Statistical Analysis of Genetic Data:

Interactions and Multivariate Outcome

Ann-Sophie Buchardt

PhD Thesis by:

Ann-Sophie Buchardt
Section of Biostatistics
Department of Public Health
University of Copenhagen
Denmark
ann-sophie@buchardt.net

Assessment committee:

Associate professor Jørgen Holm Petersen (chairperson)
University of Copenhagen, Dennmark

Associate professor Ulrich Halekoh
University of Southern Denmark, Denmark

Professor Lars-Gustav Snipen
Norwegian University of Life Sciences, Norway

Academic Advisors:

Professor Claus Thorn Ekstrøm
University of Copenhagen

Professor Haja Kadarmideen
Technical University of Denmark

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen on December 1, 2022.

Preface

This thesis constitutes my PhD dissertation at the University of Copenhagen. It has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen. The work included here was carried out between August 2016 and December 2022, under the supervision of Prof. Claus Thorn Ekstrøm (University of Copenhagen) and Prof. Haja Kadarmideen (Technical University of Denmark). It was funded through Styrelsen for Forskning og Innovation and the Section of Biostatistics, Department of Public Health, University of Copenhagen.

The thesis is broadly concerned with statistical genetics, with contributions to the methodologies of interactions and multivariate outcome. Motivating applications are primarily to be found in genetics fields.

Acknowledgements

I would like to thank my supervisor, Prof. Claus Thorn Ekstrøm, for introducing me to the problems discussed in this thesis, for being a constant source of knowledge and inspiration, for being an active and constructive co-author on all manuscripts, and for thoughtful guidance and support throughout the process.

I would also like to thank my other co-author, Prof. Xiang Zhou, for our collaboration. I also owe Nicolai Meinshausen thanks for functioning as a very friendly and accessible contact during my stay at the Seminar für Statistik at ETH Zürich and to the other PhD Fellows and the staff of the seminar in general, for their hospitality.

At the Section of Biostatistics, I would like to thank in particular Prof. Esben Budtz-Jørgensen for dedication and support. I also owe thanks to my office-mates and the other PhD Fellows down the hall. In general, I would like to thank everybody at the Section of Biostatistics at the University of Copenhagen for the various ways in which they have contributed to this thesis.

I thank my family and friends who in one way or another have encouraged me during the last six years.

Finally, my sincerest thanks to my husband Kristian for his loving support.

Copenhagen, December 2022
Ann-Sophie Buchardt

Til William & Emily

Summary

Improvement of feed efficiency (FE) is an important goal in pig production. Multi-omics experiments including genomics and systematic integration of omics data and modelling (systems biology) is a very promising approach to detect causal genetic and regulatory mechanisms underlying FE. The practical motivation for this PhD project is to investigate if genomic and metabolomic profiles are different in genetically efficient and inefficient pigs for feed utilisation in a systems biology experiment. Thus, we are in search of testable biomarkers for FE. This motivates the development and application of statistical methods for genomic prediction and selection methods for FE given the genetic architecture and biological information.

In Chapter 1, we give an overview of the research project and its contributions, and the main subjects of study are introduced. The Manuscripts I–III are self-contained manuscripts containing the main results of the research project.

It is characteristic for genomic or multi-omics data that the number, p, of features is (much) larger than the number, N, of observations, and when analysing such data we are concerned not only with identifying relevant features but also with identifying more complex relationships such as interactions, e.g., gene-gene or gene-environment interactions. However, the introduction of interactions rapidly increases the complexity of the statistical methodology, since the total number of possible pairwise interactions is $\binom{p}{2} = \frac{1}{2}p(p-1)$. For example, when a microarray is used to detect the expression of thousands of genes at the same time, the number of potential pairwise gene-gene interactions easily exceeds a million. Fitting a regression model to such data is computationally challenging and expectedly very time consuming. Furthermore, it is not clear exactly how to consistently include features which are interacting in regularised regression models which are useful when $p > N$. In Manuscript I we develop a computationally tractable two-step procedure for identifying pairwise interactions in a linear regression model subject to lasso regularisation under different assumptions of hierarchy: No hierarchy allowing for all interactions being included at once; weak hierarchy allowing for the inclusion of only interactions between at least one pre-selected main effect and another main effect; strong hierarchy allowing for the inclusion of only interactions between pre-selected constituent main effects. We have motivated our approach by modelling pairwise interactions for quantitative variables and experimented with explicitly applying (and not applying) penalties on the main effects and interactions, thereby obtaining interpretable models. We compare our method with existing approaches on real data as well as simulated data and find that under the assumption of uncorrelated features in a “reasonably” sized data set our method outperforms existing methods in computational cost.

When analysing genomic data or multi-omics data, information on a set of multiple, simultaneously measured traits is often collected in populations sampled for genome-wide association studies (GWASs). While single-trait GWASs have found numerous novel loci associated with complex diseases, simultaneously measured traits are often correlated and, therefore, require a joint modelling approach. However, applying the tools of quantitative genetics to high-dimensional, highly correlated datasets presents considerable analytical and computational challenges. These are the subject of Manuscript II, where we propose a method for utilising polygenic scores (PGSs) as a means for inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits in a multivariate GWAS. By estimating clusters of correlated traits that share some genetic component we are able to analyse the data combined in clusters and increase precision and power and gain computational advantages. We compare the method with fully parametric multivariate techniques on simulated data using R and illustrate the utility of the method by examining a heterogeneous stock mouse data set from the Wellcome Trust Centre for Human Genetics. We demonstrate that the method successfully identifies clusters and increases precision, power and computational efficiency.

In Manuscript III we extend the method proposed in Manuscript I to the multiple outcome setting. One of the major challenges in GWASs is the lack of power to detect main effects, and power to detect interactions is even more challenging. However, when multiple potentially correlated traits are available a joint modelling approach can be used to increase precision and power. We propose a method for selecting pairwise interactions in a multivariate regression model which involves only a subset of the features and interactions. The procedure uses (sparse) group lasso and adaptive lasso for screening for interactions and fitting multivariate hierarchical pairwise interaction models. This way interactions (and features) are simultaneously estimated and selected while hierarchical restrictions are taken into account. This ensures interpretability of the model (in terms of interactions) and should improve computing time. We consider the scenario which assumes that there is an unknown subset of the features which affect the outcome, but this subset is not preserved across all components of the outcome. As a special case, we consider the scenario which assumes that there is an unknown subset of the features which affect the outcome, and this same subset is preserved across all components of the outcome.

Resumé

Forbedring af foderudnyttelse (FDU) er et vigtigt mål i kommerciel svineproduktion. Multiomics eksperimenter, herunder genetik og systematisk integration af omicsdata og modellering (systembiologi), er meget lovende fremgangsmåder til at påvise kausale genetiske og regulatoriske mekanismer, der ligger til grund for FDU. Den praktiske motivation for dette ph.d.-projekt er at undersøge, om genetiske og metaboliske profiler er forskellige i genetisk effektive og ineffektive grise for FDU i et systembiologisk eksperiment. Vi leder dermed efter testbare biomarkører for FDU. Dette motiverer udviklingen og anvendelsen af statistiske metoder for genetisk prædiktions- og selektionsmetoder for FDU, givet den genetiske arkitektur og tilgængelig biologisk information.

I Kapitel 1 giver vi et overblik over forskningsprojektet og dets hovedbidrag, og vi introducerer de hovedemner, som projektet omhandler. Manuskripterne I–III er selvstændige manuskripter og indeholder projektets hovedresultater.

Det er karakteristisk for genetisk data eller multi-omics data, at antallet, p, af kovariater er (meget) større end antallet, N, af observationer, og når man analyserer denne slags data, er man ikke kun interesseret i at identificere relevante kovariater men også i at identificere mere komplekse relationer såsom vekselvirkninger, for eksempel gen-gen eller gen-miljø vekselvirkninger. Ved at introducere vekselvirkninger øger man dog kompleksiteten af de statistiske metoder betragteligt, da det samlede antal af mulige parvise vekselvirkninger er $\binom{p}{2} = \frac{1}{2}p(p-1)$. Hvis, for eksempel, en microarray er anvendt til at påvise genekspressionen af tusindvis af gener på samme tid, så vil det potentielle antal parvise vekselvirkninger nemt overstige en million. At fitte en regressionmodel til den slags data er beregningsmæssigt udfordrende og forventeligt meget tidskrævende. Det er desuden ikke klart, præcist hvordan man konsistent inkluderer kovariater som vekselvirker i en regulariseret regressionsmodel, som er anvendelig når $p > N$. I Manuskript I udvikler vi en beregningsmæssigt håndterbar to-trinsprocedure til at identificere parvise vekselvirkninger i en lineær regressionsmodel underlagt lasso regularisering under forskellige antagelser af hierarki: Intet hierarki som tillader alle vekselvirkninger at indgå; svagt hierarki som kun tillader vekselvirkninger mellem mindst en forudbestemt hovedeffekt og en anden hovedeffekt; og stærkt hierarki som kun tillader vekselvirkninger mellem tilhørende forudbestemte hovedeffekter. Vores tilgang er motiveret af at

modellere parvise vekselvirkninger for kvantitative variable og eksperimentere med eksplicit at pålægge (og ikke pålægge) regulariseringer på hovedeffekterne og vekselvirkningerne for dermed at opnå fortolkelige modeller. Vi sammenligner vores metode med eksisterende fremgangsmåder på rigtigt data såvel som simuleret data og finder, at vores metode, under antagelsen af at kovariaterne i et “rimeligt” stort datasæt er ukorrelerede, overgår eksisterende metoder hvad angår beregningsmæssige omkostninger.

Når man analyserer genetisk data eller multi-omics data, er der ofte indsamlet information fra flere samtidigt målte træk. Selvom enkelt-træks genome-wide association studier (GWAS'er) har fundet flere nye loci, associeret med komplekse sygdomme, så er træk, som er målt samtidig, ofte korrelerede, og en simultan modelleringsmetode er derfor krævet. At anvende metoderne for kvantitativ genetik på højdimensionelle, stærkt korrelerede datasæt, medfører dog betragtelige analytiske og beregningsmæssige udfordringer. Dette er emnet i Manuskript II, hvor vi foreslår en metode til at anvende polygenic scores (PGSs) til at udlede genetiske relationer mellem flere, samtidigt målte og potentielt korrelerede træk i et multivariat GWAS. Ved at estimere clusters af korrelerede træk, der deler nogle genetiske komponenter, er vi i stand til at analysere data kombineret i clusters og derved øge præcision og power og opnå beregningsmæssige fordele. Vi sammenligner metoden med fuldt parametriske multivariate teknikker på simuleret data ved at anvende R og illustrerer anvendelsen af metoden ved at undersøge heterogent musedatasæt fra the Wellcome Trust Centre for Human Genetics. Vi demonstrerer, at metoden med succes identificerer clusters og øger præcision, power og beregningsmæssig effektivitet.

I Manuskript III udvider vi metoden foreslået i Manuskript I til rammerne for multiple outcome. En af de store udfordringer i GWAS'er er den manglende power til at påvise hovedeffekter, og power til at påvise vekselvirkninger er endnu mere krævende. Når multiple, potentielt korrelerede træk er tilgængelige, kan en samlet modellering dog anvendes til at øge præcision og power. Vi foreslår en metode til at udvælge parvise vekselvirkninger i en multivariat regressionsmodel, som kun involverer en delmængde af kovariaterne og vekselvirkningerne. Proceduren anvender *(sparse) gruppe lasso* og *adaptive lasso* til at screene for vekselvirkninger og fitte en multivariat hierarkisk parvis vekselvirkningsmodel. På den måde bliver vekselvirkninger (og hovedeffekter) samtidigt estimeret og udvalgt, mens der bliver taget højde for hierarkiske restriktioner. Dette sikrer fortolkning af modellen (hvad angår vekselvirkninger) og burde forbedre beregningstiden. Vi betragter scenariet, der antager, at der er en ukendt delmængde af kovariaterne, der påvirker outcome, men denne delmængde er ikke bevaret henover komponenterne af outcome. Som et specialtilfælde betragter vi scenariet, der antager, at der er en ukendt delmængde af kovariaterne, der påvirker outcome, og at denne samme delmængde er bevaret henover alle komponenter af outcome.

Contents

Chapter 1

Introduction

The objective of statistical genetics is the development and application of statistical methods for drawing inferences from genetic data. In the following sections we propose a line of research along two subfields of statistical genetics. The subfields consider the analysis of interactions of high-dimensional features, e.g., gene-gene or gene-environment interactions, and of multivariate outcome, e.g., from a multivariate genome-wide association study (GWAS). In the study of interactions we focus, in particular, on regularised regressions, which are extremely convenient when the number of features measured is larger than the number of observations. In the endeavour to extend the framework to a multivariate outcome we were inspired to study the prospects of inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits in a simpler framework: we consider the standard linear regressions, where the exploitation of polygenic scores and various estimation techniques for linear regression allow us to draw inference.

Both the move from additive effects to interaction effects and the move from univariate to multivariate outcome represents a major increase in mathematical complexity. But not only that; with the increasing sophistication of measuring devices, extraordinary amounts of multiomics data are available today. Thus, the question of efficient computation is of key importance in integrative statistical genetics.

In this introductory chapter, we give an overview of the contributions of the thesis. The first step towards this is an introduction to the data generated as part of the FEEDOMICS study, which motivated the thesis. This is given in Section 1.2. The second step is a discussion of the objectives of the thesis, presented in Section 1.3, where a description of the research project is laid out in the context of motivational data examples. The results of the thesis can be divided into two subjects:

- Regularised regression with hierarchical interactions
- Multivariate modelling with polygenic scores

These subjects are covered separately in Section 1.4 and Section 1.5, where we both outline the basis for and results obtained under each headline, as well as discuss our findings and apply the developed methodology to evaluate the FEEDOMICS data. In Section 1.6 the goal of Section 1.4 is extended to a multivariate setup, and we outline the background and framework of the subject as well as discuss our findings. In Section 1.7 we review perspectives for future research based on the results obtained here. Section A summarises the manuscripts.

As mentioned in the summary, the thesis consists of an introduction and three manuscripts, each of which comprises a stand-alone scientific contribution. Table 1.1 gives an overview of the publication status of each manuscript at the time of this writing.

Manuscript I	Under revision in *Statistical Applications in Genetics and Molecular Biology*
Manuscript II	Submitted to the *Journal of the Royal Statistical Society*
Manuscript III	In preparation for *Statistical Applications in Genetics and Molecular Biology*

Table 1.1: *Publication status of manuscript included in the thesis at the time of this writing.*

In Manuscript II, we propose a method for inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits in a multivariate GWAS. As our main contribution we present an alternative to modelling a completely unstructured correlation matrix for the correlation among traits, via an approximation of the problem using polygenic scores (PGSs).

Manuscript I and Manuscript III are concerned with methods for identifying pairwise interactions in penalised regression models when the number of features measured is larger than the number of observations and the outcome is univariate and multivariate, respectively. While similar in their aim, Manuscript III presents particular difficulties due to, among other things, the potentially different relevance of different features across outcome components. In both Manuscript I and Manuscript III, different assumptions of hierarchy are taken into consideration by using a two-step procedure.

In all three manuscripts the algorithms provided have been implemented in R, and they are available online. See Buchardt (2022b) for the implementations related to Manuscript I and Manuscript III and Buchardt (2022a) for the implementation related to Manuscript II.

1.1 A short introduction to genetics

This subsection covers a selection of terms encompassing the spectrum of genetics and other biological terminology encountered in this thesis. As a brief reference guide intended for the less experienced genetics audience, fundamental concepts and connections are presented and related for an improved comprehension of the motivation and application of the statistical methods developed. The material is derived from Wikipedia articles, specifically the series on genetics (Genetics, 2021) and biology (Biology, 2021).

1.1.1 Genetics

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms. We refer to Figure 1.1 for an illustration of the relation between the following terms.

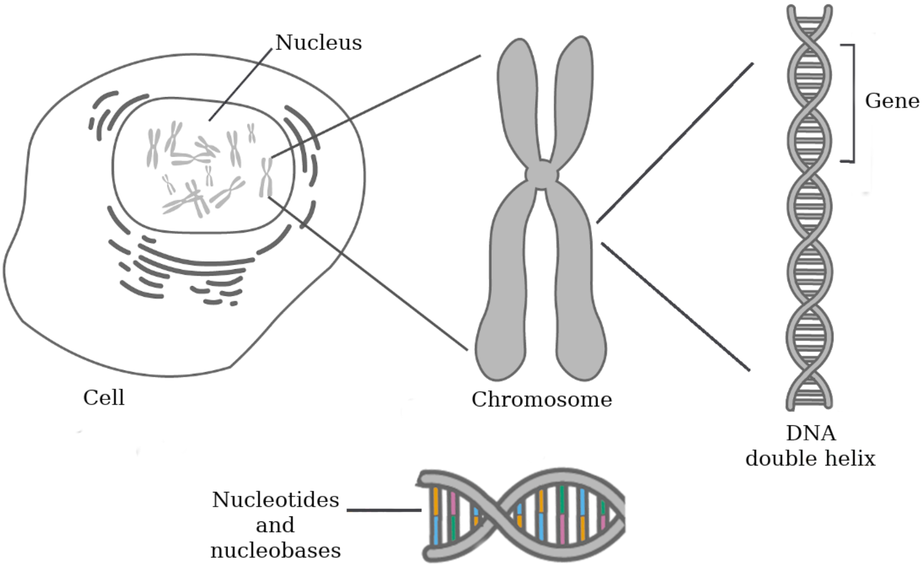


Figure 1.1: *Diagram of genome.*

Nucleotides are organic molecules consisting of a nucleoside and a phosphate. They serve as monomeric units of the nucleic acid polymers – deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), both of which are essential biomolecules within all life-forms on Earth.

DNA is a molecule composed of two polynucleotide chains (a sequence of nucleotides joined together) that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. Each nucleotide found in DNA is composed of one of four nitrogen-containing *nucleobases* (cytosine (C), guanine (G), adenine (A) or thymine (T), a sugar called deoxyribose, and a phosphate group. They function as the fundamental units of the genetic code.

A *chromosome* consists of a single, very long DNA helix on which thousands of genes are encoded. The region of the chromosome at which a particular gene is located is called its *locus*. Each locus contains one allele of a gene. An *allele* is one of two, or more, forms of a given gene variant. Alleles can come in different extremes of size. At the higher end, it can be up to several thousand base-pairs long. A base-pair is a fundamental unit of double-stranded nucleic acids consisting of two nucleobases bound to each other by hydrogen bonds. At the lowest possible size an allele can be a *single nucleotide polymorphism (SNP)*. Members of a population may have different alleles at the locus, each with a slightly different gene sequence. For example, at a specific base position in the human genome, the G nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by an A. This means that there is a SNP (a variation of a single nucleotide at a specific position in the genome) at this specific position, and the two possible nucleotide variations – G or A – are said to be the alleles for this specific position. This is illustrated in Figure 1.2. SNPs pinpoint differences in our susceptibility to a wide range of diseases, and the severity of illness and the way the body responds to treatments are also manifestations of genetic variations caused by SNPs. The *major allele* is the most common allele for a given SNP. The *minor allele* is the less common allele for a SNP.

The *genome* is all genetic information of an organism. The genome includes both the organism's ensemble of genes (the coding regions) and the non-coding DNA, as well as mitochondrial DNA and chloroplast DNA. A *gene* is a basic unit of heredity and a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein. The transmission of genes to an organism's offspring is the basis of the inheritance of phenotypic traits. These genes make up different DNA sequences called genotypes. Genotypes along with environmental and developmental factors determine what the phenotypes will be. The *genotype* is the individual organism's unique set of all the genes, and *genotyping* is the process of determining the genotype at specific positions within the genome of an individual. The genotype of a diploid organism (an organism with paired chromosomes, one from each parent) at a single locus on the DNA is described by the words *homozygous*, *heterozygous*, *hemizygous*, and *nullizygous*. Homozygous describes a genotype consisting of two identical alleles at a given locus, heterozygous describes a genotype consisting of two different alleles at a locus, hemizygous describes a genotype consisting of only a single copy of a particular gene in an otherwise diploid organism, and nullizygous refers to an otherwise-diploid organism in which both copies of the gene are missing.

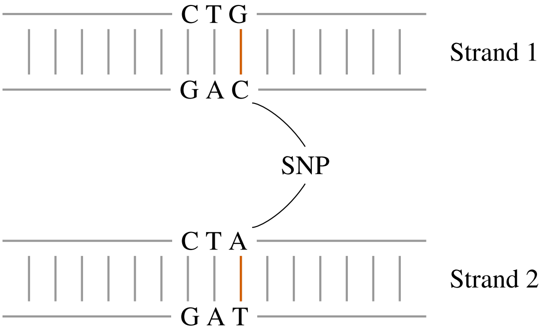


Figure 1.2: *DNA strand 1 differs from DNA strand 2 at a single base-pair location.*

1.1.2 Metabolomics

Metabolomics is the scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates and products of cell metabolism.

Metabolism is a set of chemical reactions in organisms with three main purposes: the conversion of the energy in food to energy available to run cellular processes; the conversion of food to building blocks for proteins, lipids, nucleic acids, and some carbohydrates; and the elimination of metabolic wastes (excrements).

Metabolites are intermediate or end products of metabolism. They are small molecules that take part in metabolic processes.

1.2 Phenotypic, genomic, and metabolomic data from Danish pigs

This PhD project is part of a larger research project with the acronym FEEDOMICS funded by the Danish research councils. The overall objective of the FEEDOMICS project is to improve the feed efficiency (FE) in Danish pig production. Feed costs are the single largest expense of raising pigs and the underlying hypothesis of the FEEDOMICS project is that genomic, transcriptomic and metabolic profiles are substantially different in extremely efficient and extremely inefficient pigs for feed utilisation. The materials for the project consist of data obtained from a collaboration between the Pig research centre (VSP) of the Danish Agriculture & Food Council (L&F), Danish Crown (DC) Herning and the University of Copenhagen.

While we have access to phenotypic, metabolomic, and genomic data as well as a pedigree database, we focus our presentation on the phenotypic and omics datasets, which have motivated our research project.

We consider data from a randomised controlled trial, tracking 112 randomly chosen pigs from two breeds, 62 DanBred Durocs (Durocs) and 50 DanBred Landraces (Landraces). Before the test phase begins the pigs arrive weighing approximately 7 kg and are acclimated. The test phase begins when the pigs weigh approximately 30 kg and when the pigs weigh approximately 100 kg, they are slaughtered.

Phenotypic data A list of the relevant variables in the phenotypic dataset are presented in Table 1.2. In the following analyses we focus primarily on the feed efficiency index (and meat percentage and scanning weight), which we use as outcome variable(s).

Variable name	Type	Description	Missing
`ID`	numeric	Unique identifier of each pig.	0
`scanning_date`	character	Scanning date (4 categories).	0
`scanning_weight`	numeric	Scanning weight (kg).	0
`kg_feed_consumed`	numeric	Feed consumed (kg).	0
`daily_gain`	numeric	Daily gain (g/day).	4
`Feed efficiency`	numeric	Feed efficiency index.	0
`meat%`	numeric	Meat %.	0
`race`	character	Race (Duroc or Landrace)	0

Table 1.2: *Variables in phenotype data.*

In Figure 1.3 we display violin plots for the continuous features. The violin plots have been supplemented with median and interquartile range information. What is most noteworthy is that whereas the estimated density of the scanning weight is somewhat similar for the two races, even though the density for Landrace pigs is bimodal and more heavy-tailed, the estimated density of the meat percentage indicates that a higher meat percentage is associated with the Landraces.

Genomic data We have genomic data for 112 complete-case pigs, including 62 Duroc and 50 Landrace pigs. The pigs were genotyped using the GGP Porcine HD array, specifically

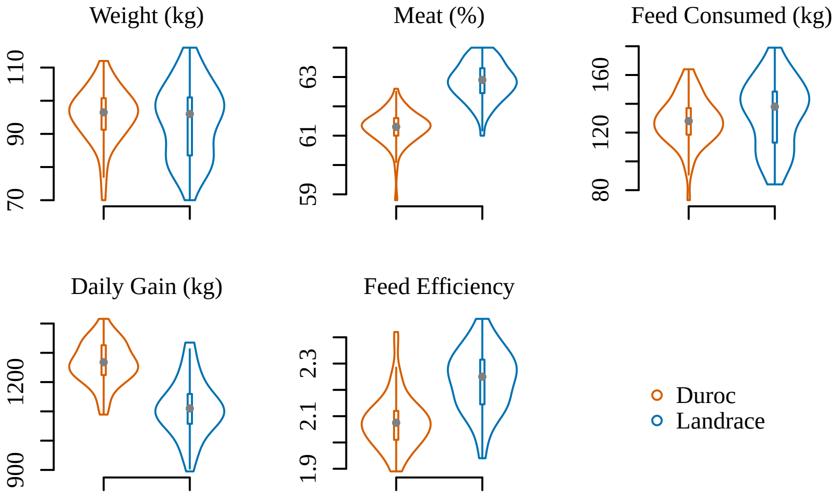


Figure 1.3: *Violin plots of the continuous variables from the phenotype table, showing that the distribution of several of the lab test variables is skewed, which suggests that a log transform will be beneficial. The distribution of the scanning weight and the feed consumed of the Landraces both appear bimodal.*

chosen for optimal chromosomal spacing and high minor allele frequency values for use in most commercial breeding lines. Sampling, sequencing, quality control, mapping and read quantification is described in more detail in Carmelo and Kadarmideen (2020).

For the Durocs, we have access to 31,079 SNPs after SNPs with no variation have been removed; for the Landraces we have access to 35,913 SNPs after SNPs with no variation have been removed. In the normal cell of a pig, or any mammal, chromosomes occur in distinct pairs. In contrast to humans, who have 23 chromosome pairs, pigs have 19 chromosome pairs (18 autosomes and the sex chromosome) for a total of 38 chromosomes. In Figure 1.4 we show the number of SNPs per chromosome for the Durocs (blue) and Landraces (orange). We observe that the numbers are not uniform across chromosomes, but they conform with the chromosome lengths (Ye et al., 2013). In Figure 1.5 we show the (smoothed) density of the SNP positions for each chromosome for the Durocs (blue) and Landraces (orange). We observe that, with the exception of the sex chromosome (number 19), the observed SNPs are generally uniformly positioned within the chromosomes.

Metabolomic data We have metabolite data for 109 complete-case pigs, including 59 Duroc and 50 Landrace. Blood samples were collected from each pig at two time points, once at the start of the FE testing phase where the pigs weigh approximately 30 kg, and again 45 days later. Non-targeted metabolomics analysis was performed on blood plasma using liquid

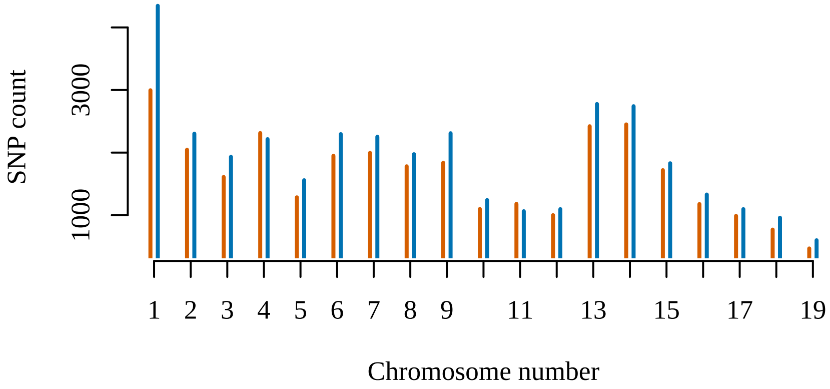


Figure 1.4: *Number of SNPs per chromosome for Durocs (orange) and Landraces (blue).*

chromatography–mass spectrometry (LC-MS), and 729 metabolites were identified in the data, yielding two data sets, `metabols_1` and `metabols_2`.

In Figure 1.6 we show barplots of the average metabolite intensities stratified on races for sample 1 (upper panel) and sample 2 (lower panel). For an initial overview, we have highlighted some of the metabolites with high intensities across race and sample.

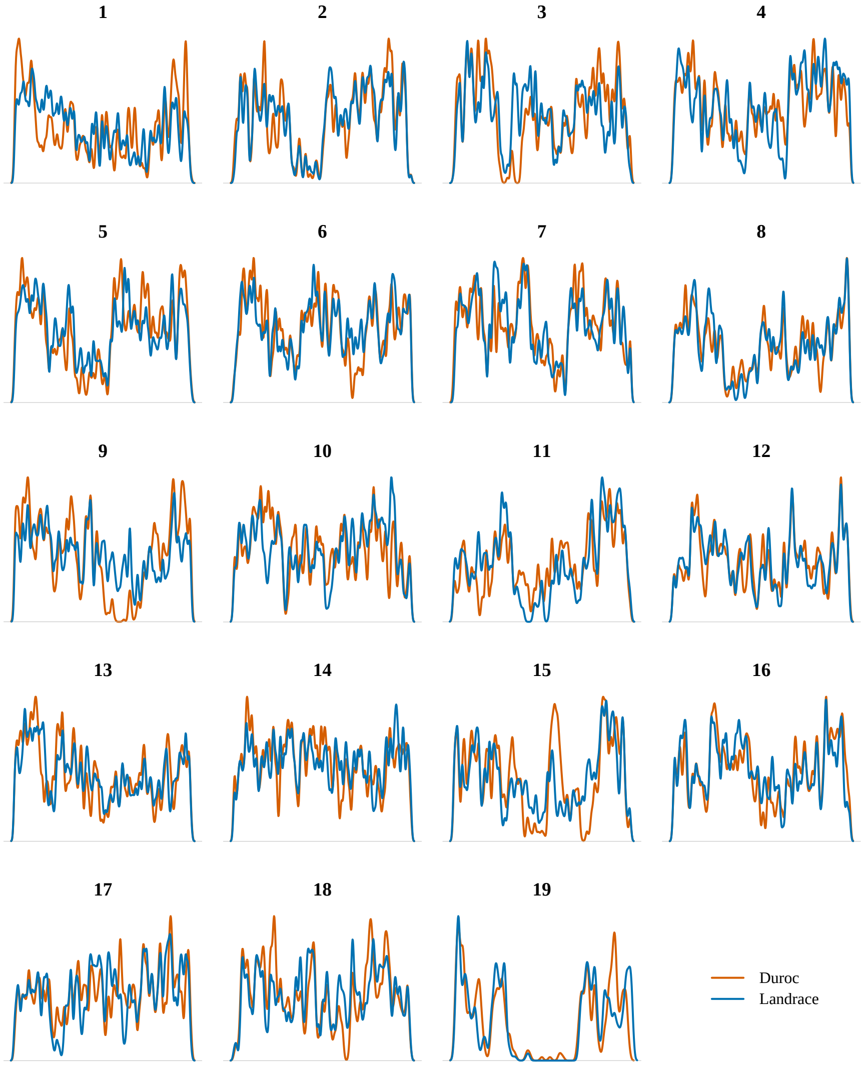


Figure 1.5: *SNP density for each chromosome for Durocs (orange) and Landraces (blue).*

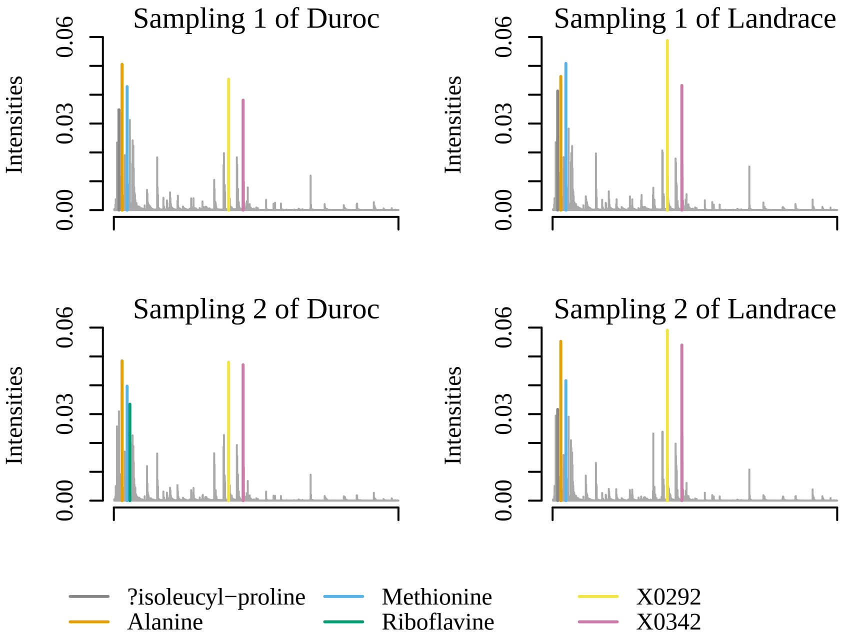


Figure 1.6: *Barplot of average (over pigs) normalised (per pig) metabolite intensities for Durocs (left panel) and Landraces (right panel) from sample 1 (upper panel) and sample 2 (lower panel).*

1.3 Objectives and description of the research project

The aim of statistical genetics is to develop statistical methods for understanding the genetic basis of diseases and traits in humans and animals. The methods involve large-scale datasets from candidate-gene approaches, which focus on associations between genetic variation within pre-specified genes of interest, to genome-wide association studies (GWASs or GWA studies), which search the entire genome for common genetic variation.

A GWAS is an observational study of a genome-wide set of genetic variants in different individuals with the aim of detecting potential variants associated with a phenotypic trait. For example, a GWAS may focus on associations between SNPs and a trait like a major human disease or another distinct variant of a phenotypic characteristic of any other organism – such as the FE of pigs. Often, genetic variants are tested for Hardy-Weinberg equilibrium (HWE) before testing for association, and a genetic model for the association test is specified. The HWE, also known as the Hardy–Weinberg principle, model, theorem, or law, states that, in the absence of other evolutionary influences (e.g., migration, mutation, selection, and inbreeding), allele and genotype frequencies will remain in their equilibrium state over the generations. In the simplest case of a single locus with two alleles denoted A and a with frequencies $f(A) = \pi$ and $f(a) = 1 - \pi$, respectively, the expected genotype frequencies under random mating are $f(AA) = \pi^2$ for the AA homozygotes, $f(aa) = (1 - \pi)^2$ for the aa homozygotes, and $f(Aa) = 2\pi(1 - \pi)$ for the heterozygotes Aa and aA. In the absence of selection, mutation, genetic drift, or other forces, allele frequencies π and $1 - \pi$ are constant between generations, so equilibrium is reached. Hardy-Weinberg equilibrium (HWE) is achieved in one generation of random mating. We refer to Hartl and Clark (1982) for a comprehensive treatment of the assumptions that underlie HWE. The principle is important in the context of genetic association studies for various reasons. For example, disequilibrium may be the result of genotyping error, e.g., a confusion of heterozygotes and homozygotes, and tests for HWE may help to detect (gross) genotyping error. In the FEEDOMICS data at hand, we do, however, know of a certain degree of inbreeding, and, therefore, HWE is most likely not reached at least for those loci which are influenced by inbreeding. Therefore, we omit testing for HWE.

A common statistical method for identifying SNPs associated with a quantitative trait is single SNP association testing with linear regression assuming an additive model. That is, under the assumption that we have N observations of a univariate outcome $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ (the trait) and one associated feature $\mathbf{x} \in \{0, 1, 2\}^N$ (a SNP where genotypes are coded as '0' (homozygote major, aa), '1' (heterozygote, Aa and aA), and '2' (homozygote minor, AA)), a univariate simple linear regression model on the form

$$y_i = \beta_0 + x_i\beta + \varepsilon_i, \tag{1.1}$$

is fitted. Here $\boldsymbol{\beta} = (\beta_0, \beta)^\top \in \mathbb{R}^2$ are unknown regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N) \in \mathbb{R}^N$ is a vector of zero-mean Gaussian random errors – an unobserved random variable adding noise to the linear relationship between the outcome and features:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_N\left(\mathbf{0}_N, \sigma^2\mathbf{I}_{N\times N}\right),$$

where $\mathbf{0}_N \in \mathbb{R}^N$ is a vector of zeros, $\mathbf{I}_{N\times N}$ is the $N \times N$ identity matrix, and $\sigma^2 > 0$ is the same for all errors, under the assumption that errors of the outcome are homoscedastic (they have the same finite variance) and uncorrelated (reflecting the assumption, that the underlying observations are independent).

The combination of encoding of $\mathbf{x}$ and linearity of the model corresponds to the assumption that the trait is additive in SNP effect, that is, having two minor alleles (AA coded as '2') rather than having no minor alleles (aa coded as '0') is twice as likely to affect the outcome in a certain direction as is having just one minor allele (Aa or aA coded as '1') rather than no minor alleles. This is unlike the dominance model which assumes that having one or two minor alleles rather than having no minor allele is twice as likely to affect the outcome in a certain direction and the recessive model which assumes that having two minor alleles rather than having no or one minor allele is twice as likely to affect the outcome in a certain direction. In Figure 1.7 we show scatterplots of FE against SNP number 52044 for the pooled data (left) and a altered version of the data (right). Specifically, for the altered version, we have added 0.2 to the FE of pigs with no minor alleles and subtracted 0.05 to the FE of pigs with just one minor allele. The alterations are purely for illustrative purposes. We have added a linear regression line and group means. Bearing in mind that eyeballing whether three points are on a line is not reliable, it could be argued that the left plot indicates that the additive (minor allele) model is reasonable, while the right plot indicates a dominant model. Whether the linearity and additivity assumption is reasonable for all SNPs from a GWAS is questionable. It is, however, not doable, to investigate the assumption in the case of thousands or millions of SNPs, and in the following we make the assumption without further investigations.

In the context of GWAS with, say, p SNPs, p univariate simple linear regression analyses are performed and t-tests are used to test whether to reject the null hypothesis that there is no linear relationship between a given SNP and the trait. After p-values are calculated for all SNPs, Manhattan plots are often used to display significant SNPs. In this context, a Manhattan plot shows the negative logarithm of the p-value for each SNP as a function of genomic location. Due to the hundreds of thousands to millions of SNPs to be tested, the p-value threshold for significance needs to be corrected for multiple testing. Bonferroni correction is a simple approach, which we use in Example 1.3.1 below, where the local significance level is the desired overall significance level α one aims to control divided by the number of tests performed. For example, in the case of $\alpha = 0.05$ and a GWAS where a microarray of one million SNPs is employed, which is typical in human SNP arrays, the threshold is 5×10^{-8}.

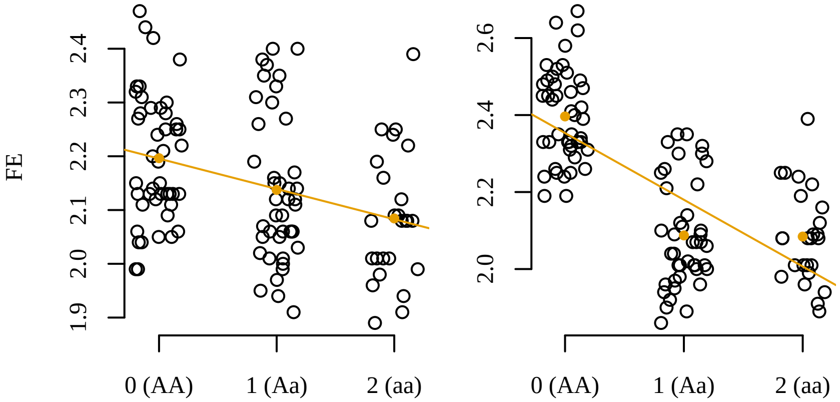


Figure 1.7: *Scatterplot of FE plotted against SNP number 52044 (left) and an altered version of FE plotted against SNP number 52044 (right).*

Example 1.3.1. As an example, we consider the data introduced in Section 1.2. We are interested in comparing the DNA of participating pigs of different races, for a particular trait, in this case FE. If one type of the variant (one allele) is more frequent in feed efficient pigs, the variant is said to be associated with the FE. The associated SNPs are then considered to mark a region of the pig genome that may influence the FE.
For each race, Duroc and Landrace, we fit 31,079 and 35,913 univariate simple linear regression models, respectively, carry out t-tests, and calculate p-values. The resulting Manhattan plots for the Durocs (top) and Landraces (bottom) are shown in Figure B.1 in Appendix B. The SNPs with the most significant association stand out on the plot, since the strongest statistical associations have the smallest p-values and their negative logarithms will be the greatest. The dashed horizontal line represents the threshold for genome-wide significance from the Bonferroni correction method. The dotted horizontal line represents the conventional threshold 5×10^{-8}. We observe only a few strongly associated SNPs for the Landraces (indicated by ○) above the dashed line and none for the Durocs, see Table C.1 in Appendix C for a comprehensive list. ◊

The standard GWAS method based on simple linear regressions has several obvious limitations; in particular, insufficient sample size and control for multiple testing are common problems. Standard variable-by-variable hypothesis testing strategies are based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low. If multiple hypotheses are tested, the probability of producing false positives increases. On the other hand, significance rules, such as Bonferroni correction, can be conservative and might increase the probability of producing false negatives, that is, reduce the statistical power.

Multiple linear regression is an extension of simple linear regression, which allows us to explain the variation in the outcome that can be attributed to variation in the features and the relative contribution of each of the (multiple) features to the total variance explained. However, standard multiple linear regression models result in only non-zero estimates which makes interpretation of the results difficult when the number, p, of features is large. Furthermore, if the number, N, of observations is not much larger than p, the OLS estimates may exhibit a lot of variability resulting in almost sure overfitting. Therefore, they are likely to yield poor predictions. Finally, if $p > N$ the OLS estimates are not unique in which case the method cannot be used at all. However, in the case of $p > N$ it is unlikely that all features play an important role, anyway. Therefore, regularised regression methods have been suggested as an alternative method (also) in GWAS. Different types of regularised regression methods exist with different (dis)advantages. Here, we focus on lasso or ℓ_1-regularised regression and generalised regularisations, such as the group lasso.

There are many alternatives to OLS and regularisation, for example, subset selection and dimension reduction. The subset selection methods use least squares to fit a linear model that contains a subset of the predictors. While they yield simple and easily interpretable models it turns out that shrinking the coefficient estimates via regularisation, can significantly reduce their variance (James et al., 2013). The dimension reduction methods transform the features and then fit linear regression models by means of least squares using the transformed features. While these methods reduce the computing time and storage space required, the new features are not easily interpretable.

While the majority of GWASs used to focus on the effect of individual SNPs, the growing sample size of GWASs has facilitated the discovery of gene-environment interactions (GxE) and even gene-gene interactions (GxG), and empirical evidence shows that such interactions may be an important genetic component underlying complex traits and diseases (Lee et al., 2018; Schrode et al., 2018; Dong et al., 2017). Gene-gene interactions, or *epistasis*, is a phenomenon in genetics in which the expression of one gene is affected by the expression of one or more independently inherited genes. In other words, there is an interaction, either across genes or within them, leading to non-linear effects. Thus, there is biological motivation for considering interactions, when analysing genetic data.

A variety of methods can be used to test for statistical interaction between features that encode the genotype and an outcome. One approach to modelling interaction effects involves regression. However, fitting regression models with interactions is challenging even for a moderate-size number, p, of measured features, as the total number of possible pairwise interactions is $\binom{p}{2} = \frac{1}{2}p(p-1)$. For example, the genetic FEEDOMICS dataset with approximately 30 thousand SNPs results in approximately 450 million potential pairwise interactions.

In addition, exhaustive genome-wide searches for interactions are difficult, due to a lack of power in most datasets and the substantial memory and computing time needed. The methodology proposed here is the following: If a subset of active features can be identified, imposing hierarchical restrictions (Bien et al., 2013) in a stepwise manner, may sufficiently reduce the number of interactions to be included in a regularised regression model. Our proposal in this direction is given in Manuscript I and a general overview of the subject is given in Section 1.4. An overview of a multivariate analogue is given in Section 1.6 and a theoretical proposal is given in Manuscript III.

The motivation for multivariate modelling is illustrated by the simple case of three outcome components and one feature:

Example 1.3.2. Assume that we are interested in the impact of the race of a pig on the FE (`Feed efficiency`), the meat percentage (`meat %`), and the weight (`scanning_weight`). Pairwise scatterplots of the dataset are shown in Figure 1.8: They indicate a pattern of the race where the race simultaneously influences `Feed efficiency` and `meat %` but not necessarily `scanning_weight`. This encourages a joint analysis of `Feed efficiency` and `meat %` and an individual analysis of `scanning_weight`. ◊

Traditional association methods for quantitative traits are mostly designed for use with independent traits. However, genetic association studies in practice often involve multiple traits resulting from a common disease mechanism, and samples for such studies are often stratified based on some trait outcomes. Using information from multiple, simultaneously measured and potentially correlated quantitative traits in a common statistical analysis has many potential advantages (Galesloot et al., 2014), such as increased precision and power (Schmitz et al., 1998) and computational advantages. In such situations, statistical methods using only one of the traits may be inefficient and lead to under-powered tests for detecting genetic associations. On the other hand, estimation and testing procedures for evaluating the shared-association of a genetic marker on the joint distribution of multiple traits may increase the statistical and computational complexity of the analyses massively. In Manuscript II we develop a method for understanding the genetic relationship among multiple, simultaneously measured, and potentially correlated quantitative traits. The features may be inaccessible since our methods work provided that we have access to associated PGSs (polygenic scores) or other summary statistics presumed to be indicative of the correlation structure of the traits. Here, we are interested in identifying potential clusters of the traits that share some genetic component, and we wish to be able to predict the outcome for new data points. Thus, we are interested in the pairwise "relatedness" of the traits in terms of genetic variation, and for that reason we consider PGSs, which marginally serves as a prediction for a trait when taking into account variation in multiple genetic variants.

Linear mixed models (LMMs) have been used in GWASs to model correlated traits. Because of the computational demands involved in the estimation of the parameters of an LMM, and due to the complications arising from misspecification of the covariance structure and the challenges of an unconstrained covariance matrix, we began considering how the problem could be simplified. Our hope was to be able to identify clusters of multiple, simultaneously measured quantitative traits sharing genetic characteristics and to predict the outcome for new data points. Our proposal in this direction is given in Manuscript II. A general overview of the subject of multivariate modelling with PGSs is given in Section 1.5.

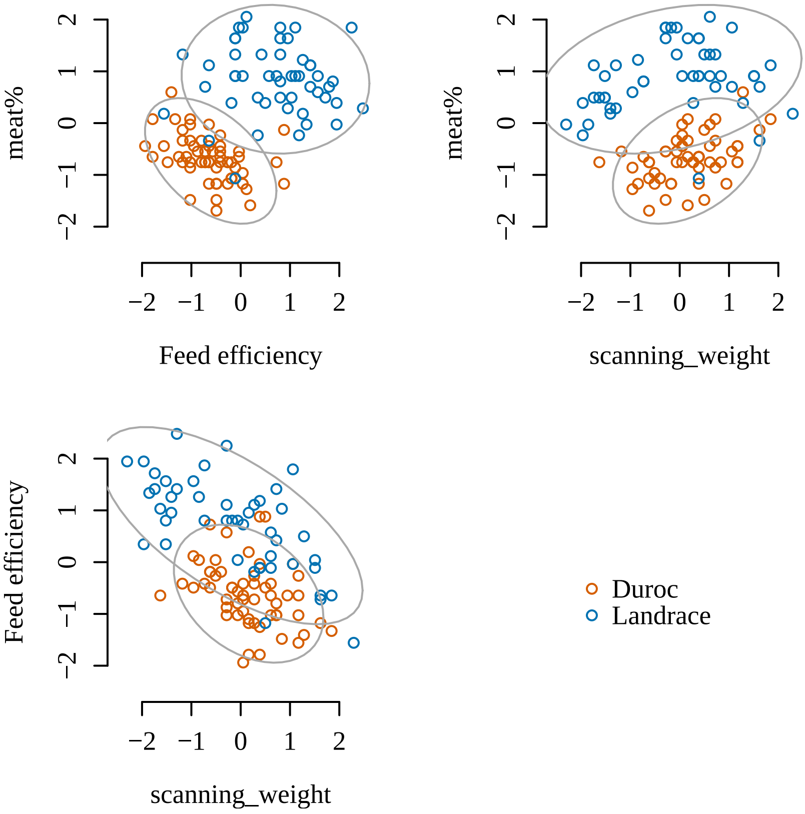


Figure 1.8: *Pairwise scatterplots of pig data with three traits plotted against each other and coloured according to the value of the single feature, the race of the pig. We have added 90% confidence ellipses for each group, with the centre being the group mean, the shape the group covariance matrix, and the radius the square root of the value of the χ^2-distribution with two degrees of freedom at 0.1.* Top left*: This plot suggests that the race influences both the* `Feed efficiency` *and* `meat %`*, since it reveals two clusters of observations complying with the race in* both *the direction of* `Feed efficiency` *and* `meat %`. Top right*: This plot reveals a more faint clustering in the direction of* `meat %` *suggesting that the race influences* `meat %` *and not* `scanning_weight`*.* Bottom left*: This plot reveals no clear relation between* `Feed efficiency`*,* `scanning_weight`*, and race.*

In the following sections, background, overviews, and discussions of our progress in the research objectives are given, and applications to the FEEDOMICS data are presented.

1.4 Hierarchical interactions and regularised regression

In Manuscript I we develop an intuitive method for identifying hierarchical interactions using regularised regression, with the aim of obtaining interpretable models. In Section 1.4.1 and Section 1.4.2 we give a more detailed account of existing methods which form the basis of Manuscript I, in Section 1.4.3 we outline the results of Manuscript I, and in Section 1.4.4 we discuss our results and review perspectives for future research based on the material covered.

In order to properly understand the context of our work, we begin by motivating and introducing hierarchical interaction models in a $p < N$ context and regularised regression method when interactions are not a concern. After doing so, we outline our work on regularised regression with hierarchical interactions.

1.4.1 Hierarchical interaction models

As argued in the introductory Section 1.3 biological and statistical incentive for considering interactions in analyses of genetic data exists. In fact, wrongly omitting interaction terms may reduce the predictive power of models, bias the estimates, and lead to wrong conclusions. When to include interactions in a multivariate regression model is, however, not straightforward. Interpretations of traditional regression analyses do not encourage including an interaction in a model without the corresponding main effects, since main effects can be viewed as deviations from the mean and interactions as deviations from the main effects (Bien and Tibshirani, 2014). Furthermore, it can be argued that large main effects are more likely to lead to important interactions than small (Cox, 1984). Finally, including “too many” interaction terms may unnecessarily complicate the model and its interpretation and is likely to be computationally infeasible. This motivates the notion of *hierarchy*.

Our work with interactions and hierarchical assumptions concerns the development of these notions for regularised regression models. As we study the genomes of pigs to assess the FE, we hope that not all of the 21,640 genes in a pig are directly involved in the process that leads to higher FE. This prompts an underlying assumption of simplicity, and, in statistical terms, one form of simplicity is sparsity. We say that a sparse statistical model is one in which only a relatively small number of parameters (or features) play an important role. The starting point is the *univariate multiple linear regression model*. For N observations of a univariate centred outcome $\boldsymbol{y} = (y_1, \ldots, y_N)$ and p associated features $\mathbf{x}_j = (x_{1j}, \ldots, x_{Nj})$, $j = 1, \ldots, p$, the model takes the form

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{1.2}$$

for all $i = 1, \ldots, N$, where β_j, $j = 1, \ldots, p$ are unknown parameters for the main effects and ϵ_i, $i = 1, \ldots, N$, are error variables. A popular method for estimating the parameters in a linear regression model is the *ordinary least squares* (OLS) method, in which we find the parameters $\boldsymbol{\beta}$ which minimise the *residual sum of squares* (RSS):

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 .$$

If $p > N$ the system is overdetermined, since the number of equations is smaller than the number of unknowns, so there is an infinite set of solutions that make the objective function equal to zero, and these solutions almost surely overfit the data. This motivates the notion of *regularised regression* which involves fitting a model involving all p features where coefficients are shrunken towards (or estimated to be exactly) zero relative to the least squares estimates.

We assume that we have N observations of the univariate outcome $\boldsymbol{y}$ and p associated features $\boldsymbol{X}$ with pairwise interactions. Here, we consider the situation where the outcome is assumed to be quantitative ($\boldsymbol{y} \in \mathbb{R}^N$) and the error distribution is assumed to be Gaussian. The generalised linear model (GLM) (Nelder and Wedderburn, 1972) is a generalisation of ordinary linear regression which, via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value, allows for the outcome to have an error distribution which is not Gaussian. While interaction effects are simply imposed through the link function, it should be noted, that the presence of both the link function and the interaction terms further complicate the interpretation of the relationship between the outcome and the features.

If the features are categorical, they are assumed to be represented by dummy variables, and the pairwise interaction term is formed by multiplying the corresponding variables. In this case, two features are said to interact if the effect of one feature depends on the level(s) of the other. Similarly, a categorical and a quantitative feature are said to interact, if the partial slope corresponding to the quantitative variable depends on the level of the categorical variable. Estimates and hypothesis tests are calculated similarly for both cases (Ekstrøm and Sørensen, 2014). If the features are quantitative, $\boldsymbol{X} \in \mathbb{R}^{N \times p}$, the pairwise interaction term is formed by multiplying the two corresponding features. In this case, testing and interpretation is more complex. A comprehensive source on the treatment of such interactions in multiple regression is found in Aiken and West (1991).

The *pairwise interaction model* takes the form

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{k=1}^{p}\sum_{j<k} x_{ij}x_{ik}\Theta_{jk} + \epsilon_i, \tag{1.3}$$

where $\beta_1, \ldots, \beta_p$ are unknown parameters for the main effects, $\boldsymbol{\Theta} \in \mathbb{R}^{p \times p}$ are unknown parameters for the pairwise interactions, and $\boldsymbol{\varepsilon}$ is a random error variable. We let $\boldsymbol{\Theta}$ represent a symmetric matrix, i.e., $\boldsymbol{\Theta} = \boldsymbol{\Theta}^\top$. The strict inequality in the interaction summation precludes over-parametrisation arising from the inclusion of the same effect twice, e.g. including both $x_{ij}x_{ik}$ and $x_{ik}x_{ij}$.

When aiming to interpret pairwise interaction parameters in a linear regression model it is necessary to consider the concept of hierarchy. For example, are we to include (both) main effects when associated pairwise interactions are estimated to be non-zero?

In the following we present model restrictions in a form which makes it possible to specify a regularised regression procedure which produces sparse interaction models that honour these different hierarchical restrictions. We define the following hierarchical, anti-hierarchical, and non-hierarchical restrictions, replicated from Manuscript I:

Strong hierarchy There are interactions only among pairs of non-zero main effects,

$$\mathrm{H_S}: \quad \boldsymbol{\Theta}_{jk} \neq 0 \quad \Rightarrow \quad \boldsymbol{\beta}_j \neq 0 \text{ and } \boldsymbol{\beta}_k \neq 0.$$

Weak hierarchy Each interaction has at least one of its main effects present,

$$\mathrm{H_W}: \quad \boldsymbol{\Theta}_{jk} \neq 0 \quad \Rightarrow \quad \boldsymbol{\beta}_j \neq 0 \text{ or } \boldsymbol{\beta}_k \neq 0.$$

Anti-hierarchy Interactions are only among pairs of main effects which are not present,

$$\mathrm{H_A}: \quad \boldsymbol{\Theta}_{jk} \neq 0 \quad \Rightarrow \quad \boldsymbol{\beta}_j = 0 \text{ and } \boldsymbol{\beta}_k = 0.$$

Pure interactions There are no main effects present, only interactions,

$$\mathrm{H_I}: \quad \boldsymbol{\beta}_j = 0 \quad \forall j = 1, \ldots, p.$$

Pure main effects There are no interactions present, only main effects,

$$\mathrm{H_M}: \quad \boldsymbol{\Theta}_{jk} = 0 \quad \forall j,k = 1, \ldots, p.$$

No hierarchy There are no restrictions to the presence of main effects and interactions, $\mathrm{H_N}$.

While the methods developed in Manuscript I allow for all of the above mentioned hierarchies an anti-hierarchical or pure interactions structure is not computationally tractable, and often conceptually unrealistic. A pure main effects structure is irrelevant in our search for interactions. Therefore, we make no further mention of these structures in this section.

In general, we may denote by $\mathcal{M}_\mathrm{H}$ and $\mathcal{I}_\mathrm{H}$ the sets of main effects and interactions, respectively, to be included in the model *subject to one of the hierarchical restrictions*. Then, the restrictions above result in the following *hierarchical pairwise interaction model*:

$$y_i = \sum_{j \in \mathcal{M}_{\mathrm{H}}} x_{ij}\beta_j + \sum_{k=1}^{p} \sum_{j<k} \mathbb{1}_{\{(j,k)\in\mathcal{I}_{\mathrm{H}}\}} x_{ij} x_{ik} \Theta_{jk} + \epsilon_i. \tag{1.4}$$

1.4.2 Regularised regression

When we talk about fitting a regularised regression model we refer to a model in which only a relatively small number of features are active, that is, a model which involves only a subset of the features. The lasso (least absolute shrinkage and selection operator) is a regression analysis method (Tibshirani, 1996) which imposes an ℓ_1 penalty on their size, thereby, shrinking the regression coefficients towards (or estimated to be exactly) zero relative to the least squares estimates.

We assume that we have N observations of a univariate outcome $\boldsymbol{y} \in \mathbb{R}^N$ and p associated features $\boldsymbol{X}$. The objective of *lasso* or *ℓ_1-regularised regression* is to solve an optimisation problem of the form

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{minimise}} \quad \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \\ &\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t, \end{aligned} \tag{1.5}$$

where t is a parameter determined e.g., by cross-validation, before the estimation process is initialised. Selecting the value of t is a trade-off between the fit to data and the predictive power of the model. We recognise the objective function to be minimised over the parameter vector $\boldsymbol{\beta}$ as a residual sum of squares (RSS), that is, the lasso coefficients minimise a penalised RSS. In the left panel of Figure 1.9 we illustrate the estimation for the lasso regression in the case of two features. The grey ellipses are contours of the RSS function centred at the point $\hat{\boldsymbol{\beta}}^{\mathrm{O}}$ which corresponds to the (unconstrained) OLS estimate. The area bounded by the orange lines is the constraint region $|\beta_1| + |\beta_2| \leq t$ and the lasso method selects the point on the contour which first hits the constraint region when expanding the contours from the centre $\hat{\boldsymbol{\beta}}^{\mathrm{O}}$. For comparison, we illustrate the estimation for the *ridge* regression in the right panel. It solves an optimisation problem very similar to (1.5):

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{minimise}} \quad \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \\ &\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq t^2. \end{aligned} \tag{1.6}$$

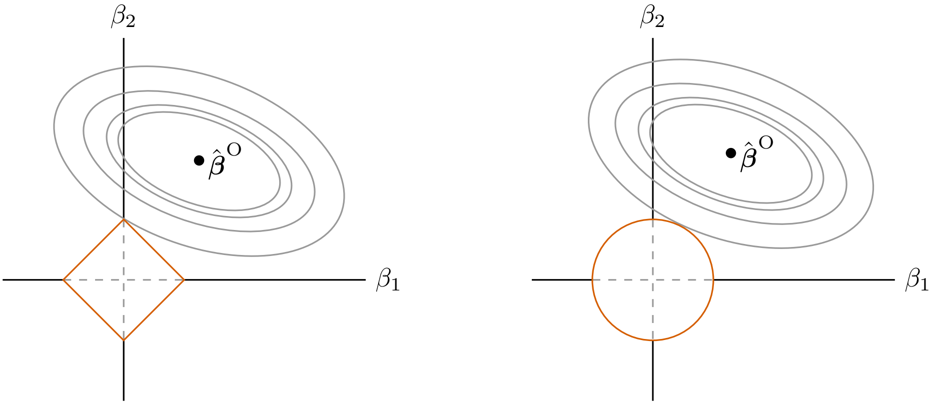

β_2
$\hat{\boldsymbol{\beta}}^{\mathrm{O}}$
β_1
β_2
$\hat{\boldsymbol{\beta}}^{\mathrm{O}}$
β_1

Figure 1.9: *Illustration of estimation via lasso regression (left) and ridge regression (right). The grey ellipses are contours of the residual-sum-of-squares function centred at the point* $\hat{\boldsymbol{\beta}}^{\mathrm{O}}$ *which corresponds to the OLS estimate and the area bounded by the orange lines is the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively.*

We can write the lasso problem in the Lagrangian form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{1.7}$$

By Lagrangian duality, there is a one-to-one correspondence between the constrained problem (1.5) and the Lagrangian form (1.7).

