Statistical Methods in Neuroimaging Genetics:
Pathways Sparse Regression
and Cluster Size Inference

A thesis presented for the degree of
Doctor of Philosophy of the University of London
and the Diploma of Imperial College
by

Matthew Silver

Department of Mathematics
Imperial College
180 Queen’s Gate, London SW7 2BZ

JANUARY 25, 2013
Dedicated to my mother, Ruth...

...and to my existing family - Bryher, Morwenna, Rosie, Eloïse and my father Dave.
I certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Matt Silver
Copyright

Copyright in the text of this thesis rests with the Author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the Author and lodged in the doctorate thesis archive of the college central library. Details may be obtained from the Librarian. This page must form part of any such copies made. Further copies (by any process) of copies made in accordance with such instructions may not be made without the permission (in writing) of the Author. The ownership of any intellectual property rights which may be described in this thesis is vested in Imperial College, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the University, which will prescribe the terms and conditions of any such agreement. Further information on the conditions under which disclosures and exploitation may take place is available from the Imperial College registry.
Acknowledgements

I am indebted to my Ph.D supervisor, Dr Giovanni Montana, for his advice and guidance. Thanks to him, I have made the journey from a position of scant understanding, to one of at least moderate understanding of the world of statistics applied to the analysis of large biological datasets. I am also grateful to the Wellcome Trust, who generously funded my studies and research.

In addition I would like to acknowledge contributions from external collaborators in three of the studies described in this thesis:

1. For the Alzheimer’s pathways study described in Section 3.5, image pre-processing and analysis was carried out by Eva Janousova, at the time visiting Imperial College from the Institute of Biostatistics and Analyses, Masaryk University, Brno in the Czech Republic, and by Dr Xue Hua and Professor Paul Thompson from the Laboratory of Neuro Imaging, Department of Neurology at the UCLA School of Medicine, USA. ADNI Genotype data was extracted from imputed genotypes used in two previous studies (Vounou et al., 2011; Stein et al., 2010a).

2. For the HDL-C study described in Section 4.3, non-imputed genotype data was provided by Professor YY Teo and Cheng Peng from the Department of Statistics and Applied Probability, University of Singapore.

3. For the study investigating control of false positive rates using cluster size inference in Chapter 6, which began as a final project for an MSc in Bioinformatics and Theoretical Systems Biology at Imperial College, extensive help and support was provided by Dr Thomas Nichols at the Department of Statistics & Warwick Manufacturing Group, University of Warwick, UK.

Finally, I would like to thank my immediate family, Bryher, Morwenna, Rosie and Eloïse, for their boundless patience and support throughout.

Matt Silver
ADNI Data Use

Neuroimaging and genotype data used in a number of studies in this thesis were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of any part of this thesis. A complete listing of ADNI investigators can be found at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

ADNI is funded by a National Institutes of Health Grant U01 AG024904, and by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; Bioclinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.
Abstract

In the field of neuroimaging genetics, brain images are used as phenotypes in the search for genetic variants associated with brain structure or function. This search presents a formidable statistical challenge, not least because of the very high dimensionality of genotype and phenotype data produced by modern SNP (single nucleotide polymorphism) arrays and high resolution MRI. This thesis focuses on the use of multivariate sparse regression models such as the group lasso and sparse group lasso for the identification of gene pathways associated with both univariate and multivariate quantitative traits.

The methods described here take particular account of various factors specific to pathways genome-wide association studies including widespread correlation (linkage disequilibrium) between genetic predictors, and the fact that many variants overlap multiple pathways. A resampling strategy that exploits finite sample variability is employed to provide robust rankings for pathways, SNPs and genes. Comprehensive simulation studies are presented comparing one proposed method, pathways group lasso with adaptive weights, to a popular alternative. This method is extended to the case of a multivariate phenotype, and the resulting pathways sparse reduced-rank regression model and algorithm is applied to a study identifying gene pathways associated with structural change in the brain characteristic of Alzheimer’s disease. The original model is also adapted for the task of ‘pathways-driven’ SNP and gene selection, and this latter model, pathways sparse group lasso with adaptive weights, is applied in a search for SNPs and genes associated with elevated lipid levels in two separate cohorts of Asian adults.

Finally, in a separate section an existing method for the identification of spatially extended clusters of image voxels with heightened activation is evaluated in an imaging genetic context. This method, known as cluster size inference, rests on a number of assump-
tions. Using real imaging and SNP data, false positive rates are found to be poorly con-
trolled outside of a narrow range of parameters related to image smoothness and activation
thresholds for cluster formation.
# Table of contents

## Abstract

vi

## List of Publications

5

## 1 Introduction

6

1.1 Gene association mapping ............................................. 6
1.2 The gene pathways approach ........................................... 11
1.3 Imaging genetics ....................................................... 16
1.3.1 Cluster size inference .............................................. 21
1.4 Mathematical notation and nomenclature ........................... 22
1.5 Abbreviations .......................................................... 24

## 2 Identifying pathways associated with a univariate quantitative trait: ‘Pathways Group Lasso with Adaptive Weights’

25

2.1 Penalised regression approaches in gene mapping .................. 26
2.1.1 Ridge regression .................................................... 28
2.1.2 Variable selection with the lasso ................................ 31
2.2 The group lasso for pathway selection .............................. 36
2.2.1 The problem of overlapping pathways ............................ 39
2.3 Group lasso estimation algorithm .................................... 41
2.3.1 Block coordinate descent for non-orthogonal groups .......... 42
2.3.2 Taylor approximation of penalty .................................. 44
2.3.3 Use of pathway ‘active set’ ....................................... 46
2.3.4 Efficient computation of block residuals ........................ 48
2.4 Selection bias and pathway weighting ............................... 49
2.5 Pathway ranking ....................................................... 52
2.5.1 Ranking performance measures ................................... 55
2.6 P-GLAW Simulation Study ............................................ 57
2.6.1 Genotype and pathways data ..................................... 57
2.6.2 Simulation framework .............................................. 59
2.6.3 Results .............................................................. 62
2.7 Discussion .............................................................. 71
6 False positives in neuroimaging genetics using cluster size inference with Voxel-based morphometry data

6.1 Imaging data ............................................. 177
   6.1.1 ADNI images .................................... 177
   6.1.2 Simulated images ................................ 178
6.2 Genotype data .......................................... 178
6.3 Statistical inference ................................... 180
   6.3.1 Non-stationary cluster size inference ......... 181
6.4 Results .................................................. 183
   6.4.1 ADNI images .................................... 183
   6.4.2 Simulated images ................................ 186
6.5 Discussion .............................................. 188
6.6 Conclusion .............................................. 194

A Review of Existing Pathways Methods

A.1 Univariate-based methods ............................ 196
A.2 Multivariate methods .................................. 200
List of Figures

1.1 Schematic representation of points of common genetic variation between individuals .............................................. 7
1.2 Pairwise correlation (LD) patterns between 420 SNPs on a 500kb stretch of human chromosome 5 .................................. 8
1.3 Schematic of genes acting together in a functional pathway ................. 12
1.4 Schematic illustration of the SNP to pathway mapping process ............. 14
1.5 Imaging genetic analysis .......................................................... 18
1.6 MULM in imaging genetic analysis ............................................. 22

2.1 Lasso coefficient estimation with two predictors ................................. 30
2.2 Group lasso - schematic illustration of distribution of causal SNPs .......... 37
2.3 The problem of overlapping pathways ........................................ 40
2.4 SNP to pathway mapping .......................................................... 58
2.5 Frequency distribution of ADNI SNPs by number of pathways they map to 59
2.6 Distributions of $|C|$ across 500 MC simulations for each of the 6 scenarios tested ............................................................... 61
2.7 Application of bias-adjusted weighting procedure to the data used in the simulation study ........................................... 63
2.8 Ranking performance using adapted weights .................................... 65
2.9 ROC curves ............................................................................. 68
2.10 Box plots of distribution of ranking power, $p_{100}$, across 500 simulations . 69
2.11 Distribution of the power-adjusted, normalised, weighted ranking score, $R$, across 500 simulations ........................................ 70

3.1 Reduced-rank regression model ....................................................... 78
3.2 Reduced-rank regression latent factors .......................................... 79
3.3 PsRRR Simulation study 1 ............................................................ 89
3.4 PsRRR simulation study 2 ............................................................. 91
3.5 Mapping SNPs to pathways .......................................................... 97
3.6 ADNI/KEGG pathway sizes and SNP overlaps ................................. 98
3.7 Parallel computing framework for PsRRR estimation ......................... 100
LIST OF FIGURES

3.8 Sample mean (left) and standard deviation (right) of slope coefficients for the 2 subject groups .................................................. 101
3.9 Imaging signature characteristic of AD ........................................ 102
3.10 3D multi-dimensional scaling plot illustrating the spread of imaging signatures across ADs and CNs ............................... 103
3.11 Measure of extent to which genes previously linked to AD are enriched in highly-ranked pathways .................................. 107

4.1 Sparsity patterns enforced by the group lasso and sparse group lasso ...... 113
4.2 SGL vs Lasso. Comparison of power to detect 5 causal SNPs ............. 122
4.3 SGL vs Lasso. Distribution over 500 MC simulations of power to detect 5 causal SNPs .............................................................. 123
4.4 Two pathways with partially overlapping causal SNPs .................... 125
4.5 SGL Simulation Study with overlapping pathways .......................... 129
4.6 SGL-CGD vs SGL-BCGD performance, measured across 2000 MC simulations ................................................................. 131
4.7 SP2 dataset. SNP to pathway mapping ........................................ 138
4.8 SiMES dataset. SNP to pathway mapping .................................... 138
4.9 Empirical and null pathway selection frequency distributions for all 185 KEGG pathways with the SP2 dataset .............................. 141
4.10 Empirical and null SNP selection frequency distributions with the SP2 dataset ................................................................. 142
4.11 SP2 dataset: Scatter plots comparing empirical and null selection frequencies ................................................................. 144
4.12 SiMES dataset: Scatter plots comparing empirical and null selection frequencies ................................................................. 148
4.13 Selection frequency distributions for the SiMES dataset .................. 149
4.14 Comparison of top-k SP2 and SiMES pathway rankings .................. 154
4.15 Comparison of top-k SP2 and SiMES gene rankings ........................ 157

6.1 Non-stationary image simulation ................................................ 179
6.2 Accuracy of estimation of c, the theoretical number of clusters under RFT . 189
6.3 Distribution of voxel-wise FWHM for ADNI images smoothed with 6mm (left) and 12mm (right) Gaussian smoothing kernels ............... 191
6.4 12mm VBM image non-stationarity ............................................ 195
# List of Tables

1.1 SNP association test in a case-control study ........................................... 9
2.1 Scenarios tested in simulation study ....................................................... 60
2.2 GL estimation times in seconds ............................................................. 62
2.3 Selected ranking performance measures for P-GLAW and GG ................. 66
2.4 Proportion of 500 simulations with $R < 10$ and $R < 3$ ....................... 71
3.1 Empirical phenotypic $SNR$ values .......................................................... 88
3.2 Empirical phenotypic $SNR$ values for Simulation study 2 ..................... 90
3.3 Available scans at 6, 12 and 24 months for the ADNI-1 dataset (down-  
loaded on February 28, 2011) ............................................................... 95
3.4 AD genes included in this study ............................................................. 99
3.5 Top 30 pathways, ranked by pathway selection frequency ..................... 104
3.6 Top 30 SNPs and genes, respectively ranked by SNP and gene selection  
frequency, using lasso sRRR ............................................................... 106
4.1 Simulation Study 2: Mean number of pathways and SNPs selected by each  
model at each effect size, $\gamma$, across 2000 MC simulations. ..................... 130
4.2 Genotype and phenotype information corresponding to the SP2 and SiMES  
datasets used in the study. ................................................................. 136
4.3 SNP and gene to pathway mappings for the SP2 and SiMES datasets. .... 139
4.4 Separate combinations of the P-SGLAW regularisation parameters, $\lambda$ and  
$\alpha$ used for analysis of the SP2 dataset ............................................. 140
4.5 SP2 dataset: Pearson correlation coefficients ($r$) and p-values .............. 143
4.6 SP2 dataset: Top 30 pathways, ranked by pathway selection frequency .... 146
4.7 SP2 dataset: Top 30 SNPs and genes ranked by SNP and gene selection  
frequency .......................................................................................... 147
4.8 SiMES dataset: Pearson correlation coefficients ($r$) and p-values .......... 148
4.9 SiMES dataset: Top 30 pathways, ranked by pathway selection frequency  . 150
4.10 SiMES dataset: Top 30 SNPs and genes ranked by SNP and gene selection  
frequency .......................................................................................... 151
4.11 Consensus set of pathways for SP2 and SiMES datasets with $k = 25$ .... 155
4.12 Summary of genes analysed and ranked in SP2 and SiMES datasets. .... 157
4.13 Consensus set of genes for SP2 and SiMES datasets with $k = 244$ . . . . . 158
4.14 Summary of SNPs analysed and ranked in SP2 and SiMES datasets. . . . . 159
4.15 Top 30 SNPs ranked in both SP2 and SiMES datasets, ranked in order of mean ranking, $\psi^{SNP}$ (4.28). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 160
6.1 FWER-corrected results - real (ADNI) images . . . . . . . . . . . . . . . . 184
6.2 FDR-corrected results - real (ADNI) images . . . . . . . . . . . . . . . . . 185
6.3 Results - simulated images . . . . . . . . . . . . . . . . . . . . . . . . . . 187
A.1 Key features of reviewed pathways-based genetic association methods . . . 204
List of Publications

Much of the material to be found in this thesis is covered in the following published papers:


Chapter 1

Introduction

1.1 Gene association mapping

Since the publishing of the first draft of the human genome, a huge amount of effort has been expended in the hunt for genetic variation associated with human disease. Much of this work has focused on the identification of single points of variation, known as single nucleotide polymorphisms or SNPs, where common variation is observed in the human population (see Figure 1.1). Parallel advances in sequencing technology and the detailed cataloguing of SNPs for diverse human populations have led to the identification of 100s of SNPs affecting an array of common diseases and traits (Green et al., 2011; Altshuler, Daly, and Lander, 2008).

In modern genome-wide association studies (GWAS), 100s or 1000s of individuals are genotyped at up to a million or more SNPs across the genome. This ‘hypothesis-free’ approach to searching for genetic variants in GWAS has been greatly aided by the ability to exploit patterns of widespread correlation or linkage disequilibrium (LD) between genetic markers belonging to unrelated individuals. Different human sub-populations can be characterised by these distinctive LD patterns, which have a block-like structure, arising from the relatively few sexual recombination events that have occurred throughout recent human history (see Figure 1.2). For a given human sub-population, the presence of these LD blocks or haplotypes enables single ‘tagging’ SNPs to be used as proxies for other, ungenotyped variants, dramatically decreasing the number of variants that need to be genotyped
Figure 1.1: Schematic representation of points of common genetic variation between individuals. Letters ACGT represent single nucleotides that make up the human genome. Extended stretches of nucleotides can be grouped into genes, broadly classified as functional units within the genome. Single nucleotide polymorphisms or SNPs, are genetic loci at which there is common variation between individuals in the human population, typically defined as variants whose minor ‘allele’ (here marked in red) occurs with a frequency > 1%. An individual has two copies of each allele, and so may possess 0, 1 or 2 minor alleles. SNPs may occur within genes (SNP 1) or between genes (SNP 2).
1.1 Gene association mapping

(The International HapMap Consortium, 2005). Variants identified using this approach will however typically have only an indirect relationship to the true causal variant(s) that they tag (Hirschhorn and Daly, 2005). For this reason, this approach may also come at the cost of reduced power, compared to fine-mapping studies that include all SNPs in a genomic region of interest. The presence of LD can also lead to inflated numbers of false positives, due to spurious associations arising from SNPs in LD with causal SNPs.

Genetic variation between subjects is generally measured by recording the number of minor alleles possessed by each individual, a SNP’s minor allele being defined as the less common variant in the sample (variants marked in red in Figure 1.1). Since individuals inherit two copies of each allele, one from the mother and one from the father, each subject may possess zero, one or two minor alleles. Thus in a study with $N$ subjects genotyped at $P$ SNPs, we can denote the minor allele count for SNP $j$ observed on individual $i$ as $x_{ij} \in \{0, 1, 2\}, i = 1, \ldots, N; j = 1, \ldots, P$.

The standard statistical approach is then one of mass univariate testing, where the hypothesis that minor alleles are associated with a particular trait or phenotype is tested, one SNP at a time. In a case-control study for example, we might test the hypothesis that minor alleles are more likely to be found in cases than controls by testing for independence of rows and columns in a $2 \times 2$ contingency table, using a $\chi^2$ test (see Table 1.1). Different
Table 1.1: Association test for SNP $j$ in a case-control study. (a) $d_0, d_{12}$ represent observed number of cases with $x_j = 0$, or $x_j \in \{1, 2\}$, respectively. $c_0, c_{12}$ represent the same for controls. (b) Expected values under the null, i.e. where rows and columns in (a) are independent. The hypothesis that minor alleles are associated with disease status is tested by observing that $\sum (\text{Observed} - \text{Expected})^2 / \text{Expected} \sim \chi^2_1$ under the null.

<table>
<thead>
<tr>
<th></th>
<th>$x_j$</th>
<th>0, 1, 2</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>$d_0$</td>
<td>$d_{12}$</td>
<td>D</td>
</tr>
<tr>
<td>controls</td>
<td>$c_0$</td>
<td>$c_{12}$</td>
<td>C</td>
</tr>
<tr>
<td>total</td>
<td>$n_0$</td>
<td>$n_{12}$</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$x_j$</th>
<th>0, 1, 2</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>$n_0D/N$</td>
<td>$n_{12}D/N$</td>
<td>D</td>
</tr>
<tr>
<td>controls</td>
<td>$n_0C/N$</td>
<td>$n_{12}C/N$</td>
<td>C</td>
</tr>
<tr>
<td>total</td>
<td>$N$</td>
<td>$N$</td>
<td>N</td>
</tr>
</tbody>
</table>

genetic models, assuming for example recessive or dominant genetic effects can be tested by changing the column categories. An alternative strategy for testing associations in case-control studies is to code phenotypes as $y_i \in \{0, 1\}$ for cases and controls respectively, and use a logistic regression model, $\logit(\pi_i) = \beta_0 + \beta_1 x_{ij}$, where $\logit(\pi_i) = \log(\pi_i / (1 - \pi_i))$ and $\pi_i$ is the disease risk of the $i$th individual. Here we assume an additive genetic model, whereby disease risk increases in proportion with the number of minor alleles. We can test the null hypothesis that there is no association between disease status, $H_0 : \hat{\beta}_1 = 0$, using a likelihood-ratio test. The use of a logistic regression model has the advantage of allowing easy incorporation of model covariates, such as age or sex.

The search for SNPs, known as ‘quantitative trait loci’ (QTL), influencing quantitative traits, i.e. those that vary continuously, for example blood pressure or blood lipid levels, is also gaining momentum as a potentially more powerful way to study the underlying causes of complex disease (Plomin, Haworth, and Davis, 2009). For quantitative traits, the association of SNP minor alleles with a phenotype is established using a quantitative trait test of association (QTT). For example by assuming a linear model $y_i = \beta_0 + x_{ij}\beta_j + \epsilon_i$, we can perform a t-test on the SNP regression coefficient $\beta_j$ to test the null hypothesis $H_0 : \beta_j = 0$. Once again this model can easily be extended to include covariates.

A major limitation of the mass univariate testing approach in GWAS arises from the very large number of SNPs, and hence hypotheses that must be tested. In order for the number of false positives to be adequately controlled, some correction to the single-SNP threshold for significance, $\alpha$, corresponding to the probability that a single test will yield a false positive association, must be applied. If we perform $P$ such tests across the genome,
and further assume that each test is independent, the corresponding chance, genome-wide, for such a false positive finding is then $\alpha_{GW} = 1 - (1 - \alpha)^P$. This leads to the familiar Bonferroni correction, $\alpha_{GW} = \alpha/P$, which at the $\alpha = 0.05$ level yields a significance threshold, $\alpha_{GW} = 10^{-7}$ for a typical GWAS with $P = 500,000$. Such a stringent threshold for the achieving of genome-wide significance substantially reduces the power to detect associations (Altshuler, Daly, and Lander, 2008). The presence of widespread correlation between markers due to LD means that tests are in fact not independent, so that the Bonferroni correction is likely to be too conservative. One solution is to approximate the false positive rate by comparing empirical p-values with a null distribution generated from the same dataset, but with multiple phenotype permutations. Here LD patterns are preserved while the phenotype-genotype association is broken (Balding, 2006). Another approach is to increase power by accepting a higher risk of false positives by controlling the expected number of false discoveries as a proportion of positive tests, using the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

In contrast to univariate, ‘one SNP at a time’ methods, multivariate or multilocus methods allow all SNPs to be considered in the model at the same time, which can aid the identification of weak signals while diminishing the importance of false ones. Multilocus or haplotype methods test the association of haplotypes, rather than individual SNPs, with phenotypes, and are thus able to implicitly exploit dependencies between SNPs, while reducing the number of tests to be performed (Duggal et al., 2008; Balding, Waldron, and Whittaker, 2006). Another approach is to extend the previous, single SNP regression model described earlier, to a multiple regression model, in which all SNPs are modelled together. Thus for a quantitative trait we have the multiple linear regression (MLR) model, $y_i = \beta_0 + \sum_{j=1}^{P} x_{ij} \beta_j + \epsilon_i$. The use of multiple regression models in GWAS, while theoretically attractive, is however problematic in practice. Firstly, in GWAS, the number of SNPs is typically much greater than the sample size, so that estimates for the SNP regression coefficients $\beta_j, j = 1, \ldots, P$, are not uniquely defined. Secondly, even where SNP coefficients are estimable, LD or multicollinearity between SNPs renders such estimates unreliable (Hastie, Tibshirani, and Friedman, 2008).

One solution is to use some form of regularisation or constraint on the coefficient vector $\beta = (\beta_1, \ldots, \beta_P)$. Types of regularisation include ridge regression (Hoerl and Kennard,
1970), lasso regression (Tibshirani, 1996) and the elastic net (Zou and Hastie, 2005), each of which provides coefficient estimates with different characteristic properties (see Section 2.1). The lasso is a particularly attractive option for GWAS, since the lasso constraint imposes sparse estimates for $\beta$, in that many individual coefficient estimates are precisely zero. For this reason, the lasso, and related approaches such as stepwise regression (Hesterberg et al., 2008) are often described as performing ‘model selection’, since they highlight a subset of important predictors. Since its first application in GWAS (Wu et al., 2009), the lasso and related sparse regression models have been used to tackle a variety of problems in biomarker discovery. Of particular note in the present context is their use in imaging genetics for the identification of SNPs associated with high-dimensional neuroimaging phenotypes (Bunea et al., 2011; Kohannim et al., 2012a; Vounou, Nichols, and Montana, 2010).

1.2 The gene pathways approach

Gene variants identified in GWAS have so far uncovered only a relatively small part of the known heritability of most common diseases (Visscher et al., 2012). This has focused attention on the need to develop new methodological approaches. A number of factors that might explain this ‘missing heritability’ have been suggested, including the failure of many current models to capture the presence of gene-gene and gene-environment interactions, of multiple SNPs with small effect and of rare variants (Visscher et al., 2012; Manolio et al., 2009; Goldstein, 2009).

One promising approach uses prior information on functional structure present within the genome. This is motivated by the observation that in many cases disease states are likely to be driven by multiple genetic variants of small to moderate effect, mediated through their interaction in molecular networks or pathways (Figure 1.3), rather than by the effects of a few, highly penetrant mutations (Schadt, 2009). Where this assumption holds, the hope is that by considering the joint effects of multiple variants acting in concert, pathways gene association studies (PGAS) will reveal aspects of a disease’s genetic architecture that would otherwise be missed when considering variants individually (Wang, Li, and Hakonarson, 2010; Fridley and Biernacka, 2011). Other potential benefits of PGAS include:
1.2 The gene pathways approach

The gene pathways approach

Figure 1.3: Schematic of genes acting together in a functional pathway. Coloured nodes represent genes, and edges represent possible interactions between genes. Note that pathways can interact, and genes may belong to multiple pathways.

- the ability to accommodate genetic heterogeneity, where causal markers accumulate within genes or pathways, but vary between individuals, so that marginal effects are reduced across the sample as a whole (Holmans et al., 2009)
- the ability to compare results across studies genotyping different variants, that may nonetheless be mapped to common pathways (Cantor, Lange, and Sinsheimer, 2010; Ma and Kosorok, 2010)
- better elucidation of disease mechanisms by providing a biological interpretation of association results (Cantor, Lange, and Sinsheimer, 2010)
- identification of targets for drug or other therapeutic interventions through the identification of disease implicated pathways (Hirschhorn, 2009).

First developed in the context of gene expression studies (Mootha et al., 2003), pathways-based methods have more recently been extended to the analysis of GWAS data (Holmans et al., 2009; Luo et al., 2010; Lango Allen et al., 2010; Lambert et al., 2010). This has led
to the identification of putative causal pathways for a number of diseases including Parkinson’s Disease (Lesnick et al., 2007), Crohn’s Disease (Wang et al., 2009b) and rheumatoid arthritis (Eleftherohorinou et al., 2011).

PGAS methods rely on prior information mapping SNPs to functional networks or pathways. Since pathways are typically defined as groups of interacting genes, SNP to pathway mapping is a two-part process, requiring both the mapping of genes to pathways, and of SNPs to genes. A consistent strategy for this mapping process has however yet to be established, a situation compounded by a lack of agreement on what constitutes a pathway in the first place (Cantor, Lange, and Sinsheimer, 2010).

The number and size of databases devoted to classifying genes into pathways is growing rapidly, as is the range and diversity of gene interactions considered (see for example http://www.pathguide.org/). Databases such as those provided by KEGG (http://www.genome.jp/kegg/pathway.html), Reactome (http://www.reactome.org/) and Biocarta (http://www.biocarta.com/) classify pathways across a number of functional domains, for example apoptosis, cell adhesion or lipid metabolism; or crystallise current knowledge on specific disease-related molecular reaction networks. Strategies for pathways database assembly range from a fully-automated text-mining approach, to that of careful curation by experts. Inevitably therefore, there is considerable variation between databases, in terms of both gene coverage and consistency (Soh et al., 2010), so that the choice of database(s) will itself influence results in PGAS.

The mapping of SNPs to genes adds a further layer of complexity, since although many SNPs may occur within gene boundaries, in a typical GWAS, the vast majority of SNPs will reside in inter-genic regions (see Figure 1.1). In an attempt to include variants potentially residing in functionally significant regions lying outside gene boundaries, SNPs may be mapped to nearby genes using various distance thresholds. Various values for SNP to gene mapping distances, measured in thousands of nucleotide base pairs (kb), have been suggested in the literature, ranging from mapping SNPs to genes only if they fall within a specific gene, to the attempt to encompass upstream promoters and enhancers by extending the range to 10, 20 or even 500kb and beyond (Wang et al., 2009b; Eleftherohorinou et al., 2009; Cantor, Lange, and Sinsheimer, 2010). This process is illustrated schematically in Figure 1.4. Notable features of the SNP to pathway mapping process include the fact that
1.2 The gene pathways approach

genomes (and therefore SNPs) may map to more than one pathway, and also that many SNPs and genes do not currently map to any known pathway (Fridley and Biernacka, 2011).

![Diagram of SNP to pathway mapping process](image)

Figure 1.4: Schematic illustration of the SNP to pathway mapping process. (i) Genes (green circles) are mapped to pathways using information on gene-gene interactions (top row), obtained from a gene pathways database. Many genes do not map to any known pathway (unfilled circles). Also, some genes may map to more than one pathway. (ii) Genes that map to a pathway are in turn mapped to genotyped SNPs within a specified distance. Many SNPs cannot be mapped to a pathway since they do not map to a mapped gene (unfilled squares). Note SNPs may map to more than one gene. Some SNPs (orange squares) may map to more than one pathway, either because they map to multiple genes belonging to different pathways, or because they map to a single gene that belongs to multiple pathways.

In common with standard statistical analysis methods in GWAS (see section 1.1), the majority of existing PGAS methods begin with a univariate test of association, in which individual SNPs are scored according to their degree of association with disease status or a quantitative trait. Various techniques are then used to combine these univariate statistics into pathway scores. For example, the GenGen method (Wang, Li, and Bucan, 2007) first ranks all genes according to the value of the highest-scoring SNP within 500kb of each gene, using the SNP’s $\chi^2$ statistic. Pathway significance is then assessed by determining the degree to which high-ranking genes are over-represented in a given gene set, in comparison with the genomic background. The PLINK toolkit (Purcell et al., 2007) features a ‘set-based test’, in which pathway significance is measured by taking the average, marginal p-value of a pre-determined maximum number of ‘uncorrelated’ SNPs within the pathway. Here, uncorrelated SNPs are defined as those whose pairwise linkage disequilibrium (LD) is below a certain threshold value. As a final step, where more than one pathway is consid-
Few multivariate PGAS methods that jointly model genetic predictors currently exist, and at the time of writing, to the best of our knowledge there are none that model a quantitative response. Just as with gene or SNP mapping, a multivariate or multilocus approach to pathway mapping has the potential to increase power by fully accounting for LD structure within SNP data (Chen et al., 2010; Hoggart et al., 2008). For example, the PoDA method (Braun and Buetow, 2011) assesses pathway significance by computing a multilocus distance measure between cases and controls, from all SNPs mapped to a pathway. This is expected to emphasise epistatic interactions between SNPs, over purely marginal effects on disease status. A number of penalized regression techniques that allow prior information on the relationship between SNP markers to be incorporated into the model selection process have recently been proposed. For example, Zhou et al. (2010) group SNPs into genes, and utilise a useful property of the group lasso (Yuan and Lin, 2006) to aid the detection of rare variants within genes. The GRASS method (Chen et al., 2010) uses a combination of lasso and ridge regression to assess the significance of association between a candidate pathway and a dichotomous (case-control) phenotype. Comparisons with other PGAS methods using simulated data suggests that this regularised approach may be more powerful. Finally, Zhao et al. (2011) use a combination of PCA and lasso regression to identify pathways associated with disease status. A review of existing PGAS methods is presented in Appendix A.

In Chapter 2, we describe a group-sparse, multiple regression model, with SNPs grouped into pathways, to identify causal pathways associated with a quantitative trait. Our method, which we call ‘Pathways Group Lasso With Adaptive Weights’ (P-GLAW), includes all pathways together in a single regression model, since we expect to gain a better measure of the relative importance of different pathways by ensuring they compete against each other. Other notable features of our method include an adaptive pathway weighting procedure that accounts for factors biasing pathway selection, and the use of a subsampling procedure for the ranking of important pathways. Our method additionally takes account of the presence of SNPs overlapping multiple pathways and uses a novel combination of techniques to optimise model estimation, making it fast to run, even on whole genome datasets. We conclude Chapter 2 with a comparison study in which we compare our method with an alternative,
commonly used pathways method based on univariate SNP statistics, using real pathways and genotype data. We show that our method demonstrates high sensitivity and specificity for the detection of important pathways, showing the greatest relative gains in performance where marginal SNP effect sizes are small.

In identifying pathways associated with a quantitative trait, a natural follow-up question is to ask which SNPs and/or genes are driving pathway selection? One way to answer this question is by conducting a two-stage analysis, in which we search for important SNPs in pathways identified in an initial pathway-mapping step. Such an approach is taken by Eleftherohorinou et al. (2009), who use the lasso to identify SNPs in previously identified significant pathways. We adopt a two-step approach to SNP and gene ranking in the method and application study that we describe in Chapter 3. However, since the assumption here is that few SNPs in a pathway are contributing to any putative pathway signal, a potentially more elegant approach is to perform simultaneous, pathway and SNP selection in a single model. An existing sparse regression model, the sparse group lasso (SGL) (Simon et al., 2012), generates the required ‘dual-level’ sparsity, although it has yet to be applied to pathway and SNP selection. In Chapter 4, we develop an SGL-based sparse selection method, which we call ‘Pathways Sparse Group Lasso With Adaptive Weights’ (P-SGLAW), and examine whether ‘pathways-driven SNP selection’, that is the incorporation of information on the interaction of SNPs within gene pathways into a variable selection model, can boost the power to detect SNPs associated with a quantitative trait. In simulation studies we show that where causal variants are enriched in a particular pathway, our proposed method can aid the identification of SNPs, compared with simple lasso selection that disregards pathways information. We conclude Chapter 4 with an application study in which we investigate pathways, SNPs and genes associated with levels of HDL-cholesterol in two Asian cohorts.

1.3 Imaging genetics

In the emerging field of neuroimaging genetics, scans of the living brain obtained using PET, MRI or other imaging modalities, are used to extract phenotypes in genetic association studies (Glahn, Thompson, and Blangero, 2007). The rationale here is that the use
of heritable imaging signatures of disease, known as endophenotypes, may increase the power to detect causal variants, when compared for example with case-control phenotypes, since gene effects are expected to be more penetrant at the imaging level, rather than at the diagnostic level (Meyer-Lindenberg and Weinberger, 2006; Hibar et al., 2011a). In some respects this use of quantitative endophenotypes in imaging genetics follows similar trends in the wider field of complex trait genetics, where the search for quantitative trait loci (QTL) influencing quantitative traits is gaining momentum as a potentially more powerful way to study the underlying causes of complex disease (Plomin, Haworth, and Davis, 2009).

To date, neuroimaging genetic studies have been used to study genetic mechanisms underpinning a wide range of psychiatric and neurodegenerative disease including depression, schizophrenia and Alzheimer’s disease, and have also been used to study genetic effects on cognition in healthy populations (see Bigos and Weinberger (2010) for a review). Aside from identifying genetic factors influencing brain structure and function, the endophenotype approach can also enable the mapping of genetic effects across the brain, potentially highlighting specific brain regions where genetic effects may be concentrated (Thompson, Martin, and Wright, 2010).

Neuroimaging genetic analysis involves the search for genomic variants, for example SNPs or copy number variations, associated with quantitative measures derived from brain scans (Figure 1.5). Typical neuroimaging-derived quantitative measures include those describing the distribution of tissue types such as grey or white matter in the brain, variations in the size of anatomical structures such as cortical volume, or measures of brain connectivity (Thompson, Martin, and Wright, 2010). The search for significant associations however presents a particular statistical challenge, due to the very high dimensionality of both genotype and imaging data (Poline et al., 2010). For example, an imaging genetic study may involve a search for causal genetic associations between upwards of 500,000 SNPs and more than 100,000 voxels\(^1\). The simplest approach is one of ‘mass univariate linear modelling’ (MULM), in which each genetic predictor is tested for association against each voxel-wise phenotype (see Figure 1.6, left). However, this approach has a number of drawbacks.

---

\(^1\)Strictly speaking, the word voxel describes a single 3D picture element in a neuroimaging scan. We use the term interchangeably to describe any voxel-wise quantity (phenotype) derived from such an image. For example a scalar quantity measuring grey matter intensity or change in brain volume at a single voxel.
Firstly, the potentially huge number of hypotheses tested under a MULM framework requires the application of a very large multiple testing correction, leading to a corresponding loss of power. Many imaging genetic studies reduce the scale of this problem by decreasing the number of tested hypotheses. This can be achieved for example by narrowing the search to a set of candidate genes (Hariri et al., 2002); focusing on specific ‘regions of interest’ (ROIs) obtained by parcelling the brain into different anatomical areas (Joyner et al., 2009); or modelling single imaging measures, derived for example by taking an average intensity value across a specific ROI (Potkin et al., 2009b). Each of these strategies involves either additional assumptions about putative associations, or the discarding of potentially informative data. In contrast, a ‘hypothesis-free’ approach was taken by Stein et al. (2010b), who conducted the first brain-wide, genome-wide imaging genetics study in an investigation of SNPs associated with changes in brain structure in patients with Alzheimer’s Disease. To reduce the potentially huge amount of tested hypotheses ($10^{10}$ voxel x SNP comparisons), only the most significant SNP at each voxel was assessed, combined with
an adjustment to the multiple testing threshold to account for dependencies between SNPs. This latter adjustment follows from the fact that correlation between SNPs arising from the LD structure present in genotype data means that voxel x SNP tests are not strictly independent. The effective number of tests is therefore reduced, so that a Bonferroni-corrected threshold is likely to be over conservative. In the genetic domain, such adjustments are well established (Johnson et al., 2010). Similar adjustments accounting for voxel-voxel dependencies arising from spatial correlations in the imaging domain have also been developed (Worsley, 2003), and these are discussed further in Chapter 6.

Aside from the multiple testing problem, a second potential drawback of the MULM approach is that by treating both gene variants and voxel-wise phenotypes as independent, the inherent structure in both is ignored. With genotypes, this structure may manifest as complex patterns of LD, or from functional associations of SNPs in genes and pathways, as discussed previously. With phenotypes, structure may arise from correlations between brain regions sharing structural and/or functional features, and there is evidence that many such features are under strong genetic influence (Peper et al., 2007; Chiang et al., 2009). As with the case of genotypes discussed previously (Section 1.1), the fact that multiple phenotypes may be influenced by many of the same genetic factors means that their joint modelling can be expected to boost power (Hibar et al., 2011a; Vounou et al., 2011).

Most multilocus approaches in imaging genetics involve either the modelling of a single summary phenotypic measure, or the performing of multiple independent multilocus tests of association against multiple phenotypes. For example, Bralten et al. (2011) used a multilocus, gene-based test of association to investigate the influence of the $SORL1$ gene on hippocampal volume. Multiple linear regression models for the joint modelling of SNPs have also been employed. Problems of multicollinearity with genome-wide SNP data mean that this is often combined with some form of dimensionality reduction on the SNP data, for example using PCA to obtain orthogonal sets of predictors (Hibar et al., 2011b). This approach has the potential disadvantage of making the results of any such analysis difficult to interpret or replicate, since any genetic factors identified then consist of principal components summarising the effect of multiple predictors. As discussed previously, sparse regression techniques offer an alternative solution. Penalised regression models applied to genotype estimation for the modelling of imaging-derived univariate phenotypes include
1.3 Imaging genetics

the use of ridge regression (Kohannim et al., 2011), the lasso (Kohannim et al., 2012a) and the elastic net (Kohannim et al., 2012b), with the first reporting a boost in power compared to a univariate analysis in which each SNP is considered separately.

