haley, 2000] Knott, S. A. and Haley, C. S. (2000). Multitrait least squares for quantitative trait loci detection. *Genetics*, 156:899–911. [14](#), [18](#), [20](#)

REFERENCES

- [Korte et al., 2012] Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*. [25](#), [26](#), [27](#)
- [Ku et al., 2010] Ku, C. S., Loy, E. Y., Pawitan, Y., and Chia, K. S. (2010). The pursuit of genome-wide association studies: where are we now? *Journal of Human Genetics*, 55(4):195–206. [2](#), [8](#), [9](#), [10](#), [15](#)
- [Lauc et al., 2010] Lauc, G., Essafi, A., Huffman, J. E., Hayward, C., Knežević, A., Kattla, J. J., Polasek, O., Gornik, O., Vitart, V., Abrahams, J. L., Pučić, M., Novokmet, M., Redžić, I., Campbell, S., Wild, S. H., Borovečki, F., Wang, W., Kolcic, I., Zgaga, L., Gyllensten, U., Wilson, J. F., Wright, A. F., Hastie, N. D., Campbell, H., Rudd, P. M., and Rudan, I. (2010). Genomics meets glycomics—the first GWAS study of human N-glycome identifies HNF1 as a master regulator of plasma protein fucosylation. *PLoS genetics*, 6(12):e1001256. [133](#), [134](#), [141](#), [148](#), [156](#)
- [Legarra A. et al., 2005] Legarra A., López-Romero P., and Ugarte E. (2005). Bayesian model selection of contemporary groups for BLUP genetic evaluation in Latxa dairy sheep. *Livestock Production Science*, 93(3):205–212. [102](#)
- [Li et al., 2006] Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., and Churchill, G. A. (2006). Structural model analysis of multiple quantitative traits. *PLoS genetics*, 2(7):e114. [18](#), [20](#)

REFERENCES

- [Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835. [25](#), [99](#), [106](#), [135](#), [154](#)
- [Listgarten et al., 2012] Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526. [160](#)
- [Liu et al., 2009] Liu, J., Pei, Y., Papasian, C. J., and Deng, H.-W. (2009). Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic Epidemiology*, 33(3):217–227. [161](#)
- [Logsdon et al., 2010] Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11:58. [87](#), [160](#)
- [Maesschalck et al., 2000] Maesschalck, R. D., Jouan-Rimbaud, D., and Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18. [56](#)
- [Mangin et al., 1998] Mangin, B., Thoquet, P., and Grimsley, N. (1998). Pleiotropic QTL analysis. *Biometrics*, 54(1):88–99. [18](#)
- [Manolio, 2010] Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176. [9](#)

REFERENCES

- [Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753. 1, 11
- [Marchini, 2009] Marchini, J. (2009). A review of statistical issues in the analysis of genome-wide association studies : population structure, genotype imputation and assessing statistical significance. 44
- [Marchini, 2010] Marchini, J. (2010). SNPTTEST v2: Technical details. <http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.v2.pdf>. 38, 169
- [Marchini and Howie, 2010] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511. 87
- [McCarthy et al., 2008] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369. 10.1038/nrg2344. 4, 6
- [McPherson et al., 2007] McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., and Cox, D. R. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316:1488–1491. 10.1126/science.1142447. 15

REFERENCES

- [Metcalf, 2007] Metcalf, D. (2007). Concise review: hematopoietic stem cells and tissue stem cells: current concepts and unanswered questions. *Stem Cells*, 25(10):2390–2395. [66](#), [67](#)
- [Mitchell and Beauchamp, 1988] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83(404):1023–1036. [84](#), [87](#)
- [O’Reilly et al., 2012] O’Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M.-R., and Coin, L. J. M. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*, 7(5):e34861. [2](#), [17](#), [20](#), [112](#), [114](#)
- [Pompanon et al., 2005] Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6(11):847–846. [5](#)
- [Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909. [22](#)
- [Price et al., 2010] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463. [22](#)
- [Pruim et al., 2010] Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., and Willer, C. J.

REFERENCES

- (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337. [75](#)
- [Ragoussis, 2009] Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, 10(1):117–133. PMID: 19453250. [5](#)
- [Rakitsch et al., 2012] Rakitsch, B., Lippert, C., Topa, H., Borgwardt, K., Honkela, A., and Stegle, O. (2012). A mixed model approach for joint genetic analysis of alternatively spliced transcript isoforms using RNA-Seq data. *arXiv.org*, q-bio.GN. [26](#), [27](#)
- [Rekaya et al., 2003] Rekaya, R., Weigel, K., and Gianola, D. (2003). Bayesian estimation of parameters of a structural model for genetic covariances between milk yield in five regions of the United States. *Journal of Dairy Science*, 86(5):1837–1844. [102](#)
- [Rothman et al., 2010] Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962. [19](#), [20](#)
- [Segura et al., 2012] Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*. [160](#)
- [Sellke et al., 2001] Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of ρ Values for Testing Precise Null Hypotheses. *The American Statistician*, 55(1):62–71. [44](#)