The method translates each coefficient by a constant factor λ, truncating at zero: Since the function $|\beta_j|$ is not differentiable in zero, lasso has the ability to letting coefficients equal zero. The ℓ_1 penalty $\sum_{j=1}^{p} |\beta_j|$ makes the solution non-linear in the y_i, and it has no closed form expression. The penalty parameter λ can be determined by a model validation technique such as cross-validation, and selecting the value of λ translates to selecting a proper amount of regularisation and is, as such, a trade-off between data fitting and sparsity.

For computational reasons it is recommended that quantitative features are centred before the optimisation problem is solved, see Aiken and West (1991), such that each column has mean zero, that is, $\frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0$. If the features are not measured in the same units, it is also recommended that the features are scaled such that each column has unit variance, that is, $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$. Otherwise the lasso solution depends on the scale since lasso puts constraints on the size of the coefficients associated to each feature. For the sake of simplicity we assume in the following that the outcome values are centred before the optimisation problem is solved, that is, $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i = 0$, and the intercept term is omitted from the model. In linear regression models, this condition is not a restriction, since given an optimal solution, $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}})$, obtained from the centred data, an optimal solution, $(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}})$, for the uncentred data is easily recovered:

$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}_0 = \bar{y} - \sum_{j=1}^{p} \bar{\boldsymbol{X}}_j \tilde{\boldsymbol{\beta}}_j$, where $\bar{y}$ and $\bar{\boldsymbol{X}}_j$, $j = 1, \ldots, p$, are the original means.

Example 1.4.1. We consider as outcome, $\mathbf{y}$, the FE and as associated features, $\mathbf{X}$, the 31,079 and 35,913 SNPs for the Durocs and Landraces, respectively. We perform lasso regression on each race separately using the R package glmnet (Friedman et al., 2010). In Figure 1.10 we display the cross-validation curve (dots), and upper and lower standard deviation curves (grey bars) along the λ sequence. Two selected λ's are indicated by the vertical dotted lines. One is λ_{min}, which is the value of λ that gives minimum mean cross-validated error. The other is λ_{1se}, which gives the most regularised model such that the error is within one standard error of the minimum. That is, where λ_{min} selects the *best model* in terms of mean squared error, λ_{1se} selects the more parsimonious model with error no more than one standard error above the error of the *best model*. Often, a more conservative model is desired, and in that case λ_{1se} used, as it acknowledges the fact that the cross-validation curves are estimated with error, and errs on the side of parsimony (Hastie et al., 2009). However, we ought to be liberal when including main effects to allow more interactions to be considered and avoid them being disregarded because of low power. Therefore, λ_{min} is used as shrinkage and selection criterion in this example as well as in the later proposed methods. We observe that 32 SNPs are selected for the Durocs and 48 SNPs are selected for the Landraces at λ_{min}. A comprehensive list of selected SNPs is given in Table C.1 in Appendix C. We observe that notably more SNPs are selected by the lasso than the GWAS method used in Example 1.3.1. In the Manhattan plots in Figure B.1 in Appendix B we indicate by $+$ SNPs selected by the lasso. $\diamondsuit$

1.4.3 A framework for hierarchical regularised regression

We are now ready to describe our efforts in developing a method for identifying hierarchical interactions using regularised regression. Our work is detailed in Manuscript I.

Fitting regression models with interactions is challenging even for a moderate-size number of measured features p, as the total number of possible pairwise interactions is $\binom{p}{2} = \frac{1}{2}p(p-1)$. For the genetic dataset presented in Section 1.2 with approximately 25 thousand SNPs, this results in more than 300 million potential pairwise interactions. Including all these interactions in a regression model is, of course, not meaningful. Therefore, we are interested in "sufficiently" reducing the number of interactions to be included in a regularised regression model by imposing hierarchical restrictions in a stepwise manner. The first step in this direction is extending the lasso optimisation problem to the pairwise interaction model (1.3), such that a penalty is imposed on the interaction coefficients, thereby shrinking them towards (or estimated to be exactly) zero. We choose to treat main effects and interactions equally and impose the same penalty on main effects and interactions, thereby keeping down the number

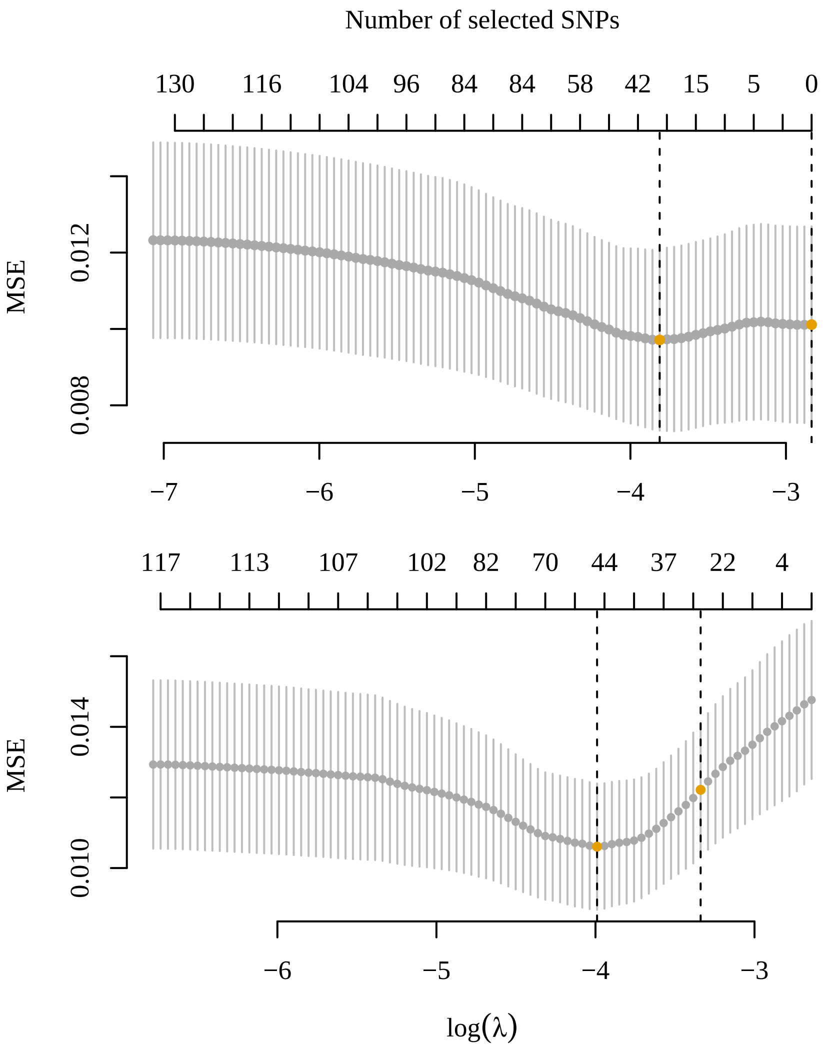


Figure 1.10: *Cross-validation curves (dots), and upper and lower standard deviation curves (grey bars) along the λ sequence (bottom axis) and corresponding sequence of number of selected SNPs (upper axis) for the Durocs (top) and Landraces (bottom).* λ_{min} *and* λ_{1se} *are indicated by the orange dots and vertical dotted lines.*

of hyperparameters. The optimisation problem takes the form

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}}{\text{minimise}} \left\{ q(\boldsymbol{\beta},\boldsymbol{\Theta}) + \lambda\sum_{j=1}^{p}|\beta_j| + \lambda\sum_{k=1}^{p}\sum_{j<k}|\Theta_{jk}| \right\},$$

where $q(\boldsymbol{\beta}, \boldsymbol{\Theta})$ denotes the loss function

$$q(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{k=1}^{p} \sum_{j<k} x_{ij}x_{ik}\Theta_{jk} \right)^2 .$$

The remaining challenge is now to impose hierarchical restrictions in the lasso optimisation problem for the pairwise interaction model. Our starting point is the *adaptive lasso* proposed by Zou (2006), where adaptive weights are used for penalising different coefficients in the ℓ_1 penalty. The adaptive lasso solves an optimisation problem of the form

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimise}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \gamma_j |\beta_j| \right\}, \tag{1.8}$$

where the penalty parameter $\lambda \geq 0$ is determined by cross-validation and the quantity $\boldsymbol{\gamma} \in \mathbb{R}^p$, $\gamma_j \geq 0, j = 1, 2, \ldots, p$, is a penalty modifier. Inspired by this, we extend the definition of the adaptive lasso to the framework of interactions. For the pairwise interaction model (1.3) the adaptive lasso estimates are determined by solving the optimisation problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\Theta} \in \mathbb{R}^{p \times p}}{\text{minimise}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{k=1}^{p} \sum_{j<k} x_{ij}x_{ik}\Theta_{jk} \right)^2 \right.$$
$$\left. + \lambda \sum_{j=1}^{p} \gamma_j |\beta_j| + \lambda \sum_{k=1}^{p} \sum_{j<k} \tau_{jk} |\Theta_{jk}| \right\}.$$

Now, defining the penalty modifiers $\gamma_j \geq 0$ and $\tau_{jk} \geq 0, j, k = 1, 2, \ldots, p$, is key for achieving our objective of constructing a procedure which ensures interpretation in terms of hierarchy. While, typically, the primary motivation for the adaptive lasso compared to the usual lasso is that it enjoys the oracle properties (Zou, 2006), we are going to take advantage of the actual adaptiveness, that is, the possibility of using different weights for penalising coefficients. To obtain hierarchical interactions, we make the following observations. The method works for $\lambda = 0$ whenever $p < N$. However, our procedure is targeted towards datasets with $p \gg N$, and therefore, we consider only $\lambda > 0$. When $\gamma_j = 0$, the jth regularisation term disappears. That is, feature j is not penalised but always included in the model. When $\gamma_j = \infty$ feature j is always excluded, and for all $j = 1, 2, \ldots, p$ for which γ_j are equal to the same constant value, the corresponding features are equally penalised. Similar observations can be made for τ_{jk}. By these observations, we are able to formally define, for all $j = 1, \ldots, p$,

$$\gamma_j = \mathbb{1}_{\{j \notin \mathcal{M}_{\mathrm{H}}\}},$$

where $\mathcal{M}_{\mathrm{H}}$ is the index set of features to be included in the model *subject to* one of the hierarchical restrictions, $\mathrm{H_S}$ or $\mathrm{H_W}$, and for all $j, k = 1, \ldots, p$,

$$\tau_{jk} = \mathbb{1}_{\{(j,k) \in \mathcal{I}_{\mathrm{H}}\}},$$

where $\mathcal{I}_{\mathrm{H}}$ is the index set of pairwise interactions (pairs of features) to be included in the model *subject to* one of the hierarchical restrictions, $\mathrm{H_S}$ or $\mathrm{H_W}$, analogously to (1.4).

Finally, it remains to estimate the sets $\mathcal{M}_{\mathrm{H}}$ and $\mathcal{I}_{\mathrm{H}}$. The following two-step procedure, replicated in a slightly shortened form from Manuscript I, is our proposed method.

Step 1 Assume the pure main effect model,

$$y_i = \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_i,$$

for which the lasso estimates are determined by solving the optimisation problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimise}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\},$$

for some $\lambda_1 > 0$. Define by $\mathcal{S} = \{k : \beta_k \neq 0\}$ the set of main effects, that is, the set of non-zero coefficients, which is estimated by $\mathcal{S}^{(\lambda_1)} = \{k : \hat{\beta}_k^{(\lambda_1)} \neq 0\}$, that is, the set of non-zero estimates, corresponding to the main effects selected by lasso.

Step 2 Define by $\mathcal{M}_{\mathrm{H}}^{(\lambda_1)}$ and $\mathcal{I}_{\mathrm{H}}^{(\lambda_1)}$ the sets of main effects and interactions, respectively, to be included in the model subject to one of the hierarchical restrictions, $\mathrm{H_S}$ or $\mathrm{H_W}$, for a given value of λ_1. Then, assume the pairwise interaction model,

$$y_i = \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j + \sum_{k=1}^{p}\sum_{j<k} \mathbb{1}_{\left\{(j,k)\in\mathcal{I}_{\mathrm{H}}^{(\lambda_1)}\right\}} x_{ij}x_{ik}\Theta_{jk} + \boldsymbol{\varepsilon}_i,$$

for which the adaptive lasso estimates are determined by solving the optimisation problem

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}}{\text{minimise}} \Bigg\{ & \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{k=1}^{p}\sum_{j<k} \mathbb{1}_{\left\{(j,k)\in\mathcal{I}_{\mathrm{H}}^{(\lambda_1)}\right\}} x_{ij}x_{ik}\Theta_{jk} \right)^2 \\ & + \lambda_2 \sum_{j\notin\mathcal{M}_{\mathrm{H}}^{(\lambda_1)}} |\beta_j| + \lambda_2 \sum_{k=1}^{p}\sum_{j<k} \mathbb{1}_{\left\{(j,k)\in\mathcal{I}_{\mathrm{H}}^{(\lambda_1)}\right\}} |\Theta_{jk}| \Bigg\}, \end{aligned}$$

for some $\lambda_2 > 0$.
We are finally able to identify the set $\mathcal{S}_{\mathrm{H}} = \{k : \boldsymbol{\beta}_k \neq 0\}$ of main effects relevant under the model and estimated by $\mathcal{S}_{\mathrm{H}}^{(\lambda_2)} = \{k : \hat{\beta}_k^{(\lambda_2)} \neq 0\}$ and the set $\mathcal{R}_{\mathrm{H}} = \{j,k : \Theta_{jk} \neq 0\}$ of interactions relevant under the model and estimated by $\mathcal{R}_{\mathrm{H}}^{(\lambda_2)} = \{j,k : \hat{\Theta}_{jk}^{(\lambda_2)} \neq 0\}$.

This concludes our exposition of the motivation and framework of Manuscript I. We have focused on aspects relating to the primary contribution: the proposal of a method for identifying pairwise interactions in a regularised regression model under different assumptions of hierarchy using the penalty modifier of the adaptive lasso. Example 1.4.2 shows how the method is used in a concrete case.

Example 1.4.2. We apply the two-step procedure to the SNP data using our `R` package `ilasso` (Buchardt and Ekstrøm, 2021) and apply the method separately for the two breeds (Duroc and Landrace). In the first step of the procedure the same 32 and 48 SNPs are selected for the Durocs and Landraces, respectively, as by the lasso in Example 1.4.1. In the second step, under the assumption of strong hierarchy, no interactions are identified for neither Durocs nor Landraces. It is not unreasonable to attribute part of the lack of identified interactions to low statistical power in this case, and we refer to Manuscript I for a real data example in which interactions are in fact identified by the two-step procedure. ◊

1.4.4 Discussion of the two-step procedure

The majority of GWASs used to focus on the effect of individual SNPs. However, the growing sample size of GWASs has facilitated the discovery of gene-environment and gene-gene interactions, and empirical evidence shows that such interactions might be an important genetic component underlying complex traits and diseases.

How to systematically include interacting features in regularised regression models is not clear. Including interactions in regularised regression models under hierarchical restrictions will facilitate simultaneous feature selection, parameter estimation, and interpretation of the results.

We have proposed a two-step regularised regression procedure for identification of pairwise interactions, when strong or weak hierarchy is assumed. The procedure uses lasso and adaptive lasso to screen for interactions and fitting hierarchical pairwise interaction models. Bien et al. (2013) propose a penalised regression method which produces sparse estimates under weak or strong hierarchy by adding a set of convex constraints to lasso. The method is implemented in the R package `hierNet`, but the implemented algorithm is not suitable for large scale problems. Other methods make use of the group-lasso and overlapped group-lasso to select interactions and enforce hierarchy (Lim and Hastie, 2015; Meier et al., 2008), and then others use stepwise procedures (Bickel et al., 2010; Park and Hastie, 2008) and two-step procedures which always enforces strong hierarchy (Wu et al., 2010).

Through simulation studies (see Manuscript I) we have shown that the computing time for our method is considerably reduced compared to that of `hierNet` and is highly satisfactory even for fairly large data. Furthermore, the two-step procedure works very well, both in terms of false discovery rate and recall with respect to interactions, under the assumption of strong hierarchy whether the truth is strong or weak hierarchy. Under the assumption of weak or no hierarchy the recall with respect to interactions is good, however, the false discovery rate is high. There is no consequence other than unnecessary computing time of imposing a hierarchical restriction, when no interactions are present. This was demonstrated in Example 1.4.2 where no interactions were identified, and the same SNPs were identified as by the usual lasso. The above observations led us to the recommendation of a purpose driven choice of hierarchy if the underlying hierarchical structure of data is unknown. Specifically, if controlling the expected proportion of false interaction discoveries is the main concern, assuming strong hierarchy is recommended. If discovering the true interaction is prioritised but the false discovery rate of interactions is still a concern, assuming weak hierarchy is recommended. If discovering the true interaction is prioritised and neither the false discovery rate of interactions nor the computing time is any concern, assuming no hierarchy is recommended.

Utilising the (adaptive) lasso proved very useful for simultaneous estimation and variable selection under hierarchical restrictions. However, the two-step procedure has two major drawbacks. First, it does not necessarily possess the oracle property; that is, we have not shown that it performs (in terms of feature selection) as well as the true underlying model when the number of observations goes to infinity. Secondly, it is somewhat indifferent to grouping effects, that is, the selection among a set of strong but correlated features. The adaptive lasso yields consistent estimators of the parameters, that is, it is consistent for variable selection as it includes only the correct subset of variables and for model selection as the MSE of the parameters is low, while retaining the attractive convexity property of the lasso (Zou, 2006). However, these optimality properties are likely no longer true, when using cross-validation on the same data twice. It would be of interest to understand under which conditions leaving some data out of the selection in the first round to be used for cross-validation in the second round is reasonable (in terms of power). Nonetheless, our efforts have aided in setting up a framework which may be helpful for future research efforts. In particular, our current work may be a stepping stone towards application of the two-step procedure in the case of correlated features: The doubly regularised technique *elastic net* proposed by Zou and Hastie (2005) is able to select groups of highly correlated features, but it does not necessarily possess the oracle property. However, the adaptive elastic net proposed by Zou and Zhang (2009) inherits some of the desirable properties of the adaptive lasso and elastic net. In particular, when the correlations among the features is high, the adaptive elastic net can significantly improve the prediction accuracy and has the oracle property under certain regularity conditions. Here, we have treated the choosing of the regularisation parameter primarily as a means to an end. Choosing the right penalty parameter is, however, difficult and the cross-validated choices often include too many features. It would be of interest to apply, e.g., stability selection (Meinshausen and Bühlmann, 2010) for choosing the right penalty parameters in the two-step procedure.

1.5 Multivariate modelling with polygenic scores

In Manuscript II we propose a method for utilising polygenic scores (PGSs) as a means for inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits in a multivariate GWAS. Here, the term "multivariate" refers to multiple outcome components or responses. In Section 1.5.1 we give a more detailed account of existing methods which form the basis of Manuscript II, in Section 1.5.2 we outline the results of Manuscript II, and in Sections 1.5.3 and 1.5.4 we discuss our results and review perspectives for future research based on the material covered.

In order to properly understand the context of our work, we begin by giving a presentation of some fundamental statistical assumptions and an introduction to polygenic scores. After doing so, we outline our work on using polygenic scores to infer a genetic relationship among multiple traits.

1.5.1 Statistical assumptions and polygenic scores

As argued in the introductory Section 1.3 there is a statistical incentive for considering information from multiple traits in analyses of genetic data. In fact, statistical methods using only one of simultaneously measured traits may be inefficient and lead to under-powered tests for detecting genetic associations (Schmitz et al., 1998). Procedures that simultaneously evaluate genetic associations on the joint distribution of multiple traits may, however, immensely increase the statistical and computational complexity of the analyses. More recently methods using GWAS summary statistics have been proposed, and our work falls into this category as we utilise PGSs for inferring genetic correlation between traits.

We focus on linear regression models, which are often suitable when the outcome is quantitative, and when the error distribution is approximately Gaussian. We consider N independent samples, $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^q$, of a q-dimensional random variable distributed according to a multivariate normal distribution and define a matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^\top \in \mathbb{R}^{N \times q}$ from the q outcome components. We also define a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times p}$ of N samples of p associated features. We assume that the data generating process can be described by a linear regression model of the form

$$\mathbf{Y} = \mathbf{M} + \mathbf{XC} + \mathbf{E},$$

where $\mathbf{M} = \mathbf{1}_N \boldsymbol{\mu}^\top \in \mathbb{R}^{N \times q}$ is a matrix of intercepts with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^\top$, $\mathbf{C} \in \mathbb{R}^{p \times q}$ is a matrix of regression coefficients, and $\mathbf{E} \in \mathbb{R}^{N \times q}$ are independent Gaussian random errors. However, we are interested in a more constrained scenario, where we may not have access to data on SNP-level; we have access to *polygenic scores (PGSs)* which are linear transformations between multiple genetic variants to scores that *summarise* the estimated effect of, e.g., SNPs. Typically they are calculated as a weighted sum of trait-associated alleles, which are one of two, or more, forms of a given gene variant. That is, they are constructed from the "weights" derived from a GWAS, or from some form of machine learning algorithm. Thus, a PGS reflects an estimated genetic predisposition for a given trait without taking environmental factors into account, and it can be used as a numeric predictor for that trait.

PGSs are widely used and available in many fields such as animal breeding and plant breeding, and they are also used in human genetics. In Manuscript II we allow for a set of PGSs to be provided from an unrelated GWAS or a public resource. For the purpose of completeness, we present a simple method of construction here.

In general, the weights used for the PGSs are estimated using some form of regression analysis, and since the total number of features p is usually larger than the sample size N, one cannot use OLS for estimating the parameters simultaneously. Various methodologies deal with this problem as well as how to generate the weights of the SNPs and how to determine which $s \leq p$ features should be included (e.g., Euesden et al. (2014); Wray et al. (2014)). To keep the setup simple we define, for each trait $l = 1, \ldots, q$, weights $w_{1l}, \ldots, w_{pl}$ as the marginal effects on the trait $\mathbf{y}_l$, estimated separately from a univariate simple linear regression model of the form (1.1). This way, we have to do $p \cdot q$ separate univariate simple linear regression analyses but we avoid making dimensionality reduction of the SNPs before computing the PGSs. That is, for each outcome component $\mathbf{y}_l \in \mathbb{R}^N$, $l = 1, \ldots, q$, and each feature $\mathbf{x}_j \in \mathbb{R}^N$, $j = 1, \ldots, p$, we estimate a univariate simple linear regression model of the form

$$\mathbf{y}_l = w_{0l} + \mathbf{x}_j w_{jl} + \mathbf{e}_l,$$

where $w_{0l} \in \mathbb{R}$ is the intercept, $w_{jl} \in \mathbb{R}$ is a regression coefficient, and $\mathbf{e}_l \in \mathbb{R}^N$ is a vector of independent Gaussian random errors. For ease of notation, we label all regression coefficients as elements in a $p \times q$ matrix $\mathbf{W}$, which is not to be mistaken for a matrix of coefficients from a multivariate multiple linear regression.

Finally, we define the PGS for each outcome component $l = 1, \ldots, q$ and each individual $i = 1, \ldots, N$ by

$$z_{il} = \sum_{j=1}^{p} x_{ij} \hat{w}_{jl},$$

where $\hat{w}_{jl}$ is the maximum likelihood estimate of w_{jl}.

In the following we assume that a set of PGSs has been provided, and we use this set of estimated PGSs as numeric predictors (features). That is, we assume that we have access to N samples of exactly one feature per trait, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N) \in \mathbb{R}^{N \times q}$. Therefore, we consider applying a statistical model, in particular, a linear regression model of the form

$$\vec{\mathbf{Y}} = \vec{\mathbf{Z}}\vec{\mathbf{B}} + \vec{\mathbf{E}}, \tag{1.9}$$

where $\vec{\mathbf{Y}} = \text{vec}(\mathbf{Y}^\top) \in \mathbb{R}^{Nq}$ denote the row-wise vectorisation of $\mathbf{Y}$ obtained by stacking the columns of the matrix $\mathbf{Y}^\top$, that is,

$$\vec{\mathbf{Y}} = \left(y_{11}, y_{12}, \ldots, y_{1q}, y_{21}, y_{22}, \ldots, y_{2q}, \ldots, y_{N1}, y_{N2}, \ldots, y_{Nq}\right)^\top.$$

Furthermore, $\vec{\mathbf{B}} = (\xi_1, \beta_1, \xi_2, \beta_2, \ldots, \xi_q, \beta_q)^\top \in \mathbb{R}^{2q}$ is a vector of unknown intercepts $\xi_1, \ldots, \xi_q$ and regression coefficients $\beta_1, \ldots, \beta_q$ and $\vec{\mathbf{Z}} \in \mathbb{R}^{Nq \times 2q}$ is a corresponding design matrix of features, such that the first column of $\vec{\mathbf{Z}}$ is a column of ones and the second corresponds to the row-wise vectorisation of $\mathbf{Z}$. $\vec{\mathbf{E}} \in \mathbb{R}^{Nq}$ is a vector of Gaussian random errors, that is,

$$\vec{\mathbf{E}} \sim \mathcal{N}_{Nq}\left(\mathbf{0}_{Nq}, \boldsymbol{\Omega}\right),$$

where $\mathbf{0}_{Nq} \in \mathbb{R}^{Nq}$ is a vector of zeros and $\boldsymbol{\Omega} \in \mathbb{R}^{Nq \times Nq}$ is a general covariance matrix. Possible structures for $\boldsymbol{\Omega}$ may be a multiple of the identity, a diagonal, or a general positive definite matrix. As mentioned, standard multiple linear regression models assume that the underlying observations are independent, which is reflected in the assumption that errors of the outcome are independent, that is, $\boldsymbol{\Omega} = \sigma_r^2 \mathbf{I}_{Nq \times Nq}$ for some $\sigma_r^2 > 0$. Since we are interested in potentially clustered traits with unequal variability across clusters, we assume that the outcome components are both correlated and heteroscedastic. Therefore, we assume that the covariance matrix, $\boldsymbol{\Omega} \in \mathbb{R}^{Nq \times Nq}$, is a block diagonal matrix of the form

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{R} & 0 & \cdots & 0 \\ 0 & \mathbf{R} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{R} \end{bmatrix}, \tag{1.10}$$

where $\mathbf{R} \in \mathbb{R}^{q \times q}$ are the residual covariances among traits within each individual $i = 1, \ldots, N$.

If the among-trait and within-individual correlation, $\boldsymbol{\Omega}$, of the errors is known and $\boldsymbol{\Omega} \neq \mathbf{I}$, the generalised least squares (GLS) estimator is the best linear unbiased estimator of the unknown parameters $\vec{\mathbf{B}}$ in a linear regression model when there is a certain known degree of correlation between the residuals. In these cases, ordinary least squares (OLS) and weighted least squares (WLS) can be statistically inefficient (in terms of mean squared error), or even give misleading inferences (Baltagi, 2008). However, in our case, the covariance matrix $\boldsymbol{\Omega}$ is unknown. Therefore, the methodology proposed is the following: If the structure of $\boldsymbol{\Omega}$ can be approximated, then we can get a consistent (in terms of structure) estimate using the feasible generalised least squares (FGLS) estimator. This is one of two fundamental ideas of Manuscript II and it is further detailed in Section 1.5.2 below. The other fundamental idea of Manuscript II is that the structure of $\mathbf{\Omega}$ can be approximated by means of PGSs.

1.5.2 A framework for multivariate modelling with PGSs

We are now ready to describe our efforts in developing a method for utilising PGSs as a means for inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits and for utilising the inferred clusters for simultaneous modelling of clustered traits.

Our initial motivation for considering PGSs was the presumption that much of the variability observed in a trait is attributable to genetic differences, i.e., heredity. Therefore, the traits may be correlated via the PGS, and it may be reasonable to assume that the underlying independence relations for the traits can be approximated by the relations observed in the PGSs.

Graphs give a powerful way of representing independence relations and computing conditional probabilities among a set of random variables, and, in order to understand the context of our work, we introduce some graphical terminology and properties. For finitely many random variables $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_q)$ with index set $\mathbf{V} := \{1, \ldots, q\}$, an undirected graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ on a set of nodes $\mathbf{V}$ is a pair $(\mathbf{V}, \mathcal{E})$ with $\mathcal{E} \subseteq \mathbf{V} \times \mathbf{V}$, where the elements of $\mathcal{E}$ are referred to as edges. Two nodes i and j are adjacent if either $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$. A *path* in $\mathcal{G}$ is a sequence of (at least two) distinct nodes $i_1, \ldots, i_m$, such that there is an edge between i_k and i_{k+1} for all $k = 1, \ldots, m-1$. A graph $\mathcal{G}_g = (\mathbf{V}_g, \mathcal{E}_g)$ is called a *subgraph* of $\mathcal{G}$ if $\mathbf{V}_g = \mathbf{V}$ and $\mathcal{E}_g \subseteq \mathcal{E}$, and a *connected component* of a graph is a maximal subgraph in which there exists a path between any two nodes.

Now, let $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ be an undirected graph with $\mathbf{V} := \{1, \ldots, q\}$ and corresponding random variables $\mathbf{Z} = (\mathbf{Z}_\alpha)_{\alpha \in \mathbf{V}}$ representing the system of PGSs. Then the PGS for a certain trait is represented by a node, and an edge represents a pair of traits which are "related" in terms of PGSs. In Figure 1.11 we show an example of a system of five PGSs in an undirected graph with the lines indicating two connected components and a singleton.

In order to formalise the "relatedness" we recall the notion of conditional independence.

Definition 1.5.1 ((Lauritzen, 1996), p. 28)**.** If Z_1, Z_2, and Z_3 are random variables with joint distribution P and joint density with respect to a product measure, we say that Z_1 *is*

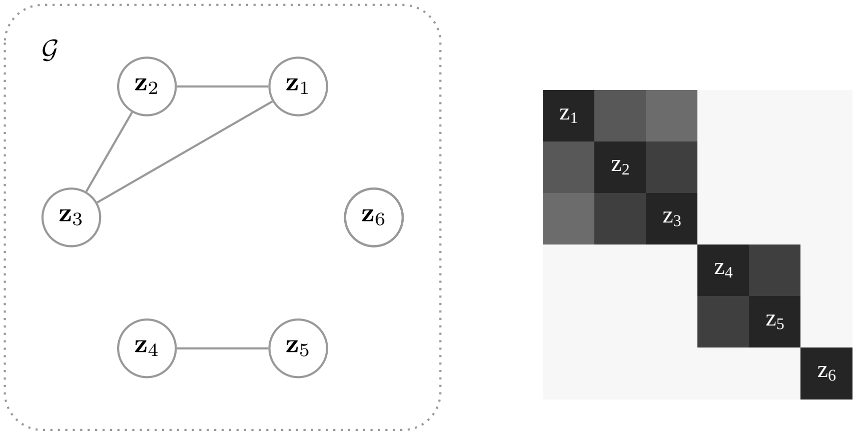


Figure 1.11: *Example of a system of six PGSs in an undirected graph (left). The solid lines indicate two connected components; one containg $\mathbf{z}_1, \mathbf{z}_2$, and $\mathbf{z}_3$, another containg $\mathbf{z}_4$ and $\mathbf{z}_5$. The heatmap (right) is generated using the sparse precision matrix, where dark colours indicate large values and light colours represent values closer to zero. A connection between two nodes i and j indicates a non-zero value for the corresponding element in the estimated precision matrix.*

conditionally independent of Z_2 given Z_3 under P, and we write $Z_1 \perp\!\!\!\perp Z_2 \mid Z_3$, if,

$$f_{Z_1 Z_2 | Z_3}(z_1, z_2 \mid z_3) = f_{Z_1|Z_3}(z_1 \mid z_3) f_{Z_2|Z_3}(z_2 \mid z_3),$$

holds P almost surely.

Next, we recall a result which will help make clustering of traits sharing genetic characteristics a possibility. We assume that $\mathbf{Z}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ assumed to be regular such that the precision matrix $\mathbf{P} = \boldsymbol{\Sigma}^{-1}$ is well defined.

Proposition 1.5.1 ((Lauritzen, 1996), Proposition 5.2)**.** Assume that $\mathbf{Z} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is regular. Then it holds for $i, j \in \mathbf{V}$ with $i \neq j$ that

$$\mathbf{Z}_i \perp\!\!\!\perp \mathbf{Z}_j \mid \mathbf{Z}_{\mathbf{V} \setminus \{i,j\}} \Leftrightarrow p_{ij} = 0,$$

where $\mathbf{P} = \{p_{ij}\}_{i,j \in \mathbf{V}} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix of the distribution.

In words, Proposition 1.5.1 states that for a matrix of random variables following a multivariate Gaussian distribution, the ijth component of the corresponding precision matrix is zero if and only if the variables i and j are conditionally independent, given the others. To better understand how precision matrices translate to connected graphical models, consider the mock model of Figure 1.11 with six PGSs. The heatmap (right) is generated using the precision matrix, where dark colours indicate large values and light colours represent values closer to zero.

A connection between two nodes i and j corresponds to a non-zero value for the corresponding element in the estimated precision matrix. The precision matrix is divided into two diagonal blocks of sizes three and two, respectively, corresponding to the two connected components of the graph (left); one containg $\mathbf{z}_1, \mathbf{z}_2$, and $\mathbf{z}_3$, another containg $\mathbf{z}_4$ and $\mathbf{z}_5$. The two blocks in the precision matrix consist of large non-zero values and the remaining elements are zero.

Now, under the assumption that the traits are correlated via the PGS, connected component of the graph $\mathcal{G}$ represents a cluster of related traits in terms of genetic characteristics. Therefore, summing up, a plan for achieving our objective of identifying clusters of traits is to approximate the precision matrix, $\mathbf{P}$, of the PGSs and further sparsity. This way, we are able to identify blocks of zeros in the precision matrix, and, thus, potential connected components, which correspond to clusters of traits sharing some genetic characteristics. Since we aim at estimating a sparse version of the precision matrix of the PGSs, it makes sense to use the *graphical lasso*, proposed by Friedman et al. (2007), which imposes an ℓ_1 penalty for the estimation of the precision matrix, to increase its sparsity.

For computational reasons the PGSs should be centred before the graphical lasso optimisation problem is solved, such that each column has mean zero. That is, $\frac{1}{N}\sum_{i=1}^{N} z_{il} = 0,\ l = 1,\dots,q$. We denote by $\mathbf{\Sigma}$ the $q \times q$ positive-definite covariance matrix of the matrix, $\mathbf{Z}$, of scaled and centred PGSs. From this we estimate a sparse precision matrix using a lasso (ℓ_1) penalty via the graphical lasso. For a precision matrix $\mathbf{P}$ and an empirical correlation matrix $\mathbf{S}$ the graphical lasso maximises the penalised log-likelihood

$$\log\left(\det\left(\mathbf{P}\right)\right) - \operatorname{tr}\left(\mathbf{SP}\right) - \rho \|\mathbf{P}\|_1$$

over non-negative definite matrices $\mathbf{P}$. Here $\det$ denotes the determinant, tr denotes the trace, and $\|\mathbf{P}\|_1 = (\sum_{i=1}^{q} |p_i|)$ is the ℓ_1 norm.

In order to show how to infer connected components, we recall the definition of an adjacency matrix.

Definition 1.5.2 ((Peters et al., 2018), Definition B.3)**.** We can represent a directed graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ over q nodes with a binary $q \times q$ matrix $\mathbf{A}$ (taking values 0 or 1):

$$a_{ij} = 1 \Leftrightarrow (i, j) \in \mathcal{E}.$$

$\mathbf{A}$ is called the *adjacency matrix* of $\mathcal{G}$.

If the graph is undirected, that is, all of its edges are bidirectional, the adjacency matrix is symmetric. In words, the adjacency matrix of an undirected graph is a symmetric square matrix with zeros on the diagonal, and the non-zero elements (usually ones) of the matrix represent pairs of nodes which are adjacent.

We are able to generate an adjacency matrix from the estimated sparse precision matrix, $\hat{\mathbf{P}}_{ij}$ by letting $A_{ij} = 1$ if $\hat{\mathbf{P}}_{ij} \neq 0$ for $i \neq j$ and zero otherwise.

Example 1.5.1. The result of applying the graphical lasso to the genetic data for the Durocs is shown in Figure 1.12 and for the Landraces in Figure 1.13, for four different values of the penalty parameter ρ increasing from top left to bottom right. $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, and $\mathbf{z}_4$ corresponds to PGSs computed for `scanning_weight`, `kg_feed_consumed`, `Feed efficiency`, and `meat%`, respectively. We observe that for the smallest value of ρ the graphs are fully connected, which corresponds to one cluster of the PGSs. When ρ increases, the estimated graphs becomes more sparse. For example, for the Durocs, $\rho = 0.3$ yields one connected component of three PGSs ($\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_4$), corresponding to a cluster of the traits `scanning_weight`, `kg_feed_consumed`, and `meat%` and a singleton containing `Feed efficiency`. For the Landrace, $\rho = 0.3$ yields one connected component of three PGSs ($\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_3$), corresponding to a cluster of the traits `scanning_weight`, `kg_feed_consumed`, and `Feed efficiency` and a singleton containing `meat%`. ◊

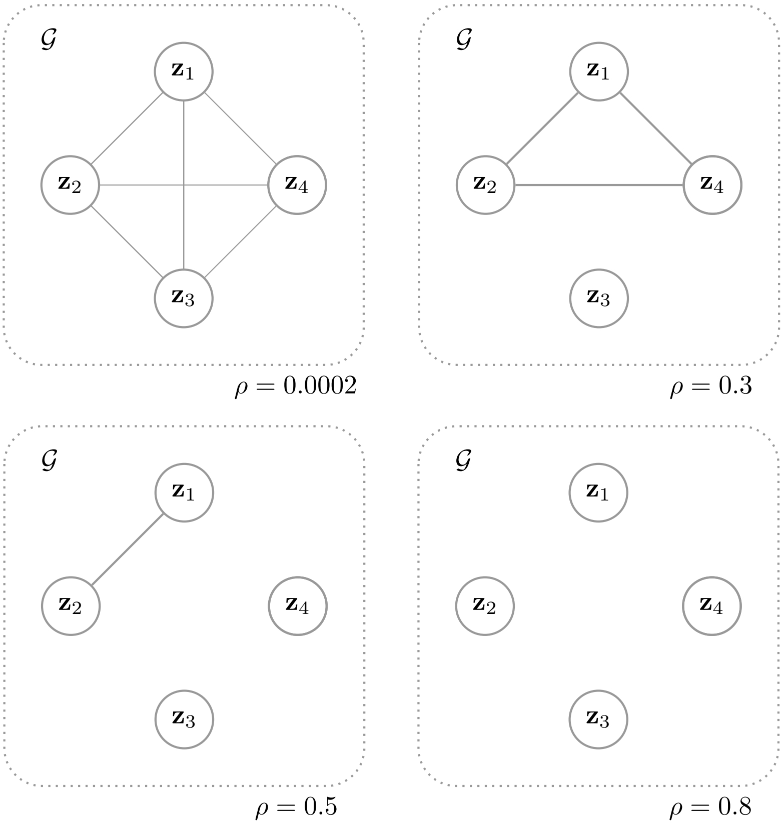


Figure 1.12: *Genetic data for Durocs: Undirected graphs from the graphical lasso with different (increasing from top left to bottom right) values of the penalty parameter ρ.*

We note that a sufficiently small penalisation, $\rho \to 0$, most likely results in a dense version of the precision matrix, corresponding to a fully connected graph with no conditionally inde-

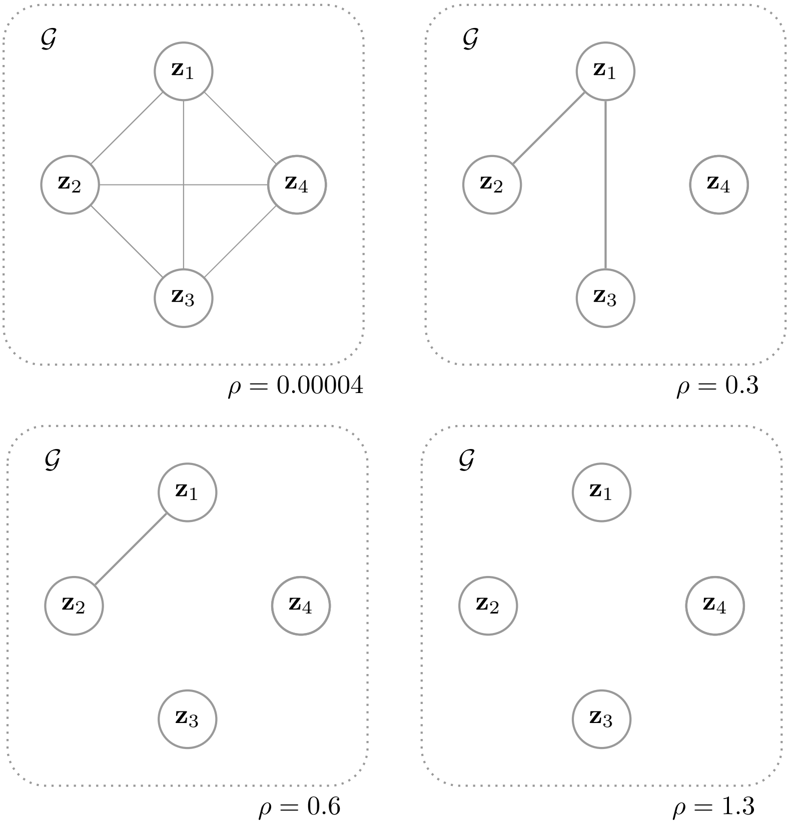


Figure 1.13: *Genetic data for Landraces: Undirected graphs from the graphical lasso with different (increasing from top left to bottom right) values of the penalty parameter ρ.*

pendent random variables. Similarly, a sufficiently large penalisation, $\rho \to \infty$, most likely results in a sparse version of the precision matrix, corresponding to a graph with no edges and only conditionally independent random variables. What is considered a good choice of regularisation parameter, however, depends on whether the goal is prediction accuracy, increasing precision of estimates, gain in power or recovering the right model for interpretation purposes.

In conclusion, by using the graphical lasso we are able to obtain an approximation of the precision matrix, $\hat{\mathbf{P}}_{(\rho)}$, of the PGSs at a grid of values for the regularisation parameter ρ. From these, we construct corresponding adjacency matrices, $\hat{\mathbf{A}}_{(\rho)}$, each of which represent a set of connected components $C_{(\rho)}$.

Let us outline what has been achieved so far. We have defined the notion of PGSs, we have introduced some graphical terminology and defined the notion of conditional independence in this context, and we have introduced the graphical lasso, resulting in sparse precision matrices. The idea behind these concepts is the following. Consider a set of traits and associated PGSs.

We are able to represent the PGSs and their relations by a graph and approximate a sparse precision matrix of the PGSs using the graphical lasso. This suggests that, for a given level of regularisation, the PGSs are represented by connected components. This is what the notion of conditional independence given in Definition 1.5.1 and the property Proposition 1.5.1 allow us to formalise, and Example 1.5.1 shows this in action.

The conclusion is that by approximating the precision matrix of the PGSs and further sparsity, we are able to identify blocks of zeros in the precision matrix, and, thus, potential connected components, which correspond to clusters of traits sharing some genetic characteristics. Next, we outline some fundamental concepts for estimating the parameters in the linear regression model (1.9), and when used in combination, they allow us to exploit the information gained from clustering the traits for simultaneously modelling traits which share some genetic characteristics.

Ordinary least squares A popular method for estimating the parameters in a linear regression model is the *ordinary least squares* (OLS) method, in which we find the parameters $\mathbf{b} \in \mathbb{R}^{2\times 1}$ which minimise the *residual sum of squares* (RSS):

$$\mathrm{RSS}(\mathbf{b}) = (\mathbf{y} - \mathbf{Z}\mathbf{b})^\top (\mathbf{y} - \mathbf{Z}\mathbf{b}) .$$

From here onwards, a constant term is included in the matrix of features $\mathbf{Z}$, such that the first column of $\mathbf{Z}$ is a column of ones allowing estimation of the intercept while the following columns contain the features associated with the corresponding trait values. Since the function $\mathrm{RSS}(\mathbf{b})$ is quadratic in $\mathbf{b}$ with positive-definite Hessian, it possesses a unique global minimum at $\hat{\mathbf{b}}^{\mathrm{O}}$, given by the closed form expression

$$\hat{\mathbf{b}}^{\mathrm{O}} = \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{y}.$$

Later, we exploit the property that the OLS estimator is consistent when the features are exogenous (uncorrelated with the random errors in the model).

Weighted least squares Weighted least squares (WLS) is a generalisation of OLS in which knowledge of the variance of observations is incorporated into the regression. WLS is also a special case of generalised least squares (GLS). WLS is a technique for estimating the unknown parameters in a linear regression model when there is no degree of correlation between the residuals, that is, when all the off-diagonal entries of $\boldsymbol{\Omega}$ (the among-trait and within-individual correlation of the errors in (1.9)) are zero, but the variances of the observations (along the covariance matrix diagonal) are unequal (heteroscedasticity). In this case a *weighted sum of squares* (WSS) is minimised:

$$\mathrm{WSS}(\mathbf{b}) = (\mathbf{y} - \mathbf{Z}\mathbf{b})^\top \mathbf{W} (\mathbf{y} - \mathbf{Z}\mathbf{b}) ,$$

where w_i is the weight of observation $i = 1, \ldots, N$. Under the assumption that observations are uncorrelated, the weights should, ideally, be equal to the reciprocal of the variance of the measurement. The solution is given by the closed form expression

$$\hat{\mathbf{b}}^{\mathrm{W}} = \left(\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}\mathbf{y},$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$ is the diagonal matrix of the weights $w_1, \ldots, w_N$.

Feasible generalised least squares If the among-trait and within-individual correlation, $\boldsymbol{\Omega}$, of the errors in (1.9) is unknown, it is possible to get a consistent (in terms of structure) estimate using the feasible generalised least squares (FGLS) estimator.

The FGLS procedure consists of two steps: First, a linear regression model of the form

$$\mathbf{y} = \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon},$$

is estimated by OLS, and the residuals,

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}^{\mathrm{O}},$$

are used to build an estimator, $\hat{\boldsymbol{\Omega}}^{\mathrm{O}}$ of the covariance matrix, $\boldsymbol{\Omega}$, of the errors:

$$\hat{\boldsymbol{\Omega}}^{\mathrm{O}} = \mathrm{cov}\left(\hat{\mathbf{u}}\right).$$

Second, using the estimator of the covariance matrix of the errors, the unknown regression coefficients are estimated by WLS using $\hat{\boldsymbol{\Omega}}^{\mathrm{O}}$ as the weights:

$$\hat{\mathbf{b}}^{\mathrm{F}(1)} = \left(\mathbf{Z}^{\top}\left(\hat{\boldsymbol{\Omega}}^{\mathrm{O}}\right)^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\left(\hat{\boldsymbol{\Omega}}^{\mathrm{O}}\right)^{-1}\mathbf{y}.$$

The procedure is iterated with the first iteration given by

$$\begin{aligned}
\hat{\mathbf{u}}^{\mathrm{F}(1)} &= \left(\mathbf{y} - \hat{\mathbf{y}}^{\mathrm{F}(1)}\right)\\
\hat{\boldsymbol{\Omega}}^{\mathrm{F}(1)} &= \mathrm{cov}\left(\hat{\mathbf{u}}^{\mathrm{F}(1)}\right)\\
\hat{\mathbf{b}}^{\mathrm{F}(2)} &= \left(\mathbf{Z}^{\top}\left(\hat{\boldsymbol{\Omega}}^{\mathrm{F}(1)}\right)^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\left(\hat{\boldsymbol{\Omega}}^{\mathrm{F}(1)}\right)^{-1}\mathbf{y}
\end{aligned}$$

where $\hat{\mathbf{y}}^{\mathrm{F}(1)} \in \mathbb{R}^{N}$ are the fitted values from the regression:

$$\hat{\mathbf{y}}^{\mathrm{F}(1)} = \mathbf{Z}\hat{\mathbf{b}}^{\mathrm{F}(1)}.$$

This estimation of $\boldsymbol{\Omega}$ is iterated to convergence and we obtain estimates $\hat{\boldsymbol{\Omega}}^{\mathrm{F}}$ and $\hat{\mathbf{b}}^{\mathrm{F}}$. The standard error, SE, of the estimated coefficients are given by

$$\mathrm{SE} = \sqrt{\mathrm{diag}\left(\left(\mathbf{Z}^{\top}\left(\hat{\boldsymbol{\Omega}}^{\mathrm{F}}\right)^{-1}\mathbf{Z}\right)^{-1}\right)}.$$

If the estimator of the covariance matrix of the errors is a consistent estimator, like the OLS, then the FGLS estimator of the unknown regression coefficients is a consistent estimator (Baltagi, 2008).

The geneJAM method We are now ready to outline the geneJAM method, which estimates clusters of correlated traits that share some genetic component and analyse data combined in these clusters. An algorithmic overview, replicated from Manuscript II, of the procedure is shown in Algorithm 1.

Using the graphical lasso we estimate sparse precision matrices, $\hat{\mathbf{P}}_{(\rho)}$, of the PGSs at a grid of values for the regularisation parameter ρ. For each tried value of ρ, we generate from each $\hat{\mathbf{P}}_{(\rho)}$ a corresponding adjacency matrix from which we conclude a clustering represented by a graph $\mathcal{G}_{(\rho)}$, that is, we infer a set of $C_{(\rho)}$ connected components. We assume that the residual covariances among traits within each individual, $\mathbf{R}$, exhibit this clustering as well. Then, we use the information gained from the clusterings to estimate the unknown parameters in (1.9) via FGLS. For the first step in the FGLS estimation of $\mathbf{R}$ we estimate a linear regression model of the form

$$\mathbf{y}_l = \mathbf{Z}_l \mathbf{b}_l + \mathbf{e}_l,$$

for each trait separately, $l = 1, \ldots, q$. Here, $\mathbf{Z}_l \in \mathbb{R}^{N \times 2}$ are design matrices, where a constant term is included, such that the first column of $\mathbf{Z}_l$ is a column of ones allowing estimation of the intercept while the following column contains the feature associated with the corresponding trait value. $\mathbf{e}_l \in \mathbb{R}^N$ is a vector of independent Gaussian random errors. The OLS estimates are

$$\hat{\mathbf{b}}_l^{\mathrm{O}} = \left(\mathbf{z}_l^\top \mathbf{Z}_l\right)^{-1} \mathbf{Z}_l^\top \mathbf{y}_l,$$

for $l = 1, \ldots, q$, and the corresponding estimated residuals are

$$\hat{\mathbf{u}}_l = \left(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}^{\mathrm{O}}\right)_l.$$

As mentioned, we assume that the structure of the residual among-trait within-individual covariance can be approximated by the structure of the covariance of the PGS which is represented by connected components in the graph $\mathcal{G}_{(\rho)}$. Therefore, for each value of the regularisation parameter ρ, we construct an estimate, $\hat{\mathbf{R}}^{\mathrm{O}}_{(\rho)}$, of the covariance of the errors by computing the covariance of the OLS estimated residuals in the same connected component:

$$\hat{\mathbf{R}}^{\mathrm{O}}_{g(\rho)} = \mathrm{cov}\left(\hat{\mathbf{U}}_{g(\rho)}\right),$$

where $\hat{\mathbf{R}}^{\mathrm{O}}_{g(\rho)}$ and $\hat{\mathbf{U}}_{g(\rho)}$ are sub-matrices of $\hat{\mathbf{R}}^{\mathrm{O}}_{(\rho)}$ and $\hat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_q) \in \mathbb{R}^{N \times q}$, respectively, corresponding to the connected components $g(\rho) = 1, \ldots, C_{(\rho)}$ at a given value of ρ. From these we construct an estimate, $\hat{\mathbf{\Omega}}^{\mathrm{O}}_{(\rho)}$, of $\mathbf{\Omega}$ of the form (1.10). In the second step in the FGLS estimation, we build the FGLS estimator, $\hat{\vec{\mathbf{B}}}^{\mathrm{F}}_{(\rho)} \in \mathbb{R}^{2q}$, using WLS:

$$\hat{\vec{\mathbf{B}}}^{\mathrm{F}}_{(\rho)} = \left(\vec{\mathbf{Z}}^{\top}\left(\hat{\mathbf{\Omega}}^{\mathrm{O}}_{(\rho)}\right)^{-1}\vec{\mathbf{Z}}\right)^{-1}\vec{\mathbf{Z}}^{\top}\left(\hat{\mathbf{\Omega}}^{\mathrm{O}}_{(\rho)}\right)^{-1}\vec{\mathbf{Y}}.$$

The procedure is iterated to convergence, see Algorithm 1, and we obtain estimates $\hat{\mathbf{\Omega}}_{(\rho)}$ and $\hat{\vec{\mathbf{B}}}_{(\rho)}$.

Summing up, we have approximated a sparse precision matrix of the PGSs via the graphical lasso. Thus, for a given level of regularisation, we are able to identify blocks of zeros in the precision matrix, and, therefore, potential connected components, which correspond to clusters of traits sharing some genetic characteristics. Using this clustering structure and FGLS we are able to build an estimator of the unknown parameters $\vec{\mathbf{B}}$ and the unknown among-trait and within-individual correlation, $\mathbf{\Omega}$, of the errors in (1.9). That is, we have, through the use of PGSs, achieved our objective of identifying clusters of traits and analysing the data combined in these clusters.

In practice, choosing the value of the regularisation parameter ρ is an important issue as it controls the conditional independence structure of the PGSs, and, thereby, the clustering of the traits. Since we aim at optimising the precision, we use the standard error of the estimated coefficients to tune the geneJAM method. In particular, we are interested in optimising the precision when traits are correlated, and therefore, we use the standard error of traits in connected components of size strictly larger than one. The standard error, $\mathrm{SE}_{(\rho)}$, of the estimated coefficients are given by

$$\mathrm{SE}_{(\rho)} = \sqrt{\mathrm{diag}\left(\left(\vec{\mathbf{Z}}^{\top}\left(\hat{\mathbf{\Omega}}_{(\rho)}\right)^{-1}\vec{\mathbf{Z}}\right)^{-1}\right)}.$$

Thus, for a given value, ρ, of the regularisation parameter, we compute the standard error $\mathrm{SE}_{(\rho)} \in \mathbb{R}^{q}$ and, for all traits which are in clusters of size strictly larger than one, we compute the average standard error (SE),

$$\bar{\mathrm{SE}}_{(\rho)} = \frac{1}{\left|\mathcal{G}_{(\rho)}\right|}\sum_{l\in\mathcal{G}_{(\rho)}} \mathrm{SE}_{(\rho)l},$$

where $\mathcal{G}_{(\rho)}$ is the index set of traits in connected components of size strictly larger than one and $\left|\mathcal{G}_{(\rho)}\right|$ is the cardinality of $\mathcal{G}_{(\rho)}$. We use the smallest value of $\bar{\mathrm{SE}}_{(\rho)}$, over all tried values of ρ, as our selection criterion for the regularisation.

This concludes our exposition of the motivation and framework of the geneJAM method proposed in Manuscript II. We have focused on aspects relating to the primary contribution: utilising polygenic scores to identify clusters of traits, and, thereby, enabling analyses of data combined (in said clusters). Example 1.5.2 and Example 1.5.3 show how the method is used in concrete cases.

Algorithm 1: The geneJAM algorithm

```
for $l = 1, \dots, q$ do
    Compute $\hat{\mathbf{b}}_l^{\mathrm{O}}$ and $\hat{\mathbf{u}}_l$
end
Centre PGSs $\mathbf{Z}$ per column
Compute empirical covariance matrix $\boldsymbol{\Sigma}$ of centred $\mathbf{Z}$
if no sequence of regularisation parameters $\rho_r$, $r = 1, \dots, R$, is provided then
    Specify length $R$ of sequence of regularisation parameter $\rho_r$
    Specify ratio $\delta$ between regularisation parameters $\rho_r$, $r = 1, \dots, R$
    Choose a sequence of regularisation parameters $\rho_r$, $r = 1, \dots, R$, with maximal
    value, $\rho_R$, defined by the maximum column sum of $\boldsymbol{\Sigma}$ and the rest of the sequence
    determined by $R$ and $\delta$.
end
for $r = 1, \dots, R$ do
    Estimate sparse precision matrices $\hat{\mathbf{P}}^{(\rho_r)}$ to obtain $C^{(\rho_r)}$ clusters
    for $g(\rho) = 1, \dots, C^{(\rho_r)}$ do
        Compute $\hat{\Omega}_{g(\rho)}^{\mathrm{O}}(\rho)$
    end
    Construct $\hat{\boldsymbol{\Omega}}^{\mathrm{O}}$
    Compute estimate $\hat{\vec{\mathbf{B}}}^{\mathrm{F}}$
    Define $\hat{\boldsymbol{\Omega}}^{\mathrm{F}(1)} = \hat{\boldsymbol{\Omega}}^{\mathrm{O}}$
    Initialise $t \to 1$
    repeat
        Compute $\hat{\mathbf{u}}^{\mathrm{F}(t)}$
        for $g(\rho) = 1, \dots, C^{(\rho_r)}$ do
            Compute $\hat{\Omega}_{g(\rho)}^{\mathrm{F}(t)}$
        end
        Construct $\hat{\boldsymbol{\Omega}}^{\mathrm{F}(t)}$
        Compute estimate $\hat{\vec{\mathbf{B}}}^{\mathrm{F}(t+1)}$
    until convergence
end
```

Example 1.5.2. We apply the geneJAM method to the SNP data for the races separately. We include only the four traits scanning_weight, kg_feed_consumed, Feed efficiency, and meat%. We omit individuals with missing trait and obtain 62 Durocs and 50 Landraces. For each quantitative trait we generate PGSs as described in Section 1.5.1. The geneJAM method is applied at a suitable sequence of the regularisation parameters ρ.

In Figure 1.14 we show diagnostics plots for the Durocs (a) and Landraces (b). In the left panel we show the average SE plotted against the values of ρ used in the fits. The orange coloured dot indicates the minimum average SE and corresponding regularisation parameter $\hat{\rho}_{\min}$. We observe that the average SE curve attains a minimum at $\hat{\rho}_{\min}$. In the right panel we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$. Grey squares represent estimated edges and white space represent no edges. For both races, we observe that the best precision is obtained when all traits are clustered together. This is indication that the traits share enough genetic characteristics for a joint analysis of all of them to be beneficial in terms of precision and power (Schmitz et al., 1998).

In Figure 1.15 we visualise the covariance structure between traits for the Durocs (a) and Landraces (b). Dark colours indicate large values and light colours indicate small values. In the left panel we show the structure observed in data and in the right panel we show the structure estimated by the geneJAM method. We observe very similar structures all around, that is, a stronger relation between scanning_weight and kg_feed_consumed. This may be indication of a stronger genetic relationship between these two traits. ◊

Now, the application of the geneJAM method goes beyond the PGSs. In fact, any summary statistic sharing the characteristics of PGSs can be used. This includes metabolite scores as we demonstrate in the following example.

Example 1.5.3. We also apply geneJAM to the dataset of 729 metabolites. We include the four traits scanning_weight, kg_feed_consumed, Feed efficiency, and meat% and generate metabolomic scores in the same way PGS are produced, see Section 1.5.1.

Figures B.2–B.5 in Appendix B show visualisations of the estimated adjacency matrices at the tried sequence of values of ρ. In Figure 1.16 we show diagnostics plots for the Durocs from the first (a) and second (b) sample and for the Landraces from the first (c) and second (d) sample. We observe that for the second sample from the Durocs and both samples from the Landraces, the minimum average SE is obtained when all traits are clustered together. This is indication that the traits share enough metabolomic characteristics for a joint analysis to be beneficial in terms of precision and power (Schmitz et al., 1998). For the first sample from the Durocs the minimum average SE is obtained when the first and second trait (scanning_weight and kg_feed_consumed) are clustered and the remaining traits are singleton clusters. This is indication that, in the first sample for the Durocs, the traits scanning_weight and kg _feed_consumed share enough metabolomic characteristics for a joint analysis of these two traits to have an impact in terms of precision and power (Schmitz et al., 1998), but nothing is gained by analysing the other traits combined.