A number of methods that are able to jointly consider genotypes and multivariate phenotypes in a single multivariate model have been proposed. For example, multivariate latent variable models such as canonical correlation analysis (CCA), parallel independent component analysis (pICA) and partial least squares (PLS) have been used in an attempt to identify latent factors that capture some of the relationship between genetic and phenotypic domains. CCA and pICA attempt to identify orthogonal (CCA) or independent (pICA) latent factors – specifically linear combinations of variables from each domain – that maximise the correlation between imaging and genetic data. Similarly, PLS identifies orthogonal latent factors that maximise between-domain covariance. These methods have been applied to the analysis of imaging genetic data (Hardoon et al., 2009; Calhoun, Liu, and Adali, 2009; Liu et al., 2009; Jagannathan et al., 2010). Studies so far conducted have used relatively small pools of SNPs and voxels, and concerns have been raised that the power to identify latent factors may be substantially reduced with genome-wide data (Hibar et al., 2011). In addition, as with PCA, these latent variable models identify linear combinations of predictors (SNPs or voxels), making it hard to assess the importance of individual predictors or compare results across studies (Hibar et al., 2011a; Vounou, Nichols, and Montana, 2010).

Recently, variations on some of these multivariate latent variable models have been introduced that enforce sparse selection of variables within latent factors. These models, such as sparse CCA (Parkhomenko, Trichler, and Beyene, 2009; Witten, Tibshirani, and Hastie, 2009; Chen and Liu, 2011) and sparse PLS (Chun and Kele, 2010) have recently been compared in an imaging genetic context (Le Flochcor et al., 2012). A closely related model, sparse reduced-rank regression (sRRR) (Vounou, Nichols, and Montana, 2010) builds on a previous technique in multivariate regression known as reduced-rank regression (Izenman, 2008), and identifies latent factors linking genotype and phenotype predictors by constraining the rank of the matrix of estimated regression coefficients. sRRR imposes an additional regularisation penalty on the resulting coefficient estimates, for example using a lasso penalty, that enforces sparse solutions. In extensive simulations using
realistic imaging and genotypic data, sRRR demonstrated increased power to detect true genotype-phenotype associations, compared with a MULM approach (Vounou, Nichols, and Montana, 2010).

In Chapter 3, we describe a method that extends the P-GLAW method for pathways selection described in Chapter 2 to accommodate a multivariate quantitative phenotype. Our proposed ‘Pathways sparse Reduced-Rank Regression’ (PsRRR) model builds on the sRRR model, but performs pathway selection through the imposition of a group lasso penalty on genotype coefficients. We demonstrate proof of principle of the efficacy of our proposed method in simulation studies, and conclude Chapter 3 with an investigation of pathways associated with longitudinal structural change in Alzheimer’s disease, using a multivariate imaging phenotype.

1.3.1 Cluster size inference

Despite recent interest in multivariate methods, MULM remains the most common approach to imaging genetics analysis. MULM techniques are also used widely for the mapping of human brain function, for example in analysing fMRI to identify areas of heightened activation during task-related activities (Ashburner et al., 2006). The primary output of a MULM analysis is then a list of test statistics, describing the extent of association between each SNP or task at each voxel. These are often displayed as a statistical map (see Figure 1.6, middle and right).

The structural and functional architecture of the brain means that association signals may often extend across neighbouring voxels to form clusters of heightened activation. For this reason, statistical tests based on contiguous regions of heightened activation have been developed. These can offer increased power over voxel-wise tests, where for example activation at each voxel in a cluster does not on its own achieve statistical significance, but inference based on cluster-wise activation does (Poline et al., 1997). Cluster size inference works by first thresholding the statistical image at a particular significance level, and then comparing the size of any resulting super-threshold clusters with that expected by chance using random field theory (RFT) (Friston et al., 1996). Since the number of clusters is typically much smaller than the number of voxels, this method has the advantage of requir-
1.4 Mathematical notation and nomenclature

Throughout this thesis we use the terms predictor, genotype, genetic variant and SNP (single nucleotide polymorphism) interchangeably. Likewise for the terms quantitative trait, response and phenotype, and the terms pathway and group.

We assume all observations of genotypes and phenotypes are from a random sample of $N$ unrelated individuals, drawn from the same population. Matrices are denoted by bold capital letters, and vectors by bold lower case. We additionally adopt the following notation for observed genotypes and phenotypes, and for gene pathways:
Introduction 23

Genotypes

We assume that minor allele counts at $P$ SNPs, $x_1, \ldots, x_P$, are recorded for $N$ subjects and denote by $x_{ij} \in \{0, 1, 2\}$, the observed minor allele count for SNP $j$ on individual $i$. These observed minor allele counts are arranged in an $(N \times P)$ SNP genotype matrix, $X$. We denote by $x_j = (x_{1j}, \ldots, x_{Nj})'$, the column vector of minor allele counts corresponding to SNP $j$, so that $X = (x_1, \ldots, x_P)$, and by $x_i = (x_{i1}, \ldots, x_{iN})$ the vector of minor allele counts corresponding to the $i$th row of $X$.

Phenotypes

We assume $N$ observations of a $Q$-dimensional multivariate quantitative trait or phenotype, $y_1, \ldots, y_Q$. The observed values for phenotype $q$ are arranged in an $(N \times 1)$ column vector, $y_q$, and the $Q$ phenotypes are arranged in an $(N \times Q)$ phenotype matrix $Y = (y_1, \ldots, y_Q)$. In the case that $Q = 1$, we drop the subscript and denote the vector of observations for the corresponding univariate phenotype $y$ by $y$.

Pathways

We frequently assume that SNPs may be mapped to $L$ groups or gene pathways, $G_l \subset \{1, \ldots, P\}$, $l = 1, \ldots, L$, using prior information extracted from a pathways database. We denote by $P_l$, the number of SNPs in group $G_l$. Groups are sometimes assumed to be disjoint, so that $G_l \cap G_{l'} = \emptyset$ for any $l \neq l'$. We state clearly where this is or is not the case. For simplicity, in the text (but not in equations) we occasionally refer to ‘pathway $l’$, by which we mean the pathway indexed by $l$, that is $G_l$.

For convenience, we index the SNPs in group $G_l$ by $l_1, \ldots, l_{P_l}$, while noting that each index has a unique one-to-one mapping to the original set of SNP indices, so that $l_z \in \{1, \ldots, P\}$ for $z = 1, \ldots, P_l$. We denote by $X_l = (x_{l_1}, \ldots, x_{l_{P_l}})$ the $(N \times P_l)$ matrix of minor allele counts corresponding to SNPs in pathway $G_l$.

Throughout the text, we denote the cardinality, or number of elements in a set $S$, by $|S|$. Finally we denote the $\ell_q$ norm of a $p$-dimensional vector $v$ as $||v||_q = (\sum_{j=1}^p |v_j|^q)^{\frac{1}{q}}$, although we occasionally denote the $\ell_1$ norm as $|v|$ for simplicity.
1.5 Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCGD</td>
<td>block coordinate gradient descent</td>
</tr>
<tr>
<td>CGD</td>
<td>coordinate gradient descent</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>(f)MRI</td>
<td>(functional) magnetic resonance imaging</td>
</tr>
<tr>
<td>FWER</td>
<td>family-wise error rate</td>
</tr>
<tr>
<td>FWHM</td>
<td>full-width at half-maximum</td>
</tr>
<tr>
<td>GL</td>
<td>group lasso</td>
</tr>
<tr>
<td>GWAS</td>
<td>genome-wide association study/studies</td>
</tr>
<tr>
<td>LD</td>
<td>linkage disequilibrium</td>
</tr>
<tr>
<td>MAF</td>
<td>minor allele frequency</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MLR</td>
<td>multiple linear regression</td>
</tr>
<tr>
<td>MULM</td>
<td>mass univariate linear modelling</td>
</tr>
<tr>
<td>OLS</td>
<td>ordinary least squares</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PET</td>
<td>positron emission tomography</td>
</tr>
<tr>
<td>P-GLAW</td>
<td>pathways group lasso with adaptive weights</td>
</tr>
<tr>
<td>PGAS</td>
<td>pathways gene association study/studies</td>
</tr>
<tr>
<td>P-SGLAW</td>
<td>pathways sparse group lasso with adaptive weights</td>
</tr>
<tr>
<td>PsRRR</td>
<td>pathways sparse reduced-rank regression</td>
</tr>
<tr>
<td>QTL</td>
<td>quantitative trait locus/loci</td>
</tr>
<tr>
<td>QTT</td>
<td>quantitative trait test</td>
</tr>
<tr>
<td>ROI</td>
<td>region of interest</td>
</tr>
<tr>
<td>RFT</td>
<td>random field theory</td>
</tr>
<tr>
<td>SGL</td>
<td>sparse group lasso</td>
</tr>
<tr>
<td>SGL-CGD</td>
<td>sparse group lasso - coordinate gradient descent</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>sRRR</td>
<td>sparse reduced-rank regression</td>
</tr>
<tr>
<td>VBM</td>
<td>voxel-based morphometry</td>
</tr>
</tbody>
</table>
Chapter 2

Identifying pathways associated with a univariate quantitative trait: ‘Pathways Group Lasso with Adaptive Weights’

We now turn to the problem of developing a multiple regression model for the identification of biological pathways associated with a quantitative trait. As outlined in Section 1.2, our assumption is that where causal SNPs are enriched in a pathway, an approach that includes all SNPs in a single, sparse multiple regression model will have increased power, compared to a more conventional approach in which SNPs are considered one at a time. To this end, we use a modified version of a sparse regression model known as the group lasso (GL) with SNPs grouped into pathways. We believe this is the first method able to jointly model the genome-wide, group-wise association of gene variants with a quantitative trait.

We face a number of challenges in applying GL to SNP and pathway data for the identification of implicated pathways. These include the fact that pathways overlap, since many SNPs map to multiple pathways; the problem of selection bias, that is the tendency of the model to select pathways having specific statistical properties irrespective of their association with the phenotype; and the problem of assessing pathway importance in a finite sample. In addition, the potentially very large size of SNP datasets makes the development of strategies for computationally efficient model estimation a necessity. In following sections we outline our approach to tackling each of these challenges.
2.1 Penalised regression approaches in gene mapping

The chapter is arranged as follows. We begin in Section 2.1, with a review of the most common penalised regression models used in gene mapping – ridge regression and the lasso and variations thereof. In Section 2.2 we describe the group lasso model, and explain its relevance to the selection of pathways in PGAS. We also describe our strategy for dealing with the fact that pathways overlap. In Section 2.3, we describe an estimation algorithm for the GL with overlaps, along with a number of strategies we employ to speed up the estimation process for the large datasets in PGAS. In Section 2.4 we outline our approach to reducing potential bias in pathway selection using an adaptive weight tuning procedure, and in Section 2.5 we explain a method for ranking pathways in order of importance, using a subsampling procedure that exploits finite sample variability.

In Section 2.6, we evaluate our method, which we call ‘Pathways Group Lasso with Adaptive Weights’ (P-GLAW) in a simulation study, using real genetic and pathway data with quantitative phenotypes simulated under an additive genetic model. We feel the use of real genotype and pathway data is crucial, so as to capture the distributions of gene size and number within a pathway, together with the complex SNP LD patterns and overlaps between pathways, all of which may have a significant effect on pathway ranking performance. We consider a range of scenarios with different causal SNP distributions and effect sizes, and compare our method with a popular, existing PGAS method based on combining multiple, univariate SNP-phenotype association statistics. To our knowledge, this is the first such power study comparing multivariate and univariate approaches using real SNP and pathway data. We conclude the chapter with a discussion in Section 2.7.

2.1 Penalised regression approaches in gene mapping

We return here to the problem, introduced in section 1.1, of estimating $P$ SNP regression coefficients associated with a univariate, quantitative trait in a multiple linear regression model. In what follows, we assume that the response vector, $y$ is mean-centred, so that $\sum_{i=1}^{N} y_i = 0$. We additionally assume that the observed SNP allele count vectors are mean-centred and of unit length, so that $\sum_{i=1}^{N} x_{ij} = 0$ and $\sum_{i=1}^{N} x_{ij}^2 = 1$ for $j = 1, \ldots, P$. 
The standard, multiple linear regression model is given by

\[ y = X\beta + \epsilon, \]

where the errors, \( \epsilon_i, i = 1, \ldots, N \) are assumed to be independent and identically distributed, with variance \( \sigma^2 \). When the design matrix, \( X \), is of full rank, an unbiased estimate for the regression coefficient vector, \( \beta \) is obtained by minimising the residual sum of squares \( ||y - X\beta||^2 \) to give

\[ \hat{\beta}_{OLS} = (X'X)^{-1}X'y \]

with variance

\[ \text{Var}(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}. \]

Two problems immediately arise in the context of gene mapping. Firstly, multicollinearity due to LD between SNP predictors can make the inverse \( (X'X)^{-1} \) very sensitive to small changes in the data. This makes the resulting estimates for individual SNP coefficients unreliable, undermining the original goal of identifying predictors with functional significance (Malo, Libiger, and Schork, 2008). Secondly, in gene mapping we are frequently faced with the situation where there are more predictors than subjects. In this case \( X \) is rank deficient, \( X'X \) is not invertible, and no unique estimates for \( \beta \) exist.

One possible solution is to take a model selection approach, in which we attempt to pick a subset of SNPs that best predicts the response. This strategy is used in a number of stepwise variable selection methods in which subsets of predictors are successively either included or excluded from the model, depending on their effect on the model’s goodness of fit. The discrete nature of these approaches however makes them unreliable, since at each step the model is optimised for a subset of predictors only. More exhaustive model selection methods such as all subsets regression also suffer from bias, and are computationally feasible only for relatively small numbers of predictors (Hesterberg et al., 2008).

In contrast, continuous shrinkage or penalised regression methods have been shown to provide more reliable coefficient estimates than the discrete methods described above (Hastie, Tibshirani, and Friedman, 2008). These methods work by adding some form of
2.1 Penalised regression approaches in gene mapping

regularisation to the estimation of $\beta$. This typically involves the imposition of a constraint or penalty on $\beta$, so that

$$\hat{\beta} = \arg \min_{\beta} \| y - X\beta \|^2_2 \quad \text{subject to} \quad P(\beta) < c$$  (2.2)

where the penalty function $P(\beta)$ is some function of the regression coefficients, and $c > 0$. An equivalent expression is

$$\hat{\beta} = \arg \min_{\beta} \left\{ \| y - X\beta \|^2_2 + \lambda P(\beta) \right\}$$  (2.3)

with $\lambda > 0$, where $\lambda$ is a tunable parameter controlling the amount of regularisation or penalisation to be applied. Note there is a one-to-one correspondence between $c$ and $\lambda$ (Hastie, Tibshirani, and Friedman, 2008).

A common choice for $P(\beta)$ is a function of the $\ell_\gamma$-norm of the coefficient vectors,

$$P(\beta) = \| \beta \|_\gamma^\gamma = \sum_{j=1}^{P} |\beta_j|^\gamma$$  (2.4)

where $\gamma > 0$. Depending on the choice of $\gamma$, the resulting estimates for the SNP coefficients exhibit some potentially useful properties. Moreover, with $\gamma \geq 1$, the penalty (2.4) has the attractive property of being convex (strictly convex when $\gamma > 1$), and as we shall see, this makes the resulting convex optimisation problem (2.3) amenable to efficient solutions.

We now consider the two paradigm cases in penalised regression, with $\gamma = 1$ and $\gamma = 2$.

2.1.1 Ridge regression

The case with $\gamma = 2$ is known as ridge regression (Hoerl and Kennard, 1970). The ridge estimator (2.2) for $\beta$ then satisfies

$$\hat{\beta}^{\text{ridge}} = \arg \min_{\beta} \| y - X\beta \|^2_2 \quad \text{subject to} \quad \sum_{j=1}^{P} \beta_j^2 < c$$  (2.5)
To better understand some of the properties of the ridge estimator, we begin by considering the simple case of two predictors. The optimisation (2.5) then has a simple geometric interpretation, illustrated in Fig. 2.1. In the figure, dashed lines represent lines of increasing sum of squared residual error, \( ||y - X\beta||_2^2 \) as \( \hat{\beta} \) moves away from its optimum value \( \hat{\beta}^{OLS} = \arg \min_\beta ||y - X\beta||_2^2 \). The ridge penalty ensures that \( \hat{\beta}^{ridge} \) is constrained to lie within the region bounded by the blue circle, \( \beta_1^2 + \beta_2^2 = c \). Estimates for \( \hat{\beta}_1^{ridge} \) and \( \hat{\beta}_2^{ridge} \) satisfying (2.5) then correspond to the blue point in Fig. 2.1. The effect of the ridge penalty is thus to shrink coefficient estimates, compared to their OLS values, with the amount of shrinkage increasing with decreasing \( c \) (or equivalently increasing \( \lambda \)). This shrinkage property can be very useful in the case of correlated predictors. For example, where two predictors are highly correlated, OLS estimation can result in large variations in estimated coefficients, where a large positive coefficient for one predictor is cancelled by a similarly large negative coefficient for the other. Such variations are constrained by the ridge penalty, which exhibits a ‘grouping effect’ that ensures correlated predictors have similar coefficients, potentially making results easier to interpret. This shrinkage property can also lead to superior prediction accuracy through a bias-variance trade off (Hastie, Tibshirani, and Friedman, 2008; Malo, Libiger, and Schork, 2008).

An equivalent form for the ridge optimisation (2.5) is

\[
\hat{\beta}^{ridge} = \arg \min_\beta \left\{ ||y - X\beta||_2^2 + \lambda \sum_{j=1}^P \beta_j^2 \right\}
\]

with solution

\[
\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1}X'y
\]

(2.6)

where \( I \) is the \( (P \times P) \) identity matrix. By comparing (2.6) with (2.1), we see how ridge regression enables the estimation of coefficients in the \( P > N \) case, where \( X \) is rank deficient, since the addition of a positive constant to the diagonal of \( X'X \) renders the problem non-singular, even when \( X'X \) is not of full rank (Hastie, Tibshirani, and Friedman, 2008).

Several recent studies have exploited properties of the ridge estimator highlighted above in the identification of important SNPs in GWAS, including in an imaging genetics context (Sun et al., 2009; Kohannim et al., 2011).
2.1 Penalised regression approaches in gene mapping

Figure 2.1: Coefficient estimation with two predictors, $\beta_1$ and $\beta_2$. Dashed lines represent contours of constant sum of squared residual error, $||y - X\beta||^2_2$. With ridge regression, the constraint is $\beta_1^2 + \beta_2^2 < c$ (blue circle) and the estimated coefficients $\hat{\beta}_{\text{ridge}}$ correspond to the blue dot. With lasso regression, the constraint is $|\beta_1| + |\beta_2| < c$ (red diamond) and the estimated coefficients $\hat{\beta}_{\text{lasso}}$ correspond to the red dot.
2.1.2 Variable selection with the lasso

The case with $\gamma = 1$ is known as lasso regression (Tibshirani, 1996). The lasso estimator for $\beta$ satisfies

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} ||y - X\beta||_2^2 \text{ subject to } \sum_{j=1}^{P} |\beta_j| < c. \quad (2.7)$$

Once again considering the simple case with two predictors illustrated in Fig. 2.1, with the lasso $\beta$ is constrained to lie within the region bounded by the red diamond, $|\beta_1| + |\beta_2| = c$. Estimates for $\hat{\beta}_1^{\text{lasso}}$ and $\hat{\beta}_2^{\text{lasso}}$ then correspond to the red point in Fig. 2.1. We see that as $c$ approaches zero (or equivalently $\lambda$ moves away from zero), the contours of the OLS estimation are increasingly likely to intersect the red diamond at one of the axes, meaning that one of the coefficients, $\beta_1$ or $\beta_2$ is set to zero. The lasso thus performs a form of variable or model selection, in the sense that a single variable is ‘selected’ by the model.

This example readily extends to the $P > 2$ case, with the lasso producing sparse estimates for $\hat{\beta}_{\text{lasso}}$, in the sense that multiple predictors have their coefficients, $\hat{\beta}_j^{\text{lasso}}$, set to zero. The degree of sparsity is controlled by $c$ or $\lambda$, such that an increasing number of predictors have zero coefficients as $c$ is reduced, or equivalently $\lambda$ is increased. We can then think of the lasso as selecting a subset of predictors, $\hat{S} = \{j : \beta_j \neq 0, j = 1, \ldots, P\}$.

In high-dimensional, large $P$ settings typical encountered in GWAS and gene expression analysis, variable selection methods using penalised regression are attractive for a number of reasons. Firstly, a parsimonious model that selects a small number of important predictors is easier to interpret. In any case, in SNP mapping, we expect only a small set, $S \subset \{1, \ldots, P\}$, of all predictors to be truly associated with the phenotype, so that the assumptions underlying sparse models such as the lasso are likely to hold. Secondly, as with ridge regression, the shrinkage property can improve a model’s capacity to predict outcomes with new data, reducing variance at the expense of slightly increased bias (Hastie, Tibshirani, and Friedman, 2008). Finally, in contrast to the discrete subset selection methods described previously, penalised regression models such as the lasso perform continuous variable selection, and thus tend to select the set of true predictors with greater reliability (Hesterberg et al., 2008).
2.1 Penalised regression approaches in gene mapping

An important issue with all penalised regression models is the need to choose a value for the regularisation parameter, $\lambda$. For sparse regression models such as the lasso this determines the number of variables selected by the model. A common strategy is to use $K$-fold cross validation to choose a $\lambda$ value that minimises the prediction error between training and test datasets. One drawback of this approach is that it focuses on optimising the size of the set, $\hat{S}$, of selected variables that minimises the cross validated prediction error. Since the variables in $\hat{S}$ will vary across each fold of the cross validation, this procedure may not be a good means of establishing the importance of a unique set of variables (Vounou et al., 2011). Other approaches include the use of information-theoretic metrics such as the Akaike or Bayesian information criteria (AIC and BIC), that trade model goodness of fit against model complexity (degrees of freedom) (Zou, Hastie, and Tibshirani, 2007). Data resampling or bootstrapping techniques have also been demonstrated to improve model consistency, in the sense that the true model support is selected with a high probability (Bach, 2008a; Meinshausen and Bühlmann, 2010). We expand on this latter approach in following sections.

We now turn to the problem of estimating coefficients with lasso penalised regression. An equivalent form for the lasso optimisation (2.7) is

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \frac{1}{2} || y - X\beta ||^2_2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\} \quad (2.8)$$

We first consider the solution for a single predictor, $j$. The lasso estimator then minimises

$$f_j(\beta_j) = \frac{1}{2} \sum_{i=1}^{N} (y_i - x_{ij}\beta_j)^2 + \lambda |\beta_j|$$

$$= \frac{1}{2} \sum_{i=1}^{N} y_i^2 - \beta_j \sum_{i=1}^{N} (x_{ij}y_i) + \frac{1}{2} \beta_j^2 + \lambda |\beta_j| \quad (2.9)$$

(since $\sum_{i=1}^{N} x_{ij}^2 = 1$). Differentiating with respect to $\beta_j$, and noting from (2.1), that the unconstrained OLS estimate for a single predictor, $j$, $\hat{\beta}_{j,\text{OLS}} = \sum_i x_{ij}y_i$, we see that (2.9) is minimised when

$$\hat{\beta}_j = \hat{\beta}_{j,\text{OLS}} - \lambda \frac{d|\hat{\beta}_j|}{d\hat{\beta}_j}.$$
When $\hat{\beta}_j \neq 0$, this has solutions $\hat{\beta}_j = \hat{\beta}_j^{OLS} - \lambda \text{sign}(\hat{\beta}_j)$ \footnote{sign($z$) has the value $+1$ when $z > 0$ and $-1$ when $z < 0$.}. When $\hat{\beta}_j = 0$, the derivative $d|\hat{\beta}_j|/d\beta_j$ is not defined. We instead form the subgradient or subdifferential, $d\hat{\beta}_j/d\beta_j \in [-1,1]$ (Bertsekas, 1999), so that $\lambda \geq |\hat{\beta}_j^{OLS}|$. More concisely, the effect of the lasso can then be summarised as follows:

$$
\hat{\beta}_j^{lasso} = S(\hat{\beta}_j^{OLS}, \lambda), \quad j = 1, \ldots, P
$$

where

$$
S(\hat{\beta}_j^{OLS}, \lambda) = \begin{cases} 
\hat{\beta}_j^{OLS} - \lambda & \text{if } \hat{\beta}_j^{OLS} > 0 \text{ and } |\hat{\beta}_j^{OLS}| > \lambda \\
\hat{\beta}_j^{OLS} + \lambda & \text{if } \hat{\beta}_j^{OLS} < 0 \text{ and } |\hat{\beta}_j^{OLS}| > \lambda \\
0 & \text{if } |\hat{\beta}_j^{OLS}| \leq \lambda 
\end{cases}
$$

(2.10)

Here we see clearly that the effect of the lasso is to shrink larger OLS coefficients towards zero. OLS coefficients below a size threshold determined by $\lambda$ are set to zero. For this reason the lasso is often described as performing ‘soft thresholding’ of OLS estimates, and $S(\hat{\beta}_j^{OLS}, \lambda)$ is known as the soft thresholding operator.

For multiple, uncorrelated predictors, that is where $X'X = I_P$, lasso estimates for each predictor are obtained by applying the above soft thresholding procedure to each predictor in turn, so that

$$
\hat{\beta}_j^{lasso} = S(\hat{\beta}_j^{OLS}, \lambda), \quad j = 1, \ldots, P.
$$

Where this is not the case, that is where predictors are correlated, we can instead solve the lasso estimation (2.7) by taking the partial derivatives with respect to each $\beta_j$, that is by minimising

$$
f(\beta_j) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{k \neq j} x_{ik}\hat{\beta}_k - x_{ij}\hat{\beta}_j)^2 + \lambda \sum_{k \neq j} |\hat{\beta}_k| + \lambda |\beta_j| 
$$

(2.11)

with respect to $\beta_j$, where $\hat{\beta}_k, (k \neq j)$ are the current coefficient estimates for all predictors other than $j$. However, estimates for $\beta$ cannot then be obtained directly, since each $\beta_j$ depends on all other estimates $\hat{\beta}_k, k \neq j$. A number of estimation algorithms to solve the
2.1 Penalised regression approaches in gene mapping

lasso in the non-orthogonal case have been proposed (Tibshirani, 1996; Efron et al., 2004).
The method of coordinate descent (Friedman et al., 2007) is particularly attractive due to
its simplicity and speed.

Coordinate descent algorithms (and related methods such as gradient descent (Kim and
Kim, 2004)) offer a means of solving convex optimisation problems in which the estimation
is solved for each predictor in turn, while keeping current estimates for all other coefficients
fixed at their current values, in an iterative fashion (Tseng, 1988). Thus by comparing (2.11)
with the single predictor case (2.9), we see that the coordinate wise update for \( \beta_j \) is the same
as the lasso estimation for a single predictor, but with the OLS estimate, \( \beta_j \overset{OLS}{\rightarrow} \) replaced
with the ‘partial residual’ \( \sum_{i=1}^{N} x_{ij} (y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k) \), so that

\[
\hat{\beta}_j \overset{\text{lasso}}{\leftarrow} S \left( \sum_{i=1}^{N} x_{ij} (y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k), \lambda \right).
\]

It is convenient to express the partial residual in terms of the current full residual, \( \hat{r} = \sum_{i=1}^{N} (y_i - \sum_{k=1}^{P} x_{ik} \hat{\beta}_k) \), so that,

\[
\sum_{i=1}^{N} x_{ij} (y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k) = \sum_{i=1}^{N} x_{ij} (\hat{r}_i + x_{ij} \hat{\beta}_j)
= \sum_{i=1}^{N} (x_{ij} \hat{r}_i) + \hat{\beta}_j.
\]  \hspace{1cm} (2.12)

A final expression for the coordinate-wise lasso update for predictor \( j \) is then given by

\[
\hat{\beta}_j \overset{\text{lasso}}{\leftarrow} S \left( \sum_{i=1}^{N} (x_{ij} \hat{r}_i) + \hat{\beta}_j, \lambda \right).
\]

Coordinate descent then proceeds by initialising \( \hat{\beta} \) (e.g. using \( \hat{\beta} = 0 \)), and updating each
component \( j = 1, \ldots, P, 1, \ldots, P, \ldots \) until some convergence criterion is met.

Lasso regression has been used for the identification of genetic predictors associated
with both quantitative and dichotomous (case-control) traits (Shi et al., 2011; Wu et al.,
2009). For the latter, Wu et al., 2009 use a log-likelihood loss function \( L(\beta) \) derived from
a logistic regression model with a lasso penalty to analyse two-way and higher-order SNP-
SNP interactions in a genome-wide dataset. For subject $i$, disease status is labelled as $y_i = 1$ for cases and $y_i = 0$ for controls, and estimates for $\beta$ are then obtained by solving

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ L(\beta) + \lambda \sum_{j=1}^{P} |\beta_j| \right\}$$

where

$$L(\beta) = -\sum_{i=1}^{N} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)$$

(2.13)

and $\pi_i = Pr(y_i=1)$ is given by

$$\frac{e^{x_i'.\beta}}{1 + e^{x_i'.\beta}}.$$ 

The choice of $\lambda$ is driven by the desired number of SNPs to be identified. SNP-SNP and higher order interactions are analysed only for those SNPs previously selected by the model. Hoggart et al. (2008) propose a similar method for SNP selection in a Bayesian context. They use a Laplace or double exponential prior that performs model selection in a manner similar to the lasso.

In contrast to ridge regression, the lasso is not strictly convex. This means that where for example two identical predictors are included in the model, the lasso does not give a unique solution, whereas the ridge penalty ensures that identical predictors have identical coefficients. In practice, where predictors, or groups of predictors are highly correlated, the lasso tends to select one predictor at random, rather than grouping coefficients together, as with ridge regression (Zou and Hastie, 2005). This is an example of inconsistency in model selection, meaning that the lasso may not asymptotically select the true model (Zhao and Yu, 2006). To improve consistency, Zou (2006) propose the adaptive lasso, a two-step procedure in which the regularisation parameter, $\lambda$, is allowed to vary, with $\lambda_j, (j = 1, \ldots, P)$, depending on OLS estimates for each variable. Bootstrap or resampling techniques that attempt to recover the true model by considering multiple subsamples of the data have also been found to improve consistency (Chatterjee and Lahiri, 2011; Meinshausen and Bühlmann, 2010), and we apply this approach in methods that we develop below, and in following chapters.

In gene mapping, where for example multiple SNPs in tight LD might tag a causal
2.2 The group lasso for pathway selection

variant, the tendency of the lasso to select one of multiple correlated predictors may not be
desirable (Cho et al., 2009a). One solution proposed by Zou and Hastie (2005) is the elastic
net, which combines lasso and ridge penalties in a single model, enabling the selection of
groups of correlated variables, for example groups of SNPs in strong LD. This model has
been used in a number of recent gene mapping studies (Cho et al., 2009a; Do et al., 2011;
Guzzetta, Jurman, and Furlanello, 2010). A drawback of this approach is its incorporation
of an extra regularisation parameter, potentially requiring more computationally intensive
methods for the tuning of parameters.

Other interesting extensions of the lasso include the fused lasso (Tibshirani et al., 2005)
which penalises differences between adjacent, ordered predictors, and the graphical lasso
which estimates undirected graphs for continuous data by applying a lasso penalty to the
inverse covariance matrix (Friedman, Hastie, and Tibshirani, 2008).

2.2 The group lasso for pathway selection

We now turn to the description of a penalised regression model able to identify gene path-
ways associated with a univariate quantitative trait, $y$. We initially consider the situation
where SNPs are partitioned into $L$ mutually exclusive pathways, or groups. Each group $G_l$,
for $l = 1, \ldots, L$, is a subset of $\{1, 2, \ldots, P\}$ of cardinality $P_l$, such that $G_l \cap G_{l'} = \emptyset$ for
any $l \neq l'$. We denote by $G = \{1, \ldots, P\}$, the set of all SNP indices. For convenience, we
denote the indices of SNPs in $G_l$ by $l_1, l_2, \ldots, l_{P_l}$.

In general, where $P$ is large, we expect only a small proportion of SNPs to be causal, in
the sense that they influence the phenotype. We further assume that this set $S \subset \{1, \ldots, P\}$
of causal SNPs will tend to be enriched within functional groups, or gene pathways. This
latter assumption is illustrated schematically in Fig. 2.2, where causal SNPs (marked in
grey) tend to accumulate within a small number of causal pathways, while the majority of
pathways contain no causal SNPs. A model generating such a sparsity pattern is said to be
group-sparse, in that SNPs affecting $y$ are to be found in a set $C \subset \{1, \ldots, L\}$ of causal
gene pathways (groups), with $|C| \ll L$, where $|C|$ denotes the cardinality of $C$.

We assume that $y$ can be optimally predicted, in the least squares sense, by a linear
combination of allele counts corresponding to SNPs in $S$. If we denote the parameter
Figure 2.2: Group-sparse distribution of causal SNPs. The set $S \subset \{1, \ldots, P\}$ of causal SNPs influencing the phenotype are represented by boxes that are shaded grey. Causal SNPs are assumed to occur within a set $C$ of causal pathways. Here $C = \{2, 3\}$. Note that the particular distribution of causal SNPs may vary for each individual, $i = 1, \ldots, N$. The group sparsity assumption is that $|C| \ll L$.

A penalised regression model enforcing group sparsity is the group lasso (GL) (Yuan and Lin, 2006). This imposes a constraint on the $\ell_2$ norm of the group or pathway coefficient vectors, so that the penalty function (2.4) is given by

$$P(\beta) = \sum_{l=1}^{L} w_l ||\beta_l||_2$$

where $w_l$ is a pathway weighting factor that allows the constraint to vary between groups. The corresponding group lasso estimator is then given by

$$\hat{\beta}^{GL} = \arg\min_{\beta} \left\{ \frac{1}{2}||y - X\beta||_2^2 + \lambda \sum_{l=1}^{L} w_l ||\beta_l||_2 \right\}.$$  

(2.15)
2.2 The group lasso for pathway selection

The group lasso can be thought of as imposing an intermediate or ‘hybrid’ $\ell_1/\ell_2$ penalty. Sparsity at the pathway level is encouraged through the imposition of an $\ell_1$ lasso penalty on $||\beta_l||_2$, which ensures that SNPs belonging to pathways not selected by the model have zero regression coefficients. For selected pathways, i.e. those with $\hat{\beta}_l \neq 0$, SNP coefficients tend to shrink, through the imposition of a ridge-type penalty on $||\beta_l||_2$. All SNPs within selected pathways are retained by the model, that is $\hat{\beta}_j \neq 0$, for $j = l_1, \ldots, l_P$. The degree of group sparsity is controlled by the regularisation parameter, $\lambda$, such that the number of pathways selected by the model increases with decreasing $\lambda$. For a given $\lambda$, the group sparsity pattern is determined both by the data ($y$ and $X$), and by the distribution of pathway weights, $w = (w_1, \ldots, w_l)$, such that an increase in $w_l$ means that pathway $l$ is less likely to be selected, whereas a decrease in $w_l$ will have the opposite effect. The GL optimisation problem (2.15) is convex, and as with the lasso can be solved using group or block-wise coordinate descent methods (see Section 2.3.1). Meier et al. (2008) describe a variation of the GL model described above for a binary response (applicable for example to case-control data) that utilises a logistic loss function (2.13) in place of the least squares loss in (2.15).

The GL with its group sparsity property falls into a wider class of models imposing structured sparsity, using prior information on the relationship between variables (Jenatton and Bach, 2011). This extra information can be encoded in the form of group mappings, as with GL, or for example as graphs or trees describing hierarchical relationships that can be used to drive variable selection (Zhao, Rocha, and Yu, 2009; Huang, Zhang, and Metaxas, 2011; Jenatton and Bach, 2011; Kim and Xing, 2012). Alongside the use of functional information, such as the grouping of SNPs and genes into gene pathways considered here, other choices for driving structured sparsity include the use of spatial information and connectivity maps in neuroimaging applications (Jenatton et al., 2011; Kong et al., 2010). A further example of hierarchical selection is described by Bien, Taylor, and Tibshirani (2012), who use a hierarchical, sparse interaction model in which interacting pairs of variables are selected only if both variables are marginally important.
2.2.1 The problem of overlapping pathways

The group lasso in its original formulation rests on the assumption that groups are disjoint. Problems arise however in the situation where pathways overlap, that is where one or more SNPs map to more than one pathway, so that \( G_l \cap G_{l'} \neq \emptyset \) for some \( l \neq l' \). This situation is commonplace when mapping SNPs to pathways (see Fig. 1.4). Firstly, where groups overlap, the penalty term (2.14) is no longer separable into groups, and convergence using coordinate descent is no longer guaranteed (Tseng and Yun, 2009). Secondly, if we wish to be able to select pathways independently, GL is unable to do this. We illustrate this last point using a simple example in Fig. 2.3 A, where we consider three pathways, \( G_1, G_2 \) and \( G_3 \), two of which overlap. As a consequence of this, pathway parameter vectors \( \beta_1 \) and \( \beta_2 \) also overlap, since they have a number of SNPs in common (shaded dark grey). If a shared SNP is selected (i.e. it has a non-zero coefficient), then both pathways to which it belongs (\( G_1 \) and \( G_2 \)) are also selected, since their corresponding pathway parameter vectors have non-zero \( \ell_2 \) norms. The GL regression model thus does not meet our requirements, since in order to be able to rank pathways in order of importance, we wish to be able to distinguish overlapping pathways and select them independently. Conversely, where shared SNPs have zero coefficients, for example in the case that \( G_1 \) is not selected in the model, then these SNPs will have zero coefficients in each and every pathway to which they belong (here \( G_1 \) and \( G_2 \)). Hence SNPs retained in the model are necessarily drawn from the complement of a union of (unselected) pathways. We instead require retained SNPs to be drawn from a union of (selected) pathways, so that a SNP driving selection in one pathway may still have a zero coefficient in another.