REFERENCES

- [Shriner and Yi, 2009] Shriner, D. and Yi, N. (2009). Deviance information criterion (DIC) in Bayesian multiple QTL mapping. *Computational Statistics and Data Analysis*, 53(5):1850–1860. [102](#)
- [Solovieff et al., 2013] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495. [14](#)
- [Soranzo et al., 2009] Soranzo, N., Spector, T. D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., Salo, P., Voight, B. F., Burns, P., Laskowski, R. A., Xue, Y., Menzel, S., Altshuler, D., Bradley, J. R., Bumpstead, S., Burnett, M.-S., Devaney, J., Döring, A., Elosua, R., Epstein, S. E., Erber, W., Falchi, M., Garner, S. F., Ghorri, M. J. R., Goodall, A. H., Gwilliam, R., Hakonarson, H. H., Hall, A. S., Hammond, N., Hengstenberg, C., Illig, T., König, I. R., Knouff, C. W., McPherson, R., Melander, O., Mooser, V., Nauck, M., Nieminen, M. S., O'Donnell, C. J., Peltonen, L., Potter, S. C., Prokisch, H., Rader, D. J., Rice, C. M., Roberts, R., Salomaa, V., Sambrook, J., Schreiber, S., Schunkert, H., Schwartz, S. M., Serbanovic-Canic, J., Sinisalo, J., Siscovick, D. S., Stark, K., Surakka, I., Stephens, J., Thompson, J. R., Völker, U., Völzke, H., Watkins, N. A., Wells, G. A., Wichmann, H.-E., Van Heel, D. A., Tyler-Smith, C., Thein, S. L., Kathiresan, S., Perola, M., Reilly, M. P., Stewart, A. F. R., Erdmann, J., Samani, N. J., Meisinger, C., Greinacher, A., Deloukas, P., Ouwehand, W. H., and Gieger, C. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium.

REFERENCES

- Nature Genetics*, 41(11):1182–1190. [2](#), [16](#), [45](#), [49](#), [63](#), [64](#), [68](#), [69](#), [70](#), [72](#), [77](#), [80](#), [153](#)
- [Spiegelhalter et al., 2002] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. [102](#), [103](#)
- [Stearns, 2010] Stearns, F. W. (2010). One hundred years of pleiotropy: a retrospective. *Genetics*, 186:767–773. [15](#), [21](#), [27](#)
- [Stephens, 2013] Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, 8(7):e65245. [2](#), [49](#)
- [Stephens and Balding, 2009] Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–90. [44](#)
- [The 1000 Genomes Consortium, 2010] The 1000 Genomes Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073. 10.1038/nature09534. [8](#)
- [The International HapMap Consortium, 2007] The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861. 10.1038/nature06258. [8](#)
- [Titsias and Lazaro-Gredilla, 2011] Titsias, M. and Lazaro-Gredilla, M. (2011). Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learn-

REFERENCES

- ing. *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 1–9. [88](#), [181](#), [189](#)
- [Verzilli et al., 2005] Verzilli, C. J., Stallard, N., and Whittaker, J. C. (2005). Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. *Genetic Epidemiology*, 28(4):313–325. [19](#), [20](#)
- [Visscher et al., 2012] Visscher, P. M., Brown, M. A., McCarthy, M., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7 – 24. [10](#)
- [Voight and Pritchard, 2005] Voight, B. F. and Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS genetics*, 1(3):e32. [22](#)
- [Wagner, 2008] Wagner, G. P. (2008). Pleiotropic scaling of gene effects and the ‘cost of complexity’. *Nature*, 452:470–472. 10.1038/nature06756. [14](#)
- [Wagner and Zhang, 2011] Wagner, G. P. and Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12(3):204–213. 10.1038/nrg2949. [13](#), [15](#)
- [Wang, 2007] Wang, C. (2007). Variational Bayesian Approach to Canonical Correlation Analysis. *Neural Networks, IEEE Transactions on*, 18(3):905–910. [18](#), [20](#)
- [Wichmann et al., 2005] Wichmann, H., Gieger, C., and Illig, T. (2005). KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 61(S 01):26–30. [12](#), [63](#)

REFERENCES

- [Wilson et al., 1998] Wilson, P. W. F., D’Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847. 10.1161/01.CIR.97.18.1837. [12](#)
- [WTCCC, 2007] WTCCC (2007). Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, 447:661–678. 10.1038/nature05911. [7](#), [9](#), [44](#)
- [Yang et al., 2010] Yang, Q., Wu, H., Guo, C.-Y., and Fox, C. S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*, 34(5):444–454. [17](#), [20](#)
- [Yin and Li, 2011] Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 5(4):2630–2650. [19](#), [20](#)
- [Yu et al., 2006] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208. [23](#)
- [Zeggini et al., 2007] Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., and Lango, H. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341. 10.1126/science.1142364. [15](#)

REFERENCES

- [Zhang and Deng, 2010] Zhang, F. and Deng, H.-W. (2010). Correcting for cryptic relatedness in population-based association studies of continuous traits. *Human Heredity*, 69(1):28–33. [22](#)
- [Zhao et al., 2007] Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., and Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS genetics*, 3(1):e4. [23](#)
- [Zhou and Stephens, 2012] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824. [25](#), [26](#)
- [Zhou and Stephens, 2014] Zhou, X. and Stephens, M. (2014). Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature Methods*, 11:407–409. [26](#), [27](#), [159](#)
- [Zuk et al., 2012] Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198. [2](#)