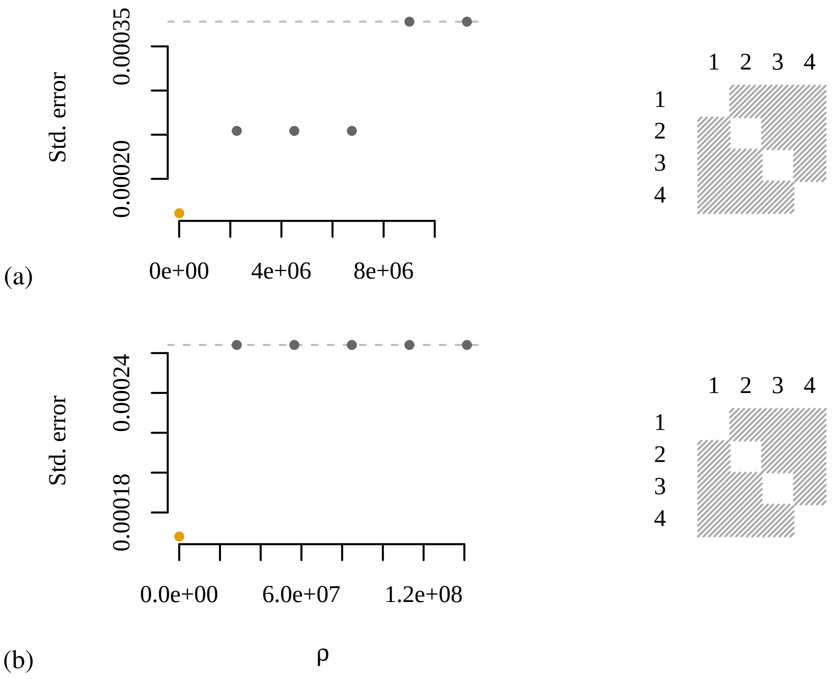


Figure 1.14: *Diagnostics plots of the SNP data from the Durocs (a) and Landraces (b).* Left panel: *average SE curve as a function of ρ.* Right panel: *adjacency matrix corresponding to $\hat{\rho}_{\min}$. Grey squares represent estimated edges and white space represent no edges.*

In Figure 1.17 we visualise the covariance structure between traits for the Durocs (a) and Landraces (b). Dark colours indicate large values and light colours indicate small values. In the left panel we show the structure observed in data and in the centre and right panel we show the structure estimated by the geneJAM method for the first and second sample, respectively. Similarly to Example 1.5.2, we observe very similar structures all around, i.e., a stronger relation between `scanning_weight` and `kg_feed_consumed`. This may be indication of a stronger metabolomic relationship between these two traits. ◊

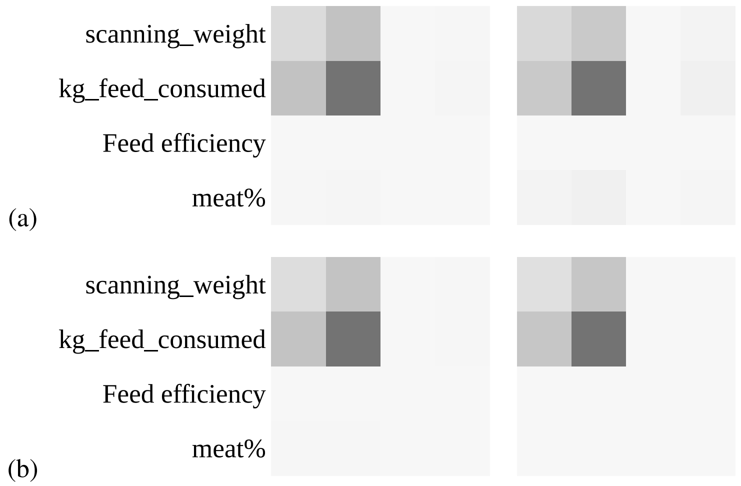


Figure 1.15: *Visualisation of the covariance structure for the SNP data from the Durocs (a) and Landraces (b). Dark colours indicate large values and light colours indicate small values.* Left: *Structure observed in the data.* Right: *Structure estimated by the geneJAM method.*

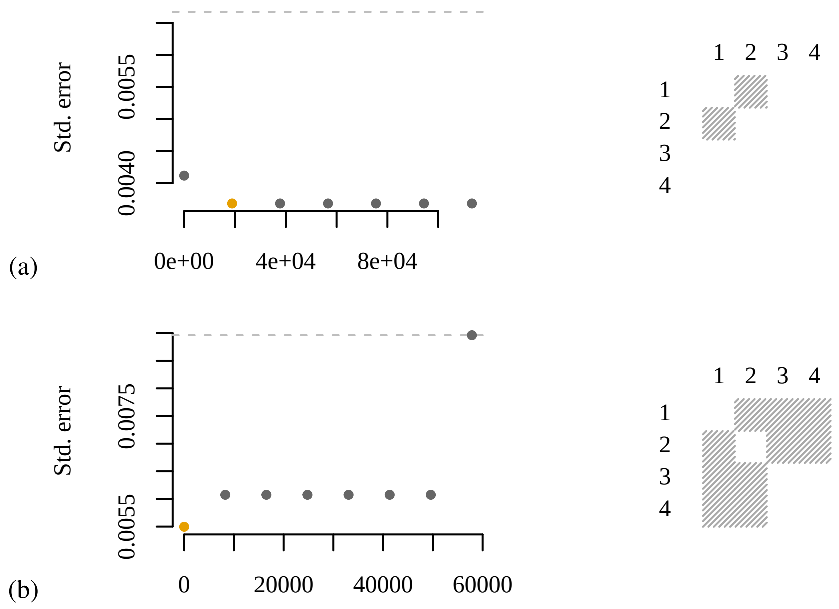


Figure 1.16: *(Continued on the next page.)*

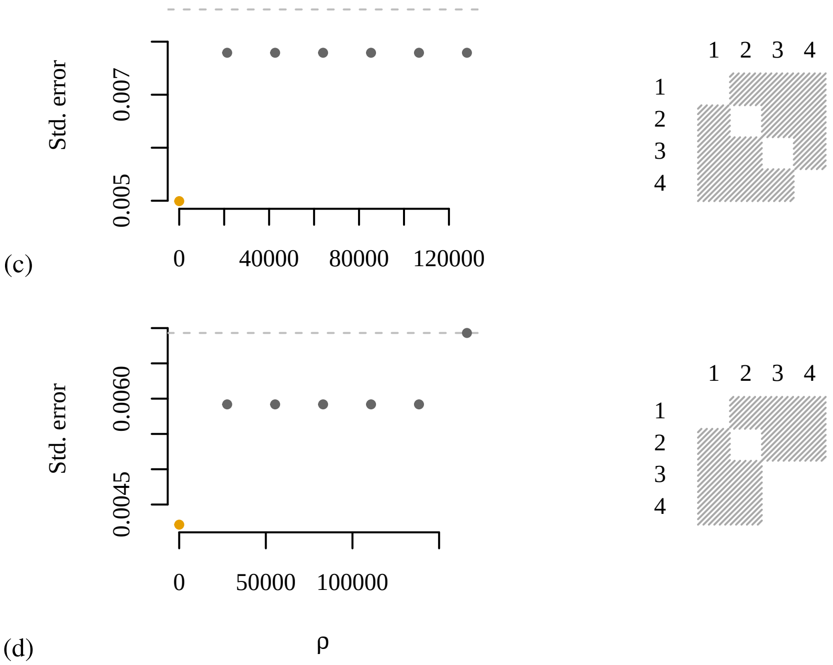


Figure 1.16: *(Continued from the previous page.) Diagnostics plots of the metabolite data for the Durocs from the first (a) and the second (b) sample and for the Landraces from the first (c) and second (d) sample.* Left panel: *average SE curve as a function of ρ.* Right panel: *adjacency matrix corresponding to $\hat{\rho}_{\min}$. Grey squares represent estimated edges and white space represent no edges.*

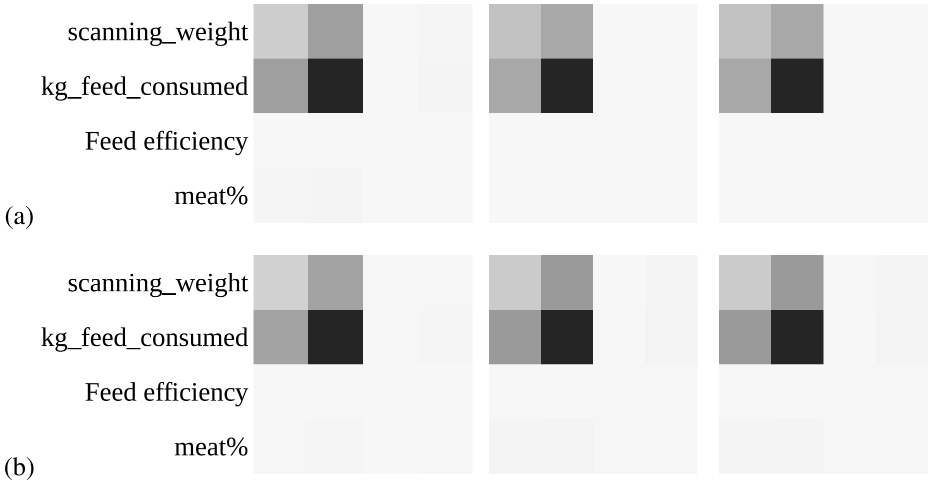


Figure 1.17: *Visualisation of the covariance structure for the metabolite data for the Durocs (a) and the the Landraces (b). Dark colours indicate large values and light colours indicate small values.* Left panel: *Structure observed in the data.* Centre panel: *Structure estimated by the geneJAM method for the first sample.* Right panel: *Structure estimated by the geneJAM method for the second sample.*

1.5.3 Discussion of the geneJAM method

Multivariate GWASs have gained attention in genetic studies as they offer several advantages over analysing each trait in a separate GWAS (Galesloot et al., 2014).

How to simultaneously analyse a large number of simultaneously measured and potentially correlated traits in a multivariate GWAS in a computationally efficient way is not clear. Identifying clusters of correlated traits that share some genetic component will facilitate analysis of the data combined in clusters and increase precision and power and provide computational advantages.

We have proposed a versatile method for finding clusters of correlated outcomes that share some genetic component by means of PGSs: by estimating the precision matrix of the PGSs, we are able to approximate clusters of traits that share some genetic component. From this we induce the structure of the error covariance matrix of a regression model and use the FGLS estimator for estimating the unknown parameters in a linear regression model when there is a certain unknown degree of correlation between the residuals. Existing algorithms for covariance estimation include the traditional restricted maximum likelihood (REML) method and the recent method of moments (MoM). Compared to REML, MoM approaches are computationally efficient and require only GWAS summary statistics. However, MoM approaches can be statistically inefficient, often yielding inaccurate covariance estimates (Zhou, 2017). In Manuscript II we used the REML method (specifically, the `lmer` function in the `lme4` package (Bates et al., 2015)) to estimate the parameters in a linear mixed-effects model (LMM) with an unconstrained covariance structure for the random effects as reference when assessing the precision and computational performance of our method. Compared to both the simple linear regression and the LMM the geneJAM method was superior in terms precision of the estimates of traits on which there are genetic effects. Furthermore, the computing time of the geneJAM method was highly satisfactory even for large data (`lme4` cannot handle a number of traits $q > 2^6$). To improve the computation time of the LMM we could constrain the covariance for the random effects to reduce the number of parameters but this would require a lot of (data assisted) assumptions.

Key advantages of our framework are that there are no prior assumptions on the structure nor the sizes of the clusters of traits, information on PGS level is sufficient when data on SNP level is not available, and, following the clustering and modelling, both clusters and model fit are readily available. Moreover, the computing time is highly satisfactory even for large data, and compared to both the simple linear regression and the multilevel models the geneJAM method is superior in terms precision of the estimates of traits on which there are genetic effects. Finally, if (independent) environmental factors are available, they are easily adjusted for in the regression models.

In practice, a requirement of a more biological nature is that a certain amount of the observed variability in the traits needs to be due to heredity for the clustering of the PGSs to be a reasonable approximation of the clustering of the traits.

1.5.4 Perspectives: related individuals

Underlying the framework outlined above is the assumption that individuals are independent. However, many genetic association studies of complex traits sample related individuals since samples with relatedness can offer advantages in terms of power over samples which include only unrelated individuals (Teng and Risch, 1999). Furthermore, including relatives in sequencing studies increases power for identifying rare variants with extremely low frequencies (Kazma and Bailey, 2011). To ensure validity of the association results, familial correlations must be taken into account when related individuals are included in a genetic association study. Therefore, it would be of interest to consider extensions of the geneJAM method to the case of related individuals.

In the following, we give a rough idea of how the matrix normal distribution, introduced by de Waal and Luttrell (1985), could be used extend the geneJAM method to allow for among-individual and within-trait correlation.

The matrix normal distribution is a generalisation of the multivariate normal distribution to matrix-valued random variables. If we define a matrix from the N observations of the q outcome components, $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q) \in \mathbb{R}^{N\times q}$, the outcome follows a matrix normal ($\mathcal{MN}$) distribution

$$\mathbf{Y} \sim \mathcal{MN}_{N\times q}\left(\mathbf{M}, \mathbf{U}, \mathbf{V}\right), \tag{1.11}$$

with general location matrix $\mathbf{M} = \mathbf{1}_N \boldsymbol{\mu}^\top \in \mathbb{R}^{N\times q}$. The scale matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$ specifies the row (among-individual and within-trait) covariances for each trait, and, the assumption of the geneJAM method that observations within traits are uncorrelated, that is, rows are independent, is equivalent to the assumption that $\mathbf{U} = \mathbf{I}_{N\times N}$. The scale matrix $\mathbf{V} \in \mathbb{R}^{q\times q}$ models the (among-trait and within-individual) systematic and residual covariances.

Under the assumption that $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_q) \in \mathbb{R}^{N\times q}$ is a matrix of Nq independent samples of the features (PGSs) from the standard normal distribution, so that

$$\mathbf{Z} \sim \mathcal{MN}_{N\times q}\left(\mathbf{0}, \mathbf{I}_{N\times N}, \mathbf{I}_{q\times q}\right),$$

and letting $\mathbf{Y} = \mathbf{M} + \mathbf{AZB}$,

$$\mathbf{Y} \sim \mathcal{MN}_{N\times q}\left(\mathbf{M}, \mathbf{AA}^\top, \mathbf{B}^\top\mathbf{B}\right),$$

where $\mathbf{A}$ and $\mathbf{B}$ can be chosen by Cholesky decomposition or a similar matrix square root operation.

Standard multiple linear regression models assume that the underlying observations are independent. Under this assumption, $\mathbf{A}$ is the $N \times N$ identity matrix with $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_{N\times N}$ and the *multivariate linear regression* model takes the form

$$\mathbf{Y} = \mathbf{M} + \mathbf{Z}\mathbf{B} + \mathbf{E}, \tag{1.12}$$

where $\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_q) \in \mathbb{R}^{q\times q}$ is a matrix of unknown regression coefficients such that $\mathbf{V} = \mathbf{B}\mathbf{B}^\top$, and $\mathbf{E} \in \mathbb{R}^{N\times q}$ is a matrix of Gaussian random errors. The assumption that observations are independent is reflected in the assumption that errors of the outcome are independent, that is,

$$\mathbf{E} \sim \mathcal{MN}_{N\times q}\left(\mathbf{0}_{N\times q}, \sigma_r^2\mathbf{I}_{N\times N}, \sigma_c^2\mathbf{I}_{q\times q}\right),$$

where $\mathbf{0}_{N\times q}$ is an $N \times q$ matrix of zeros, $\sigma_r^2 > 0$ is the among-individual and within-trait variance, and $\sigma_c^2 > 0$ is the among-traits and within-individual residual variance.

This is essentially a matrix normal representation of the statistical assumptions of the proposed geneJAM method. To allow for among-trait correlation, the geneJAM assumes that $\mathbf{E} \sim \mathcal{MN}_{N\times q}(\mathbf{0}_{N\times q}, \sigma_r^2\mathbf{I}_{N\times N}, \mathbf{\Omega})$, where $\mathbf{\Omega} \in \mathbb{R}^{q\times q}$ is a general covariance matrix. Possible structures for $\mathbf{\Omega}$ may be a multiple of the identity, a diagonal, or, more importantly, a general positive definite matrix.

As mentioned earlier, imposing some structure on the covariance matrix $\mathbf{\Omega}$ is crucial to the consistency of the FGLS estimator. We approached this by approximating a clustering of the PGSs and assuming this to be indicative of the structure of the covariance of the traits and, therefore, of the errors.

It would be of interest to also allow for among-individual and within-trait correlation by assuming a general $\mathbf{A} \neq \mathbf{I}$ and that $\mathbf{E} \sim \mathcal{MN}_{N\times q}(\mathbf{0}_{N\times q}, \mathbf{\Psi}, \mathbf{\Omega})$ for some general covariance matrix $\mathbf{\Psi} \in \mathbb{R}^{q\times q}$.

1.6 Multivariate modelling and interactions

In Manuscript III we extend the method, developed in Manuscript I, for identifying hierarchical interactions using regularised regression to a multivariate setup. In Section 1.6.1 and Section 1.6.2 we give a more detailed account of existing methods which form the basis of Manuscript III, in Section 1.6.3 we outline the results of Manuscript III, and in Section 1.6.4 we discuss our results and review perspectives for future research based on the material covered.

Our work with interactions and hierarchical assumptions concerns the development of these notions for multivariate regularised regression models. Our goal is to estimate the outcome components separately but to use, via regularisation, information across components to infer subsets of relevant features and interactions. In order to properly understand the context of our work, we begin by giving a presentation of multivariate hierarchical interaction models in a $p < N$ context and of multivariate regularised regression methods when interactions are not a concern. After doing so, we outline our work on multivariate regularised regression with hierarchical interactions.

1.6.1 Multivariate hierarchical interaction models

The linear regression model for a single outcome is generalised by the multivariate regression model for multiple outcome components simply by assuming a linear regression model for each. Thus, we assume that we have N observations of the multivariate outcome $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_q) \in \mathbb{R}^{N \times q}$ and p associated features $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ with pairwise interactions. As for the univariate setup, the pairwise interaction term is formed simply by multiplying the two corresponding features whether quantitative or categorical (represented by dummy variables). We also assume that each outcome component is centred before the optimisation problem is solved, and the intercept term is omitted in the following.

Thus, for each observation $i = 1, \dots, N$ and each outcome $l = 1, \dots, q$ the *multivariate pure main effect model* takes the form

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + e_{il}, \tag{1.13}$$

and the *multivariate pairwise interaction model* takes the form

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + \sum_{k=1}^{p} \sum_{j<k} x_{ij} x_{ik} \Theta_{jkl} + e_{il} \tag{1.14}$$

where $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p)^\top \in \mathbb{R}^{p \times q}$ with $\mathbf{b}_j = (b_{1j}, \dots, b_{qj})$, $j = 1, \dots, p$, are unknown regression parameters for the main effects, $\boldsymbol{\Theta} \in \mathbb{R}^{p \times p \times q}$ are unknown parameters for the pairwise interactions, and $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_q)^\top \in \mathbb{R}^{N \times q}$ with $\mathbf{e}_l = (e_{1l}, \dots, e_{Nl})^\top$, $l = 1, \dots, q$, are Gaussian errors. For each $l = 1, \dots, q$ we let $\boldsymbol{\Theta}_l \in \mathbb{R}^{p \times p}$ represent a symmetric matrix, i.e. $\boldsymbol{\Theta}_l = \boldsymbol{\Theta}_l^\top$, such that the strict inequality in the interaction summation precludes overparametrisation arising from the inclusion of the same effect twice, e.g., including both $x_{ij} x_{ik}$ and $x_{ik} x_{ij}$.

As mentioned, when aiming to interpret pairwise interactions in a (multivariate) linear regression model it is necessary to consider the concept of hierarchy. In the following, we present model restrictions in a form which makes it possible to specify a regularised regression procedure, which produces multivariate sparse interaction models that honour these restrictions. The hierarchical constraint of the pairwise interactions is defined for all outcome components simultaneously, that is, exactly one of the following hierarchical, anti-hierarchical, and non-hierarchical restrictions, replicated from Manuscript III, is assumed for all $l = 1, \dots, q$:

Strong hierarchy There are interactions only among pairs of non-zero main effects,

$$\mathrm{H_S}: \quad \Theta_{jkl} \neq 0 \quad \Rightarrow \quad b_{jl} \neq 0 \text{ and } b_{kl} \neq 0.$$

Weak hierarchy Each interaction has at least one of its main effects present,

$$\mathrm{H_W}: \quad \Theta_{jkl} \neq 0 \quad \Rightarrow \quad b_{jl} \neq 0 \text{ or } b_{kl} \neq 0.$$

Anti-hierarchy Interactions are only among pairs of main effects that are not present,

$$\mathrm{H_A}: \quad \Theta_{jkl} \neq 0 \quad \Rightarrow \quad b_{jl} = 0 \text{ and } b_{kl} = 0.$$

Pure interactions There are no main effects present, only interactions,

$$\mathrm{H_I}: \quad b_{jl} = 0 \quad \forall j = 1, \dots, p.$$

Pure main effects There are no interactions present, only main effects,

$$\mathrm{H_M}: \quad \Theta_{jkl} = 0 \quad \forall j, k = 1, \dots, p.$$

No hierarchy There are no restrictions to the presence of main effects and interactions, $\mathrm{H_N}$.

In general, we may denote by $\mathcal{I}_\mathrm{H}$ the index set of pairwise interactions (pairs of features) to be included in the model subject to one of the hierarchical restrictions. Then, the restrictions above result in the following *hierarchical pairwise interaction model*:

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + \sum_{k=1}^{p} \sum_{j<k} \mathbb{1}_{\{(j,k) \in \mathcal{I}_\mathrm{H}\}} x_{ij} x_{ik} \Theta_{jkl} + e_{il}. \tag{1.15}$$

1.6.2 Multivariate regularised regression

We are interested in identifying a subset of relevant features by fitting a regularised regression model for a multivariate setup.

As suggested by Hastie et al. (2015), we approach ℓ_1 regularised regression in the multivariate outcome setting by considering a special case of the univariate *group lasso* (Yuan and Lin,

2006) with p groups of equal size q. To better understand the multivariate outcome setting we begin by introducing the group lasso in the univariate outcome setting: For N samples of a univariate outcome, $\boldsymbol{y} \in \mathbb{R}^N$, we assume that we have $J \leq p$ groups of features, such that, for $j = 1, \ldots, J$, the matrix $\mathbf{Z}_j \in \mathbb{R}^{N \times p_j}$ represents the p_j features in group j. We recommend scaling each outcome component if they are not measured in the same units and each feature if these are not measured in the same units. Otherwise the lasso solution depends on the scale. We denote by $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_J) \in \mathbb{R}^{N \times p}$ the collection of features, that is, the features are grouped by columns, and we require that $\sum_{j=1}^{J} p_j = p$, with p the total number of features, such that there are no overlapping groups. We wish to model the outcome from the collection of grouped features, and we assume a linear model,

$$y_i = \sum_{j=1}^{J} z_{ij} \mathbf{b}_j + \epsilon_i,$$

where the vector $\mathbf{b}_j \in \mathbb{R}^{p_j}$ represents a collection of p_j unknown parameters corresponding to the group of features $\mathbf{Z}_j$, $j = 1, \ldots, J$. Please note, that in order to avoid confusion, we use $\mathbf{Z}_j$ and $\mathbf{b}_j$ to represent groups of features and their coefficients, rather than the $\mathbf{x}_j$ and $\boldsymbol{\beta}_j$, which we use for single features.

The univariate group lasso solves, for all $j = 1, \ldots, J$, the convex optimisation problem

$$\begin{aligned} \underset{\mathbf{b}_j \in \mathbb{R}^{p_j}}{\text{minimise}} \quad & \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{J} z_{ij} \mathbf{b}_j \right)^2 \\ \text{subject to} \quad & \sum_{j=1}^{J} \|\mathbf{b}_j\|_2 \leq t, \end{aligned} \tag{1.16}$$

where $\|\mathbf{b}_j\|_2$ is the Euclidean norm of the vector $\mathbf{b}_j$ and $t \geq 0$ is a pre-specified tuning parameter which controls the amount of shrinkage applied to the estimates. We can write the univariate group lasso problem in the Lagrangian form

$$\underset{\mathbf{b}_j \in \mathbb{R}^{p_j}}{\text{minimise}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{J} z_{ij} \mathbf{b}_j \right)^2 + \lambda \sum_{j=1}^{J} \|\mathbf{b}_j\|_2 \right\}, \tag{1.17}$$

for some penalty parameter $\lambda \geq 0$, and for every value of λ there is a t (and for every value of t there is a λ) such that solving (1.16) and (1.17) results in the same estimates. The penalty parameter λ is determined by a model validation technique such as cross-validation, and selecting the value of λ translates to selecting a proper amount of regularisation and is, as such, a trade-off between data fitting and sparsity. In the following we exclude the trivial case that $\lambda = 0$.

We observe that the group optimisation problem reduces to the ordinary univariate lasso, when $p_j = 1$ such that all groups are singletons and $\|\mathbf{b}_j\|_2 = |\mathbf{b}_j|$. We also note that, depending on

$\lambda > 0$, all elements of the vector $\hat{\boldsymbol{b}}_j$ will be zero *or* non-zero simultaneously – corresponding to selecting all or none of the features in group j. Hence, the general univariate group lasso allows for sparsity with respect to selection of groups but not with respect to the selection of features within groups. While the univariate sparse group lasso (Simon et al., 2013) allows for both sparsity of groups and within each group, it is not relevant here. In Figure 1.18 we display the constraint region for the univariate group lasso (left), the univariate sparse group lasso (centre), and the univariate lasso (right) regressions in the case of three features divided into two groups yielding two group coefficients $\mathbf{B}_1 = (B_1, B_2) \in \mathbb{R}^2$ and $\mathbf{B}_2 = B_3 \in \mathbb{R}$.

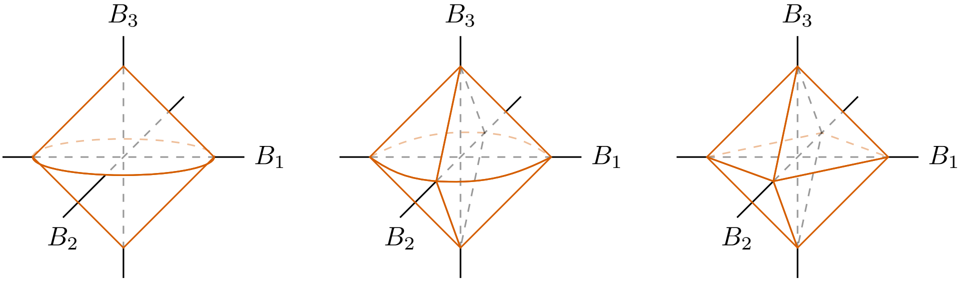


Figure 1.18: *The group lasso ball (left) in* $\mathbb{R}^3$ *compared to the sparse group lasso ball (centre) and the* ℓ_1 *ball (right) in the case of two groups with coefficients* $\mathbf{B}_1 = (B_1, B_2) \in \mathbb{R}^2$ *and* $\mathbf{B}_2 = B_3 \in \mathbb{R}$.

We are now ready to approach the multivariate outcome setting, which we motivate first by the two examples Example 1.6.1 and later by Example 1.6.3. In Example 1.6.1 we illustrate the assumption that there is an unknown subset of the features which are relevant and this same subset is *preserved* across all q components of the outcome. In Example 1.6.3 we illustrate the more general assumption that there is an unknown subset of the features which are relevant but this subset is *not preserved* across all components of the outcome.

Example 1.6.1. As mentioned before, the overall objective of the FEEDOMICS project is to improve the FE in the Danish pig production and in this context we are also interested in the traits weight and meat percentage. Furthermore, we are interested in the effect of race. That is, we consider an associated binary feature representing the race (Duroc or Landrace). The assumption, that the relevance of features is preserved across outcome components means that if, e.g., race is relevant, it affects both meat percentage and feed consumption but not necessarily with the same effect size. We illustrate this in the two scatterplots in Figure 1.19. The plots suggest that meat percentage and feed consumption are strongly associated as there seems to be a positive linear correlation between the two variables. In the left plot, data points have been coloured according to race (orange for Duroc, blue for Landrace). The plot indicates that race is not relevant for either of the traits; there is no obvious association between the race of the pigs and the levels of weight and feed consumption. In the right plot, we have, for illustrative purposes, coloured the data points artificially. That is, for weights less than 95 kg and feed consumption less than 130 kg the data points are coloured orange, otherwise they are coloured blue. Had this been the actual representation of the race (e.g., orange for Duroc, blue for Landrace), we would have concluded that Landraces were associated with both higher scanning weight and larger feed consumption compared to Durocs. That is, race would seem relevant, and would appear relevant across both components of the outcome. ◊

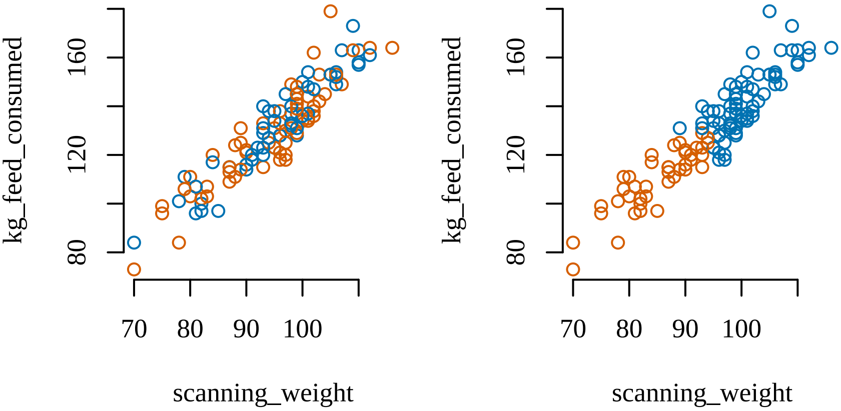


Figure 1.19: *Scatterplots suggesting that the two traits are strongly associated as there seem to be a positive linear correlation between them.* Left*: The data points are coloured according to the race (orange for Duroc, blue for Landrace) and suggest that race is not relevant for either of the traits.* Right*: The data points are artificially coloured, but if we pretend that they represent the race (again, orange for Duroc, blue for Landrace), they would suggest that race is relevant for both weight and feed consumption.*

The pure main effect model (1.13) can be seen as a coupled collection of q standard regression problems in $\mathbb{R}^p$, each sharing the same features, in which column l, $\mathbf{b}_l \in \mathbb{R}^q$, of $\mathbf{B}$ is the coefficient vector for the lth problem. If we define the groups by the rows $\mathbf{b}_j \in \mathbb{R}^q$, $j = 1, \ldots, p$, from the full matrix of parameters $\mathbf{B} \in \mathbb{R}^{p \times q}$, the problem is a special case of the general group lasso (1.17), in which we have $J = p$ groups of equal size $p_j = q$, $j = 1, \ldots, p$. Under the assumption that the same set of features are relevant for the modelling across all q components of the outcome variable, that is, the collection of q outcome components are regressed on the same $N \times p$ matrix of features, we approach the multivariateness by solving q separate lasso problems, one for each column of the $p \times q$ regression matrix $\mathbf{B}$. Thus, for the multivariate pure main effect model (1.13), the objective of the multivariate lasso is to solve the regularised least-squares problem

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \sum_{i=1}^{N} \sum_{l=1}^{q} \left(y_{il} - \sum_{j=1}^{p} x_{ij} b_{jl} \right)^2 + \lambda \sum_{j=1}^{p} \|\mathbf{b}_j\|_2 \right\}, \tag{1.18}$$

for some penalty parameter $\lambda > 0$. This way, the collection of q outcome components are regressed on the same $N \times p$ matrix of features and the q linear regression problems are coupled together via the regularisation constraint which ensures that all features are equally penalised across outcome components. It should be noted that, in general, the formulation of (1.17) implies that all groups are equally penalised, which makes larger groups more likely to be selected. The reason is, that if any of the features in a group is selected, then the whole group is selected. Rescaling the parameters is one way to ensure that small groups are not overwhelmed by large groups in the selection (Liu et al., 2013). It is, however, not a problem when an analogue to (1.17) is used to approach the multivariateness as in (1.18), where the p "groups" are constructed to be of equal size q.

Example 1.6.2. The `R` package `glmnet` provides implementations of multivariate Gaussian models to be fit under the assumption that the same subset is selected for all outcome components, using a group lasso penalty on the coefficients for each variable as in (1.18), see Friedman et al. (2010). As an example, we look at the pig data and consider four outcome components: `scanning_weight`, `kg_feed_consumed`, `Feed efficiency`, and `meat%`. In Figure 1.20 we display the cross-validation curve (dots), and upper and lower standard deviation curves (grey bars) along the λ sequence for the Durocs (top) and Landraces (bottom). The value of λ that gives minimum mean cross-validated error, λ_{min}, and λ_{1se}, which gives the most regularised model such that the error is within one standard error of the minimum are indicated by vertical dotted lines and orange dots. We observe that for the Durocs no SNPs are selected at λ_{min}. For the Landraces 38 SNPs are selected at λ_{min}. A comprehensive list of selected SNPs is given in Table C.1 in Appendix C. We note that there is an overlap of SNPs selected by the multivariate lasso and the lasso method used in Example 1.4.1. In the Manhattan plots in Figure B.1 in Appendix B we indicate by × SNPs selected by the multivariate lasso. ♢

In general, the group lasso is a grouping of parameters in the penalty and the multivariate lasso (1.18) is a grouping of parameters across outcome components. The effect of such a grouping is to couple the regularisation of the parameters across the outcome components, which may be a good idea, if all the outcome components can be predicted from roughly the same set of features. We are, however, interested in extending the framework to the more general assumption that there is an unknown subset of the features which are relevant but this subset is *not preserved* across all q components of the outcome. We motivate this assumption by the following example.

Example 1.6.3. We continue Example 1.6.1. The assumption, that the relevance of features is not preserved across outcome components means that, e.g., race affects meat percentage but not feed consumption. We illustrate this in the two scatterplots in Figure 1.21. In the left plot, data points have been coloured according to race (orange for Duroc, blue for Landrace). The plot indicates that race is not relevant for either of the traits. In the right plot, we have, for illustrative purposes, coloured the data points artificially. That is, for weights less than 95 kg the data points are coloured orange, otherwise they are coloured blue – regardless of the quantity of feed consumed. Had this been the actual representation of the race (e.g., orange for Duroc, blue for Landrace), we would have concluded that Landraces were associated with

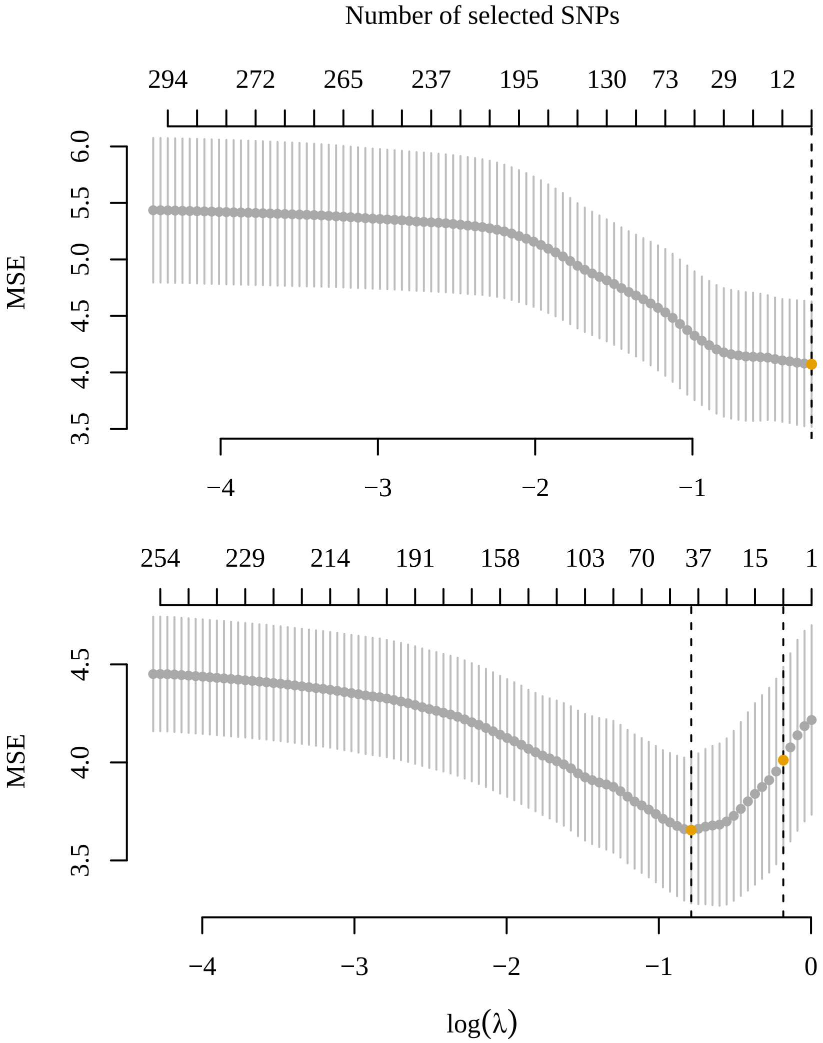


Figure 1.20: *Cross-validation curve (dots) for the multivariate lasso regression, and upper and lower standard deviation curves (grey bars) along the λ sequence (bottom axis) and corresponding sequence of number of selected SNPs (upper axis) for the Durocs (top) and Landraces (bottom). λ_{min} and λ_{1se} are indicated by the orange dots and vertical dotted lines.*

higher scanning weight but not with larger feed consumption compared to Durocs. That is, race would seem relevant, but the relevance would not be preserved across the components of the outcome since it would be relevant for weight but not for feed consumption. ◊

As mentioned, the nature of the multivariate lasso implies that when a feature is selected by the procedure, the coefficients for that feature will be non-zero for the entire collection of

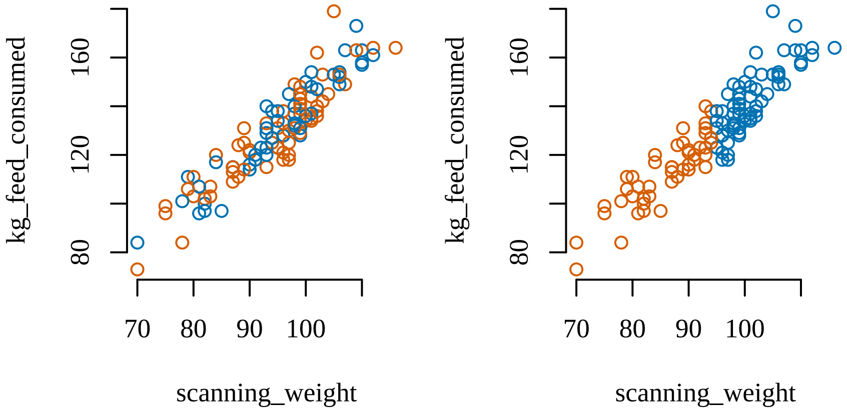


Figure 1.21: *Scatterplots suggesting that the two traits are strongly associated as there seem to be a positive linear correlation between them.* Left*: The data points are coloured according to the race (orange for Duroc, blue for Landrace) and suggest that race is not relevant for either of the traits.* Right*: The data points are artificially coloured, but if we pretend that they represent the race (again, orange for Duroc, blue for Landrace), they suggest that race is relevant for weight but not for feed consumption.*

outcome components. It may, however, be more appropriate to have sparsity with respect to the parameters for the entire collection of outcome components as well as for components within the collection of grouped components, since all features are not necessarily relevant for all outcome components. That is, we would like sparsity both with respect to selection of features for the entire outcome and with respect to selection of features for each outcome component. The additional "within-group" sparsity is achieved by setting up an analogue to the multivariate sparse group lasso (MSGLasso) proposed by Li et al. (2015). The objective of the MSGLasso is to solve the penalised optimisation problem,

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \sum_{j=1}^{p}\sum_{l=1}^{q} \lambda_{jl}|b_{jl}| + \sum_{j=1}^{p} \lambda_j \left\|\mathbf{b}_j\right\|_2 \right\}, \tag{1.19}$$

where the ℓ_2 penalty term aims to shrink irrelevant groups to zero and the ℓ_1 penalty term aims to shrink irrelevant entries within a relevant group to zero. Thus, the multivariate sparse group lasso makes a compromise between the group lasso ($\lambda_{jl} = 0$) and the ordinary lasso ($\lambda_j = 0$). In Figure 1.22 we illustrate difference between the lasso (top), the group lasso (middle) and the sparse group lasso (bottom) in the case of three groups of unequal size.

In (1.19), the collection of q outcome components are regressed on the same $N \times p$ matrix of features. When $\lambda_j > 0$ and $\lambda_{jl} = 0$ for all $j = 1, \ldots, p$ and $l = 1, \ldots, q$, the problem reduces to the group lasso. This corresponds to the assumption that the same set of features are relevant for the modelling across all components of the outcome. When $\lambda_{jl} > 0$, the achieved within-

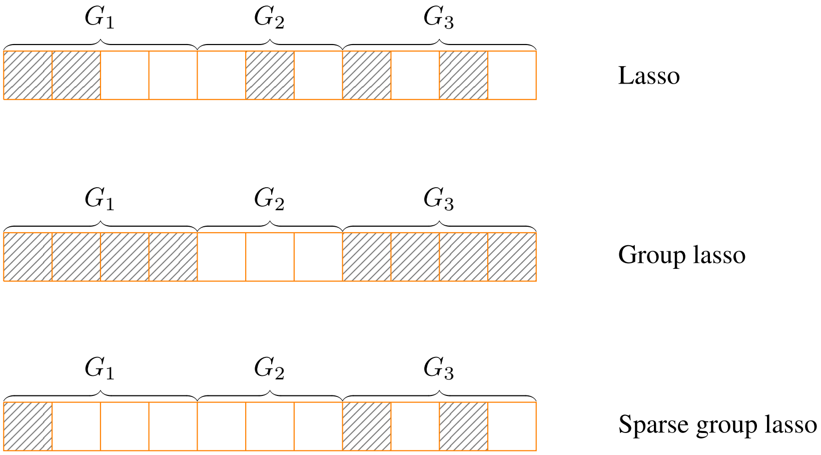


Figure 1.22: *Sparsity patterns enforced by the lasso (top), group lasso (middle), and sparse group lasso (bottom) assuming that data can be divided into three groups. The set of selected (groups of) features are shaded grey. Lasso selects features regardless of the group structure; group lasso selects features on a group basis; sparse group lasso selects features on a group basis while also selecting individual features within each selected group.*

group sparsity allows for the features to be differently penalised across outcome components, corresponding to the assumption that there is an unknown subset of the features which are relevant, but this subset is *not preserved* across all components of the outcome.

1.6.3 A framework for multivariate hierarchical regularised regression

We are now ready to describe our efforts in developing a method for identifying hierarchical interactions using multivariate regularised regression. Our work is detailed in Manuscript III.

We are interested in identifying a subset of relevant features, such that imposing hierarchical restrictions in a stepwise manner, sufficiently reduce the number of interactions to be included in a multivariate regularised regression model.

The first step in this direction is to set up an analogue to the MSGLasso optimisation problem (1.19) to the pairwise interaction model (1.14), such that a penalty is imposed on the interaction coefficients, thereby shrinking them towards (or estimated to be exactly) zero.

We define collections of interactions by the rows $\mathbf{\Theta}_{jk} \in \mathbb{R}^q$, $j, k = 1, \ldots, p$ of the full stack of matrices of parameters $\mathbf{\Theta} \in \mathbb{R}^{p \times p \times q}$. The optimisation problem takes the form

$$\underset{\mathbf{B} \in \mathbb{R}^{p\times q}, \mathbf{\Theta} \in \mathbb{R}^{p\times p\times q}}{\text{minimise}} \left\{ L(\mathbf{B}, \mathbf{\Theta}) + \sum_{j=1}^{p}\sum_{l=1}^{q} \lambda_{jl}|b_{jl}| + \sum_{j=1}^{p} \lambda_j \left\|\mathbf{b}_j\right\|_2 \right.$$
$$\left. + \sum_{k=1}^{p}\sum_{j=1}^{k-1}\sum_{l=1}^{q} \tau_{jkl}|\Theta_{jkl}| + \sum_{j=1}^{p} \tau_j \left\|\mathbf{\Theta}_j\right\|_2 \right\},$$

where $L(\mathbf{B}, \mathbf{\Theta})$ denotes the loss function

$$L(\mathbf{B}, \mathbf{\Theta}) = \frac{1}{2Nq} \sum_{i=1}^{N}\sum_{l=1}^{q} \left(y_{il} - \sum_{j=1}^{p} x_{ij} b_{jl} - \sum_{k=1}^{p}\sum_{j=1}^{k-1} x_{ij}x_{ik}\Theta_{jkl} \right)^2 .$$

Now, defining the penalty parameters is key for achieving our objective of constructing a procedure which ensures interpretation in terms of hierarchy. To obtain hierarchical interactions, we make the following observations. When $\lambda_j = 0$, feature j is not penalised for any component of the outcome and, therefore, always included in the model for the entire collection of outcome components; when $\lambda_j = \infty$ feature j is always excluded for all outcome components; and for all $j = 1, 2, \ldots, p$ for which λ_j are equal to the same constant value, the corresponding feature is equally penalised for the entire collection of outcome components. Similarly, when $\lambda_{jl} = 0$, feature j is not penalised for the lth outcome component but always included in the model; when $\lambda_{jl} = \infty$ feature j is always excluded for the lth outcome component; and for all $j = 1, 2, \ldots, p$ and $l = 1, \ldots, q$ for which λ_{jl} are equal to the same constant value, the corresponding features are equally penalised for the corresponding outcome components. The interpretations of the penalties τ_j and τ_{jkl} on the interactions are analogous to those of λ_j and λ_{jl}, respectively. By these observations, we are able to define

$$\lambda_{jl} = \lambda_2\alpha_2\mathbb{1}_{\{(j,l)\notin\mathcal{M}_e\}}, \quad \lambda_j = \lambda_2(1-\alpha_2)\mathbb{1}_{\{j\notin\mathcal{M}_g\}},$$
$$\tau_{jkl} = \lambda_2\alpha_3\mathbb{1}_{\{(j,k,l)\in\mathcal{I}_e\}}, \quad \tau_j = \lambda_2(1-\alpha_3)\mathbb{1}_{\{j\in\mathcal{I}_g\}},$$

where $\mathcal{M}_e$ and $\mathcal{I}_e$ define the index sets of main effects and interactions, respectively, to be included in the model subject to one of the hierarchical restrictions for each outcome component. Similarly, $\mathcal{M}_g$ and $\mathcal{I}_g$ define the index sets of the main effects and interactions, respectively, to be included in the model for the entire collection of outcome components subject to one of the hierarchical restrictions.

Solving the strong and weak multivariate hierarchical lasso outlined above is not computationally tractable even for a moderate number, p, of features and even less so when p is large and $p \gg N$. Therefore, we propose in Manuscript III a two-step procedure. As the first step in the procedure we include only the main effects and apply the "usual" MSGLasso for variable selection. As the second step in the procedure we include main effects and interactions in accordance with step one and subject to one of the restrictions, and we apply a variation of the MSGLasso with the penalties determined by the given hierarchical restriction.

The following, replicated in a shortened form from Manuscript III, is our proposed method.

Step 1 We assume the pure main effect model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

for which the MSGLasso estimates are determined by solving the optimisation problem

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \lambda_1 \sum_{j=1}^{p} \left(\sum_{l=1}^{q} \alpha_1 |b_{jl}| + (1-\alpha_1) \|\mathbf{b}_j\|_2 \right) \right\},$$

where we have defined $\lambda_{jl} = \lambda_1\alpha_1$ and $\lambda_j = \lambda_1(1-\alpha_1)$ with $\alpha_1 \in [0,1]$. We obtain the estimates $\hat{\mathcal{M}}_e$, $\hat{\mathcal{I}}_e$, $\hat{\mathcal{M}}_g$, and $\hat{\mathcal{I}}_g$.

Step 2 We assume the pairwise interaction model, for all $i = 1, \ldots, N$ and $l = 1, \ldots, q$,

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + \sum_{k=1}^{p} \sum_{j<k} \mathbb{1}_{\{(j,k,l)\in\hat{\mathcal{I}}_e\}} x_{ij} x_{ik} \Theta_{jkl} + e_{il},$$

for which the MSGLasso estimates are determined by solving the optimisation problem

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times q}, \boldsymbol{\Theta}\in\mathbb{R}^{p\times p\times q}}{\text{minimise}} \left\{ L(\mathbf{B}, \boldsymbol{\Theta}) + \lambda_2 \sum_{j=1}^{p} \left(\sum_{l=1}^{q} \alpha_2 \mathbb{1}_{\{(j,l)\notin\hat{\mathcal{M}}_e\}} |b_{jl}| + (1-\alpha_2)\mathbb{1}_{\{j\notin\hat{\mathcal{M}}_g\}} \|\mathbf{b}_j\|_2 \right.\right.$$
$$\left.\left. + \sum_{k=1}^{j-1} \sum_{l=1}^{q} \alpha_3 \mathbb{1}_{\{(j,k,l)\in\hat{\mathcal{I}}_e\}} |\Theta_{jkl}| + (1-\alpha_3)\mathbb{1}_{\{j\in\hat{\mathcal{I}}_g\}} \|\boldsymbol{\Theta}_j\|_2 \right) \right\},$$

where $L(\mathbf{B}, \boldsymbol{\Theta})$ denotes the loss function

$$L(\mathbf{B}, \boldsymbol{\Theta}) = \frac{1}{2Nq} \sum_{i=1}^{N} \sum_{l=1}^{q} \left(y_{il} - \sum_{j=1}^{p} x_{ij} b_{jl} - \sum_{k=1}^{p} \sum_{j<k} \mathbb{1}_{\{(j,k,l)\in\hat{\mathcal{I}}_e\}} x_{ij} x_{ik} \Theta_{jkl} \right)^2,$$

and $\lambda_{jl} = \lambda_2\alpha_2\mathbb{1}_{\{(j,l)\notin\hat{\mathcal{M}}_e\}}$, $\lambda_j = \lambda_2(1-\alpha_2)\mathbb{1}_{\{j\notin\hat{\mathcal{M}}_g\}}$, $\tau_{jkl} = \lambda_2\alpha_3\mathbb{1}_{\{(j,k,l)\in\hat{\mathcal{I}}_e\}}$, and $\tau_j = \lambda_2(1-\alpha_3)\mathbb{1}_{\{j\in\hat{\mathcal{I}}_g\}}$.

This concludes our exposition of the background and framework of Manuscript III. We have focused on aspects relating to the primary contribution: the proposal of a method for identifying pairwise interactions in a multivariate regularised regression model under different assumptions of hierarchy using the penalty parameters of the MSGLasso.

1.6.4 Discussion of the multivariate two-step procedure

The majority of GWASs used to focus on the effect of individual SNPs. However, the growing sample size of GWASs has facilitated the discovery of gene-environment and gene-gene interactions, and empirical evidence shows that such interactions may be an important genetic component underlying complex traits and diseases. Moreover, multivariate GWASs have gained attention in genetic studies as they offer several advantages over analysing each trait in a separate GWAS (Galesloot et al., 2014).

How to systematically include interacting features in multivariate regularised regression models is not clear. Including interactions in multivariate regularised regression models under hierarchical restrictions will facilitate simultaneous feature selection, parameter estimation, and interpretation of the results.

We have proposed a two-step multivariate regularised regression procedure for identification of pairwise interactions, when strong or weak hierarchy is assumed. The procedure uses a variation of the MSGLasso to screen for interactions and fitting multivariate hierarchical pairwise interaction models. The main contribution of our procedure is that interactions (and features) are simultaneously estimated and selected while hierarchical restrictions are taken into account. This ensures interpretability of the model (in terms of interactions) and should improve computing time. Importantly, information across all outcome components is used, via regularisation, to infer subsets of relevant features and interactions.

In addition the framework allows for direct specification of whether the handling of features (and interactions) should be equal *or* different across outcome components.

We are not aware of any alternative methods that simultaneously estimate and select interactions under hierarchical restrictions for a multivariate outcome.

Future work will consider the issue of correlation among features as it is somewhat unrealistic to totally neglect this. For example, in highly dense genetic maps, the correlations between SNPs positioned close to each other become higher as the genetic maps become denser. A starting point may be to use the functionality of the MSGLasso allowing for (arbitrary) group structures as intended.

In Manuscript III, we have presented the theoretical framework and a proof-of-concept simulation study. The obvious next step is more detailed simulation studies as well as application on real data. For example, to understand the consequence of a wrong hierarchical assumption it would be interesting to simulate datasets with different (non-)hierarchical structures and apply the multivariate two-step procedure under different hierarchical assumptions. Furthermore, to understand the consequence of not requiring hierarchy it would be interesting to compare our method on simulated datasets to the usual (multivariate) lasso for the pairwise interaction model under the assumption of no hierarchy. In addition, it would be of interest to assess the selection abilities of the method in real data.

We conclude with some remarks on the regularisation parameters. In Manuscript III, we choose to treat main effects and interactions equally and impose the same penalty on main effects and interactions. This way, we keep down the number of hyperparameters, but it might result in main effects being overwhelmed by interactions in the selection. Therefore, it could be argued that main effects should be less penalised than interactions. Moreover, we treat the regularisation parameters primarily as a means to an end. In general, choosing the "right" penalty parameters is difficult. Both cross-validation and stability selection are possible strategies.

1.7 Directions for future research

In this section, we outline some possible topics for further research. Some of the main challenges that remain are concerned with correlated features (Manuscript I and Manuscript III) and correlated individuals (Manuscript II). In the following we give a summary of perspectives for future research based on the results obtained in the manuscripts and the material covered in the previous sections. More details are found in Sections 1.4.4, 1.5.3, 1.5.4, and 1.6.4 as well as in the corresponding manuscripts.

Correlated features. The two-step procedure proposed in Manuscript I is not suited for data with highly correlated features, since the lasso performs erratically when features are correlated. It would be of interest to expand the procedure to the situations where features are correlated. In particular, it would be of interest to adjust the two-step procedure with the adaptive elastic net.

Oracle property of two-step procedure. As mentioned in Manuscript I, it is likely that the oracle property of the adaptive lasso no longer holds, when using cross-validation on the same data twice. As mentioned, we suspect that this reuse may be necessary in many practical cases due to lack of power. In order to retain the attractive oracle property, it would be of interest to understand when leaving some data out of the selection in the first round to be used for cross-validation in the second round is reasonable (in terms of power).

Tuning of the two-step procedure. In Manuscript I the process of choosing the regularisation parameter is seen, primarily, as a means to an end. Choosing the right penalty parameter is, however, difficult and cross-validated choices often include too many features. It would be of interest to apply other tuning methods, such as stability selection (Meinshausen and Bühlmann, 2010), for choosing the right penalty parameters in the two-step procedure.

Related individuals. The geneJAM method proposed in Manuscript II does not allow for correlated measurements on each individual, and as a consequence it is not appropriate to investigate the behaviour of related individuals. Nonetheless, our efforts have aided in setting up a framework which may be helpful for future research efforts. In particular, our current work may be a stepping stone towards application in the case of related individuals by means of the matrix normal distribution (Section 1.5.4).

Non-Gaussian error distributions. In Manuscript II, we consider the situation where the outcome is assumed to be quantitative ($\mathbf{Y} \in \mathbb{R}^{N \times q}$) and the error distribution is assumed to be multivariate Gaussian. A useful property of the multivariate Gaussian distribution is that the ijth component of the corresponding precision matrix is zero if and only if the variables i and j are conditionally independent, given the others. The geneJAM method exploit this property, since, by approximating the precision matrix of the PGSs and further sparsity, we are able to identify blocks of zeros in the precision matrix, and, thus, potential connected components, which correspond to clusters of traits sharing some genetic characteristics. It would be of interest to extend the method to non-Gaussian error distributions for example in the case of a binary outcome representing, e.g., the risk of a complex diseases. In this case a polygenic (risk) score can be used to reflect an individual's estimated genetic predisposition.

Correlated features and multivariate outcome. It is not clear how suited the two-step procedure for multivariate outcome proposed in Manuscript III is for data with highly correlated features. It would be of interest to to understand under which conditions highly correlated features can be accurately selected. In particular, it would be of interest to exploit the functionality of the MSGLasso allowing for arbitrary group structures.

Oracle property of the multivariate two-step procedure. As for the two-step procedure proposed in Manuscript I, it is likely that the oracle property of the MSGLasso no longer holds when using cross-validation on the same data twice. It would be of interest to improve the procedure to retain the oracle property.

Simulation and application of the multivariate two-step procedure. In Manuscript III, we have presented a theoretical framework and a proof-of-concept simulation study. It would be of general interest to conduct more numerical experiments to assess the method. In particular, it would be of interest to apply the method on more complex simulated data and on real data to evaluate the selection abilities and modelling performance (in terms of computing time).

Bibliography

Aiken, L. S. and West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications, Inc., 1 edition.

Baltagi, B. H. (2008). *Econometrics*. Berlin: Springer, 4 edition.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2010). Hierarchical selection of variables in sparse high-dimensional regression. *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D Brown*, 6:56–69.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bien, J. and Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.6.

Biology (2021). Biology — Wikipedia, the free encyclopedia. [Online; accessed September 8, 2021].

Buchardt, A.-S. (2022a). geneJAM: Joint regression analysis of multiple traits based on genetic relationships. `https://github.com/abuchardt/geneJAM`.

Buchardt, A.-S. (2022b). iLasso: Identifying interactions via hierarchical lasso regularisation. `https://github.com/abuchardt/ilasso`.

Buchardt, A.-S. and Ekstrøm, C. T. (2021). Identifying interactions via hierarchical lasso regularisation.

Carmelo, V. A. O. and Kadarmideen, H. N. (2020). Genome regulation and gene interaction networks inferred from muscle transcriptome underlying feed efficiency in pigs. *Frontiers in Genetics*, 11:650.

Cox, D. R. (1984). Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1–24.

de Waal, F. B. M. and Luttrell, L. M. (1985). The formal hierarchy of rhesus macaques: An investigation of the bared-teeth display. *American Journal of Primatology*, 9(2):73–85.

Dong, S.-S., Hu, W.-X., Yang, T.-L., Chen, X.-F., Yan, H., Chen, X.-D., Tan, L.-J., Tian, Q., Deng, H.-W., and Guo, Y. (2017). Snp-snp interactions between wnt4 and wnt5a were associated with obesity related traits in han chinese population. *Scientific Reports*, 7.

Ekstrøm, C. T. and Sørensen, H. (2014). *Introduction to Statistical Data Analysis for the Life Sciences*. CRC Press, 2 edition.

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2014). PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9):1466–1468.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLOS ONE*, 9(4):1–8.

Genetics (2021). Genetics — Wikipedia, the free encyclopedia. [Online; accessed September 8, 2021].

Hartl, D. L. and Clark, A. G. (1982). *Principles of population genetics*. Sinauer Associates, Inc. Publishers, 4 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag New York, 2 edition.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. Chapman and Hall/CRC Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Kazma, R. and Bailey, J. N. (2011). Population-based and family-based designs to analyze rare variants in complex diseases. *Genetic Epidemiology*, (3):41–47.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series.

Lee, K.-Y., Leung, K.-S., Tang, N. L. S., and Wong, M.-H. (2018). Discovering genetic factors for psoriasis through exhaustively searching for significant second order snp-snp interactions. *Scientific Reports*, 8.

Li, Y., Nan, B., and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–63.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Liu, J., Huang, J., Ma, S., and Wang, K. (2013). Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics*, 14(2):205–219.

Meier, L., Geer, S. V. D., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50.

Peters, J., Janzing, D., and Scholkopf, B. (2018). *Elements of Causal Inference*. The MIT Press.

Schmitz, S., Cherny, S. S., and Fulker, D. W. (1998). Increase in power through multivariate analyses. *Behavior Genetics*, 28(5):357–363.