A number of strategies to deal with the problem of overlapping groups have been proposed (Mosci et al., 2010; Liu and Ye, 2010). One method, motivated by the analysis of gene expression data, is to duplicate SNP predictors in \( X \), so that SNPs belonging to more than one pathway can enter the model separately (see Fig. 2.3 B) (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011). The process works as follows. An expanded design matrix is formed from the column-wise concatenation of the \( L, (N \times P_l) \) sub-matrices, \( X_{G_l} \), where \( X_{G_l} = \{ x_{ij} : i = 1, \ldots, N, j \in G_l \} \), to form the expanded design matrix \( X^* = [X_{G_1}, X_{G_2}, \ldots, X_{G_L}] \) of size \( (N \times P^*) \), where \( P^* = \sum_l P_l \). The correspond-
2.2 The group lasso for pathway selection

Figure 2.3: The problem of overlapping pathways, with three pathways, $G_1$, $G_2$, and $G_3$, two of which overlap. A: Standard formulation. Pathway parameter vectors $\beta_1$ and $\beta_2$ overlap, since they have SNPs in common (shaded dark grey). Where an overlapping SNP has a non-zero coefficient, only $G_3$, can be selected independently. B: Formulation with duplicated SNPs. An expanded parameter vector, $\beta^*$, is created by duplicating overlapping SNPs (dotted line). SNPs in $\beta^*_1$ and $\beta^*_2$ now enter the model separately, so that pathways can be selected independently.

ing $P^* \times 1$ parameter vector, $\beta^*$, is formed by joining the $L, (P_l \times 1)$ pathway parameter vectors, $\beta^*_l$, so that $\beta^* = [\beta^*_1, \beta^*_2, \ldots, \beta^*_L]'$. Pathway mappings with SNP indices in the expanded variable space are reflected in updated groups $G^*_1, \ldots, G^*_L$. The GL estimator (2.15), adapted to account for overlapping groups, is then given by

$$\hat{\beta}^{GL^*} = \arg\min_{\beta} \left\{ \frac{1}{2} ||y - X^*\beta^*||_2^2 + \lambda \sum_{l=1}^L w_l ||\beta^*_l||_2 \right\}.$$  

With this overlap expansion, the model is then able to perform pathway selection in the way that we require, and the corresponding optimisation problem is amenable to solution using block coordinate descent (see Section 2.3.1). In the following sections we assume that both $\beta$ and $X$ have been expanded using the method described here, but we omit the * superscript for clarity. For this reason, when referring to the group lasso with overlapping groups we generally reference the estimator (2.15), but make it clear that the necessary variable expansion has been performed.
In the case of non-overlapping groups, the consistency properties of the group lasso have been investigated for a range of different scenarios for \( N, L \) and \( P \) (Bach, 2008b; Wei and Huang, 2010; Huang and Zhang, 2010). As with the lasso, while the GL model is often able to accurately recover true, underlying group sparsity in the data, variations such as the adaptive group lasso (Wang and Leng, 2008; Wei and Huang, 2010), that allow \( \lambda \) to vary between groups have been advocated to improve model consistency. In the case of overlapping groups, Percival (2012) has recently investigated the theoretical properties of the GL using the variable expansion strategy described above. They found that the GL’s ability to recover group-wise sparse signals depends to some extent on the complexity of the overlap structure. For this reason, we consider it important to evaluate the performance of our method with real data, and we do this using simulations over a range of scenarios in Section 2.6.

### 2.3 Group lasso estimation algorithm

When the penalty in the optimisation (2.15) is block-separable, that is the groups indexed by 1, \ldots, \( L \) are disjoint, the resulting convex optimisation problem is amenable to solution using block coordinate descent (BCD) (Friedman et al., 2007). The BCD algorithm offers an efficient strategy for solving convex optimisation problems with grouped variables, that guarantees convergence provided that the block-separability criterion is met (Tseng and Yun, 2009). As with the lasso and coordinate descent (2.11), BCD works by breaking down the optimisation into a series of single variable problems, solving the optimisation for each variable (here SNP) in turn, while holding all the others fixed, until a suitable minimum based on some stopping criterion is reached. Where variables are grouped, as in GL, estimates are obtained for each pathway parameter vector, \( \beta_i \) in turn, while holding constant the current estimates for all other pathway parameter vectors, \( \hat{\beta}_m, (m \neq l) \), and then cycling through each pathway until convergence.

Yuan and Lin (2006) derive a method for solving GL under the assumption that the group design matrices, \( \mathbf{X}_{\bar{g}_i} \) are orthogonal, that is \( \mathbf{X}_{\bar{g}_i}' \mathbf{X}_{\bar{g}_i} = \mathbf{I} \). This assumption does not hold in our case, so in the next section we derive a solution for GL in the case of non-orthogonal groups. We additionally find that GL estimation using BCD can be slow,
2.3 Group lasso estimation algorithm

particularly for the large datasets common to PGAS, and so in the following sections propose a number of strategies for speeding up parameter estimation.

Apart from the BCD algorithm that we describe here, alternative estimation methods for GL include ‘global’ and ‘group-wise’ descent algorithms in which $\beta$ or $\beta_l$ is updated in a single step, respectively (Yuan and Lin, 2006; Meier et al., 2008; Kim and Kim, 2006; Roth and Fischer, 2008). Each method varies in terms of its underlying assumptions and computational burden, with the latter depending on the dimensions of the datasets being considered (Foygel and Drton, 2010).

2.3.1 Block coordinate descent for non-orthogonal groups

We assume that the penalty in (2.15) is block-separable, so that in our context, the SNP duplication strategy described in Section 2.2.1 has been applied. BCD then proceeds as follows. We begin by considering a single pathway $l$. We collect the observed, standardised minor allele counts for SNP $j$ in the column vector $x_j = (x_{1j}, x_{2j}, \ldots, x_{Nj})'$. From (2.15), an estimate $\hat{\beta}_l$, for a single block $l$ is then obtained by minimising

$$f(\beta_l) = \frac{1}{2}||\hat{r}_l - \sum_{j=l_1}^{l_{P_l}} x_j \beta_j||_2^2 + \lambda w_l ||\beta_l||_2$$

(2.16)

with respect to $\beta_l$, where $\hat{r}_l = y - \sum_{m\neq l} X_m \hat{\beta}_m$. The vector $\hat{r}_l$ is the ‘partial residual’ vector for pathway $l$, based on the current estimates, $\hat{\beta}_m, m \neq l$, of the other pathway parameter vectors.

Estimates for each $\beta_j$ are then obtained by taking partial derivatives with respect to $\beta_j$, that is by setting

$$\frac{\partial f(\beta_l)}{\partial \beta_j} = 0 \quad \text{for } j = l_1, \ldots, l_{P_l}. \quad (2.17)$$

Ignoring the penalty term, the partial derivative with respect to $\beta_j$ is

$$\frac{\partial}{\partial \beta_j} \frac{1}{2}||\hat{r}_l - \sum_j x_j \beta_j||_2^2 = -x_j'(\hat{r}_l - \sum_j x_j \beta_j).$$
We denote the partial derivative of the penalty term, by

\[ s_j = \frac{\partial}{\partial \beta_j} ||\beta||_2 \]

so that (2.17) can be written as

\[-x_j'(\hat{r}_l - \sum_j x_j \beta_j) + \lambda w_l s_j = 0 \quad j = l_1, \ldots, l_{P_l}. \tag{2.18}\]

We first consider the case where \( \beta_l = 0 \), that is \( \beta_j = 0 \), for \( j = l_1, \ldots, l_{P_l} \). In this case \( ||\beta||_2 \) is not differentiable. We instead form the \( P_l \) sub-differentials, \( s_j \in [-1, 1] \) (Bertsekas, 1999), so that

\[ \sum_j s_j^2 \leq 1. \tag{2.19}\]

The system of equations (2.18) can then be written

\[ s_j = \frac{1}{\lambda w_l} x_j' \hat{r}_l \quad j = l_1, \ldots, l_{P_l} \]

and using (2.19), we see that \( \beta_l = 0 \) when

\[ \sum_j s_j^2 = \frac{1}{\lambda^2 w_l^2} \sum_j (x_j' \hat{r}_l)^2 \leq 1. \tag{2.20}\]

Note that for all groups to have an equal chance of being selected by the model, that is for (2.20) to be unbiased with respect to group size, a weight, \( w_l = \sqrt{P_l} \), as proposed by Yuan and Lin (2006), can be applied.

Finally, since

\[ \sum_j (x_j' \hat{r}_l)^2 = ||X_l' \hat{r}_l||_2^2 \]

we can rewrite (2.20) as

\[ (\sum_j s_j^2)^{\frac{1}{2}} = \frac{1}{\lambda w_l} ||X_l' \hat{r}_l||_2 \leq 1, \]

so that if \( \beta_l = 0 \)
2.3 Group lasso estimation algorithm

\[ \|X_i \hat{r}_l\|_2 \leq \lambda w_l. \quad (2.21) \]

When \( \beta_l \neq 0 \), the objective function (2.16) can be minimised numerically using coordinate descent within group \( l \), as a series of one-dimensional estimations over \( \beta_j, j = l_1, \ldots, l_{P_l}, l_1, \ldots, l_{P_1}, \ldots \), until convergence of \( \beta_l \). At each step, the one-dimensional optimisation for \( \beta_j \) is given by

\[ \hat{\beta}_j = \arg \min_{\beta_j} \left\{ \frac{1}{2} \| \hat{r}_l - \sum_{k \in G_l, k \neq j} X_k \hat{\beta}_k - x_j \beta_j \|_2^2 + \lambda w_l \left( \sum_{k \in G_l, k \neq j} \hat{\beta}_k^2 + \beta_j^2 \right)^{\frac{1}{2}} \right\}. \quad (2.22) \]

Once group \( l \) has converged, the estimation proceeds for each group, \( l = 1, \ldots, L, 1, \ldots, L, \ldots \) until global convergence of \( \beta \). Friedman, Hastie, and Tibshirani (2010) suggest a golden section search over \( \beta_j \), combined with parabolic interpolation for the 1d optimisation (2.22). The number of such estimations depends on \( L \) and \( P^* \). This can make the GL optimisation prohibitively slow, particularly for the large \( P \) typically found in PGAS. For this reason, we next describe three strategies for speeding up the estimation.

2.3.2 Taylor approximation of penalty

One means of speeding up the estimation for \( \beta_j \) is to use a linear or quadratic approximation of the GL \( \ell_2 \) penalty (Zou and Li, 2008; Fan and Li, 2001), enabling the replacement of the multi-step numerical optimisation over \( \beta_j \) with a one-step calculation. Breheny and Huang (2009) propose the use of a Taylor approximation for a range of different estimation problems with grouped variables and we adopt this approach for our GL estimation problem.

For convenience, we rewrite the 1d optimisation (2.22) as

\[ f(\beta_l | \hat{\beta}_k, k \neq j) = \frac{1}{2} \| \hat{r} + x_j \hat{\beta}_j - x_j \beta_j \|_2^2 + \lambda w_l \Gamma(\beta_l | \hat{\beta}_k), \quad (2.23) \]

where \( \hat{r} = y - \sum_i X_i \hat{\beta}_i \) is the total residual, using the current estimates of all SNP coefficients; and \( \Gamma(\beta_l | \hat{\beta}_k) = (c + \beta_j^2)^{\frac{1}{2}} \), with \( c = \sum_{k \neq j} \hat{\beta}_k^2 \).

We now consider the first order Taylor expansion of \( \Gamma(\beta_l | \hat{\beta}_k) \) as a function of \( x = \beta_j^2 \),
about the point \( a = \hat{\beta}_j^2 \),

\[
\Gamma(x) \simeq \Gamma(a) + \Gamma'(a)(x - a).
\]

Now

\[
\Gamma(x) = (c + x)^{\frac{1}{2}}
\]

and \( \Gamma'(a) = \frac{1}{2(c + a)^{\frac{1}{2}}} \)

so that

\[
\Gamma(x) \simeq (c + a)^{\frac{1}{2}} + \frac{x - a}{2(c + a)^{\frac{1}{2}}}
\]

Substituting \( a = \hat{\beta}_j^2 \), and noting that \((c + a)^{\frac{1}{2}} = \|\hat{\beta}_l\|_2\), where \( \hat{\beta}_l \) denotes the current estimate of \( \beta_l \), this gives

\[
\Gamma(\beta_j^2) \simeq \hat{\beta}_l + \beta_j^2 - \hat{\beta}_j^2 + \frac{\lambda w_1}{\|\hat{\beta}_l\|_2}
\]

Substituting this expression in (2.23), we have

\[
f(\beta_l|\hat{\beta}_k, k \neq j) = \frac{1}{2}||\hat{x} + x_j\hat{\beta}_j - x_j\beta_j||_2^2 + \lambda w_1[\hat{\beta}_l + \frac{\beta_j^2 - \hat{\beta}_j^2}{2\|\hat{\beta}_l\|_2}]
\]

Differentiating with respect to \( \beta_j \) gives

\[
\frac{\partial f(\beta_l)}{\partial \beta_j}\bigg|_{\hat{\beta}_k,k\neq j} = -x_j'(\hat{\beta}_j + \hat{x}_j) + \lambda w_1 \frac{\hat{\beta}_j}{\|\hat{\beta}_l\|_2}
\]

since \( \sum_i x_{ij}^2 = x_j'x_j = 1 \). Rearranging terms and setting the partial derivative equal to zero, we see that the minimum is achieved when

\[
\beta_j = \frac{\hat{x}_j'\hat{x} + \hat{\beta}_j}{1 + \lambda'} \quad \text{where} \quad \lambda' = \frac{\lambda w_1}{\|\hat{\beta}_l\|_2}
\]

(2.24)

Where the current estimate \( \|\hat{\beta}_l\|_2 = 0 \), that is when group \( l \) first enters the estimation, we set \( \|\hat{\beta}_l\|_2 \) to be a small positive quantity, enabling \( \beta_j \) in (2.24) to be estimated.
BCD proceeds by obtaining estimates for each $\beta_j, j = l_1, \ldots, l_{P_l}, 1, \ldots, l_{P_l}, \ldots$ until convergence within the block, and for each pathway, $l = 1, \ldots, L, 1, \ldots, L, \ldots$ in turn, until a stopping criterion indicating a global minimum of (2.15) has been satisfied. The estimation process is summarised in Box 2.1.

**Box 2.1 GL estimation algorithm using BCD**

1. initialise $\hat{\beta} \leftarrow 0$

   block coordinate descent

2. repeat

3. for pathway $l = 1, 2, \ldots, L$

4. $\hat{r}_l \leftarrow y - \sum_{m \neq l} X_m \hat{\beta}_m$

5. if $||X_l \hat{r}_l||_2 \leq \lambda w_l$

6. $\hat{\beta}_l \leftarrow 0$

7. else

8. coordinate descent within block

9. repeat

10. for $j = l_1, \ldots, l_{P_l}$

11. estimate $\beta_j$ using (2.24)

12. until $\beta_l$ converges

13. $\hat{\beta}_l \leftarrow \beta_l$

14. until $\hat{\beta}$ converges

15. $\hat{\beta}_{GL} \leftarrow \hat{\beta}$

### 2.3.3 Use of pathway ‘active set’

Sparse model estimation with large datasets can be fast, because for example the lasso soft-thresholding operator, (2.10), or the GL group selection criterion (2.21), sets most coefficients to zero, so reducing the need for subsequent coefficient estimation through coordinate descent. However, for GL with multiple, large overlapping groups, the need for
the repeated calculation of (2.21) to establish whether or not a particular group can enter the estimation carries a substantial computational burden. This problem motivates another strategy providing substantial gains in computational efficiency for a range of sparse regression problems. This active set strategy relies on the pre-selection of a subset of ‘potentially active’ predictors, or groups of predictors that are likely to be selected by the model at a given $\lambda$ (Tibshirani et al., 2010; Roth and Fischer, 2008). The optimisation can then be run over this reduced set of variables, with a subsequent check to ensure that no other predictors should have been included in the first place. The active set procedure offers potentially dramatic speed up in execution times, particularly for very large datasets such as those found in PGAS, due to the reduced number of computations that need to be performed. In addition there are substantial savings in the amount of working memory required to store data during processing, which can also lead to big reductions in computation times with large datasets where memory is constrained.

For the GL, we begin by considering the inequality (2.21). For groups to enter the model we require

$$||X_l^\prime \hat{r}_l||_2 > \lambda w_l$$

so that, at the first iteration, with $\beta$ initialised to zero, a group $G_l$ enters the model if

$$||X_l^\prime y||_2 > \lambda w_l$$

We define the ‘active set’ $A$ of potentially active groups that satisfy (2.26) as

$$A = \{m \in G : ||X_m^\prime y||_2 > \lambda w_m\}$$

and additionally define

$$\lambda_{max} = \min_\lambda : ||X_l^\prime y||_2 \leq \lambda w_l$$

as the smallest $\lambda$ value for which the active set is empty. Note that provided $\lambda$ is close to $\lambda_{max}$, then $|A| \ll L$.

Once one or more groups enter the model, not all $\beta_l$ will be zero, so that the inequality
2.3 Group lasso estimation algorithm

(2.25) then determines which groups may enter or leave the model during block coordinate descent. The active set procedure rests on the observation that in practice, the final set of groups selected by the model rarely includes any groups not in \( \mathcal{A} \) (Tibshirani et al., 2010). We can therefore perform the full estimation on \( \mathcal{A} \), followed by a check of the inequality (2.25), to see if any additional groups not in \( \mathcal{A} \) can enter the model. If there are no additional groups, then we have the full solution. If not, then we run the full estimation again, with the additional groups satisfying (2.25) added to \( \mathcal{A} \). A summary of the active set algorithm is given in Box 2.2.

Box 2.2 Active set algorithm for a single \( \lambda \) value
1. Form the active set, \( \mathcal{A} = \{ m \in \mathcal{G} : ||X_m'y||_2 > \lambda w_m \} \)
2. Initialise \( \beta \leftarrow 0 \)
3. Solve the GL estimation at \( \lambda \), using only the groups in \( \mathcal{A} \):
   \[
   \hat{\beta} = \min_{\beta} \frac{1}{2} ||y - \sum_{m \in \mathcal{A}} X_m \beta_m ||^2_2 + \lambda \sum_{m \in \mathcal{A}} w_m ||\beta_m||_2
   \]
4. Compute the revised active set on the full dataset:
   \[
   \mathcal{A}^+ = \{ z \in \mathcal{G} : ||X_z'\hat{r}_z||_2 > \lambda w_z \}
   \]
5. if \( \mathcal{A}^+ / \mathcal{A} = \emptyset \)
6. \( \hat{\beta}^{GL} \leftarrow \hat{\beta} \)
7. STOP
8. else
9. \( \mathcal{A} \leftarrow \mathcal{A}^+ \)
10. goto 2.

2.3.4 Efficient computation of block residuals

A further, large computational burden results from the repeated calculation of the residuals \( r_l \) and \( r \) in (2.21) and (2.24). The computational overhead for these calculations is substantial, both because of the size of the expanded design matrix \( (N = 743 \text{ and } P^* = 66,085 \text{ in the simulation study described in section 2.6, substantially larger for genome wide data),} \)
and because of the iterative nature of the BCD algorithm, meaning that a very large number of calculations are performed. We therefore achieve one further substantial gain in computational efficiency by noting that since the blocks are separable, during BCD only the single block residual, \( h_l = y - X_l \beta_l \), changes between iterations \( j = 1, \ldots, l_1, \ldots, l_P, \ldots \) within block \( l \), and between iterations \( l = 1, \ldots, L, 1, \ldots, L, \ldots \) across blocks. We therefore only need update \( h_l \) at each iteration, with \( r \) and \( r_l \) updated using computationally inexpensive matrix subtractions and additions.

### 2.4 Selection bias and pathway weighting

PGAS methods derived from univariate SNP statistics are subject to various biasing factors that can influence pathway ranking under the null, where no SNPs influence the phenotypic trait, \( y \). These factors vary from method to method, but may include the number and size of genes in a pathway, as well as LD between SNPs and genes. Such biasing factors are generally corrected through the use of permutation procedures. For example, the ‘GenGen’ method (Wang et al., 2009b), measures the degree to which pathways are enriched with high ranking genes, and is subject to potential bias due to variation in the number of SNPs mapped to a gene, and to differences in LD between SNPs mapped to different genes. The bias correction procedure begins by forming multiple datasets through permutation of phenotype labels. For each permuted dataset, gene scores are generated from univariate SNP statistics, and a pathway enrichment score is calculated. A normalised (bias-corrected) pathway enrichment score is then derived by comparing the distribution of pathway scores under the null with the score obtained from the unpermuted data.

Regression-based methods are similarly prone to bias, and once again the use of permutation has been proposed to correct for this, along with dimensionality reduction to extract non-redundant information. For example, with the GRASS method (Chen et al., 2010), genetic information within each gene is first summarised as ‘eigenSNPs’, obtained through PCA. The biasing effect of gene size is once again accounted for through the generation of a null distribution, formed by permuting phenotype labels.

With the GL under the null, pathway selection will be influenced by pathway size (i.e. the number of SNPs within a pathway), since the accumulation of spurious associ-
2.4 Selection bias and pathway weighting

ations in larger pathways will give rise to larger $||\beta||_2$ in (2.15). In addition, variation in
dependencies between SNPs within pathways, and to a lesser extent between pathways will
give rise to corresponding variations in $||\beta||_2$ where spurious associations arise in regions
of high LD.

One way to correct for biases arising from variations in the statistical properties of dif-
ferent pathways or groups is through the use of appropriate group weights $w = (w_1, \ldots, w_L)$,
ensuring the group penalty (2.14) to vary between groups. For example, as noted by Yuan
and Lin (2006), one possible choice for the pathway weighting would be

$$w_l = \sqrt{P_l}$$

(2.28)

which ensures that groups of different size are penalised equally, and so have an equal
chance of being selected by the model, other things being equal (see (2.20)). In principle,
we could follow this strategy and attempt to account for other, additional factors that may
also bias pathway selection. However, there are a number of problems with this approach.
Consider for example the biasing effect of dependencies between SNPs within a pathway.
Where causal SNPs tag, or reside within large blocks with strong LD, the pathway signal
will be amplified, increasing the chance that such pathways will be selected by the model,
compared with other pathways where LD is low. This biasing effect will further depend on
the distribution of LD within the pathway, which will in turn depend on other factors such
as the number and size of pathway genes. The precise form of any additional term(s) that
should be added to (2.28) to account for this bias is thus unclear. Even if we were able to
identify a list of potential biasing factors, and formulate bias-correcting weight adjustments
for each, we are still faced with the problem that there may be other, unknown factors
that contribute to the bias. We therefore adopt a ‘hypothesis-free’ approach to adjusting
pathway weights, which makes no assumptions about those factors which might influence
pathway selection.

Consider pathway selection under the GL model (2.15), with $\lambda$ tuned to select $M$ path-
ways. We begin with the case $M = 1$. When there is no selection bias, and assuming no
genetic association, a pathway $\mathcal{G}_l$ should be randomly selected by the model according to a
uniform distribution, namely with probability $\Pi_l = 1/L$, for $l = 1, \ldots, L$. However, when
biasing factors are present this is generally not the case, and the empirical probability distribution describing pathway selection, $\Pi^*(w)$ will not be uniform. Here the dependence on the weight vector $w$ has been made explicit, since with $\lambda$ tuned to select a single pathway, and for given $y$ and $X$, $w$ alone determines the frequency distribution. A measure of distance between these two distributions can be obtained by computing their Kullback-Leibler (KL) divergence

$$D = \sum_l \Pi_l^*(w) \log \frac{\Pi_l^*(w)}{\Pi_l}$$

(2.29)

where $\Pi_l^*(w)$ is the empirical probability for the selection of pathway $G_l$ under the assumption of no genetic associations. When GL pathway selection is unbiased, we expect this distance to be approximately zero. Our strategy consists in adaptively adjusting $w$ in order to minimise $D$.

Our adaptive weighting procedure is an iterative one, whereby at each iteration $\tau$ we first update the previous weight vector $w^{(\tau-1)}$, and then re-estimate $\Pi^*(w^{(\tau)})$ by fitting the GL model $Z$ times, each with a random permutation of the response in order to create $Z$ null data sets. $\Pi_l^*(w^{(\tau)})$ is then the frequency at which pathway $G_l$ is selected across the $Z$ null data sets at iteration $\tau$. The algorithm is initialised at iteration $\tau = 0$ by using an initial weight vector $w^{(0)}$, for instance the standard size weighting (2.28). This procedure is then repeated until $D$ reaches some suitably small value.

From (2.29), a reduction in $D$ can be obtained by reducing the magnitude of the difference, $d_l = \Pi_l^*(w) - \Pi_l$, for all $l$. As each $|d_l|$ approaches zero, the ratio, $\Pi_l^*(w)/\Pi_l$, approaches one, so that the contribution of pathway $G_l$ to $D$ is decreased. With this in mind, at each iteration, we adjust pathway weights according to the following formula,

$$w_l^{(\tau)} = w_l^{(\tau-1)} \left[ 1 - \text{sign}(d_l)(\eta - 1)L^2d_l^2 \right] \quad 0 < \eta < 1$$

(2.30)

where the parameter $\eta$ controls the maximum amount by which each $w_l$ can be reduced in a single iteration, in the case that pathway $G_l$ is selected with zero frequency. The weighting update equation has the following desirable properties. When $0 \leq \Pi_l^* < \Pi_l$,

---

2Alternatively, in a simulation study where the null distribution of the response is known (as in section 2.6), the $Z$ models can be fitted after sampling a response from that null distribution.
2.5 Pathway ranking

i.e. \(-\frac{1}{L} \leq d_l < 0\), \(w_l\) is decreased, up to a maximum factor \(\eta\) when \(\Pi_l^* = 0\), increasing the chance that group \(l\) is selected. When \(\Pi_l^* > \Pi_l\), i.e. \(d_l > 0\), \(w_l\) is increased, decreasing the chance that group \(l\) is selected. Finally, when \(\Pi_l^* = \Pi_l\), i.e. \(d_l = 0\), \(w_l\) is unchanged. The square in the weight adjustment factor ensures that large values of \(|d_l|\) result in relatively large adjustments to \(w_l\).

The estimation of \(\Pi^*\) when \(M > 1\), that is where more than one pathway is selected by the model, is computationally infeasible even for a small value of \(M\), since we would need to estimate the empirical joint probability distribution that \(M\) pathways are selected. However, we expect that many of the factors biasing pathway selection when \(M = 1\) will similarly affect this joint probability distribution. Under this assumption, we estimate the optimal weight vector \(\mathbf{w}\) only in the \(M = 1\) case. Simulation studies (see section 2.6.3) indicate that this data-driven adaptive waiting scheme is able to substantially increase power and specificity compared with the standard weighting (2.28), even when \(M > 1\), indicating that this assumption holds in practice.

Note that under the null, where more than one pathway is selected by the model, we don’t expect pathway selection probabilities to be uniform, since the presence of overlapping SNPs means that pathways are not independent.

Finally, we note that despite the need for multiple MC simulations over multiple iterations, our proposed bias-adjusted weighting strategy is fast, since it relies on fitting the GL model with \(\lambda\) tuned to select a single pathway only, ensuring that the active set (Section 2.3.3) is very small, and model estimation time for each of the \(Z\) model fits is minimal.

2.5 Pathway ranking

With most variable selection methods, a choice for the regularisation parameter, \(\lambda\), must be made, since this determines the number of variables selected by the model. Common strategies include the use of cross validation to choose a \(\lambda\) value that minimises the prediction error between training and test datasets (Hastie, Tibshirani, and Friedman, 2008). One drawback of this approach is that it focuses on optimising the size of the set, \(\hat{C}\), of selected pathways (more generally, selected variables) that minimises the cross validated prediction error. Since the variables in \(\hat{C}\) will vary across each fold of the cross validation,
this procedure is not in general a good means of establishing the importance of a unique set of variables, and can give rise to the selection of too many variables (Vounou et al., 2011; Meinshausen and Bühlmann, 2010). For the lasso, alternative approaches, based on data subsampling or bootstrapping have been shown to improve model consistency, in the sense that the correct model is selected with a high probability (Bach, 2008a; Meinshausen and Bühlmann, 2010; Chatterjee and Lahiri, 2011). These methods work by recording selected variables across multiple subsamples of the data, and forming the final set of selected variables either as the intersection of variables selected at each model fit, or by assessing variable selection frequencies. Examples of the use of such approaches can be found in a number of recent gene mapping studies involving model selection using either the lasso or elastic net (Cho et al., 2010; Eleftherohorinou et al., 2011; Motyer et al., 2011; Vounou et al., 2011). Motivated by these ideas, we adopt a resampling strategy in which we calculate pathway selection frequencies by repeatedly fitting the model over $B$ subsamples of the data, at a fixed value for $\lambda$. Each random subsample of size $N/2$ is drawn with replacement. Our motivation here is to exploit knowledge of finite sample variability obtained by subsampling, to achieve better estimates of pathway importance. With this approach, which in some respects resembles the ‘pointwise stability selection’ strategy of Meinshausen and Bühlmann (2010), selection frequencies provide a direct measure of confidence in the selected pathways in a finite sample.

This resampling strategy also allows us to rank pathways in order of their strength of association with the phenotype, so that we expect pathways in $C$ will achieve a high ranking, whereas those in $C'$ will be ranked low. The presentation of results as ranked lists of pathways or SNPs is common in genetic association mapping, and this approach has the added advantage of allowing us to make direct comparisons with alternative pathway methods that use p-values as a ranking criterion (see Section 2.6.3).

We denote the set of selected pathways at subsample $b$ by

\[
\hat{C}^{(b)} = \{ l : \hat{\beta}_l^{(b)} \neq 0 \} \quad b = 1, \ldots, B
\]

where $\hat{\beta}_l^{(b)}$ is the estimated SNP coefficient vector for pathway $l$ at subsample $b$. The
2.5 Pathway ranking

The selection probability for pathway \( l \) measured across all \( B \) subsamples is then

\[
\pi_{l_{\text{path}}} = \frac{1}{B} \sum_{b=1}^{B} I_{l(b)} \quad l = 1, \ldots, L
\]

where the indicator function, \( I_{l(b)} = 1 \) if \( l \in \hat{C}(b) \), and 0 otherwise. Pathways are ranked in order of their selection probabilities, \( \pi_{l_{\text{path}}} \geq \ldots \geq \pi_{L_{\text{path}}} \).

Computation time increases rapidly with \( M \), the number of selected pathways, so that with the number, \( |C| \), of causal pathways unknown, the choice of \( M \) is driven by the number of causal pathways we expect to identify within computational constraints. For the simulation study described in Section 2.6, we use \( B = 100 \) subsamples, and at each subsample we perform a line search over \( \lambda \), to ensure that \( M \geq M_{\text{min}} \) pathways are selected. This procedure is described in Box 2.3. We note that since typically \( M \ll L \), some \( \pi_{l_{\text{path}}} \) may be zero. Such pathways are classified as unranked.

**Box 2.3** Line search procedure for tuning \( \lambda \) to select \( M \geq M_{\text{min}} \) pathways

1. \( \lambda_{\text{max}} \leftarrow \min_{\lambda} : ||X_l^T y||_2 \leq \lambda w_l, l = 1, \ldots, L \) (2.27)

2. \( M \leftarrow 0; c \leftarrow 0.8^\dagger \)

3. while \( M < M_{\text{min}} \)

4. \( \lambda \leftarrow c\lambda_{\text{max}} \)

5. \( A \leftarrow \{ m \in G : ||X_m^T y||_2 > \lambda w_m \} \)

6. if \( |A| \geq M_{\text{min}}^\dagger \)

7. obtain \( \hat{\beta}^{GL}(\lambda, A) \), as described in box 2.2 (starting at step 2.)

8. \( M \leftarrow |\{ l \in G : ||\hat{\beta}_l|| > 0 \}| \)

\[^\dagger\] The value of \( c \) is chosen for computational convenience. A value close to 1 ensures as few pathways are selected by the model as possible, thus speeding up the estimation. However, a value too close to 1 means that the decrease in \( \lambda \) at each iteration is small, meaning that many iterations may have to be performed before \( M \) reaches the desired range.

\[^\dagger\dagger\] This step is introduced for computational efficiency, since during BCD, the number of selected groups rarely decreases.
2.5.1 Ranking performance measures

In this section we describe 3 separate ranking performance measures that we use to evaluate the performance of our method in a simulation study described in Section 2.6. One complicating factor is the issue of overlapping pathways, making the effective number of causal pathways, $|C|$, dependent on the degree to which SNPs in $S$ overlap multiple pathways. In addition, with any method based on variable selection, the possibility that causal pathways are unranked, i.e. they are not selected by the model, must be taken into account.

Consider the situation where the set $S$ of causal SNPs, with cardinality $|S| > 1$, is known. We may choose to define $C$ in its most restricted sense as the set of pathways that contain all members of $S$, or alternatively $C$ might include all pathways containing one or more SNPs belonging to $S$. In either case $|C|$ will depend on the degree to which SNPs in $S$ overlap multiple pathways. This in turn depends on the particular distribution of causal SNPs with respect to overlapping genes. The need to accommodate this variability in $|C|$ in part motivates our formulation of the ranking measures described below.

We propose three separate ranking measures that capture different aspects of ranking performance, and focus on the top 100 ranked pathways only. We do this firstly because in any method attention is inevitably focused on the highest ranking pathways (or alternatively those with the highest statistical significance in a hypothesis testing framework). Also, since in a simulation study we compare the performance of our variable selection method which identifies a limited number of pathways against an alternative method that scores all pathways, some suitable cutoff in rank must be chosen.

We denote the set of ranked causal pathways by $\hat{C}^* = \{k \in C : \pi_{path}^k > 0\}$, and their rankings by $r_{k_1}, r_{k_2}, \ldots$, ranked in order of their respective selection frequencies, $\pi_{k_1}^{path} > \pi_{k_2}^{path} > \ldots$. We further denote by $\hat{C}^*_{100} = \{k \in \hat{C}^* : r_k \leq 100\}$, cardinality $|\hat{C}^*_{100}|$, the set of causal pathways falling in the top 100 ranks. Our three proposed ranking measures are as follows:

1. **Highest causal pathway rank**, $r_{k_1}$, that is the single highest rank achieved by any pathway in $\hat{C}^*_{100}$. This lies in the range $1 \leq r_{k_1} \leq 100$, and is only defined for

---

The * superscript is used to distinguish $\hat{C}^*$, from $\hat{C}$, the set of all ranked pathways.
2.5 Pathway ranking

\[ |\hat{C}^*_100| \geq 1. \]

2. **Ranking power**, \( p_{100} \), defined as

\[
p_{100} = \frac{|\hat{C}^*_100|}{|C|}
\]  

(2.31)

with \( 0 \leq p_{100} \leq 1 \). \( p_{100} = 0 \) when no causal pathways are ranked in the top 100 (\( \hat{C}^*_100 = \emptyset \)), and \( p_{100} = 1 \) when all causal pathways are ranked in the top 100 (\( \hat{C}^*_100 = C \)).

3. **Power-adjusted, normalised, weighted ranking score**, \( R \). This takes account of the actual rankings, \( r_{k1}, \ldots \) as well as the ranking power, \( p_{100} \). We begin by defining a normalised, weighted ranking score,

\[
R^* = \frac{\sum_{k \in \hat{C}^*_100} r_k^{\frac{1}{2}}}{\sum_{k=1}^{\left|\hat{C}^*_100\right|} k^{\frac{1}{2}}}
\]

Here the square root places greater emphasis on highly-ranked causal pathways. The denominator is a normalising factor which represents the minimum possible weighted ranking score, with \( r_{k1} = 1, r_{k2} = 2, \ldots \), ensuring that \( R^* \) attains its minimum value of 1 when the pathways in \( \hat{C}^*_100 \) are optimally ranked. Higher values of \( R^* \) indicate suboptimal ranking. \( R^* \) takes no account of the possibility that not all causal pathways are ranked, i.e. \( \hat{C}^*_100 \neq C \). To do this we form the adjusted measure

\[
R = \begin{cases} 
R^*/p_{100} & \text{if } p_{100} > 0 \\
R_0 & \text{if } p_{100} = 0
\end{cases}
\]

(2.32)

\( R \) thus attains a minimum value of 1 when all causal pathways are optimally ranked, and the value \( R_0 \) when no causal pathways are ranked.

\textsuperscript{4}We assume that \( |C| \leq 100 \), which is always true in the simulation study described in Section 2.6
2.6 P-GLAW Simulation Study

We assess the power of our proposed P-GLAW method in a simulation study using real genotype and pathway data, with simulated, quantitative phenotypes generated under an additive genetic model from SNPs within a single, randomly selected causal pathway. The presence of overlapping SNPs means that the actual number of causal pathways is then typically greater than one. We additionally compare our method’s performance with an alternative, univariate-based method commonly used in gene set analysis. Computation times for both methods increase with \( P \), and because of this, and the large number of scenarios and simulations tested, we restrict this analysis to SNPs on a single chromosome.

2.6.1 Genotype and pathways data

We use genotypes obtained from the Alzheimer’s Disease Neuroimaging Initiative, ADNI (www.loni.ucla.edu/ADNI), derived from the Illumina Human 610-Quad BeadChip. Subjects comprise a mix of healthy controls, those diagnosed as having mild cognitive impairment, and those with AD. After removing variants with a call rate < 95\%, minor allele frequency (MAF) < 0.1, and significant deviation from Hardy-Weinberg equilibrium (\( p < 5.7 \times 10^{-7} \)), 448,294 SNPs remain. In this study we use genotype data from \( N = 743 \) subjects, and consider only SNPs from chromosome 1 (33,850 SNPs).

Popular databases used for the mapping of genes to biological pathways include the Kyoto Encyclopedia of Genes and Genomes (KEGG, www.genome.jp/kegg/pathway.html) and BioCarta (www.biocarta.com/genes/index.asp). For this study we use data on ‘canonical pathways’ from the Molecular Signals Database (MSigDB, www.broadinstitute.org/gsea/msigdb/index.jsp), which is a commonly-used, curated collection of pathways obtained from multiple sources. At the time of writing this comprised 880 pathways mapped to 6,804 genes. 2,382 human gene locations on chromosome 1, corresponding to assembly GRCh36 are obtained using Ensembl’s biomart API (www.biomart.org). ADNI-genotyped SNPs on chromosome 1 are then mapped to annotated genes within 10kb (20,399 SNPs mapped to 2,096 genes), and these remaining genes and SNPs are then mapped to pathways using MSigDB (8,102 SNPs mapped to 778 pathways). Thus we see that the majority of chromosome 1 SNPs fail to map to any pathway, but that the majority of annotated path-
ways map to at least 1 SNP on this chromosome. Finally, small (< 10 SNPs) and identical pathways are removed. After all pre-processing we are left with a total of \( P = 8,078 \) SNPs mapped to 551 pathways (max: 1,059; min: 10; mean: 120 ± 142 SNPs per pathway). All SNP to pathway mapping and filtering was performed using bespoke code written in Python. The mapping and filtering process is illustrated in Fig. 2.4.