Schrode, N., Ho, S.-M., Yamamuro, K., Dobbyn, A., Huckins, L., Matos, M. R., Cheng, E., Deans, P. J. M., Flaherty, E., Barretto, N., Topol, A., Alganem, K., Abadali, S., Gregory, J., Hoelzli, E., Phatnani, H., Singh, V., Girish, D., Aronow, B., Mccullumsmith, R., Hoffman, G. E., Stahl, E. A., Morishita, H., Sklar, P., and Brennand, K. J. (2018). Synergistic effects of common schizophrenia risk variants. *Nature genetics*, 51(10):1475–1485.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22.

Teng, J. and Risch, N. (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. ii. individual genotyping. *Genome research*, 9(3):234–241.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58(1):267–288.

Valdar, W., Solberg, L., Gauguier, D., Heyes, S., Klenerman, P., Cookson, W., Taylor, M., Rawlins, J., Mott, R., and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38:879–87.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., and Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10):1068–1087.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*, 34(3):275–285.

Ye, R.-S., Xi, Q.-Y., Qi, Q., Cheng, X., Chen, T., Li, H., Kallon, S., Shu, G., Wang, S.-B., Jiang, Q.-Y., and Zhang, Y.-L. (2013). Differentially expressed mirnas after gnrh treatment and their potential roles in fsh regulation in porcine anterior pituitary cell. *PLOS ONE*, 8(2):1–11.

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

Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The Annals of Applied Statistics*, 11(4):2027–2051.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751.

A Summary of manuscripts

In this section we summarise the manuscripts of the thesis.

Manuscript I

"Identifying Interactions via Hierarchical Lasso Regularisation"

In Manuscript I we propose a novel method for identifying hierarchical interactions using regularised regression. It is typical for genomic or multi-omics data that the number, p, of features measured is larger than the number, N, of observations, and when analysing such data it is of interest to not only identify relevant features but also more complex relationships such as gene-gene interactions. However, the introduction of interactions into a model rapidly increases the complexity of said model, since the total number of possible pairwise interactions is $\binom{p}{2} = \frac{1}{2}p(p-1)$. For example, when using 500K single nucleotide polymorphism microarrays there are approximately 125 billion pairwise interactions. Fitting a regression model to such data is computationally challenging, potentially extremely time consuming, and mathematically distorts the relation between the number of features and the number of observations. The goal is to obtain both computationally tractable and interpretable models in terms of effect size as well as feature and interaction selection. We develop an intuitive 2-step procedure for identifying interactions when $p \gg N$ by imposing hierarchical restrictions. In the first step we use lasso (Tibshirani, 1996) and include all features to select the most promising main effects. In the second step, we use adaptive lasso (Zou, 2006) and include main effects and interactions according to hierarchical restrictions to select the most promising terms. The method has potential applications to cross-omics studies and genome-wide studies of gene-gene interactions (epistatis) and, as a special case, to studies of gene-environment interactions. The utility of the method is illustrated in the manuscript on both simulated and real data. Data from the Wellcome Trust Centre for Human Genetics (Valdar et al., 2006) were analysed and several SNP-SNP interactions were identified. Simulations were used to assess the false discovery rate of the interactions under different hierarchical assumptions on the data generating process and the method, and we found that under the right hierarchical restrictions, the cross-validated choices of the penalty parameter succeed in detecting the true interaction. The consequence of no hierarchical assumption is an immense amount of false discoveries. To assess the modelling performance in terms of computing time the method was compared to the `HierNet` function (Bien and Tibshirani, 2014) implemented in R, proposed by Bien et al. (2013), and we demonstrated that the computing time is considerably reduced by our method and is highly satisfactory even for fairly large data.

Manuscript II

"Joint Regression Analysis of Multiple Traits Based on Genetic Relationships"

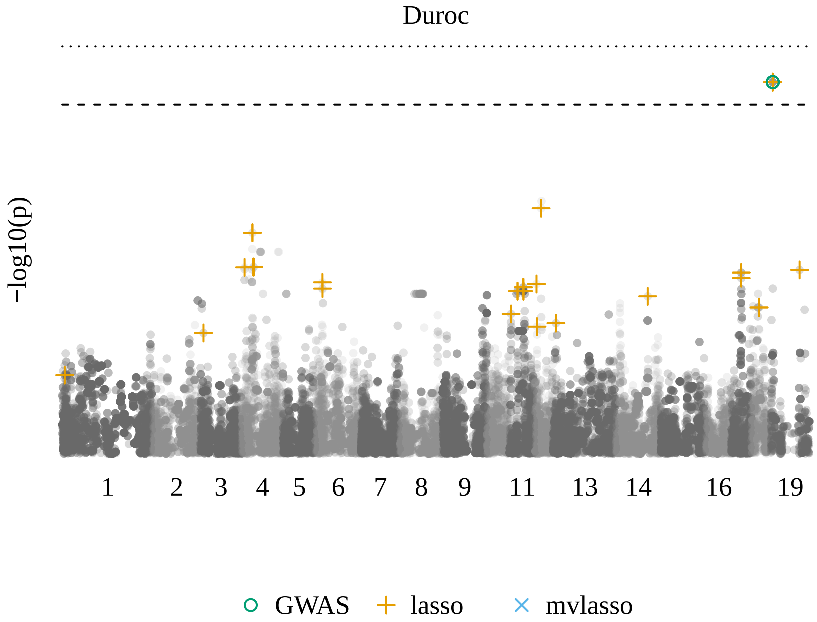
In Manuscript II we propose a novel clustering and estimation method using polygenic scores (PGSs) for inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits in a multivariate GWAS. PGSs are widely available and employed in fields such as animal breeding, plant breeding, and human genetics for predicting and understanding genetic architectures. Based on the presumption that much of the variability observed in a trait is attributable to genetic differences, i.e., heredity, we develop the geneJAM method for estimating clusters of correlated traits that share some genetic component and jointly analyse data in these clusters. By approximating the precision matrix of the PGSs and further sparsity via graphical lasso (Friedman et al., 2007), we identify blocks of zeros in the precision matrix, and, thus, potential connected components, which correspond to clusters of traits sharing some genetic characteristics. Using this clustering structure together with feasible generalised least squares (FGLS) we are able to build estimators of the unknown parameters in a linear regression model and the unknown among-trait and within-individual correlation of the errors. The method has potential applications to many biology studies with traits embedded in clusters sharing genetic characteristics. Data from the Wellcome Trust Centre for Human Genetics (Valdar et al., 2006) were analysed and clusters of traits were identified. The method was compared with the fully parametric linear regressions and multilevel models on simulated data, and we demonstrate that computing time is highly satisfactory even for large data, and compared to both the simple linear regression and the multilevel models the geneJAM method is superior in terms of precision of the estimates of traits on which there are genetic effects.

Manuscript III

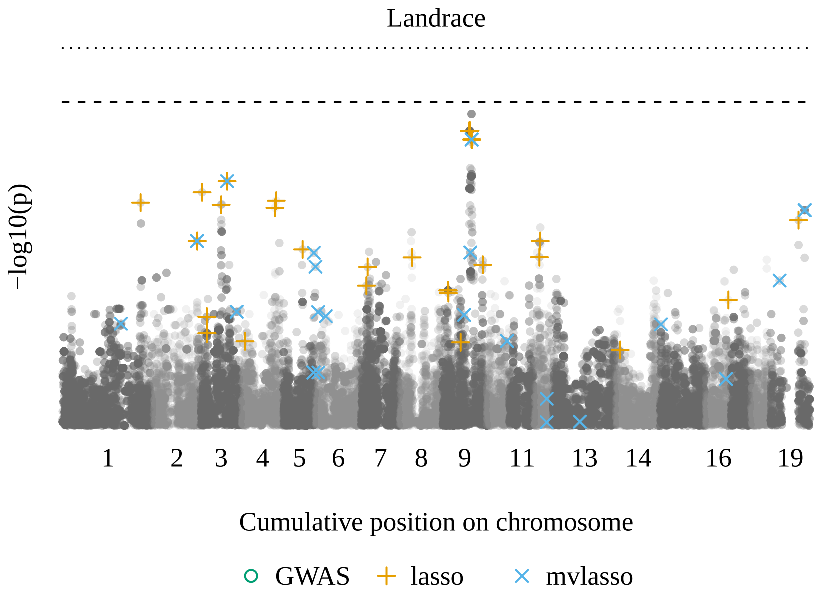
"Identifying Hierarchical Interactions via Multivariate Sparse Group Lasso Regularisation"

In Manuscript III we propose a novel method for identifying hierarchical interactions using multivariate sparse group lasso (MSGLasso) proposed by Li et al. (2015). The method is analogous to the two-step procedure proposed in Manuscript I and extends the framework to the even more complex setup of a multivariate outcome, where a joint analysis of all traits may increase power. We develop a 2-step procedure for identifying interactions by imposing hierarchical restrictions, and we consider the scenario where an unknown subset of the features and interactions are relevant, but this subset is not preserved across the components of the outcome. As a special case, we consider the scenario where an unknown subset of the features and interactions are relevant, and this same subset is preserved across all components of the outcome. In the first step of the procedure we use MSGLasso and include all features to select the most promising main effects for each outcome component. In the second step, we include main effects and interactions and use a variation of MSGLasso with carefully chosen regularisation parameters to impose hierarchical restrictions. This way, the most promising (for each outcome component) features and interactions are selected. The method has potential applications to genomic or multi-omics data where the number of features measured is likely to be much larger than the number of observation, samples often contain multiple, potentially correlated, and simultaneously measured traits, and it is of interest to identify interactions, e.g., gene-gene or gene-environment interactions. The utility of the method is illustrated in the manuscript on simulated data. Simulations were used to assess the false discovery rate of the interactions under the strong hierarchical assumption on the data generating process and the method, and we found that the cross-validated choices of the penalty parameter succeed in detecting the true interactions with a very low false discovery rate.

B Figures

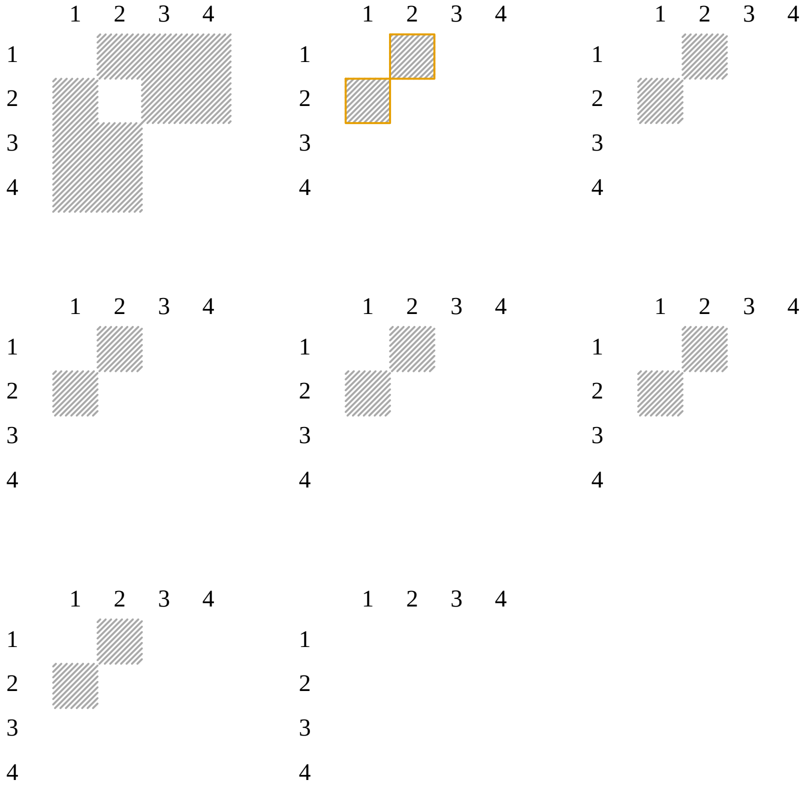


(a) *Manhattan plots for the Durocs.*



(b) *Manhattan plots for the Landraces.*

Figure B.1: *The dashed horizontal line represents the threshold for genome-wide significance from the Bonferroni correction method. The dotted horizontal line represents the conventional threshold* 5×10^{-8}. ○ *indicates SNPs selected by the GWAS, see Example 1.3.1.* + *indicates SNPs selected by the usual lasso, see Example 1.4.1 and the first step of the two-step procedure, see Example 1.4.2.* × *indicates SNPs selected by the multivariate lasso.*



	1	2	3	4
1		▨	▨	▨
2	▨		▨	▨
3	▨	▨		
4	▨	▨		

	1	2	3	4
1		▨		
2	▨			
3				
4				

	1	2	3	4
1		▨		
2	▨			
3				
4				

	1	2	3	4
1		▨		
2	▨			
3				
4				

	1	2	3	4
1		▨		
2	▨			
3				
4				

	1	2	3	4
1		▨		
2	▨			
3				
4				

	1	2	3	4
1		▨		
2	▨			
3				
4				

	1	2	3	4
1				
2				
3				
4				

Figure B.2: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for the first metabolomic sample of the Durocs. Grey squares represent estimated edges and white space represent no edges. Orange borders represent adjacency matrix corresponding to* $\hat{\rho}_{\min}$.

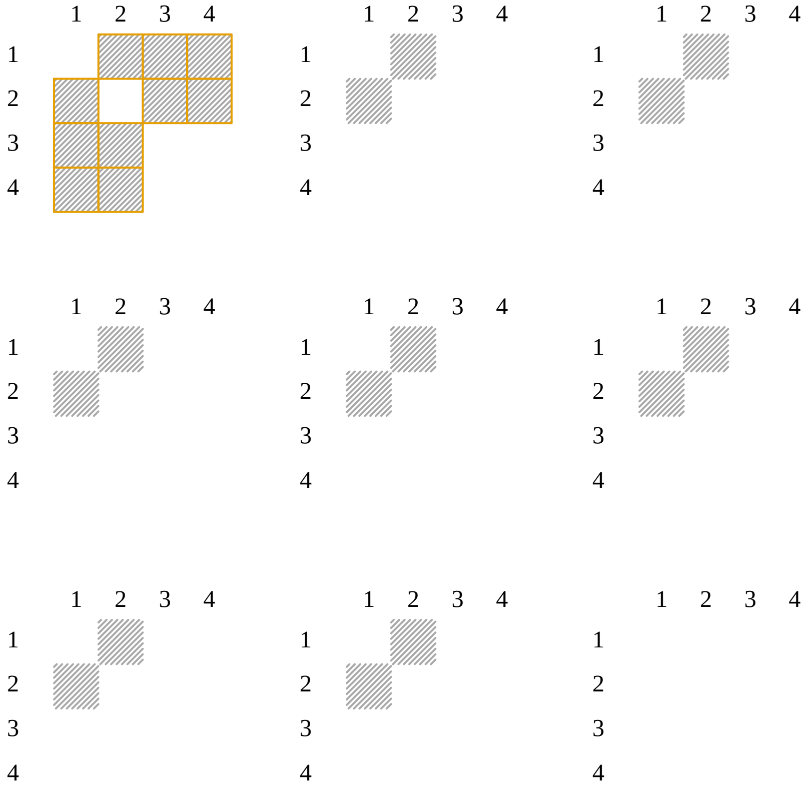


Figure B.3: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for the second metabolomic sample of the Durocs. Grey squares represent estimated edges and white space represent no edges. Orange borders represent adjacency matrix corresponding to* $\hat{\rho}_{\mathrm{min}}$.

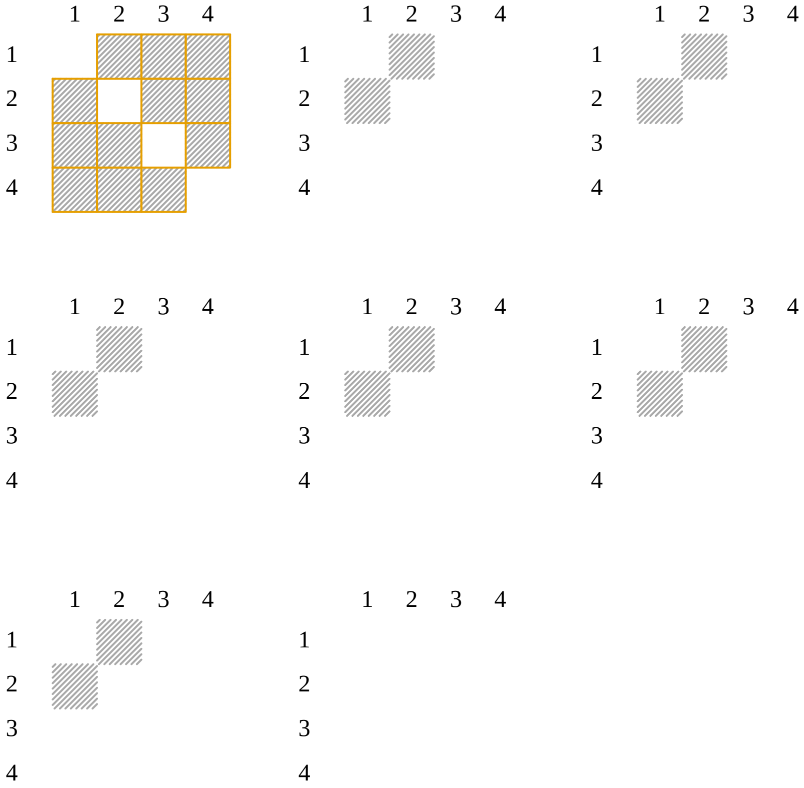


Figure B.4: *Visualisation of estimated adjacency matrices at the tried sequence of values of* ρ *for the first metabolomic sample of the Landraces. Grey squares represent estimated edges and white space represent no edges. Orange borders represent adjacency matrix corresponding to* $\hat{\rho}_{\mathrm{min}}$.

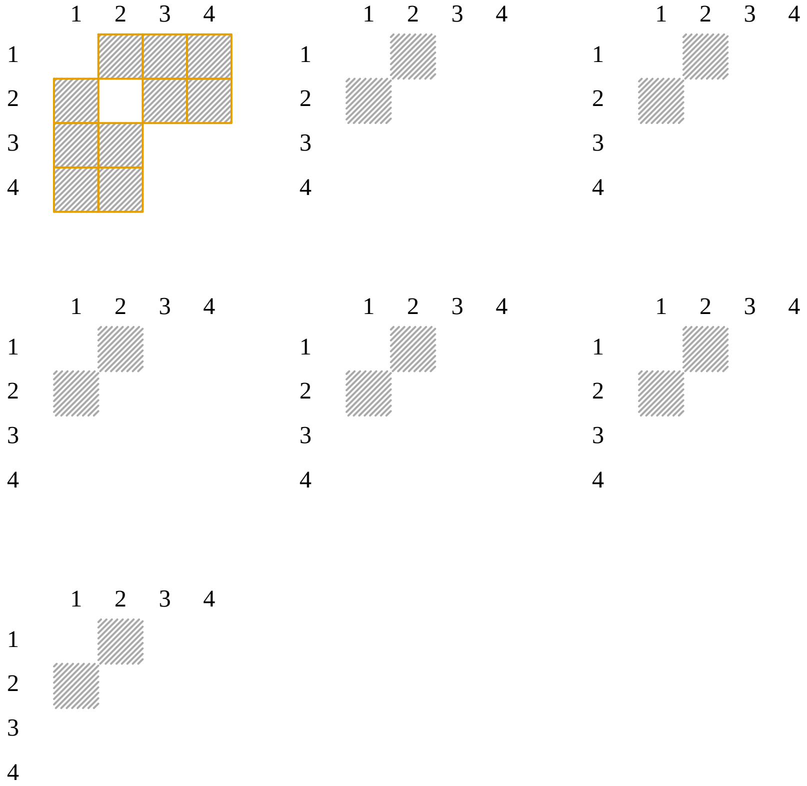


Figure B.5: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for the second metabolomic sample of the Landraces. Grey squares represent estimated edges and white space represent no edges. Orange borders represent adjacency matrix corresponding to $\hat{\rho}_{\min}$.*

C Tables

SNPID	Race	Chromosome	Position	GWAS	lasso	mvlasso
30362	Duroc	1	8803333		●	
29277	Duroc	3	11597405		●	
3471	Duroc	4	9617724		●	
21918	Duroc	4	36793710		●	
3724	Duroc	4	36811567		●	
3739	Duroc	4	39536242		●	
3747	Duroc	4	41360877		●	
3748	Duroc	4	41452288		●	
50049	Duroc	6	24222649		●	
59222	Duroc	6	24389455		●	
42902	Duroc	11	5464980		●	
8790	Duroc	11	27815254		●	
25474	Duroc	11	27829151		●	
8875	Duroc	11	47689807		●	
47357	Duroc	11	48200175		●	
52162	Duroc	11	48299203		●	
8880	Duroc	11	48376222		●	
61714	Duroc	12	5900678		●	
30624	Duroc	12	7651064		●	
61923	Duroc	12	21742885		●	
9525	Duroc	13	9580136		●	
60030	Duroc	14	108751907		●	
44572	Duroc	17	34347812		●	
28557	Duroc	17	34384763		●	
44574	Duroc	17	34497668		●	
13820	Duroc	18	26687448		●	
28869	Duroc	18	26745288		●	
29057	Duroc	19	13152222	●	●	
14071	Duroc	19	13190474	●	●	
44885	Duroc	19	13245808	●	●	
14913	Duroc	19	13448866	●	●	
15414	Duroc	19	106199952		●	
20297	Landrace	1	202481802			●
16478	Landrace	1	270659673		●	
29722	Landrace	2	151237469		●	●
2749	Landrace	2	151248688		●	●

(Continued on next page)

SNPID	Race	Chromosome	Position	GWAS	lasso	mvlasso
2752	Landrace	2	151433869		●	
30660	Landrace	3	6069989		●	
39701	Landrace	3	22551656		●	
40769	Landrace	3	22843902		●	
2863	Landrace	3	22855839		●	
2861	Landrace	3	22952343		●	
2858	Landrace	3	22983641		●	
58981	Landrace	3	23156713		●	
3075	Landrace	3	72191624		●	
40866	Landrace	3	92827526		●	●
59499	Landrace	3	126039599			●
21608	Landrace	3	126210935			●
3481	Landrace	4	10597928		●	
41246	Landrace	4	114290737		●	
36411	Landrace	4	118826911		●	
22756	Landrace	5	66837858		●	
5196	Landrace	5	104263450			●
5197	Landrace	5	104300278			●
5198	Landrace	5	104338721			●
5192	Landrace	5	104693932			●
5189	Landrace	5	104937103			●
22884	Landrace	5	105544703			●
33983	Landrace	5	111371828			●
52484	Landrace	6	9277653			●
31231	Landrace	6	9952971			●
22987	Landrace	6	35605675			●
5839	Landrace	7	18668395		●	
5875	Landrace	7	23022559		●	
31066	Landrace	8	42378737		●	
53405	Landrace	9	18490706		●	
50464	Landrace	9	18519025		●	
15547	Landrace	9	18599832		●	
51386	Landrace	9	18740429		●	
57831	Landrace	9	18765671		●	
18537	Landrace	9	18939093		●	
31844	Landrace	9	20062541		●	
56053	Landrace	9	62513471		●	
42599	Landrace	9	74724089			●
7897	Landrace	9	92626084		●	
52927	Landrace	9	92976421		●	

(Continued on next page)

SNPID	Race	Chromosome	Position	GWAS	lasso	mvlasso
42617	Landrace	9	93098631		●	
7899	Landrace	9	93125776		●	
7900	Landrace	9	93148647		●	
7902	Landrace	9	93743533		●	
50527	Landrace	9	93851347		●	
7905	Landrace	9	93901993		●	
30878	Landrace	9	95164137		●	
42619	Landrace	9	95957104			●
53117	Landrace	9	95973005			●
54769	Landrace	9	100612906		●	●
24827	Landrace	9	100627098		●	●
61047	Landrace	9	100949090		●	●
31177	Landrace	9	100971561		●	●
31415	Landrace	9	100985297		●	●
8115	Landrace	9	139376932		●	
42872	Landrace	10	70144318			●
25316	Landrace	10	70158066			●
59583	Landrace	12	16459010		●	
9273	Landrace	12	19513657		●	
33842	Landrace	12	41853982			●
16358	Landrace	12	42035618			●
17816	Landrace	13	93356864			●
10056	Landrace	13	93567191			●
26703	Landrace	14	14392171		●	
27588	Landrace	15	2136441			●
28308	Landrace	16	69860082			●
13053	Landrace	16	69865172			●
49644	Landrace	16	78120686		●	
14138	Landrace	19	38708793			●
19073	Landrace	19	104170699		●	
29196	Landrace	19	125159088			●
14294	Landrace	19	125186910			●
56840	Landrace	19	125199118			●
14295	Landrace	19	125294839			●

Table C.1: *Overview of SNPs selected by the standard GWAS method, the lasso, the two-step procedure, and the multivariate lasso.* (Continued)

Manuscript I

Identifying Interactions via Hierarchical Lasso Regularisation

ANN-SOPHIE BUCHARDT AND CLAUS THORN EKSTRØM

Identifying Interactions via Hierarchical Lasso Regularisation

ANN-SOPHIE BUCHARDT*, CLAUS THORN EKSTRØM

Section of Biostatistics, Department of Public Health, University of Copenhagen,

Øster Farimagsgade 5, 1014 København K, Denmark

Corresponding author: *asbu@sund.ku.dk

Abstract

Penalised regression models such as the lasso are powerful methods, which use sparsity to do feature selection when the number of features measured is larger than the number of observations. This is typical for genomic or multi-omics data, and when analysing such data we are concerned not only with identifying relevant features but also with identifying more complex relationships such as gene-gene interactions. It is, however, not clear exactly how to systematically include interacting features in penalised regression models.

We consider methods for identifying pairwise interactions in a linear regression model under different assumptions of hierarchy.

We approach the problem by using a two-step procedure. In the first step lasso includes only the main effects and selects the most promising. Next, adaptive lasso includes main effects and interactions according to hierarchical restrictions and selects the most promising terms. The approach is motivated by modelling pairwise interactions for qualitative variables and experimenting with explicitly applying penalties on the main effects and interactions, thereby obtaining interpretable models. We compare the methods with existing techniques on simulated data using R and illustrate the utility of the methods by examining a data set of a heterogeneous stock of mice and identifying several gene-gene interactions.

Keywords: hierarchical; interactions; lasso; regularised regression; sparsity

1 Introduction

The existence of, e.g., gene-gene or gene-environment interactions has gained attention in genetic studies but the introduction of interactions into a model rapidly increases the complexity of said model, see Hu et al. (2014) and Cornelis et al. (2012). If the number of measured features is p, the total number of possible pairwise interactions is $\binom{p}{2} = \frac{1}{2}p(p-1)$. Thus, when using 500K single nucleotide polymorphism microarrays, say, there are approximately 125 billion pairwise interactions. Fitting a regression model to such data is computationally challenging, potentially extremely time consuming, and mathematically distorts the relation between the number of features, p, and the number of observations, N, as the design matrix no longer has full rank. The objective of this paper is to propose a computationally efficient method for selecting pairwise interactions in a linear model which involves only a subset of the features.

Interpretations of traditional regression analyses do not encourage including an interaction in a model without the corresponding main effects. For example, Lim and Hastie (2015) argue:

> 'Since main effects [...] can be viewed as deviations from the global mean, and interactions are deviations from the main effects, it rarely make sense to have interactions without main effects.'

When deciding on a model where more than two features are present and give rise to multiple interactions we face the additional challenge of deciding which interactions to consider. According to Cox (1984):

> 'One general principle that can be used in such cases is that large component main effects are more likely to lead to appreciable interactions than small components.'

Likewise, Bien et al. (2013) argue:

> '[R]ather than looking at all possible interactions, it may be useful to focus our search on those interactions that have large main effects.'

This supports the opinion that an interaction should not enter a model without the corresponding main effects and further expresses the view that larger main effects are of more practical importance.

It is, however, possible that the magnitude of the true effect of a feature, is unrelated to both the direction and magnitude of the true associated interaction effects: two treatments, A and B, may be ineffective when given alone but helpful or detrimental when given in combination. For example, according to Leekha et al. (2011), in certain cases of serious infections, the addition of gentamicin to penicillin has been shown to be bactericidal, whereas penicillin alone is only bacteriostatic and gentamicin alone has no significant activity.

Our work is motivated by the wish to do cross-omics comparisons and to understand synergy in genetic interactions, that is, when combinations of genes carry more information than the sum of information provided by individual genes. While hierarchical assumptions may be fulfilled for some degrees of synergy in gene pairs, they are not fulfilled when each individual gene does not carry any information on the phenotype under study, while a simultaneous consideration of the two genes produces an association with the phenotype.

We propose a simple two-step penalised regression procedure for identification of pairwise interactions, such as high-synergy gene pairs, when *strong or weak hierarchy* is assumed and suggest the usual lasso for identification of pairwise interactions in situations where hierarchical assumptions do not seem reasonable. We compare the results of feature selection by the hierarchical and non-hierarchical lasso on data sets which exhibit different hierarchical properties.

Our method is inspired by the *hierarchical interaction models* for which the inclusion of interactions depends on the significance of the constituent main effects.

Different penalised regression methods have already been formulated for hierarchical interaction models. For example, Bien et al. (2013) propose a procedure which produces sparse estimates while satisfying either weak or strong hierarchy by adding a set of convex constraints to lasso. The method is implemented in the R package `hierNet`, but the implemented algorithm is not suitable for large scale problems. Other methods make use of the *group-lasso* and *overlapped group-lasso* to select interactions and enforce hierarchy, see, e.g., Lim and Hastie (2015) and Meier et al. (2008), and then others use stepwise procedures where the "best" variable is iteratively added or removed thereby enforcing the strong hierarchy restriction, see, e.g., Bickel et al. (2010) and Park and Hastie (2008). Wu et al. (2010) propose a two-step procedure not too different from the one proposed in this paper, but their method always enforces strong hierarchy and they provide no information on the scalability of the method.

In Section 2 we introduce the pairwise interaction model, the concept of hierarchical sparsity, and our framework and two-step procedure for finding pairwise interactions. We study our method and other techniques on real data, as well as simulated data, in Section 3. The two-step procedure is implemented using the R package `glmnet`; the R code is part of our package `ilasso` which is available online, see Buchardt (2021). We conclude with a discussion and recommendations in Section 4.

2 Methods

In this section we present a two-step penalised regression procedure which enables the detection of pairwise interactions.

A set of categorical features is said to interact if the effect of one feature depends on the level(s) of other feature(s) in the set. Similarly, a categorical and a quantitative feature are said to interact, if the partial slope corresponding to the quantitative variable depends on the level of the categorical variable. Estimates and hypothesis tests are calculated similarly for both cases, and interpretations are fairly clear, see Ekstrøm and Sørensen (2014). In the case of interactions between quantitative features, testing and interpretation is more complex. A comprehensive source on the treatment of such interactions in multiple regression is found in Aiken and West (1991). The method developed in this paper is applicable to all three cases; for exposition purposes, all our examples regard interactions between categorical features.

We focus in the methods section on linear regression models which are suitable when the outcome is quantitative, and ideally when the error distribution is Gaussian. The form of technique in our methodology is easily generalised to other regression models, and we will be illustrating the utility of our method by examining interactions on both quantitative and binary outcome. See Hastie et al. (2015) for a discussion on generalisations of simple linear models and the lasso which are suitable for different types of outcome.

The *full model* includes an additive effect of the features as well as an interaction between the two and takes the form

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\Theta_{12} + \epsilon_i,$$

for all observations $i = 1, \dots, N$ where β_0 is the intercept, β_1 and β_2 are unknown parameters for the main effects, Θ_{12} is a parameter for the pairwise interaction, and ϵ_i are Gaussian error variable. We construct an interaction variable from the product of the original two features; in the case of binary features the interaction variable estimates the difference in means between the two levels. The additive linear regression model or *pure main effect model* is one approach for modelling the additive effect of two features on a quantitative outcome and disregard potential interactions. It corresponds to the assumption $\Theta_{12} = 0$. The *pure interaction model* is an approach for describing the situation in which neither of the two features has any effect on the outcome singly but a combination of the two does. It corresponds to the assumption $\beta_1 = \beta_2 = 0$. Finally, the *partial model* includes an additive effect of only one of the features but also the combination of them both corresponds to $\beta_1 = 0$ *or* $\beta_2 = 0$.

Now, selecting which model to fit to a given data set is not straight forward. Often, only the pure main effect model is considered, but without including the interaction term, any main effect found is potentially partly due to marginal effects of an interaction.

The full model is said to honour so-called strong hierarchy and the partial models are said to honour weak hierarchy. We look into these hierarchical restrictions later, but for now we acknowledge the difficult task of deciding on a suitable hierarchy regardless of the data at hand.

An additional challenge involves the parametrisation of the model, since it is possible that different parametrisations result in different hierarchies. An example is presented in Table 1. This challenge is not specific to our method. It is a general problem arising when interactions are included in regression models, and theoretical and simulation studies of the consequences of incorrect parametrisations is beyond the scope of this paper.

When searching over all features and pairs of features simultaneously, it is possible that main effects mask interaction effects and vice versa. Imposing hierarchical restrictions in a stepwise manner, by searching for prominent main effects first, then only considering corresponding interactions, does not necessarily imply that interactions are easily found, since the underlying hierarchical structure is most likely unknown and different from the assumed hierarchical structure. In this paper we propose a two-step procedure, which is useful even when the underlying hierarchical structure is unknown and tackles the issue of masked effects in the case of categorical features.

	$\mathbf{x}_2 = 0$	$\mathbf{x}_2 = 1$
$\mathbf{x}_1 = 0$	0	0
$\mathbf{x}_1 = 1$	0	1

	$\mathbf{x}_2 = 0$	$\mathbf{x}_2 = 1$
$\mathbf{x}_1 = 1$	0	1
$\mathbf{x}_1 = 0$	0	0

Table 1: *As an example of different parametrisations resulting in different hierarchies, consider the simple case of a two-factor design. We assume that we have N observations of a univariate outcome $\mathbf{y} \in \mathbb{R}^N$, e.g., the severity of a disease, and two associated binary features $\mathbf{X} \in \{0,1\}^{N\times 2}$, e.g., indicating whether either of two single nucleotide polymorphisms (SNPs) are present. We show two different parametrisations of a possible sample of the average severity of disease, $E[\mathbf{y}]$, as a function of the combination of the SNPs $\mathbf{x}_1$ and $\mathbf{x}_2$.* Left: *The binary features are encoded using a dummy variable representation: they take the value 0 to indicate the absence of the SNP and 1 to indicate the presence of the SNP. If we use '0' as the reference level, the sample suggests a pure interaction model with $\Theta_{12} = 1$, that is, $y_i = x_{i1}x_{i2}$.* Right: *Instead let $\mathbf{x}_1$ take the value 1 to indicate the absence of the SNP and 0 to indicate the presence of the SNP. Now, if we still use '0' as the reference level, the sample suggests a partial model with $\beta_2 = 1$ and $\Theta_{12} = -1$, that is, $y_i = x_{i1} - x_{i1}x_{i2}$.*

Since our interest in interactions goes beyond the two-factor design, we extend the pairwise interaction model to p features in the next section.

2.1 Pairwise interaction model

We assume that we have N observations of the univariate outcome $\mathbf{y} \in \mathbb{R}^N$ and p associated features $\mathbf{X}$ with pairwise interactions. If the features are quantitative, $\mathbf{X} \in \mathbb{R}^{N\times p}$, the pairwise interaction term is formed simply by multiplying the two corresponding features. For computational reasons we recommend centring quantitative features before the optimisation problem is solved, see Aiken and West (1991), such that each column has mean zero. If the features are not measured in the same units, we also recommend scaling the features such that each column has unit variance. Otherwise the lasso solution depends on the scale since lasso puts constraints on the size of the coefficients associated to each feature. If the features are categorical, they are assumed to be represented by dummy variables, and the pairwise interaction term is formed by multiplying the corresponding variables. For the sake of simplicity we assume that the outcome values are centred before the optimisation problem is solved and the intercept term is omitted from the model. In linear regression models, this condition is not a restriction, since given an optimal solution, $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, obtained from the centred data, an optimal solution, $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$, for the uncentred data is easily recovered: $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and $\tilde{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \tilde{\beta}_j$, where $\bar{y}$ and $\bar{x}_j$, $j = 1, \ldots, p$, are the original means.

Thus, we consider the linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{k=1}^{p} \sum_{j<k} x_{ij}x_{ik}\Theta_{jk} + \epsilon_i, \tag{1}$$

for $i = 1, \ldots, N$. Here $\beta_1, \ldots, \beta_p$ are unknown parameters for the main effects, $\boldsymbol{\Theta} \in \mathbb{R}^{p \times p}$ are unknown parameters for the pairwise interactions, and ϵ_i is a random error variable. We let $\boldsymbol{\Theta}$ represent a symmetric matrix, i.e., $\boldsymbol{\Theta} = \boldsymbol{\Theta}^\top$. The strict inequality in the interaction summation precludes over-parametrisation arising from the inclusion of the same effect twice, e.g., including both $x_{ij}x_{ik}$ and $x_{ik}x_{ij}$.

Based on the model (1) we aim to select a subset of the p main effects and the $\frac{1}{2}p(p-1)$ interactions – we refer to the variables in this subset as the *relevant* variables. In pursuit of this goal we establish the notions of *hierarchy* and *sparsity* in the next sections.

2.2 Hierarchy

We already introduced the challenges connected to model selection – in the sense of hierarchical restrictions – as far as interactions are concerned. In this section we present model restrictions in a form which makes it is possible to specify a penalised regression procedure which produces sparse interaction models that honour these restrictions. This specification enables a data assisted and purpose driven choice of hierarchy, which is what we recommend and discuss in more detail in Section 4. We define the following hierarchical, anti-hierarchical, and non-hierarchical restrictions:

Strong hierarchy There are interactions only among pairs of non-zero main effects,

$$\mathrm{H_S}: \quad \Theta_{jk} \neq 0 \quad \Rightarrow \quad \beta_j \neq 0 \text{ and } \beta_k \neq 0.$$

Weak hierarchy Each interaction has at least one of its main effects present,

$$\mathrm{H_W}: \quad \Theta_{jk} \neq 0 \quad \Rightarrow \quad \beta_j \neq 0 \text{ or } \beta_k \neq 0.$$

Anti-hierarchy Interactions are only among pairs of main effects which are not present,

$$\mathrm{H_A}: \quad \Theta_{jk} \neq 0 \quad \Rightarrow \quad \beta_j = 0 \text{ and } \beta_k = 0.$$

Pure interactions There are no main effects present, only interactions,

$$\mathrm{H_I}: \quad \beta_j = 0 \quad \forall j = 1, \dots, p.$$

Pure main effects There are no interactions present, only main effects,

$$\mathrm{H_M}: \quad \Theta_{jk} = 0 \quad \forall j, k = 1, \dots, p.$$

No hierarchy There are no restrictions to the presence of main effects and interactions, $\mathrm{H_N}$.

As an example, recall the simple case of a two-factor design with N observations of the centred outcome $\mathbf{y} \in \mathbb{R}^N$ and two associated binary features $\mathbf{X} \in \{0,1\}^{N \times 2}$. In this case, the restrictions above result in the following models:

$$\begin{aligned}
\mathrm{H_S}: &\quad y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \mathbb{1}_{\{\beta_1 \neq 0 \wedge \beta_2 \neq 0\}} x_{i1}x_{i2}\Theta_{12} + \epsilon_i; \\
\mathrm{H_W}: &\quad y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \mathbb{1}_{\{\beta_1 \neq 0 \vee \beta_2 \neq 0\}} x_{i1}x_{i2}\Theta_{12} + \epsilon_i; \\
\mathrm{H_A}: &\quad y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \mathbb{1}_{\{\beta_1 = \beta_2 = 0\}} x_{i1}x_{i2}\Theta_{12} + \epsilon_i; \\
\mathrm{H_I}: &\quad y_i = x_{i1}x_{i2}\Theta_{12} + \epsilon_i; \\
\mathrm{H_M}: &\quad y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i; \\
\mathrm{H_N}: &\quad y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\Theta_{12} + \epsilon_i.
\end{aligned}$$

Next, we introduce methods for estimating the parameters in pairwise interaction models, which honour one of the hierarchical, anti-hierarchical, or non-hierarchical restrictions.

2.3 Sparsity

The *ordinary least squares* (OLS) method is a popular method for estimating the parameters in a linear model. If N is not much larger than p, the OLS estimates may exhibit a lot of variability resulting in almost sure over-fitting and are, therefore, likely to yield poor predictions. Finally, if $p > N$ the OLS estimates are not well-defined in which case the method cannot be used at all. However, in the case of $p > N$ it is unlikely that all features are relevant, anyway. This motivates the notion of *sparsity* and fitting a *sparse model* which involves only a subset of the features. In order for us to obtain such a subset we can regularise the estimation process.

For the usual linear regression model with no interactions,

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,$$

the *lasso* (least absolute shrinkage and selection operator), introduced by Tibshirani (1996), is an ℓ_1-regularised regression method, which shrinks the regression coefficients by imposing a penalty on their size. The objective of lasso is to solve an optimisation problem of the form

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{minimise}} && \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \\ &\text{subject to} && \sum_{j=1}^{p}|\beta_j| \leq t, \end{aligned} \tag{2}$$

where $t \geq 0$ is a pre-specified tuning parameter which controls the amount of shrinkage applied to the estimates. We can write the lasso problem in the Lagrangian form

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimise}}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}, \tag{3}$$

for some penalty parameter $\lambda \geq 0$, and for every value of λ there is a t (and for every value of t there is a λ) such that solving (2) and (3) results in the same estimates. The penalty parameter λ can be determined by a model validation technique such as cross-validation, and selecting the value of λ translates to selecting a proper amount of regularisation and is, as such, a trade-off between data fitting and sparsity. Hence, the lasso coefficients minimise a penalised residual sum of squares, and, since the absolute value function is not differentiable in zero, lasso has the ability to letting coefficients equal zero, thus resulting in a model which involves only a subset of the features. The lasso ℓ_1-penalty $\sum_{j=1}^{p}|\beta_j|$ makes the solution non-linear in y_i, and it has no closed form expression.

Now, fitting the usual lasso ℓ_1-penalty on the joint set of main effects and interactions corresponds to the assumption of no hierarchy, H_N. That is, for the pairwise interaction model (1), the lasso problem takes the form

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}}{\text{minimise}}\left\{q(\boldsymbol{\beta},\boldsymbol{\Theta}) + \lambda\sum_{j=1}^{p}|\beta_j| + \lambda\sum_{k=1}^{p}\sum_{j<k}|\Theta_{jk}|\right\},$$

where $q(\boldsymbol{\beta},\boldsymbol{\Theta})$ denotes the loss function

$$q(\boldsymbol{\beta},\boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{k=1}^{p}\sum_{j<k} x_{ij}x_{ik}\Theta_{jk}\right)^2.$$

In order for us to take into account actual hierarchical structures we can make use of the *group lasso penalty*, which results in *structured sparsity*, see Lim and Hastie (2015), or we can add a set of convex constraints to lasso as suggested by Bien et al. (2013) who introduce what they call the *all-pairs lasso*: To produce models that are guaranteed to be hierarchical Bien et al. (2013) build hierarchy into the lasso optimisation problem as a constraint on $\boldsymbol{\Theta}$, thereby obtaining the *strong hierarchical lasso* defined as

$$\begin{array}{lll} \underset{\boldsymbol{\beta}^{\pm}\in\mathbb{R}^p,\boldsymbol{\Theta}\in\mathbb{R}^{p\times p}}{\text{minimise}} & q(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-, \boldsymbol{\Theta}) + \lambda\mathbf{1}^\top(\boldsymbol{\beta}^+ + \boldsymbol{\beta}^-) + \lambda\sum_{k=1}^{p}\sum_{j<k}|\Theta_{jk}| & \\ \text{subject to} & \boldsymbol{\Theta} = \boldsymbol{\Theta}^\top, \quad \left.\begin{array}{l} \sum_{j=1}^{p}|\boldsymbol{\Theta}_j| \leq \beta_j^+ + \beta_j^- \\ \beta_j^+ \geq 0, \quad \beta_j^- \geq 0 \end{array}\right\} & \text{for } j = 1,\ldots,p, \end{array}$$

where $\boldsymbol{\Theta}_j$ denotes the jth row (and column, by symmetry) of $\boldsymbol{\Theta}$. Dropping the assumption $\boldsymbol{\Theta} = \boldsymbol{\Theta}^\top$ of symmetry of $\boldsymbol{\Theta}$ yields the *weak hierarchical lasso.*

The R package `hierNet` provides implementations of the strong and weak hierarchical lasso, both for Gaussian and logistic losses, see Bien and Tibshirani (2014). Currently, `hierNet` is, however, only “reasonably fast for moderate sized problems (100-200 variables)”, as is stated in the R documentation. Thus, when it comes to solving large-scale problems, there is still room for improvement.

In this paper we propose an approximation, by means of a variant of the *adaptive lasso* proposed by Zou (2006), which is computationally tractable and suited for taking into account both strong and weak hierarchical structures. While the method allows for an anti-hierarchical structure, it is not computationally tractable, often conceptually unrealistic, and we make no further mention of it in this paper. For the usual linear regression model with no interactions the adaptive lasso solves an optimisation problem of the form

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimise}}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\gamma_j|\beta_j|\right\}, \tag{4}$$

where the penalty parameter $\lambda \geq 0$ is determined by cross-validation and the quantities $\gamma_j \geq 0$, $j = 1, 2, \ldots, p$, is a penalty modifier: when $\gamma_j = 0$, feature j is never penalised; when $\gamma_j = \infty$ feature j is always excluded; for all $j = 1, 2, \ldots, p$ for which γ_j are equal to the same constant value, the corresponding features are equally penalised.

2.4 Two-step procedure

Solving the strong and weak hierarchical lasso outlined above is very time consuming, and in this paper we propose a simple two-step procedure for selecting relevant features and pairwise interactions according to one of the hierarchical restrictions, $\mathrm{H_S}$or $\mathrm{H_W}$, which were introduced in Section 2.2.

As the first step in our procedure we include only the main effects and apply the usual lasso for variable selection. As the second step in the procedure we include main effects and interactions in accordance with step one and subject to one of the restrictions $\text{H}_\text{S}$or H_W, and we apply a variation of the adaptive lasso with the penalty modifier depending on the given hierarchical restriction.

Formally, we define the procedure as follows:

Step 1 Define by $\mathcal{M}$ and $\mathcal{I}$ the sets of indexes for all the main effects and interactions, respectively, and assume the pure main effect model,

$$y_i = \sum_{j \in \mathcal{M}} x_{ij}\beta_j + \epsilon_i,$$

for which the lasso estimates are determined by solving the optimisation problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimise}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j \in \mathcal{M}} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j \in \mathcal{M}} |\beta_j| \right\},$$

where the penalty parameter $\lambda_1 \geq 0$ is determined by cross-validation. Define by $\mathcal{S} = \{k : \beta_k \neq 0\}$ the set of relevant main effects, that is, the set of non-zero coefficients, which is estimated by $\mathcal{S}^{(\lambda_1)} = \{k : \hat{\beta}_k^{(\lambda_1)} \neq 0\}$, that is, the set of non-zero estimates, corresponding to the main effects selected by lasso.

Step 2 Define by $\mathcal{M}_\text{H}$ and $\mathcal{I}_\text{H}$ the sets of main effects and interactions, respectively, to be included in the model subject to one of the hierarchical restrictions, $\text{H}_\text{S}$or H_W. For example, when the model is subject to the restriction of strong hierarchy, $\mathcal{M}_\text{H} = \mathcal{S}$ and $\mathcal{I}_\text{H} = \{j, k : \beta_j \in \mathcal{S} \wedge \beta_k \in \mathcal{S}\}$, which are estimated by $\mathcal{M}_\text{H}^{(\lambda_1)} = \mathcal{S}^{(\lambda_1)}$ and $\mathcal{I}_\text{H}^{(\lambda_1)} = \{j, k : \beta_j \in \mathcal{S}^{(\lambda_1)} \wedge \beta_k \in \mathcal{S}^{(\lambda_1)}\}$, respectively.

Then, assume the pairwise interaction model,

$$y_i = \sum_{j \in \mathcal{M}} x_{ij}\beta_j + \sum_{k \in \mathcal{I}_\text{H}^{(\lambda_1)}} \sum_{\substack{j<k \\ j \in \mathcal{I}_\text{H}^{(\lambda_1)}}} x_{ij}x_{ik}\Theta_{jk} + \epsilon_i,$$

for which the adaptive lasso estimates are determined by solving the optimisation problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\Theta} \in \mathbb{R}^{p \times p}}{\text{minimise}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j \in \mathcal{M}} x_{ij}\beta_j - \sum_{k \in \mathcal{I}_\text{H}^{(\lambda_1)}} \sum_{\substack{j<k \\ j \in \mathcal{I}_\text{H}^{(\lambda_1)}}} x_{ij}x_{ik}\Theta_{jk} \right)^2 \right.$$

$$\left. + \lambda_2 \sum_{j \in \mathcal{M}} \gamma_j |\beta_j| + \lambda_2 \sum_{k \in \mathcal{I}_\text{H}^{(\lambda_1)}} \sum_{\substack{j<k \\ j \in \mathcal{I}_\text{H}^{(\lambda_1)}}} |\Theta_{jk}| \right\},$$

where the penalty parameter $\lambda_2 \geq 0$ is determined by cross-validation and the penalty modifier $\gamma_j \geq 0, j = 1, 2, \ldots, p$, is defined as $\gamma_j = \mathbb{1}_{\{j \notin \mathcal{M}_{\mathrm{H}}^{(\lambda_1)}\}}$ to ensure that the main effects selected in the first step are never (or always) penalised in the second step. For example, when the model is subject to the restriction of strong or weak hierarchy, $\gamma_j = \mathbb{1}_{\{j \notin \mathcal{M}_{\mathrm{H}}^{(\lambda_1)}\}} = \mathbb{1}_{\{j \notin \mathcal{S}^{(\lambda_1)}\}} = \mathbb{1}_{\{j \notin \{k : \hat{\beta}_k^{(\lambda_1)} \neq 0\}\}}$, to ensure that the main effects selected in the first step are never penalised in the second step. That is, the main effects selected in the second step are the same as those selected in the first step, thereby, ensuring that the same main effects honour the assumed hierarchical restriction in both steps.

We are finally able to identify the set $\mathcal{S}_{\mathrm{H}} = \{k : \beta_k \neq 0\}$ of main effects relevant under the model and estimated by $\mathcal{S}_{\mathrm{H}}^{(\lambda_2)} = \{k : \hat{\beta}_k^{(\lambda_2)} \neq 0\}$ and the set $\mathcal{R}_{\mathrm{H}} = \{j, k : \Theta_{jk} \neq 0\}$ of interactions relevant under the model and estimated by $\mathcal{R}_{\mathrm{H}}^{(\lambda_2)} = \{j, k : \hat{\Theta}_{jk}^{(\lambda_2)} \neq 0\}$.

Let us explain the reasoning behind the stepwise nature of the method and the modified (and non-) usage of the adaptive lasso. Given unlimited data, time, and memory we could include all main effects and pairwise interactions at once and use the adaptive lasso (4) with weights on the form $\gamma_j = 1/\|\hat{\beta}_j\|^{\nu}$, where $\hat{\beta}_j$ is the OLS estimate and $\nu > 0$. With this choice of weights, the adaptive lasso enjoys the oracle properties, see Zou (2006), and yields consistent variable selection. In practice, however, computational limitations make dimensionality reduction necessary when interactions are of interest, even for small- and moderate-sized data sets. Hence the reason for step one. In general, the lasso as used in step one is not consistent for variable selection and may include too many variables. Ordinarily, this suggests using the adaptive lasso, but in finite sample settings we ought to be liberal when including main effects to allow more interactions to be considered and avoid them being disregarded because of low power. Hence the reason for applying the usual lasso in step one. In step two we use a modification of the adaptive lasso (4) where the weights are defined as an indicator function. This does not retain the attractive oracle properties, but it ensures our goal of interpretation in terms of hierarchy and it allows our approach to work as a screening tool for finite sample sizes.

For categorical features the stepwise nature of the procedure has a great advantage. As previously mentioned, it is likely that any interactions will be seen as main effects in a marginal analysis, such as the first step of the procedure. On the other hand, in a simultaneous analysis of main effects and interactions, such as the second step of the procedure, different parametrisations may lead to different hierarchies. Furthermore, correlation or, even, collinearity between an interacting main effect and the corresponding interaction variable may occur. This makes the interpretation of the main effects difficult, but it does not reduce the predictive power of the model in a simultaneous analysis. These are, for us, useful properties, since, in a marginal analysis, both main effects of a pairwise interaction are included regardless of the parametrisation and we can use the main effects selected as indicators of potential interactions. In a succeeding simultaneous analysis the size of the main effects which were selected in the marginal analysis may change, but it is possible to ensure that they are included – regardless of the parametrisation. It should be noted that these properties imply that the procedure is not suitable when the truth is anti-hierarchy, since, in that case, the main effects corresponding to a relevant interaction are likely to be selected in the first step, resulting in the interaction not being included in the second step.

Furthermore, we would like to point out another concern arising from the stepwise nature, specifically the using of cross-validation twice on the same data. The main concern is the risk of overfitting. Overfitting models will generally have poor predictive performance, as they can exaggerate minor fluctuations in the data. This, again, may result in too many variables being included, but, as we argued above, we wish to be liberal when including main effects to allow more interactions to be considered and avoid them being disregarded because of low power. If, however, power is not a problem in the data at hand, we do suggest leaving, say, 50% of data out of the selection in the first round to be used for cross-validation in the second round.

Finally, we would like to make it clear, why the main effects are re-estimated in the second step. We are interested in obtaining interpretable models in terms of effect size as well as feature selection. Simultaneous estimation of main effects and interactions is key to correct interpretation of data, as omission of certain terms from a model changes estimates of the other parameters.

In the next section we investigate the efficacy of the procedure through empirical studies and compare the performance of our method to the application of lasso on the pure main effect model, the pure interaction model, and the full model as well as to the application of the strong and weak hierarchical lasso.

3 Results

We illustrate the utility of the two-step procedure described in Section 2.4 on both simulated and real data. The motivation for both sets of examples is to understand the performance of the two-step procedure on one of the important problems in statistical genetics: genotype by genotype interactions (G×G) in genomic selection. First, we use simulations to assess the false discovery rate of the interactions under different hierarchical assumptions on the data generating process and the method. Next, we assess the modelling performance of the two-step procedure in terms of computing time by using three simulation scenarios corresponding to two different hierarchies: weak and strong. Here, the goal is to show that our method is faster than other methods implemented in R. While the method allows for anti-hierarchical restrictions as well as strong and weak hierarchy, we perceive the anti-hierarchical structure primarily as a mathematical construction with little practical relevance. Therefore, there will be no discussion of interactions honouring anti-hierarchy in the applications.

Finally, we will assess the selection abilities of our method in a heterogeneous stock mouse data set, see Valdar et al. (2006), from the Wellcome Trust Centre for Human Genetics (WTCCC).

3.1 Simulated data

We wish to study the advantages and disadvantages of the two-step procedure under the different hierarchical restrictions introduced in Section 2.2. We compare our method to the usual lasso applied to the joint set of main effects and interactions, which corresponds to the assumption of no hierarchy. We also compare the method to the strong and weak hierarchical lasso implemented in the `hierNet` package where the penalty parameter is determined by cross-validation with 10 folds using the `hierNet.cv()` function.

3.2 Sampling procedure

Since the efficacy of a procedure depends on the true model generating the data we simulate different set-ups such that each of the different restrictions are tried as the ground truth, and we apply our method to each scenario separately.

The sampling procedure goes as follows: each sample is generated with $N = 200$ observations $\mathbf{y}$ and $p = 600$ associated 3-way categorical features $\mathbf{X} \in \{0, 1, 2\}^{200\times600}$ resulting in 179700 potential pairwise interactions. The choice of sample size is a trade-off between replication of real-world studies, precision in the assessment of performance, and time-wise costs. Where applicable, there are one or two main effects and/or one interaction in the data generating process. In practice, simulating several independent interactions would produce similar results when each of these interactions are considered in turn. From the features we generate observations by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is standard Gaussian noise vector. We create a sparse problem by letting $\beta_j = 0$ for all $j = 1, \ldots, p$, except for $\beta_1, \beta_2, \beta_3$, and Θ_{12}, for which different values are tried.

We fit the models in the two-step procedure using the R package `glmnet` by Friedman et al. (2010). The penalty parameters λ_1 and λ_2 are determined by cross-validation, which we perform using the `cv.glmnet()` function with 10 folds. The function fits a lasso penalised linear model using the `glmnet()` function, which it runs 10+1 times; the first to get a sequence of values for the penalty parameter, and then the remainder to compute the fit with each of the folds omitted. The cross-validated error is accumulated, and the mean cross-validated error (CVM) and standard deviation over the folds are computed. Since the folds in `cv.glmnet()` are selected at random the results of the procedure are random, and we reduce this randomness by running `cv.glmnet()` $R = 10$ times and average the error curves. That is, from each iteration, we obtain a sequence of values for the penalty parameter and a corresponding sequence of values for the CVM. From the cross-validation fit we choose the set of non-zero coefficients (corresponding to the discoveries) at the value of the penalty parameter corresponding to the minimum point-wise average of the CVM. Finally, to reduce the residual variation we repeat the sampling of the outcome as well as the fitting using the two-step procedure $B = 100$ times. Thus, in total the four methods (the two-step procedure assuming either strong or weak hierarchy and the usual lasso) are evaluated on about 20000 simulations each.

In order for us to compare the performance of the methods we use the false-discovery rate of the interactions (iFDR), defined as

$$\text{iFDR} = \frac{V^{(i)}}{V^{(i)} + S^{(i)}},$$

where $V^{(i)}$ is the number of false discoveries of interactions and $S^{(i)}$ is the number of true discoveries of interactions. Please note that the false-discovery rates are defined in a manner that returns a zero if the number of false discoveries is zero. If there are no false discoveries nor true discoveries, the false-discovery rates are not provided.

Under the assumption of no hierarchy, $B = 100$ repetitions of the simulations and the model fitting (the $R = 10$ times repeated application of the usual lasso with 10-fold cross-validation to the joint set of main effects and interactions) is a very lengthy operation. Thus, we are interested in eliminating features from the problem to reduce the computational load, and, to this end, we apply a *screening rule*, which, according to Hastie et al. (2015), has the potential to speed up the computation considerably while still providing the exact numerical solution. In particular, we apply the *global strong rule*, which discards feature j whenever

$$\left|\mathbf{x}_j^\top \mathbf{y}\right| < 2\lambda - \lambda_{\max},$$

where $\lambda_{\max} = \max_j |\langle \mathbf{x}_j, \mathbf{y} \rangle|$, that is the variable which has the largest absolute inner-product. Finally, we apply the usual lasso on the features left after screening using the `glmnet` package. The penalty parameter is determined by 10-fold cross-validation, which we perform using the `cv.glmnet()` function.

We present the mean iFDR (solid lines) with corresponding point-wise approximate confidence intervals which we compute as the mean iFDR plus/minus twice the standard error of the mean. We also present the average number of true interaction discoveries (dotted lines) conditioned on having discovered a true interaction. We refer to this number as the *recall* with respect to interactions. Please note that the average number of false interaction discoveries, which complete the the picture is not displayed. Since the simulated data honour strong, weak, or no hierarchy with one interaction effect, the desired number of true interaction discoveries is one. That is, the dotted lines should be close to one. Since we aim to control the expected proportion of interaction discoveries which are false, the desired value for the mean iFDR is zero. That is, the solid lines should to be close to zero.

3.2.1 Strong Hierarchy

In this section we present the result of applying the two-step procedure under different hierarchical assumptions when the data generating process honours strong hierarchy. Thus, we consider simulations for different effect sizes of the features $\mathbf{x}_1$ and $\mathbf{x}_2$ and the interaction between the two, that is for different given non-zero values of β_1, β_2, and Θ_{12}. We compare the results to results obtained by applying the usual lasso on the joint set of main effects and interactions, that is by assuming no hierarchy.