More than 80\% of SNPs are observed to overlap more than 1 pathway, with around 20\% overlapping 10 or more pathways and 2\% overlapping 60 or more (see Fig. 2.5). After variable expansion to account for overlapping pathways (see Section 2.2.1), we have \( P^* = 66,085 \) SNPs.

![Figure 2.4: SNP to pathway mapping.](image-url)
2.6.2 Simulation framework

We begin by adjusting the pathway weight vector, $\mathbf{w}$, using the bias-adjusted adaptive weighting procedure described in Section 2.4. We use 10 iterations (i.e. $\tau = 1, \ldots, 10$), and calculate the empirical, null pathway selection distribution $\Pi^*$, using $Z = 40,000$ simulations, each with response vector $\mathbf{y}$ sampled from a standard normal distribution, $\mathcal{N}(0, 1)$, since many quantitative traits are expected to be normally distributed.

For the simulation of a SNP-dependent response, we begin by drawing $|S|$ SNPs from a single, randomly selected causal pathway, $G_\phi$, according to some specified distribution (see below), and then form the set $C$, of causal pathways that contain all the members of $S$. We thus chose to define $C$ in its most restricted sense, rather than for example including pathways that contain one or more SNPs in $S^5$. Note that the number, $|C|$ of causal pathways will vary according to the particular pattern of overlaps corresponding to genes mapping to SNPs in $S$.

For each MC simulation we generate a univariate quantitative phenotype,

$$y_i = \sum_{k \in S} \zeta_k x_{ik} + \epsilon \quad i = 1, \ldots, N$$

Figure 2.5: Frequency distribution of ADNI SNPs by number of pathways they map to. SNPs are mapped to genes within 10kbp. The data set consists of 8,078 SNPs and 551 pathways.

\footnote{It would also be interesting to consider the more general case. However, here we have a particular interest in identifying small, distributed signals, that are typically missed in SNP-wise GWAS.}
where $\zeta_k$ is the allelic effect per minor allele due to causal SNP $k$. Setting $w_k = \zeta_k x_k$, we define the effect size of SNP $k$ as $\delta_k = \frac{E(w_k)}{E(y)}$ for $k \in \mathcal{S}$, and set $\epsilon \sim \mathcal{N}(1, \sigma^2)$ so that $\delta_k = 0$ when $\zeta = 0$. We also record the average SNP effect size as a proportion of total phenotypic variance, $ES_k = \text{Var}(w_k)/\text{Var}(y)$, and the mean proportionate change in response per minor allele, $E(\zeta_k)$. For our simulations we control $\delta_k$, and set $\zeta_k$ accordingly, so that effect size is independent of SNP MAF, whereas $\zeta_k$ and $ES_k$ are MAF-dependent.

The power and specificity of any PGAS method is likely to depend on a range of factors including the number and distribution of causal SNPs, and the size of their phenotypic effect (Wang, Li, and Hakonarson, 2010; Fridley and Biernacka, 2011). We therefore assess the performance of our method across 6 different scenarios in which we vary each of these factors. Furthermore, we test each scenario over 500 MC simulations to account for variation in causal SNP MAFs, gene size and number within pathways, and LD patterns within and between causal pathways. The list of scenarios tested is presented in Table 2.1.

| scenario | $|\mathcal{S}|$ | $\delta_k$ | distribution | description |
|----------|----------------|------------|--------------|-------------|
| (a)      | 10             | 0.005      | random from $G_\phi$ | $|\mathcal{S}|$ large; $\delta_k$ large; random distribn |
| (b)      | 3              | 0.005      | random from $G_\phi$ | $|\mathcal{S}|$ small; $\delta_k$ large; random distribn |
| (c)      | 3              | 0.005      | random from single gene in $G_\phi$ | $|\mathcal{S}|$ small; $\delta_k$ large; single gene |
| (d)      | 10             | 0.001      | random from $G_\phi$ | $|\mathcal{S}|$ large; $\delta_k$ small; random distribn |
| (e)      | 3              | 0.001      | random from $G_\phi$ | $|\mathcal{S}|$ small; $\delta_k$ small; random distribn |
| (f)      | 3              | 0.001      | random from single gene in $G_\phi$ | $|\mathcal{S}|$ small; $\delta_k$ small; single gene |

Table 2.1: Scenarios tested in simulation study. For scenarios (c) and (f), in the rare event that a gene has less than 3 SNPs, all SNPs within the gene are selected.

First, we consider scenarios where the number of causal SNPs is small ($|\mathcal{S}| = 3$) or large ($|\mathcal{S}| = 10$).

Secondly, we consider two different SNP effect sizes. We choose values for $\sigma^2_\epsilon$ and $\delta_k$ to mimic effect sizes obtained in recent association studies, focusing particularly on the smallest reported effect sizes. Park et al. (2010) review GWAS for a number of quantitative traits (height, Crohn’s disease and breast, prostate and colorectal cancers) and report values for $ES_k$ ranging from 0.02 to 0.0004. Cho et al. (2009) report values for $\zeta_k$ for 8 quantitative traits in a large GWAS ranging from 1.6 to 0.006. A recent neuroimaging genetic study measuring genetic effects on a variety of traits related to brain structure reports significant
values for $\zeta_k$ of around 0.07 (Joyner et al., 2009). We set $\sigma = 0.2$, and test $\delta_k = 0.005$ and 0.001, which gives values for $ES_k = 0.001$ and 0.00004 and $E(\zeta_k) = 0.01$ and 0.002 respectively.

Finally, we also vary the distribution of causal SNPs with respect to genes and associated LD blocks, since we expect this to affect performance in our regression model (see Section 2.4). The particular distribution of causal SNPs is also expected to the affect the sensitivity of other pathways methods, for example where pathway scores are derived in a two-step process that begins with the calculation of gene association scores (Wang, Li, and Bucan, 2007). The distributions of $|C|$, the number of causal pathways for each scenario described in Table 2.1, are shown in Fig. 2.6.

Figure 2.6: Distributions of $|C|$ across 500 MC simulations for each of the 6 scenarios described in Table 2.1. Causal pathways are defined as those containing all SNPs in $S$. Where SNPs are distributed within a single gene (scenarios (c) and (f)), the number of causal pathways tends to be larger, since a single gene can map to multiple pathways. Where causal SNPs are distributed randomly across $G_{\phi}$ (scenarios (a), (b), (d), and (e)), this number tends to be smaller, particularly where the number of causal SNPs is large (scenarios (a) and (d)).
2.6 P-GLAW Simulation Study

2.6.3 Results

We begin with an investigation of the effect of our proposed methods for speeding up the GL estimation algorithm (see Sections 2.3.2, 2.3.3 and 2.3.4). We first note that GL estimation times will depend on the sample size \(N\) and the number of SNPs \(P\), which will in turn affect the number of mapped pathways \(L\) and the number of SNPs in the regression model after expansion to account for overlaps \(P^*\). Estimation times will further depend on the number of groups selected \(M\), and the amount of signal present, since these affect convergence times. For illustrative purposes, in Table 2.2 we compare execution times using our accelerated methods, with ‘standard’ block coordinate descent (described in Section 2.3.1), for a single model fit with a null response, and for \(M = 10\). Estimation times are seen to be substantially reduced across a range of values for \(N\) and \(P\), dramatically so for larger datasets.

We next illustrate the application of the bias-correcting strategy described in Section 2.4. The pathway weight vector, \(w\), is adjusted over 10 iterations, each with \(Z = 40,000\) MC simulations with the response vector \(y\) sampled from \(\mathcal{N}(0, 1)\). Fig. 2.7 (c) shows how the weight adjustment factor \(w^{(r)}/w^{(r-1)}\), (see (2.30)), varies with \(d_l\) across all pathways at a single iteration. Fig. 2.7 (a) and (b) show the observed, empirical distribution, \(\Pi^*\), using the standard size weighting (2.28), and the adapted weights (2.30) after 10 iterations, respectively. The corresponding KL divergence measure, \(D\) (see (2.29)), is observed to
Figure 2.7: Application of bias-adjusted weighting procedure to the data used in the simulation study. $Z = 40,000$, with a different null response, $y \sim N(0, 1)$, at each MC simulation. $\eta = 0.98$. (a) Empirical pathway selection frequency distribution, $\Pi^*$, with standard, pathway size weighting, $w_l = \sqrt{P_l}$. $D = 2.24$. Dotted horizontal line shows the expected distribution, $\Pi_l = 1/L \approx 0.002$. (b) $\Pi^*$ with bias-adjusted weights after 10 iterations. $D = 0.12$. (c) Variation of weighting adjustment factor $w^{(\tau)}/w^{(\tau-1)}$ with $d_l$ at a single iteration, with $\eta = 0.98$. Each point represents the adjustment to a single $w_l$, $l = 1, \ldots, L$. (d) Decrease in K-L divergence, $D$, over 10 iterations.
reduce steadily over the 10 iterations (Fig. 2.7 (d)), illustrating how the proposed weight adjustment procedure reduces pathway selection bias.

Next, we compare performance using the bias-adjusted weights with that obtained using the standard size weighting $w_l = \sqrt{P_l}$. We find the adjusted weighting scheme offers a considerable improvement in ranking performance for all ranking measures, and illustrate this in Fig. 2.8 for a single scenario (scenario (a)) using the ranking performance measures described in section 2.5.1. Fig. 2.8 (a) shows the first ranking measure ($r_{k_1}$) as a ROC curve, in which we show the proportion of simulations with $r_{k_1} \leq z$, for ranks $z = 1, 2, \ldots, 100$. We plot $z$ on the horizontal axis as a false positive rate (FPR), so that FPR = $(z - 1)/L$.

At a FPR of 0.05, we see that the adapted weighting scheme shows a more than 2 fold increase in power (from 0.29 to 0.62) over the standard pathway size weighting (2.28), indicating 62% of MC simulations have $r_{k_1} \leq 28$, compared with 29% for the standard size weighting. The distribution of $p_{100}$ across 500 MC simulations is illustrated as a box-plot in Fig. 2.8 (b). Here we see that the adapted weighting scheme offers a clear and substantial improvement in GL’s capacity to rank a high proportion of causal pathways in the top 100 ($p = 2.03 \times 10^{-50}$ that the two population $p_{100}$ CDFs are equal using a two-sample Kolmogorov-Smirnov (KS) test). GL with the standard weighting scheme performs particularly poorly with 55% of simulations failing to rank any causal pathway in any simulation, compared with 18% for the adapted weighting scheme. Finally, Fig. 2.8 (c) shows the distribution of the $R$ ranking measure across 500 simulations under the two weighting schemes. Once again we see that the adaptive weighting scheme demonstrates improved ranking performance over the standard size weighting scheme, with the distribution of $R$ scores skewed towards lower values for the former, indicating that causal pathways tend to be ranked higher.

For the remainder of this section, we describe results obtained using the bias-adjusted pathway weighting scheme described above. We assess P-GLAW ranking performance across the full range of scenarios described in Table 2.1, and compare these with pathway rankings obtained using the ‘GenGen’ (GG) method proposed by Wang, Li, and Bucan (2007). GG is a widely-used, GSEA-type PGAS method that measures pathway enrichment using genes scores derived from univariate SNP statistics. Studies using GG include searches for implicated pathways in Crohn’s disease (Wang et al., 2009b), autism spectrum
Figure 2.8: Ranking performance using adapted weights, compared with standard pathway size weighting, \( w_l = \sqrt{P_l} \). Ranking measures are described in Section 2.5.1. Results for a single scenario (scenario (a)) are illustrated, corresponding to \(|S| = 10 \) and \( \delta_k = 0.005 \), with causal SNPs selected at random from a single randomly selected pathway. All measures over 500 MC simulations. (a) ROC curves illustrating power to identify at least one causal pathway in the top 100. (b) Distribution of ranking power, \( p_{100} \), across 500 simulations. This is the proportion \( |\hat{C}_{100}|/|C| \) of causal pathways that are ranked in the top 100. Notches indicate 95% confidence intervals for the true median. (c) Distribution of the power-adjusted, normalised, weighted ranking score, \( R \), across 500 simulations, with \( R_0 = 100 \). The final ‘50+’ column includes simulations for which no causal pathway was ranked in the top 100.
disorders (Wang et al., 2009a), breast cancer (Menashe et al., 2010) and Alzheimer’s disease (Lambert et al., 2010). GG begins by scoring each SNP according to its association with the phenotype. SNPs are then mapped to genes within a specified distance, and each gene is scored according to its most significant mapped SNP. The enrichment of highly-ranked genes in a given pathway is then compared with those across all pathways, to obtain a pathway enrichment score. For GenGen we use identical source data (genotypes, phenotypes, SNP to gene, and gene to pathway mappings), and rank pathways by their normalised enrichment score, NES(\(l\)) (see A.1), determined from 1,000 permutations (the GG default settings). MC simulations for P-GLAW and GG are performed in parallel across 50 (P-GLAW) and 500 (GG) processors respectively, on a high-performance computing cluster. As described above for the comparison of alternative weighting schemes, results for the comparison study are presented in the form of \( r_{k_1} \) ROC curves (Fig. 2.9), \( p_{100} \) boxplots (Fig. 2.10) and \( R \) bar graphs (Fig. 2.11). Selected ranking measures are presented in numerical form in Tables 2.3 and 2.4.

<table>
<thead>
<tr>
<th>scen.</th>
<th>ROC power, FPR = 0.05</th>
<th>median ( p_{100} )</th>
<th>propn. ( p_{100} = 0 )</th>
<th>KS 2 sample test ( p_{100} ) cdfs the same</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>0.62 P-GLAW 0.35 GG 1.76</td>
<td>0.60 P-GLAW 0.60 GG 1.00</td>
<td>0.18 P-GLAW 0.26 GG 0.70</td>
<td>( p = 0.0082 )</td>
</tr>
<tr>
<td>(b)</td>
<td>0.61 P-GLAW 0.33 GG 1.84</td>
<td>0.33 P-GLAW 0.11 GG 3.00</td>
<td>0.21 P-GLAW 0.45 GG 0.46</td>
<td>( p = 9.6 \times 10^{-25} )</td>
</tr>
<tr>
<td>(c)</td>
<td>0.81 P-GLAW 0.54 GG 1.49</td>
<td>0.35 P-GLAW 0.20 GG 1.73</td>
<td>0.06 P-GLAW 0.23 GG 0.25</td>
<td>( p = 2.5 \times 10^{-25} )</td>
</tr>
<tr>
<td>(d)</td>
<td>0.44 P-GLAW 0.18 GG 2.37</td>
<td>0.33 P-GLAW 0.00 GG ( \infty )</td>
<td>0.30 P-GLAW 0.62 GG 0.48</td>
<td>( p = 7.7 \times 10^{-27} )</td>
</tr>
<tr>
<td>(e)</td>
<td>0.59 P-GLAW 0.27 GG 2.18</td>
<td>0.33 P-GLAW 0.01 GG 37.33</td>
<td>0.23 P-GLAW 0.50 GG 0.46</td>
<td>( p = 9.2 \times 10^{-28} )</td>
</tr>
<tr>
<td>(f)</td>
<td>0.79 P-GLAW 0.45 GG 1.74</td>
<td>0.31 P-GLAW 0.14 GG 2.31</td>
<td>0.06 P-GLAW 0.31 GG 0.20</td>
<td>( p = 3 \times 10^{-38} )</td>
</tr>
</tbody>
</table>

Table 2.3: Selected ranking performance measures for P-GLAW and GG for the 6 scenarios described in Table 2.1. ROC power, FPR = 0.05: proportion of 500 MC simulations with \( r_{k_1} \leq 28 \) corresponding to a FPR of 0.05. median \( p_{100} \): median of \( p_{100} \) distribution across 500 MC simulations. Proportion with \( p_{100} = 0 \): proportion of 500 MC simulations with no causal pathway in the top 100 ranks. KS 2 sample test: two-sample Kolmogorov-Smirnov test of the hypothesis that the P-GLAW and GG \( p_{100} \) population cdfs are the same.

Beginning with the ROC curves illustrating the \( r_{k_1} \) ranking measure (Fig. 2.9 and ‘ROC power’ column in Table 2.3), P-GLAW consistently demonstrates increased power and specificity across all of the top 100 ranks illustrated. The relative gain in power for P-GLAW is greater at the smallest effect size for each equivalent scenario, (a) vs. (d), (b)
vs. (e), and (c) vs. (f). At the smaller effect size, where causal SNPs are distributed randomly within causal pathways, power increases where the number of causal SNPs is fewer ((d) vs. (e)). Finally, maximum power is achieved for both methods where causal SNPs are located within a single gene ((c) and (f)).

Turning to the distributions of the $p_{100}$ ranking measure (Fig. 2.10, and ‘median $p_{100}$’ column in Table 2.3), P-GLAW again outperforms GG across all scenarios. For example, the null hypothesis that the two population cdfs are equal is rejected at the $\alpha = 0.05$ level (Table 2.3, final column), as is the null hypothesis that the two sample medians are the same (see Fig. 2.10, notches indicate 95% confidence intervals for median), except for scenario (a) where median $p_{100}$ is not significantly different for the two methods. Excluding scenario (a) where both methods perform relatively well, P-GLAW median $p_{100}$ is consistent across each scenario, and is maintained from the larger to the smaller effect size. This is in marked contrast to GG, where this measure shows a large decrease at the smaller effect size, although the decrease is less marked when causal SNPs are located within a single gene. A similar pattern persists for both P-GLAW and GG if we consider the proportion of simulations with $p_{100} = 0$, i.e. where no causal pathways are found in the top 100 ranks, except for P-GLAW in the case where causal SNPs are located in a single gene, where this measure is particularly low (Table 2.3, 4th column).

The final series of plots (Fig. 2.11), illustrate the distributions of the power-adjusted, normalised, weighted ranking score $R$ (2.32). These distributions once again follow the trends in ranking performance highlighted above, but they offer a more nuanced view, in the sense that while this measure takes power into account, it is also sensitive to the actual causal pathway rankings. Here we see that P-GLAW distributions are skewed towards lower $R$ values compared with GG, indicating that causal pathways tend to be ranked higher with P-GLAW. This is borne out if we focus on the proportion of simulations with $R < 10$ (Table 2.4, ‘$R < 10$’). Here we see once again that proportionate gains in ranking performance for P-GLAW over GG are largest for the smallest effect size ((a)-(c) vs. (d)-(f)). This table also gives results for the proportion of simulations showing near optimal ranking of causal pathways ($R < 3$), although the very small frequencies suggest that little can be inferred from these.
2.6 P-GLAW Simulation Study

Figure 2.9: ROC curves illustrating proportion of simulations with \( r_{k_1} \leq z \), for ranks \( z = 1, 2, \ldots, 100 \). Power is average across 500 simulations. False positive rate = \( \frac{z - 1}{L} \). Scenarios corresponding to the higher SNP effect size (\( \delta_k = 0.005 \)) are presented in the left-hand column, with the equivalent scenarios at the lower effect size (\( \delta_k = 0.001 \)) on the right.
Figure 2.10: Box plots of distribution of ranking power, $p_{100}$, across 500 simulations. This is the proportion, $|\hat{C}|_{100}/|C|$, of causal pathways that fall in the top 100 ranks. Notches indicate 95% confidence intervals for the true median.
Figure 2.11: Distribution of the power-adjusted, normalised, weighted ranking score, $R$, across 500 simulations. The final ‘50+’ column includes simulations for which no causal pathway was ranked in the top 100, i.e. $\hat{C}_{100} = \emptyset$ with $R_0 = 100$ (see (2.32)).
Table 2.4: Proportion of 500 simulations with $R < 10$ and $R < 3$ for the 6 scenarios described in Table 2.1.

2.7 Discussion

We have described a penalised regression-based strategy (P-GLAW) that exploits functional structure within genotypes to identify biological pathways associated with a continuous trait. We use the group lasso, with all mapped SNPs and pathways in a single regression model, and use a novel combination of methods including bias-adjusted group weights, a subsampling scheme to rank pathways, and a number of adjustments to the estimation algorithm designed to accelerate model estimation and make the analysis of large scale datasets computationally feasible. An important feature of our method is its ability to accommodate the presence of overlapping pathways. On the assumption that causal SNPs are enriched within a biological pathway, we find in a simulation study that our proposed method shows relative gains in both power and specificity across a range of scenarios, compared with an alternative pathways method (GG), based on univariate SNP statistics, that we use as a benchmark. We believe this is the first such study evaluating the performance of a GL-based method for pathway selection using real SNP and pathway data across a range of realistic scenarios.

One key motivation for a pathways-based approach is the desire to harness the joint effects of those SNPs or genes with relatively small effect size, that typically fail to achieve genome-wide significance in GWAS (Baranzini et al., 2009). We hypothesise that the advantages inherent in a multivariate approach to modelling SNP effects will increase power to detect these, and in our simulation study we therefore focus on scenarios with causal SNPs that exhibit effect sizes at or below the limits of those found in recent GWAS. To evaluate the performance of each method considered here, we devise three separate rank-
One factor affecting power is the ‘genetic architecture’ of the disease in question, that is the number and distribution of SNP effects across causal pathways (Wang, Li, and Hakonarson, 2010). For example, causal SNPs may be distributed across many genes in a pathway, or restricted to a single gene. Since PGAS methods vary in the way that they combine the effects of individual SNPs, the specific genetic architecture is expected to impact power for different methods in different ways (Wang et al., 2009b; Holmans et al., 2009). GG uses genes scores corresponding to the most significant SNP associated with a gene to establish pathway significance. This has the advantage of reducing redundant information arising from SNPs in LD with a causal SNP within a single gene, but may lead to reduced power where causal variants reside in distinct LD blocks within a gene (Wang, Li, and Bucan, 2007).

An important, related factor that we find has received little attention is the issue of overlapping pathways, and the consequent effect on PGAS performance. Aside from the issue of how to handle overlapping predictors in our regression model, the precise distribution of causal SNPs with respect to genes that overlap multiple pathways will affect the number of pathways that are considered to be ‘causal’, and we expect this to affect ranking performance for different methods in different ways. Most existing PGAS methods consider pathways individually. In certain circumstances this approach may be advantageous, for example where the goal is to identify as many pathways harbouring potentially causal genes as possible. However, a potential disadvantage of this approach is that it can fail to distinguish between pathways having a similar effect (Skarman et al., 2012), and it has been suggested that overlapping genes may render permutation-based procedures for the control of false positives less reliable (Wang, Li, and Bucan, 2007). One proposed solution is to reduce or down-weight the contribution of overlapping genes (Tarca et al., 2012), although this seems hard to justify on biological grounds. We instead force all pathways to compete in a single regression model, with the motivation that this will tend to distinguish important pathways from overlapping pathways with a smaller association signal. Ultimately the choice of PGAS model will depend on the goals of the analysis.

To explore some of these issues we investigate a variety of different genetic architectures, in which we vary both the number and distribution of causal SNPs within causal
pathways. In general, we find that P-GLAW performance is reasonably consistent across the range of causal SNP distributions and effect sizes considered. Additionally, our method is generally able to outperform the benchmark (GG). GG performance at the smaller effect size is particularly weak, so that P-GLAW shows the largest gains in relative performance here.

An insight into some of those factors affecting ranking performance for the two methods considered here is afforded by considering some of the ranking measures in more detail. Starting with the highest ranking causal pathway measure \( r_{k_1} \), as expected we find that this decreases for each scenario at the smaller effect size. However, at the smaller effect size this measure is observed to increase for both methods as the number of causal SNPs is decreased, markedly so when the reduced number of causal SNPs are concentrated in a single gene. Since the effect size for each causal SNP is held constant, this seems counterintuitive, since the pathway signal is reduced when there are fewer causal SNPs. In addition, for the reasons described above, for GG this signal may be further reduced where causal SNPs reside within a single gene. The explanation is likely to be that the effective number of causal pathways tends to increase as the number of causal SNPs is reduced, increasing the probability that a single causal pathway is ranked high (see Fig. 2.6). The number of causal pathways is at its largest when causal SNPs are concentrated in a single gene. This once again highlights the importance of considering the distribution of overlaps when interpreting PGAS results, and in particular the presence of so-called ‘hub genes’, that have the potential to dis-regulate large numbers of pathways (Kim, Wuchty, and Przytycka, 2011).

Where the pathway signal is strongest (scenario (a)), both methods tend to rank a high proportion of causal pathways in the top 100 (high \( p_{100} \)), although the proportion of MC simulations in which GG fails to rank any causal pathways (that is the proportion of simulations with \( p_{100} = 0 \)) is relatively high. On this measure of ranking power and in marked contrast to P-GLAW, GG performs relatively poorly across all other scenarios, particularly at the smaller effect size. This would seem to confirm our intuition that a multivariate method that retains all SNPs in a pathway may be more sensitive to smaller SNP effects. Finally, we note that P-GLAW is generally relatively insensitive to variation in the number and distribution of SNPs within causal pathways, as might be expected from the smoothing properties of the GL \( \ell_2 \) penalty, which ensures that all SNPs within a selected pathway are
2.7 Discussion

retained in the model (Zhou et al., 2010). The need to account for factors such as variation in LD, gene and pathway size is a feature common to all PGAS methods. A range of approaches, often used in combination, have been proposed to correct for these biasing factors, including the use of gene scores that summarise SNP statistics (Holmans et al., 2009), and permutation of phenotypes (Wang et al., 2009b). Dimensionality reduction techniques have also been advocated for the control of redundant information (Chen et al., 2010; Zhu and Li, 2011; Ballard, Cho, and Zhao, 2010). For P-GLAW, we propose a method that adjusts the distribution of pathway weights according to the observed bias in pathway selection frequencies across multiple MC simulations under the null. We find in a simulation study that our proposed bias correction method does substantially increase power and specificity, indicating that pathway selection bias is decreased. One potential disadvantage of our approach is that it takes no account of the variation in biasing factors across a pathway. It would be interesting to compare the relative merits of our approach against alternative bias-reduction methods, for example the use of PCA for within-pathway dimensionality reduction. However, we consider the retention of all SNPs in the regression model to be a potentially attractive feature of our approach, as it affords the possibility of the simultaneous identification of causal SNPs driving pathway selection, a feature that we explore in Chapter 4.

In situations where predictors, or groups of predictors are correlated, both the lasso and group lasso can demonstrate problems with consistency, that is they may be unable to identify the true set of causal predictors or groups, even for large sample sizes (Zhao and Yu, 2006; Wei and Huang, 2010). A range of strategies for improving model consistency have been proposed (Bach, 2008b; Wang et al., 2010; Wei and Huang, 2010), and for the lasso, the use of bootstrap or other subsampling techniques has been advocated (Chatterjee and Lahiri, 2011). For the lasso, a subsampling approach has been shown to offer a number of advantages including reduced sensitivity to the choice of $\lambda$, and much improved control of false positives (compared to alternative methods such as cross-validation), although this may come at the cost of a small reduction in power (Meinshausen and Bühlmann, 2010). We adopt a subsampling approach for the ranking of pathways and demonstrate promising group selection performance with the real dataset considered here.

We pay considerable attention to the need to develop fast algorithms for solving the
GL, a problem that is particularly acute when using regression models with GWAS data. Using a combination of techniques, we develop a GL estimation algorithm that can quickly solve the GL using whole genome data. However, the very large number of simulations and scenarios considered in our simulation study, and the relatively slow performance of the benchmark method mean that we restrict the analysis to mapped SNPs from a single chromosome.

We note that phenotypes in our simulation study are generated under an additive linear model for allelic effects, so that phenotypic effects increase linearly with minor allele dosage. This assumption of additive linear SNP effects is built into both the P-GLAW and GG models, in the former through the SNP allele codings in the genotype design matrix, and in the latter through the particular model used to generate the univariate SNP scores. For both methods alternative models, for example capturing dominant or recessive effects, can easily be accommodated.

Further discussion is reserved for Chapter 5, where we highlight issues arising from a comparison of the P-GLAW method with others discussed in following chapters.
Chapter 3

Identifying pathways associated with a multivariate quantitative trait: ‘Pathways Sparse Reduced-rank Regression’

We now turn to the problem of identifying gene pathways associated with a multivariate quantitative trait (MQT). We do this by incorporating the group sparsity constraint on genotype coefficients used in our previous P-GLAW model, into a multivariate sparse reduced-rank regression model, previously developed for the identification of SNPs (Vounou, Nichols, and Montana, 2010). As with P-GLAW, our proposed ‘Pathways Sparse Reduced-Rank Regression’ (PsRRR) method includes all SNPs in a single model, and accounts for potential biasing factors such as dependencies between phenotypes and SNPs using an adaptive, weight-tuning procedure.

A primary motivation for extending the previous model to the case of a multivariate response is our interest in analysing high-dimensional neuroimaging genetic datasets, although we note that the model and algorithms we describe here can be applied to any type of MQT. In neuroimaging genetics, the use of high-dimensional endophenotypes offers a potentially more power approach to identifying gene variants influencing brain structure and function, and can enable the voxel-wise mapping of genetic effects across the brain
Previous work has demonstrated that a sparse reduced-rank regression (sRRR) approach that exploits the multivariate nature of the phenotype can be more powerful than a mass-univariate linear modelling approach in which each phenotype is regressed against each SNP (Vounou, Nichols, and Montana, 2010).

In the following section we describe the PsRRR model in detail, before outlining an algorithm for coefficient estimation in Section 3.2. In Section 3.3 we explain how our previous subsampling strategy is used for pathway ranking with PsRRR, and introduce a method for additionally ranking SNPs and genes that may be driving pathway selection. We describe two simulation studies demonstrating the efficacy of our proposed method for both pathway and voxel selection in Section 3.4. In Section 3.5 we describe an application study, in which we identify a number of gene pathways associated with voxelwise structural change in subjects with Alzheimer’s Disease. We conclude with a discussion in Section 3.6.

3.1 The pathways sparse reduced-rank regression model

We consider the problem of identifying gene pathways associated with an MQT or phenotype, $y_1, \ldots, y_Q$. The observed values for phenotype $q$, measured for $N$ unrelated individuals, are arranged in an $(N \times 1)$, mean-centred response vector $y_q$, and the $Q$ phenotypes are arranged in an $(N \times Q)$ response matrix $Y = (y_1, \ldots, y_Q)$. Notation for genotypes and pathways is as described in the previous chapter.

If we denote by $C = (C_1, \ldots, C_Q)$, a $(P \times Q)$ matrix of regression coefficients, then we can model the multivariate response using a multivariate, multiple linear regression (MMLR) model, as

$$Y = XC + E$$

where $E$ is an $(N \times Q)$ matrix of error terms, with zero mean, possibly correlated columns. A least squares estimate for $C$ may be obtained by generalising the multiple least squares optimisation to include a multivariate response, that is by minimising the residual sum of squares

$$M_{MMLR} = \text{Tr}\{(Y - XC)(Y - XC)\}' \}. \quad (3.2)$$

Where $N > P$ and the design matrix $X$ is of full rank, the least squares estimates are given
by $\hat{C} = (X'X)^{-1}X'Y$. Note that the $(P \times 1)$ column vectors $\hat{C}_1, \ldots, \hat{C}_Q$ of $\hat{C}$ are then just the least squares estimates of the regression of each $y_q$ on $X$, that is

$$\hat{C}_q = \arg \min_{C_q} \|y_q - XC_q\|^2_2 \quad q = 1, \ldots, Q. \quad (3.3)$$

As was noted in the previous case with a univariate response (Section 2.1), multicollinearity and/or rank deficiency in the genotype matrix $X$ make the direct application of this model to the case of high-dimensional genomic datasets problematic. Furthermore, since the estimation (3.3) is equivalent to performing $Q$ independent regressions, it takes no account of the multivariate nature of $Y$. Ideally, we would like to exploit this in our estimation procedure to boost power (Breiman and Friedman, 1997; Vounou, Nichols, and Montana, 2010).

These limitations are addressed in reduced-rank regression (RRR) (Izenman, 2008), by restricting the rank of the coefficient matrix $C$. Specifically we impose the constraint that $C$ has rank $r < \min(P, Q)$, and rewrite $C$ as $C = BA$, where $A$ and $B$ both have (full) rank $r$. The reduced rank form of (3.1) is then given by

$$Y = XBA + E \quad (3.4)$$

where $B$ and $A$ are $(P \times r)$ and $(r \times Q)$ matrices of regression coefficients respectively relating to genotypes and phenotypes (Fig. 3.1).

![Figure 3.1: Reduced-rank regression model.](image-url)
This model has an interesting interpretation as exposing $r$ hidden or latent factors, which capture the major part of the relationship between $Y$ and $X$. If we denote by $B_{(k)}$, the $k$th column of $B$, then we see that the products $XB_{(k)}$, $k = 1, \ldots, r$, represent $r$ linear combinations of the $P$ predictor variables. Similarly, the $r$ row vectors, $A_{(k)}$, $k = 1, \ldots, r$, represent the transformation of each of these back to the dimensions of $Y$, so that they can predict the response. The linear combinations $XB_{(k)}$ and $YA'_{(k)}$ thus represent a reduced set of $r$ (latent) factors that capture the relationship between response and predictors, reduced in the sense that this set has dimensionality $r < \min(P, Q)$ (see Figure 3.2).

Figure 3.2: Reduced-rank regression latent factors.

We now consider the rank-1 RRR model which captures the first, main set of genotype and phenotype latent factors describing the association between $X$ and $Y$. With $r = 1$, we rewrite (3.4) as

$$Y = Xba + E$$

(3.5)

where $b$ and $a$ are $(P \times 1)$ and $(1 \times Q)$ coefficient vectors respectively relating to genotypes and phenotypes. Least squares estimates for $b$ and $a$ are then obtained by minimising the
3.1 The pathways sparse reduced-rank regression model

rank-1 equivalent of (3.2),

\[ M^{RR}_1 = \text{Tr}\{(Y - Xba)\Gamma(Y - Xba)\} \]  

(3.6)

where \( \Gamma \) is a given \((q \times q)\) positive definite matrix of weights. The choice of \( \Gamma \) reflects how we deal with correlation between the responses \( y_1, \ldots, y_q \) in the least squares optimisation. Such correlations can be exploited by setting \( \Gamma \) to be the inverse of the estimated covariance of the responses. In the context of imaging genetics for example, where a voxel-wise multivariate response may be derived from structural MRI, spatial correlations between phenotypes are expected in part to reflect common genetic variation. However, the calculation of \( Y'Y \) is computationally very intensive, and is in any case likely to be inaccurate for small sample sizes, so we instead use the simplifying approximation \( \Gamma = I_q \), effectively assuming the responses to be uncorrelated (Vounou, Nichols, and Montana, 2010; Vounou et al., 2011).

In sparse reduced-rank regression (sRRR), Vounou, Nichols, and Montana (2010) extend the RRR model by imposing a regularisation penalty on \( b \) and/or \( a \), to obtain sparse estimates for genotype and/or phenotype coefficient vectors respectively. In our proposed ‘Pathways Sparse Reduced-Rank Regression’ (PsRRR) model, we obtain group-sparse estimates for \( b \) by imposing an additional group lasso penalty on (3.6). As before, we do this under the assumption that only a small proportion of SNPs will be ‘causal’, in the sense that they exhibit phenotypic effects, and further assume that causal SNPs will tend to be enriched within functional groups, or gene pathways.

As with the previous P-GLAW model, we assume that all \( P \) SNPs may be mapped to \( L \) groups, \( G_l \subset \{1, \ldots, P\} \), \( l = 1, \ldots, L \), and further assume that groups are disjoint or non-overlapping, so that in the case of overlapping pathways, a suitable expansion of the genotype matrix \( X \) and coefficient vector \( b \) has been performed (Section 2.2.1). We denote the rank-1 vector of SNP regression coefficients by \( b = (b_1, \ldots, b_P) \), and the corresponding vector of coefficients for SNPs mapping to \( G_l \) by \( b_l = (b_{l1}, b_{l2}, \ldots, b_{Pl}) \).

Group-sparse solutions to the rank-1 RRR model (3.5) are then obtained by minimising
the penalised least squares problem

\[ M_{\ell_2}^{PsRR_{1:R}} = \frac{1}{2} \text{Tr}\{(Y - Xba)(Y - Xba)\}'\} + \lambda \sum_{l=1}^{L} w_l ||b_l||_2 \]  

(3.7)

with respect to \( b \) and \( a \). This corresponds to the OLS optimisation (3.6), but with an additional group-wise penalty on the \( \ell_2 \)-norm of the pathway coefficient vectors \( ||b_l||_2, l = 1, \ldots, L \); a regularisation parameter \( \lambda \); and an additional group weighting parameter \( \omega_l \) that can vary from group to group. As with the GL penalty in the previous P-GLAW model (Section 2.2), depending on the value of \( \lambda \), this penalty has the effect of setting multiple pathway SNP coefficient vectors, \( b_l = 0, l \subset \{1, \ldots, L\} \), thereby enforcing group sparsity. Pathways with non-zero coefficient vectors then form the set \( \hat{C} \) of selected pathways, so that

\[ \hat{C}(\lambda) = \{l : \hat{b}_l \neq 0\}. \]

Finally, if we assume that only a small proportion of phenotypes (or voxels), \( Q \), will be affected by SNPs residing in causal pathways, then we can additionally perform voxel selection, by imposing an additional sparsity constraint on the phenotype coefficient vector. For example, with an \( \ell_1 \) lasso-type constraint on \( a \), the minimisation (3.7) becomes

\[ M_{\ell_1, \ell_2}^{PsRR_{1:R}} = \frac{1}{2} \text{Tr}\{(Y - Xba)(Y - Xba)\}'\} + \lambda_1 \sum_{l=1}^{L} w_l ||b_l||_2 + \lambda_2 ||a||_1. \]  

(3.8)

Note that the additional penalty entails an extra regularisation parameter, \( \lambda_2 \). The set \( \hat{Q} \) of selected voxels is then given by

\[ \hat{Q}(\lambda_1, \lambda_2) = \{q : \hat{a}_q \neq 0\} \]

where \( \hat{a}_q \) are the estimated coefficients of the rank-1 phenotype coefficient vector \( a \), respectively mapping to phenotype, \( q = 1, \ldots, Q \).
3.2 Model estimation

We now describe an efficient algorithm for estimating genotype and phenotype coefficients by minimising the group-sparse, rank-1 penalised least squares problem (3.7). Expanding (3.7), and noting that the first term $YY'$ does not depend on $b$ or $a$, solutions satisfy

$$\hat{b}, \hat{a} = \arg \min_{b,a} \left\{ \frac{1}{2}(-2a'Yb + aa'b'X'Xb) + \lambda \sum_{l=1}^{L} w_l ||b_l||_2 \right\}.$$  

(3.9)

This optimisation is bi-convex in $a$ and $b$, in the sense that it is convex in $a$ for fixed $b$ and vice versa (Gorski, Pfeuffer, and Klamroth, 2007). This makes it amenable to solution using coordinate descent (Friedman et al., 2007), and a global solution can then be obtained by iteratively estimating one coefficient vector ($b$ or $a$), while holding the other fixed at its current value, until convergence (Chen and Chan, 2012).

Thus, for fixed $b = \hat{b}$, and with the additional constraint that $\hat{b}'\hat{b} = 1$, we estimate $\hat{a}$ as

$$\hat{a} = \arg \min_{a} \left\{ \frac{1}{2}(-2a'Y\hat{b} + aa'\hat{b}'X'X\hat{b}) + \lambda \sum_{l=1}^{L} w_l ||\hat{b}_l||_2 \right\}.$$  

Differentiating and setting to zero gives

$$\hat{a} = \frac{\hat{b}'X'Y}{\hat{b}'X'X\hat{b}}.$$  

Similarly, for fixed $a = \hat{a}$, and with the additional constraint that $\hat{a}\hat{a}' = 1$, we have

$$\hat{b} = \arg \min_{b} \left\{ \frac{1}{2}(-2\hat{a}Y'Xb + b'X'Xb) + \lambda \sum_{l=1}^{L} w_l ||b_l||_2 \right\}.$$  

(3.10)

This is precisely the standard group lasso optimisation (2.15), but with the univariate response vector, $y$, replaced by its rank-1 equivalent, $Y\hat{a}'$. (3.10) can thus be solved using block coordinate descent (BCD) as described in Box 2.1. The equivalent BCD estimation algorithm with multivariate $Y$ is presented in Box 3.1.

As with the effect of the GL penalty in the previous P-GLAW model, as $\lambda$ increases, fewer groups (or pathways) are selected by the model (Box 3.1, step 5), while for selected
Box 3.1 $\Omega(\hat{a}, Y, X, \lambda)$: Rank-1 PsRRR estimation of $b$ using BCD

1. $b \leftarrow \hat{b}$ (initialise to current estimate)  
   block coordinate descent
2. repeat
3. for $l = 1, 2, \ldots, L$
4.  
5. if $\|X'_l r_l\|_2 \leq \lambda w_l$
6.  
7. else
   coordinate descent within block
8. repeat
9. for $j = l_1, \ldots, l_{P_l}$
10.  
11.  
12. until $b_l$ converges
13. until $b$ converges
14. $\Omega \leftarrow b$

pathways with $b_l \neq 0$, estimated SNP coefficients, $b_j, j = l_1, \ldots, l_{P_l}$, tend to shrink towards zero (Box 3.1, step 11).

The full PsRRR estimation algorithm is presented in Box 3.2. The addition of a phenotype vector sparsity constraint (3.8) is easily accommodated by adding a coordinate descent procedure to the estimation of $a$ in Box 3.2 step 5, as described in Section 2.1.2.

As with P-GLAW, before proceeding with the full model estimation, we apply an initial weight tuning step, to account for factors that may bias pathway selection. To do this we follow the adaptive weighting strategy described previously (see Section 2.4), but with $R$ PsRRR estimations performed over multiple permutations of the rows of the phenotype matrix, $Y$.

Estimates for $b$ and $a$ respectively represent the first (rank 1) latent factors that are expected to capture the strongest signal of association between gene pathways and the pheno-
3.3 Pathway, gene and SNP ranking

Box 3.2 Rank-1 PsRRR full estimation algorithm using coordinate descent

1. initialise:
   \( a_0 \leftarrow 1/||1||_2; \ b_0 \leftarrow 1/||1||_2 \)
   \( \lambda \leftarrow c\lambda_{max} \)
   where \( \lambda_{max} = \min \{ \lambda : ||X_l^T Y a_0'||_2 = \lambda w_l, \ l = 1, \ldots, L \} \)
   and \( 0 < c < 1 \)

2. repeat:
   3. \( b \leftarrow \Omega(a_0, Y, X, \lambda) \) (from Box 3.1)
   4. \( b \leftarrow b/||b||_2 \) (normalise)
   5. \( a \leftarrow \frac{b'X'Y}{b'X'xb} \)
   6. \( a \leftarrow a/||a||_2 \) (normalise)
   7. \( b_0 \leftarrow b; a_0 \leftarrow a \)
   8. until \( b \) and \( a \) converge

In principle, it is possible to capture further latent factors of diminishing importance, by iteratively repeating the procedure described above, after regressing out the effects of previous factors (Vounou, Nichols, and Montana, 2010). With PsRRR however, the estimation of further ranks is complicated by the fact that pathways overlap, by the typically large number of SNPs in selected pathways, and by the need to recalibrate the group weights at each step. For this reason we consider only the first latent factor in the application study described in Section 3.5.

3.3 Pathway, gene and SNP ranking

3.3.1 Pathway ranking

As with P-GLAW, we use a resampling strategy to rank pathways by repeatedly fitting the model over \( B \) subsamples, each of size \( N/2 \), drawn without replacement, at a fixed value for \( \lambda \). We denote the set of selected pathways at subsample \( b \) by

\[ \hat{C}^{(b)} = \{ l : \hat{b}_l^{(b)} \neq 0 \} \quad b = 1, \ldots, B. \]
The selection probability for pathway $l$ measured across all $B$ subsamples is then

$$
\pi_{l}^{path} = \frac{1}{B} \sum_{b=1}^{B} I_{l}^{(b)} \quad l = 1, \ldots, L
$$

where the indicator function, $I_{l}^{(b)} = 1$ if $l \in \hat{C}^{(b)}$, and 0 otherwise. Pathways are ranked in order of their selection probabilities, $\pi_{1}^{path} \geq \ldots \geq \pi_{L}^{path}$.

### 3.3.2 SNP and gene ranking

The PsRRR model is designed to identify important pathways which may contain multiple genetic markers with varying effect sizes. However, it is still interesting to establish which SNPs and genes are most predictive of the response amongst those mapped to the set $\hat{C}^{(b)}$ of selected pathways at subsample $b$. Note that these are not necessarily the SNPs and genes that are driving the selection of any particular pathway in the PsRRR model.

To do this, we perform a second level of variable selection using sRRR with a lasso penalty (Vounou et al., 2011). In the unexpanded variable space, we first form the reduced $(N \times Z^{(b)})$ matrix $X_{\hat{C}^{(b)}}$, with columns $\{x_{j} : j \in \bigcup_{l \in \hat{C}^{(b)}} G_{l}\}$ corresponding to all SNPs in pathways selected at subsample $b$. Sparse estimates for the corresponding SNP coefficient vector, $\beta$, and rank-1 phenotype vector $\alpha$ then satisfy the equivalent of (3.9) with a lasso penalty, namely

$$
\hat{\beta}, \hat{\alpha} = \arg \min_{\beta, \alpha} \left\{ \frac{1}{2}(-2\alpha Y'X_{\hat{C}^{(b)}}\beta + \alpha \alpha'X'_{\hat{C}^{(b)}}X_{\hat{C}^{(b)}}\beta) + \lambda \| \beta \|_{1} \right\}.
$$

We denote the set of SNPs selected at sample $b$ by $\hat{S}^{(b)}$, and further denote the set of selected genes to which the SNPs in $\hat{S}^{(b)}$ are mapped by $\hat{\varphi}^{(b)} \subset \Phi$, where $\Phi = \{1, \ldots, G\}$ is the set of gene indices corresponding to all $G$ mapped genes. Using the same strategy as for pathway ranking, we obtain an expression for the selection probability of SNP $j$ across $B$ subsamples as

$$
\pi_{j}^{SNP} = \frac{1}{B} \sum_{b=1}^{B} I_{j}^{(b)}
$$

where the indicator function, $I_{j}^{(b)} = 1$ if $j \in \hat{S}^{(b)}$, and 0 otherwise. A similar expression
3.4 Simulation studies

for the selection probability for gene $g$ is

$$\pi_{g}^{\text{gene}} = \frac{1}{B} \sum_{b=1}^{B} J_{g}^{(b)}$$

where the indicator function, $J_{g}^{(b)} = 1$ if $g \in \hat{\phi}^{(b)}$, and 0 otherwise. SNPs and genes are then ranked in order of their respective selection frequencies.

3.4 Simulation studies

In this section we describe two simulation studies to demonstrate proof of principle of the efficacy of our proposed PsRRR model and estimation algorithm. In the first study we confirm the intuition that the power to detect causal pathways will depend on both SNP effect size, and the proportion of affected phenotypes for a multivariate trait. In a second simulation study, we focus on the convergence of the phenotype coefficient vector in PsRRR. To illustrate this we impose an additional lasso penalty on $a$ (3.8), so that pathway and voxel selection are performed simultaneously. We then explore the relationship between pathway and voxel selection power, while again varying genetic effect size and the proportion of affected phenotypes.

3.4.1 Data simulation

We simulate $P = 5000$ genetic markers for $N = 400$ individuals. Marker frequencies for each SNP are sampled independently from a multinomial distribution following a Hardy Weinberg equilibrium frequency distribution, so that for SNP $j$ with minor allele frequency $m_j$, $p(x_{ij}=0) = m_j^2$; $p(x_{ij}=1) = 2m_j(1 - m_j)$ and $p(x_{ij}=2) = m_j^2$. SNP minor allele frequencies are sampled from a uniform distribution $U(0.1, 0.5)$.

SNPs are distributed equally between 50 non-overlapping pathways, each containing 100 SNPs. Since pathways are non-overlapping, of equal size, and SNP genotypes are independent, these factors cannot bias pathway selection. While some pathway selection bias may still be present due to variation in $\{m_j\}$, we make the simplifying assumption that this will be negligible, and use a uniform pathway weighting vector $w = 1$. 
A baseline multivariate phenotype, \( y_q, q = 1, \ldots, 1000 \) is sampled from the multivariate normal distribution \( \mathcal{M}(\mu, \Sigma) \), with mean vector \( \mu = (\mu_1, \ldots, \mu_{1000}) \) sampled from \( \mathcal{N}(10, 1) \) and \( \Sigma = I_{1000} \). To generate SNP effects, we first select a single pathway, \( G_l \), at random. From this pathway we randomly select 5 SNPs to from the set \( S \subset G_l \) of causal SNPs affecting one or more phenotypes. At each MC simulation we generate a genetic effect in a fixed proportion, \( \rho \), of all phenotypes, and denote the set of affected phenotypes by \( Q \subset \{1, \ldots, Q\} \), so that \( |Q| = 1000\rho \). Affected phenotype \( q \) is then adjusted so that

\[
y_q^* = y_q + w_q
\]

where

\[
w_q = \delta_q \sum_{k \in S} \zeta_k x_k.
\]

Here \( \delta_q \) controls the overall additive genetic effect on phenotype \( y_q \) due to all casual SNPs in \( S \), and \( \zeta_k \) determines the contribution from causal SNP \( k \), with \( \sum_{k \in S} \zeta_k = 1 \). In our simulations we maintain a constant overall genetic effect size,

\[
\gamma = \frac{\mathbb{E}(w_q)}{\mathbb{E}(y_q)}
\]

across all affected phenotypes, so that \( \gamma \) represents the proportionate increase in the mean value of \( y_q \) due to all genetic effects. We also set \( \zeta_k = 1/5 \), for \( k \in S \), so that the contribution from each causal SNP allele is equal. This enables us to determine \( \delta_q \) for a given \( \gamma \) as

\[
\delta_q = \frac{5\gamma \mathbb{E}(y_q)}{2 \sum_{k \in S} m_k}.
\]

Note that for constant \( \gamma \), the proportionate effect on the mean value of \( y_q \) due to SNP \( k \) is MAF dependent, and is given by \( 2\delta_q m_k / \mathbb{E}(y_q) \).

Finally, we denote the mean signal to noise ratio by

\[
SNR = \frac{1}{Q} \sum_{q \in Q} \frac{\text{Var}(w_q)}{\text{Var}(y_q)}.
\]
3.4 Simulation studies

3.4.2 PsRRR Simulation study 1: Pathway selection

In the first simulation study, we investigate how the ability of PsRRR to correctly identify a single causal pathway depends on both the overall genetic effect size, \( \gamma \), and the proportion of affected phenotypes, \( \rho \). We use the estimation procedure described in Section 3.2, with a sparsity constraint imposed on the genotype coefficient vector, \( b \), only. Each scenario (\( \gamma / \rho \) combination) is tested over 500 MC simulations. For each simulation we randomly select 5 causal SNPs from a single causal pathway, and generate a multivariate phenotype according to the simulation procedure described in the preceding section. We consider a range of values for both \( \gamma \) and \( \rho \), and present the corresponding values for the observed mean signal to noise ratio (3.11) across 500 MC simulations in Table 3.1. SNR is observed to increase both with increasing \( \gamma \) and with increasing \( \rho \).

<table>
<thead>
<tr>
<th>( \gamma )</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>( 2.31 \times 10^{-06} )</td>
<td>( 4.51 \times 10^{-06} )</td>
<td>( 7.05 \times 10^{-06} )</td>
<td>( 1.41 \times 10^{-05} )</td>
<td>( 1.64 \times 10^{-05} )</td>
</tr>
<tr>
<td>0.003</td>
<td>( 2.07 \times 10^{-05} )</td>
<td>( 4.10 \times 10^{-05} )</td>
<td>( 6.31 \times 10^{-05} )</td>
<td>( 8.31 \times 10^{-05} )</td>
<td>( 1.04 \times 10^{-04} )</td>
</tr>
<tr>
<td>0.005</td>
<td>( 5.79 \times 10^{-05} )</td>
<td>( 1.15 \times 10^{-04} )</td>
<td>( 1.69 \times 10^{-04} )</td>
<td>( 2.32 \times 10^{-04} )</td>
<td>( 2.93 \times 10^{-04} )</td>
</tr>
<tr>
<td>0.007</td>
<td>( 1.11 \times 10^{-04} )</td>
<td>( 2.23 \times 10^{-04} )</td>
<td>( 3.42 \times 10^{-04} )</td>
<td>( 4.62 \times 10^{-04} )</td>
<td>( 5.73 \times 10^{-04} )</td>
</tr>
<tr>
<td>0.009</td>
<td>( 1.86 \times 10^{-04} )</td>
<td>( 3.79 \times 10^{-04} )</td>
<td>( 5.64 \times 10^{-04} )</td>
<td>( 7.58 \times 10^{-04} )</td>
<td>( 9.55 \times 10^{-04} )</td>
</tr>
</tbody>
</table>

Table 3.1: Empirical phenotypic SNR values (3.11) corresponding to the range of overall genetic effect size (\( \gamma \)) and proportion of affected phenotypes (\( \rho \)) used in Simulation study 1. Each figure is average SNR across 500 MC simulations.

At each simulation, the regularisation parameter, \( \lambda \) is tuned so that a single pathway is selected by the model. Results are shown in Fig. 3.3. It is clear that power increases both with increasing genetic effect size on each affected phenotype, and also as the number of affected phenotypes as a proportion of all phenotypes increases. Comparison with the empirical mean signal to noise ratios presented in Table 3.1 shows that the power to detect a causal pathway closely follows SNR.
3.4.3 PsRRR Simulation study 2: Simultaneous pathway and voxel selection

In the second simulation study we perform simultaneous pathway and voxel selection by imposing an additional lasso constraint on the phenotype coefficient vector (3.8). The simulation and model estimation procedures are as described for the previous simulation study, except that for each MC simulation, sparse estimates for the phenotype coefficient vector are estimated using the coordinate descent procedure described in Section 2.1.2. This replaces the non-sparse $a$ update, described in Box 3.2 step 5. Once again, the group-sparsity (pathway) regularisation parameter $\lambda_1$ is tuned so that a single pathway is selected at each MC simulation. We additionally set the lasso (phenotype) regularisation parameter $\lambda_2 = 0.3\lambda_{\text{SNP}}^{\text{max}}$ (see (3.8)), where $\lambda_{\text{SNP}}^{\text{max}}$ is the smallest value for $\lambda_2$ at which no SNPs are selected (i.e. the equivalent of (2.27)). This results in an average of 240 voxels selected at each simulation.

For each scenario, the number of affected phenotypes, $|Q| = \rho Q$ varies, as does the
number of selected voxels, $|\hat{Q}|$, across MC simulations. For this reason we concentrate on selection precision, or the number of correctly selected voxels as a proportion of $|\hat{Q}|$. At MC simulation $z$, this is given by

$$V_z(\gamma, \rho) = \frac{|\hat{Q}_z \cap Q_z|}{|\hat{Q}_z|},$$

where $|\hat{Q}_z|$ and $|Q_z|$ are respectively the number of selected voxels and the number of affected voxels at simulation $z$. We then report the average value of this quantity across all 500 MC simulations,

$$V(\gamma, \rho) = \frac{1}{500} \sum_{z=1}^{500} V_z(\gamma, \rho),$$

so that $V = 1$ when there is maximum precision (no false positives), and $V = 0$ when the precision is zero, that is all selected voxels are false positives.

$SNR$ values for the range of scenarios explored in this simulation study are presented in Table 3.2, and results are presented in Figure 3.4.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>0.0025</th>
<th>0.005</th>
<th>0.01</th>
<th>0.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>$1.46 \times 10^{-05}$</td>
<td>$2.97 \times 10^{-05}$</td>
<td>$4.41 \times 10^{-05}$</td>
<td>$5.79 \times 10^{-05}$</td>
</tr>
<tr>
<td>0.2</td>
<td>$5.85 \times 10^{-05}$</td>
<td>$1.17 \times 10^{-04}$</td>
<td>$1.78 \times 10^{-04}$</td>
<td>$2.36 \times 10^{-04}$</td>
</tr>
<tr>
<td>0.3</td>
<td>$2.34 \times 10^{-04}$</td>
<td>$4.71 \times 10^{-04}$</td>
<td>$7.08 \times 10^{-04}$</td>
<td>$9.42 \times 10^{-04}$</td>
</tr>
<tr>
<td>0.4</td>
<td>$2.34 \times 10^{-04}$</td>
<td>$4.71 \times 10^{-04}$</td>
<td>$7.08 \times 10^{-04}$</td>
<td>$9.42 \times 10^{-04}$</td>
</tr>
<tr>
<td>0.5</td>
<td>$2.34 \times 10^{-04}$</td>
<td>$4.71 \times 10^{-04}$</td>
<td>$7.08 \times 10^{-04}$</td>
<td>$9.42 \times 10^{-04}$</td>
</tr>
</tbody>
</table>

Table 3.2: Empirical phenotypic $SNR$ values for Simulation study 2.

The left hand plot in Figure 3.4 shows variation in pathway selection power, which is the proportion of simulations in which the correct causal pathway is selected. The right hand plot shows voxel selection precision, $V(\gamma, \rho)$. Note that by chance, the expected precision is just $\rho$, the proportion $|Q_z|/Q$ of causal phenotypes. This expected value is indicated by the dashed line.

Pathway selection power (Figure 3.4, left-hand plot) is broadly in line with that observed in Simulation study 1, where there was no voxel selection. Note however that the range of effect sizes explored in this simulation study stretches beyond those used in the previous study, to include larger pathway effects. This is because voxel selection per-
Figure 3.4: Simulation study 2. Left hand: proportion of 500 MC simulations in which the true causal pathway is selected. Right hand: Voxel selection performance. This is the proportion of selected voxels that are ‘affected’, i.e. that exhibit genetic effects, averaged over 500 MC simulations (see (3.12)). The dashed line represents the value of this measure expected by chance.

Performance lags behind pathway selection performance, in the sense that pathway selection power is observed to be high, even when voxel selection performance is relatively poor. For example with $\gamma = 0.005$ (green lines in Figure 3.4), pathway selection power is relatively high where $\rho \geq 0.4$, and yet the proportion of correctly selected voxels, $V$, is no greater than that expected by chance.

With real data, relative pathway and voxel selection performance is expected to depend on numerous factors including, for example, the number of voxels, SNPs and pathways, and the extent of correlations between voxels and between SNPs. Both the preceding simulation studies are presented primarily as a proof of principle of the efficacy of our proposed algorithm, as well as providing a simple illustration that pathway selection power will depend on maximising the signal to noise ratio in the phenotype. In the context of identifying pathways in an imaging genetics study, this entails selecting an imaging phenotype that is likely to maximise the phenotypic signal to noise ratio of the genetic effects under investigation.
3.5 PsRRR Application study: Gene pathways implicated in Alzheimer’s disease

Identification of gene pathways implicated in Alzheimer’s disease using longitudinal imaging phenotypes

A growing list of genetic variants has now been associated with greater susceptibility to develop early and late-onset Alzheimer’s disease (AD), with the $\text{APOE}^{\epsilon 4}$ allele consistently identified as having the largest effect (for an up to date list see www.alzgene.org). SNP association studies have been augmented by numerous studies considering differential gene expression across various brain regions, again focusing on differences between cases and controls (Zou et al., 2010). Recently, case-control susceptibility studies have been augmented by studies using neuroimaging phenotypes. The rationale here is that the use of heritable imaging signatures (endophenotypes) of disease may increase the power to detect causal variants, since gene effects are expected to be more penetrant at this level (see Section 1.3). This approach has been used to identify genes associated with a range of AD-associated imaging phenotypes including measures of hippocampal volume (Stein et al., 2012), cortical thickness (Burrgren et al., 2008) and longitudinal changes in brain structure (Vounou et al., 2011).

AD is a moderate to highly heritable condition, yet as with many common heritable diseases, association studies have to date identified gene variants explaining only a relatively modest amount of known AD heritability (Braskie, Ringman, and Thompson, 2011). We therefore perform a pathways analysis in the hope that we can reveal aspects of the disease’s genetic architecture that may not be amenable to standard methods. A longer term goal is that studies such as these will lead to a deeper understanding of the underlying mechanisms by which gene mutations impact disease etiology, potentially playing an important role in the translation of basic AD biology into therapy and patient care (Sleegers et al., 2010).

At least two recent studies have performed pathways analyses using dichotomous (case-control) phenotypes (Lambert et al., 2009; Jones et al., 2010). In the following sections we describe what we believe to be the first AD pathways study using a multivariate imaging endophenotype. Specifically, we apply the PsRRR method described earlier in this chapter to a pathways analysis of the ADNI cohort (www.loni.ucla.edu/ADNI), comparing genome-wide SNP data with voxel-wise tensor-based morphometry (TBM) maps describing longi-
tudinal structural changes that are characteristic of AD. As in our previous P-GLAW simu-
lution study (Section 2.6), we map SNPs to pathways from the KEGG pathways database,
a curated collection of functional gene pathways representing current knowledge of mole-
cular interaction and reaction networks (http://www.genome.jp/kegg/pathway.html). Our
method is however able to accommodate alternative sources of information for the group-
ing of SNPs and genes, for example using gene ontology (GO) terms, or information from
protein interaction networks (Wu, Feng, and Stein, 2010; Jensen and Bork, 2010).

In this study we use a high-dimensional phenotype describing structural change relative
to baseline over three time points in subjects with AD, and in healthy controls. From this
we extract an imaging endophenotype that is highly characteristic of AD in our sample by
using a stringent statistical threshold to exclude voxels that do not discriminate between
AD and CN. Our main objective here is not to build a robust statistical classifier for AD,
but instead to produce a quantitative phenotype having maximal sample variability between
AD and CN for the subsequent gene mapping stage of our analysis.

The study is presented as follows. We begin in section 3.5.1 with a description of the
voxel-wise TBM maps used in the analysis, and in section 3.5.2 we outline how we use
these maps to generate an imaging signature characteristic of structural change in AD, that
is able to discriminate between AD patients and controls in our sample. In section 3.5.3 we
describe the genotype data used in the study, together with quality control procedures, and
in section 3.5.4 we explain how this genotype data is mapped to gene pathways. We discuss
our strategies for addressing the significant computational challenge of fitting a regression-
based model with such high dimensional datasets in section 3.5.5. Pathway, SNP and gene
ranking results are presented in section 3.5.6, and we conclude with a discussion in section
3.6.

3.5.1 Imaging data

Image pre-processing was carried out by Xue Hua and Paul Thompson from the Laboratory
of Neuro Imaging, Department of Neurology at the UCLA School of Medicine, USA.

Longitudinal brain MRI scans (1.5 Tesla) were downloaded from the ADNI public
database (http://www.loni.ucla.edu/ADNI/Data/). Serial brain MRI scans ($N = 1446$; see
Table 3.3) were analysed from 200 probable AD patients and 232 healthy elderly controls (CN). AD and CN subjects were scanned at screening and followed up at 6, 12, and 24 months. Additional scans for CN subjects at 36 and 48 months are not included in this study. All subjects were scanned with a standardised 1.5T MP-RAGE protocol developed for ADNI (Jack et al., 2008). The typical acquisition parameters were repetition time (TR) of 2400 ms, minimum full echo time (TE), inversion time (TI) of 1000 ms, flip angle of 8°, 24 cm field of view, 192 × 192 × 166 acquisition matrix in the x−, y−, and z− dimensions, yielding a voxel size of 1.25 × 1.25 × 1.2 mm³, later reconstructed to 1 mm isotropic voxels. Image correction steps included gradwarp (Jovicich et al., 2006), B1-correction (Jack et al., 2008), N3 bias field correction (Sled, Zijdenbos, and Evans, 1998), and phantom-based geometrical scaling (Gunter et al., 2006).

Linear registration (9-parameter) was used to align the longitudinal scan series of each subject, and then the mutually aligned time-series was registered to the International Consortium for Brain Mapping template (ICBM-53) (Mazziotta et al., 2001). Brain masks that excluded skull, other non-brain tissues, and the image background were generated automatically using a parameter-less robust brain extraction tool (ROBEX) (Iglesias et al., 2011).

Individual Jacobian maps representing voxelwise structural brain change were created by warping the skull-stripped, globally registered and scaled follow-up scan to match the corresponding screening scan. We used a non-linear, inverse consistent, elastic intensity-based registration algorithm (Leow et al., 2005), which optimises a joint cost function based on mutual information (MI) and the elastic energy of the deformation. Colour-coded maps of the Jacobian determinants were created to illustrate regions of ventricular/CSF expansion (i.e., with det \( J(r) > 1 \)), or brain tissue loss (i.e., with det \( J(r) < 1 \)) (Frackowiak et al., 2003; Chung et al., 2001; Freeborough and Fox, 1998; Riddle et al., 2004; Thompson et al., 2000; Toga, 1999) over time. These longitudinal maps of tissue change were also spatially normalised across subjects by non-linearly aligning all individual Jacobian maps to an average group template known as the minimal deformation target (MDT), for regional comparisons and group statistical analyses.

The study was conducted according to the Good Clinical Practice guidelines, the Declaration of Helsinki and U.S. 21 CFR Part 50-Protection of Human Subjects, and Part 56-Institutional Review Boards. Written informed consent was obtained from all participants.
before experimental procedures, including cognitive tests, were performed.

Table 3.3: Available scans at 6, 12 and 24 months for the ADNI-1 dataset (downloaded on February 28, 2011)

<table>
<thead>
<tr>
<th></th>
<th>Screening</th>
<th>6Mo</th>
<th>12Mo</th>
<th>24Mo</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>200</td>
<td>165</td>
<td>144</td>
<td>111</td>
</tr>
<tr>
<td>CN</td>
<td>232</td>
<td>214</td>
<td>202</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>432</td>
<td>379</td>
<td>346</td>
<td>289</td>
</tr>
</tbody>
</table>

At screening:

<table>
<thead>
<tr>
<th>Group</th>
<th>age (years)</th>
<th>N male</th>
<th>N female</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>75.7±7.7</td>
<td>103</td>
<td>97</td>
</tr>
<tr>
<td>CN</td>
<td>76.0±5.0</td>
<td>120</td>
<td>112</td>
</tr>
</tbody>
</table>

3.5.2 Phenotype extraction

Phenotype extraction was carried out by Eva Janousova, at the time visiting Imperial College from the Institute of Biostatistics and Analyses, Masaryk University, Brno in the Czech Republic.

For the study, we include 253 individuals (99 AD, 154 CN) with longitudinal maps at all three time points (6, 12 and 24 months), who have also been genotyped by ADNI.

To maximise the power to detect causal pathways, we seek a phenotype which is highly representative of those structural changes in the brain that are characteristic of AD. One way to do this is to use prior knowledge on regions of interest (ROI) to extract a univariate quantitative measure as a disease signature (Potkin et al., 2009b). We instead use a voxel-wise, data-driven approach to produce a multivariate disease signature that may present a stronger signal for the detection of genetic effects (Vounou et al., 2011).

A previous imaging genetic study on the same ADNI cohort measured structural change relative to baseline at a single time point only. In that study an AD-specific phenotype was produced using a sparse linear classifier to select a subset of voxels that minimised the CN/AD classification error (Vounou et al., 2011). In the present study where we incorporate two additional timepoints, we instead begin by fitting a linear regression with an
intercept term, where the dependent variable is the voxel value (change relative to baseline at screening), and the independent variable is time. The regression coefficient for the slope thus gives a summary measure of tissue change over time at each voxel. To obtain a phenotype that is maximally discriminative between CN and AD in our sample, we remove all voxels where the difference in the slopes is not significantly different from zero, by performing an analysis of variance (ANOVA), with sex and age as covariates. Finally we select the most discriminative voxels whose ANOVA p-values exceed a level of 0.05, with a Bonferroni correction for multiple testing. Once again, the use of an ultra-conservative significance threshold ensures that our phenotypic disease signature is maximally discriminative between CN and AD in our sample. The final set of phenotypes used in the study then corresponds to the voxel-wise slope coefficients for all 253 subjects at the selected voxels, corrected for sex and age (see Section 3.5.6).

3.5.3 Genotype data

ADNI Genotypes for the 253 subjects in the present study were extracted from imputed genotype data used in two previous studies (Vounou et al., 2011; Stein et al., 2010a). ADNI genotyping is performed using the Human610-Quad Bead-Chip, which includes 620,901 SNPs and copy number variations (see Saykin et al. (2010) for details). SNPs defining the APOE\(\epsilon4\) variant are not included in the original genotyping chip, but have been genotyped separately by ADNI. These were added to the final genotype dataset. Subjects were unrelated, and all of European ancestry, and had previously passed screening for evidence of population stratification using the procedure described in Stein et al. (2010a). 78,874 non-autosomal SNPs are excluded from the study, as are SNPs with a genotyping rate < 95% (42,680 SNPs excluded), a Hardy-Weinberg equilibrium p-value < 5 \times 10^{-7} (873 SNPs excluded), and a minor allele frequency < 0.1 (64,204 SNPs excluded). 434,271 SNPs remained after all SNP filtering steps described above.

3.5.4 SNP to pathway mapping

For our AD pathways study, we proceed as follows. A list of 21,004 human gene chromosomal locations, corresponding to human genome assembly GRCH36 was obtained using
Ensembl’s BioMart API (www.biomart.org). SNPs were then mapped to any gene within 10k base pairs. This resulted in 211,106 SNPs being mapped to 18,405 genes. While the majority of known genes did map to at least one SNP in our study, approximately half of the SNPs passing QC were not located within 10kbp of a known gene. For pathway mapping, we used the KEGG canonical pathway gene sets obtained from from the Molecular Signatures Database v3.0 (http://www.broadinstitute.org/gsea/msigdb/index.jsp), which contains 186 gene sets, mapping to a total of 5,267 distinct genes, with many genes mapping to more than one pathway. Note that only around 25% of all known genes map to a pathway in this dataset. We map all SNPs within 10kbp of one or more of the 5,267 pathway-mapped genes to the pathway(s) concerned. Finally, we exclude the largest pathway, by number of mapped SNPs, (‘Pathways in Cancer’) that is highly redundant, in that it contains multiple other pathways as subsets. This results in 66,162 SNPs mapped to 4,425 genes and 185 pathways (see Fig. 3.5).

Figure 3.5: Mapping SNPs to pathways
The distribution of pathway sizes in terms of the number of SNPs that they map to is illustrated in Fig. 3.6 (left). Pathway sizes range from 57 to 5,111 SNPs (mean 949). The distribution of overlapping SNPs, that is the number of pathways to which each SNP is mapped, is illustrated in Fig. 3.6 (right). This ranges from 1 to 45 pathways (mean 2.65).

Note that following the above procedure, some genes previously implicated in AD imaging genetic studies do not map to any pathways, and thus are not included in the analysis. For example, in this study, 12 out of 30 genes highlighted in the review by Braskie, Ringman, and Thompson (2011) are mapped to pathways. The remaining 18 genes are excluded because they do not feature in any KEGG pathway. Also note that since SNPs are mapped to all genes within a range of 10kbp, AD implicated SNPs may map to more than one gene, and its corresponding pathway(s). This is the case for example with a number of SNPs mapping to the APOE and TOMM40 genes. This information is summarised in Table 3.4.
Table 3.4: AD genes included in this study. 12 out of 30 genes previously linked to AD imaging endophenotypes (Braskie, Ringman, and Thompson, 2011) that are included in this study are listed in the left hand column. These are genes that (a) map to a KEGG pathway and (b) have a genotyped SNP within 10kbp. The right hand column shows neighbouring genes that map to one or more SNPs mapping to the respective AD implicated gene.

<table>
<thead>
<tr>
<th>Implicated gene</th>
<th>Mapped genes in study</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOMM40</td>
<td>TOMM40 APOE PVRL2</td>
</tr>
<tr>
<td>ACE</td>
<td>ACE</td>
</tr>
<tr>
<td>EPHA4</td>
<td>EPHA4</td>
</tr>
<tr>
<td>CCR2</td>
<td>CCR2 CCR5</td>
</tr>
<tr>
<td>APOE</td>
<td>TOMM40 APOE PVRL2</td>
</tr>
<tr>
<td>FAS</td>
<td>FAS</td>
</tr>
<tr>
<td>CHRNB2</td>
<td>ADAR CHRNB2</td>
</tr>
<tr>
<td>EFNA5</td>
<td>EFNA5</td>
</tr>
<tr>
<td>LDLR</td>
<td>LDLR</td>
</tr>
<tr>
<td>CR1</td>
<td>CR1 CR2</td>
</tr>
<tr>
<td>GRIN2B</td>
<td>GRIN2B</td>
</tr>
<tr>
<td>IL8</td>
<td>IL8</td>
</tr>
</tbody>
</table>

### 3.5.5 Computational Issues

All computer code for the analysis was written in the open source Python programming language, using Numpy and SciPy modules which are optimised for efficient operation with large matrices. Execution of the PsRRR estimation algorithm nonetheless presents a considerable computational burden, both in terms of processor time and memory use. We therefore implement the full range of strategies designed to increase computational efficiency for the GL described earlier, namely a Taylor approximation of the group lasso penalty (Section 2.3.2), efficient computation of block residuals (2.3.4), and the use of a pathway active set (2.3.3). The latter in particular leads to substantial gains in computational efficiency and a large reduction in memory requirements, resulting from the very much reduced size of \( X \) in \( \Omega(a, Y, X, \lambda) \) (Box 3.1).

The need to fit a large number of PsRRR models over multiple subsamples of the data for pathway ranking presents another major drain on computational resources. However, the fact that model estimations for each subsample are entirely independent presents an opportunity for performing multiple model fits in parallel. We implement such a strategy using a computer cluster, in which a single client node distributes subsamples across mul-
3.5 PsRRR Application study: Gene pathways implicated in Alzheimer’s disease

Multiple CPU cores (see Figure 3.7). Parallel computations and client-server communication are implemented in Parallel Python (http://www.parallelpython.com/). All high-bandwidth processor and memory intensive operations are carried out on the server nodes. Only low-bandwidth instructions and results pass between client and servers. The computational framework is highly scalable, so that the user can specify the number of server nodes and CPUs, with the software determining how model estimations are distributed between these. The resulting reduction in computation time due to parallelisation is considerable. For example, in the AD study described here, total execution time (excluding weight tuning) with $B = 1000$ subsamples was $6\frac{1}{2}$ hours, whereas total execution time if each job were run serially would be approximately $10\frac{1}{2}$ days.

![Figure 3.7: Parallel computing framework for PsRRR estimation. Coefficient estimation for multiple data subsamples is carried out by CPUs (‘workers’) distributed across a computing cluster. Computing jobs are distributed by a single python process residing on the server node. All workers reside on one of 8 client nodes. High-bandwidth, data intensive operations are carried out within each client node. Only low-bandwidth instructions and results pipes pass between server and client.](image-url)
3.5.6 Results

*AD associated phenotypes*

An imaging signature characteristic of AD was created using the procedure described in section 3.5.2. As described previously, we begin by computing a linear least-squares fit of the longitudinal structural change across 3 time points at each voxel. An illustration of average slope coefficients, and their variation between subjects, is shown in Fig. 3.8. Increased expansion of ventricular volumes is clear in all subjects, but this increase is most marked in AD patients, where ventricular volumes expand by an average 1.2% per year (white regions in left hand part of Fig. 3.8). AD patients also show the most variation in structural change over time.

![Figure 3.8: Sample mean (left) and standard deviation (right) of slope coefficients for the 2 subject groups. Slope coefficients represent a linear approximation of change in brain volume over time. Scales represent $10\times$ percentage change in voxel volume per year, so that for example a slope coefficient of 12 (white areas in left hand plot) is equivalent to an average yearly increase in voxel volume of 1.2%.](image)

A statistical image showing the corresponding ANOVA p-values, a measure of the extent to which each voxel is able to discriminate between ADs and CNs, is shown in the top row of Fig. 3.9. From the $Q^* = 2,153,231$ voxels in this image, we extract a final set of $Q = 148,023$ voxels whose p-values exceed a Bonferroni-corrected threshold of $0.05/Q^*$. This final set of voxels that are most discriminative between ADs and CNs, are highlighted in yellow in the bottom row of Fig. 3.9. These $Q$ voxels constitute the phenotype for each
3.5 PsRRR Application study: Gene pathways implicated in Alzheimer’s disease

Figure 3.9: Imaging signature characteristic of AD. Top: Statistical image showing p-values ($-\log_{10}$ scale) obtained from an ANOVA on the linear structural change over 3 time points, corrected for age and sex, to discriminate between AD and CN subjects. Bottom: The final set of $Q = 148,023$ selected voxels with p-values exceeding a Bonferroni-corrected threshold $\alpha_B = 0.05/2153231, (-\log_{10} \alpha_B = 7.6)$ are highlighted in yellow.
subject used in the study. We provide a further indication of the discriminatory power of the selected voxels by visualising the Euclidean distances between subjects using the selected voxels in a 3D multi-dimensional scaling plot in Fig. 3.10. The relatively small overlap between CD and AD subjects indicates that our chosen disease signature is indeed discriminative between the two groups. As expected we also see evidence of greater variability in the AD group, compared with CN.