In Figure 1 we show the mean iFDR (solid lines) with corresponding point-wise approximate confidence intervals as a function of $\beta_1 = \beta_2$ when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. The lines are coloured according to the true value of Θ_{12}. In Figure 2 we show the recall with respect to interactions (dotted lines) as a function of $\beta_1 = \beta_2$ when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. Again, the lines are coloured according to the true value of Θ_{12}.

We observe that under the strong hierarchical model, $\mathrm{H_S}$, the mean iFDR is close to zero for all values of the two true main effects and all values of the true interaction. This indicates that when we do discover an interaction it is likely to be the correct one. We also observe that the recall with respect to interactions is almost constant in the true main effects and increases as the size of the true interactions increases. It appears that whenever the value of the true interaction is greater than or equal to 1, it is discovered more than 80 % of the times. Under the weak hierarchical model, $\mathrm{H_W}$, both the mean iFDR and the recall with respect to interactions is almost constant in the true main effects and increase as the size of the true interactions increases. It appears that the mean iFDR is over 70 % regardless of the value of the two true main effects and the true interaction. This suggests that more false interactions are discovered when the hierarchical restriction is relaxed. The relaxation does, however, not affect the recall with respect to interactions much. Under the assumption of no hierarchy, $\mathrm{H_N}$, the mean iFDR is greater than 80 % regardless of the value of the two true main effects and the true interaction. However, the number of true discoveries is one whenever the true size of the interaction is greater than 0.2 or the values of the two true main effects is greater than 0.4, and we conclude that, when disregarding the hierarchical structure, the one true interaction is likely to be among the many false discoveries. The values of the mean iFDR are very close when the size of the true interaction is greater than 0.2. Therefore, only the solid line representing the mean iFDR for $\Theta_{12} = 2.2$ is visible. The values of the recall with respect to interactions are very close when the size of the true interaction is greater than 0.2. Therefore, only the orange dotted line representing the recall with respect to interactions for $\Theta_{12} = 2.2$ is visible.

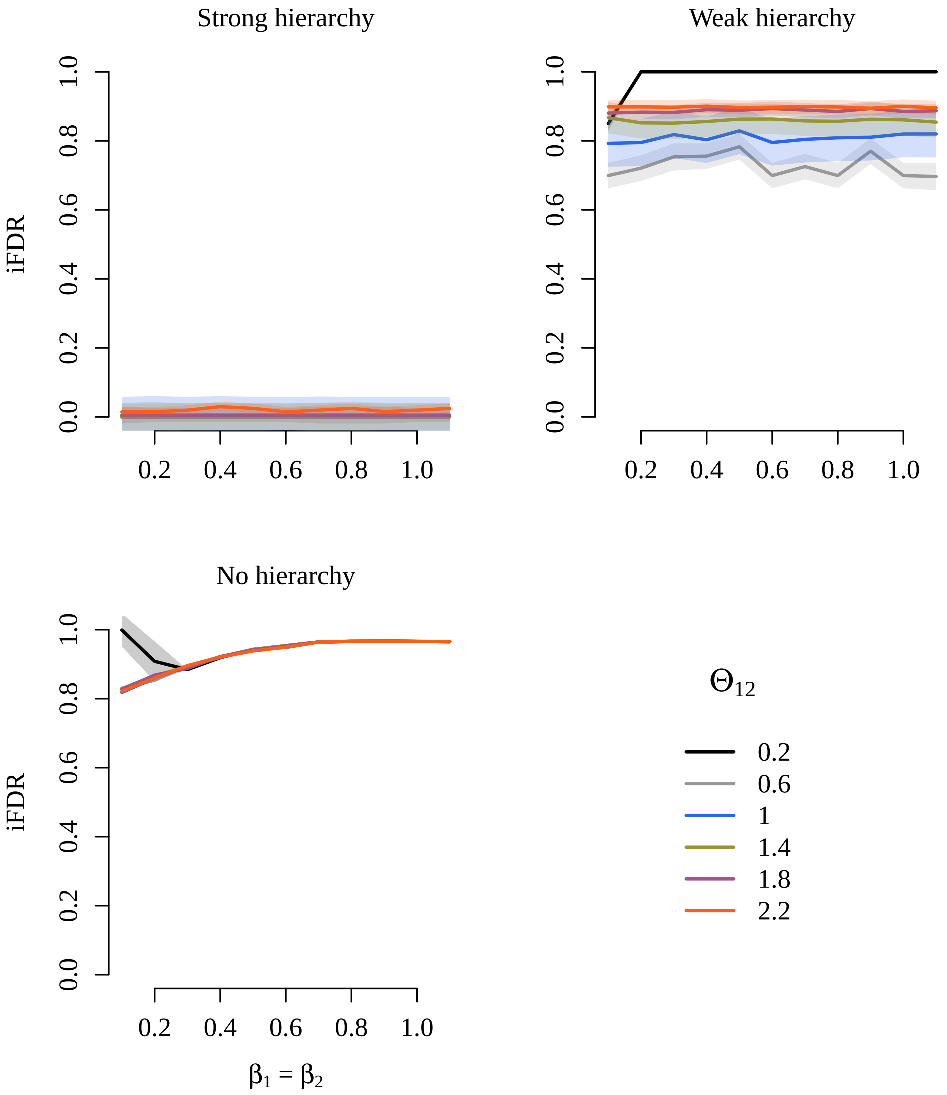


Figure 1: *The data generating process honours strong hierarchy. Mean* iFDR *(solid lines) as a function of $\beta_1 = \beta_2$ when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. Lines are coloured according to the true value, Θ_{12}, of the interaction. The lines should be close to zero.*

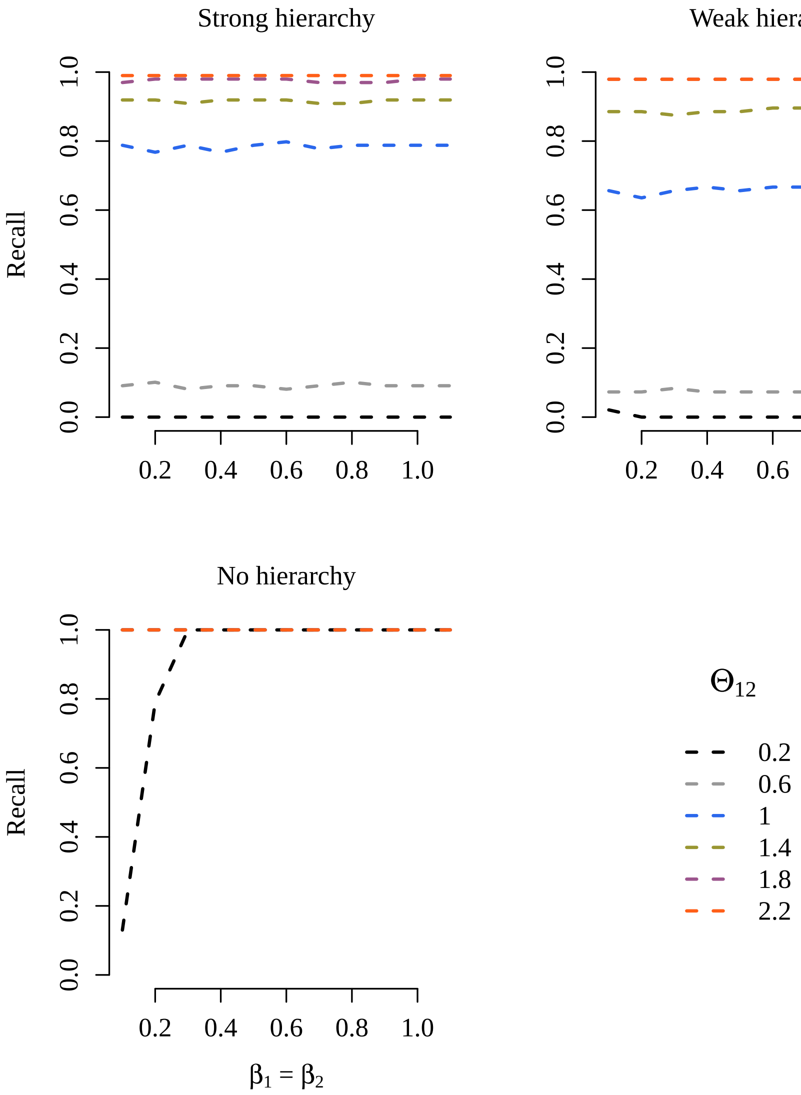


Figure 2: *The data generating process honours strong hierarchy. Recall with respect to interactions (dotted lines) as a function of $\beta_1 = \beta_2$ when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. Lines are coloured according to the true value, Θ_{12}, of the interaction. The lines should to be close to one.*

3.2.2 Weak Hierarchy

In this section we present the result of applying the two-step procedure under different hierarchical assumptions when the data generating process honours weak hierarchy. Thus, we consider simulations for different effects of the feature $\mathbf{x}_1$ and the interaction between the $\mathbf{x}_1$ and $\mathbf{x}_2$, that is for different given non-zero values of β_1 and Θ_{12}. We compare the results to results obtained by assuming no hierarchy.

In Figure 3 we show the mean iFDR (solid lines) with corresponding point-wise approximate confidence intervals as a function of β_1 when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. The lines are coloured according to the true value of Θ_{12}. In Figure 4 we show the recall with respect to interactions (dotted lines) as a function of β_1 when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. Again, the lines are coloured according to the true value of Θ_{12}.

We observe that the results are very similar to those obtained when the data generating process honours strong hierarchy: under $\mathrm{H_S}$ the mean iFDR is approximately zero for all values of the two true main effects and all values of the true interaction and whenever the value of the true interaction is greater than 1, it is discovered more than 75 % of the times. More false interactions are discovered when the hierarchical restriction is relaxed and $\mathrm{H_W}$ is assumed, and the recall with respect to interactions is decreased a bit in the value of the true interaction. Under $\mathrm{H_N}$ the results are almost identical to those obtained when the data generating process honours strong hierarchy.

We see that whether the data generating process honours strong or weak hierarchy with one interaction, the procedure performs very similarly under $\mathrm{H_S}$ and $\mathrm{H_W}$. Essentially, the only difference is, that under $\mathrm{H_S}$ the process is likely to discover one more false interaction when the data generating process honours weak hierarchy instead of strong hierarchy.

3.2.3 Other hierarchical restrictions

In Appendix A we discuss the result of applying the two-step procedure under different hierarchical assumptions when the data generating process honours the pure interaction set-up. We refrain from the pure main effect set-up as it results in a constant iFDR equal to one and zero discoveries of true interactions.

3.2.4 Comparison of methods

In this section we compare the performance – in terms of computing time, discoveries, and prediction error – of the two-step procedure and the `hierNet` package. Due to the long computation time of `hierNet` we cannot do as thorough an investigation of the technique

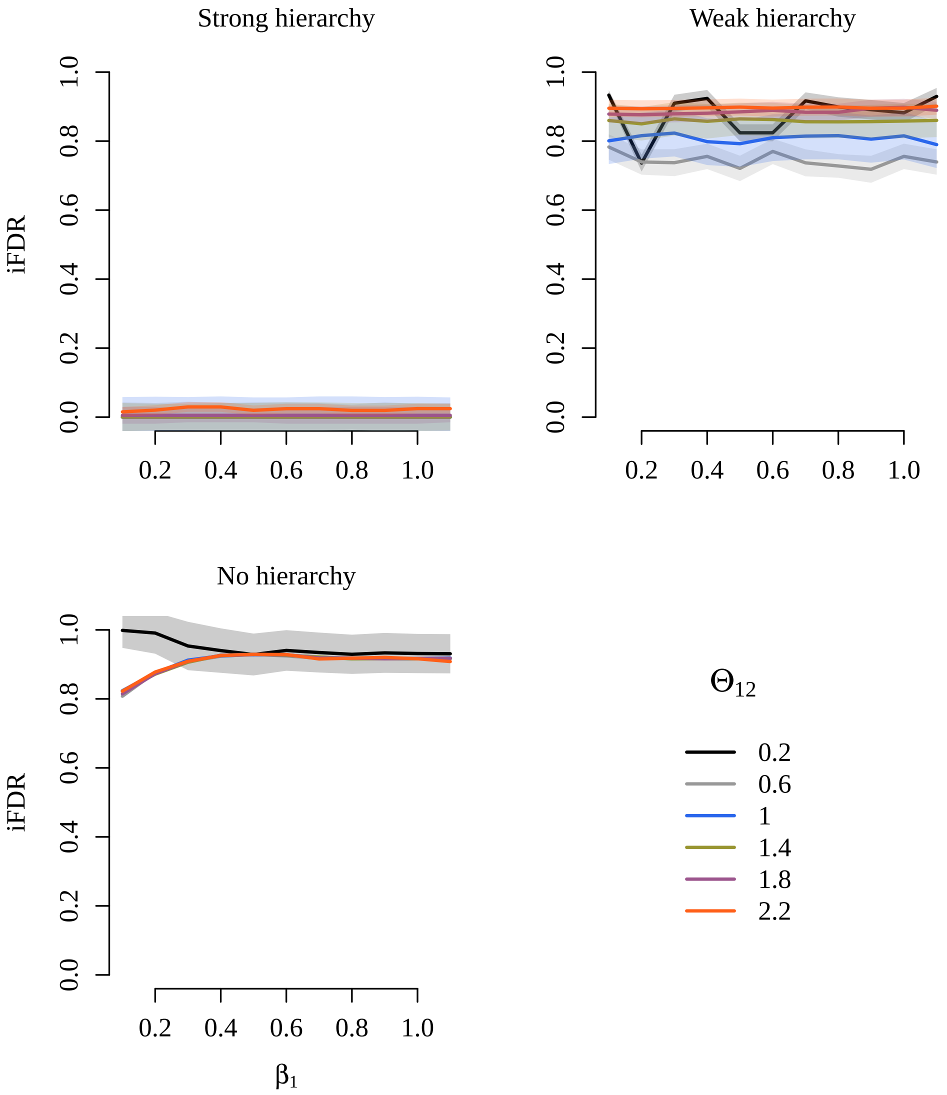


Figure 3: *The data generating process honours weak hierarchy. Mean iFDR (solid lines) as a function of $\beta_1 = \beta_2$ when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. Lines are coloured according to the true value, Θ_{12}, of the interaction. The lines should be close to zero.*

as we did of the two-step procedure in the previous sections. For example, fitting one sparse interaction model subject to the strong hierarchy restriction using cross-validation to estimate the regularisation parameter to a data set with $N = 200$ observations and $p = 600$ continuous

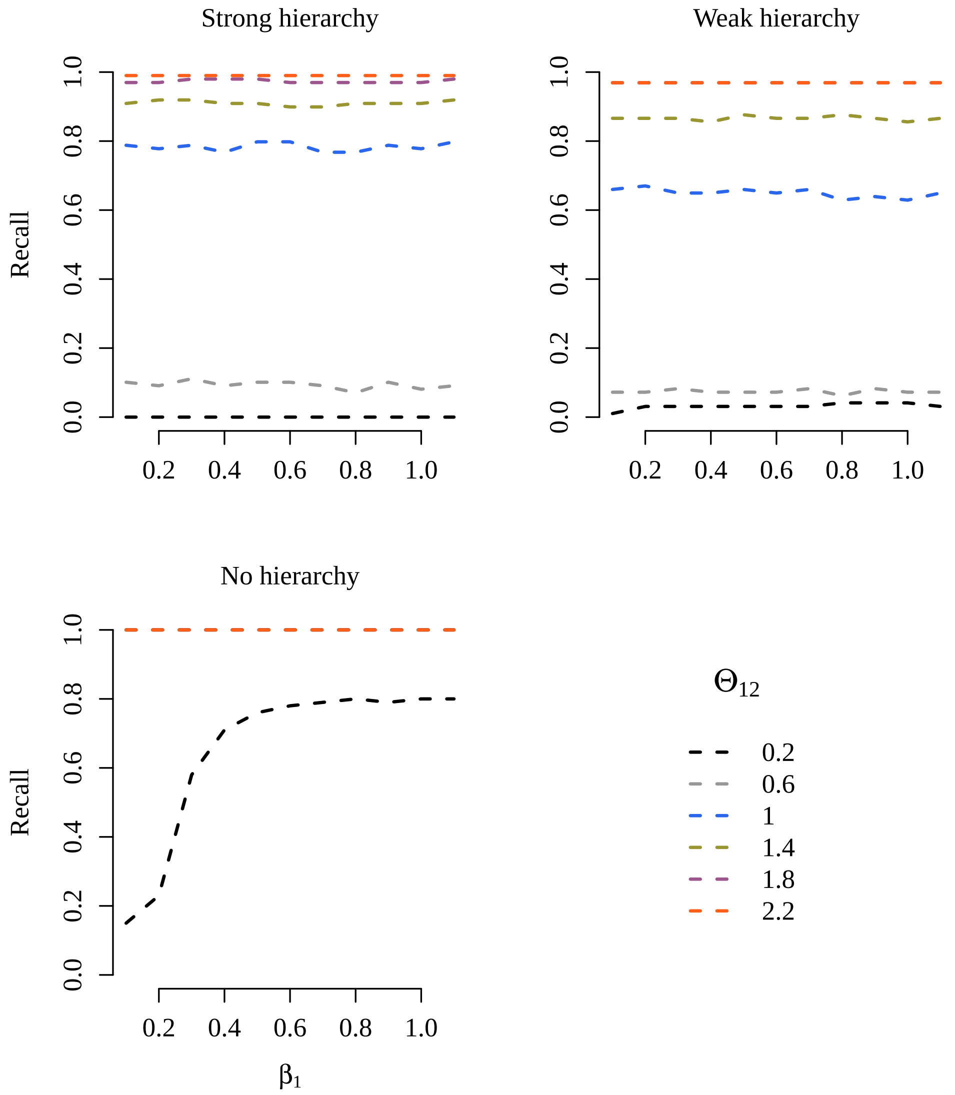


Figure 4: *The data generating process honours weak hierarchy. Recall with respect to interactions (dotted lines) as a function of* $\beta_1 = \beta_2$ *when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. Lines are coloured according to the true value,* Θ_{12}*, of the interaction. Dotted lines should to be close to one.*

features takes more than nine hours on an Intel Core i5-5200U processor.

Thus, we simulate two data sets, as described in Section 3.1, honouring strong and weak hierarchy respectively, only each sample is generated with $N = 100$ observations and $p = 200$

categorical features, and we consider only the case of $\beta_1 = 1$, $\beta_2 = 1$ (strong hierarchy), and $\Theta_{12} = 1$. For each data set we run strong and weak hierarchical lasso using the `hierNet.path()` and `hierNet.cv()` functions. We also run the two-step procedure using the `cv.glmnet()` function.

The resulting discoveries of interactions are summarised in Table 2. We observe that regardless of the true underlying hierarchy and the assumed hierarchy both techniques discover the true interaction. Furthermore, we observe that `hierNet` always discovers substantially more false interactions than the two-step procedure.

		Strong		Weak	
		`hierNet`	`2-step`	`hierNet`	`2-step`
Truth	Strong	1/11	1/1	1/17	1/11
	Weak	1/11	1/1	1/17	1/11
	Anti	1/20	0/0	1/39	1/22

Table 2: *Number of true/total discoveries of interactions for the* `hierNet` *technique and two-step procedure using the* `cv.glmnet()` *function when the data generating process honours strong, weak, or anti-hierarchy (rows) and the model assumes strong or weak hierarchy (columns).*

The resulting computation times in seconds are summarised in Table 3. We observe that the two-step procedure appears to be approximately 400 times faster than `hierNet`, regardless of the true underlying hierarchy and the assumed hierarchy. Even for fairly large data, the computation time is very satisfactory for the two-step procedure: for a data set honouring strong hierarchy with $N = 200$ observations and $p = 1000$ features, that is, approximately 500 thousand pairwise interactions, the computation time of the two-step procedure assuming strong hierarchy is 1.7 seconds when using the `cv.glmnet()` function. In comparison, for a data set with no more than $N = 50$ observations and $p = 100$ features, that is, approximately 5000 pairwise interactions, the computation time of the `hierNet` technique is 1.8 minutes.

		Strong		Weak	
		`hierNet`	`2-step`	`hierNet`	`2-step`
Truth	Strong	2971.86	0.93	382.60	0.94
	Weak	2817.00	0.91	299.66	0.88
	Anti	2898.10	0.88	319.64	1.22

Table 3: *Computing time (in seconds) for the* `hierNet` *technique and two-step procedure using the* `cv.glmnet()` *function when the data generating process honours strong, weak, or anti-hierarchy (rows) and the model assumes strong or weak hierarchy (columns).*

As suggested by Kyung et al. (2010), there is no consensus on a statistically valid method of calculating standard errors for lasso predictions. With this in mind as well as the added variance of cross-validation for both methods, we present comparisons of predictions errors of the two methods in the following. Thus, we simulate two data sets, as described in Section 3.1, honouring strong and weak hierarchy respectively, but to compute the MSPE over out-of-sample data, we need to run the procedures several times, and this is very time consuming with the `hierNet` package. Therefore, each sample is generated with $N = 1000$ observations (divided into ten folds) and only $p = 30$ categorical features. Again, we consider only the case of $\beta_1 = 1$, $\beta_2 = 1$ (strong hierarchy), and $\Theta_{12} = 1$. Box plots of the MSPEs are shown in Figure 5. Blue and orange colours indicate strong and weak hierarchical assumption, respectively; solid and striped indicate `hierNet` and the two-step procedure, respectively. We observe that whether the truth is strong (left panel) or weak (right panel) hierarchy, the results are very similar.

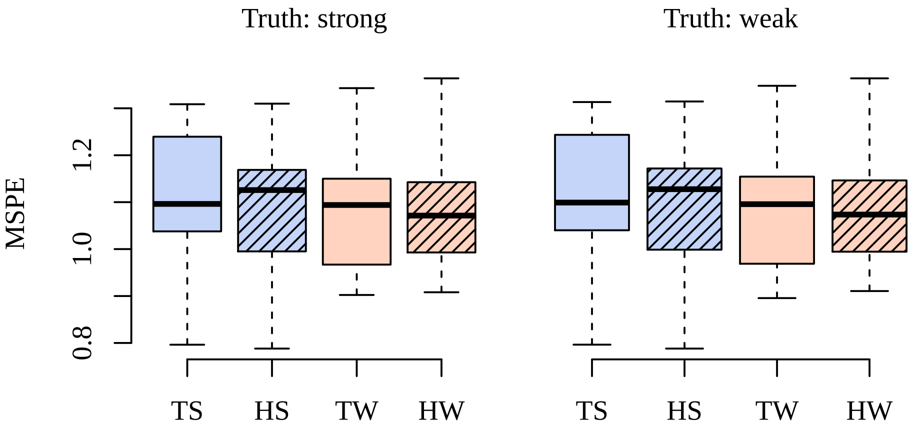


Figure 5: *Boxplots of prediction error when the data generating process honours strong (left panel) or weak (right panel) hierarchy. Blue and orange colours indicate strong and weak hierarchical assumption, respectively; solid and striped indicate `hierNet` and the two-step procedure, respectively.*

3.3 Real data

We apply our method to the heterogeneous stock of mice data set available from the Wellcome Trust Centre for Human Genetics (Valdar et al., 2006).

The heterogeneous stock of mice consists of 1,904 individuals from 85 families, all descended from eight inbred progenitor strains, see Valdar et al. (2006). Please note that the aim of the example is to demonstrate proof of concept; the proposed method is used as a screening tool for identifying potential interactions. This way we can subsequently investigate the associations using more complicated nested models taking into account the the biological relationships of related subjects. The data contains 129 quantitative traits which are classified into six broad categories including behaviour, diabetes, asthma, immunology, haematology, and biochemistry.

A total of 12,226 autosomal SNPs were available for all mice. We omit individuals with missing phenotype, and, as Crawford et al. (2018), for "individuals with missing genotypes, we imputed missing values by the mean genotype of that SNP in their family. All polymorphic SNPs with minor allele frequency above 1% in the training data were used for prediction". Furthermore, the SNPs are dichotomised, with '0' indicating the more common allele and '1' indicating the less common alleles.

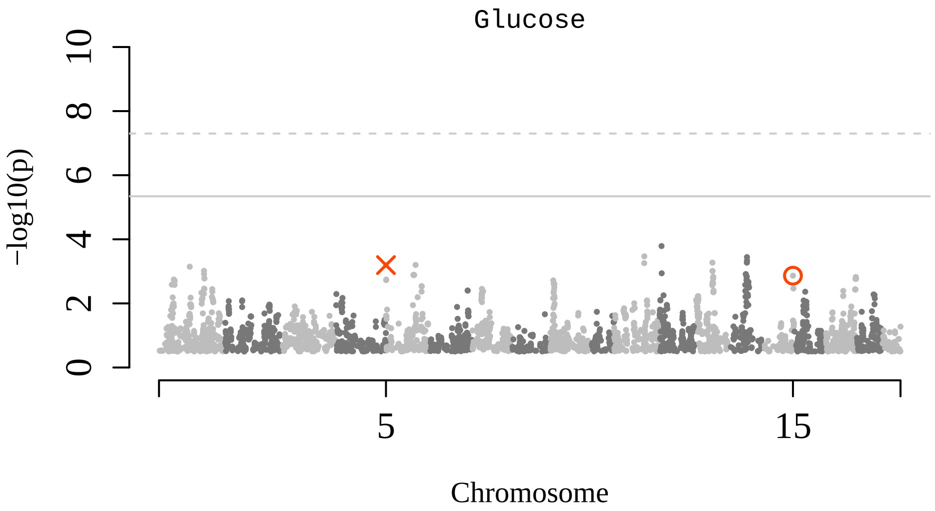
For each quantitative trait we are interested in detecting gene-gene interactions (G×G), that is, when the effect of one gene is dependent on another gene. The two-step procedure is applied on each outcome separately with 10 repetitions of 10-fold cross validation in both steps under the model assumption of strong hierarchy, $\texttt{H}_\texttt{S}$. We have chosen strong hierarchy, since, for a substantial number of the traits, the number of main effects, selected in the first step, results in more interactions than computationally feasible. For example, for the third trait (`AdrenalMeanWeight`), 59 main effects are selected in the first step. This causes R to crash, since, under the assumption of weak hierarchy, the number of interactions to be included in the second step is approximately 900 millions, whereas, under the assumption of strong hierarchy, the number of interactions to be included is approximately 18 millions which is doable in R on our computer.

For the biochemistry outcome `Glucose` with $N = 1499$ non-missing observations, $p = 10996$ SNPs have non-zero variation and are tried in the model. $|\mathcal{M}_{\text{H}_\text{S}}^{\lambda_1}| = 27$ SNPs are selected as main effects in the first step and $|\mathcal{M}_{\text{H}_\text{S}}^{\lambda_2}| = 41$ main effects and $|\mathcal{I}_{\text{H}_\text{S}}^{\lambda_2}| = 1$ interaction are selected in the second step. The interacting SNPs are `CEL-5_3149134` and `rs3657663`. For the diabetes outcome `Insulin_75` with $N = 1553$ non-missing observations, $p = 10984$ SNPs have non-zero variation and are tried in the model. $|\mathcal{M}_{\text{H}_\text{S}}^{\lambda_1}| = 108$ SNPs are selected as main effects in the first step and $|\mathcal{M}_{\text{H}_\text{S}}^{\lambda_2}| = 108$ main effects and $|\mathcal{I}_{\text{H}_\text{S}}^{\lambda_2}| = 1$ interaction are selected in the second step. The interacting SNPs are `CEL-3_74975193` and `rs4172689`. For the asthma outcome `EnhancedPause` with $N = 1499$ non-missing observations, $p = 10996$ SNPs have non-zero variation and are tried in the model. $|\mathcal{M}_{\text{H}_\text{S}}^{\lambda_1}| = 54$ SNPs are selected as main effects in the first step and $|\mathcal{M}_{\text{H}_\text{S}}^{\lambda_2}| = 67$ main effects and $|\mathcal{I}_{\text{H}_\text{S}}^{\lambda_2}| = 2$ interaction are selected in the second step. SNP `rs3708061` interacts with both `rs4164782` and `rs4164966`. The computation time is approximately 10–15 minutes for each step on a standard computer.

In Table B1 we provide a summary table which lists all significant SNPs detected by the two-step procedure for the three traits `Glucose`, `Insulin_75`, and `EnhancedPause`.

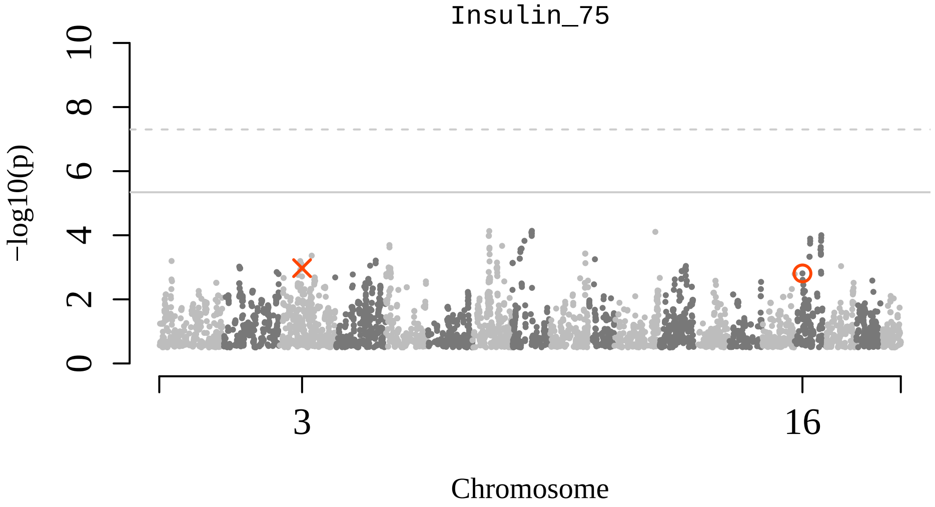
Columns list the three traits, and, for each trait, SNPs which are included in the sets $\mathcal{M}_{\text{HS}}^{(\lambda_1)}$, $\mathcal{M}_{\text{HS}}^{(\lambda_2)}$, $\mathcal{I}_{\text{HS}}^{(\lambda_2)}$, defined in Section 2.4, are indicated by bullets. Rows listed in bold indicate the SNPs between which an interaction is found.

In Figure 6 we show Manhattan plots of genome-wide association studies for `Glucose`, `Insulin_75`, and `EnhancedPause`. The figure shows that no SNP is significant with the commonly used genome-wide significance p-value threshold of 5×10^{-8} (dashed line) nor with the Bonferroni correction testing each individual hypothesis at $\alpha = 0.05/p$ (solid line). The main effects corresponding to the interactions identified by the two-step procedure are highlighted in the figure. We observe that for `Glucose` SNPs on chromosome number 5 and 15 are interacting, for `Insulin_75` SNPs on chromosome number 3 and 16 are interacting, and for `EnhancedPause` two SNPs on chromosome 16 are interacting on of which is also interacting with a SNP on chromosome 4.



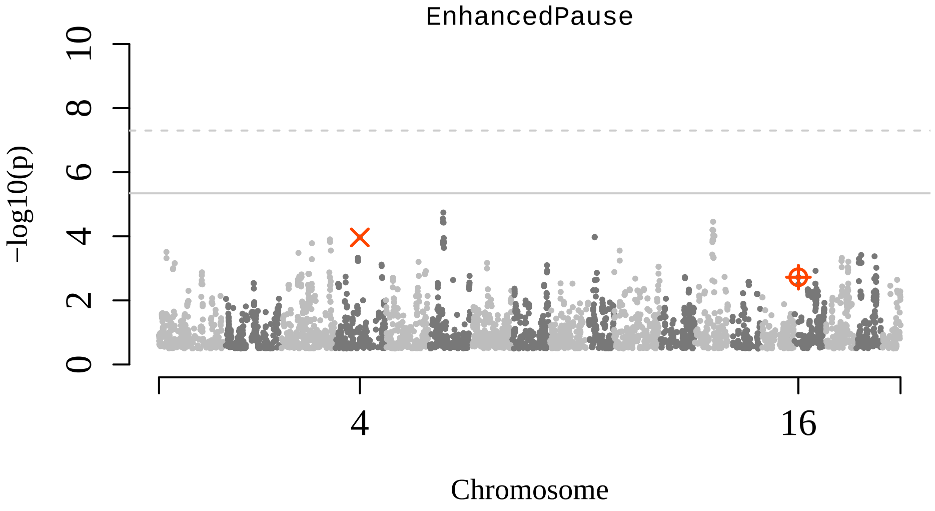
(a)

Figure 6: *(Continued on the next page.)*



(b)

Figure 6: *(Continued on the next page.)*



(c)

Figure 6: *Manhattan plots of genome-wide association studies for* `Glucose` *(a),* `Insulin_75` *(b), and* `EnhancedPause` *(c). Cumulated genomic positions are displayed along the first axis, with the negative logarithm (to base 10) of the association p-value for each SNP displayed on the second axis. Thus, each dot on the Manhattan plots represents a SNP. Chromosomes on which interacting SNPs are identified are highlighted. We observe that no SNP is significant with the commonly used genome-wide significance p-value threshold of* 5×10^{-8} *(dashed line) nor with the Bonferroni correction testing each individual hypothesis at* $\alpha = 0.05/p$ *(solid line). Nevertheless, significantly associated interactions are identified by the two-step procedure; the corresponding main effects are highlighted by different symbols in the figure. Please note that for* `EnhancedPause` *(bottom) the two selected interactions share a main effect, thus, only three SNP are highlighted.*

4 Discussion

In this paper, we have proposed a two-step procedure, which uses lasso and adaptive lasso for screening for interactions and fitting pairwise interaction models. A key advantage of our framework is that the two-step procedure discovers substantially fewer false interactions than the `hierNet`. Furthermore, the hierarchical restrictions are no cause for additional computing time compared to both the strong and weak hierarchical lasso as well as the usual lasso for the pairwise interaction model under the assumption of no hierarchy. On the contrary, the computing time is considerably reduced and is highly satisfactory even for fairly large data.

To understand the consequence of a wrong hierarchical assumption we simulated in Section 3.1 data sets with different (non-)hierarchical structures and applied the two-step procedure under different hierarchical assumptions. From the figures 1–4 we observed that both in terms of false discovery rate and recall with respect to interactions the two-step procedure works very well under the assumption of strong hierarchy whether the truth is strong or weak hierarchy. Under the assumption of weak or no hierarchy the recall with respect to interactions is good, however, the false discovery rate is high.

To understand the consequence of not requiring hierarchy we compared our method to the usual lasso for the pairwise interaction model under the assumption of no hierarchy and found that, regardless of the true hierarchical structure, the usual lasso tends to discover the true interaction but also many false interactions.

We conclude that, whatever the true data-generating hierarchy, the strong hierarchical assumption helps lowering the fals discovery rate while retaining a high recall with respect to interactions compared to the assumption of weak or no hierarchy.

The specification of a penalised regression procedure which produces sparse interaction models that honour hierarchical restrictions enables a data assisted choice of hierarchy, as comparing the results and regularisation paths of different model assumptions may help in shedding light on the underlying hierarchical structure of data. The short computing time of the two-step procedure is very advantageous in this regard.

If the underlying hierarchical structure of data is completely unknown we recommend a purpose driven choice of hierarchy as follows: If controlling the expected proportion of interaction discoveries which are false is the main concern, assuming strong hierarchy is recommended. If discovering the true interaction is prioritised but the iFDR is still a concern, assuming weak hierarchy is recommended. If discovering the true interaction is prioritised and neither the iFDR nor the computing time is any concern, assuming no hierarchy is recommended.

Finally, computational limits may restrict the choice of hierarchy. For example, the assumption of strong hierarchy can be computationally feasible in situations where weak hierarchy is not, since the number of interactions to be included in the model is usually smaller under the assumption of strong hierarchy than under the assumption of weak hierarchy.

In this paper, we see the choosing of the regularisation parameter primarily as a means to an end. Choosing the right penalty parameter is, however, difficult and, as is seen in Section 3, the cross-validated choices often include too many features. Meinshausen and Bühlmann (2010) show that choosing the right amount of regularisation is much less critical for the stability path used for stability selection than for the regularisation path used for cross-validation and that there is a better chance of selecting truly relevant variables with stability selection. Thus, it would be interesting to apply stability selection for choosing the right penalty parameters in the two-step procedure.

Our work has potential applications to cross-omics studies and genome-wide studies of G×G, that is, epistatic pairs where the effect of one gene (locus) is dependent on the presence of a *modifier gene*. It is possible to extend our framework to the case of multi-way interactions and, thereby, to the application to studies of, e.g., higher-order epistatic interactions. Increasing the order of interactions does, however, pose quite the challenge: given p features, the number of terms in a linear model which includes every features and every possible interaction is $\sum_{j=1}^{p} \binom{n}{j} = 2^p$. Thus, even the case where only pairwise and three-way interactions are allowed is computationally demanding. Furthermore, the presence of potential higher-order interactions has serious implications for the interpretation of the model and dramatically increases the chance of finding false-positive discoveries.

As a special case, our work has potential applications to studies of genotype by environment interactions (G×E), that is, when two different genotypes respond to environmental variation in different ways. In studies such as genetic epidemiology, presumably, a small-scale number of environmental factors and a large-scale number of genes will be available. In this case, penalisation of the environmental factors may be undesirable, and, with a slight modification, the two-step procedure is capable of taking this into account. For example, in order to select G×E when the model is subject to the restriction of weak hierarchy, $\mathrm{H_W}$, one could skip step one of the two-step procedure, define $\mathcal{M}_{\mathrm{H_W}}$ by the set of environmental factors, and define $\mathcal{I}_{\mathrm{H_W}}$ as the set of pairwise interactions between genes and environmental factors. In order to select G×E when the model is subject to the restriction of strong hierarchy, $\mathrm{H_S}$, one could include genes in step one of the two step procedure, estimate $\mathcal{M}_{\mathrm{H_S}}^{(\lambda_1)}$ by the union of genetic effects selected in step one and the environmental factors, and estimate $\mathcal{I}_{\mathrm{H_S}}$ by the set of pairwise interactions between genetic effects selected in step one and the environmental factors.

Lasso is a very useful technique for simultaneous estimation and variable selection. However, it has two major drawbacks. First, it does not possess the oracle property; that is, it does not perform as well as if the true underlying model was given in advance. Secondly, it is somewhat indifferent to grouping effects, that is, the selection among a set of strong but correlated features. Even though the adaptive lasso yields consistent estimates of the parameters while retaining the attractive convexity property of the lasso, see Zou (2006), these optimality properties are likely no longer true, when using cross-validation on the same data twice. As previously mentioned, we suspect that this reuse may be necessary in many practical cases due to lack of power. However, our method is ad hoc in nature, and our focus is on interpretability and scalability rather than unbiased and oracle properties of the estimators. As the lasso, the adaptive lasso does not capture any grouping effect. The doubly regularised technique *elastic net* proposed by Zou and Hastie (2005) encourages grouping effect, but it does not necessarily possess the oracle property. However, the adaptive elastic net proposed by Zou and Zhang (2009) inherits some of the desirable properties of the adaptive lasso and elastic net, and, in particular, it has the oracle property under certain regularity conditions. Hence, we intend to pursue the application of the two-step procedure with the adaptive elastic net in the case of correlated features. In the case of dependant features we should, however, beware that an observed interaction may be *spurious*; that is, an interaction between two features, X_1 and X_2, say, may be detected in the sample even when there is no true interaction in the population. The reason is that when the correlation between X_1 and X_2 increases so does the correlation between X_1X_2 and X_1, which results in an overlap between the variance explained by X_1X_2 and the variance explained by X_1. Dependence between gene and environment, for example, can arise in several ways including mediation, pleiotropy, and confounding, and several examples of gene-environment interaction under gene-environment dependence have recently been published, see, e.g., Johnson et al. (2014) and Nickels et al. (2013). Caution in reporting interactions for genetic markers is, however, urged by Dudbridge and Fletcher (2014) who show that under gene-environment dependence, a statistical interaction can be present between a marker and environment even if there is no interaction between the causal variant and the environment.

Several two-stage multiple testing procedures have been proposed to detect gene-environment interactions in GWASs, and in Dai et al. (2012) they discuss general properties of some of these procedures and prove the asymptotic independence of various filtering and testing statistics. While these methods use marginal association as a filter for interaction our method includes all features simultaneously via lasso, such that each coefficient represents the additional effect of adding the corresponding variable to the model, conditional on the effects of all other variables in the model. This way, selection (or not) of a feature in the first step is based on information from all other features when using our two-step procedure.

Appendices

A Other hierarchical restrictions

A.1 Pure Interaction

In this section we present the result of applying the two-step procedure under different hierarchical and anti-hierarchical assumptions when the data generating process honours the pure interaction set-up. Thus, we consider simulations for different effects of the interaction between the $\mathbf{x}_1$ and $\mathbf{x}_2$, that is for different given non-zero values of Θ_{12}. We compare the results to results obtained by applying the usual lasso on the joint set of main effects and interactions.

In Figure A1 we show the mean iFDR (solid line) with corresponding point-wise approximate confidence intervals as a function of Θ_{12} when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively. In Figure A2 we show the sensitivity with respect to interactions (dotted line) as a function of Θ_{12} when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively.

We observe that under H_S the mean iFDR is 0 when the value of the true interaction is greater than 0.6. When the value of the true interaction is less than 0.6 the average number of true as well as false interaction discoveries is zero resulting in the mean iFDR being undefined. The average number of true interaction discoveries is increasing in the size of the true interaction, and for values greater than 1, the true interaction is discovered more than 80 % of the times. Under H_W the mean iFDR is, in general, greater than 80 % and the average number of true interaction discoveries is increasing in the size of the true interaction, and for values greater than 1.2, the true interaction is discovered more than 80 % of the times Under H_N the mean iFDR is greater than 80 % and the average number of true interaction discoveries is 100 % whenever the size of the true interaction is greater than 0.6.

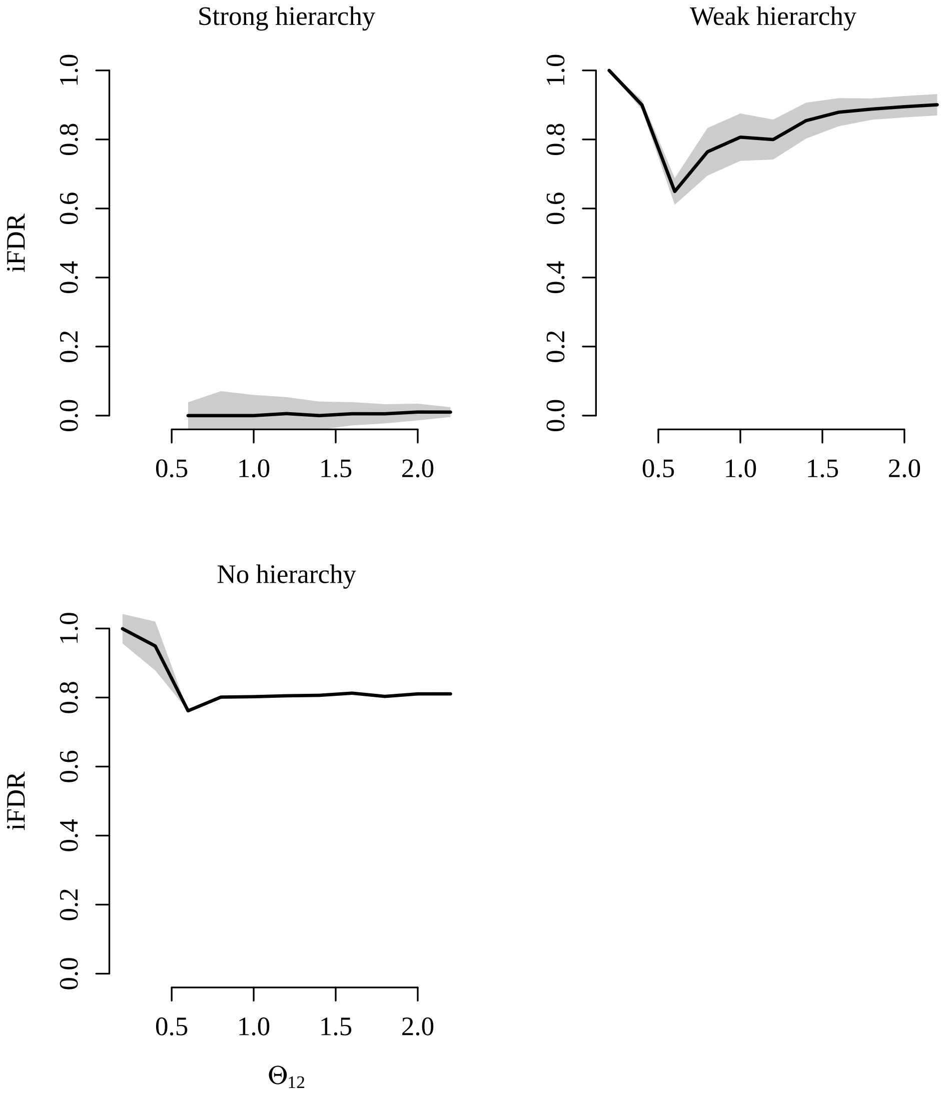


Figure A1: *The data generating process honours the pure interaction set-up. Mean iFDR (solid lines) as a function of* Θ_{12} *when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively.*

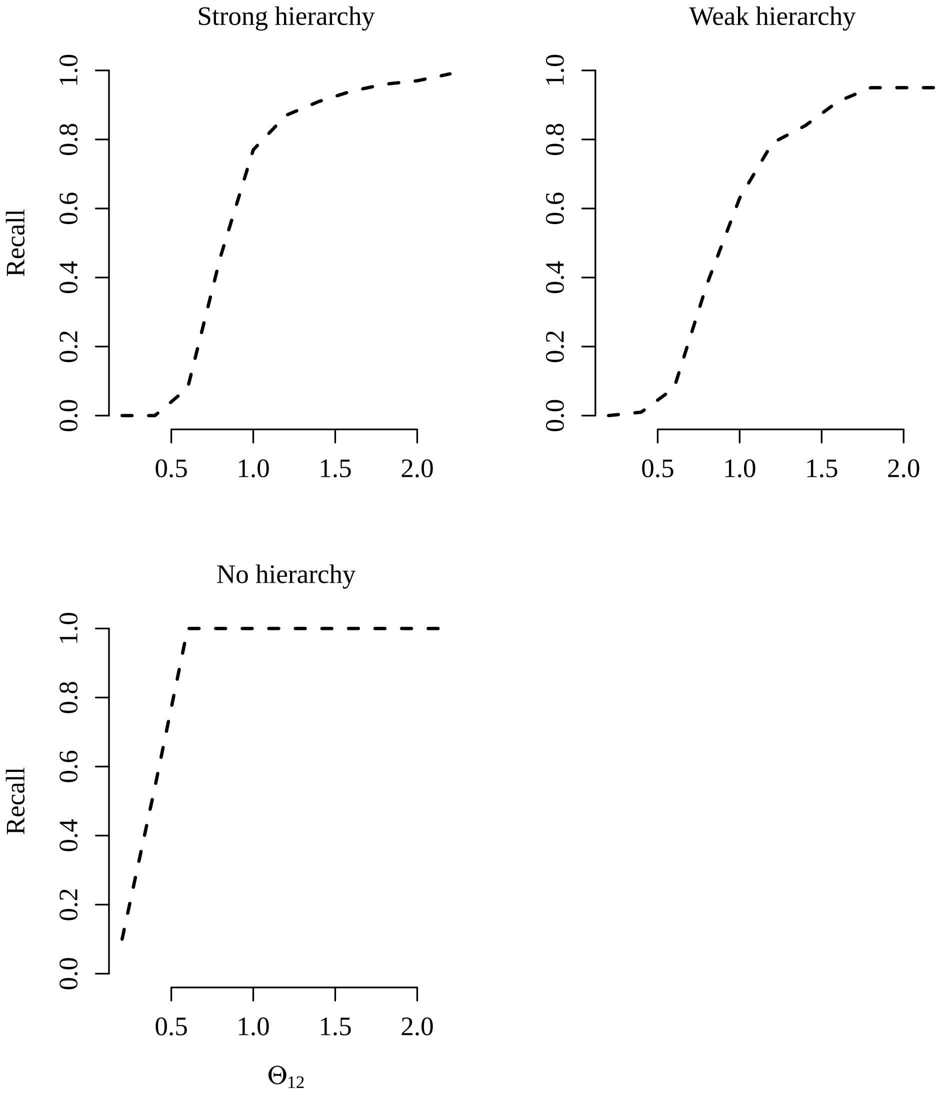


Figure A2: *The data generating process honours the pure interaction set-up. Sensitivity with respect to interactions (dotted lines) as a function of* Θ_{12} *when the model is constrained by strong, weak, or no hierarchy, corresponding to the three plots respectively.*

B Tables

		Glucose			Insulin_75			EnhancedPause		
SNP	Chr	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$
rs3697012	1								●	
rs3687812	1								●	
rs6253968	1							●	●	
rs3655881	1				●	●				
rs6282096	1				●	●				
rs13475806	1	●	●							
rs3685569	1							●	●	
rs3677272	1	●	●							
mCV23641317	1	●	●							
rs3675140	1							●	●	
rs13475988	1	●	●							
rs3696200	1				●	●				
rs13476107	1				●	●				
rs3718090	1							●	●	
rs13476125	1	●	●							
rs13476152	1				●	●				
rs13476237	1				●	●				
UT_1_176.817447	1				●	●				
rs8242509	1				●	●				
rs3680295	1								●	
mCV24786469	1				●	●				
rs3715385	2				●	●				
rs6305730	2				●	●				
rs13476557	2				●	●				
rs13476570	2	●	●							
rs13476681	2				●	●				
rs13476682	2				●	●				
rs6323360	2							●	●	
rs13476687	2							●	●	
rs3697020	2				●	●				
gnf02.126.027	2				●	●				
rs3696248	2				●	●				
rs13476895	2				●	●				
rs13476925	2				●	●				
rs3022956	3								●	
rs3672601	3				●	●				
rs3700657	3							●	●	
rs13477175	3				●	●				
rs3699288	3				●	●				
rs3693395	3				●	●				
CEL-3_74975193	**3**				●	●	●			
UT_3_93.299511	3							●	●	
rs6224522	3							●	●	
rs3720007	3				●	●				
rs13477276	3				●	●				
rs13477280	3							●	●	

		`Glucose`			`Insulin_75`			`EnhancedPause`		
SNP	Chr	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$
rs3675845	3				●	●		●	●	
rs3726226	3				●	●				
rs13477334	3				●	●				
mCV23446189	3				●	●				
rs13477469	3							●	●	
rs13477532	4				●	●				
rs13477543	4	●	●							
rs13477643	4							●	●	
rs3699688	4				●	●				
rs3022985	4				●	●				
rs3708061	**4**							●	●	●
rs3686220	4				●	●				
rs3659791	4				●	●				
rs13477861	4				●	●				
rs13477862	4				●	●				
rs13477863	4				●	●				
rs6399039	4				●	●				
rs13478014	4				●	●				
rs13478069	4				●	●				
CEL-5_3149134	**5**	●	●	●						
rs3666313	5				●	●				
rs6234642	5				●	●			●	
rs3724053	5				●	●				
rs13478143	5				●	●				
rs4225096	5				●	●				
CEL-5_24211033	5								●	
rs13478157	5							●	●	
rs3700261	5							●	●	
rs3680434	5							●	●	
rs3723011	5	●	●							
rs3680521	5	●	●							
UT_5_94.545323	5	●	●							
CEL-5_93652588	5	●	●							
rs13478430	5	●	●							
rs3710018	5							●	●	
rs8265855	5				●	●				
rs3700374	5							●	●	
rs13478567	5							●	●	
rs3659848	5				●	●				
mCV25009162	5							●	●	
CEL-6_56528034	6							●	●	
rs6299256	6							●	●	
rs6334723	6	●	●							
rs13479058	6				●	●				
gnf06.139.257	6							●	●	
rs3694031	7				●	●				
rs6226457	7	●	●							
rs6252200	7		●							
rs6295036	7	●	●							
rs13479253	7		●							

		Glucose			Insulin_75			EnhancedPause		
SNP	Chr	$\mathcal{M}_{\mathrm{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\mathrm{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\mathrm{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\mathrm{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\mathrm{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\mathrm{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\mathrm{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\mathrm{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\mathrm{HS}}^{(\lambda_2)}$
rs3719301	7							●	●	
rs6405142	7							●	●	
rs3676254	7				●	●				
mCV25303361	7				●	●				
rs13479411	7				●	●				
rs6224196	7				●	●				
rs13479450	7		●							
rs3673653	7		●							
rs3711721	7		●							
mCV23009137	7				●	●				
rs3719258	7				●	●				
rs3686436	7				●	●				
rs3723092	8				●	●				
rs6360601	8				●	●				
rs13479622	8							●	●	
gnf08.031.589	8				●	●				
rs13479690	8				●	●				
CEL-8_45424252	8				●	●				
rs3667475	8								●	
rs3691327	8				●	●				
rs13480026	8							●	●	
UT_9_35.918713	9	●	●							
rs6289553	9				●	●				
rs3723670	9							●	●	
rs3699230	9							●	●	
rs3673457	9								●	
rs13480385	9				●	●				
rs13480387	9		●							
rs3665498	9				●	●				
rs3720706	9				●	●				
rs3658244	9				●	●				
rs13480467	10				●	●				
rs3659676	10							●	●	
rs3706825	10							●	●	
rs3672342	10				●	●				
rs13480797	10				●	●				
rs3676616	10		●							
rs3724116	11							●	●	
rs3703198	11							●	●	
rs3715495	11							●	●	
rs6379880	11	●	●							
gnf11.117.476	11				●	●				
rs6354903	11				●	●				
rs6349487	11				●	●				
rs3661724	11							●	●	
rs3678982	11							●	●	
CEL-11_121219118	11							●	●	
rs3693796	11				●	●				
rs3709002	12	●	●							
rs6209157	12	●	●							

		Glucose			Insulin_75			EnhancedPause		
SNP	Chr	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$
rs13481358	12				●	●				
rs13481455	12				●	●				
rs13481522	12								●	
rs13481525	12				●	●				
rs13481527	12				●	●				
rs3687032	12				●	●				
rs13481537	12		●							
rs3662628	12				●	●				
gnf12.084.566	12				●	●				
rs3719282	12				●	●				
rs3716095	12				●	●				
rs13481574	12				●	●				
rs3670410	12				●	●				
rs6176416	12				●	●				
rs13481656	12				●	●				
rs4229602	12				●	●				
rs13481895	13	●	●							
rs6212230	13							●	●	
rs3688040	13							●	●	
rs13481902	13							●	●	
gnf13.082.462	13							●	●	
rs6175633	14				●	●				
rs3683585	14				●	●				
rs13482239	14	●	●							
UT_14_62.125092	14								●	
rs4230463	14								●	
rs3697682	14								●	
rs6351929	14							●	●	
rs3718611	14							●	●	
rs13482396	14				●	●				
rs13459176	15		●							
rs13459177	15		●							
CEL-15_9687257	15		●							
rs13482427	15		●							
rs13482498	15				●	●				
CEL-15_28692470	15				●	●				
CEL-15_28995144	15				●	●				
rs3702361	15				●	●				
gnf15.062.808	15				●	●				
mCV22701834	15				●	●				
rs13482712	15				●	●				
rs3657663	**15**	●	●	●						
rs13482752	15				●	●				
rs4164782	**16**							●	●	●
rs4164966	**16**							●	●	●
rs4165029	16							●	●	
rs4172689	**16**				●	●	●			
rs4177963	16	●	●							
rs4187596	16				●	●				
rs4199362	16							●	●	

		Glucose			Insulin_75			EnhancedPause		
SNP	Chr	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_1)}$	$\mathcal{M}_{\text{HS}}^{(\lambda_2)}$	$\mathcal{I}_{\text{HS}}^{(\lambda_2)}$
rs4204016	16				●	●				
rs3680665	16				●	●				
rs6409750	17		●							
rs6314185	17				●	●				
rs13483026	17							●	●	
UT_17_64.236785	17	●	●							
rs3664306	17		●							
rs3717212	17							●	●	
CEL-17_92336158	17	●	●							
CEL-18_5565618	18								●	
rs13483200	18				●	●				
rs13483236	18							●	●	
mCV22331476	18				●	●				
rs3677707	18							●	●	
rs13483394	18	●	●							
rs3704084	18							●	●	
rs13483605	19				●	●				
rs3676974	19							●	●	

Table B1: *Summary table listing all significant SNPs detected by the two-step procedure for the three traits* `Glucose`*,* `Insulin_75`*, and* `EnhancedPause`*. Columns list the three traits, and, for each trait, SNPs which are included in the sets* $\mathcal{M}_{\text{HS}}^{(\lambda_1)}$, $\mathcal{M}_{\text{HS}}^{(\lambda_2)}$, $\mathcal{I}_{\text{HS}}^{(\lambda_2)}$ *are indicated by bullets. Rows listed in bold indicate the SNPs between which an interaction is found.*

Bibliography

Aiken, L. S. and West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications, Inc., 1 edition.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2010). Hierarchical selection of variables in sparse high-dimensional regression. *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D Brown*, 6:56–69.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bien, J. and Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.6.

Buchardt, A.-S. (2021). iLasso: Identifying interactions via hierarchical lasso regularisation. `https://github.com/abuchardt/ilasso`.

Cornelis, M. C., Tchetgen, E. J. T., Liang, L., Qi, L., Chatterjee, N., Hu, F. B., and Kraft, P. (2012). Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology*, 175(3):191–202.

Cox, D. R. (1984). Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1–24.

Crawford, L., Wood, K. C., Zhou, X., and Mukherjee, S. (2018). Bayesian approximate kernel regression with variable selection. *Journal of the American Statistical Association*, 113:1710–1721.

Dai, J. Y., Kooperberg, C., Leblanc, M., and Prentice, R. L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, 99(4):929–944.

Dudbridge, F. and Fletcher, O. (2014). Gene-environment dependence creates spurious gene-environment interaction. *American Journal of Human Genetics*, 95(3):301–307.

Ekstrøm, C. T. and Sørensen, H. (2014). *Introduction to Statistical Data Analysis for the Life Sciences*. CRC Press, 2 edition.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. Chapman and Hall/CRC Press.

Hu, J. K., Wang, X., and Wang, P. (2014). Testing gene-gene interactions in genome wide association studies. *Genetic Epidemiology*, 38(2):123–134.

Johnson, N., Dudbridge, F., Orr, N., Gibson, L., Jones, M. E., Schoemaker, M. J., Folkerd, E. J., Haynes, B. P., Hopper, J. L., Southey, M. C., et al. (2014). Genetic variation at cyp3a is associated with age at menarche and breast cancer risk: a case-control study. *Breast Cancer Research*, 16(3):R51.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–412.

Leekha, S., Terrell, C. L., and Edson, R. S. (2011). General principles of antimicrobial therapy. *Mayo Clinic Proceedings*, 86(2):156–167.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Meier, L., Geer, S. V. D., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Nickels, S., Truong, T., Hein, R., Stevens, K., Buck, K., Behrens, S., Eilber, U., Schmidt, M., Häberle, L., Vrieling, A., et al. (2013). Evidence of gene–environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLOS Genetics*, 9(3):1–14.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58(1):267–288.

Valdar, W., Solberg, L., Gauguier, D., Heyes, S., Klenerman, P., Cookson, W., Taylor, M., Rawlins, J., Mott, R., and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38:879–87.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*, 34(3):275–285.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751.

Manuscript II

Joint Regression Analysis of Multiple Traits Based on Genetic Relationships

Ann-Sophie Buchardt, Xiang Zhou and Claus Thorn Ekstrøm

Joint Regression Analysis of Multiple Traits Based on Genetic Relationships

ANN-SOPHIE BUCHARDT[1*], XIANG ZHOU[2], CLAUS THORN EKSTRØM[1]

[1]*Section of Biostatistics, Department of Public Health, University of Copenhagen,*
Øster Farimagsgade 5, 1014 København K, Denmark

Corresponding author: *asbu@sund.ku.dk

[2]*Department of Biostatistics, School of Public Health, University of Michigan,*
1415 Washington Heights, Ann Arbor, Michigan 48109

Abstract

Polygenic scores (PGSs) are widely available and employed in genomic data analyses for predicting and understanding genetic architectures. We propose a novel clustering and estimation method using PGSs for inferring a genetic relationship among multiple, simultaneously measured and potentially correlated traits in a multivariate GWAS.