![Figure 3.10: 3D multi-dimensional scaling plot illustrating the spread of imaging signatures across ADs and CNs. Imaging signatures correspond to selected voxels only.](image)

**Pathway, SNP and gene rankings**

We use the PsRRR algorithm described in section 3.1 to identify KEGG pathways associated with the AD-discriminative longitudinal phenotypes described in the preceding section. Pathways are ranked in order of importance using the resampling strategy described in section 3.3, with $B = 1000$ subsamples. Each subsample is balanced to maintain a constant AD:CN ratio. We use $\lambda = 0.8\lambda_{\text{max}}$, which results in the selection of an average of 7 pathways at each subsample (min 1, max 15, SD = 2.3). Pathway ranking results are presented in Table 3.5.

SNPs and genes are ranked using sRRR with a lasso penalty on the SNP coefficient vector, as described in section 3.3. Lasso selection is performed on pathways selected at each subsample in the pathways analysis described above, so that once again $B = 1000$. The number of SNPs, $Z^{(b)}$, included in the lasso model at subsample $b$ varies according to
### Table 3.5: Top 30 pathways, ranked by pathway selection frequency.

<table>
<thead>
<tr>
<th>Rank</th>
<th>KEGG pathway name</th>
<th>$\pi_{path}$</th>
<th>Size (# SNPs)</th>
<th>Lasso selected genes in pathway</th>
<th>Known AD genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Insulin signaling pathway</td>
<td>0.524</td>
<td>1517</td>
<td>HK2 PIK3R3 PIK3CG ACACA G6PC</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Vascular smooth muscle contraction</td>
<td>0.456</td>
<td>3236</td>
<td>PRKCB ADCYS ADCY2 PIKCA MYLK PLCB1</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Melanogenesis</td>
<td>0.331</td>
<td>1638</td>
<td>PRKCB ADCYS ADCY2 PRKCA GNAI1 WNT2 PLCB1</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Focal adhesion</td>
<td>0.232</td>
<td>4009</td>
<td>PRKCB PRKCA PIK3R3 MYLK PIK3CG COLS3 RELN ACTN1</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Gap junction</td>
<td>0.180</td>
<td>2350</td>
<td>PRKCB ADCYS ADCY2 PRKCA GNAI1 PLCB1</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Huntingtons disease</td>
<td>0.155</td>
<td>1980</td>
<td>PLCB1 DNAI2 UQCHR</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Purine metabolism</td>
<td>0.154</td>
<td>2096</td>
<td>ADCY8 ADCY2 ALLC</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Pyruvate metabolism</td>
<td>0.153</td>
<td>456</td>
<td>ACACA</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Propanoate metabolism</td>
<td>0.152</td>
<td>471</td>
<td>ACACA</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Amyotrophic lateral sclerosis als</td>
<td>0.151</td>
<td>865</td>
<td>TOMM40</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Chemokine signaling pathway</td>
<td>0.145</td>
<td>2769</td>
<td>PRKCB ADCYS ADCY2 PIK3R3 PIK3CG GNAI1 PLCB1 XCLI ITK GNG2 GRK5 CCR2 IL8</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Phosphatidylinositol signaling system</td>
<td>0.138</td>
<td>2067</td>
<td>PRKCB PRKCA PIK3R3 PIK3CG DGKA DGKB PLCB1 DGKI</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Citrate cycle tca cyclc</td>
<td>0.137</td>
<td>210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Glycosphingolipid biosynthesis globo series</td>
<td>0.135</td>
<td>227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>Alzheimers disease</td>
<td>0.127</td>
<td>2500</td>
<td>PLCB1 APOE UQCHR</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>Complement and coagulation cascades</td>
<td>0.119</td>
<td>783</td>
<td>CR1</td>
<td>APOE FAS GRIN2B</td>
</tr>
<tr>
<td>17.</td>
<td>Steroid biosynthesis</td>
<td>0.113</td>
<td>153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Jak stat signaling pathway</td>
<td>0.106</td>
<td>1311</td>
<td>PIK3R3 PIK3CG</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>Ecm receptor interaction</td>
<td>0.104</td>
<td>1969</td>
<td>COLS3 RELN</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>Tight junction</td>
<td>0.103</td>
<td>3332</td>
<td>PRKCB PRKCA GNAI1 ACTN1 YES1</td>
<td></td>
</tr>
<tr>
<td>21.</td>
<td>Glycerolipid metabolism</td>
<td>0.102</td>
<td>877</td>
<td>DGKA DGKB DGK1</td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>Calcium signaling pathway</td>
<td>0.096</td>
<td>5111</td>
<td>PRKCB ADCYS ADCY2 PRKCA MYLK PLCB1</td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>Toll like receptor signaling pathway</td>
<td>0.096</td>
<td>712</td>
<td>PIK3R3 PIK3CG</td>
<td>IL8</td>
</tr>
<tr>
<td>24.</td>
<td>Leishmania infection</td>
<td>0.090</td>
<td>620</td>
<td>PRKCB CR1</td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>Lysosome</td>
<td>0.089</td>
<td>1111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td>Fc gamma r mediated phagocytosis</td>
<td>0.080</td>
<td>1976</td>
<td>PRKCB PRKCA PIK3R3 PIK3CG</td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>Neurotrophin signaling pathway</td>
<td>0.075</td>
<td>1689</td>
<td>PIK3R3 PIK3CG</td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td>Glycophospholipid metabolism</td>
<td>0.071</td>
<td>1047</td>
<td>DGKA DGKB DGK1</td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td>Renal cell carcinoma</td>
<td>0.071</td>
<td>840</td>
<td>PIK3R3 PIK3CG</td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>Wnt signaling pathway</td>
<td>0.070</td>
<td>2023</td>
<td>PRKCB PRKCA WNT2 PLCB1</td>
<td></td>
</tr>
</tbody>
</table>

1Top 30 ranked genes in this pathway, using lasso selection (see Table 3.6). 2Previously identified AD genes in the pathway (see Table 3.4).
the number and size (in terms of the number of mapped SNPs) of selected pathways. $Z^{(b)}$ ranges from a minimum of 227, to a maximum of 19,642 (mean = 8400; SD = 3000). As with pathway ranking, we use $\lambda = 0.8\lambda_{\text{max}}$, which results in the selection of an average of 11.5 SNPs at each subsample (min 1, max 56, SD = 11.7). SNP and gene ranking results are presented in Table 3.6.

We first consider the pathway ranking results in Table 3.5. Under the null, where there is no association between phenotypes and genotypes, and with a single pathway selected by the model at each subsample, the expected pathway selection frequency distribution is uniform, with, $\pi_{\text{path}}^l = 1/185 \approx 0.005$. With an average of 7 pathways selected at each subsample, as is the case here, and assuming pathways are independent, the corresponding pathway selection frequency distribution under the null is also uniform, with, $\pi_{\text{path}}^l = 7/185 \approx 0.038$. However, where more than one pathway is selected by the model, pathway selection probabilities will not be uniform, since the presence of overlapping SNPs means that pathways are not independent. Instead, selection probabilities will reflect the pattern of overlaps corresponding to the distribution of causal SNPs (or spurious associations under the null). Indeed, even when relatively few SNPs or genes are associated with the phenotype, we can expect multiple pathways to harbour genetic effects since many SNPs and genes overlap multiple pathways. For this reason the figure of 0.038 should be seen only as a guide threshold to signify pathway importance, and while we report pathway selection frequencies, $\pi_{\text{path}}^l$, our main focus is on pathway rankings. To aid interpretation of pathway rankings, for each pathway we list those genes in the pathway that are ranked in the top 30 genes, selected by lasso selection (see Table 3.6).

In the final column of Table 3.5 we list genes in the top ranked pathways that have previously been linked to AD imaging phenotypes in the review by Braskie, Ringman, and Thompson (2011). Both the number of such genes affecting phenotypes in this study, and the extent to which these genes may drive pathway selection are unknown. It is nevertheless interesting to consider whether these genes are significantly enriched amongst high-ranking pathways. To do this we calculate an average ranking for each ‘AD gene’ by taking the average rank achieved by all pathways containing the gene in question. We then derive an AD gene enrichment score by summing average AD gene ranks across all AD genes. A lower score thus indicates pathways containing AD genes tend to be ranked high. We com-
### 3.5 PsRRR Application study: Gene pathways implicated in Alzheimer’s disease

Table 3.6: Top 30 SNPs and genes, respectively ranked by SNP and gene selection frequency, using lasso sRRR. Note the APOE gene is selected at a lower frequency than the APOϵ4 SNP, since the allele is often selected in a pathway where it is mapped to the TOMM40 gene only.

<table>
<thead>
<tr>
<th>Rank</th>
<th>SNP</th>
<th>$\pi^SNP$</th>
<th>Mapped gene(s)</th>
<th>Gene</th>
<th>$\pi^{gene}$</th>
<th># mapped SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs4788426</td>
<td>0.451</td>
<td>PRKCB</td>
<td>PRKCB</td>
<td>0.451</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>rs11074601</td>
<td>0.429</td>
<td>PRKCB</td>
<td>ADCY8</td>
<td>0.411</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>rs263264</td>
<td>0.411</td>
<td>ADCY8</td>
<td>ADCY2</td>
<td>0.392</td>
<td>106</td>
</tr>
<tr>
<td>4</td>
<td>rs13189711</td>
<td>0.392</td>
<td>ADCY2</td>
<td>HK2</td>
<td>0.302</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>rs680545</td>
<td>0.302</td>
<td>HK2</td>
<td>PRKCA</td>
<td>0.290</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>rs4622543</td>
<td>0.290</td>
<td>PRKCA</td>
<td>PIK3R3</td>
<td>0.267</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>rs9896483</td>
<td>0.274</td>
<td>PRKCA</td>
<td>MYLK</td>
<td>0.234</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>rs1052610</td>
<td>0.267</td>
<td>PIK3R3</td>
<td>PIK3CG</td>
<td>0.207</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>APOE</td>
<td>0.251</td>
<td>TOMM40 APOE</td>
<td>COL5A3</td>
<td>0.174</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>rs1254403</td>
<td>0.234</td>
<td>MYLK</td>
<td>GNAI1</td>
<td>0.167</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td>rs4730205</td>
<td>0.207</td>
<td>PIK3CG</td>
<td>ACACA</td>
<td>0.164</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td>rs889130</td>
<td>0.174</td>
<td>COL5A3</td>
<td>G6PC</td>
<td>0.163</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>rs6973616</td>
<td>0.167</td>
<td>GNAI1</td>
<td>DGKA</td>
<td>0.160</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>rs9906543</td>
<td>0.164</td>
<td>ACACA</td>
<td>CR1</td>
<td>0.154</td>
<td>21</td>
</tr>
<tr>
<td>15</td>
<td>rs2229611</td>
<td>0.163</td>
<td>G6PC</td>
<td>TOMM40</td>
<td>0.152</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>rs10876862</td>
<td>0.160</td>
<td>DGKA</td>
<td>WNT2</td>
<td>0.137</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>rs772700</td>
<td>0.160</td>
<td>DGKA</td>
<td>DGKB</td>
<td>0.131</td>
<td>200</td>
</tr>
<tr>
<td>18</td>
<td>rs12734030</td>
<td>0.154</td>
<td>CR1</td>
<td>PLCB1</td>
<td>0.128</td>
<td>218</td>
</tr>
<tr>
<td>19</td>
<td>rs11117959</td>
<td>0.154</td>
<td>CR1</td>
<td>APOE</td>
<td>0.127</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>rs650877</td>
<td>0.154</td>
<td>CR1</td>
<td>RELN</td>
<td>0.117</td>
<td>160</td>
</tr>
<tr>
<td>21</td>
<td>rs11118131</td>
<td>0.154</td>
<td>CR1</td>
<td>DGKI</td>
<td>0.112</td>
<td>49</td>
</tr>
<tr>
<td>22</td>
<td>rs6691117</td>
<td>0.142</td>
<td>CR1</td>
<td>ACTN1</td>
<td>0.110</td>
<td>41</td>
</tr>
<tr>
<td>23</td>
<td>rs677066</td>
<td>0.142</td>
<td>CR1</td>
<td>ALLC</td>
<td>0.108</td>
<td>18</td>
</tr>
<tr>
<td>24</td>
<td>rs2239956</td>
<td>0.137</td>
<td>WNT2</td>
<td>XCL1</td>
<td>0.086</td>
<td>7</td>
</tr>
<tr>
<td>25</td>
<td>rs4719392</td>
<td>0.131</td>
<td>DGKB</td>
<td>ITK</td>
<td>0.084</td>
<td>27</td>
</tr>
<tr>
<td>26</td>
<td>rs6077420</td>
<td>0.128</td>
<td>PLCB1</td>
<td>DNAI2</td>
<td>0.077</td>
<td>16</td>
</tr>
<tr>
<td>27</td>
<td>rs7777178</td>
<td>0.126</td>
<td>DGKB</td>
<td>GNG2</td>
<td>0.076</td>
<td>31</td>
</tr>
<tr>
<td>28</td>
<td>rs12699607</td>
<td>0.122</td>
<td>DGKB</td>
<td>GRK5</td>
<td>0.074</td>
<td>56</td>
</tr>
<tr>
<td>29</td>
<td>rs7796440</td>
<td>0.122</td>
<td>DGKB</td>
<td>UQCRH</td>
<td>0.071</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>rs1872837</td>
<td>0.120</td>
<td>HK2</td>
<td>YES1</td>
<td>0.068</td>
<td>11</td>
</tr>
</tbody>
</table>
pare this empirically derived score with the distribution of scores obtained by permuting pathway rankings 100,000 times. The null distribution of this enrichment score (obtained by permutation), and the empirically observed value are compared in Fig. 3.11. Finally, we compute a p-value for the null hypothesis that the empirically observed enrichment score has arisen by chance, as the proportion of enrichment scores obtained through permutation that are lower than the observed value. This gives a value $p = 0.0051$, indicating that AD genes are highly over-represented amongst top ranking pathways, compared to what would be expected by chance.

![Graph showing distribution of enrichment scores](image-url)

Figure 3.11: Measure of extent to which genes previously linked to AD are enriched in highly-ranked pathways. The histogram shows the distribution of AD gene enrichment scores obtained when permuting pathway rankings 100,000 times. The vertical black line indicates the observed AD gene enrichment score using the true pathway rankings obtained in the study. From this we derive a p-value indicating the probability that the empirical AD gene enrichment score could arise by chance as $p = 0.0051$. AD-linked genes are those identified in Braskie, Ringman, and Thompson (2011).

### 3.6 Discussion

We have described a method for the identification of gene pathways associated with a multivariate quantitative trait (MQT). Here, we extended previous work modelling a uni-
3.6 Discussion

variate response (Chapter 2), where we showed that a multilocus, group-sparse modelling approach can demonstrate increased power to detect causal pathways, when compared to conventional approaches that begin by modelling individual SNP-phenotype associations. We applied our method in an AD gene pathways study using imaging endophenotypes, but our method is not restricted to the case of biological pathways or imaging phenotypes, and can be applied to any data in which we seek to identify sparse groups of predictors affecting a multivariate response.

To the best of our knowledge, few other multilocus methods for the identification of biological pathways currently exist, and of those that do, none are able to accommodate a multivariate phenotype. While a methodological study comparing the various approaches would be interesting, as has been noted by others, a lack of benchmark datasets with validated pathways makes comparison between methods difficult (Chen et al., 2010; Khatri, Sirota, and Butte, 2012). The GRASS method (Chen et al., 2010) and the method proposed by Zhao et al. (2011) use sparse regression techniques to measure pathway significance. These methods are currently implemented for case-control data, and univariate phenotypes only. Each method makes different assumptions about the distribution of important SNPs and genes affecting the phenotype. GRASS assumes sparsity at the SNP level within each pathway gene, and retains all genes in each pathway considered. Zhao’s method assumes sparsity at the gene level. In contrast, our PsRRR method assumes sparsity only at the pathway level (although we subsequently perform SNP and gene selection as a second step in selected pathways), and considers all pathways together in a single model. As such, each method is expected to perform differently, depending on the true distribution of causal SNPs and genes. GRASS and Zhao’s methods also use a pre-processing dimensionality reduction step on SNPs within each gene using PCA. While this has been shown to be advantageous in certain circumstances (Wang and Abbott, 2008), we elect to retain original SNP genotypes in our model, since this facilitates sparse SNP selection. A further feature that distinguishes our method from GRASS and Zhao’s method is that we include all pathways together in a single regression model. By doing this we hope to gain a better measure of the relative importance of different pathways by ensuring they compete against each other.

In any method modelling effects on an MQT, the use of a multivariate disease signature
that is characteristic of the disease under investigation is important. This is especially so in the case of high-dimensional imaging phenotypes, where a poorly characterised imaging signature with low signal to noise ratio may show no advantage over a simple ROI average-based approach (Vounou et al., 2011). In our application study we extract an AD imaging phenotype that is highly discriminative of subjects with the disease, compared to controls, by excluding voxels at which the fitted slopes, measuring structural change over 3 time points, are not significantly different between the two groups. The subsequent pathway and gene mapping stages will clearly depend on the particular choice of phenotype, so that a different phenotype may well highlight different genetic effects. An analysis of the sensitivity of our gene mapping procedure to the choice of phenotype is however beyond the scope of the present study.

We use a resampling strategy to rank pathways by selection frequency across multiple $\frac{N}{2}$ subsamples of the data. Similar strategies have been advocated for improving consistency with the lasso (Bach, 2008a; Chatterjee and Lahiri, 2011), and here this strategy is designed to provide a robust measure of the relative importance of individual pathways in a finite sample. In some respects our approach resembles the ‘pointwise stability selection’ strategy proposed by Meinshausen and Bühlmann (2010). For the latter, a theoretical bound for determining a selection frequency threshold that controls the expected number of false positives has been derived. However, this rests on an exchangeability assumption regarding the selection of noise variables that does not apply here, since the variables under selection are groups of variables (pathways) that are functionally related, and overlap in terms of the genes that they contain.

In principle our method enables the voxel-wise mapping of pathway effects across the brain, through the analysis of the phenotype coefficient vector $a$. In Section 3.4.3 we demonstrated in a simple simulation study that our method is able to identify important voxels, through the imposition of an additional lasso penalty on $a$ to enforce sparse voxel selection. Alternative sparse penalties such as those associated with the elastic net (Carroll et al., 2009) or indeed GL may be particularly appropriate here, as they could account for spatial or other structural features present in the imaging domain, although both would require the tuning of addition regularisation parameters. Although we do not perform voxel selection in our AD application study, we note that this would make an interesting exten-
Discussion

Of the top-ranking pathways identified in our AD application study (see Table 3.5), functions associated with many of the top 10 ranked pathways have been linked to aspects of AD biology described in the literature. Beginning with the top 2 ranked pathways, numerous studies suggest links between disruption to the insulin signalling pathway and AD (Liu et al., 2011; Liao and Xu, 2009; Monte and Wands, 2005; Biessels and Kappelle, 2005; Steen et al., 2005), and to the role of vascular smooth muscle dysfunction in AD-associated neurodegeneration (Zlokovic, 2011). Other functions previously associated with AD biology among high-ranking pathways include those related to focal adhesion, gap junctions, chemokine signalling and phosphatidylinositol signalling (Caltagarone, Jing, and Bowser, 2007; Nakase and Naus, 2004; Xia and Hyman, 1999; Kim et al., 2003; Huber, Egleon, and Davis, 2001; Ravetti et al., 2010). Ideally we would like to validate these results using an independent dataset. However, at the time of writing no other datasets with similar endophenotypes were available.

In order to better elucidate which genes may be driving pathway selection in our AD application study, we performed a follow up analysis designed to identify SNPs and genes in selected pathways that are separately associated with the phenotype (see Table 3.6). These rankings are derived from lasso selection of SNPs within all selected pathways at each subsample, irrespective of their groupings within pathways. They are therefore expected to capture larger, independent signals of association, and not necessarily all the salient signals within a particular pathway that may be driving pathway selection. In particular, the group lasso is designed to detect distributed signals that are unlikely to coincide with those captured using lasso selection. From this analysis, it is clear that the lipid kinase genes *PIK3R3/PIK3CG*, and the calcium-activated, phospholipid-dependent genes *PRKCA/PRKCB* play a role in driving selection of many pathways in the top 30 ranks. All these genes have previously been linked in gene expression studies with β-amyloid plaque formation in the AD brain (Liang et al., 2008). Aside from the previously validated AD endophenotype-related genes *TOMM40, CR1* and *APOE* (Shen et al., 2010; Lambert et al., 2009; Biffi et al., 2010), other genes occurring in the top 10 ranking pathways, include *ADCY2, ACTN1, ACACA* and *GNAII*, all of which have been associated with AD related...
changes in hippocampal gene expression (Taguchi et al., 2005; Ravetti et al., 2010, supporting information). Along with APOE and TOMM40, ADCY2 was also highlighted in a previous study searching for SNPs associated with AD-associated structural change (Vounou et al., 2011). This latter study was on the same ADNI cohort, but unlike the current study was not pathways-driven, and used phenotypes describing structural change measured at a single time point (relative to baseline) only.

The major AD risk and phenotype-related gene APOE, and risk allele APOEɛ4 are respectively ranked 19 and 9. In our study the APOE gene maps to a single pathway, the KEGG Alzheimer’s disease pathway, and this pathway is selected in \( \approx 13\% \) of subsamples. Notably, in all subsamples in which the KEGG Alzheimer’s disease pathway is selected, the APOEɛ4 allele is the sole selected SNP, confirming the known large marginal effect of this allele on AD phenotypes. The higher ranking of the APOEɛ4 SNP, relative to the APOE gene, reflects the fact that this SNP also maps to the TOMM40 gene, which occurs in a number of other pathways selected by the model. This may affect the Alzheimer pathway’s ranking, as may the fact that selection of this pathway is driven by the presence of this single, strong APOEɛ4 signal, and as explained above, the model is designed to identify distributed signals across a pathway.

Our model rests on a number of assumptions, and as a consequence will fail to detect a number of different association signals. For example, while our model implicitly accommodates the fact that SNPs and genes interact within functional pathways, we do not explicitly model interaction effects. Also, we make the simplifying assumption that voxel-wise measures are uncorrelated. In reality, the phenotype will exhibit a complex correlation structure which will affect the association signal. Vounou, Nichols, and Montana (2010) have demonstrated that even under this simplifying assumption, significant gains in power can be achieved by modelling a multivariate phenotype, compared to a mass univariate modelling approach. Finally, our model is founded on the assumption that causal SNPs tend to accumulate within functional pathways, and as such is not designed to identify significant marginal effects, as evidenced for example by the failure of our model to rank the high-risk APOE gene highly in our AD application study. For this last reason, any pathways analysis should be seen as being complementary to conventional GWAS approaches.
Chapter 4

Pathways-driven SNP selection: ‘Pathways Sparse Group Lasso with Adaptive Weights’

We now consider an extension to the previous P-GLAW model (Chapter 2) that uses pathways information to drive the identification of SNPs associated with a univariate quantitative trait. The motivation here is that the use of prior information on putative gene interactions within pathways will increase power to identify causal SNPs, compared to alternative methods that disregard such information. The GL is not well-suited to this task, since it performs non-sparse, continuous shrinkage of SNP coefficients within a selected pathway. These SNP coefficients are themselves an unreliable measure of SNP importance, in part because of LD, and also because the majority of SNPs within a selected pathway are unlikely to be associated with the phenotype under consideration. We instead seek a model with the following properties:

- sparsity at the pathway level, so that only pathways containing multiple SNPs with possibly small or moderate marginal effects are selected by the model, and
- sparsity at the SNP level within selected pathways, so that only those SNPs driving pathway selection are selected by the model
A suitable sparse regression model enforcing the required ‘dual-level’ sparsity pattern is the sparse group lasso (SGL). SGL is a comparatively recent development in sparse modelling, and in simulations has been shown to accurately recover dual-level sparsity, in comparison to both GL and lasso (Friedman, Hastie, and Tibshirani, 2010; Simon et al., 2012). SGL has been used for the identification of rare variants in a case-control study by grouping SNPs into genes (Zhou et al., 2010); for the identification of genomic regions whose copy number variations have an impact on RNA expression levels (Peng et al., 2010); and to model geographical factors driving climate change (Chatterjee et al., 2011).

A schematic illustration of the two types of sparsity pattern generated by the GL and SGL is given in Figure 4.1. As with GL, SGL can be seen as fitting into a wider class of structured-sparsity inducing models that use prior information on relationships between predictors to enforce different sparsity patterns (Zhao, Rocha, and Yu, 2009; Huang, Zhang, and Metaxas, 2011; Jenatton and Bach, 2011).

![Figure 4.1: Sparsity patterns enforced by the group lasso and sparse group lasso. The set $S \subset \{1, \ldots, P\}$ of causal SNPs influencing the phenotype are represented by boxes that are shaded grey. Causal SNPs are assumed to occur within a set $C \subset \{1, \ldots, L\}$ of causal pathways, $G_1, \ldots, G_L$. Here $C = \{2, 3\}$. The group lasso enforces sparsity at the group or pathway level only, whereas the sparse group lasso additionally enforces sparsity at the SNP level.

In passing, we note that in principle we could attempt to identify important SNPs by conducting a two-stage analysis, in which we first identify important pathways, and then in a second step search for SNPs within selected pathways (Eleftherohorinou et al., 2009; Eleftherohorinou et al., 2011). This is the strategy we adopted in the AD study described in the previous chapter, where we perform a second round of lasso selection after the initial pathway selection step. There are however a number of problems with this approach.
Firstly, as discussed previously, highlighted SNPs are then not necessarily those that were driving pathway selection in the first step of the analysis. Secondly, an implicit (and reasonable) assumption is that only a small number of SNPs are driving pathway selection, so that a sparse regression model that assumes this is preferable. Finally, a model imposing dual-level sparsity is able to perform simultaneous pathway and SNP selection, and is thus simpler to implement.

This chapter is organised as follows. In Section 4.1 we introduce the SGL model, and in Section 4.1.1 we describe an efficient estimation algorithm, using block coordinate gradient descent (BCGD), for the case of non-overlapping groups. In Section 4.1.2, we describe a simulation study illustrating superior group (pathway) and variable (SNP) selection performance in the case that the true supporting model is group-sparse. In Section 4.2 we extend the previous model to the case of overlapping groups, using the same procedure of duplicating overlapping variables as we described previously for GL. In principle, we can then solve this model using the BCGD estimation algorithm described for the non-overlapping case. However, we argue in Section 4.2.1 that this approach does not give us the outcome we require. For this reason we describe a modified estimation algorithm using coordinate gradient descent, that treats each group as independent, and in Section 4.2.2 we demonstrate in a simulation study that this new algorithm is able to identify the correct SNPs and pathways with improved sensitivity and specificity. In Sections 4.2.3 and 4.2.4 we explain how the weight tuning and pathway ranking strategies previously described for our GL-based models can be applied to SGL. We call the resulting method for pathways-driven SNP selection ‘Pathways-Sparse Group Lasso with Adaptive Weights’ or P-SGLAW. In Section 4.3 we apply the P-SGLAW method in a study looking at pathways, SNPs and genes associated with high-density lipoprotein cholesterol levels in two separate cohorts of Asian adults. We conclude this chapter with a discussion in Section 4.4.

### 4.1 The sparse group lasso model

We begin by considering the case of non-overlapping groups. For a univariate response vector, $y$, with the SGL (Simon et al., 2012), sparse estimates for the SNP coefficient
vector, $\beta$ are given by

$$\hat{\beta}^{SGL} = \arg \min_{\beta} \left\{ \frac{1}{2} ||y - X\beta||^2_2 + (1 - \alpha)\lambda \sum_{l=1}^{L} w_l ||\beta_l||_2 + \alpha \lambda ||\beta||_1 \right\} \quad (4.1)$$

with $|\alpha| \leq 1$. Note that this is equivalent to the group lasso estimator (2.15), with the addition of an extra $\ell_1$ penalty. When $\alpha = 0$, (4.1) reduces to the group lasso. With $\alpha > 0$, the additional penalty on the $l_1$-norm of $\beta$ encourages sparsity at the SNP level. $\alpha$ then controls the distribution of the regularisation between groups (pathways) and SNPs, such that as $\alpha$ approaches 0 from above, greater sparsity at the group level is encouraged over sparsity at the SNP level.

### 4.1.1 Model estimation

For the estimation of $\beta$ we proceed, as with the group lasso, by noting that the optimisation (4.1) is convex, and (in the case of non-overlapping groups) that the penalty is block-separable, so that we can obtain a solution using block, or group-wise coordinate descent (Tseng and Yun, 2009). For a single group, $l$, the minimising function corresponding to (4.1) is given by

$$f(\beta_l) = \frac{1}{2} ||y - X\beta||^2_2 + (1 - \alpha)\lambda w_l ||\beta_l||_2 + \alpha \lambda ||\beta_l||_1. \quad (4.2)$$

An optimal solution for SNP coefficient $\beta_j$ is then derived from the subgradient equations

$$-X_j^T(\hat{r}_l - \sum_{k\neq j} X_k \hat{\beta}_k - X_j \hat{\beta}_j) + (1 - \alpha)\lambda w_l s_j + \alpha \lambda t_j = 0 \quad j = l_1, \ldots, l_P, \quad (4.3)$$

where as before (Section 2.3.1) $\hat{\beta}_k, k \neq j$ are the current estimates for other SNP coefficients in group $l$, and the group partial residual, $\hat{r}_l = y - \sum_{m \neq l} X_m \hat{\beta}_m$. Here $s_j$ and $t_j$ are
4.1 The sparse group lasso model

the respective subgradients of $||\beta_l||_2$ and $|\beta_j|$, with

$$s_j = \begin{cases} \frac{\beta_j}{||\beta_l||_2} & \text{if } ||\beta_l||_2 \neq 0 \\ \in [-1,1] & \text{if } ||\beta_l||_2 = 0 \end{cases}$$

$$t_j = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \in [-1,1] & \text{if } \beta_j = 0. \end{cases} \quad (4.4)$$

If $\beta_l = 0$, that is group $l$ is not selected by the model, then from (4.3)

$$-X'_j\hat{r}_l + (1 - \alpha)\lambda w_l s_j + \alpha \lambda t_j = 0, \quad j = l_1, \ldots, l_P. \quad (4.5)$$

Substituting $a = X'_l\hat{r}_l$ gives

$$a_j = (1 - \alpha)\lambda w_l s_j + \alpha \lambda t_j, \quad j = l_1, \ldots, l_P$$

so that

$$s_j^2 = \frac{1}{(1 - \alpha)^2 \lambda^2 w_l^2} (a_j - \alpha \lambda t_j)^2, \quad j = l_1, \ldots, l_P,$$

and

$$\sum_j s_j^2 = \frac{1}{(1 - \alpha)^2 \lambda^2 w_l^2} \sum_j (a_j - \alpha \lambda t_j)^2.$$  

From (4.4), when $\beta_l = 0$, $||s||_2 = (\sum_j s_j^2)^{\frac{1}{2}} \leq 1$, so that

$$\sum_j (a_j - \alpha \lambda t_j)^2 \leq (1 - \alpha)^2 \lambda^2 w_l^2. \quad (4.6)$$

Also from (4.4), one further condition when $\beta_l = 0$ is that $t_j \in [-1,1]$. The values, $\hat{t}_j$ that minimise the left hand size of (4.6) are therefore given by

$$\hat{t}_j = \begin{cases} \frac{a_j}{\alpha \lambda} & \text{if } |\frac{a_j}{\alpha \lambda}| \leq 1 \\ \text{sign}(\frac{a_j}{\alpha \lambda}) & \text{if } |\frac{a_j}{\alpha \lambda}| > 1. \end{cases}$$
Substituting for $a_j$, we can then write the values for $a_j - \alpha \lambda t_j$ that minimise the left hand side of (4.6) as

$$a_j - \alpha \lambda t_j = \begin{cases} 0 & \text{if } |X_j' \hat{r}_l| \leq \alpha \lambda \\ \text{sign}(X_j' \hat{r}_l)(|X_j' \hat{r}_l| - \alpha \lambda) & \text{if } |X_j' \hat{r}_l| > \alpha \lambda \end{cases}$$

$$= S(X_j' \hat{r}_l, \alpha \lambda)$$

for $j = l_1, \ldots, l_P$, where

$$S(X_j' \hat{r}_l, \alpha \lambda) = \text{sign}(X_j' \hat{r}_l)(|X_j' \hat{r}_l| - \alpha \lambda)_+$$

(4.7)
is the usual lasso soft thresholding operator (2.10). Finally, we can now rewrite the condition for $\hat{\beta}_l = 0$, (4.6) as

$$||S(X_l' \hat{r}_l, \alpha \lambda)||_2 \leq (1 - \alpha) \lambda w_l,$$

(4.8)

where the vector $S(X_l' \hat{r}_l, \alpha \lambda) = [S(X_{l_1}' \hat{r}_l, \alpha \lambda), \ldots, S(X_{l_P}' \hat{r}_l, \alpha \lambda)]$. Note that with $\alpha = 0$, this reduces to the GL group selection criterion (2.21).

In the case that $\beta_l \neq 0$, that is group $l$ is selected by the model, from (4.3) and (4.4) we see that $\beta_j = 0$ when

$$-X_j'(\hat{r}_l - \sum_{k \neq j} X_k \hat{\beta}_k) \leq |\alpha \lambda|.$$  

(4.9)

For completeness, we rewrite the criterion for selecting pathway $l$ from (4.8) as

$$||S(X_l' \hat{r}_l, \alpha \lambda)||_2 > (1 - \alpha) \lambda w_l$$

(4.10)

and the criterion for selecting SNP $j$ in selected pathway $l$ from (4.9) as

$$|X_j' \hat{r}_{l,j}| > \alpha \lambda$$

(4.11)

where $\hat{r}_{l,j} = \hat{r}_l - \sum_{k \neq j} X_k \hat{\beta}_k$ is the SNP partial residual, obtained by regressing out the current estimated effects of all other predictors in the model, except for predictor $j$.

A number of methods for the estimation of $\beta_l$ in the case that $||\beta_l||_2 \neq 0$ have been
4.1 The sparse group lasso model

proposed (Friedman, Hastie, and Tibshirani, 2010; Foygel and Drton, 2010; Liu and Ye, 2010; Simon et al., 2012). A complicating factor is the discontinuities in the first (and second) derivatives of \( s_j \) at \( ||\beta_l||_2 = 0 \), that is where \( ||\beta_l||_2 \) first moves away from zero, and of \( t_j \) when \( \beta_j = 0 \). As with GL, Friedman, Hastie, and Tibshirani (2010) describe a numerical method using coordinate descent, by combining a golden search over \( \beta_j \) with parabolic interpolation. However we find this too computationally intensive for the large datasets we wish to analyse. Simon et al. (2012) propose an accelerated, block gradient descent method in which \( \beta_l \) is iteratively updated in a single step along the line of steepest descent of the block objective function until convergence. We instead use a block, coordinate-wise gradient descent (BCGD) method that uses a Newton update, similar to that proposed by Zhou et al. (2010), and we describe this below.

To update \( \beta_j \) from its current estimate, \( \hat{\beta}_j \), we note from (4.3) and (4.4) that if \( \hat{\beta}_j \neq 0 \), the subgradient equation for predictor \( j \) is given by

\[
\partial_j = -X_j'(\hat{r}_l - X_l\hat{\beta}_l) + (1 - \alpha)\lambda w_l \frac{\hat{\beta}_j}{||\beta_l||_2} + \alpha \lambda \cdot \text{sign}(\hat{\beta}_j).
\]

(4.12)

We then descend along the gradient at \( \hat{\beta}_j \) towards the minimum using Newton’s method. The Newton update, \( \hat{\beta}_j^* \), is then given by

\[
\hat{\beta}_j^* = \hat{\beta}_j - \frac{\partial_j}{\partial_j'}
\]

where

\[
\partial_j' = 1 + \frac{(1 - \alpha)\lambda w_l}{||\beta_l||_2} \left( 1 - \frac{\hat{\beta}_j^2}{||\beta_l||_2^2} \right)
\]

(4.13)

is the derivative of (4.12) evaluated at \( \hat{\beta}_j \). The update (4.13) is repeated until convergence.

We must also deal with the case where \( \hat{\beta}_j = 0 \). Here we adopt a slightly different strategy, since the partial derivative, \( t_j \) of \( \beta_j \) is not continuous. We avoid this discontinuity by testing the ‘directional derivatives’, \( \partial_j^+ \) and \( \partial_j^- \), respectively representing the partial derivatives at \( \beta_j = 0 \) in the direction of increasing and decreasing \( \beta_j \). Recalling that we are dealing with the case \( ||\beta_l||_2 \neq 0 \), at \( \beta_j = 0 \) the group penalty term in (4.12) disappears. That is, once a group is selected by model it becomes easier for each SNP coefficient to
move away from zero. The two directional derivatives are then given by

\[
\partial_j^+ = -X_j'(\hat{r}_l - X_l\hat{\beta}_l) + \alpha \lambda \\
\partial_j^- = -X_j'(\hat{r}_l - X_l\hat{\beta}_l) - \alpha \lambda.
\]  

(4.14)

Since the minimising function (4.2) is convex, there are three possible outcomes, and we substitute for \(\partial_j\) in (4.13) accordingly:

\[
\partial_j \leftarrow \begin{cases} 
\partial_j^- & \text{if } \partial_j^- > 0 \text{ and } \partial_j^+ > 0 \\
\partial_j^+ & \text{if } \partial_j^- < 0 \text{ and } \partial_j^+ < 0 \\
0 & \text{if } \partial_j^- < 0 \text{ and } \partial_j^+ > 0
\end{cases}
\]  

(4.15)

In the third case, \(f(\beta_l)\) is increasing either side of \(\beta_j = 0\), so that \(\hat{\beta}_j\) must remain at zero. We can then proceed with the standard Newton update (4.13).

Finally, since the Newton update may occasionally overstep the minimum (where \(\partial_j = 0\)), a simple remedy proposed by Zhou et al. (2010) is to check that \(f(\beta_l)\) is decreasing at each iteration. If this is not the case, then the step size in (4.13) is halved. The complete algorithm for SGL estimation using BCGD is presented in Box 4.1.