Using graphical lasso, we estimate a sparse covariance matrix of the PGSs and obtain clusters of traits sharing genetic characteristics. We use the clusters to specify the structure of the error covariance matrix of a generalised least squares (GLS) model and use the feasible GLS estimator for estimating a linear regression model with a certain unknown degree of correlation between the residuals.

The method suits many biology studies well with traits embedded in some genetic functioning groups and facilitates developement of the PGS research. We compare the method with fully parametric techniques on simulated data and illustrate the utility of the methods by examining a heterogeneous stock mouse data set from the Wellcome Trust Centre for Human Genetics. We demonstrate that the method successfully identifies clusters of traits and increases precision, power and computational efficiency.

Keywords: correlated traits, joint analysis; multiple traits; multivariate GWAS; polygenic scores

1 Introduction

The motivation for this paper is the problem of jointly analysing multiple, simultaneously measured quantitative traits in a genome-wide association study (GWAS), i.e., a multivariate GWAS. Multivariate GWASs have gained attention in genetic studies as it offers several advantages over analysing each trait in a separate GWAS (Galesloot et al., 2014). While single-trait GWASs have found numerous genetic variants associated with complex diseases (MacArthur et al., 2016), traits are often correlated, for example due to pleiotropic genes and, therefore, a joint modelling approach can be used to increase precision and power (Schmitz et al., 1998). Furthermore, these variants typically have a small effect and correspond to a small fraction of truly associated variants, therefore they have limited predictive power (Yang et al., 2010; Dudbridge, 2013).

While a univariate GWAS cannot exploit potential correlation between traits, carrying out a multivariate genome-wide association (GWA) analysis may vastly increase the statistical and computational complexity of the analyses. For example, an unconstrained correlation matrix for q traits requires $q(q-1)/2$ correlation parameters. Such large numbers of parameters require very large data sets and present considerable computational challenges.

Different multivariate GWA methods have been formulated for simultaneously testing marker associations with multiple traits influenced by pleiotropic genes (Zhou and Stephens, 2012), (Kang et al., 2008), and (Zhou and Stephens, 2014). In the case of correlated traits, the OLS estimator is unbiased, but inefficient. Instead we may use the *feasible generalised least squares (FGLS)* estimator. The feasible estimator is, provided the covariance matrix is consistently estimated, asymptotically more efficient for large samples under heteroskedasticity. A further challenge is that the total number p of features (e.g., SNPs) is usually larger than the sample size N when working with genomic data. In this case, the design matrix does not have full rank and the OLS estimator cannot be computed. Various methodologies deal with this problem as well as how to generate the weights of the SNPs and how to determine which $s \leq p$ features should be included (e.g., Euesden et al. (2014); Wray et al. (2014)). To meet these challenges we introduce *polygenic scores (PGSs)*. Our goal is to use PGSs to both perform dimensionality reduction and identify clusters of traits to approximate the *structure* of the error covariance matrix of a linear regression model. We can then use the FGLS estimator for estimating the unknown parameters.

PGSs are linear transformations of multiple genetic variants to scores that *summarise* the estimated effect of, e.g., SNPs. Often they are calculated as a weighted sum of SNPs, that is, they are constructed from the "weights" derived from a GWAS, or from some form of machine learning algorithm. This way, a PGS reflects an estimated genetic predisposition for a given trait without taking environmental factors into account and can be used as a numeric predictor for that trait.

The objective of this paper is to propose a computationally efficient method for inferring a genetic relationship among a large number of simultaneously measured and potentially correlated traits in a multivariate GWAS. The method simultaneously analyses multiple correlated quantitative traits in clusters that share some genetic component under the assumption that the trait values in a cluster follow a multivariate normal distribution.

We propose a versatile tool called a joint analysis of multiple traits based on genetic relationships (*geneJAM*), which implements the method of finding clusters of correlated outcome components (traits) that share some genetic component by means of PGSs. The PGS method (Dudbridge, 2013) aggregates the effects of variants across the genome to estimate heritability, infer genetic overlap between traits, and predict phenotypes based on the genetic profile. The approach is motivated by utilising the widely available PGSs as numeric predictors to identify conditional independences of traits giving rise to clusters of traits, and, thereby, being able to analyse the data combined in clusters and increase precision and power and gain computational advantages. By estimating a sparse version of the precision matrix of the PGSs, approximate zeroes induce conditional independence which enables us to infer clusters of traits sharing genetic characteristics. Having identified the clusters, we are able to jointly analyse the traits, thereby increasing precision and power. We propose a simple approach: having estimated the precision matrix of the PGSs, we are able to approximate clusters of traits that share some genetic component. Thus, we are able to specify the structure of the error covariance matrix of a linear regression model when there is a certain unknown degree of correlation between the residuals and use the feasible generalised least squares (FGLS) estimator for estimating the unknown parameters.

In Section 2 we introduce the conceptual framework of the geneJAM method and the theoretical foundation including linear regression models, the concept of PGSs, conditional independence, FGLS, and sparsity of correlation matrices.

We study our method and other techniques on real data, as well as simulated data, in Section 3. Specifically, we analyse a heterogeneous stock mouse data set (Valdar et al., 2006), from the Wellcome Trust Centre for Human Genetics. The methods are implemented using the R packages `MESS` by Ekstrøm (2019) and `glasso` by Friedman et al. (2018); the R code is part of our package `geneJAM` which is available online, see (Buchardt, 2022). We conclude with a discussion and recommendations in Section 4.

2 Method

We focus on linear regression models, which are suitable when the outcome is quantitative, and ideally when the error distribution is Gaussian. We consider N independent samples,

$\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^q$, of a q-dimensional random variable distributed according to a multivariate normal distribution. We define a matrix, $\mathbf{Y}$, from the N observations of the q outcome components, $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^\top \in \mathbb{R}^{N \times q}$. We assume that the data generating process can be described by a linear regression model of the form

$$\mathbf{Y} = \mathbf{M} + \mathbf{X}\mathbf{C} + \mathbf{E}, \tag{1}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times q}$ is a matrix of N samples of p features, $\mathbf{M} = \mathbf{1}_N \boldsymbol{\mu}^\top \in \mathbb{R}^{N \times q}$ is a matrix of intercepts with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^\top$, $\mathbf{C} \in \mathbb{R}^{p \times q}$ is a matrix of regression coefficients, and $\mathbf{E} \in \mathbb{R}^{N \times q}$ are independent Gaussian random errors.

In the following we assume that a set of PGSs has been provided – one for each outcome – and we use this set of estimated PGSs as numeric predictors (features). That is, we assume that we have access to N samples of exactly one feature per trait, $\mathbf{z}_1, \ldots, \mathbf{z}_N \in \mathbb{R}^q$.

We let $\vec{\mathbf{Y}} = \mathrm{vec}(\mathbf{Y}^\top) \in \mathbb{R}^{Nq}$ denote the row-wise vectorisation of $\mathbf{Y}$ obtained by stacking the columns of the matrix $\mathbf{Y}^\top$, that is,

$$\vec{\mathbf{Y}} = (y_{11}, y_{12}, \ldots, y_{1q}, y_{21}, y_{22}, \ldots, y_{2q}, \ldots, y_{N1}, y_{N2}, \ldots, y_{Nq})^\top .$$

Now, we consider a linear regression model on the form

$$\vec{\mathbf{Y}} = \vec{\mathbf{Z}}\vec{\mathbf{B}} + \vec{\mathbf{E}}, \tag{2}$$

where $\vec{\mathbf{B}} \in \mathbb{R}^{2q}$ denotes the row-wise vectorisation of a matrix $\mathbf{B} \in \mathbb{R}^{2 \times q}$ of unknown intercepts and regression coefficients such that $\vec{\mathbf{B}} = (\xi_1, \beta_1, \xi_2, \beta_2, \ldots, \xi_q, \beta_q)^\top$. $\vec{\mathbf{Z}} \in \mathbb{R}^{Nq \times 2q}$ is a corresponding design matrix of features and $\vec{\mathbf{E}} \in \mathbb{R}^{Nq}$ is a vector of Gaussian random errors. To allow for among-trait correlation, we assume that

$$\vec{\mathbf{E}} \sim \mathcal{N}_{Nq}\left(\mathbf{0}_{Nq}, \boldsymbol{\Omega}\right),$$

where $\mathbf{0}_{Nq} \in \mathbb{R}^{Nq}$ is a vector of zeros and $\boldsymbol{\Omega} \in \mathbb{R}^{Nq \times Nq}$ is a general covariance matrix. Possible structures for $\boldsymbol{\Omega}$ may be a block diagonal matrix, a diagonal matrix, or, more importantly, a general positive definite matrix. This flexibility is essential to our method, since the outcome that our method is targeted towards, is supposed to be both correlated and heteroscedastic across outcome components. In particular, we assume that the covariance matrix, $\boldsymbol{\Omega}$, is a block diagonal matrix on the form

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{R} & 0 & \cdots & 0 \\ 0 & \mathbf{R} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{R} \end{bmatrix}, \tag{3}$$

where $\mathbf{R} \in \mathbb{R}^{q \times q}$ are the residual covariances among traits within each individual $i = 1, \ldots, N$. If the covariance matrix is known, the generalised least squares (GLS) estimator is the best linear unbiased estimator of the unknown parameters $\vec{\mathbf{B}}$ in a linear regression model when there is a certain known degree of correlation between the residuals. In these cases, ordinary least squares (OLS) and weighted least squares (WLS) can be statistically inefficient, or even give misleading inferences (Baltagi, 2008). In our case, the among-trait and within-individual correlation, $\boldsymbol{\Omega}$, of the errors is unknown, but we can get a consistent (in terms of structure) estimate using the feasible generalised least squares (FGLS) estimator. In general terms, the FGLS procedure consists of two steps: First, a linear regression model is estimated by OLS, and the residuals are used to build an estimator of the covariance matrix of the errors. Second, using the estimator of the covariance matrix of the errors, the unknown regression coefficients are estimated using WLS, which generalises OLS and allows the covariance matrix of the errors to be different from an identity matrix. If the estimator of the covariance matrix of the errors is a consistent estimator, like the OLS, then the FGLS estimator of the unknown regression coefficients is a consistent estimator (Baltagi, 2008). The FGLS estimator is discussed in more detail in Section 2.1.

Our initial motivation for considering PGSs is the presumption that much of the variability observed in a trait is attributable to genetic differences. Therefore, the traits may be correlated via the PGS, and it may be reasonable to assume that the underlying independence relations for the traits can be approximated by the relations observed in the PGSs. Let us briefly elaborate on the interplay between the structure of the constants $\mathbf{C}$, which reflect the true population coefficients, the expected correlation structure of the PGSs, and the expected independence structure of the among-trait and within-individual correlation. Figure 1 illustrates a few examples of cross-trait associations where circles represent features (SNPs) and squares represent outcome (components). If there are no genetic effects on SNP level (a), $\mathbf{C}$ is a matrix of zeros. In this case there is also no effect on PGS level and there is no (genetically determined) among-trait and within-individual correlation, that is, $\boldsymbol{\Omega}$ is a diagonal matrix and we can use the OLS estimator. If each SNP affects at most one trait (b*), the corresponding PGSs are independent and, again, there is no (genetically determined) among-trait and within-individual correlation. If there is a single cross-trait association where one SNP, say, $\mathbf{x}_1$, affects multiple traits, say, $\mathbf{y}_1$ and $\mathbf{y}_2$, $\mathbf{C}$ is a matrix of zeros except for the entries c_{11} and c_{21}. In this case there is an effect on PGS level. Furthermore, the PGSs $\mathbf{z}_1$ and $\mathbf{z}_2$ are correlated and there is (genetically determined) among-trait and within-individual correlation. That is, $\boldsymbol{\Omega}$ is a block diagonal matrix and we cannot use the OLS estimator. This argument can be further extrapolated for the case of multiple cross-trait associations where multiple SNPs affect multiple traits as exemplified in (d) and (e). In such cases, more PGSs will be correlated and the among-trait and within-individual correlation structure will be more complex. That is, $\boldsymbol{\Omega}$ will be even "further away" from the identity matrix, and a joint analysis of the correlated traits will have an even greater impact in terms of precision and power (Schmitz et al., 1998).

2.1 The geneJAM method

We provide a method which identifies blocks in the correlation matrix of PGSs via the graphical lasso with regularisation parameter ρ and models the relationship between traits and associated PGSs by linear regression. To obtain a better fit, the FGLS estimator is used to provide a consistent estimate of the covariance, $\mathbf{R}$, of the errors with a block-structure identified at a given ρ, and, then, weighted least squares (WLS) is used for estimating the unknown parameters in a linear regression model. This way, we identify clusters of potentially correlated outcome components by means of PGSs, under the assumptions that features which are marginally highly associated with a trait should carry more weight in a joint analysis of traits, and that much of the variability observed in a trait is attributable to genetic differences, i.e., heritability, such that the traits may be correlated via the PGS.

We assume that we have N observations of the multivariate outcome $\mathbf{Y} \in \mathbb{R}^{N \times q}$ with associated PGSs, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_q) \in \mathbb{R}^{N \times q}$, for each trait individually, and that the PGSs are from the same population as the data at hand but estimated from a different sample. If existing PGSs for a trait have not been provided, we refer to Section 3.1.1 for a brief introduction to generating PGSs and to Dudbridge (2013) for a comprehensive survey and a pointer to computing PGSs. We assume that the collection of PGSs can be approximated by a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

If we represent the system of PGSs as an undirected graph $\mathcal{G}$, where the PGS for a certain trait is represented by a node, then an edge represents related pairs of traits in terms of PGSs. Moreover, a *connected component* of a graph, which is a maximal sub-graph in which there exists a path between any two nodes, represents a cluster of related traits. In a probabilistic graphical model, the conditional independence structure is encoded in a graph, since the presence/absence of edges encode conditional independence relations among associated nodes (Lauritzen, 1996).

Identifying connected components of the graph $\mathcal{G}$ is one of the core functionalities of the geneJAM method: A useful property of the multivariate Gaussian distribution is that, for a matrix of random variables following a multivariate Gaussian distribution, the ijth component of the corresponding precision matrix (the matrix inverse of the covariance matrix) is zero if and only if the variables i and j are conditionally independent, given the others. This is an implication of the basic property of the multivariate Gaussian distribution being closed under conditioning (Lauritzen, 1996). We wish to exploit this property: Under the assumption that much of the variability observed in a trait is attributable to genetic differences, i.e., heritability, the traits may be correlated via the PGSs. That is, by assuming that the PGSs follow a multivariate Gaussian distribution, approximating the precision matrix, $\mathbf{P}$, of the PGSs and further sparsity, we are able to identify blocks of zeros in the precision matrix, and, thus, potential connected components, which correspond to clusters of traits sharing some genetic characteristics.

As a way of estimating a sparse precision matrix, we consider the *graphical lasso* (Friedman et al., 2007). The graphical lasso estimates a sparse precision matrix using a lasso penalty, and the regularisation path is computed at a grid of values for the regularisation parameter ρ. Thus, from the sparse precision matrix, $\hat{\mathbf{P}}_{(\rho)}$, of the PGSs estimated by graphical lasso at a given value of ρ, we obtain a clustering, that is, a set of $C_{(\rho)}$ connected components. We assume that the residual covariances among traits within each individual, $\mathbf{R}$, exhibit this clustering as well.

For the first step in the FGLS estimation of $\mathbf{R}$ we define sub-vectors $\mathbf{b}_l = (\xi_l, \beta_l)^\top \in \mathbb{R}^2$, $l = 1, \ldots, q$, of $\vec{\mathbf{B}}$ and design matrices $\mathbf{Z}_l \in \mathbb{R}^{N\times 2}$, where a constant term is included, such that the first column of $\mathbf{Z}_l$ is a column of ones allowing estimation of the intercept while the following column contains the feature (PGS) associated with the corresponding trait value. An OLS estimator of $\mathbf{b}_l$ is calculated from a linear regression model for each trait, $l = 1, \ldots, q$, separately:

$$\mathbf{y}_l = \mathbf{Z}_l \mathbf{b}_l + \mathbf{e}_l,$$

where $\mathbf{e}_l \in \mathbb{R}^N$, $l = 1, \ldots, q$, are independent Gaussian random errors $\mathbf{e}_l \sim \mathcal{N}_N(\mathbf{0}_N, \sigma_l^2 \mathbf{I}_{N\times N})$, with $\sigma_l^2 > 0$. The OLS estimator of $\mathbf{b}_l$, $l = 1, \ldots, q$, is

$$\hat{\mathbf{b}}_l^{\mathrm{O}} = \left(\mathbf{Z}_l^\top \mathbf{Z}_l\right)^{-1} \mathbf{Z}_l^\top \mathbf{y}_l.$$

From these estimates the corresponding estimated residuals,

$$\hat{\mathbf{u}}_l = \left(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}^{\mathrm{O}}\right)_l,$$

for all $l = 1, \ldots, q$, are obtained. Here, $\mathbf{Z} \in \mathbb{R}^{N\times(q+1)}$ is a corresponding design matrix of features, where a column of ones is included to allow for estimation of the intercept. Now, for each value of the regularisation parameter ρ, we construct an estimate, $\hat{\mathbf{R}}^{\mathrm{O}}_{(\rho)}$, of the covariance of the errors by computing the covariance of the OLS estimated residuals, $\hat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_q) \in \mathbb{R}^{N\times q}$, in the same connected component. That is,

$$\hat{\mathbf{R}}^{\mathrm{O}}_{g(\rho)} = \operatorname{cov}\left(\hat{\mathbf{U}}_{g(\rho)}\right),$$

where $\hat{\mathbf{R}}^{\mathrm{O}}_{g(\rho)}$ and $\hat{\mathbf{U}}_{g(\rho)}$ are the sub-matrices and of $\hat{\mathbf{R}}^{\mathrm{O}}_{(\rho)}$ and $\hat{\mathbf{U}}$, respectively, corresponding to the connected components $g(\rho) = 1, \ldots, C_{(\rho)}$ at a given value of ρ. From $\hat{\mathbf{R}}^{\mathrm{O}}_{(\rho)}$ we construct an estimate, $\hat{\mathbf{\Omega}}^{\mathrm{O}}_{(\rho)}$, of $\mathbf{\Omega}$ on the form (3). Step one of the FGLS procedure is then fulfilled.

In the second step in the FGLS estimation, we build the FGLS estimator, $\hat{\vec{\mathbf{B}}}^{\mathrm{F}}_{(\rho)} \in \mathbb{R}^{2q}$, using WLS:

$$\hat{\vec{\mathbf{B}}}^{\mathrm{F}}_{(\rho)} = \left(\vec{\mathbf{Z}}^\top \left(\hat{\mathbf{\Omega}}^{\mathrm{O}}_{(\rho)}\right)^{-1} \vec{\mathbf{Z}}\right)^{-1} \vec{\mathbf{Z}}^\top \left(\hat{\mathbf{\Omega}}^{\mathrm{O}}_{(\rho)}\right)^{-1} \vec{\mathbf{Y}}.$$

The procedure is iterated with the first iteration given by

$$\begin{aligned}
\hat{\mathbf{U}}^{\mathrm{F}(1)} &= \left(\mathbf{Y} - \hat{\mathbf{Y}}_{(\rho)}\right) \\
\hat{\mathbf{R}}^{\mathrm{F}(1)}_{g(\rho)} &= \operatorname{cov}\left(\hat{\mathbf{U}}^{\mathrm{F}(1)}_{g(\rho)}\right) \\
\hat{\mathbf{\Omega}}^{\mathrm{F}(1)}_{(\rho)} &= \operatorname{diag}\left(\hat{\mathbf{R}}^{\mathrm{F}(1)}_{1}, \ldots, \hat{\mathbf{R}}^{\mathrm{F}(1)}_{C(\rho)}\right) \\
\hat{\vec{\mathbf{B}}}^{\mathrm{F}(2)}_{(\rho)} &= \left(\vec{\mathbf{Z}}^{\top}\left(\hat{\mathbf{\Omega}}^{\mathrm{F}(1)}_{(\rho)}\right)^{-1}\vec{\mathbf{Z}}\right)^{-1}\vec{\mathbf{Z}}^{\top}\left(\hat{\mathbf{\Omega}}^{\mathrm{F}(1)}_{(\rho)}\right)^{-1}\vec{\mathbf{Y}}
\end{aligned}$$

where $\hat{\mathbf{Y}}_{(\rho)} \in \mathbb{R}^{N\times q}$ is a by-row matrix transformation of

$$\hat{\vec{\mathbf{Y}}}_{(\rho)} = \vec{\mathbf{Z}}\hat{\vec{\mathbf{B}}}^{\mathrm{F}}_{(\rho)}.$$

This estimation of $\mathbf{\Omega}$ is iterated to convergence and we obtain estimates $\hat{\mathbf{\Omega}}_{(\rho)}$ and $\hat{\vec{\mathbf{B}}}_{(\rho)}$. The standard error, $\mathrm{SE}_{(\rho)}$, of the estimated coefficients are given by

$$\mathrm{SE}_{(\rho)} = \sqrt{\operatorname{diag}\left(\left(\vec{\mathbf{Z}}^{\top}\left(\hat{\mathbf{\Omega}}_{(\rho)}\right)^{-1}\vec{\mathbf{Z}}\right)^{-1}\right)}.$$

As we discuss below, choosing the value of the regularisation parameter ρ is an important issue in practice as it controls the amount of regularisation of the precision matrix, that is, the conditional independence structure of the PGSs.

An algorithmic overview of the method is shown in Appendix A of the Supplementary Materials.

2.2 Tuning

Tuning of the regularisation parameter ρ is crucial since a large value of ρ for the graphical lasso will make all variables conditionally independent while a small value of ρ will keep most variables in one cluster. We choose the regularisation parameter for the purposes of increasing the precision of the estimates, and note that it may be preferable to err on the side of complexity and choose less sparsity to obtain larger clusters, and thus more complex models, in order to further increase power and not fail to detect clusters in the data.

Since we aim at optimising the precision, we use the standard error of the estimated coefficients to tune the geneJAM method. More specifically, for a given value of the regularisation parameter, we compute the mean of the standard errors of estimates for traits which have been clustered. Traits which, at a certain level of regularisation, are left as singletons stay as singletons as the regularisation increases, and, since there is no regularisation on singletons, they are modelled using a simple linear regression estimated by OLS. Hence, traits belonging to clusters of size one are irrelevant for the tuning. Therefore, for a given value, ρ, of the regularisation parameter, we compute the standard error $\mathrm{SE}_{(\rho)} \in \mathbb{R}^q$ and, for all traits which are in clusters of size strictly larger than one, we compute the average standard error (SE), that is,

$$\bar{\mathrm{SE}}_{(\rho)} = \frac{1}{\left|\mathcal{G}_{(\rho)}\right|} \sum_{l \in \mathcal{G}_{(\rho)}} \mathrm{SE}_{(\rho) l},$$

where $\mathcal{G}_{(\rho)}$ is the index set of traits in connected components of size strictly larger than one and $\left|\mathcal{G}_{(\rho)}\right|$ is the cardinality of $\mathcal{G}_{(\rho)}$. We use the smallest value of $\bar{\mathrm{SE}}_{(\rho)}$, over all tried values of ρ, as our selection criterion for the regularisation.

3 Results

We illustrate the utility of the geneJAM method presented in Section 2.1 on both simulated and real data. The motivation for both sets of examples is to understand the performance of the method on two important problems in statistical genetics: simultaneously measured traits and cross-trait associations.

First, we use simulations to assess the ability to cluster traits under different assumptions of cross-trait association on the data generating process and the method, see Figure 1 for different scenarios. Most complex traits are highly polygenic; they are influenced by a large number of genetic variants with moderate effects, rather than a handful of variants with large effects (Price et al., 2015). However, from the central limit theorem we know that if the PGS is based on a sum of, say, p independent features, $\mathbf{x}_1, \ldots, \mathbf{x}_p$, e.g., SNPs, with identical distributions, then the PGS of a sample approximates the Gaussian distribution. Thus, instead of simulating a large number of genetic variants with moderate effects, we simulate a few with a larger effect; corresponding to an additive aggregation, which preserves information about the mean and variance and conforms to the Gaussian distribution. The effect sizes and error distribution are chosen in each case such that the heritability, for $l = 1, \ldots, q$,

$$H_l^2 = \frac{\mathrm{V}\left[\mathbf{y}_l\right] - \mathrm{V}\left[\mathbf{e}_l\right]}{\mathrm{V}\left[\mathbf{y}_l\right]} = \frac{\sum_{j=1}^{p} \mathrm{V}\left[\mathbf{x}_j\right] c_{jl}^2}{\sum_{j=1}^{p} \mathrm{V}\left[\mathbf{x}_j\right] c_{jl}^2 + \mathrm{V}\left[\mathbf{e}_l\right]},$$

are tried in the range 0.2–0.6. This corresponds to assuming that the genetic effects explain 20–60% of the total (phenotypic) variance in the outcome, or, equivalently, that the broad-sense heritability of a trait is known to be 0.2–0.6. In particular, the noise is simulated from the standard Gaussian distribution, and the SNPs are simulated from the binomial distribution with the number of trials, n, equal to two, and success probability π. This way, in the simple case of a single locus with two alleles denoted A and a with frequencies $f(A) = \pi$ and $f(a) = 1 - \pi$, respectively, the expected genotype frequencies under random mating are $f(AA) = \pi^2$ for the AA homozygotes, $f(aa) = (1-\pi)^2$ for the aa homozygotes, and $f(Aa) = 2\pi(1-\pi)$ for the heterozygotes. These frequencies define the Hardy-Weinberg equilibrium with a minor allele frequency drawn from the continuous uniform distribution on the interval $[0.1, 0.5]$.

In order for us to assess the clustering ability of the method we use the Rand index (RI) (Rand, 1971). The Rand index measures the similarity between two clusterings. The Rand index represents the frequency of occurrence of agreements over the total pairs of traits, or the probability that two clusterings to be compared will agree on a randomly chosen pair. We apply the Rand index to the adjacency matrix $\mathbf{A}$, which is a symmetric square matrix with zeros on the diagonal representing a finite undirected graph. The elements of the matrix represent pairs of edges which are adjacent (represented by a '1') or not (represented by a '0') in the graph. We generate the adjacency matrix from the estimated sparse precision matrix, $\hat{\mathbf{P}}_{ij}$ by letting $A_{ij} = 1$ if $\hat{\mathbf{P}}_{ij} > 0$, $i \neq j$, and zero otherwise.

We also assess the modelling performance of the method in terms of precision and computing time. Here, the goal is to show that by using the geneJAM to cluster and estimate we improve the precision of the estimates compared to using simple linear regressions or mixed models, and, given a clustering of the traits, estimation using FGLS is faster than the mixed model estimation implemented in R in the `lme4` package. Finally, we compute and compare the root mean squared error (RMSE) as a general purpose error metric for the numerical predictions:

$$\text{RMSE}_l = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{il} - \hat{y}_{il})^2}, \quad l = 1, \ldots, q.$$

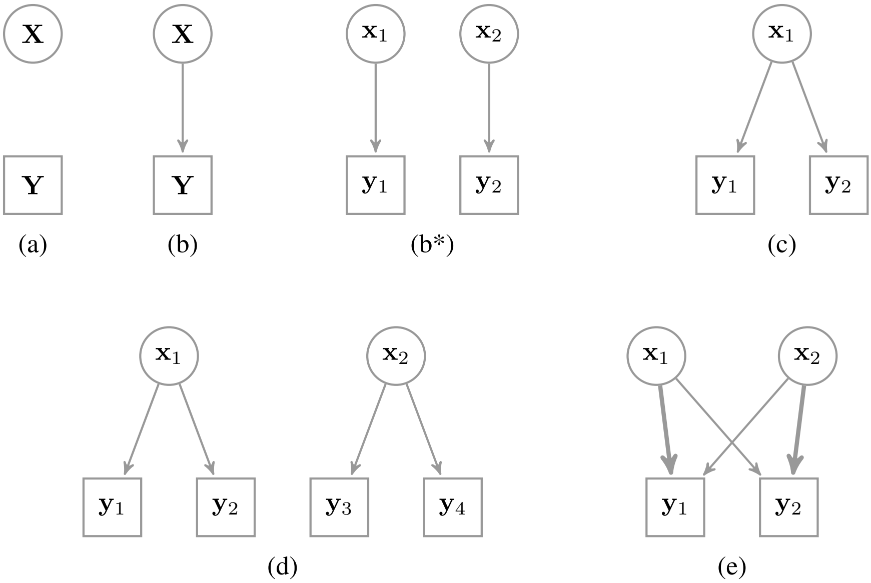
We assess the clustering and precision of our method in a heterogeneous stock mouse data set (Valdar et al., 2006), from the WTCCC.

3.1 Simulated data

Since the efficacy of a procedure depends on the true model generating the data we simulate five different set-ups, see Figure 1, such that different scenarios are tried as the ground truth, and we apply our method to each scenario separately.

3.1.1 Sampling procedure

The sampling is a two-step procedure which goes as follows: first, we generate a sample with $N = 1000$ observations of $p = 500$ features, $\boldsymbol{X}^{(0)} = \left(\mathbf{x}_1^{(0)}, \ldots, \mathbf{x}_p^{(0)}\right) \in \{0, 1, 2\}^{N \times p}$, in



(a) (b) (b*) (c)

(d) (e)

Figure 1: *Graphical illustration of different cross-trait associations.* (a), (b), *and* (b*) *are extreme cases where there are no genetic effects on SNP level (a), where all SNPs affect all traits equally (b), and where SNPs at most affects one trait (b*).* (c) *is a case of single cross-trait association where one SNP affects multiple traits.* (d) *is a case of multiple cross-trait association where multiple SNPs affect distinct clusters of traits. Finally,* (e) *is a case of overlapping cross-trait association where multiple SNPs affect the same cluster of traits.*

the Hardy–Weinberg equilibrium with a minor allele frequency of 0.3, as described above. From the features we generate observations, $\boldsymbol{Y}^{(0)} = \left(\mathbf{y}_1^{(0)}, \ldots, \mathbf{y}_q^{(0)}\right) \in \mathbb{R}^{N \times q}$, by $\boldsymbol{Y}^{(0)} = \boldsymbol{X}^{(0)}\mathbf{C} + \boldsymbol{E}^{(0)}$, where $\mathbf{C} \in \mathbb{R}^{p \times q}$ is a matrix of regression coefficients and $\boldsymbol{E}^{(0)} = \left(\mathbf{e}_1^{(0)}, \ldots, \mathbf{e}_q^{(0)}\right) \in \mathbb{R}^{N \times q}$ is a Gaussian noise matrix. We create a heteroscedastic problem by allowing different variances for the noise across the traits. We create a sparse problem by letting $c_{jl} = 0$ for all $j = 1, \ldots, p$ and $l = 1, \ldots, q$, except for a few entries, corresponding to different cross-trait association scenarios, for which different values are tried; more details are given in the subsequent sections. This way, the method is examined at different levels of broad-sense heritability. From this initial sample, $\left(\mathbf{Y}^{(0)}, \mathbf{X}^{(0)}\right)$, PGSs, $\mathbf{Z}$, are generated as follows: we estimate a univariate simple linear regression model of the form

$$\mathbf{y}_l^{(0)} = \mathbf{x}_j^{(0)} w_{jl} + \mathbf{e}_l^{(0)},$$

where the scalars $w_{jl} \in \mathbb{R}, j = 1, \ldots, p, l = 1, \ldots, q$, are regression coefficients and $\mathbf{e}_l^{(0)} \in \mathbb{R}^N, l = 1, \ldots, q$ are vectors of independent Gaussian random errors. For ease of notation and implementation, the estimated regression coefficients, $\hat{w}_{jl}, j = 1, \ldots, p, l = 1, \ldots, q$, are stored in a $p \times q$ matrix $\hat{\mathbf{W}}$, which is not to be mistaken for a matrix of coefficients from a multivariate multiple linear regression. Next, we simulate a new sample of features,

$\mathbf{X} \in \{0,1,2\}^{N\times p}$, in the Hardy–Weinberg equilibrium with a minor allele frequency of 0.3, and from these features we generate new PGSs by $\mathbf{Z} = \boldsymbol{X}\hat{\mathbf{W}}$. Finally, we simulate new observations by $\boldsymbol{Y} = \boldsymbol{X}\mathbf{C} + \boldsymbol{E}$, where $\boldsymbol{E} \in \mathbb{R}^{N\times q}$ is a Gaussian noise matrix. This way, we obtain traits and PGSs as if the PGSs were provided from an unrelated GWAS. We run the geneJAM procedure using the `geneJAM` function available online, see (Buchardt, 2022), and we repeat the sampling of the outcome as well as the running of geneJAM 100 times. Thus, in total the method is evaluated on about 50000 simulations.

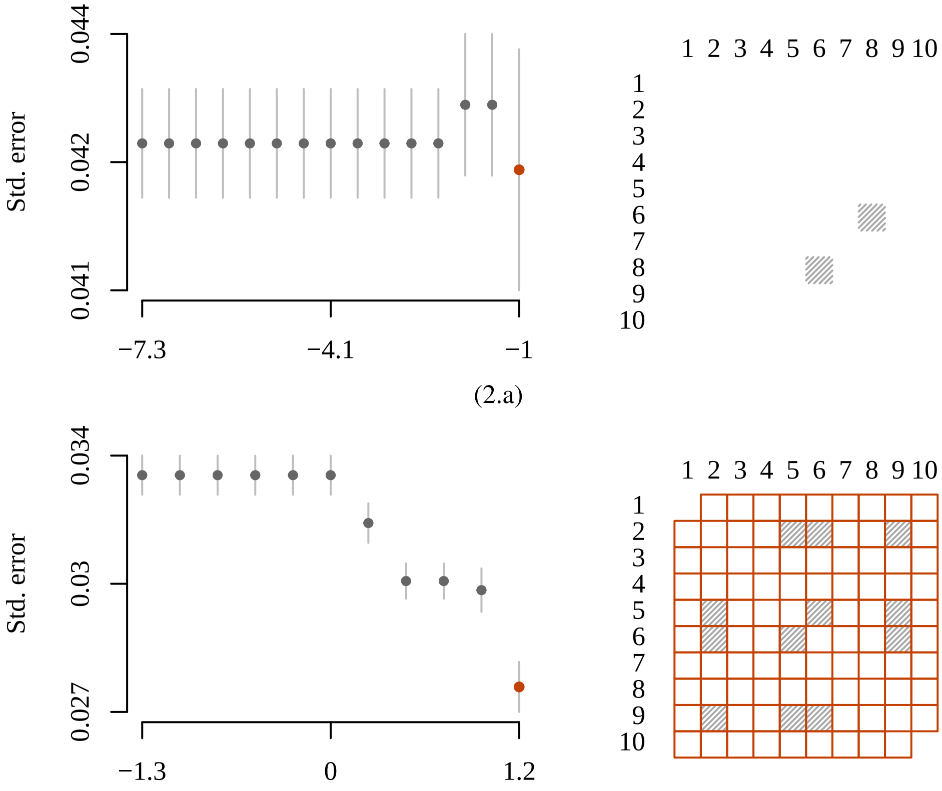
3.1.2 Precision and clustering assessment

To assess the clustering ability of geneJAM, we apply the method when the data generating process honours the different cross-trait associations on SNP level, exemplified by genetic correlations as illustrated in Figure 1. Thorough presentations are found in the supplementary material; here, we summarise our findings.

We consider simulations of i.i.d. features $\boldsymbol{X} \in \{0,1,2\}^{1000\times 500}$ and i.i.d. traits $y_{il} \sim \mathcal{N}(0,1)$, $i = 1,\dots,1000, l = 1,\dots,10$. For the scenario (a), we assume that the SNPs have no effect on the traits, that is, $c_{jl} = 0$ for all $j = 1,\dots,500, l = 1,\dots,10$. This way, genetic effects explain none of the total variance in the traits, or, equivalently, the broad-sense heritability of the traits $\mathbf{y}_1,\dots,\mathbf{y}_{10}$ is known to be $H_l^2 = 0$ for all $l = 1,\dots,10$. For the scenario (b), we assume that all SNPs have equal effect on all traits with $c_{jl} = 0.01$, for all $j = 1,\dots,500, l = 1,\dots,10$. This way, genetic effects explain 2% of the total variance in each trait, that is, $H_l^2 = 0.02$ for all $l = 1,\dots,10$. For the scenario (c) of single cross-trait association on SNP level, we assume that $c_{11} = c_{12} = 1.4$ such that $H_l^2 = 0.4$ for $l = 1,2$. For the scenario (d) of multiple cross-trait association on SNP level, we assume that $c_{11} = c_{12} = c_{23} = c_{24} = 1.4$ such that $H_l^2 = 0.4$ for $l = 1,\dots,4$. Finally, for the scenario (e) of overlapping multiple cross-trait association on SNP level, we assume that $c_{11} = c_{22} = 0.6$ and $c_{12} = c_{21} = 1.4$ such that $H_l^2 = 0.5$ for $l = 1,2$.

We present the average SE (grey dots) averaged over the 100 simulations with corresponding point-wise approximate confidence intervals which we compute as the average SE plus/minus twice the standard error of the average SE. We indicate the minimal average SE at $\hat{\rho}_{\min}$ by a coloured orange dot. We also present visualisations of the optimal clusters, by means of the estimated adjacency matrix at $\hat{\rho}_{\min}$. In Figure 2 we plot diagnostics. In the left panel we show the average SE averaged over 100 simulations with corresponding point-wise upper and lower standard deviation intervals plotted against the values of ρ used in the fits. The orange coloured dot indicates the minimum average SE and corresponding regularisation parameter $\hat{\rho}_{\min}$. In the right panel we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$ for one, randomly chosen, simulation. Grey squares represent estimated edges, white space represents no edges, and orange borders represent the true edges. In Figures 2.a and 2.b we observe that, for the scenarios (a) and (b) the average SE curve attains a minimal value (orange dot) at the largest regularisation, yielding an almost empty adjacency matrix. In the case (a), this almost corresponds to the truth of no clustered traits, and for the case (b) it corresponds quite badly with the truth of one cluster of all traits. In Figures 2.c–2.e we observe that, for the scenarios (c)–(e) the average SE curve attains a minimal value at $\hat{\rho}_{\min}$ (orange dot) and the point-wise standard deviation intervals appear narrow. For the randomly chosen simulation, we observe that the minimal average SE is obtained exactly when the true clustering is discovered, as illustrated in the right panel.

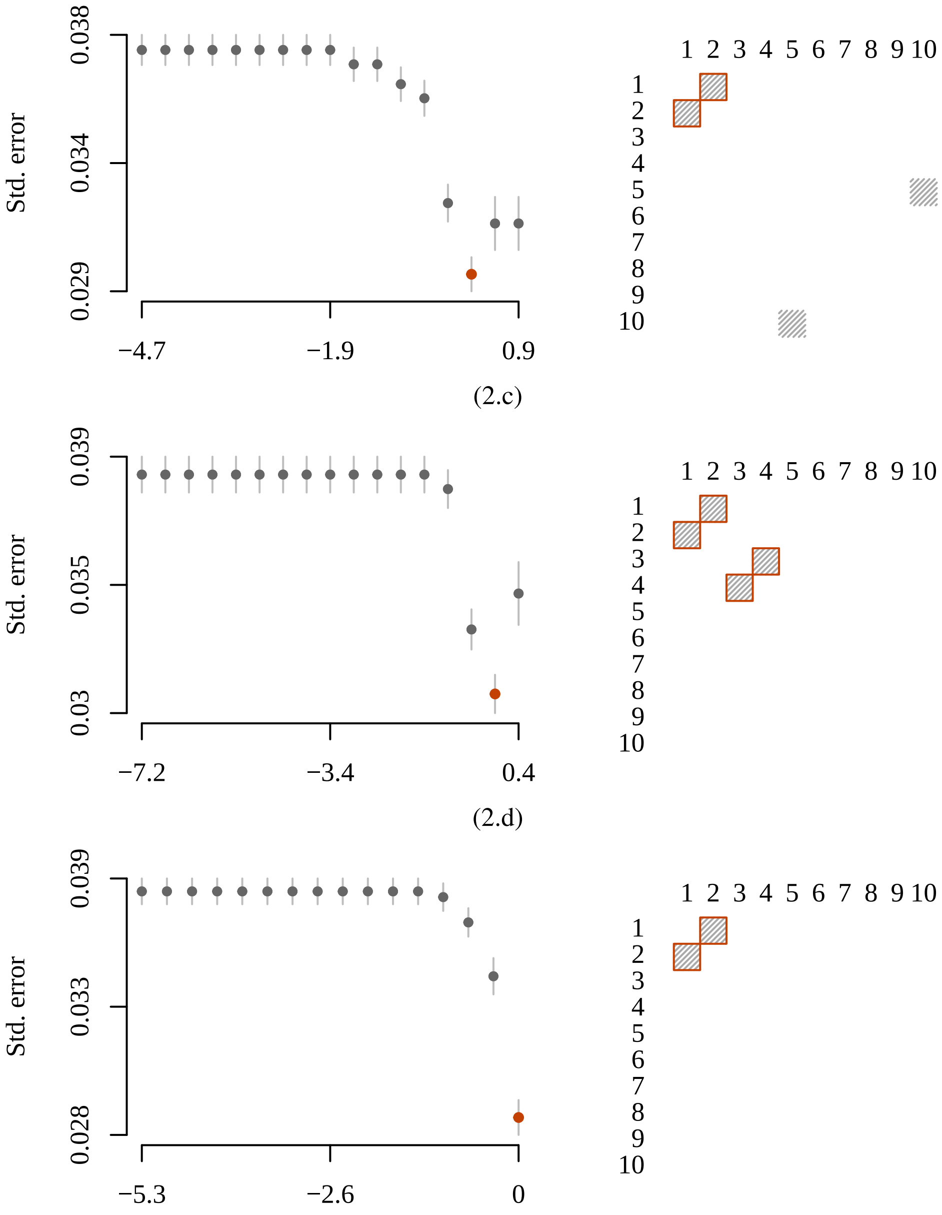
The scenarios (b)–(e) are tried with a broad-sense heritability ranging between 0.1 and 0.5. The resulting RIs are shown in Figure 3. We observe that for the scenarios (c)–(e) the RIs are greater than 0.9 (and for heritability of 0.4 or more exactly one), that is, the true clustering is discovered in almost all of the 100 simulations.



(2.a)

(2.b)

Figure 2: *(Continued on the next page.)*



(2.c)

(2.d)

(2.e)

Figure 2: *(Continued from the previous page.) Diagnostics plots for simulations of scenarios illustrated in Figure 1.* Left panel: *Average SE curve with corresponding point-wise approximate confidence intervals as a function of ρ. Orange dot indicates $\hat{\rho}_{\min}$.* Right panel: *adjacency matrix corresponding to $\hat{\rho}_{\min}$. Grey squares represent estimated edges, white space represents no edges, and orange borders represent the true edges.*

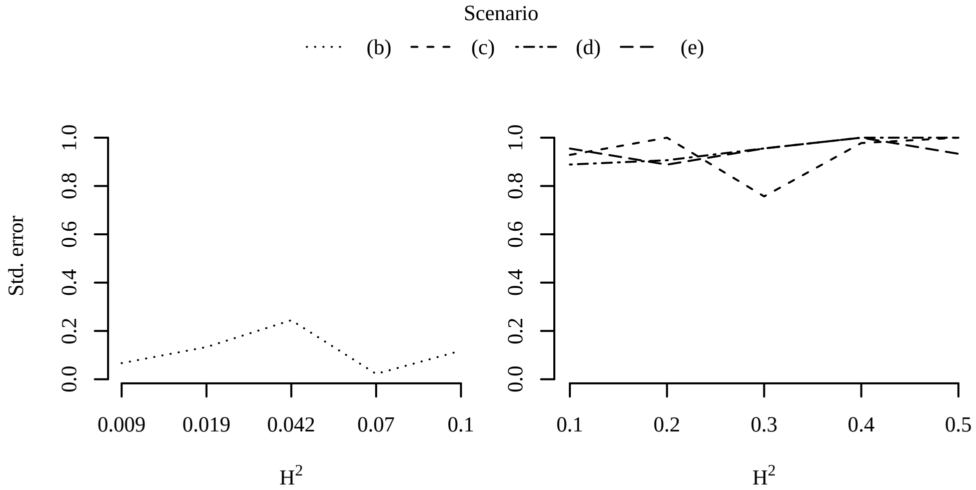


Figure 3: *Rand index, RI, for the simulated scenarios (b) (left) and (c)–(e) (right) at different values of the broad-sense heritability H^2.*

3.1.3 Comparison of methods

In this section we compare the performance – in terms of precision and computing time – of the geneJAM, simple univariate linear regressions, and linear mixed effects models.

As mentioned, standard multiple linear regression models assume that the underlying observations are independent. As an extension to the multiple linear regression models *multilevel models* or *linear mixed-effects models* (LMM) assume observations to be inhomogeneous in the sense that they are not independent but grouped. If we treat multiple simultaneously measured traits $\mathbf{Y}$ in a repeated measurement setting, then a univariate multilevel model can be used to explain the phenotypic variation: We define $\check{\mathbf{Y}} = \operatorname{vec}(\mathbf{Y}) = (\mathbf{y}_1, \ldots, \mathbf{y}_q)^\top \in \mathbb{R}^{Nq}$ as the column-wise vectorisation of $\mathbf{Y}$, that is,

$$\check{\mathbf{Y}} = (y_{11}, y_{21}, \ldots, y_{N1}, y_{12}, y_{22}, \ldots, y_{N2}, \ldots, y_{1q}, y_{2q}, \ldots, y_{Nq})^\top ,$$

not to be mistaken with the row-wise vectorisation of $\mathbf{Y}$ denoted by $\vec{\mathbf{Y}}$ and introduced in (2). Then, the *univariate multilevel model* takes the form:

$$\check{\mathbf{Y}} = \check{\mathbf{Z}}\mathbf{b} + \mathcal{U}\boldsymbol{\alpha} + \check{\mathbf{E}}, \tag{4}$$

where $\check{\mathbf{Z}} \in \mathbb{R}^{Nq \times 2q}$ is a fixed-effects design matrix, $\mathbf{b} \in \mathbb{R}^{2q}$ is a vector of regression coefficients for the fixed effects, such that each trait has a fixed intercept and slope, $\boldsymbol{\alpha} \in \mathbb{R}^N$ are person-specific regression coefficients for the random effect with random-effects design matrix $\mathcal{U} \in \mathbb{R}^{Nq \times N}$, ensuring that each individual has its own trait-specific random intercept, and $\check{\mathbf{E}} = \operatorname{vec}(\mathbf{E}) = (\mathbf{e}_1, \ldots, \mathbf{e}_q)^\top \in \mathbb{R}^{Nq}$ are trait-specific residual errors caused by nonadditive genetic variation, random environmental effects, and measurement error. We define the random-effects design matrix as an identifier of the individuals and assume that

(i) $\mathbf{e}_l \sim \mathcal{N}_N(\mathbf{0}, \sigma^2_{\mathbf{e}}\mathbf{I}_N)$ and uncorrelated for $l = 1, \dots, q$;

(ii) $\boldsymbol{\alpha} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2_{\boldsymbol{\alpha}}\mathbf{G})$;

(iii) $\mathbf{e}_1, \dots, \mathbf{e}_q$, and $\boldsymbol{\alpha}$ are independent.

Here $\mathbf{G} \in \mathbb{R}^{N\times N}$ is a general covariance matrix. Possible structures for $\mathbf{G}$ may be a block diagonal, a diagonal or a general positive definite matrix. This way, restricted maximum likelihood (REML) estimates of the parameters in models on the form (4) can be determined using the `lmer` function in the `lme4` package for R (Bates et al., 2015). We assume a common correlation among the observations from a single individual, with the correlation being the same for all individuals. Compared to the assumptions for linear regression model (2) of the geneJAM method, the assumption is analogous to the residual covariance matrix $\boldsymbol{\Omega}$ being a block diagonal with blocks corresponding to the individuals and with each block having a compound-symmetry structure. This structure has two unknown parameters, one modeling a common covariance and the other a residual variance. The form for $\mathbf{R}$ would then be as follows:

$$\mathbf{R} = \begin{pmatrix} \sigma^2_{\mathbf{e}} + \sigma^2_{\boldsymbol{\alpha}} & \sigma^2_{\mathbf{e}} & \cdots & \sigma^2_{\mathbf{e}} \\ \sigma^2_{\mathbf{e}} & \sigma^2_{\mathbf{e}} + \sigma^2_{\boldsymbol{\alpha}} & \ddots & \vdots \\ \vdots & \ddots & \sigma^2_{\mathbf{e}} + \sigma^2_{\boldsymbol{\alpha}} & \vdots \\ \sigma^2_{\mathbf{e}} & \cdots & \sigma^2_{\mathbf{e}} & \sigma^2_{\mathbf{e}} + \sigma^2_{\boldsymbol{\alpha}} \end{pmatrix}. \tag{5}$$

In comparison to the geneJAM method, this is more restrictive as the geneJAM method allows for an arbitrary structure of the residual covariances $\mathbf{R}$ among traits within each individual in (3). We use this method as reference when assessing the precision and computational performance of our method. We also compare to fitting simple linear regression models for each trait and corresponding PGS separately.

We simulate 100 data sets according to the sampling procedure described in Section 3.1.1.

We are interested in improving the precision of the estimates of traits on which there are genetic effects. We refer to this set of estimates as the *active set*. In Figure 4 we show the standard error of the estimates in the active set averaged over the 100 simulations. We observe that regardless of the scenario and heritability the geneJAM perform better than both the simple linear regression (LM) and the multilevel model (LMER). As expected, the larger the heritability the better the absolute precision of the geneJAM.

In Figure 5 we show the RMSE in the active set averaged over the 100 simulations. The figures show that, in general, the prediction performance of the methods discussed are worsening

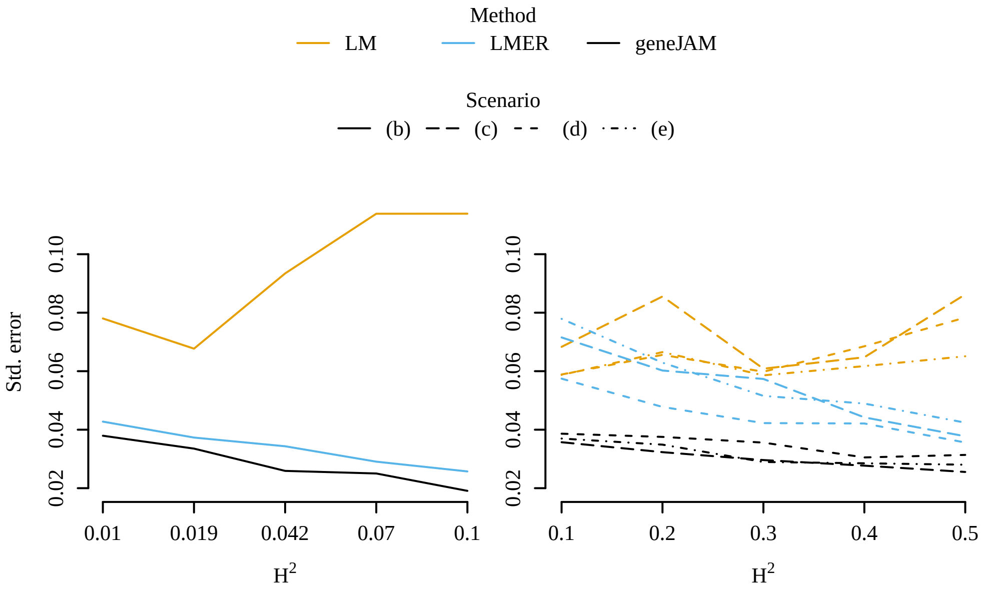


Figure 4: *Standard error of the estimates in the active set averaged over 100 simulations for the simulated scenarios (b) (left) and (c)–(e) (right) at different values of the broad-sense heritability H^2.*

when the heritability grows since the RMSE is increasing in H^2. We observe that, for low heritability, the RMSE is very similar for all three methods and all scenarios (b)–(e), that is, they are equally suited for prediction. In the scenario (b), a higher heritability results in more varying RMSEs for the three methods. Specifically, the geneJAM and the LM methods seem to be better qualified for prediction than the LMER. As expected, for the scenarios (c)–(e) the RMSE is the same for the LM and LMER methods. For the scenarios (c) and (d) the geneJAM method seems to slightly outperform the other methods in terms of predictive ability. For the scenario (e) the RMSE does not appear to differ substantially between the three methods, that is, their performance in terms of numeric prediction is very close.

In Figure 6 we visualise the covariance structure between traits for simulations of the scenarios (a)–(e) illustrated in Figure 1 corresponding to the subfigures (rows) (6.a)–(6.e), respectively. Dark colours indicate large values and light colours indicate small values. In the left panel we show the structure induced by the sampling, in the centre panel we show the structure estimated by the geneJAM method, and in the right panel we show the structure estimated by the multilevel model. We observe that, for the scenarios (c)–(e) in particular, the geneJam method appear to succefully capture the covariance structure induced by the sampling. This is not possible for the multilevel model which require a fixed structure such as (5).

To assess the computation time of the geneJAM we consider a simulation design similar to the one proposed in the previous subsection with a single cross-trait association and broad-sense

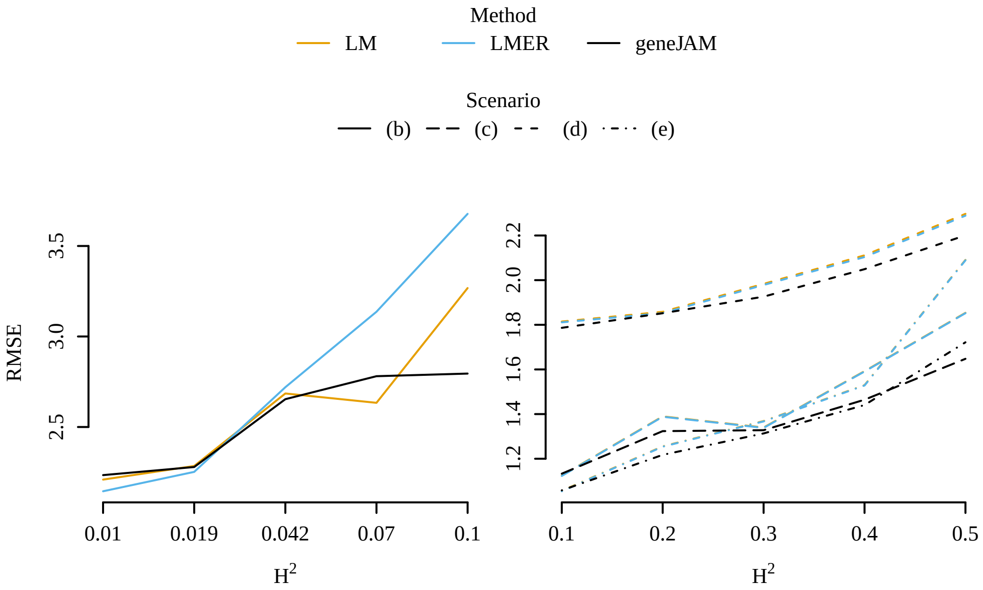
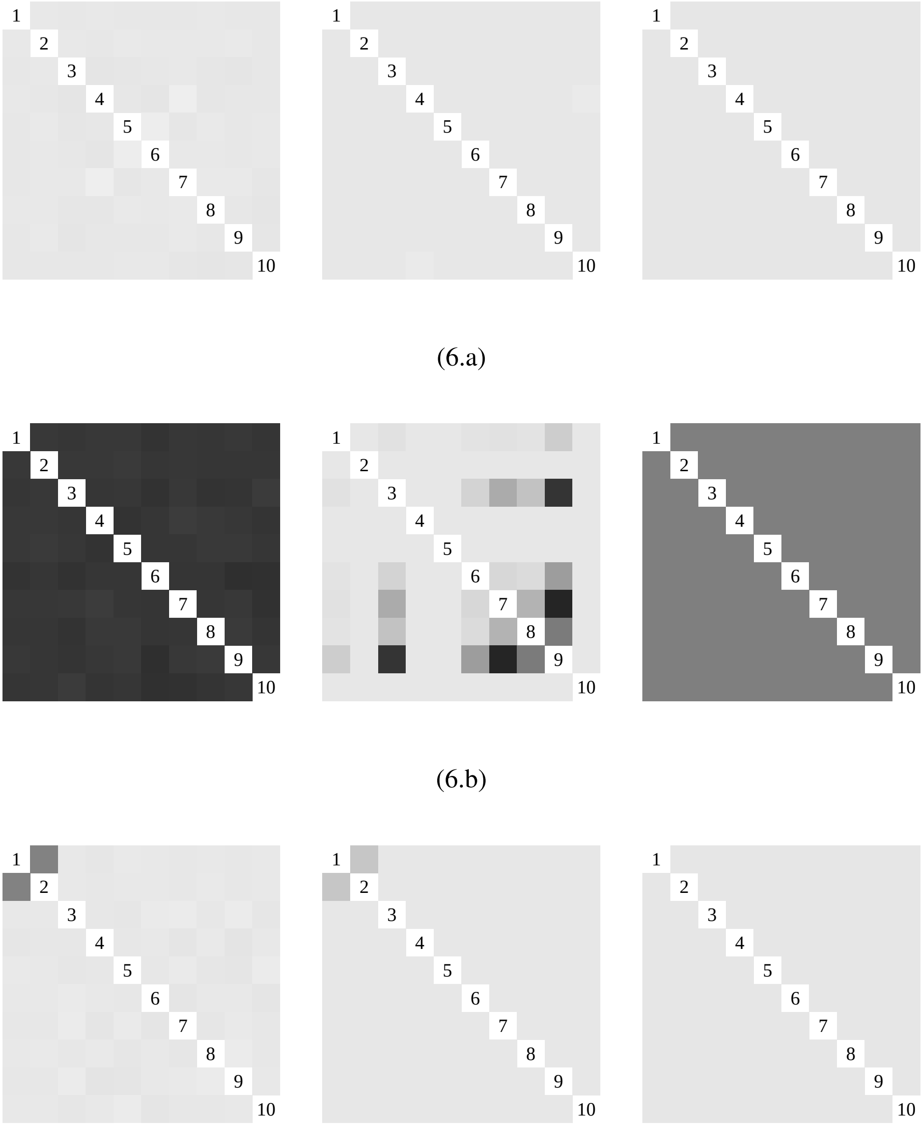


Figure 5: *Root mean squared error (RMSE) of the estimates in the active set averaged over 100 simulations for the simulated scenarios (b) (left) and (c)–(e) (right) at different values of the broad-sense heritability H^2.*

heritability $H^2 = 0.5$. We assume that the grouping is given so we do not run the graphical lasso part of the procedure. To assess the performance of the R code we use the package `microbenchmark` by Mersmann (2019).

In Figure 7 we show how the geneJAM and multilevel model scale in execution time in the number of traits q (left panel) and observations N (right panel). To assess the scaling in the number of traits each sample is generated with $N = 1000$ observations. First, we observe that the computing time of the geneJAM is almost constant in q, and fitting one FGLS takes approximately 1.2 seconds on an Intel Core i5-5200U processor. Second, we observe that, while the mixed model is faster than the geneJAM for $q < 30$, the computing time increases almost exponentially as the number of traits increases logarithmically. For $q > 30$ the mixed model is slower than geneJAM, and the `lme4` cannot handle $q > 2^6$. To assess the scaling in the number of observations each sample is generated with $q = 32$ traits. As expected, given the number of traits fixed at 32, the geneJAM is, for all values of N faster than the multilevel model. The computing time for the geneJAM increases linearly in q, whereas the computing time for the multilevel model increases faster.