One remaining practical issue is the obtaining of a value for \(\lambda_{max}\), the smallest value of \(\lambda\) at which no groups are selected by the model. We require this when setting a value for \(\lambda\), for example when using the variable ranking strategy described in Section 2.5. Noting that \(\hat{r}_l = y\) when no groups are selected, from (4.8) we obtain the smallest value, \(\lambda_{l_{\text{min}}}\), for the minimum value of \(\lambda\) at which group \(l\) is not selected as

\[
\lambda_{l_{\text{min}}} = \frac{||S(X_l'y, \alpha \lambda_{l_{\text{min}}})||_2}{(1 - \alpha)w_l}
\]  

(4.16)

We can solve this in its quadratic form by first setting an upper bound for \(\lambda\) at the point \(\lambda^*_l\), where the soft thresholding function \(S(X_l'y, \alpha \lambda) = 0\), that is when no SNPs are selected by the model. We then obtain the solution by solving

\[
||S(X_l'y, \alpha \lambda_{l_{\text{min}}})||_2^2 - (1 - \alpha)^2(\lambda_{l_{\text{min}}}^*)^2w_l^2 = 0 \quad 0 < \lambda_{l_{\text{min}}} < \lambda^*_l
\]  

(4.17)
for $\lambda_l^{\text{min}}$, where

$$\lambda_l^* = \max_j \frac{|X_j'y|}{\alpha} \quad j = l_1, \ldots, l_P.$$ 

Finally, we obtain a value for $\lambda_{\text{max}}$ as

$$\lambda_{\text{max}} = \max_l \lambda_l^{\text{min}}. \quad (4.18)$$

**Box 4.1 SGL estimation algorithm using BCGD**

1. initialise $\beta \leftarrow 0$.

2. repeat: [pathway loop]
   for pathway $l = 1, 2, \ldots, L$:
   if $\|S(X't_{il}, \alpha \lambda)\|_2 \leq (1 - \alpha)\lambda w_l$
     $\beta_l \leftarrow 0$
   else
     repeat: [SNP loop]
     for $j = l_1, \ldots, l_P$:
     if $\hat{\beta}_j = 0$
       Newton update $\beta_j^* \leftarrow \hat{\beta}_j$ using (4.15) and (4.13)
     else:
       Newton update $\beta_j^* \leftarrow \beta_j$ using (4.12) and (4.13)
     if $f(\beta_j^*) > f(\beta_l)$
       $\beta_j^* \leftarrow \frac{\beta_j^* + \beta_l}{2}$
     $\beta_j \leftarrow \beta_j^*$
     until convergence of $\beta_l$ [SNP loop]
   until convergence of $\beta$ [pathway loop]

3. $\hat{\beta}^{\text{SGL}} \leftarrow \beta$

### 4.1.2 SGL simulation study 1

Using a simple simulation study, we now test the hypothesis that where causal SNPs are enriched in a given pathway, pathway-driven hierarchical SNP selection using SGL will outperform simple lasso selection that disregards pathway information. Since we wish to perform power calculations using a Monte Carlo framework over multiple data simulations, we implement the following simple data simulation protocol:
We simulate $P = 2500$ genetic markers for $N = 400$ individuals. Marker frequencies for each SNP are sampled independently from a multinomial distribution following a Hardy Weinberg equilibrium frequency distribution (see Section 3.4.1 for details). SNP minor allele frequencies are sampled from a uniform distribution $U[0.1, 0.5]$.

SNPs are distributed equally between 50 non-overlapping pathways, each containing 50 SNPs. Since pathways are of equal size, and SNP genotypes are independent, there are no factors biasing pathway selection and we use a uniform pathway weighting vector $w = 1$.

At each MC simulation, a baseline univariate phenotype, $y$ is sampled from $N(10, 1)$. To generate genetic effects, we randomly select 5 SNPs from a single, randomly selected pathway $G_l$, to form the set $S \subset G_l$ of causal SNPs. Genetic effects are then generated as described in Section 3.4.1, but with a univariate phenotype, so that $q = 1$.

To enable a fair comparison between the two methods (SGL and lasso), we ensure that both methods select the same number of SNPs at each simulation. We do this by first obtaining the SGL solution, $\hat{S}^{SGL}$, with $\lambda = 0.85 \lambda_{\text{max}}$ and $\alpha = 0.8$, which ensures sparsity at both the pathway and SNP level. We then compute the lasso solution using coordinate descent over a range of values for the lasso regularisation penalty, $\lambda$, and choose the set $\hat{S}^{\text{lasso}}(\lambda')$ such that $|\hat{S}^{\text{lasso}}(\lambda')| = |\hat{S}^{SGL}|$

where $|\hat{S}^{SGL}|$ is the number of SNPs previously selected by SGL, and $|\hat{S}^{\text{lasso}}(\lambda')|$ is the number of SNPs selected by the lasso with $\lambda = \lambda'$. We measure performance as the mean power to detect all 5 causal SNPs over 500 MC simulations, and test a range of genetic effect sizes ($\gamma$). In a follow up study, we compare the performance of the two methods in a scenario in which pathways information is uninformative. For this we repeat the previous simulations, but with 5 causal SNPs drawn at random from all 2500 SNPs, irrespective of pathway membership. Results are presented in Figure 4.2.

Here we see that where causal SNPs are concentrated in a single causal pathway (Figure 4.2 - left), SGL demonstrates greater power (and equivalently specificity, since the total number of selected SNPs is constant), compared with the lasso, above a particular effect
4.1 The sparse group lasso model

Figure 4.2: SGL vs Lasso. Comparison of power to detect 5 causal SNPs. Each data point represents mean power over 500 MC simulations. *Left:* Causal SNPs drawn from single causal pathway. *Right:* Causal SNPs drawn at random.

size threshold (here $\gamma \approx 0.04$). Where pathway information is not important, that is causal SNPs are not enriched in any particular pathway (Figure 4.2 - right), SGL performs poorly.

To gain a deeper understanding of what is happening here, we also consider the power distributions across all 500 MC simulations corresponding to each point in the plots of Fig.4.2. These are illustrated in Figure 4.3. The top row of plots illustrates the case where causal SNPs are drawn from a single causal pathway. Here we see that there is a marked difference between the two distributions (SGL vs lasso). The lasso shows a smooth distribution in power, with mean power increasing with effect size. In contrast, with SGL the distribution is almost bimodal, with power typically either 0 or 1, depending on whether or not the correct causal pathway is selected. This clearly illustrates the power of hierarchical, pathway-driven SNP selection in the case that pathways are important. As previously found by Zhou et al. (2010) in the context of rare variants and gene selection, the joint modelling of SNPs within groups gives rise to a relaxation of the penalty on individual SNPs within selected groups, relative to the lasso. This can enable the detection of SNPs with small effect size or low MAF that are missed by the lasso, which disregards pathways information and treats all SNPs equally. Finally, where causal SNPs are not enriched in a causal pathway (bottom row of Figure 4.3), as expected SGL performs poorly. In this case
Figure 4.3: SGL vs Lasso. Distribution over 500 MC simulations of power to detect 5 causal SNPs. Each plot represents the power distribution at a single data point in Figure 4.2. The power distribution is discrete, since each method can identify 0, 1, 2, 3, 4 or 5 causal SNPs, with corresponding power 0, 0.2, 0.4, 0.6, 0.8 or 1.0. Top row: Causal SNPs drawn from single causal pathway. Bottom row: Causal SNPs drawn at random.

SGL will only select a SNP where the combined effects of constituent SNPs in a pathway are large enough to drive pathway selection.

4.2 Pathways sparse group lasso with overlaps

Where pathways overlap, we can apply the strategy described in Section 2.2.1 of expanding $X$ and $\beta$, to produce a SGL estimator (4.1) for overlapping groups as

$$\hat{\beta}_{SGL}^* = \arg \min_{\beta} \left\{ \frac{1}{2} ||y - X^* \beta^*||_2^2 + (1 - \alpha)\lambda \sum_{l=1}^{L} w_l ||\beta^*_l||_2 + \alpha \lambda ||\beta^*||_1 \right\}. \quad (4.19)$$

As before, the expanded $(N \times P^*)$ design matrix $X^*$ and $(P^* \times 1)$ coefficient vector
Pathways sparse group lasso with overlaps

$\beta^*$, (where $P^* = \sum_l P_l$), allow each SNP predictor to enter the model separately. Pathway mappings with SNP indices in the expanded variable space are reflected in updated groups $G_k^*, \ldots, G_L^*$. This variable expansion procedure confers the same advantages as with the GL case (Section 2.2.1), namely allowing overlapping pathways to be individually selected (or not selected) by the model, and also ensuring block-separability, so enabling model estimation through block coordinate descent. With SGL, variable expansion also allows overlapping SNPs to be selected (or not selected) in different pathways.

4.2.1 Model estimation

In principle, once the above variable expansion procedure has been applied, the optimisation can be solved using the estimation algorithm described in Box 4.1. However, for the purpose of pathways-driven SNP selection, the application of this algorithm presents a problem. This arises from the replication of overlapping SNP predictors in each group, $X_l^*$, that they occur.

Consider for example the simple situation where there are two pathways, $G_k^*, G_l^*$, containing sets of causal SNPs $S_k^* \subseteq G_k^*$ and $S_l^* \subseteq G_l^*$ respectively. Here the $^*$ indicates that SNP indices refer to the expanded variable space. We begin by assuming that $S_k^*$ and $S_l^*$ contain the same SNPs, so that in the unexpanded variable space, $S_k = S_l$.

We then proceed with BCGD by first estimating $\beta_k^*$. We assume that the correct SNPs are selected, so that $\{ \hat{\beta}_j^* \neq 0 : j \in S_k^* \}$, and $\hat{\beta}_j^* = 0$ otherwise. For the estimation of $\beta_l^*$, the estimated effect $\sum_{j \in S_k^*} X_j^* \hat{\beta}_j^*$, of these overlapping causal SNPs is removed from the regression, through its incorporation in the block residual $\hat{r}_l^* = y - \sum_{j \in S_k^*} X_j^* \hat{\beta}_j^*$. Since no other causal SNPs exist in pathway $G_l^*$, $X_l^* \hat{r}_l^* = 0$, so that the criterion for pathway selection, $||S(X_l^* \hat{r}_l^*, \alpha \lambda)||_2 > (1 - \alpha)\lambda w_l$ (4.10) is not met. That is $G_l^*$ is not selected.

Now consider the case where additional, non-overlapping causal SNPs, possibly with smaller effects, occur in $G_l^*$, so that in the unexpanded variable space, $S_k \subset S_l$. In other words, causal SNPs are partially overlapping (see Fig 4.4). During BCGD pathway $G_l^*$ is then less likely to be selected by the model, than would be the case if there were no

---

1This is the situation for example where multiple causal genes overlap both pathways, but one or more additional causal genes occur in $G_l$. 
overlapping SNPs, since once again the effects of overlapping causal SNPs, $S_k \cap S_l = S_k$, are removed. We provide an illustration of this effect in the simulation study described in Section 4.2.2.

For pathways-driven SNP selection, we will argue that we instead require that SNPs are selected in each and every pathway whose joint SNP effects pass a revised pathway selection threshold

$$||S(X_l^*y, \alpha\lambda)||_2 > (1 - \alpha)\lambda w_1,$$

irrespective of overlaps between pathways. This is equivalent to the previous pathway selection criterion (4.10), but with the additional assumption that pathways are independent, in the sense that they do not compete in the model estimation process. We describe a revised estimation algorithm under the assumption of pathway independence below.

We justify the strong assumption of pathway independence with the following argument. In reality, we expect that multiple pathways may simultaneously influence the phenotype, and we also expect that many such pathways will overlap, for example through their containing one or more 'hub' genes, that overlap multiple pathways (Kim, Wuchty, and Przytycka, 2011; Lehner et al., 2006). By considering each pathway independently, we aim to maximise the sensitivity of our method to detect these variants and pathways. In contrast, without the independence assumption, a competitive estimation algorithm will tend to pick out one from each set of similar, overlapping pathways, and miss potentially causal pathways and variants as a consequence. Again we illustrate this idea in the simulation study in Section 4.2.2. One potential concern is that by not allowing pathways to

![Diagram of two pathways with partially overlapping causal SNPs. Causal SNPs (marked in grey) in the set $S_k$ overlap both pathways, so that $S_k = G_k \cap G_l$. Additional causal SNPs, $S_l \cap \setminus S_k$, (marked in purple) occur in pathway $l$ only.](image)
4.2 Pathways sparse group lasso with overlaps

compete against each other, specificity (or equivalently ranking accuracy) may be reduced, since too many pathways and SNPs may be selected. We aim to avoid this, firstly by once again employing a resampling strategy to highlight the most important pathways and SNPs driving pathway selection. We also describe a heuristic approach to assessing the control of false positives by comparing empirical and null rankings in the application study in Section 4.3. We discuss the issue of specificity further in the context of results from the simulation study in Section 4.2.2.

As an aside, note that with GL, interactions between overlapping pathways are likely to be somewhat diminished. This is because all SNPs in a selected pathway are retained in the model, so that block residuals \( \hat{r}_l, l = 1, \ldots, L \) reflect effects arising from all SNPs in a selected pathway, rather than the effects of a small number of individual SNPs. Interactions between pathways arising from shared SNPs will nonetheless remain, and in fact we exploit these in our GL-based pathway selection models by allowing all pathways to compete, in order to give a robust measure of the relative importance of different pathways. In contrast, with pathways-driven SNP selection, we are additionally concerned with maximising power to identify causal SNPs, so that competition between pathways is less important. Indeed in addition to SNP selection, it is arguably desirable to highlight all potentially important causal pathways in which the combined effect of causal SNPs is sufficient to drive pathway selection.

In what follows we assume as usual that \( X \) and \( \beta \) have been expanded to account for overlaps, but we once again drop the \(^*\) notation for clarity. We precede as before by solving the block-separable optimisation (4.19) for each group or pathway in turn. However, for overlapping pathways, the assumption of pathway independence requires that each \( X_{l_i}, (l_i = 1, \ldots, L) \) is regressed against the full phenotype vector \( y \) rather than the partial residual, \( \hat{r}_l \). With this in mind, the revised subgradient equations for group \( l \) (4.3) are given by

\[ -X'_j (y - \sum_{k \neq j} X_k \hat{\beta}_k - X_j \beta_j) + (1 - \alpha) \lambda w_j \sigma_j + \alpha \lambda t_j = 0 \quad j = l_1, \ldots, l_{P_l}. \tag{4.20} \]

The estimation for group \( l \) then proceeds as described previously in Section 4.1.1, but with the partial residual \( \hat{r}_l \) replaced by \( y \), so that the group sparsity condition (4.8) for
\[ ||\hat{\beta}_l||_2 = 0 \] becomes
\[ ||S(X_l, \alpha \lambda)\|_2 \leq (1 - \alpha) \lambda w_l. \] (4.21)

As before, where group \( l \) is selected by the model, the update for \( \beta_j \), with current estimate \( \hat{\beta}_j \), is derived from the partial derivative (4.12), which under the independence assumption is given by
\[
\partial_j = -X'_j (y - X_l \hat{\beta}_l) + (1 - \alpha) \lambda w_l \frac{\hat{\beta}_j}{||\hat{\beta}_l||_2} + \alpha \lambda \cdot \text{sign}(\hat{\beta}_j),
\] (4.22)
for \( j = l_1, \ldots, l_P \). The Newton update (4.13) remains the same. When \( \hat{\beta}_j = 0 \), the revised directional derivatives (4.14) are given by
\[
\partial^+_j = -X'_j (y - X_l \hat{\beta}_l) + \alpha \lambda \\
\partial^-_j = -X'_j (y - X_l \hat{\beta}_l) - \alpha \lambda.
\] (4.23)

As before the conditions for SNP sparsity within a selected group are determined by (4.15).

The value of \( \lambda_{\text{max}} \), the smallest \( \lambda \) value at which no group is selected by the model, is determined in the same way as before, since this procedure (described in (4.16), (4.17) and (4.18)) does not depend on \( \hat{r}_l \).

Importantly, since each group is regressed independently against the phenotype vector \( y \), there is no block coordinate descent stage in the estimation, that is the revised algorithm utilises only coordinate gradient descent within each selected pathway. For this reason we use the acronym SGL-CGD for the revised algorithm, and SGL-BCGD for the previous algorithm using block coordinate gradient descent. The new algorithm is described in Box 4.2. Note that since the block coordinate descent stage is avoided, the new algorithm has the added benefit of being much faster than would otherwise be the case.

Finally, we note that for SNP selection we are interested only in the set \( \hat{S} \) of selected SNPs in the unexpanded variable space, and not the set \( S^* = \{ j^* : \beta^*_j \neq 0, j^* \in \{1, \ldots, P^*\} \} \). Since, under the independence assumption, the estimation of each \( \beta^*_l \) does not depend on the other estimates, \( \beta^*_k, k \neq l \), we do not need to record separate coefficient estimates for each pathway in which a SNP is selected. Instead we need only record the set \( \hat{S}_l, l \in \hat{C} \) of SNPs selected in each selected pathway. This has a useful practical implication,
since we can avoid the need for an expansion of $X$ or $\beta$, and simply form the complete set of selected SNPs as

$$\hat{S} = \bigcup_{l \in \mathcal{C}} \hat{S}_l.$$ 

**Box 4.2 SGL-CGD estimation algorithm**

1. initialise $\hat{\beta} \leftarrow 0$.

2. for pathway $l = 1, 2, \ldots, L$:
   - if $||S(X_l'\mathbf{y}, \alpha \lambda)||_2 \leq (1 - \alpha) \lambda w_l$
     - $\hat{\beta}_l \leftarrow 0$
   - else
     - repeat: [CGD (SNP) loop]
       - for $j = l_1, \ldots, l_{P_l}$:
         - if $\hat{\beta}_j = 0$
           - Newton update $\hat{\beta}_j^{**} \leftarrow \hat{\beta}_j$ using (4.23) and (4.13)
         - else:
           - Newton update $\hat{\beta}_j^{**} \leftarrow \hat{\beta}_j$ using (4.22) and (4.13)
     - if $f(\beta_l^{**}) > f(\hat{\beta}_l)$:
       - $\hat{\beta}_l^{**} \leftarrow \frac{\hat{\beta}_l^{**} + \hat{\beta}_j}{2}$
       - $\hat{\beta}_j \leftarrow \hat{\beta}_j^{**}$
     - until convergence

4.2.2 SGL simulation study 2

We now explore some of the issues raised in the preceding section, specifically the potential impact on pathway and SNP selection power and specificity of treating the pathways as independent in the SGL estimation algorithm. We do this in a simulation study in which we simulate overlapping pathways. The simulation scheme is specifically designed to highlight differences in pathway and SNP selection with the independence assumption (using the SGL-CGD estimation algorithm in Box 4.2) and without it (using the standard SGL-BCGD estimation algorithm in Box 4.1).
SNPs with variable MAF are simulated using the same procedure described in the previous simulation study (Section 4.1.2), but this time SNPs are mapped to 50 overlapping pathways, each containing 30 SNPs. Each pathway overlaps any adjacent (by pathway index) pathway by 10 SNPs. This overlap scheme is illustrated in Figure 4.5 (a).

As before we consider a range of overall genetic effect sizes, $\gamma$. A total of 2000 MC simulations are conducted for each effect size. At MC simulation $z$, we randomly select two adjacent pathways, $\mathcal{G}_l, \mathcal{G}_{l+1}$ where $l \in \{1, \ldots, 49\}$. From these two pathways we randomly select 10 SNPs according to the scheme illustrated in Figure 4.5 (b). This ensures that causal SNPs overlap a minimum of 1, and a maximum of 2 pathways, with $S_z \subset (\mathcal{G}_l \cap \mathcal{G}_{l-1}) \cup (\mathcal{G}_{l+1} \cap \mathcal{G}_{l+2})$. The true set of causal pathways, $\mathcal{C}$, is then given by $\{l\}$, $\{l+1\}$ or $\{l, l+1\}$ (although simulations where $|\mathcal{C}| = 1$ will be extremely rare). Genetic effects on the phenotype are generated as described previously (Section 4.1.2).

SNP coefficients are estimated for each model, SGL-BCGD and SGL-CGD, using the
same regularisation with $\lambda = 0.85\lambda_{\text{max}}$ and $\alpha = 0.85$ for both.

The average number of pathways and SNPs selected by SGL-BCGD and SGL-CGD across all 2000 MC simulations is reported in Table 4.1. As expected, for both models, the number of selected variables (pathways or SNPs) increases with decreasing effect size, as the number of pathways close to the selection threshold set by $\lambda_{\text{max}}$ (4.18) increases.

Table 4.1: Simulation Study 2: Mean number of pathways and SNPs selected by each model at each effect size, $\gamma$, across 2000 MC simulations.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>0.02</th>
<th>0.04</th>
<th>0.06</th>
<th>0.08</th>
<th>0.1</th>
<th>0.12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pathways</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGL-CGD</td>
<td>5.8</td>
<td>5.9</td>
<td>5.4</td>
<td>4.8</td>
<td>3.9</td>
<td>3.2</td>
</tr>
<tr>
<td>SGL-BCGD</td>
<td>5.8</td>
<td>5.9</td>
<td>5.4</td>
<td>4.8</td>
<td>3.9</td>
<td>3.2</td>
</tr>
<tr>
<td>SNPs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGL-CGD</td>
<td>26.6</td>
<td>27.0</td>
<td>24.8</td>
<td>22.2</td>
<td>18.5</td>
<td>15.3</td>
</tr>
<tr>
<td>SGL-BCGD</td>
<td>28.8</td>
<td>29.3</td>
<td>26.7</td>
<td>23.6</td>
<td>19.4</td>
<td>15.8</td>
</tr>
</tbody>
</table>

For each model, at MC simulation $z$ we record the pathway and SNP selection power, $|\hat{C}_z \cap C_z|/|C_z|$ and $|\hat{S}_z \cap S_z|/|S_z|$ respectively. Since the number of selected variables can vary slightly between the two models, we also record false positive rates (FPR) for pathway and SNP selection as $|\hat{C}_z \cap \setminus C_z|/|\hat{C}_z|$ and $|\hat{S}_z \cap \setminus S_z|/|\hat{S}_z|$ respectively.

The large possible variation in causal SNP distributions, causal SNP MAFs etc. make a comparison of mean power and FPR between the two methods somewhat unsatisfactory. For example, depending on effect size, a large number of simulations can have either very high, or very low pathway and SNP selection power, masking subtle differences in performance between the two methods. Since we are specifically interested in establishing the relative performance of the two methods, we instead illustrate the number of simulations at which one method outperforms the other across all 2000 MC simulations, and show this in Figure 4.6. In this figure, the number of simulations in which SGL-CGD outperforms SGL-BCGD, i.e. where SGL-CGD power $>$ SGL-BCGD power, or SGL-CGD FPR $<$ SGL-BCGD FPR, are shown in green. Conversely, the number of simulations where SGL-BCGD outperforms SGL-CGD are shown in red.

We first consider pathway selection performance (top row of Figure 4.6). For both methods, the same number of pathways are selected on average, across all effect sizes
Figure 4.6: SGL-CGD vs SGL-BCGD performance, measured across 2000 MC simulations. 

**Top row:** Pathway selection performance. (Left) green bars indicate the number of MC simulations where SGL-CGD has greater pathway selection power than SGL-BCGD. Red bars indicate where SGL-BCGD has greater power than SGL-CGD. (Right) green bars indicate the number of MC simulations where SGL-CGD has a lower FPR than SGL-BCGD. Red bars indicate the opposite.

**Bottom row:** As above, but for SNP selection performance.
4.2 Pathways sparse group lasso with overlaps

(Table 4.1). At low effect sizes, there is no difference in performance between the two methods for the large majority of MC simulations, and where there is a difference, the two methods are evenly balanced. As with the previous SGL simulation study (4.1.2), this is the region (with $\gamma \leq 0.04$) where pathway selection fairs no better than chance. With $\gamma > 0.04$, SGL-CGD consistently outperforms SGL-BCGD, both in terms of pathway selection sensitivity and control of false positives (measured by FPR).

To understand why, we turn to SNP selection performance (bottom row of Figure 4.6). At small effect sizes ($\gamma \leq 0.04$), in the small minority of simulations where the correct pathways are identified, SGL-BCGD tends to demonstrate greater power than SGL-CGD (Figure 4.6 bottom left). However, this is at the expense of lower specificity (Figure 4.6 bottom right). These difference are due to the slightly larger number of SNPs selected by SGL-BCGD (see Table 4.1), which in turn is due to the ‘screening out’ of previously selected SNPs from the adjacent causal pathway during BCGD, as described at the beginning of Section 4.2. This results in the selection of a larger number of SNPs when any two overlapping pathways are selected by the model. In the case where two causal pathways are selected, SNP selection power is then likely to be higher, although at the expense of a greater number of false positives.

When pathway effects are just on the margin of detectability ($\gamma = 0.06$), SGL-CGD is more often able to select both causal pathways, although this doesn’t translate into increased SNP selection power. This is most likely because at this effect size neither model can detect SNPs with low MAF, so that SGL-CGD is detecting the same (overlapping) SNPs in both causal pathways. Note that once again SGL-BCGD typically has a higher FPR than SGL-CGD, since more SNPs are selected from non-causal pathways.

As the effect size increases, the number of simulations in which SGL-CGD outperforms SGL-BCGD for SNP selection power grows, paralleling the former method’s enhanced pathway selection power. This is again a demonstration of the screening effect with SGL-BCGD described previously. This means that SGL-CGD is more often able to select both causal pathways, and to select additional causal SNPs that are missed by SGL-BCGD. These additional SNPs are likely to be those with lower MAF, for example, that are harder to detect with SGL-BCGD, once the effect of overlapping SNPs are screened out during estimation using BCGD. Interestingly, as before SGL-CGD continues to exhibit lower false
positive rates than SGL-BCGD. This suggests that, with the simulated data considered here, the independence assumption offers better control of false positives by enabling the selection of causal SNPs in each and every pathway to which they are mapped. In contrast, where causal SNPs are successively screened out during the estimation using BCGD, too many SNPs with spurious effects are selected.

The relative advantage of SGL-CGD over SGL-BCGD on all performance measures starts to decrease around $\gamma = 0.1$, as SGL-BCGD becomes better able to detect all causal pathways and SNPs, irrespective of the screening effect.

4.2.3 Weight tuning

SGL-CGD will be subject to the same biases arising for example from variations in the number and size of pathway genes, as was the case with the previous GL-based P-GLAW method. We therefore employ a similar weight-tuning strategy as before in which we iteratively adapt the pathway weight vector, $\mathbf{w}$ to achieve a near uniform empirical pathway selection frequency distribution, $\Pi^*$, over $R$ permutations of the phenotype vector, each with $\lambda$ tuned to select a single pathway.

For SGL-CGD with a univariate response, the weight tuning algorithm can be significantly simplified by noting that there is no need to fit the model to estimate each $\hat{\beta}_l$. This is for the following reasons. Firstly, with a univariate phenotype we don’t perform PsRRR, so that no estimate for the full genotype vector $\beta$ is required for estimating phenotype coefficients. Secondly, as discussed in the previous section, under the pathway independence assumption, each $\hat{\beta}_l$ has no influence on any other $\hat{\beta}_k$, $k \neq l$. Thus with a fixed value for $\alpha$, and with $\lambda$ tuned to select a single pathway, we need only establish which pathway enters the model first, as $\lambda$ is reduced from its maximal value, $\lambda_{\text{max}}$. From (4.18), at phenotype permutation $r$, the pathway $\hat{C}_r$ selected with permuted phenotype $y_r$ is then given by

$$\hat{C}_r = \{\arg \max_l \lambda_l^{\text{min}}\}$$

where

$$\lambda_l^{\text{min}} = \frac{||S(X_l'y_r, \lambda_l^{\text{min}})||_2}{(1 - \alpha)w_l}.$$
4.2 Pathways sparse group lasso with overlaps

Weight tuning then proceeds in exactly the same way as described under P-GLAW (Section 2.4). This ability to skip the full model estimation process results in a considerable reduction in computation time.

4.2.4 Pathway, SNP and gene ranking

As with the previous methods, we rank selected variables using a resampling strategy in which we fit the model $B$ times, each time using a random subsample of the full dataset of size $N/2$, drawn without replacement. The procedure for pathway ranking is exactly the same as that described for GL-based models.

For SNP and gene ranking, we denote the set of SNPs selected at sample $b$ (in the unexpanded variable space) by $\hat{S}^{(b)}$, where

$$\hat{S}^{(b)} = \{j : \hat{\beta}_j \neq 0, j \in G_l, l \in \{1, \ldots, L\}\},$$

and follow the same procedure for SNP and gene ranking as described at the end of Section 3.3.2. Note that in the case of SNPs that overlap multiple pathways, the same SNPs may be selected in more than one pathway.

Our proposed ‘pathways sparse group lasso with adaptive weights’ (P-SGLAW) method combines the various elements described in Section 4.2, namely estimation of the SGL model (4.19) using SGL-CGD, together with the procedure for weight tuning to remove bias, and the resampling strategy for ranking pathways, SNPs and genes in order of importance. In the next section we apply this method to an application study using real data.
4.3 P-SGLAW application study: Identification of pathways, SNPs and genes associated with high-density lipoprotein cholesterol

In this section we describe an investigation of pathways, SNPs and genes associated with high-density lipoprotein cholesterol (HDL-C) levels in two separate cohorts of Asian adults. HDL-C is sometimes referred to as ‘good cholesterol’, since high levels of HDL-C have previously been linked to a reduced incidence of cardiovascular disease (Gordon et al., 1989), and it is also thought to protect patients from the development and progression of atherosclerotic disease and associated morbidity (Toth, 2005).

4.3.1 Subjects, genotypes and phenotypes

The analysis is carried out using two datasets supplied by colleagues at the National University of Singapore, that have previously been used to search for novel variants associated with type 2 diabetes mellitus (T2D) in Asian populations. The first (discovery) cohort is from the Singapore Prospective Study Program, hereafter referred to as ‘SP2’, and the second (replication) dataset is from the Singapore Malay Eye Study or ‘SiMES’. Detailed information on both datasets can be found in Sim et al. (2011), but we briefly outline some salient features here.

Both datasets comprise whole genome data for T2D cases and controls, genotyped on the Illumina HumanHap 610 Quad array. For the present study we use controls only, since variation in lipid levels between cases and controls can be greater than the variation within controls alone. The use of both cases and controls in our analysis might then lead to a confounded analysis, where any associations could be linked to T2D status or some other spurious factor.

The SP2 dataset consists entirely of ethnic Chinese, and shows no evidence of population stratification. The SiMES dataset comprises ethnic Malays, and shows some evidence of cryptic relatedness between samples. For this reason, the first two principal components of a PCA for population structure are used as covariates in our analysis of this dataset. Again full details of the stratification analysis can be found in Sim et al. (2011) and associated supplementary information.
A summary of information pertaining to genotypes for each dataset, both before and after imputation and pathway mapping, is given in Table 4.2, along with a list of phenotypes and covariates.

Table 4.2: Genotype and phenotype information corresponding to the SP2 and SiMES datasets used in the study.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>SP2</th>
<th>Simes</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1,040</td>
<td>1,099</td>
</tr>
</tbody>
</table>

**Genotypes**

*Before imputation*

- SNPs available for analysis\(^{(1)}\): 542,297 | 557,824
- SNPs with missing genotypes\(^{(2)}\): 152,372 | 282,549

*Post imputation*

- SNPs available for analysis\(^{(3)}\): 492,639 | 515,503

**Phenotypes/covariates**

- quantitative trait (phenotype)\(^{(4)}\): HDL-C | HDL-C
- covariates: gender, age, age\(^2\), BMI\(^{(5)}\) | gender, age, age\(^2\), BMI, PC1, PC2\(^{(6)}\)

\(^{(1)}\) after first round of quality control (Sim et al., 2011) and removal of monomorphic SNPs
\(^{(2)}\) maximum 5% missing rate per SNP
\(^{(3)}\) after imputation and removal of SNPs with MAF < 0.01
\(^{(4)}\) mg/dL
\(^{(5)}\) body mass index (kg/m\(^2\))
\(^{(6)}\) principal components relating to cryptic relatedness

### 4.3.2 Genotype imputation

After the initial round of quality control, genotypes for both datasets have a maximum SNP missingness of 5%. Since our method cannot handle missing values, we perform ‘missing holes’ SNP imputation, so that all missing SNP calls are estimated against a reference panel of known haplotypes.

SNP imputation proceeds in two stages. First, imputation requires accurate estimation of haplotypes from diploid genotypes. This process, known as phasing, is performed using SHAPEIT v1 (http://www.shapeit.fr). This uses a hidden Markov model to infer haplotypes...
from sample genotypes using a map of known recombination rates across the genome (De-
laneau, Marchini, and Zagury, 2012). The recombination map must correspond to genotype
coordinates in the dataset to be imputed, so we use recombination data from HapMap phase
recombination/2008-03_rel22_B36/).

Following the primary phasing stage, SNP imputation is performed using IMPUTE
v2.2.2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html). IMPUTE uses a reference
panel of known haplotypes to infer unobserved genotypes, given a set of observed sample
haplotypes (Howie, Marchini, and Stephens, 2011). The latest version (IMPUTE 2) uses
an updated, efficient algorithm, so that a custom reference panel can be used for each
study haplotype, and for each region of the genome, enabling the full range of reference
information provided by HapMap3 (The 1000 Genomes Project Consortium, 2011) to be
used. Following IMPUTE 2 guidelines, we use HapMap3 reference data corresponding to
NCBI b36 (http://mathgen.stats.ox.ac.uk/impute/data_download_hapmap3_r2.html) which
includes haplotype data for 1,011 individuals from Africa, Asia, Europe and the Americas.
SNPs are imputed in 5MB chunks, using an effective population size ($N_e$) of 15,000, and a
buffer of 250kb to avoid edge effects, again as recommended for IMPUTE 2.

The phasing and imputation process is complex and computationally intensive. For this
reason we implement a pipeline in Python, with phasing and imputation for each chromo-
some conducted in parallel across multiple nodes in a computing cluster. This enables full
genome imputation that would otherwise take days, to be completed in a matter of hours.

4.3.3 Pathway mapping

Following imputation, SNPs for both datasets are mapped to KEGG canonical pathways
from the MSigDB database (http://www.broadinstitute.org/gsea/msigdb/index.jsp), as de-
scribed in the previous AD study (Section 3.5.4). Details of the pathway mapping process
are given in Figures 4.7 and 4.8.

Note that there is a difference in the number of SNPs available for the pathway mapping
between the two datasets, and this results in a small discrepancy in the total number of
mapped genes (SP2: 4,734 mapped genes; SiMES: 4,751). However, both datasets map
4.3 P-SGLAW application study: Gene pathways associated with HDL-C

**Figure 4.7:** *SP2 dataset.* SNP to pathway mapping.

**Figure 4.8:** *SiMES dataset.* SNP to pathway mapping.
to all 185 KEGG pathways, and a large majority of mapped genes and SNPs overlap both datasets. Detailed information on the pathway mapping process for the two datasets is presented in Table 4.3.

Table 4.3: SNP and gene to pathway mappings for the SP2 and SiMES datasets.

<table>
<thead>
<tr>
<th></th>
<th>SP2</th>
<th>SiMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SNPs mapping to pathways</td>
<td>75,389</td>
<td>78,933</td>
</tr>
<tr>
<td>Total SNPs mapping to pathways in both datasets (intersection)</td>
<td>74,864</td>
<td></td>
</tr>
<tr>
<td>Total mapped genes</td>
<td>4,734</td>
<td>4,751</td>
</tr>
<tr>
<td>Total genes mapping to pathways in both datasets (intersection)</td>
<td></td>
<td>4,726</td>
</tr>
<tr>
<td>Total mapped pathways</td>
<td>185</td>
<td>185</td>
</tr>
<tr>
<td>Minimum number of genes mapping to single pathway</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Maximum number of genes mapping to single pathway</td>
<td>63</td>
<td>63</td>
</tr>
<tr>
<td>Minimum number of SNPs mapping to single pathway</td>
<td>66</td>
<td>67</td>
</tr>
<tr>
<td>Maximum number of SNPs mapping to single pathway</td>
<td>5,759</td>
<td>6,058</td>
</tr>
<tr>
<td>Minimum number of pathways mapping to a single SNP</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum number of pathways mapping to a single SNP</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

4.3.4 Results

We perform pathways-driven SNP selection on both datasets, using the P-SGLAW method described previously. We present results for each dataset separately below.

SP2 Dataset

For the SP2 dataset we consider two separate scenarios for the regularisation parameters $\lambda$ and $\alpha$. For the two scenarios we set the sparsity parameter, $\lambda = 0.95$, but consider two values for $\alpha$, namely $\alpha = 0.95, 0.85$. We test each scenario over 1000 $N/2$ subsamples. We also compare the resulting pathway and SNP selection frequency distributions with null distributions, again over 1000 $N/2$ subsamples, but with phenotype labels permuted, so that no SNPs can influence the phenotype.

The parameter $\alpha$ controls how the regularisation penalty is distributed between the $\ell_2$ (pathway) and $\ell_1$ (SNP) norms of the coefficient vector. Each scenario therefore entails
4.3 P-SGLAW application study: Gene pathways associated with HDL-C

different numbers of selected pathways and SNPs, and this information is presented in Table 4.4.

Table 4.4: Separate combinations of the P-SGLAW regularisation parameters, $\lambda$ and $\alpha$ used for analysis of the SP2 dataset. For each $\lambda$, $\alpha$ combination, the mean ($\pm$SD) number of selected pathways and SNPs across all 1000 subsamples is reported.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha$</th>
<th>Empirical</th>
<th>Null</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>0.85</td>
<td>7.9 ± 6.1</td>
<td>9.1 ± 7.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1551 ± 1294</td>
<td>1656 ± 1401</td>
</tr>
<tr>
<td>0.95</td>
<td>0.95</td>
<td>4.8 ± 4.1</td>
<td>5.0 ± 4.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160 ± 185</td>
<td>155 ± 194</td>
</tr>
</tbody>
</table>

Comparisons of empirical and null pathway selection frequency distributions for each scenario are presented in Figure 4.9. The same comparisons for SNP selection frequencies are presented in Figure 4.10. In these plots, null distributions (coloured blue) are ordered along the $x$-axis according to their corresponding ranked empirical selection frequencies (marked in red). This is to help visualise any potential biases that may be influencing variable selection (see below).