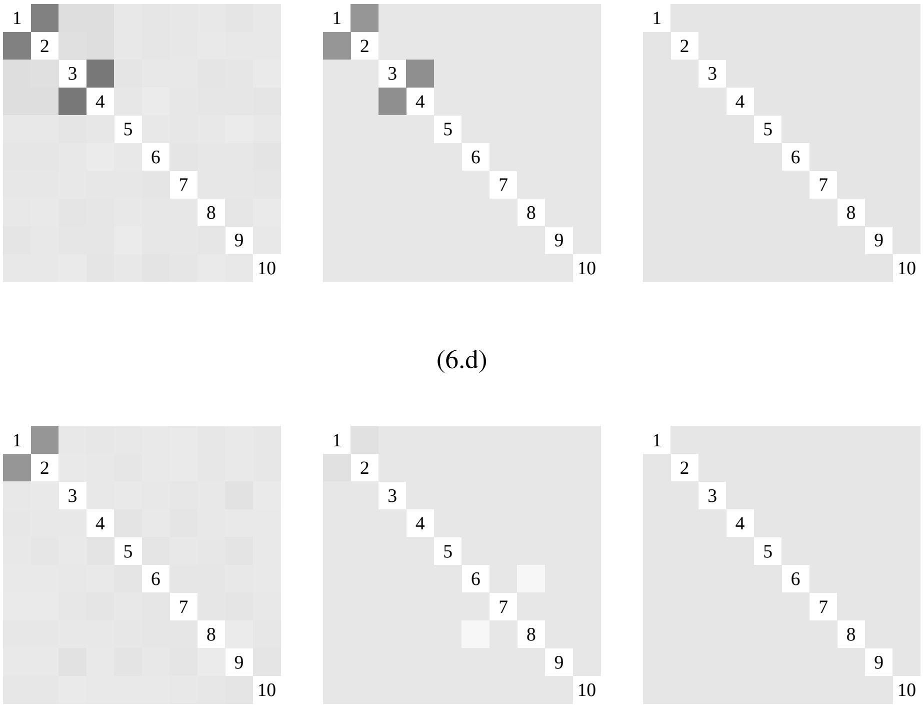


(6.a)

(6.b)

(6.c)

Figure 6: *(Continued on the next page.)*



(6.d)

(6.e)

Figure 6: *(Continued from the previous page.) Visualisation of covariance matrices for simulations of scenarios illustrated in Figure 1. Dark colours indicate large values and light colours indicate small values.* Left panel: *Structure induced by the sampling.* Centre panel: *Structure estimated by the geneJAM method.* Right panel: *Structure estimated by the multilevel model.*

3.2 Real data

We apply geneJAM to the heterogeneous stock of mice data set available from the Wellcome Trust Centre for Human Genetics (Flint and Valdar, 2008).

The heterogeneous stock of mice consists of 1,904 individuals from 85 families, all descended from eight inbred progenitor strains (Valdar et al., 2006). We consider this particular dataset not only because it contains multiple quantitative traits, but also because many of the traits are naturally associated, and thus this dataset presents a realistic mix between our simulation scenarios. Please note that the aim of the example is to demonstrate proof of concept; the

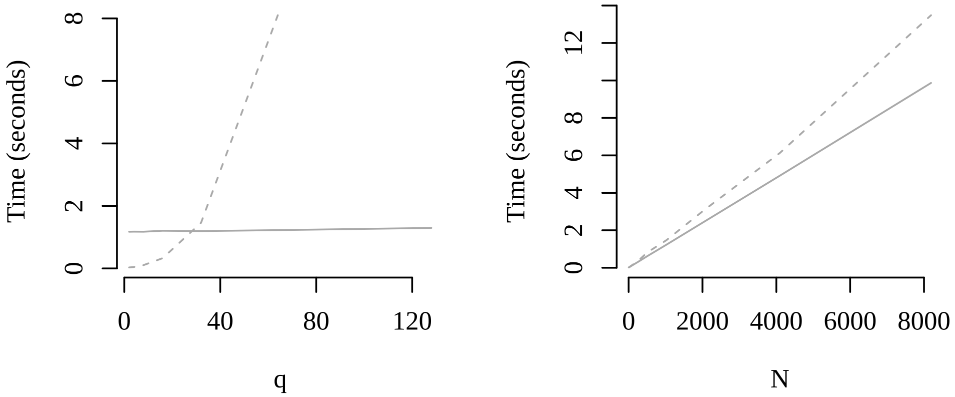


Figure 7: *Scaling (measured in seconds) of the geneJAM (solid) and multilevel model (dashed) in the number of traits q (left panel) and observations N (right panel).*

proposed method is used as a screening tool for identifying potential clusters of the traits. This way we can subsequently investigate the associations using more complicated nested models taking into account the biological relationships of related subjects. The data contain 129 quantitative traits which are classified into six broad categories including behaviour, diabetes, asthma, immunology, haematology, and biochemistry. Because of the fairly large amount of missing data in the traits, and to avoid having to discuss complicated imputation techniques, we narrow down the analysis and include only the 14 traits in the asthma category which has the fewest missing values. Development and deployment of a phenotyping protocol to collect measures on a model of asthma is described in Solberg et al. (2006) as well as a definition of the 14 traits in the asthma category. We omit individuals with missing trait and obtain $N = 1491$ observations. A total of 12,226 autosomal SNPs were available for all mice. As Crawford et al. (2018), for "individuals with missing genotypes, we imputed missing values by the mean genotype of that SNP in their family. All polymorphic SNPs with minor allele frequency above 1% in the training data were used for prediction". Furthermore, SNPs with no variation across observations are removed and we obtain $p = 10994$ SNPs.

We are interested in clustering traits sharing genetic characteristics in preparation for fitting low-dimensional multivariate linear regression models. For each quantitative trait we generate a PGS as described in Section 3.1.1. GeneJAM is applied at a suitable sequence of the regularisation parameters ρ.

In Figure 8 we show diagnostics plots. In the top row we show the average SE plotted against the values of ρ used in the fits. The orange coloured dot indicates the minimum average SE and corresponding regularisation parameter $\hat{\rho}_{\min}$. We observe that the average SE curve attains a minimum at $\hat{\rho}_{\min}$. In the bottom row we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$. Grey squares represent estimated edges and white space represents no edges. We observe that the best precision is obtained when all traits except for one are clustered together. The singleton cluster identified is the sixth trait (Pleth.base.InspiratoryTime) which is defined as the baseline inspiratory time measured by the plethysmograph (Solberg et al., 2006).

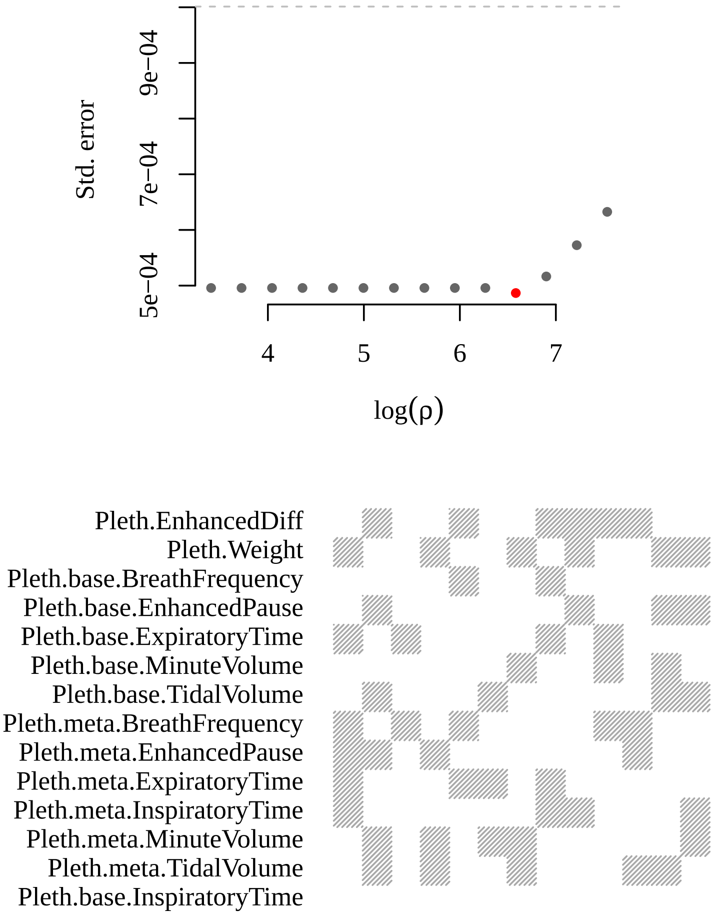


Figure 8: *Diagnostics plot for the heterogeneous stock of mice data set.* Top: *Average SE curve as a function of ρ.* Bottom: *adjacency matrix corresponding to $\hat{\rho}_{\text{min}}$. Grey squares represent estimated edges and white spaces represent no edges.*

In Figure 9 we visualise the covariance structure between traits for the heterogeneous stock of mice data set. Dark colours indicate large values and light colours indicate small values. In the top row we show the structure observed in data and in the bottom row we show the structure estimated by the geneJAM method. We observe some similar structures in the two figures, e.g., a stronger correlation between Pleth.meta.MinuteVolume and Pleth.meta.TidalVolume and between Pleth.base.MinuteVolume and Pleth.base.

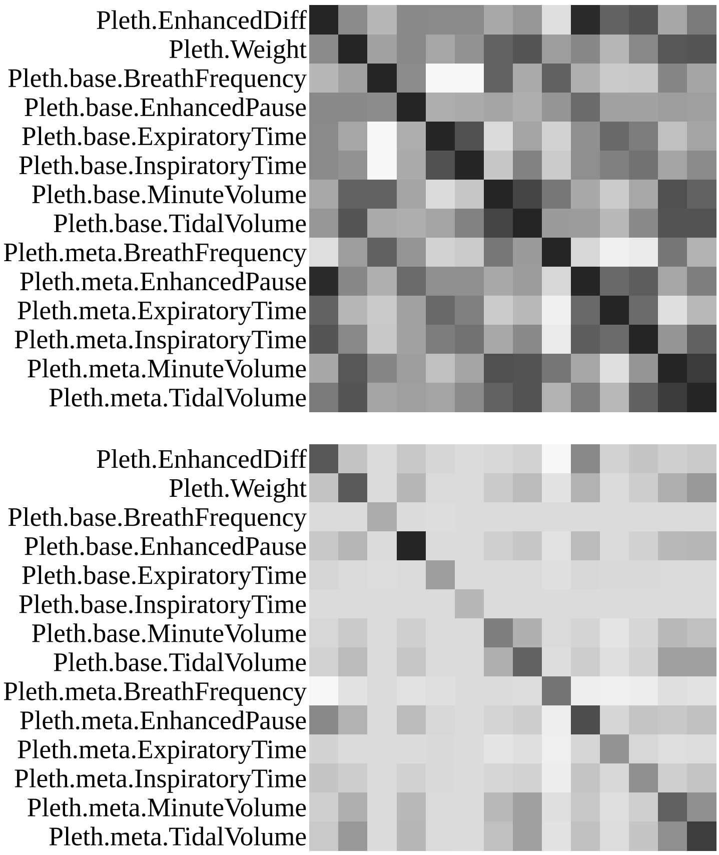


Figure 9: *Visualisation of the covariance structure for the heterogeneous stock of mice data set. Dark colours indicate large values and light colours indicate small values.* Top: *Structure observed in the data.* Bottom: *Structure estimated by the geneJAM method.*

`TidalVolume`. This is indication of two pairs of genetic relationships among the traits. We also see structures in the observed covariance which is not present in the structure estimated by the geneJAM method. There is, e.g., a stronger observed relationship between `Pleth.meta.EnhancedPause`, `Pleth.meta.ExpiratoryTime`, and `Pleth.meta.InspiratoryTime`. This may be indication of traits that are related for other reasons than genetics and the realtions are, therefore, not captured by the geneJAM method.

4 Discussion

In this paper, we have proposed a method, which uses PGSs and graphical lasso for clustering simultaneously measured traits sharing genetic characteristics in preparation for fitting low-dimensional multivariate linear regression models. Furthermore, we have proposed a procedure for choosing the optimal regularisation for the graphical lasso as well as a method for fitting the models using the FGLS estimator.

Key advantages of our framework are that there are no prior assumptions on the structure or sizes of the clusters of traits, information on PGS level is sufficient when data on SNP level is not available, and, following the clustering and modelling, both clusters and model fit are readily available. Moreover, the computing time is highly satisfactory even for large data, and compared to both the simple linear regression and the multilevel models the geneJAM is superior in terms of precision of the estimates of traits on which there are genetic effects. Finally, if (independent) environmental factors are available, they are easily adjusted for.

To understand the relevance of cluster assignment we simulated data sets with different cross-trait structures and effect sizes. In the extreme case, where no traits are affected by the SNPs, the RI is close to one. Moreover, as expected, the estimated adjacency matrix quickly reaches the true zero matrix. In the extreme case, where all traits are equally affected by all SNPs, the geneJAM does not perform very well – as would be expected; the RI decreases as the heritability increases and even for small values, the RI is only 0.11. Similarly, the estimated adjacency matrix goes from the true one matrix to the zero matrix at which the minimum average SE is (erroneously) attained. More interestingly, when a subset of features affect a subset of clustered traits and when the effect sizes are sufficiently large, the average SE attains a maximum value (not at the tails of the sequence of ρ), indicating that a range of values of the regularisation parameter exists, such that an optimal (attainable) agreement between the estimated and the true clustering is reached. This, indeed, appears to be the case, as indicated by the sequences of estimated adjacency matrices: in general, the sparsity of the matrix increases in ρ, and for a range of values of ρ the estimate is equivalent to the adjacency matrix representing the true clustering of traits. Furthermore, the larger the effect size the larger the range at which the clustering is correctly estimated.

Moreover, using the implemented tuning method we are able to determine a value of ρ at which the true clustering is identified and the average SE is minimised.

In practice, researchers may have preconceived ideas about clusters of traits sharing genetic characteristics. If no clusters are identified by the geneJAM, it is because the genetic signal is too low (or the presumption is wrong) and, in this case, there is nothing to gain by a joint analysis of the traits. It should be emphasised that traits may be correlated due to other (non-genetic) factors, in which case the geneJAM may be unable to detect the clusters. If, however, summary statistics related to the correlation (e.g., metabolite risk scores) are available, these can be used in the method instead of PGSs to obtain clusters of traits sharing other (e.g., metabolic) characteristics.

While the multilevel model is useful where multiple correlated measurements are made on each individual the geneJAM allows for correlation of traits, thereby improving the precision of the estimates of traits on which there are genetic effects. We intend to extend the geneJAM method to also allow for correlated measurements on each individual, to investigate the behaviour of related individuals.

The algorithms presented in Zhou and Stephens (2014) are computationally-efficient algorithms for fitting multivariate linear mixed models in GWASs. In practice, however, there could remain both computational and statistical barriers to applying the methods to even a quite modest number of traits, e.g., $q \approx 10$.

Supplementary Material

A Algorithmic Overview

Algorithm 1: geneJAM algorithm

```
for $l = 1, \dots, q$ do
    Compute $\hat{\mathbf{b}}_l^{\mathrm{O}}$ and $\hat{\mathbf{u}}_l$
end
Centre PGSs $\mathbf{Q}$ per column
Compute empirical covariance matrix $\boldsymbol{\Sigma}$ of centred $\mathbf{Q}$
if no sequence of regularisation parameters $\rho_r$, $r = 1, \dots, R$, is provided then
    Specify length $R$ of sequence of regularisation parameter $\rho_r$
    Specify ratio $\delta$ between regularisation parameters $\rho_r$, $r = 1, \dots, R$
    Choose a sequence of regularisation parameters $\rho_r$, $r = 1, \dots, R$, with maximal
    value, $\rho_R$, defined by the maximum column sum of $\boldsymbol{\Sigma}$ and the rest of the sequence
    determined by $R$ and $\delta$.
end
for $r = 1, \dots, R$ do
    Estimate sparse precision matrices $\hat{\mathbf{P}}^{(\rho_r)}$ to obtain $C^{(\rho_r)}$ clusters
    for $g(\rho) = 1, \dots, C^{(\rho_r)}$ do
        Compute $\hat{\Omega}_{g(\rho)}^{\mathrm{O}}(\rho)$
    end
    Construct $\hat{\boldsymbol{\Omega}}^{\mathrm{O}}$
    Compute estimate $\hat{\vec{\mathbf{B}}}^{\mathrm{F}}$
    Define $\hat{\boldsymbol{\Omega}}^{\mathrm{F}(1)} = \hat{\boldsymbol{\Omega}}^{\mathrm{O}}$
    Initialise $t \to 1$
    repeat
        Compute $\hat{\mathbf{u}}^{\mathrm{F}(t)}$
        for $g(\rho) = 1, \dots, C^{(\rho_r)}$ do
            Compute $\hat{\Omega}_{g(\rho)}^{\mathrm{F}(t)}$
        end
        Construct $\hat{\boldsymbol{\Omega}}^{\mathrm{F}(t)}$
        Compute estimate $\hat{\vec{\mathbf{B}}}^{\mathrm{F}(t+1)}$
    until convergence
end
```

B Simulated data

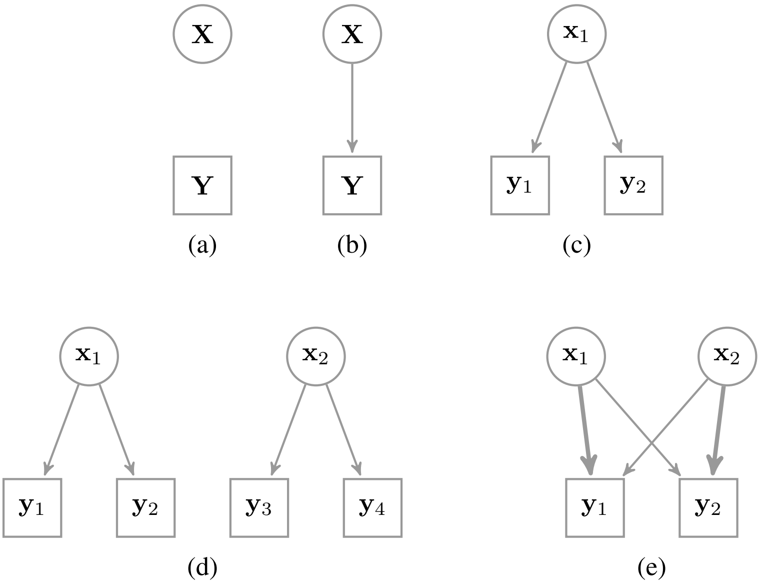


Figure B1: *Graphical illustration of different cross-trait associations.* (a) *and* (b) *are extreme cases where there are no genetic effects on SNP level (a), and where all SNPs affect all traits equally (b).* (c) *is a case of single cross-trait association where one SNP affects multiple traits.* (d) *is a case of multiple cross-trait association where multiple SNPs affect distinct clusters of traits. Finally,* (e) *is a case of overlapping cross-trait association where multiple SNPs affect the same cluster of traits.*

B.1 Extreme cases

In this section we present the result of applying the geneJAM method when there are no genetic effects on SNP level in the data generating process, as illustrated in Figure 1 (a), and when all SNPs affect all traits equally in the data generating process, as illustrated in Figure 1 (b). The simulations are generated according to the sampling procedure presented in the main manuscript.

Thus, for the first scenario, we consider simulations of i.i.d. features $\boldsymbol{X} \in \{0,1,2\}^{1000\times 500}$ and i.i.d. traits $\mathbf{Y}_{il} \sim \mathcal{N}(0,1)$, $i = 1,\ldots,1000, l = 1,\ldots,10$ on which the genes have no effect. For the second scenario, we consider simulations of i.i.d. features $\boldsymbol{X} \in \{0,1,2\}^{1000\times 500}$ and traits $\mathbf{Y} \in \mathbb{R}^{1000\times 10}$ on which all genes have equal effect, that is, for given non-zero values of $\boldsymbol{B}_{il}$, for all $i = 1,\ldots,1000, l = 1,\ldots,10$.

Table 1: *Rand index, RI, for the simulated scenarios (a)–(e) at different values of the broad-sense heritability* H^2.

	H^2	RI
(a)	0.00	0.98
(b)	0.01	0.11
	0.02	0.05
	0.04	0.04
(c)	0.17	0.93
	0.41	1.00
	0.60	1.00
(d)	0.17	0.04
	0.41	1.00
	0.60	1.00
(e)	0.24	0.96
	0.39	1.00
	0.58	1.00

In Figure B2 we plot diagnostics for the simple scenario with no effects. In the left panel we show the average SE averaged over 100 simulations with corresponding point-wise upper and lower standard deviation intervals plotted against the values of ρ used in the fits. The orange coloured dot indicates the minimum average SE and corresponding regularisation parameter $\hat{\rho}_{\min}$. We observe that the average SE curve is almost constant in ρ, but, on average, the best precision is obtained with maximal regularisation, that is, by analysing each trait separately. We do, however, also observe that the point-wise upper and lower standard deviation intervals are very wide. The RI is 0.98 for this scenario, see Table 1. In the right panel we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$ for a single, randomly chosen, simulation example. Grey squares represent estimated edges, white space represent no edges, and orange borders represent the true edges. It shows that for this random sample the minimal average SE is obtained with the maximal regularisation when two traits are clustered together.

In Figure B3 we plot diagnostics for the scenario where all SNPs affect all traits with heritability $H_l^2 = 0.01$ (top), $H_l^2 = 0.02$ (middle), and $H_l^2 = 0.04$ (bottom), $l = 1, \ldots, 10$. In the left panel we show the average SE averaged over 100 simulations with corresponding point-wise upper and lower standard deviation intervals plotted against the values of ρ used in the fits. We observe that in all three cases the average SE curve is almost constant in ρ, but, on average, the best precision is obtained with maximal regularisation (orange) at $\hat{\rho}_{\min}$, that is, using information across all traits. This is consistent with our expectations and the true clustering of the traits. In the right panel we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$. It shows that for this random sample the minimal average SE is obtained when few traits are clustered together. The RIs for the scenario (b) are 0.11, 0.05, and 0.04, for the heritability, $H_l^2 = 0.01, 0.02, 0.04$, respectively for all $l = 1, \dots, 10$, see Table 1. We see that the stronger the heritability the smaller the RI, that is, with a stronger heritability comes worse clustering performance, when all SNPs affect all traits equally in the data generating process.

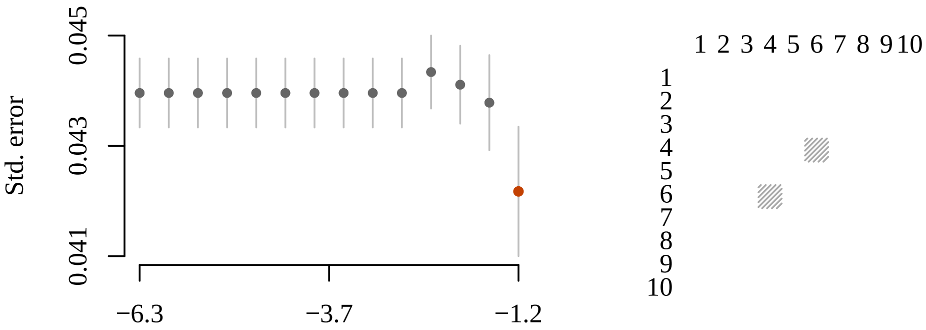


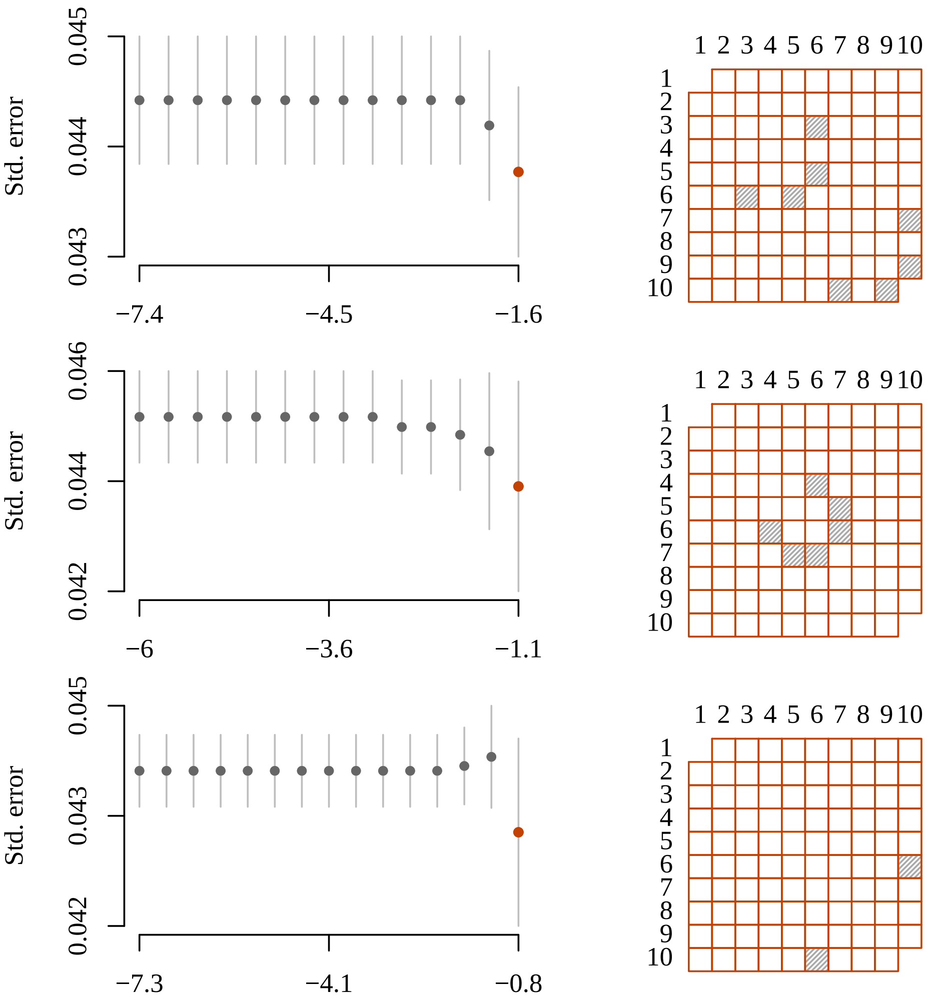
Figure B2: *Diagnostics plots for simulations of the simple scenario illustrated in Figure 1 (a).* Left panel: *Average SE curve with corresponding point-wise approximate confidence intervals as a function of ρ.* Right panel: *adjacency matrix corresponding to $\hat{\rho}_{\min}$. Grey squares represent estimated edges, white space represent no edges, and orange borders represent the true edges.*

In Figure B7–Figure B10 in the supplementary material we show, for one randomly chosen simulation of each of the two extreme scenarios, visualisations of the estimated adjacency matrices at all values of the sequences of ρ tried. As expected, the sparsity of the matrix increases in ρ.

B.2 Single cross-trait association

In this section we present the result of applying the geneJAM method when the data generating process honours single cross-trait association on SNP level, exemplified by genetic correlations as illustrated in Figure 1 (c). Thus, we consider simulations for different effect sizes of the feature $\boldsymbol{X}_1$ on two traits, $\mathbf{Y}_1$ and $\mathbf{Y}_2$, that is for different given non-zero values of $\boldsymbol{B}_{11}$ and $\boldsymbol{B}_{12}$.

In Figure B4 we plot diagnostics for the scenario (c) with $H_l^2 = 0.17$ (top), $H_l^2 = 0.41$ (middle), and $H_l^2 = 0.60$ (bottom), $l = 1, 2$. In the left panel we show the average SE averaged over 100 simulations with corresponding point-wise upper and lower standard deviation intervals plotted against the values of ρ used in the fits. We observe that, for a range of smaller values of ρ, the average SE curve is constant before it decreases to the minimal value at $\hat{\rho}_{\min}$ (orange) at which it stays. In the right panel we show, for one of the 100 simulations, a



	1	2	3	4	5	6	7	8	9	10
1										
2										
3						▨				
4										
5						▨				
6			▨		▨					
7										▨
8										
9										▨
10							▨		▨	

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4						▨				
5							▨			
6				▨			▨			
7					▨	▨				
8										
9										
10										

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										▨
7										
8										
9										
10						▨				

Figure B3: *Diagnostics plots for simulations of the scenario illustrated in Figure 1 (b) with* $H_l^2 = 0.01$ *(top),* $H_l^2 = 0.02$ *(middle), and* $H_l^2 = 0.04$ *(bottom),* $l = 1, \ldots, 10$. Left panel: *Average SE curve with corresponding point-wise approximate confidence intervals as a function of* ρ. Right panel: *adjacency matrix corresponding to* $\hat{\rho}_{\min}$. *Grey squares represent estimated edges, white space represent no edges, and orange borders represent the true edges.*

visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$. It shows that when the tried heritability is 0.41 (middle) or 0.60 (bottom), the minimal average SE is obtained exactly when the true clustering is discovered. The RIs for the scenario (c) are 0.93, 1.00, and 1.00 for the tried heritability 0.4, 0.5, and 0.6, respectively, see Table 1.

In Figure B11, Figure B12, and Figure B13 in the supplementary material we show, for simulations with heritability 0.17, 0.41, and 0.6, respectively, visualisations of the estimated

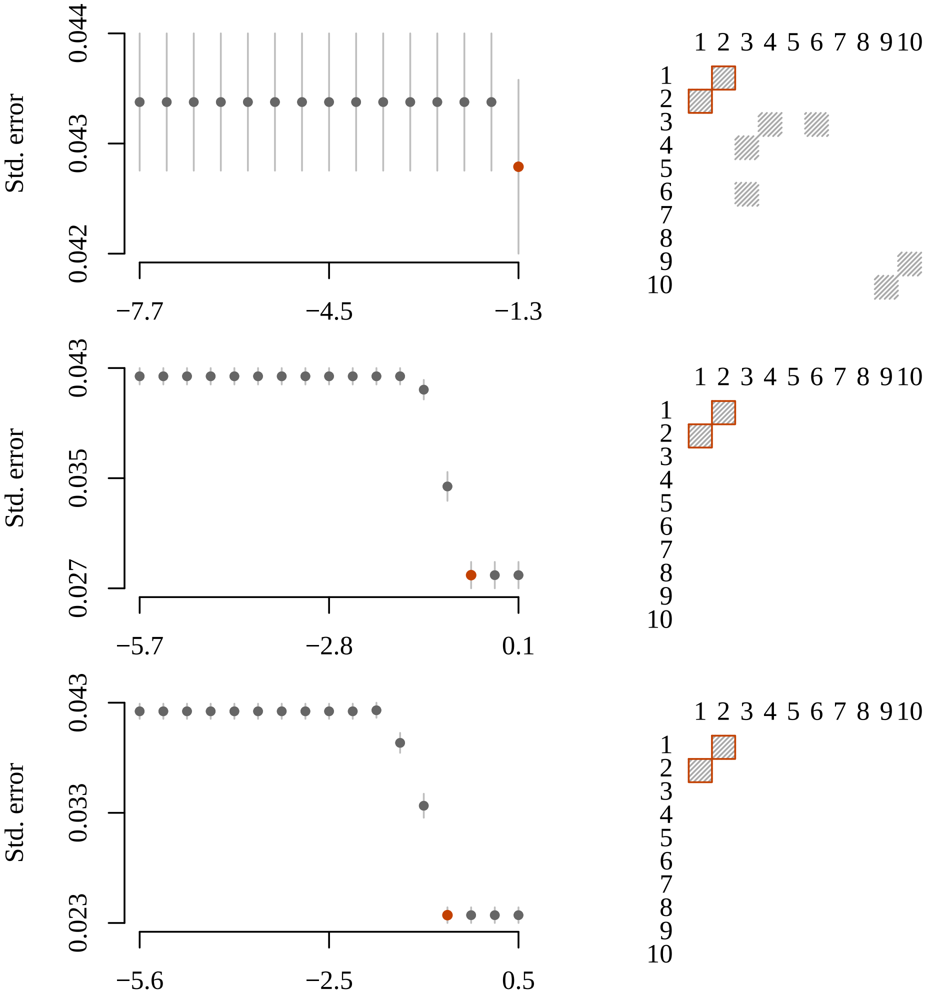


Figure B4: *Diagnostics plots for simulations of simple cross-trait associations illustrated in Figure 1 (c) with* $H_l^2 = 0.17$ *(top),* $H_l^2 = 0.41$ *(middle), and* $H_l^2 = 0.60$ *(bottom),* $l = 1, 2$. Left panel: *Average SE curve with corresponding point-wise approximate confidence intervals as a function of* ρ. Right panel: *adjacency matrix corresponding to* $\hat{\rho}_{\min}$. *Grey squares represent estimated edges, white space represent no edges, and orange borders represent the true edges.*

adjacency matrices at all values of the sequence of ρ tried. As expected, the sparsity of the matrix increases in ρ, and when $H_l^2 \geq 0.41$, $l = 1, 2$, the estimated adjacency matrix is, for a range of values of ρ, equivalent to the adjacency matrix representing the true clustering of traits. Furthermore, we observe that the larger the effect size the larger the range at which the clustering is correctly estimated.

B.3 Multiple cross-trait association

In this section we present the result of applying the geneJAM method when the data generating process honours multiple cross-trait association on SNP level, exemplified by genetic correlations as illustrated in Figure 1 (d). Thus, we consider simulations for different effect sizes of the features $\boldsymbol{X}_1$ and $\mathbf{X}_2$ each on two traits, $(\mathbf{Y}_1, \mathbf{Y}_2)$ and $(\mathbf{Y}_3, \mathbf{Y}_4)$, respectively, that is for different given non-zero values of $\boldsymbol{B}_{11}, \boldsymbol{B}_{12}, \boldsymbol{B}_{23}$, and $\boldsymbol{B}_{24}$.

In Figure B5 we show diagnostics plots similar to those of Figure B4. In the left panel we show the average SE averaged over 100 simulations with corresponding point-wise upper and lower standard deviation intervals plotted against the values of ρ used in the fits. The average SE curve attains a minimal value at $\hat{\rho}_{\min}$ (orange). In the right panel we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$. It shows that when the tried heritability is 0.41 (middle) or 0.60 (bottom) the minimal average SE is obtained exactly when the true clustering is discovered. The RIs for the scenario (d) are 0.04, 1.00, and 1.00 for the heritability 0.17, 0.41, and 0.6, respectively, see Table 1.

In Figure B14, Figure B15, and Figure B16 in the supplementary material we show, for simulations with heritability 0.17, 0.41, and 0.6, respectively, visualisations of the estimated adjacency matrices at all values of the sequence of ρ tried. As expected, the sparsity of the matrix increases in ρ, and, when $H_l^2 \geq 0.41$, $l = 1, \ldots, 4$, it is, for a range of values of ρ, equivalent to the adjacency matrix representing the true clustering of traits. Furthermore, we observe that the larger the effect size the larger the range at which the clustering is correctly estimated.

B.4 Overlapping multiple cross-trait association

In this section we present the result of applying the geneJAM method when the data generating process honours overlapping multiple cross-trait association on SNP level, exemplified by genetic correlations as illustrated in Figure 1 (e). Thus, we consider simulations for different effect sizes of the features $\boldsymbol{X}_1$ and $\mathbf{X}_2$ on two traits, $\mathbf{Y}_1$ and $\mathbf{Y}_2$, that is for different given non-zero values of $\boldsymbol{B}_{11}, \boldsymbol{B}_{12}, \boldsymbol{B}_{21}$, and $\boldsymbol{B}_{22}$.

In Figure B6 we show diagnostics plots similar to those of Figure B4 and Figure B5. In the left panel we show the average SE averaged over 100 simulations with corresponding point-wise upper and lower standard deviation intervals plotted against the values of ρ used in the fits. The average SE curve attains a minimal value at $\hat{\rho}_{\min}$ (orange). In the right panel we show a visualisation of the estimated adjacency matrix for the clustering corresponding to $\hat{\rho}_{\min}$. It shows that when the tried heritability is 0.39 (middle) or 0.58 (bottom) the minimal average SE is obtained exactly when the true clustering is discovered. The RIs for the scenario (e) are 0.96, 1.99, and 1.00 for the heritability 0.24, 0.39, and 0.58, respectively, see Table 1.

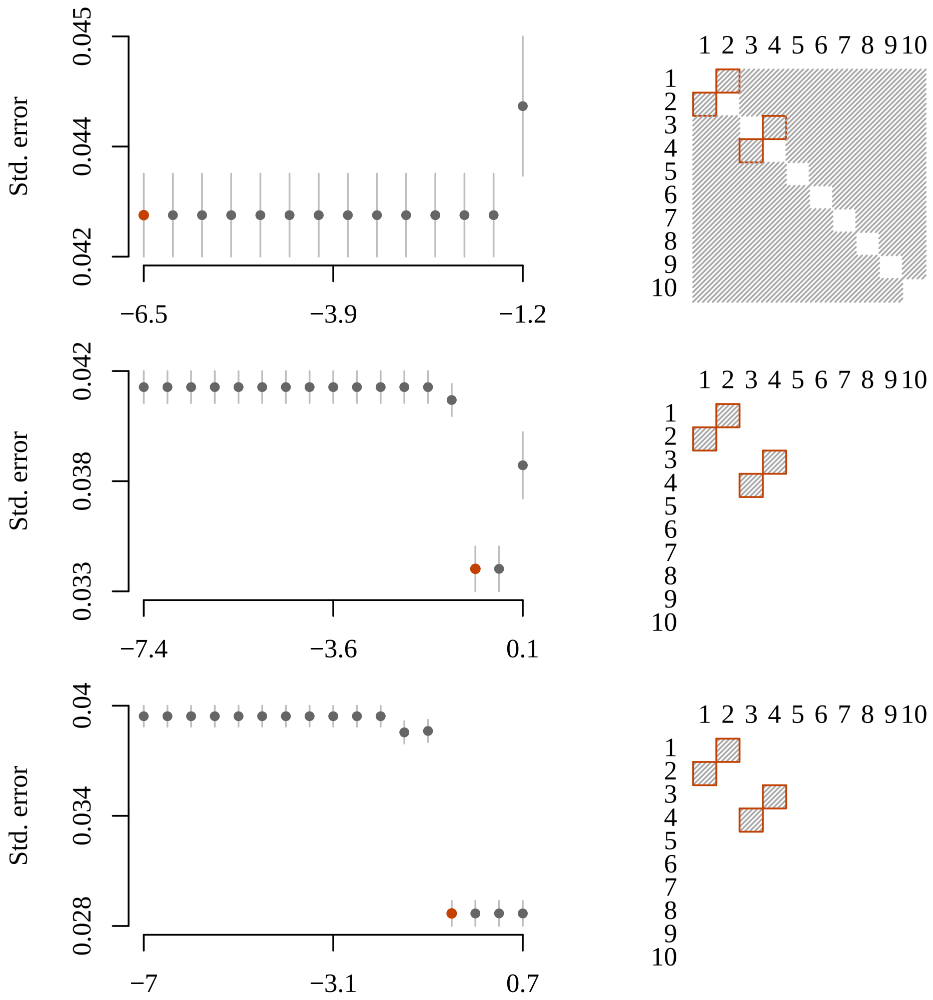


Figure B5: *Diagnostics plots for simulations of multiple cross-trait associations illustrated in Figure 1 (d) with* $H_l^2 = 0.17$ *(top),* $H_l^2 = 0.41$ *(middle), and* $H_l^2 = 0.60$ *(bottom),* $l = 1, \ldots, 4$. Left panel: *Average SE curve with corresponding point-wise approximate confidence intervals as a function of* ρ. Right panel: *adjacency matrix corresponding to* $\hat{\rho}_{\text{min}}$. *Grey squares represent estimated edges, white space represent no edges, and orange borders represent the true edges.*

In Figure B17, Figure B18, and Figure B19 in the supplementary material we show, for simulations with heritability 0.24, 0.39, and 0.58, respectively, visualisations of the estimated adjacency matrices at all values of the sequence of ρ. As expected, the sparsity of the matrix increases in ρ, and when $H_l^2 \geq 0.38$, $l = 1, 2$, for a range of values of ρ it is equivalent to the adjacency matrix representing the true clustering of traits. Furthermore, we observe that the

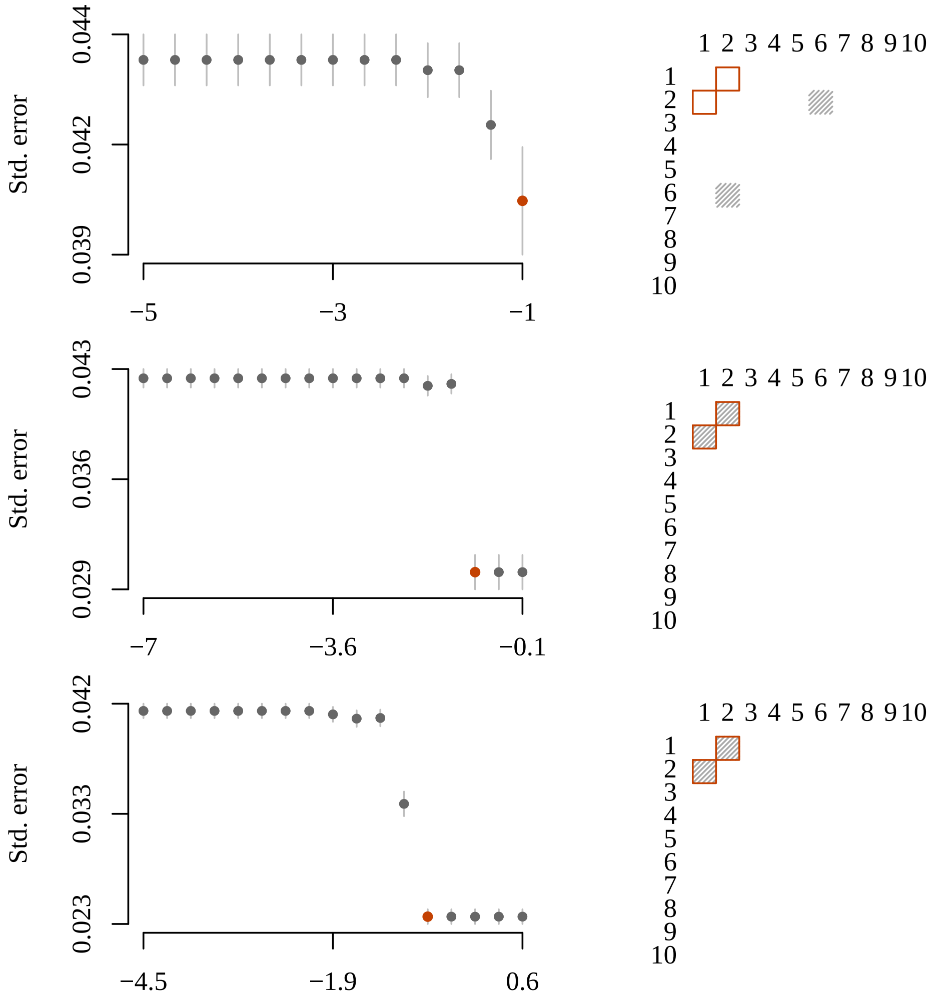


Figure B6: *Diagnostics plots for simulations of overlapping cross-trait associations illustrated in Figure 1 (e) with* $H_l^2 = 0.24$ *(top),* $H_l^2 = 0.39$ *(middle), and* $H_l^2 = 0.58$ *(bottom),* $l = 1, 2$. Left panel: *Average SE curve with corresponding point-wise approximate confidence intervals as a function of* ρ. Right panel: *adjacency matrix corresponding to* $\hat{\rho}_{\min}$. *Grey squares represent estimated edges, white space represent no edges, and orange borders represent the true edges.*

larger the effect size the larger the range at which the clustering is correctly estimated.

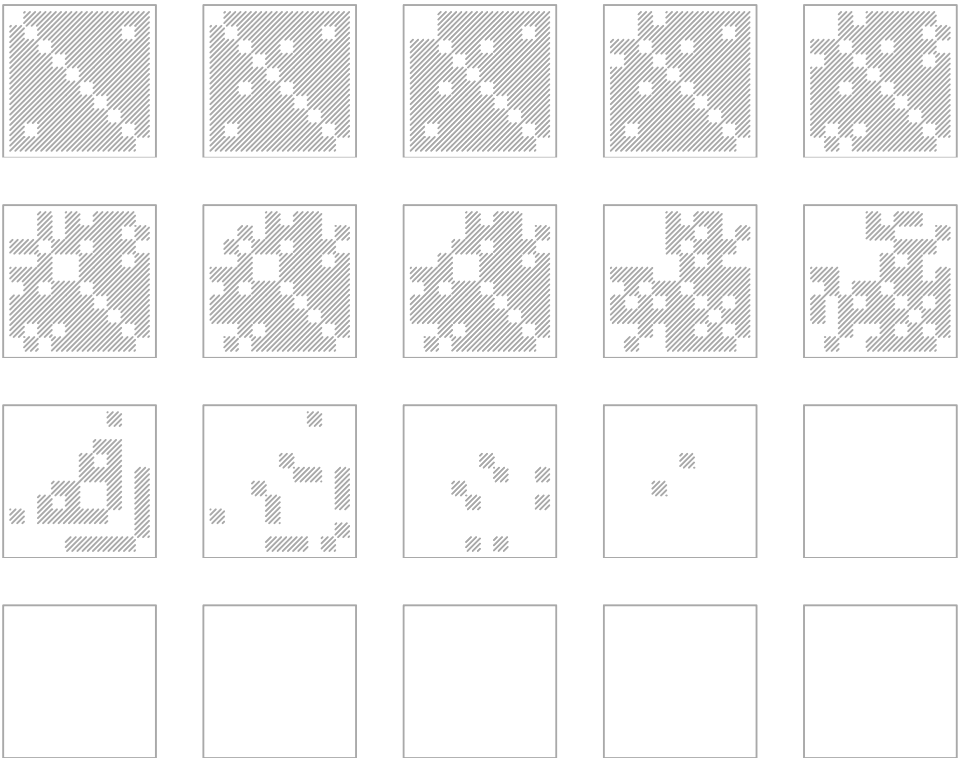


Figure B7: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (a), where the SNPs have no effect on the traits.*

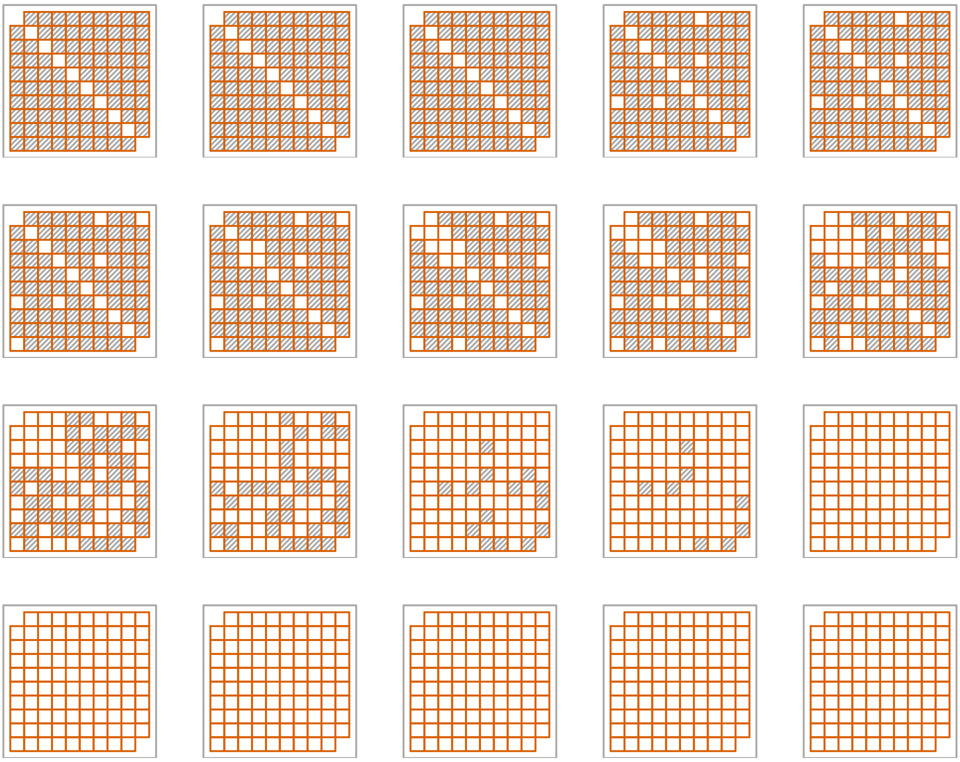


Figure B8: *Visualisation of estimated adjacency matrices at the tried sequence of values of* ρ *for simulations of the scenario illustrated in Figure 1 (b) with broad-sense heritability,* $H_l^2 = 0.01$*, of the traits* $l = 1, \ldots, 10$.

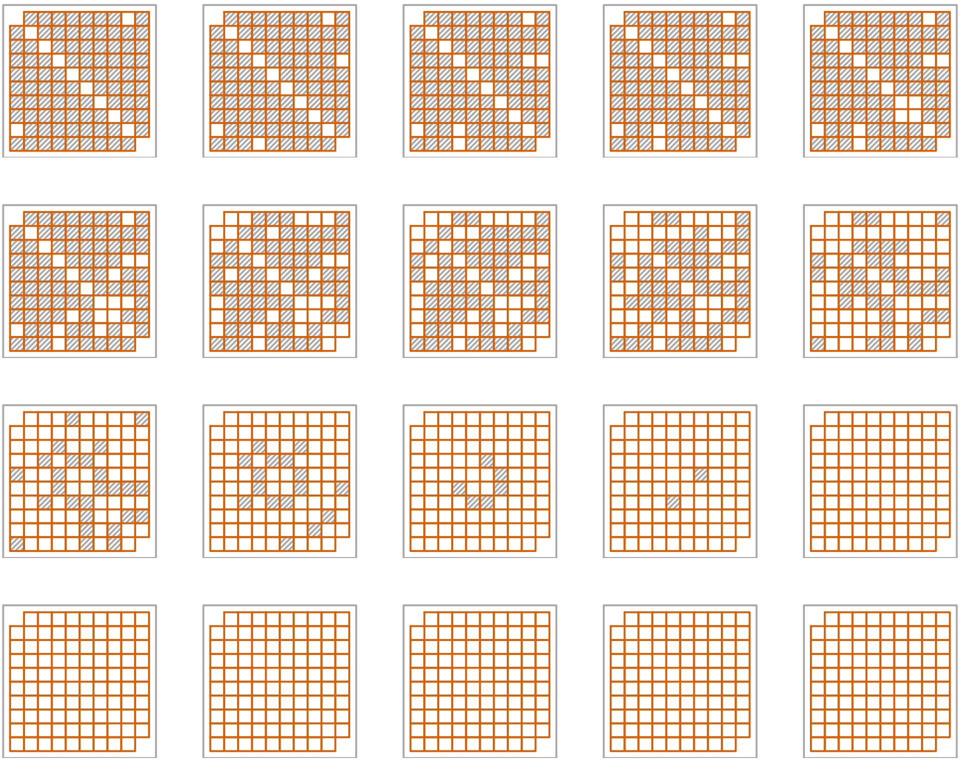


Figure B9: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (b) with broad-sense heritability, $H_l^2 =$* 0.02*, of the traits $l = 1, \ldots, 10$.*

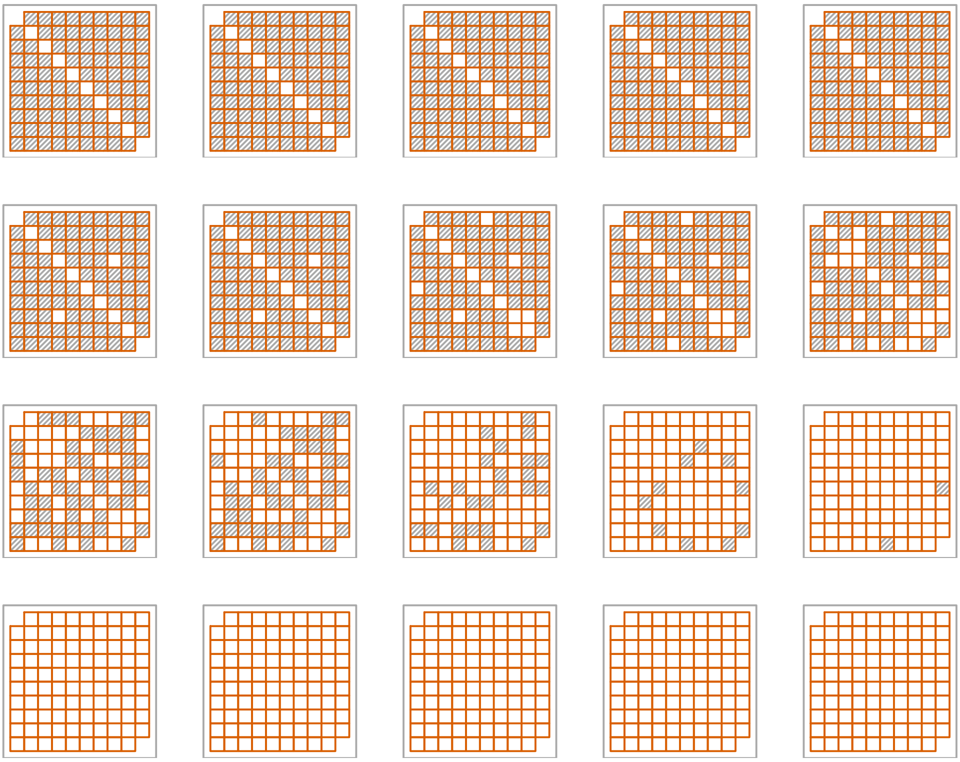


Figure B10: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (b) with broad-sense heritability, $H_l^2 = 0.04$, of the traits $l = 1, \ldots, 10$.*

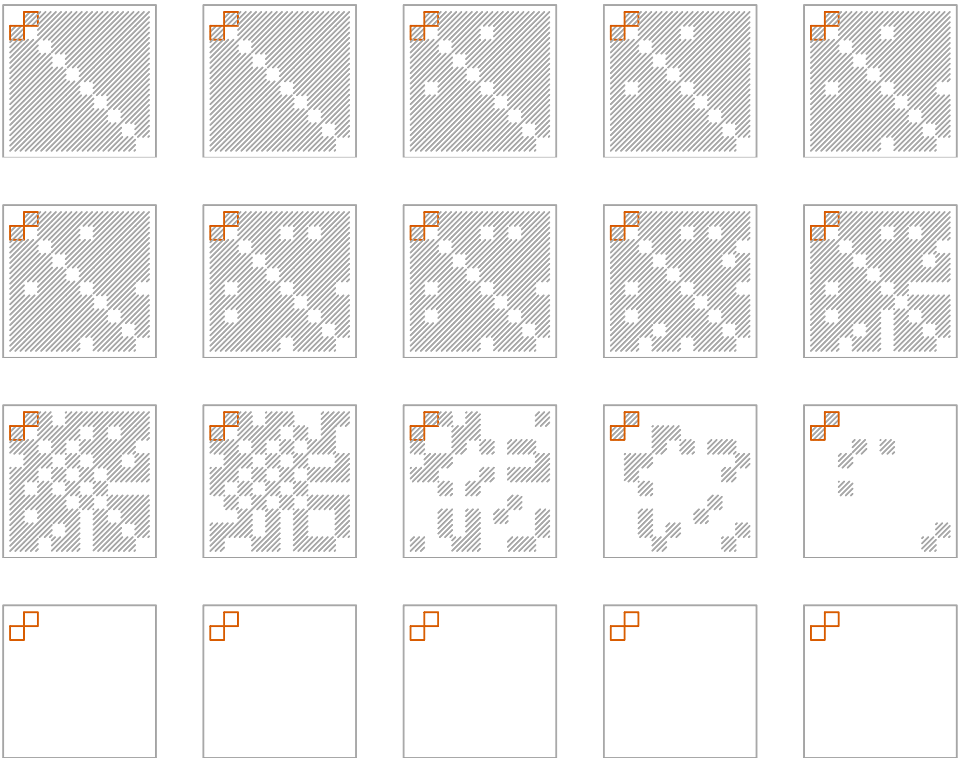


Figure B11: *Visualisation of estimated adjacency matrices at the tried sequence of values of* ρ *for simulations of the scenario illustrated in Figure 1 (c) with broad-sense heritability,* $H_l^2 = 0.17$*, of the traits* $l = 1, 2$*.*

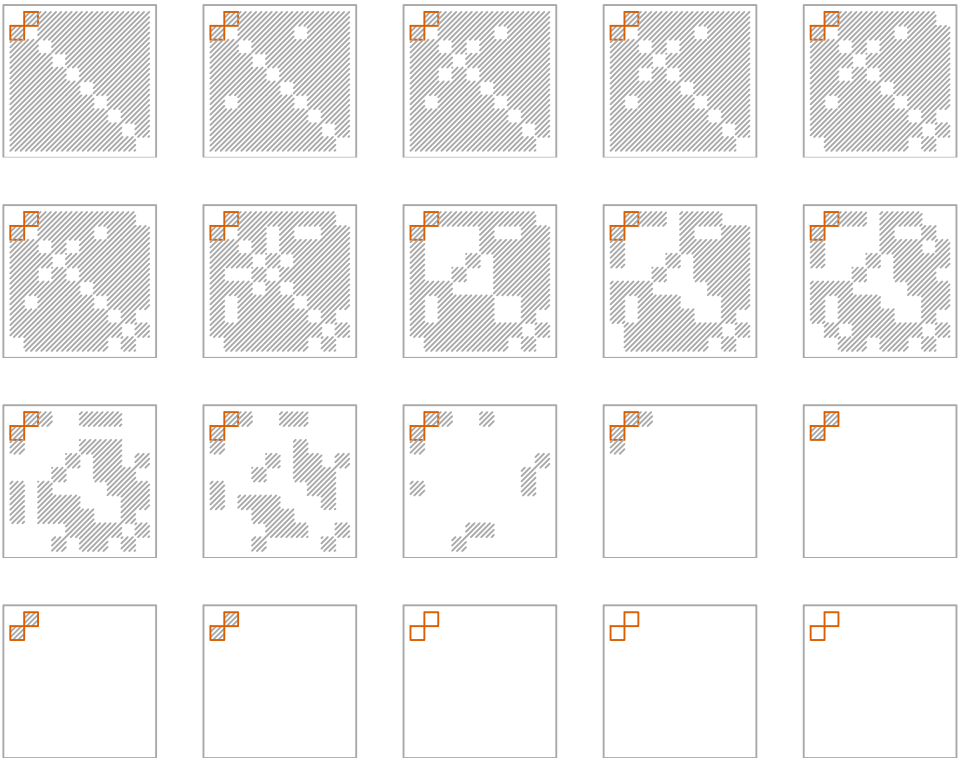


Figure B12: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (c) with broad-sense heritability, $H_l^2 = 0.41$, of the traits $l = 1, 2$.*

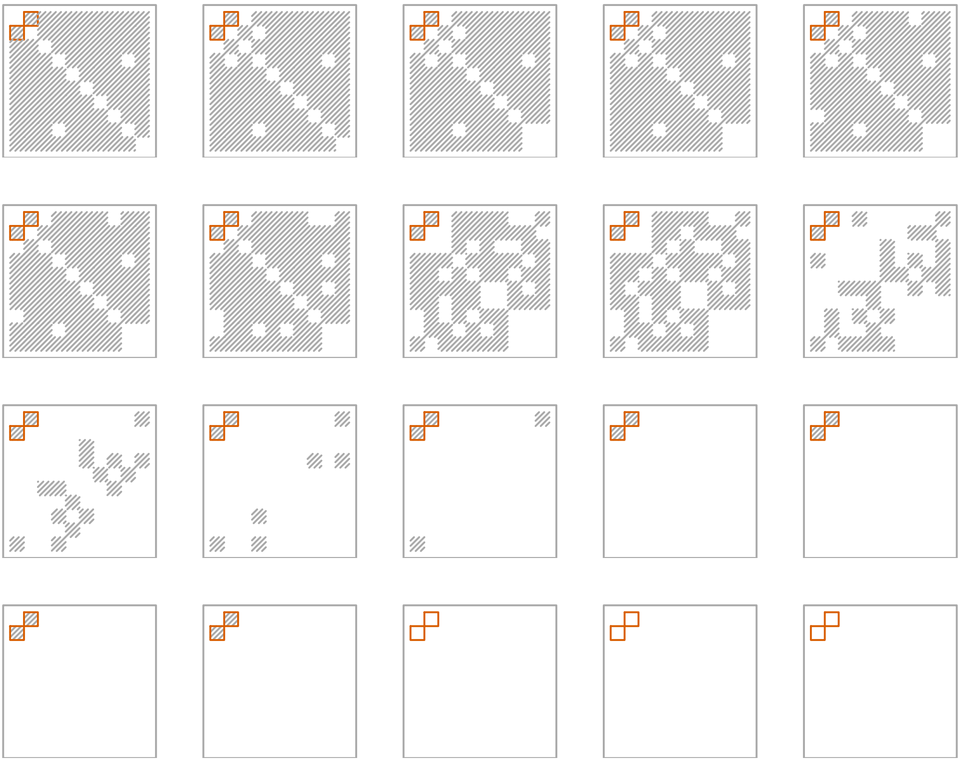


Figure B13: *Visualisation of estimated adjacency matrices at the tried sequence of values of* ρ *for simulations of the scenario illustrated in Figure 1 (c) with broad-sense heritability,* $H_l^2 = 0.60$*, of the traits* $l = 1, 2$.

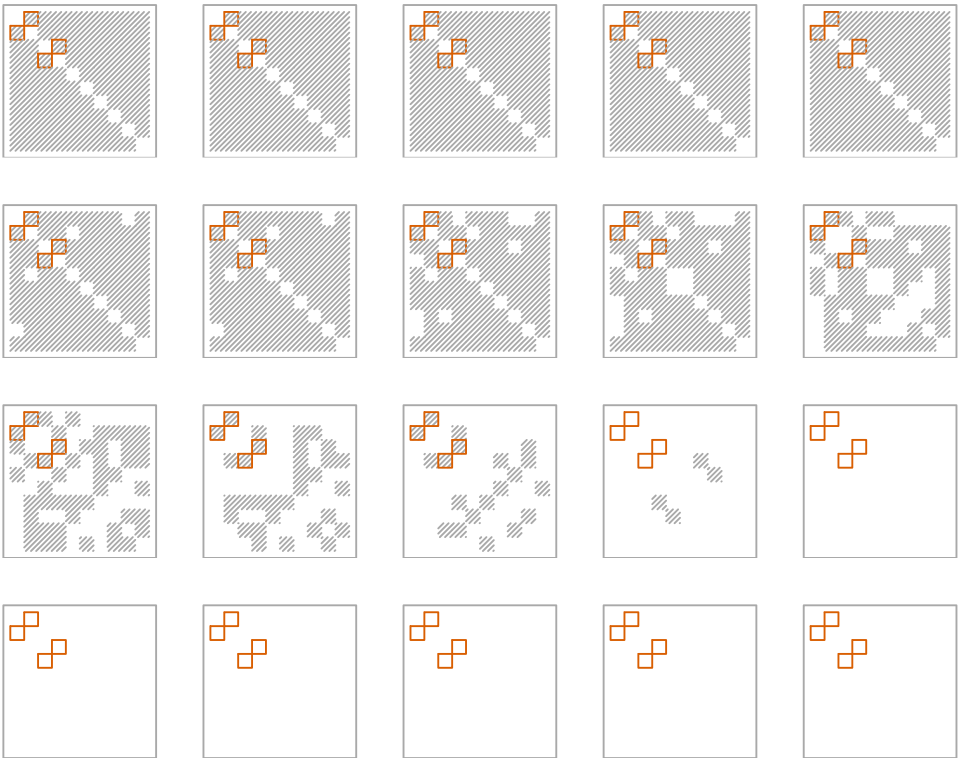


Figure B14: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (d) with broad-sense heritability, $H_l^2 = 0.17$, of the traits $l = 1, \ldots, 4$.*

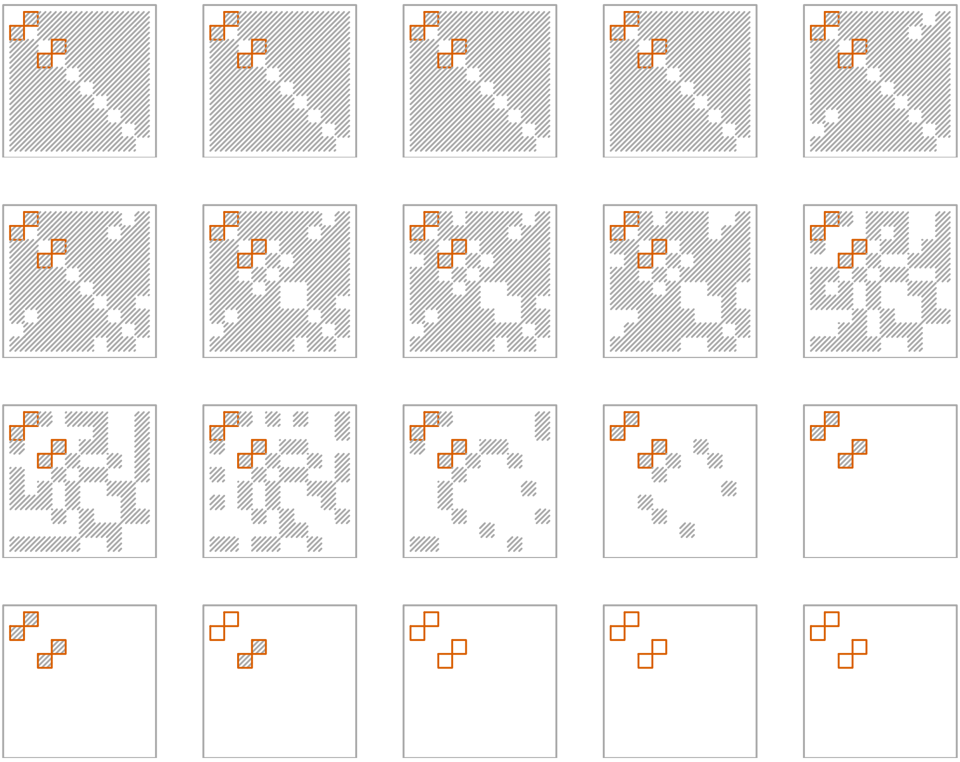


Figure B15: *Visualisation of estimated adjacency matrices at the tried sequence of values of* ρ *for simulations of the scenario illustrated in Figure 1 (d) with broad-sense heritability,* $H_l^2 = 0.41$*, of the traits* $l = 1, \ldots, 4$.

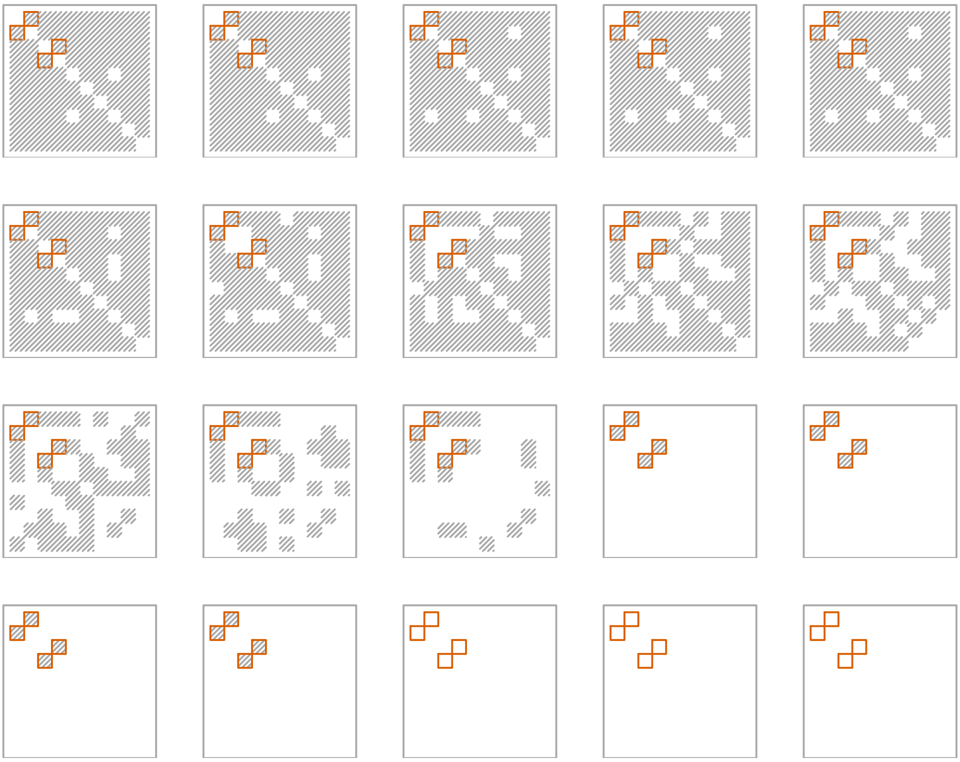


Figure B16: *Visualisation of estimated adjacency matrices at the tried sequence of values of* ρ *for simulations of the scenario illustrated in Figure 1 (d) with broad-sense heritability,* $H_l^2 = 0.60$*, of the traits* $l = 1, \ldots, 4$.

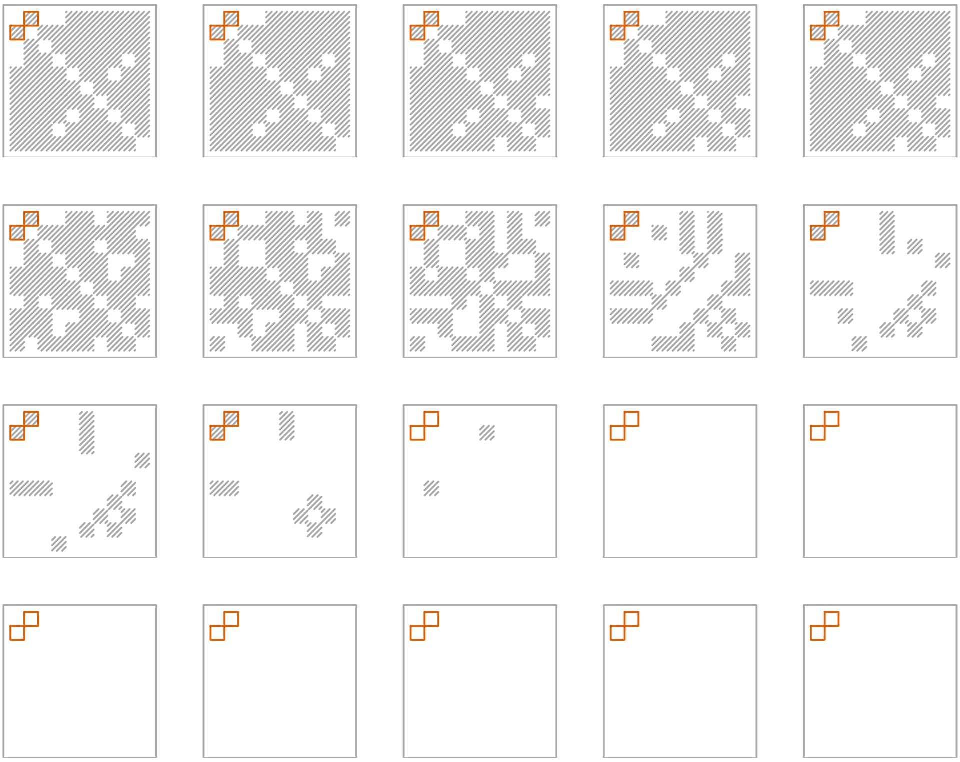


Figure B17: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (e) with broad-sense heritability, $H_l^2 = 0.24$, of the traits $l = 1, 2$.*

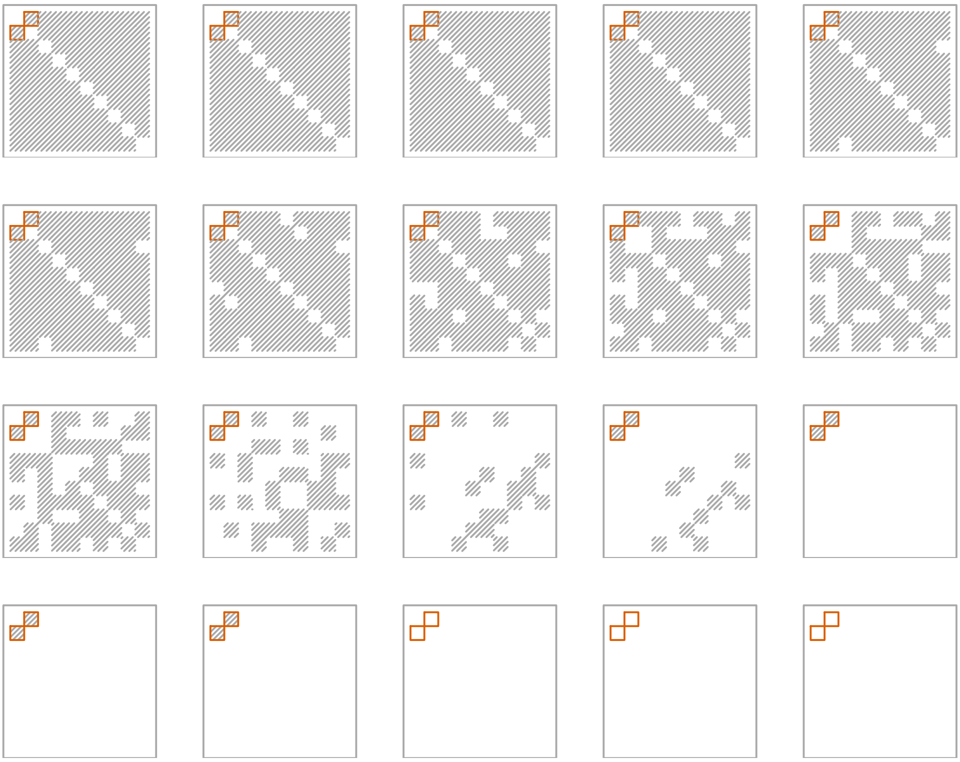


Figure B18: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (e) with broad-sense heritability,* $H_l^2 = 0.39$*, of the traits* $l = 1, 2$*.*

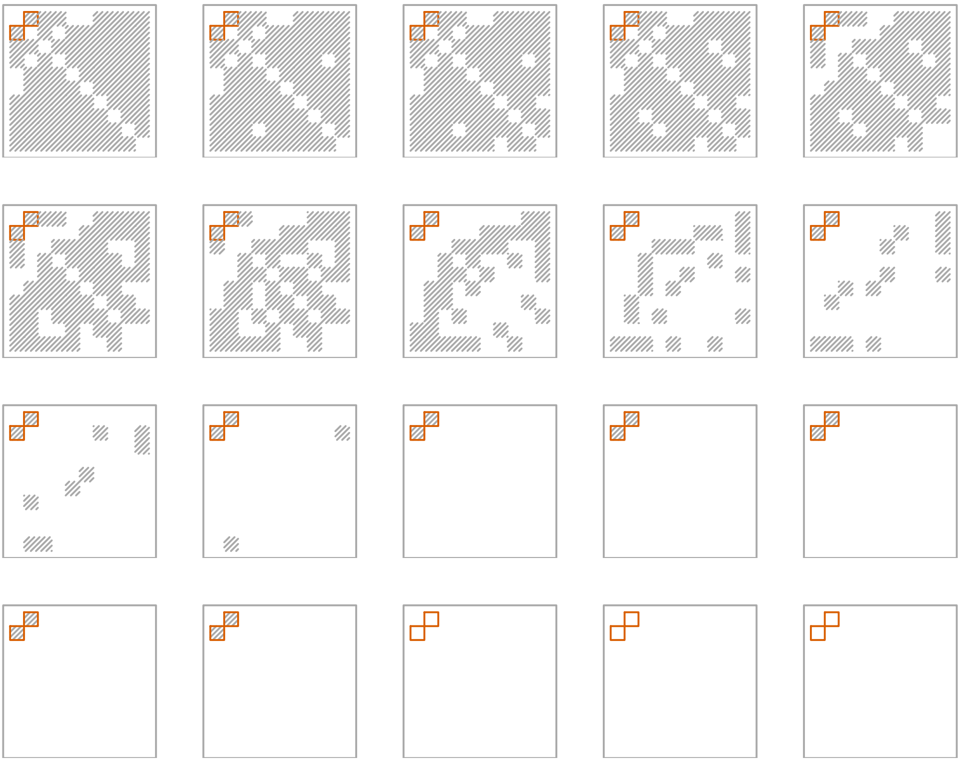


Figure B19: *Visualisation of estimated adjacency matrices at the tried sequence of values of ρ for simulations of the scenario illustrated in Figure 1 (e) with broad-sense heritability, $H_l^2 = 0.58$, of the traits $l = 1, 2$.*

Bibliography

Baltagi, B. H. (2008). *Econometrics*. Berlin: Springer, 4 edition.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Buchardt, A.-S. (2022). geneJAM: Joint regression analysis of multiple traits based on genetic relationships. `https://github.com/abuchardt/geneJAM`.

Crawford, L., Wood, K. C., Zhou, X., and Mukherjee, S. (2018). Bayesian approximate kernel regression with variable selection. *Journal of the American Statistical Association*, 113:1710–1721.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*, 9(4).

Ekstrøm, C. T. (2019). *MESS: Miscellaneous Esoteric Statistical Scripts*. R package version 0.5.5.

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2014). PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9):1466–1468.

Flint, R. M. J. and Valdar, W. (2008). The genetic architecture of complex traits in heterogeneous stock mice. [Online; accessed November 19, 2018].

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2018). *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. R package version 1.10.

Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLOS ONE*, 9(4):1–8.

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.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901.

Mersmann, O. (2019). *microbenchmark: Accurate Timing Functions*. R package version 1.4-7.

Price, A. L., Spencer, C. C. A., and Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences*, 282(1821).

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Schmitz, S., Cherny, S. S., and Fulker, D. W. (1998). Increase in power through multivariate analyses. *Behavior Genetics*, 28(5):357–363.

Solberg, L. C., Valdar, W., Gauguier, D., Nunez, G., Taylor, A., Burnett, S., Arboledas-Hita, C., Hernandez-Pliego, P., Davidson, S., Burns, P., Bhattacharya, S., Hough, T., Higgs, D., Klenerman, P., Cookson, W. O., Zhang, Y., Deacon, R. M., Rawlins, J. N. P., Mott, R., and Flint, J. (2006). A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mammalian Genome*, 17:129–146.

Valdar, W., Solberg, L., Gauguier, D., Heyes, S., Klenerman, P., Cookson, W., Taylor, M., Rawlins, J., Mott, R., and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38:879–87.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., and Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10):1068–1087.

Yang, J., Benyamin, B., B. P. McEvoy a, d. S. G., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–4.

Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409.

Manuscript III

Identifying Hierarchical Interactions via Multivariate Sparse Group Lasso Regularisation

ANN-SOPHIE BUCHARDT AND CLAUS THORN EKSTRØM

Identifying Hierarchical Interactions via Multivariate Sparse Group Lasso Regularisation

ANN-SOPHIE BUCHARDT*, CLAUS THORN EKSTRØM

Section of Biostatistics, Department of Public Health, University of Copenhagen,

Øster Farimagsgade 5, 1014 København K, Denmark

Corresponding author: *asbu@sund.ku.dk

Abstract

Penalised regression models such as lasso are powerful methods, which use sparsity to do feature selection in situations where the number of features measured is larger than the number of observations. This is typical for genomic or multi-omics data, which, in addition, may contain multiple, simultaneously measured traits. When analysing such high-dimensional and multivariate data we are concerned not only with identifying relevant features but also with identifying more complex relationships among the features such as interactions, e.g. gene-gene or gene-environment interactions. Furthermore, we are interested in a joint analysis of all traits to increase power.

It is, however, not clear exactly how to consistently include interacting features in multivariate penalised regression models.

We consider methods for identifying pairwise interactions in a multivariate linear regression model under different assumptions of hierarchy on the interactions. We assume that there is an unknown subset of the features and interactions which are relevant, but this subset is not preserved across the components of the outcome. As a special case, we consider the assumption that there is an unknown subset of the features and interactions which are relevant, and that this same subset is preserved across all components of the outcome.

We approach the problem by using a two-step procedure. In the first step lasso includes only the main effects and selects the most promising. Next, a special case of the multivariate sparse group lasso includes main effects and interactions according to hierarchical restrictions and selects the most promising terms. The approach is motivated by modelling pairwise interactions for qualitative variables and experimenting with explicitly applying penalties on the main effects and interactions, thereby obtaining interpretable models.

Keywords: multivariate outcome; regularised regression; lasso; interactions; hierarchical sparsity

1 Introduction

In Buchardt and Ekstrøm (2021), we proposed a method for identifying pairwise interactions in a univariate linear regression model under different assumptions of hierarchy. Our objective in this paper is to extend the method to the multiple outcome setting.

The existence of, e.g., gene-gene or gene-environment interactions has gained attention in genetic studies and empirical evidence shows that such interactions might be an important genetic component underlying complex traits and diseases (Lee et al., 2018; Schrode et al., 2018; Dong et al., 2017). However, the introduction of interactions into a model rapidly increases the complexity of said model (Hu et al., 2014; Cornelis et al., 2012). If the number of measured features is p, the total number of possible pairwise interactions is $\binom{p}{2} = \frac{1}{2}p(p-1)$. Thus, when using 500K single nucleotide polymorphism microarrays, say, there are approximately 125 billion pairwise interactions. Fitting a regression model to such data is computationally challenging, potentially extremely time consuming, and mathematically distorts the relation between the number of features, p, and the number of observations, N. Furthermore, joint association analysis of multiple, simultaneously measured quantitative traits in a genome-wide association study (GWAS), i.e., a multivariate GWAS, has gained attention in genetic studies as it offers several advantages over analysing each trait in a separate GWAS (Galesloot et al., 2014). While single-trait GWASs have found numerous novel loci associated with complex diseases (MacArthur et al., 2016), traits are often correlated, for example due to pleiotropic genes and, therefore, a joint modelling approach can be used to increase precision and power (Schmitz et al., 1998).

The objective of this paper is to propose a computationally efficient method for selecting pairwise interactions in a multivariate regression model which involves only a subset of the features and interactions. We consider a general setup where we assume that there is an unknown subset of the features which are relevant, but this subset is *not necessarily preserved* across all components of the outcome. By "relevant" we refer to a feature which affects at least one outcome. If a feature affects more than one outcome, we allow for the effect size to be different across the outcomes. We make the consequential assumption that that there is an unknown subset of interactions which are relevant, but this subset is not preserved across all outcomes.

When to include interactions in a multivariate (regularised) regression model is, however, not straightforward. Interpretations of traditional regression analyses do not encourage including an interaction in a model without the corresponding main effects, since main effects can be viewed as deviations from the mean and interactions as deviations from the main effects (Bien and Tibshirani, 2014). Furthermore, it can be argued that large main effects are more likely to lead to important interactions than small (Cox, 1984). Finally, including "too many" interaction terms may unnecessarily complicate the model and its interpretation and be computationally infeasible. This motivates the hierarchical restrictions that an interaction is included in a model only if one or both constituent main effects are marginally important.

In Section 2 we introduce the pairwise interaction model, the concept of hierarchical sparsity, and our framework and two-step procedure for finding pairwise interactions. We propose a method for identifying a subset of relevant features, such that imposing hierarchical restrictions in a stepwise manner, sufficiently reduce the number of interactions to be included in a multivariate regularised regression model. We study our method on simulated data, in Section 3. The two-step procedure is implemented using the R package `MSGLasso`; the R code is part of our package `ilasso` which is available online, see Buchardt (2022). We conclude with a discussion and recommendations in Section 4.

2 Methods

In this section we present a two-step regularised regression procedure which enables the identification of pairwise interactions in a multiple outcome setting.

2.1 Pairwise interaction model

We assume that we have N observations of the multivariate outcome $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q) \in \mathbb{R}^{N \times q}$ and p associated features $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ with pairwise interactions.

If the features are categorical, they are assumed to be represented by dummy variables, and the pairwise interaction term is formed by multiplying the corresponding variables. If the features are quantitative, $\mathbf{X} \in \mathbb{R}^{N \times p}$, the pairwise interaction term is formed by multiplying the two constituent features. For computational reasons we recommend centring each quantitative feature before the optimisation problem is solved (Aiken and West, 1991). If the features are not measured in the same units, we also recommend scaling each feature. Otherwise the lasso solution depends on the scale since lasso puts constraints on the size of the coefficients for feature. The methods developed in this paper are applicable to pairwise interactions between categorical features, between quantitative features, and between categorical and quantitative features. It should be noted that hypothesis tests are calculated similarly for both the cases of pairwise interactions between categorical features and between categorical and quantitative features, and interpretations are fairly clear (Ekstrøm and Sørensen, 2014). In the case of interactions between quantitative features, testing and interpretation is more complex (Aiken and West, 1991).

For the sake of simplicity we assume that the outcome values are centred before the optimisation problem is solved, that is, $\frac{1}{N}\sum_{i=1}^{N} y_{il} = 0$, for all $l = 1, \dots, q$, and the intercept term can be omitted from the model. In multivariate linear regression models, this condition is not a restriction, since given an optimal solution, $(\hat{\mathbf{b}}_0, \hat{\mathbf{B}})$, obtained from the centred data, an optimal solution, $(\tilde{\mathbf{b}}_0, \tilde{\mathbf{B}})$, for the uncentred data is easily recovered: $\tilde{\mathbf{B}} = \hat{\mathbf{B}}$ and $\tilde{\mathbf{b}}_0 = \bar{\mathbf{y}} - \sum_{j=1}^{p} \bar{x}_j \tilde{b}_j$, where $\bar{\mathbf{y}}$ and $\bar{\mathbf{x}}_j$, $j = 1, \dots, p$, are the original column means. Furthermore, if the outcome components are not measured in the same units, we also recommend scaling each component. Otherwise the lasso solution depends on the scale.

Thus, for each sample $i = 1, \dots, N$ and each outcome $l = 1, \dots, q$ the pure main effect model takes the form

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + e_{il}, \tag{1}$$

and the pairwise interaction model takes the form

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + \sum_{k=1}^{p} \sum_{j<k} x_{ij} x_{ik} \Theta_{jkl} + e_{il}, \tag{2}$$

where $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p)^\top \in \mathbb{R}^{p \times q}$ are unknown regression parameters for the main effects, $\boldsymbol{\Theta} \in \mathbb{R}^{p \times p \times q}$ are unknown parameters for the pairwise interactions, and $\boldsymbol{E} \in \mathbb{R}^{N \times q}$ are Gaussian errors. For each $l = 1, \dots, q$ we let $\boldsymbol{\Theta}_l \in \mathbb{R}^{p \times p}$ represent a symmetric matrix, i.e. $\boldsymbol{\Theta}_l = \boldsymbol{\Theta}_l^\top$, such that the strict inequality in the interaction summation precludes over-parametrisation arising from the inclusion of the same effect twice, for example, including both $x_{ij}x_{ik}$ and $x_{ik}x_{ij}$.

Based on the model (2) we aim to select a subset of the p main effects and the $\frac{1}{2}p(p-1)$ interactions – we refer to the variables in this subset as the *relevant* variables. In pursuit of this goal we establish the notions of *hierarchy* and *sparsity* in the next sections.

2.2 Hierarchy

In this section we present model restrictions in a form which makes it is possible to specify a regularised regression procedure which produces sparse interaction models that honour these restrictions. The hierarchical constraint of the pairwise interactions is defined for all outcome components simultaneously, that is, exactly one of the following hierarchical, anti-hierarchical, and non-hierarchical restrictions is assumed for all $l = 1, \dots, q$:

Strong hierarchy There are interactions only among pairs of non-zero main effects,

$$\text{H}_\text{S}: \quad \Theta_{jkl} \neq 0 \quad \Rightarrow \quad b_{jl} \neq 0 \text{ and } b_{kl} \neq 0.$$

Weak hierarchy Each interaction has at least one of its main effects present,

$$\mathrm{H_W}: \quad \Theta_{jkl} \neq 0 \quad \Rightarrow \quad b_{jl} \neq 0 \text{ or } b_{kl} \neq 0.$$

Anti-hierarchy Interactions are only among pairs of main effects that are not present,

$$\mathrm{H_A}: \quad \Theta_{jkl} \neq 0 \quad \Rightarrow \quad b_{jl} = 0 \text{ and } b_{kl} = 0.$$

Pure interactions There are no main effects present, only interactions,

$$\mathrm{H_I}: \quad b_{jl} = 0 \quad \forall j = 1, \ldots, p.$$

Pure main effects There are no interactions present, only main effects,

$$\mathrm{H_M}: \quad \Theta_{jkl} = 0 \quad \forall j, k = 1, \ldots, p.$$

No hierarchy There are no restrictions to the presence of main effects and interactions, $\mathrm{H_N}$.

Next, we introduce methods for estimating the parameters in pairwise interaction models, which honour one of the hierarchical, anti-hierarchical, or non-hierarchical restrictions.

2.3 Sparsity

The *ordinary least squares* (OLS) method for estimation is not suitable when $p > N$, in which case we prefer fitting a *sparse model*, which refer to a model which involves only a subset of the p features. In order for us to obtain such a subset we can regularise the estimation process.

The multivariate pure main effect model (1) can be seen as a coupled collection of q standard regression problems in $\mathbb{R}^p$, each sharing the same features, in which column l, $\mathbf{b}_l \in \mathbb{R}^q$, of $\mathbf{B}$ is the coefficient vector for the lth problem. Inspired by the methodology outlined for example in the book Hastie et al. (2015) we define "groups" by the rows $\mathbf{b}_j \in \mathbb{R}^q$, $j = 1, \ldots, p$, from the full matrix of parameters $\mathbf{B} \in \mathbb{R}^{p \times q}$, and the problem can be viewed as a special case of the general group lasso (Yuan and Lin, 2006), where we construct p groups of equal size q. Under the assumption that the same set of features are relevant for the modelling across all q components of the outcome variable, that is, the collection of q outcome components are regressed on the same $N \times p$ matrix of features, we approach the multivariateness by solving q separate lasso problems, one for each column of the $p \times q$ regression matrix $\mathbf{B}$. Thus, for the multivariate pure main effect model (1), the objective of the multivariate lasso is to solve the regularised least-squares problem

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \sum_{i=1}^{N} \sum_{l=1}^{q} \left(y_{il} - \sum_{j=1}^{p} x_{ij} b_{jl} \right)^2 + \lambda \sum_{j=1}^{p} \|\mathbf{b}_j\|_2 \right\}, \tag{3}$$

for some regularisation parameter $\lambda > 0$. This way, the collection of q outcome components are regressed on the same $N \times p$ matrix of features and the q linear regression problems are coupled together via the regularisation constraint which ensures that all features are equally regularised across outcome components. It should be noted, that the factor $\frac{1}{2Nq}$ appearing in (3) makes no difference in the optimum; it corresponds to a standardisation of λ and makes the regularisation comparable for different sample sizes.

The nature of the group lasso implies that when a group is selected by the procedure, all the coefficients in that group will be non-zero. For the multivariate setup, this implies that when a feature is selected by the procedure, the coefficients for that feature will be non-zero for the entire collection of outcome components. It may, however, be more appropriate to have sparsity with respect to the collection as well as components within the collection, since all features are not necessarily relevant for the entire collection of outcome components.

The additional within-group sparsity is achieved with the versatile multivariate sparse group lasso (MSGLasso) feature selection and estimation method proposed by Li et al. (2015). The method is carried out through a regularised multivariate multiple linear regression model and allows for an *arbitrary* group structure for the regression coefficient matrix. That is, the MSGLasso allows for different regularisation of groups as well as features within groups via adaptive lasso tuning parameters.

Specifically, the MSGLasso assumes that $\mathbf{B}$ contains G groups, and each group, denoted as $\mathbf{B}_g$ where $g \in \{1, \dots, G\}$, is a subset of two or more elements in $\mathbf{B}$. The group structure is denoted by $\mathcal{G} = \{\mathbf{B}_1, \dots, \mathbf{B}_G\}$. Depending on the context $\mathbf{B}$ and $\mathbf{B}_g$ denote either the set of all their elements or the numerical values of all their elements. Note that the union of all groups in $\mathcal{G}$ does not need to contain all the elements of $\mathbf{B}$, in other words, some b_{jl} may not belong to any group.

For an arbitrary group structure $\mathcal{G}$ with G groups, $\sum_{g=1}^{G} \|\mathbf{G}_g\|_2$ is the total sum of ℓ_2 norms of every group in $\mathcal{G}$, where $\|\mathbf{G}_g\|_2^2 = \sum_{b_{jl} \in \mathbf{B}_g} b_{jl}^2$.

For an arbitrary group structure $\mathcal{G}$ on $\mathbf{B}$, we denote $\{g : \mathbf{B}_g \in \mathcal{G}\}$ by $\{g \in \mathcal{G}\}$ to simplify the notation as long as it does not cause any confusion. For $j = 1, \dots, p$ and $l = 1, \dots, q$, let $\lambda_{jl} \geq 0$ be the adaptive lasso tuning parameter for b_{jl} with $\lambda_{jl} = 0$ if b_{jl} is not regularised. Let $\lambda_g \geq 0$ be the adaptive tuning parameter for group $\mathbf{B}_g \in \mathcal{G}$ with $\lambda_g = 0$ if group $\mathbf{B}_g$ is not regularised. The objective of the MSGLasso is to solve the following regularised optimisation problem for a general regularised multivariate multiple linear regression:

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \sum_{j=1}^{p} \sum_{l=1}^{q} \lambda_{jl} |b_{jl}| + \sum_{g \in \mathcal{G}} \lambda_g \|\mathbf{B}_g\|_2 \right\}, \tag{4}$$

where the ℓ_2 regularisation term aims to shrink unimportant groups to zero and the ℓ_1 regularisation term aims to shrink unimportant entries within an important group to zero.

We exclude the trivial case that $\lambda_g = 0$ for all $g \in \mathcal{G}$ and $\lambda_{jl} = 0$ for all $j = 1, \dots, p$ and $l = 1, \dots, q$. We observe that the case that $\lambda_{jl} = 0$ for all $j = 1, \dots, p$ and $l = 1, \dots, q$ and

$\lambda_g = \lambda$ for all $g \in \mathcal{G}$ with $\mathcal{G} = \{1, \ldots, p\}$ corresponds to the multivariate lasso (3) where all features are equally regularised across outcome components.

While the group structures considered in Li et al. (2015) are pre-defined by biological functions, such as gene or pathways, the group structure considered here is very simple and determined with the goal of including interactions and imposing hierarchical restrictions in mind. Therefore, we define groups (in the following referred to as collections) by the rows $\mathbf{b}_j \in \mathbb{R}^q$, $j = 1, \ldots, p$, and the problem reduces to

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \sum_{j=1}^{p} \sum_{l=1}^{q} \lambda_{jl} |b_{jl}| + \sum_{j=1}^{p} \lambda_j \|\mathbf{b}_j\|_2 \right\}.$$

In order for us to include interactions, we define collections of interactions by the rows $\boldsymbol{\Theta}_{jk} \in \mathbb{R}^q$, $j, k \in \{1, \ldots, p\}$ of the full stack of matrices of parameters $\boldsymbol{\Theta} \in \mathbb{R}^{p \times q \times q}$. Now, fitting a MSGLasso regularisation on the joint set of main effects and interactions corresponds to the assumption of no hierarchy, $\mathrm{H_N}$. That is, for the multivariate pairwise interaction model (2), the MSGLasso problem takes the form

$$\begin{aligned} \underset{\mathbf{B} \in \mathbb{R}^{p \times q}, \boldsymbol{\Theta} \in \mathbb{R}^{p \times p \times q}}{\text{minimise}} & \left\{ L(\mathbf{B}, \boldsymbol{\Theta}) + \sum_{j=1}^{p} \sum_{l=1}^{q} \lambda_{jl} |b_{jl}| + \sum_{j=1}^{p} \lambda_j \|\mathbf{b}_j\|_2 \right. \\ & \left. + \sum_{k=1}^{p} \sum_{j=1}^{k-1} \sum_{l=1}^{q} \tau_{jkl} |\Theta_{jkl}| + \sum_{j=1}^{p} \tau_j \|\boldsymbol{\Theta}_j\|_2 \right\}, \end{aligned}$$

where $L(\mathbf{B}, \boldsymbol{\Theta})$ denotes the loss function

$$L(\mathbf{B}, \boldsymbol{\Theta}) = \frac{1}{2Nq} \sum_{i=1}^{N} \sum_{l=1}^{q} \left(y_{il} - \sum_{j=1}^{p} x_{ij} b_{jl} - \sum_{k=1}^{p} \sum_{j=1}^{k-1} x_{ij} x_{ik} \Theta_{jkl} \right)^2.$$

This way, when $\lambda_j = 0$, the jth regularisation term disappears, that is, feature j is not regularised for any outcome component but always included in the model. When $\lambda_j = \infty$ feature j is always excluded for all outcome components, and for all $j = 1, 2, \ldots, p$ for which λ_j are equal to the same constant value, the corresponding features are equally regularised across the outcome. Similarly, when $\lambda_{jl} = 0$, the jlth regularisation term disappears, that is, feature j is not regularised for the lth outcome component but always included in the model. When $\lambda_{jl} = \infty$ feature j is always excluded for the lth outcome component, and for all $j = 1, 2, \ldots, p$ and $l = 1, \ldots, q$ for which λ_{jl} are equal to the same constant value, the corresponding features are equally regularised for the corresponding outcome components. The interpretation of the regularisation parameters τ_j and τ_{jkl} on the interactions is analogous to that of λ_j and λ_{jl}.

We observe that $\lambda_{jl} > 0$ (and $\tau_{jkl} > 0$) ensures the more general assumption that there is an unknown subset of the features (and interactions) which are relevant, but this subset is *not* *necessarily preserved* across the components of the outcome. On the other hand, $\lambda_{jl} = 0$ (and $\tau_{jkl} = 0$) corresponds to the assumption that there is an unknown subset of the features (and interactions) which are relevant, and this same subset is *preserved* across all q components of the outcome. Here, the term "relevant" refers to a feature which affects all outcome component, but, possibly, with different effect sizes across the components. This scenario also makes the consequential assumption that there is an unknown subset of interactions which are relevant, and this same subset is preserved across all components of the outcome. The general setup is, obviously, most flexible, and it is, possibly, the most realistic. The special case scenario is, however, suitable when all the outcome components are suspected to be predicted from roughly the same set of features.

2.4 Two-step procedure

Solving the strong and weak hierarchical lasso outlined above is very time consuming, and in this section we extend the two-step procedure of Buchardt and Ekstrøm (2021) from the univariate outcome setting to the multivariate outcome setting. That is, we propose an efficient method for selecting relevant features and pairwise interactions according to one of the hierarchical restrictions, $\mathrm{H_S}$ or $\mathrm{H_W}$, introduced in Section 2.2. While the method allows for an anti-hierarchical structure, it is not computationally tractable, often conceptually unrealistic, and we make no further mention of it in this paper.

As the first step in our procedure, we include only the main effects and apply the MSGLasso for variable selection. As the second step in the procedure we include main effects and interactions in accordance with step one and subject to one of the restrictions $\mathrm{H_S}$or $\mathrm{H_W}$, and we apply a variation of the MSGLasso where the regularisation parameters depend on the assumed hierarchical restriction.

Define by

$$\mathcal{S}_e = \{j \in \{1,\dots,p\}, l \in \{1,\dots,q\} : |b_{jl}| \neq 0\},$$

the index set of non-zero elements in $\mathbf{B}$, by

$$\mathcal{S}_g = \left\{j \in \{1,\dots,p\} : \|\mathbf{b}_j\|_2 \neq 0\right\},$$

the index set of non-zero rows in $\mathbf{B}$, by

$$\mathcal{R}_e = \left\{(j,k) \in \{1,\dots,p\}^2, l \in \{1,\dots,q\} : |\Theta_{jkl}| \neq 0\right\},$$

the index set of non-zero elements in $\mathbf{\Theta}$, and by

$$\mathcal{R}_g = \left\{j \in \{1,\dots,p\} : \|\mathbf{\Theta}_j\|_2 \neq 0\right\},$$

the index set of non-zero rows in $\mathbf{\Theta}$.

Define by $\mathcal{M}_e$ and $\mathcal{I}_e$ the index sets of main effects and interactions, respectively, to be included in the model subject to one of the hierarchical restrictions, $\mathrm{H_S}$ or $\mathrm{H_W}$, and specific to outcome components. Similarly, define by $\mathcal{M}_g$ and $\mathcal{I}_g$ the index sets of the main effects and interactions, respectively, to be included in the model for all outcome components subject to one of the hierarchical restrictions, $\mathrm{H_S}$ or $\mathrm{H_W}$. Specifically, when the model is subject to the restriction of strong hierarchy $\mathcal{M}_e = \mathcal{S}_e$, and

$$\mathcal{I}_e = \left\{(j,k) \in \{1,\ldots,p\}^2, l \in \{1,\ldots,q\} : (j,l) \in \mathcal{S}_e \wedge (k,l) \in \mathcal{S}_e\right\};$$

$\mathcal{M}_g = \mathcal{S}_g$, and

$$\mathcal{I}_g = \left(\{(j,k) \in \{1,\ldots,p\}^2 : j \in \mathcal{S}_g \wedge k \in \mathcal{S}_g\right\}.$$

When the model is subject to the restriction of weak hierarchy $\mathcal{M}_e = \mathcal{S}_e$, and

$$\mathcal{I}_e = \left\{(j,k) \in \{1,\ldots,p\}^2, l \in \{1,\ldots,q\} : (j,l) \in \mathcal{S}_e \vee (k,l) \in \mathcal{S}_e\right\};$$

$\mathcal{M}_g = \mathcal{S}_g$, and

$$\mathcal{I}_g = \left\{(j,k) \in \{1,\ldots,p\}^2 : j \in \mathcal{S}_g \vee k \in \mathcal{S}_g\right\}.$$

The procedure is defined as follows:

Step 1 We assume the pure main effect model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{E},$$

for which the MSGlasso estimates are determined by solving the optimisation problem

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\text{minimise}} \left\{ \frac{1}{2Nq} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \lambda_1 \sum_{j=1}^{p} \left(\sum_{l=1}^{q} \alpha_1 |b_{jl}| + (1-\alpha_1) \|\mathbf{b}_j\|_2 \right) \right\},$$

where we have defined $\lambda_{jl} = \lambda_1\alpha_1$ and $\lambda_j = \lambda_1(1-\alpha_1)$ with $\alpha_1 \in [0,1]$. Now, $\mathcal{S}_e$ is estimated by

$$\hat{\mathcal{S}}_e^{(\lambda_1,\alpha_1)} = \left\{ j \in \{1,\ldots,p\}, l \in \{1,\ldots,q\} : \left|\hat{b}_{jl}^{(\lambda_1,\alpha_1)}\right| \neq 0 \right\},$$

and $\mathcal{S}_g$ is estimated by

$$\hat{\mathcal{S}}_g^{(\lambda_1,\alpha_1)} = \left\{ j \in \{1,\ldots,p\} : \left\|\hat{\mathbf{b}}_j^{(\lambda_1,\alpha_1)}\right\|_2 \neq 0 \right\}.$$

$\mathcal{M}_e$, $\mathcal{I}_e$, $\mathcal{M}_g$, and $\mathcal{I}_g$ are estimated by $\hat{\mathcal{M}}_e^{(\lambda_1,\alpha_1)}$, $\hat{\mathcal{I}}_e^{(\lambda_1,\alpha_1)}$, $\hat{\mathcal{M}}_g^{(\lambda_1,\alpha_1)}$, and $\hat{\mathcal{I}}_g^{(\lambda_1,\alpha_1)}$, respectively. For simplicity, we suppress the explicit dependence on λ_1 and α_1 in the notation in the following.

Step 2 We assume the pairwise interaction model, for all $i = 1, \dots, N$ and $l = 1, \dots, q$,

$$y_{il} = \sum_{j=1}^{p} x_{ij} b_{jl} + \sum_{k=1}^{p} \sum_{j<k} \mathbb{1}_{\{(j,k,l) \in \hat{\mathcal{I}}_e\}} x_{ij} x_{ik} \Theta_{jkl} + e_{il},$$

for which the MSGLasso estimates are determined by solving the optimisation problem

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times q}, \boldsymbol{\Theta} \in \mathbb{R}^{p \times p \times q}}{\text{minimise}} \left\{ L(\mathbf{B}, \boldsymbol{\Theta}) + \lambda_2 \sum_{j=1}^{p} \left(\sum_{l=1}^{q} \alpha_2 \mathbb{1}_{\{(j,l) \notin \hat{\mathcal{M}}_e\}} |b_{jl}| + (1 - \alpha_2) \mathbb{1}_{\{j \notin \hat{\mathcal{M}}_g\}} \|\mathbf{b}_j\|_2 \right. \right.$$
$$\left. \left. + \sum_{k=1}^{j-1} \sum_{l=1}^{q} \alpha_3 \mathbb{1}_{\{(j,k,l) \in \hat{\mathcal{I}}_e\}} |\Theta_{jkl}| + (1 - \alpha_3) \mathbb{1}_{\{j \in \hat{\mathcal{I}}_g\}} \|\boldsymbol{\Theta}_j\|_2 \right) \right\},$$

where $L(\mathbf{B}, \boldsymbol{\Theta})$ denotes the loss function

$$L(\mathbf{B}, \boldsymbol{\Theta}) = \frac{1}{2Nq} \sum_{i=1}^{N} \sum_{l=1}^{q} \left(y_{il} - \sum_{j=1}^{p} x_{ij} b_{jl} - \sum_{k=1}^{p} \sum_{j<k} \mathbb{1}_{\{(j,k,l) \in \hat{\mathcal{I}}_e\}} x_{ij} x_{ik} \Theta_{jkl} \right)^2,$$

and $\lambda_{jl} = \lambda_2 \alpha_2 \mathbb{1}_{\{(j,l) \notin \hat{\mathcal{M}}_e\}}$, $\lambda_j = \lambda_2 (1 - \alpha_2) \mathbb{1}_{\{j \notin \hat{\mathcal{M}}_g\}}$, $\tau_{jkl} = \lambda_2 \alpha_3 \mathbb{1}_{\{(j,k,l) \in \hat{\mathcal{M}}_e\}}$, and $\tau_j = \lambda_2 (1 - \alpha_3) \mathbb{1}_{\{j \in \hat{\mathcal{M}}_g\}}$.

We are finally able to estimate the sets $\mathcal{S}_e$, $\mathcal{S}_g$, $\mathcal{R}_e$, and $\mathcal{R}_g$, by

$$\begin{aligned}
\hat{\mathcal{S}}_e^{(\lambda_2,\alpha_2,\alpha_3)} &= \left\{ j \in \{1, \dots, p\}, l \in \{1, \dots, q\} : \left| \hat{b}_{jl}^{(\lambda_2,\alpha_2,\alpha_3)} \right| \neq 0 \right\}, \\
\hat{\mathcal{S}}_g^{(\lambda_2,\alpha_2,\alpha_3)} &= \left\{ j \in \{1, \dots, p\} : \left\| \hat{\mathbf{b}}_j^{(\lambda_2,\alpha_2,\alpha_3)} \right\|_2 \neq 0 \right\}, \\
\hat{\mathcal{R}}_e^{(\lambda_2,\alpha_2,\alpha_3)} &= \left\{ j, k \in \{1, \dots, p\}, l \in \{1, \dots, q\} : \left| \hat{\Theta}_{jkl}^{(\lambda_2,\alpha_2,\alpha_3)} \right| \neq 0 \right\}, \\
\hat{\mathcal{R}}_g^{(\lambda_2,\alpha_2,\alpha_3)} &= \left\{ j \in \{1, \dots, p\} : \left\| \hat{\boldsymbol{\Theta}}_j^{(\lambda_2,\alpha_2,\alpha_3)} \right\|_2 \neq 0 \right\},
\end{aligned}$$

respectively.

It should be noted, that we choose to treat main effects and interactions equally and impose the same regularisation on main effects and interactions, thereby keeping down the number of parameters. This might result in main effects being overwhelmed by interactions in the selection, and it could be argued that main effects should be less regularised than interactions.

3 Proof of concept simulations

We illustrate the utility of the two-step procedure described in Section 2.4 on simulated data.

The motivation is to understand the performance of the two-step procedure on one of the important problems in statistical genetics: genotype by genotype interactions (G×G) in multi-trait genomic selection. We use simulations to assess the false discovery rate of the interactions found under the assumption of strong hierarchy for both the data generating process and the two-step procedure.

3.1 Sampling procedure

Since the efficacy of a procedure depends on the true model generating the data we simulate different set-ups such that each of the different restrictions are tried as the ground truth, and we apply our method to each scenario separately.

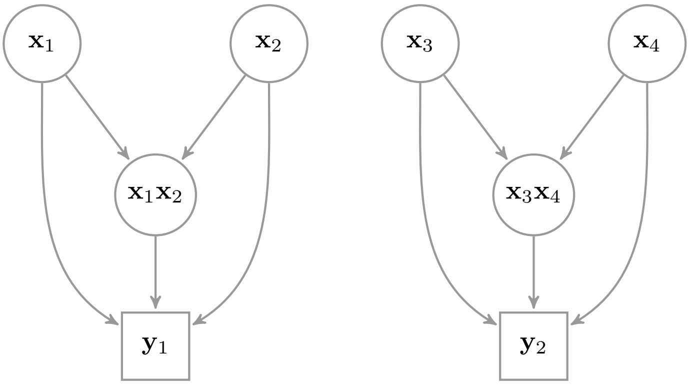


Figure 1: *Graphical illustration of data generating process where there is a main effect of* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *on* $\mathbf{y}_1$ *and an interactions between* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *affecting* $\mathbf{y}_1$ and *a main effect of* $\mathbf{x}_3$ *and* $\mathbf{x}_4$ *on* $\mathbf{y}_2$ *and an interactions between* $\mathbf{x}_3$ *and* $\mathbf{x}_4$ *affecting* $\mathbf{y}_2$.

The sampling procedure goes as follows: each sample is generated with $N = 200$ observations of $q = 10$ traits $\mathbf{Y}$ and $p = 600$ associated categorical features $\mathbf{X} \in \{0, 1, 2\}^{200\times600}$ resulting in 179700 potential pairwise interactions *per trait*. There are four main effects and two interactions in the data generating process, as illustrated in Figure 1. In practice, simulating more independent interactions would produce similar results when each of these interactions are considered in turn. From the features we generate observations by $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where $\mathbf{E}$ is a matrix of standard Gaussian noise. We create a sparse problem by letting $b_{jl} = 0$ for all $j = 1, \dots, p$ and $l = 1, \dots, q$, except for b_{11}, b_{21},b_{32}, b_{42},Θ_{121} and Θ_{342}, for which different values are tried.

We fit the models in the two-step procedure using our wrapper function `milasso()` (Buchardt and Ekstrøm, 2021) which calls functions from the R package `MSGLasso` by Li et al. (2016). The penalty parameters are determined by cross-validation, which we perform using the `MSGLasso.cv()` function with 5 folds. To reduce the residual variation we repeat the sampling of the outcome as well as the fitting using the two-step procedure $B = 100$ times.

In order for us to compare the performance of the methods we use the false-discovery rate of the interactions (iFDR) across all traits, defined as

$$\text{iFDR} = \frac{V^{(i)}}{V^{(i)} + S^{(i)}},$$

where $V^{(i)}$ is the number of false discoveries of interactions and $S^{(i)}$ is the number of true discoveries of interactions across all traits. Please note that the false-discovery rates are defined in a manner that returns a zero if the number of false discoveries is zero. If there are no false discoveries nor true discoveries, the false-discovery rates are not provided.

We present the mean iFDR with corresponding point-wise approximate confidence intervals which we compute as the mean iFDR plus/minus twice the standard error of the mean. We also present the average number of true interaction discoveries conditioned on having discovered a true interaction. We refer to this number as the *recall* with respect to interactions. Please note that the average number of false interaction discoveries, which complete the picture is not displayed. Since the simulated data honour strong hierarchy with two interaction effects, the desired number of true interaction discoveries is two. That is, the dotted lines should be close to two. Since we aim to minimise the expected proportion of interaction discoveries which are false, the desired value for the mean iFDR is zero. That is, the solid lines should to be close to zero.

In the left panel of Figure 2 we show the mean iFDR with corresponding point-wise approximate confidence intervals as a function of $\beta_{11} = \beta_{21} = \beta_{32} = \beta_{42}$ when the model is constrained by strong hierarchy. The lines are coloured according to the true value of $\Theta_{121} = \Theta_{342}$. In the right panel of Figure 2 we show the recall with respect to interactions as a function of $\beta_{11} = \beta_{21} = \beta_{32} = \beta_{42}$ when the model is constrained by strong hierarchy. Again, the lines are coloured according to the true value of $\Theta_{121} = \Theta_{342}$. We observe that, for the chosen values of coefficient, the mean iFDR is close to zero for all values of the two true main effects and all values of the true interactions. This indicates that when we do discover an interaction it is likely to be the correct one. We also observe that the recall with respect to interactions increases as the size of the true interactions increases. More specifically, we meet the goal of discovering the two interactions whenever the true size of the interactions is greater than 2.4.

Further numerical studies are needed to fully assess the performance of the method. For example, to understand the consequence of a wrong hierarchical assumption it would be interesting to simulate datasets with different (non-)hierarchical structures and apply the multivariate two-step procedure under different hierarchical assumptions. Furthermore, to understand the consequence of not requiring hierarchy it would be interesting to compare our method on simulated datasets to the usual (multivariate) lasso for the pairwise interaction model under the assumption of no hierarchy. It would also be of interest to assess the selection abilities of the method in real data.

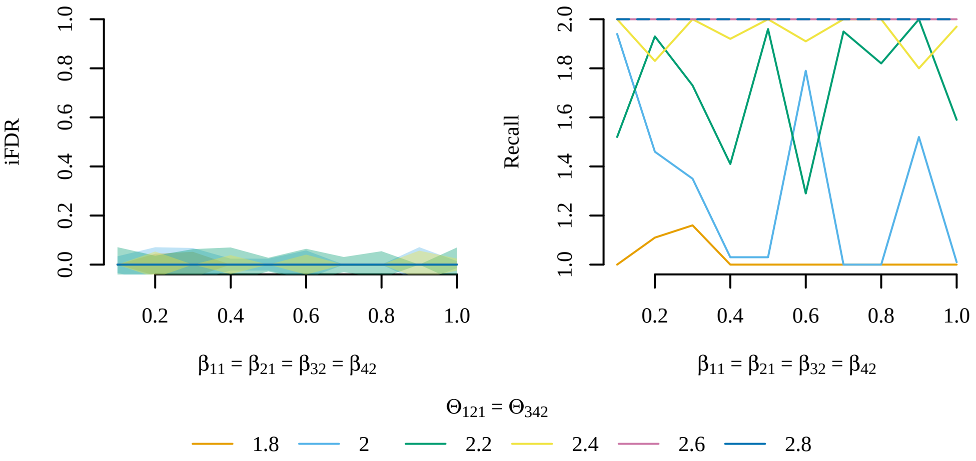


Figure 2: *The data generating process honours strong hierarchy.* Left: *Mean* iFDR *over* $B = 100$ *repeated samples and corresponding point-wise approximate confidence intervals as a function of* $\beta_{11} = \beta_{21} = \beta_{32} = \beta_{42}$ *when the model is constrained by strong hierarchy. Lines are coloured according to the true value,* $\Theta_{121} = \Theta_{342}$*, of the interactions. The lines should be close to zero.* Right: *Recall with respect to interactions as a function of* $\beta_{11} = \beta_{21} = \beta_{32} = \beta_{42}$ *when the model is constrained by strong hierarchy. Lines are coloured according to the true value,* $\Theta_{121} = \Theta_{342}$*, of the interactions. The lines should to be close to two.*

4 Discussion

In this paper, we have proposed a two-step procedure, which use a special case of the multivariate sparse group lasso (Li et al., 2015) for screening for interactions and fitting multivariate pairwise interaction models under hierarchical restrictions.

Our methodology is designed to identify pairwise interactions in a multivariate regression model which involves only a subset of the features and interactions even when data are available for a limited number of individuals and when the number of outcome components (e.g., traits) per individual and the number of features (e.g., SNPs) are large. The main contribution of our procedures is that interactions are included under hierarchical restrictions as well as simultaneously estimated and selected. We believe that this is one of the strengths of our procedure as it furthers interpretability and (expectedly) computing time. Furthermore, the framework allows for direct specification of whether the treatment of features (and interactions) should be equal or different across outcome components.

Our method could still be further refined when features are correlated. One potential solution to this issue consists in using the cluster Elastic Net framework (Price and Sherwood, 2018). As an alternative, the correlations among features may be incorporated by considering blocks of features (e.g., SNPs within chromosomes) within the model using more refined group structures of the MSGLasso, as this might improve the detection of weak associations. Other future work will also consider the oracle property, and how to improve the framework to attain these desirable properties.

In this paper, we have presented the theoretical framework and a proof-of-concept simulation study. From the simulation study we obseved that both in terms of false discovery rate and recall with respect to interactions the two-step procedure works very well under the assumption of strong hierarchy.

While we are not aware of any alternative multivariate methods that simultaneously estimate and select interactions under hierarchical restrictions, it would be of interest to compare the performance – in terms of discoveries and prediction error – of the two-step procedure and a simple marginal multiple testing procedure.

It should be noted that, in this paper, we see the choosing of the regularisation parameters primarily as a means to an end. Choosing the right regularisation parameters is, in general, difficult. Both cross-validation and stability selection are possible strategies.

Our work has potential applications to cross-omics studies and multivariate genome-wide studies of G×G, that is, epistatic pairs where the effect of one gene (locus) on, possibly, multiple traits is dependent on the presence of a *modifier gene*. It is possible to extend our framework to the case of multi-way interactions and, thereby, to the application to studies of, e.g., higher-order epistatic interactions. In practice, however, this is an unfeasible challenge in terms of tractability, computability, and interpretation. As a special case, our work has potential applications to studies of G×E. In studies such as genetic epidemiology, presumably, a small-scale number of environmental factors and a large-scale number of genes will be available. In this case, regularisation of the environmental factors may be undesirable, and, with a slight modification, the two-step procedure is capable of taking this into account.

Bibliography

Aiken, L. S. and West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications, Inc., 1 edition.

Bien, J. and Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.6.

Buchardt, A.-S. (2022). iLasso: Identifying interactions via hierarchical lasso regularisation. `https://github.com/abuchardt/ilasso`.

Buchardt, A.-S. and Ekstrøm, C. T. (2021). Identifying interactions via hierarchical lasso regularisation.

Cornelis, M. C., Tchetgen, E. J. T., Liang, L., Qi, L., Chatterjee, N., Hu, F. B., and Kraft, P. (2012). Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology*, 175(3):191–202.

Cox, D. R. (1984). Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1–24.

Dong, S.-S., Hu, W.-X., Yang, T.-L., Chen, X.-F., Yan, H., Chen, X.-D., Tan, L.-J., Tian, Q., Deng, H.-W., and Guo, Y. (2017). Snp-snp interactions between wnt4 and wnt5a were associated with obesity related traits in han chinese population. *Scientific Reports*, 7.

Ekstrøm, C. T. and Sørensen, H. (2014). *Introduction to Statistical Data Analysis for the Life Sciences*. CRC Press, 2 edition.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLOS ONE*, 9(4):1–8.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. Chapman and Hall/CRC Press.

Hu, J. K., Wang, X., and Wang, P. (2014). Testing gene-gene interactions in genome wide association studies. *Genetic Epidemiology*, 38(2):123–134.

Lee, K.-Y., Leung, K.-S., Tang, N. L. S., and Wong, M.-H. (2018). Discovering genetic factors for psoriasis through exhaustively searching for significant second order snp-snp interactions. *Scientific Reports*, 8.

Li, Y., Nan, B., and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–63.

Li, Y., Nan, B., and Zhu, J. (2016). *MSGLasso: Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure*. R package version 2.1.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901.

Price, B. S. and Sherwood, B. (2018). A cluster elastic net for multivariate regression. *Journal of Machine Learning Research*, 18(232):1–39.

Schmitz, S., Cherny, S. S., and Fulker, D. W. (1998). Increase in power through multivariate analyses. *Behavior Genetics*, 28(5):357–363.

Schrode, N., Ho, S.-M., Yamamuro, K., Dobbyn, A., Huckins, L., Matos, M. R., Cheng, E., Deans, P. J. M., Flaherty, E., Barretto, N., Topol, A., Alganem, K., Abadali, S., Gregory, J., Hoelzli, E., Phatnani, H., Singh, V., Girish, D., Aronow, B., Mccullumsmith, R., Hoffman, G. E., Stahl, E. A., Morishita, H., Sklar, P., and Brennand, K. J. (2018). Synergistic effects of common schizophrenia risk variants. *Nature genetics*, 51(10):1475–1485.

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