To interpret these results, we begin by noting from Table 4.4 that many more SNPs are selected with $\alpha = 0.85$, resulting in higher SNP selection frequencies, compared to those obtained with $\alpha = 0.95$ (see Figure 4.10). This is as expected, since a lower value for $\alpha$ implies a reduced $\ell_1$ penalty on the SNP coefficient vector, resulting in more SNPs being selected. Perhaps surprisingly, given that the $\ell_2$ group penalty $(1 - \alpha)\lambda$ is increased, the number of selected pathways is also greater. This must reflect the reduced $\ell_1$ penalty, which allows a greater number of SNPs to contribute to a putative selected pathway’s coefficient vector. This in turn increases the number of pathways that pass the threshold for selection.

This raises the question of what might be considered to be an optimal choice for the regularisation-distributional parameter $\alpha$, since different assumptions about the number of SNPs potentially influencing the phenotype may affect the resulting pathway and SNP
Figure 4.9: Empirical and null pathway selection frequency distributions for all 185 KEGG pathways with the SP2 dataset. For each scenario, pathways are ranked along the x-axis in order of their empirical pathway selection frequency, $\pi_{l_1}^{\text{path}}, \ldots, \pi_{l_L}^{\text{path}}$. (a) $\alpha = 0.85$. (b) $\alpha = 0.95$. 
4.3 P-SGLAW application study: Gene pathways associated with HDL-C

Figure 4.10: Empirical and null SNP selection frequency distributions with the SP2 dataset. For each scenario, SNPs are ranked along the $x$-axis in order of their empirical pathway selection frequency, $\pi_{j1}^{SNP} > \pi_{j2}^{SNP} > \ldots$. (a) $\alpha = 0.85$. (b) $\alpha = 0.95$. Note fewer SNPs are selected with nonzero empirical selection frequency with $\alpha = 0.95$, so that the $x$-axis range in (b) is reduced.
To answer this, we turn our attention to the pathway and SNP selection frequency distributions for each $\alpha$ value in Figures 4.9 and 4.10. At the lower value of $\alpha = 0.85$ (top plots in Figures 4.9 and 4.10), empirical pathway and SNP selection frequency distributions appear to be biased, in the sense that there is a suggestion that pathways and SNPs with the highest empirical selection frequencies also tend to be selected with a higher frequency under the null, where there is no association between genotype and phenotype. This relationship appears to be diminished with $\alpha = 0.95$, when fewer SNPs are selected by the model. We investigate this further by plotting empirical vs. null selection frequencies as a sequence of scatter plots in Figure 4.11, and we report Pearson correlation coefficients and p-values for these in Table 4.5.

### Table 4.5: SP2 dataset: Pearson correlation coefficients ($r$) and p-values for the data plotted in Figure 4.11. $n$ denotes the number of predictors considered. For SNPs, coefficients describe correlations for all predictors selected with nonzero empirical selection frequencies only, since a large number of SNPs are not selected by the model at any subsample.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.85$</th>
<th>$\alpha = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$  $r$  p-value</td>
<td>$n$  $r$  p-value</td>
</tr>
<tr>
<td>pathways</td>
<td>185  0.66  $1.3 \times 10^{-24}$</td>
<td>185  0.26  $2.9 \times 10^{-4}$</td>
</tr>
<tr>
<td>SNPs</td>
<td>62,965 0.37 0</td>
<td>30,027 0.11  $1.2 \times 10^{-84}$</td>
</tr>
</tbody>
</table>

These provide further evidence of increased correlation between empirical and null selection frequency distributions at the lower $\alpha$ value for both pathways and SNPs, again suggesting increased bias in the empirical results, in the sense that certain pathways and SNPs tend to be selected with a higher frequency, irrespective of whether or not a true signal may be present. Further qualitative evidence of reduced bias with $\alpha = 0.95$ is suggested by the clearer separation of empirical and null distributions at the higher $\alpha$ value in Figures 4.9 and 4.10. For example, the maximum empirical pathway selection frequency is reduced by a factor of 0.29 (0.35 to 0.25) as $\alpha$ is increased from 0.85 to 0.95, whereas the maximum pathway selection frequency under the null is reduced by a factor of 0.81 (0.29 to 0.054). Similarly for SNPs, the maximum empirical SNP selection frequency is reduced by a factor of 0.37 (0.52 to 0.33), whereas the maximum SNP selection frequency under the null is reduced by a factor of 0.9 (0.11 to 0.011).
4.3 P-SGLAW application study: Gene pathways associated with HDL-C

Figure 4.11: SP2 dataset: Scatter plots comparing empirical and null selection frequencies presented in Figures 4.9 and 4.10. (a) and (b): Pathway selection frequencies with $\alpha = 0.85, 0.95$ respectively. (c) and (d): SNP selection frequencies for the same $\alpha$ values. For clarity, SNP selection frequencies are plotted for the top 1000 SNPs (by empirical selection frequency) only. Corresponding correlation coefficients (for all ranked SNPs) are presented in Table 4.5. Note that pathway and SNP selection frequencies are much higher at the lower $\alpha$ value (left hand plots), since many more variables are selected (see Table 4.4).
The increased bias with $\alpha = 0.85$ is most likely due to the selection of too many SNPs, in the sense that many selected SNPs do not exhibit real phenotypic effects. These extra SNPs effectively add noise to the model, in the form of multiple weak, spurious signals. This in turn will add bias to the resulting selection frequency distributions, tending to favour, for example, SNPs that overlap multiple pathways, and the pathways that contain them. As $\alpha$ is increased, we would expect this biasing effect to be reduced, until a point where too few SNPs are selected, when there is then a risk that some of the true signal may be lost.

Note that the reduced but still significant correlations between empirical and null selection frequency distributions at $\alpha = 0.95$ in Table 4.5 are not unexpected. These may reflect the complex overlap structure between pathways, meaning that pathways (and associated SNPs) with a relatively high degree of overlap with other pathways, due for example to the presence of so called ‘hub genes’, are more likely to harbour true signals, as well as spurious ones (Lehner et al., 2006; Carter et al., 2004; Jeong et al., 2001).

Taking all the above into consideration, we choose to report results with $\alpha = 0.95$, where there is less evidence of bias due to the selection of too many SNPs. The top 30 pathways, ranked by selection frequency are presented in Table 4.6, and the top 30 ranked SNPs, together with corresponding genes to which they are mapped are presented in Table 4.7.
Table 4.6: SP2 dataset: Top 30 pathways, ranked by pathway selection frequency, $\pi_{\text{path}}$. The final column lists genes in the pathway that are in the top 30 ranked genes selected in the study (see Table 4.7). Pathways falling in the consensus set, $\Psi_{25}$, obtained by comparing pathway ranking results from both SP2 and SiMES datasets (see Table 4.11), are marked with an $^*$. 

<table>
<thead>
<tr>
<th>Rank</th>
<th>KEGG pathway name</th>
<th>$\pi_{\text{path}}$</th>
<th>Size (# SNPs)</th>
<th>top 30 ranked genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toll Like Receptor Signaling Pathway</td>
<td>0.254</td>
<td>766</td>
<td>TIRAP RAC1 IFNAR1 CD80 IL12B PIK3R1</td>
</tr>
<tr>
<td>2</td>
<td>Jak Stat Signaling Pathway</td>
<td>0.170</td>
<td>1447</td>
<td>PIA52 IL12A TPO IFNAR1 IL12B PIK3R1 IL2RA</td>
</tr>
<tr>
<td>3</td>
<td>Ubiquitin Mediated Proteolysis</td>
<td>0.165</td>
<td>1603</td>
<td>PIA52 RFWD2 PARK2</td>
</tr>
<tr>
<td>4</td>
<td>* Dilated Cardiomyopathy</td>
<td>0.103</td>
<td>3054</td>
<td>ADCY2 TGFBI PKRACB RYR2 ITGB8 ITGA1 CACNA2D3 LAMA2 CACNA1C</td>
</tr>
<tr>
<td>5</td>
<td>Cytokine Cytokine Receptor Interaction</td>
<td>0.100</td>
<td>2533</td>
<td>IL5R3A IL12B TGFBI EGFR TPO IFNAR1 IL2RA</td>
</tr>
<tr>
<td>6</td>
<td>ECM Receptor Interaction</td>
<td>0.095</td>
<td>2271</td>
<td>ITGB8 ITGA1 LAMA2</td>
</tr>
<tr>
<td>7</td>
<td>Arginine And Proline Metabolism</td>
<td>0.091</td>
<td>432</td>
<td>NOS1</td>
</tr>
<tr>
<td>8</td>
<td>Parkinson’s Disease</td>
<td>0.090</td>
<td>1320</td>
<td>PARK2</td>
</tr>
<tr>
<td>9</td>
<td>* Hypertrophic Cardiomyopathy</td>
<td>0.088</td>
<td>2819</td>
<td>TGFBI RYR2 ITGB8 ITGA1 CACNA2D3 LAMA2 CACNA1C</td>
</tr>
<tr>
<td>10</td>
<td>Small Cell Lung Cancer</td>
<td>0.068</td>
<td>1808</td>
<td>PIA52 PIK3R1 LAMA2</td>
</tr>
<tr>
<td>11</td>
<td>Natural Killer Cell Mediated Cytotoxicity</td>
<td>0.067</td>
<td>1781</td>
<td>KRAS RAC1 VAV3 VAV2 PRKCA IFNAR1 PRKACB PIK3R1</td>
</tr>
<tr>
<td>12</td>
<td>* T Cell Receptor Signaling Pathway</td>
<td>0.065</td>
<td>1541</td>
<td>KRAS VAV3 VAV2 PIK3R1</td>
</tr>
<tr>
<td>13</td>
<td>Tgf Beta Signaling Pathway</td>
<td>0.065</td>
<td>947</td>
<td>TGFBI</td>
</tr>
<tr>
<td>14</td>
<td>Ollactory Transduction</td>
<td>0.065</td>
<td>2497</td>
<td>PRKACB</td>
</tr>
<tr>
<td>15</td>
<td>* Arrhythmogenic Right Ventricular Cardiomyopathy</td>
<td>0.063</td>
<td>3726</td>
<td>RYR2 TCF7L1 ITGB8 ITGA1 CACNA2D3 LAMA2 CACNA1C</td>
</tr>
<tr>
<td>16</td>
<td>* Ppar Signaling Pathway</td>
<td>0.062</td>
<td>758</td>
<td>PRKACB</td>
</tr>
<tr>
<td>17</td>
<td>Taste Transduction</td>
<td>0.062</td>
<td>941</td>
<td>PRKACB</td>
</tr>
<tr>
<td>18</td>
<td>Type I Diabetes Mellitus</td>
<td>0.060</td>
<td>776</td>
<td>CD80 IL12B</td>
</tr>
<tr>
<td>19</td>
<td>* Ribosome</td>
<td>0.057</td>
<td>261</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>* Terpenoid Backbone Biosynthesis</td>
<td>0.056</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Neuroactive Ligand Receptor Interaction</td>
<td>0.053</td>
<td>5745</td>
<td>GRIN3A</td>
</tr>
<tr>
<td>22</td>
<td>Regulation Of Actin Cytoskeleton</td>
<td>0.053</td>
<td>3803</td>
<td>KRAS RAC1 EGFR TGFBI VAV3 ITGA1 VAV2 PIK3R1</td>
</tr>
<tr>
<td>23</td>
<td>Mismatch Repair</td>
<td>0.053</td>
<td>222</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Cell Adhesion Molecules Cami</td>
<td>0.053</td>
<td>3977</td>
<td>ITGB8 CD80</td>
</tr>
<tr>
<td>25</td>
<td>Maturity Onset Diabetes Of The Young</td>
<td>0.053</td>
<td>239</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Butanoate Metabolism</td>
<td>0.052</td>
<td>383</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Purine Metabolism</td>
<td>0.052</td>
<td>3224</td>
<td>ADCY2</td>
</tr>
<tr>
<td>28</td>
<td>P53 Signaling Pathway</td>
<td>0.052</td>
<td>598</td>
<td>RFWD2</td>
</tr>
<tr>
<td>29</td>
<td>Dorso Ventral Axis Formation</td>
<td>0.050</td>
<td>581</td>
<td>KRAS EGFR</td>
</tr>
<tr>
<td>30</td>
<td>Basal Cell Carcinoma</td>
<td>0.049</td>
<td>589</td>
<td>TCF7L1</td>
</tr>
</tbody>
</table>
Table 4.7: SP2 dataset: Top 30 SNPs and genes, respectively ranked by SNP and gene selection frequency. Genes falling in the top 30 ranks of the consensus gene set, $\Psi_{244}^{\text{gene}}$, obtained by comparing gene ranking results from both SP2 and SiMES datasets (see Table 4.13), are marked with an $^*$.  

<table>
<thead>
<tr>
<th>Rank</th>
<th>SNP</th>
<th>$\pi_{SNP}$</th>
<th>Mapped gene(s)</th>
<th>Gene</th>
<th>$\pi_{gene}$</th>
<th># mapped SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2257167</td>
<td>0.33</td>
<td>IFNAR1</td>
<td>IFNAR1</td>
<td>0.33</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>rs2254315</td>
<td>0.32</td>
<td>IFNAR1</td>
<td>IL12B</td>
<td>0.30</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>rs1041868</td>
<td>0.32</td>
<td>IFNAR1</td>
<td>PIAS2</td>
<td>0.30</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>rs7364085</td>
<td>0.31</td>
<td>IFNAR1</td>
<td>TIRAP</td>
<td>0.22</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>rs2850021</td>
<td>0.31</td>
<td>IFNAR1</td>
<td>RAC1</td>
<td>0.21</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>rs2253413</td>
<td>0.31</td>
<td>IFNAR1</td>
<td>LAMA2*</td>
<td>0.19</td>
<td>111</td>
</tr>
<tr>
<td>7</td>
<td>rs2243590</td>
<td>0.31</td>
<td>IFNAR1</td>
<td>ADCY2*</td>
<td>0.19</td>
<td>94</td>
</tr>
<tr>
<td>8</td>
<td>rs2834204</td>
<td>0.31</td>
<td>IFNAR1</td>
<td>PIK3R1</td>
<td>0.19</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>rs3181224</td>
<td>0.30</td>
<td>IL12B</td>
<td>PARK2</td>
<td>0.19</td>
<td>460</td>
</tr>
<tr>
<td>10</td>
<td>rs512868</td>
<td>0.28</td>
<td>PIAS2</td>
<td>IL2RA</td>
<td>0.19</td>
<td>55</td>
</tr>
<tr>
<td>11</td>
<td>rs641366</td>
<td>0.24</td>
<td>PIAS2</td>
<td>PRKCA*</td>
<td>0.19</td>
<td>123</td>
</tr>
<tr>
<td>12</td>
<td>rs2032215</td>
<td>0.22</td>
<td>PIAS2</td>
<td>ITGB8</td>
<td>0.18</td>
<td>27</td>
</tr>
<tr>
<td>13</td>
<td>rs8177375</td>
<td>0.22</td>
<td>TIRAP</td>
<td>TCF7L1</td>
<td>0.18</td>
<td>55</td>
</tr>
<tr>
<td>14</td>
<td>rs10893493</td>
<td>0.21</td>
<td>TIRAP</td>
<td>CD80*</td>
<td>0.18</td>
<td>21</td>
</tr>
<tr>
<td>15</td>
<td>rs4890341</td>
<td>0.20</td>
<td>PIAS2</td>
<td>GRIN3A</td>
<td>0.18</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>rs2303361</td>
<td>0.19</td>
<td>RAC1</td>
<td>PRKCB*</td>
<td>0.18</td>
<td>83</td>
</tr>
<tr>
<td>17</td>
<td>rs7873495</td>
<td>0.17</td>
<td>GRIN3A</td>
<td>CACNA1C*</td>
<td>0.17</td>
<td>180</td>
</tr>
<tr>
<td>18</td>
<td>rs1323653</td>
<td>0.17</td>
<td>IL2RA</td>
<td>TGF3</td>
<td>0.16</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>rs11762117</td>
<td>0.16</td>
<td>ITGB8</td>
<td>PRKACB</td>
<td>0.16</td>
<td>16</td>
</tr>
<tr>
<td>20</td>
<td>rs3807955</td>
<td>0.16</td>
<td>ITGB8</td>
<td>KRAS*</td>
<td>0.16</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>rs10462842</td>
<td>0.15</td>
<td>ADCY2</td>
<td>VAV3</td>
<td>0.16</td>
<td>97</td>
</tr>
<tr>
<td>22</td>
<td>rs10215885</td>
<td>0.15</td>
<td>ITGB8</td>
<td>IL5RA</td>
<td>0.15</td>
<td>38</td>
</tr>
<tr>
<td>23</td>
<td>rs3807936</td>
<td>0.15</td>
<td>ITGB8</td>
<td>ITGA1*</td>
<td>0.15</td>
<td>77</td>
</tr>
<tr>
<td>24</td>
<td>rs530205</td>
<td>0.15</td>
<td>PIAS2</td>
<td>VAV2*</td>
<td>0.15</td>
<td>85</td>
</tr>
<tr>
<td>25</td>
<td>rs3823974</td>
<td>0.15</td>
<td>ITGB8</td>
<td>EGFR*</td>
<td>0.14</td>
<td>61</td>
</tr>
<tr>
<td>26</td>
<td>rs2074425</td>
<td>0.15</td>
<td>ITGB8</td>
<td>TPO</td>
<td>0.14</td>
<td>50</td>
</tr>
<tr>
<td>27</td>
<td>rs6725799</td>
<td>0.15</td>
<td>TCF7L1</td>
<td>CACNA2D3*</td>
<td>0.14</td>
<td>283</td>
</tr>
<tr>
<td>28</td>
<td>rs2301727</td>
<td>0.15</td>
<td>ITGB8</td>
<td>RYR2*</td>
<td>0.14</td>
<td>214</td>
</tr>
<tr>
<td>29</td>
<td>rs10486391</td>
<td>0.14</td>
<td>ITGB8</td>
<td>NOS1</td>
<td>0.14</td>
<td>49</td>
</tr>
<tr>
<td>30</td>
<td>rs3779505</td>
<td>0.14</td>
<td>ITGB8</td>
<td>RFWD2</td>
<td>0.13</td>
<td>31</td>
</tr>
</tbody>
</table>
**SiMES Dataset**

For the replication SiMES dataset, we repeat the above analysis design, but consider only the ‘low bias’ scenario where $\lambda = 0.95$ and $\alpha = 0.95$. Once again we test each scenario over 1000 $N/2$ subsamples, and compare the resulting pathway and SNP selection frequency distributions with null distributions generated over 1000 $N/2$ subsamples with phenotype labels permuted. Pathway and SNP selection frequency distributions are presented in Figure 4.13. An investigation of pathway and SNP selection bias is presented in the form of scatter plots illustrating potential correlation between empirical and null selection frequencies in Figure 4.12, with corresponding Pearson correlation coefficients and p-values presented in Table 4.8. The top 30 ranked pathways, and SNPs and genes are presented in Tables 4.9 and 4.10 respectively.

![Figure 4.13](image)

**Table 4.8:** SiMES dataset: Pearson correlation coefficients ($r$) and p-values for the data plotted in Figure 4.12. $n$ denotes the number of predictors considered. For SNPs, coefficients describe correlations for all predictors selected with nonzero empirical selection frequencies only, since a large number of SNPs are not selected by the model at any subsample.

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>$r$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pathways</td>
<td>185</td>
<td>-0.094</td>
<td>0.20</td>
</tr>
<tr>
<td>SNPs</td>
<td>20,006</td>
<td>0.058</td>
<td>$2.63 \times 10^{-16}$</td>
</tr>
</tbody>
</table>
Figure 4.13: Empirical and null pathway (top) and SNP (bottom) selection frequency distributions for the SiMES dataset. $\alpha = 0.95$. For both empirical (red) and null (blue) distributions, variables (pathways and SNPs) are ranked along the $x$-axis in order of their empirical selection frequencies.
Table 4.9: SiMES dataset: Top 30 pathways, ranked by pathway selection frequency, \( \pi^{path} \). The final column lists genes in the pathway that are in the top 30 ranked genes selected in the study (i.e. genes in the top 30 gene ranking in Table 4.10). Pathways falling in the consensus set, \( \Psi_{path}^{25} \), obtained by comparing pathway ranking results from both SP2 and SiMES datasets (see Table 4.11), are marked with a *.

<table>
<thead>
<tr>
<th>Rank</th>
<th>KEGG pathway name</th>
<th>( \pi^{path} )</th>
<th>Size (# SNPs)</th>
<th>top 30 ranked genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oxidative Phosphorylation</td>
<td>0.314</td>
<td>871</td>
<td>PPA2 NDUF4 SDHB SDHC ATP6V0A4</td>
</tr>
<tr>
<td>2</td>
<td>Terpenoid Backbone Biosynthesis</td>
<td>0.260</td>
<td>158</td>
<td>PDSS2</td>
</tr>
<tr>
<td>3</td>
<td>Regulation Of Autophagy</td>
<td>0.183</td>
<td>215</td>
<td>GABARAPL1</td>
</tr>
<tr>
<td>4</td>
<td>Glycerolipid Metabolism</td>
<td>0.095</td>
<td>1074</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>* Dilated Cardiomyopathy</td>
<td>0.078</td>
<td>3177</td>
<td>ADCY2 RYR2 ITGA1 ITGA1 LAMA2 CACNA2D3 LAMA2 CACNA1C CACNA1D</td>
</tr>
<tr>
<td>6</td>
<td>* Hypertrophic Cardiomyopathy</td>
<td>0.071</td>
<td>2932</td>
<td>PRKAG2 RYR2 ITGA1 ITGA1 LAMA2 CACNA2D3 LAMA2 CACNA1C CACNA1D</td>
</tr>
<tr>
<td>7</td>
<td>* Ribosome</td>
<td>0.064</td>
<td>270</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Glutathione Metabolism</td>
<td>0.055</td>
<td>389</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>* Arrhythmogenic Right Ventricular Cardiomyopathy</td>
<td>0.053</td>
<td>3899</td>
<td>RYR2 ITGA1 ITGA1 LAMA2 CACNA2D3 LAMA2 CACNA1C CACNA1D</td>
</tr>
<tr>
<td>10</td>
<td>* T Cell Receptor Signaling Pathway</td>
<td>0.052</td>
<td>1624</td>
<td>PK7 FYN</td>
</tr>
<tr>
<td>11</td>
<td>Cardiac Muscle Contraction</td>
<td>0.047</td>
<td>1952</td>
<td>RYR2 LAMA2 CACNA2D3 CACNA1C CACNA1D</td>
</tr>
<tr>
<td>12</td>
<td>Biosynthesis Of Unsaturated Fatty Acids</td>
<td>0.047</td>
<td>282</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Lysosome</td>
<td>0.046</td>
<td>1322</td>
<td>ATP6V0A4</td>
</tr>
<tr>
<td>14</td>
<td>Apoptosis</td>
<td>0.044</td>
<td>954</td>
<td>BCL2</td>
</tr>
<tr>
<td>15</td>
<td>Pathogenic Escherichia Coli Infection</td>
<td>0.041</td>
<td>538</td>
<td>ITGA1 FYN</td>
</tr>
<tr>
<td>16</td>
<td>Metabolism Of Xenobiotics By Cytochrome P450</td>
<td>0.039</td>
<td>880</td>
<td>MGST3</td>
</tr>
<tr>
<td>17</td>
<td>Drug Metabolism Cytochrome P450</td>
<td>0.038</td>
<td>910</td>
<td>MGST3</td>
</tr>
<tr>
<td>18</td>
<td>Autoimmune Thyroid Disease</td>
<td>0.037</td>
<td>686</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Focal Adhesion</td>
<td>0.034</td>
<td>4787</td>
<td>ITGA1 ITGA1 ITGA1 LAMA2 CACNA1C CACNA1D PK7</td>
</tr>
<tr>
<td>20</td>
<td>Leishmania Infection</td>
<td>0.034</td>
<td>718</td>
<td>PK7 ITGB1</td>
</tr>
<tr>
<td>21</td>
<td>* Ppar Signaling Pathway</td>
<td>0.032</td>
<td>800</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Rna Polymerase</td>
<td>0.031</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Lysine Degradation</td>
<td>0.030</td>
<td>423</td>
<td>ALDH7A1 ALDH2</td>
</tr>
<tr>
<td>24</td>
<td>Endocytosis</td>
<td>0.030</td>
<td>3436</td>
<td>EGFR IGF1R</td>
</tr>
<tr>
<td>25</td>
<td>Gycosaminoglycan Biosynthesis Chondroitin Sulfate</td>
<td>0.029</td>
<td>727</td>
<td>EGFR IGF1R</td>
</tr>
<tr>
<td>26</td>
<td>Melanoma</td>
<td>0.028</td>
<td>1189</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Nucleotide Excision Repair</td>
<td>0.028</td>
<td>330</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Prostate Cancer</td>
<td>0.026</td>
<td>1419</td>
<td>EGFR IGF1R</td>
</tr>
<tr>
<td>29</td>
<td>Renal Cell Carcinoma</td>
<td>0.026</td>
<td>1004</td>
<td>PK7</td>
</tr>
<tr>
<td>30</td>
<td>Glycine Serine And Threonine Metabolism</td>
<td>0.026</td>
<td>268</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.10: SiMES dataset: Top 30 SNPs and genes, respectively ranked by SNP and gene selection frequency. Genes falling in the top 30 ranks of the consensus gene set, $\Psi_{244}^{\text{gene}}$, obtained by comparing gene ranking results from both SP2 and SiMES datasets (see Table 4.13), are marked with a *.

<table>
<thead>
<tr>
<th>Rank</th>
<th>SNP</th>
<th>$\pi^{SNP}$</th>
<th>Mapped gene(s)</th>
<th>Gene</th>
<th>$\pi^{\text{gene}}$</th>
<th># mapped SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2636698</td>
<td>0.31</td>
<td>PPA2</td>
<td>PPA2</td>
<td>0.31</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>rs2726503</td>
<td>0.31</td>
<td>PPA2</td>
<td>PDSS2</td>
<td>0.26</td>
<td>59</td>
</tr>
<tr>
<td>3</td>
<td>rs2713829</td>
<td>0.31</td>
<td>PPA2</td>
<td>GABARAPL1</td>
<td>0.18</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>rs2636726</td>
<td>0.31</td>
<td>PPA2</td>
<td>ATP6V0A4</td>
<td>0.15</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>rs2713834</td>
<td>0.31</td>
<td>PPA2</td>
<td>ITGB1</td>
<td>0.13</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>rs2726471</td>
<td>0.31</td>
<td>PPA2</td>
<td>CACNA1C*</td>
<td>0.11</td>
<td>186</td>
</tr>
<tr>
<td>7</td>
<td>rs2636713</td>
<td>0.31</td>
<td>PPA2</td>
<td>PRKCB*</td>
<td>0.11</td>
<td>84</td>
</tr>
<tr>
<td>8</td>
<td>rs2726516</td>
<td>0.31</td>
<td>PPA2</td>
<td>FYN</td>
<td>0.11</td>
<td>46</td>
</tr>
<tr>
<td>9</td>
<td>rs2636739</td>
<td>0.31</td>
<td>PPA2</td>
<td>BCL2*</td>
<td>0.10</td>
<td>61</td>
</tr>
<tr>
<td>10</td>
<td>rs2713861</td>
<td>0.31</td>
<td>PPA2</td>
<td>PAK7*</td>
<td>0.10</td>
<td>127</td>
</tr>
<tr>
<td>11</td>
<td>rs2636751</td>
<td>0.31</td>
<td>PPA2</td>
<td>DGKB</td>
<td>0.10</td>
<td>233</td>
</tr>
<tr>
<td>12</td>
<td>rs2298733</td>
<td>0.30</td>
<td>PPA2</td>
<td>LAMA2*</td>
<td>0.10</td>
<td>118</td>
</tr>
<tr>
<td>13</td>
<td>rs6568474</td>
<td>0.26</td>
<td>PDSS2</td>
<td>NDUFA4</td>
<td>0.10</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>rs9386622</td>
<td>0.26</td>
<td>PDSS2</td>
<td>DGKH</td>
<td>0.10</td>
<td>70</td>
</tr>
<tr>
<td>15</td>
<td>rs6924886</td>
<td>0.26</td>
<td>PDSS2</td>
<td>ADCY2*</td>
<td>0.09</td>
<td>104</td>
</tr>
<tr>
<td>16</td>
<td>rs759440</td>
<td>0.25</td>
<td>PDSS2</td>
<td>LPC</td>
<td>0.09</td>
<td>69</td>
</tr>
<tr>
<td>17</td>
<td>rs1759792</td>
<td>0.23</td>
<td>PDSS2</td>
<td>SLC8A1*</td>
<td>0.09</td>
<td>240</td>
</tr>
<tr>
<td>18</td>
<td>rs10457161</td>
<td>0.20</td>
<td>PDSS2</td>
<td>EGFR*</td>
<td>0.09</td>
<td>74</td>
</tr>
<tr>
<td>19</td>
<td>rs12821011</td>
<td>0.18</td>
<td>GABARAPL1</td>
<td>PRKAG2</td>
<td>0.09</td>
<td>118</td>
</tr>
<tr>
<td>20</td>
<td>rs11053685</td>
<td>0.18</td>
<td>GABARAPL1</td>
<td>CACNA1D</td>
<td>0.09</td>
<td>83</td>
</tr>
<tr>
<td>21</td>
<td>rs4764324</td>
<td>0.18</td>
<td>GABARAPL1</td>
<td>ITGA11*</td>
<td>0.09</td>
<td>63</td>
</tr>
<tr>
<td>22</td>
<td>rs4764327</td>
<td>0.18</td>
<td>GABARAPL1</td>
<td>IGF1R*</td>
<td>0.09</td>
<td>100</td>
</tr>
<tr>
<td>23</td>
<td>rs9373924</td>
<td>0.18</td>
<td>PDSS2</td>
<td>SDHC</td>
<td>0.09</td>
<td>9</td>
</tr>
<tr>
<td>24</td>
<td>rs10845074</td>
<td>0.18</td>
<td>GABARAPL1</td>
<td>CACNA2D3*</td>
<td>0.08</td>
<td>294</td>
</tr>
<tr>
<td>25</td>
<td>rs9320215</td>
<td>0.17</td>
<td>PDSS2</td>
<td>RYR2*</td>
<td>0.08</td>
<td>221</td>
</tr>
<tr>
<td>26</td>
<td>rs10845073</td>
<td>0.17</td>
<td>GABARAPL1</td>
<td>ITGA1*</td>
<td>0.08</td>
<td>77</td>
</tr>
<tr>
<td>27</td>
<td>rs4946826</td>
<td>0.16</td>
<td>PDSS2</td>
<td>ALDH7A1</td>
<td>0.08</td>
<td>23</td>
</tr>
<tr>
<td>28</td>
<td>rs6938393</td>
<td>0.15</td>
<td>PDSS2</td>
<td>MGST3*</td>
<td>0.08</td>
<td>40</td>
</tr>
<tr>
<td>29</td>
<td>rs13202332</td>
<td>0.13</td>
<td>PDSS2</td>
<td>ALDH2</td>
<td>0.08</td>
<td>12</td>
</tr>
<tr>
<td>30</td>
<td>rs9480754</td>
<td>0.13</td>
<td>PDSS2</td>
<td>SDHB</td>
<td>0.08</td>
<td>13</td>
</tr>
</tbody>
</table>
Comparison of ranked pathway lists

We now consider the problem of comparing rankings (pathways, genes and SNPs) obtained for each dataset. To do this we require some measure of distance between each pair of lists. Ideally this measure should place more emphasis on differences between highly-ranked variables, since we expect the association signal, and hence agreement between the ranked lists, to be strongest there. By the same reasoning, we expect there to be little or no agreement between variables at lower rankings, where selection frequencies are low. Indeed a consideration of empirical and null selection frequency distributions (Figures 4.9 (b) and 4.13, top) suggests that only the very top ranked variables are likely to reflect any true signal, so that we would additionally like our distance metric to be able to accommodate consideration of the top-\(k\) variables only, with \(k < p\), where \(p\) is the total number of variables ranked in either dataset. One complication with top-\(k\) lists is that they are partial, in the sense that unlike complete \((k = p)\) lists, a variable may occur in one list, but not the other.

In order to consider this problem, we introduce the following notation. We denote the complete set of ranked predictors by \(L = \{1, \ldots, p\}\), and begin by assuming that all variables are ranked in both datasets. We denote the rank of each variable in list 1 by \(\tau(i), i = 1, \ldots, p\), so that \(\tau(5) = 1\) if variable 5 is ranked first and so on. The corresponding ranks for list 2 are denoted by \(\sigma(i), i = 1, \ldots, p\). A suitable metric describing the distance between two top-\(k\) rankings is the Canberra distance (Jurman et al., 2008),

\[
Ca(k, \tau, \sigma) = \sum_{i=1}^{p} \frac{\left| \min\{\tau(i), k + 1\} - \min\{\sigma(i), k + 1\} \right|}{\min\{\tau(i), k + 1\} + \min\{\sigma(i), k + 1\}}.
\] (4.24)

This has the properties that we require, in that the denominator ensures more emphasis is placed on differences in the ranks of highly ranked variables in either dataset. Furthermore, this distance measure allows comparisons between partial, top-\(k\) lists, since a variable occurring in one top-\(k\) list but not the other is assigned a ranking of \(k + 1\) in the list from which it is missing. Note also that a variable \(i\) that is not in either of the top-\(k\) ranks, that is \(\tau(i), \sigma(i) > k\), makes no contribution to \(Ca(k, \tau, \sigma)\).

In order to gauge the extent to which the distance measure (4.24) differs from that
expected between two random lists, we require a value for the expected Canberra distance between two random lists, which we denote \( E[Ca(k,p)] \). Jurman et al. (2008) derive an expression for this quantity, and we use this to compute the normalised Canberra distance,

\[
Ca^*(k, \tau, \sigma) = \frac{Ca(k, \tau, \sigma)}{E[Ca(k,p)]}.
\] (4.25)

Note that this has a lower bound of 0, corresponding to exact agreement between the lists. For two random lists, the upper bound will generally be close to 1, although it can exceed 1, particularly for small \( k \), since the expected value for random lists is not necessarily the highest value.

We illustrate the variation of the normalised Canberra distance (4.25) between SP2 and SiMES pathway rankings in the left hand plot in Figure 4.14 (blue curve). We consider all possible top-\( k \) lists, \( k = 1, \ldots, 185 \) since all 185 pathways are ranked in both datasets. In the same plot, we also show

\[
Ca^*_{\pi}(k, \tau, \sigma) = \frac{1}{Z} \sum_{\pi=1}^{Z} Ca(k, \tau, \sigma^\pi) E[Ca(k,p)]
\] (4.26)

obtained by comparing empirical SP2 rankings (\( \tau \)) against \( Z = 10,000 \) permutations of the SiMES pathway rankings, \( \sigma^\pi, \pi = 1, \ldots, 10,000 \) (green curve). This latter curve confirms that the expected value, \( E[Ca(k,p)] \), is indeed a good measure of \( Ca \) in the random case where there is no agreement between rankings.

Using the same permuted rankings, \( \sigma^\pi \), we next test the null hypothesis that the observed normalised Canberra distance, \( Ca^*(k, \tau, \sigma) \), is not significantly different from that between \( \tau \) and a random list \( \sigma^\pi \), by computing a p-value as

\[
p^*(k) = \frac{1}{Z} \sum_{\pi=1}^{Z} I_{Ca^*(k, \tau, \sigma) \leq Ca^*(k, \tau, \sigma^\pi)},
\]

for \( k = 1, \ldots, 185 \). We then obtain FDR q-values using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and illustrate these for each \( k \) in the right hand plot of Figure 4.14. FDR is controlled at a nominal 5% level for \( 19 \leq k \leq 71 \), indicating that the distance between the top-\( k \) pathway rankings for both datasets is significantly different.
from the random ranking case for a wide range of possible values of \( k \). The distance \( C_a^* \) between SP2 and SiMES pathway rankings however attains its minimum value when \( k = 25 \) with \( q(25) = 0.037 \), so that on this measure, the two pathway rankings are in closest agreement when we consider the top 25 pathways in each ranked list only. Some intuitive understanding of why this might be so can be gained by considering the empirical vs. null pathway selection frequency distributions for each dataset in Figures 4.9 (b) and 4.13 (top). Here we see that the separation between empirical and null selection frequencies is most clear for values of \( k \) below around 30 for SP2, and around 15 for SiMES.

If we assume that the two pathway rankings are indeed in closest agreement when \( k = 25 \), then one means of obtaining a consensus set of important pathways is to consider their intersection,

\[
\Psi_{25}^{\text{path}} = \{ i : \tau^{-1}(i) \leq 25 \} \cap \{ j : \sigma^{-1}(j) \leq 25 \},
\]

Figure 4.14: Comparison of top-\( k \) SP2 and SiMES pathway rankings. Left hand plot: Variation of normalised Canberra distance, \( C_a^* \) with \( k \) (4.25) (blue curve). Corresponding mean values over \( Z = 10,000 \) permutations of SiMES rankings (4.26) (green curve). Right hand plot: FDR \( q \)-values (blue curve). Dotted green line shows the threshold for FDR control at the 5% level.
from which we can obtain a set of average rankings as

\[
\psi_{25}^{path} = \left\{ \frac{\tau(z) + \sigma(z)}{2} : z \in \Psi_{25}^{path} \right\}.
\]

Both the intersection set, \(\Psi_{25}^{path}\), and ordered average rankings, \(\psi_{25}^{path}\), for the two datasets under consideration are shown in Table 4.11. We additionally mark the consensus set \(\Psi_{25}^{path}\) with asterisks in Tables 4.6 and 4.9.

Table 4.11: Consensus set of pathways, \(\Psi_{25}^{path}\), for SP2 and SiMES datasets with \(k = 25\). Consensus pathways are ordered by their average rank in \(\psi_{25}^{path}\).

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Average rank ((\psi_{25}^{path}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dilated Cardiomyopathy</td>
<td>4.5</td>
</tr>
<tr>
<td>Hypertrophic Cardiomyopathy</td>
<td>7.5</td>
</tr>
<tr>
<td>T Cell Receptor Signaling Pathway</td>
<td>11.0</td>
</tr>
<tr>
<td>Terpenoid Backbone Biosynthesis</td>
<td>11.0</td>
</tr>
<tr>
<td>Arrhythmogenic Right Ventricular Cardiomyopathy</td>
<td>12.0</td>
</tr>
<tr>
<td>Ribosome</td>
<td>13.0</td>
</tr>
<tr>
<td>Ppar Signaling Pathway</td>
<td>18.5</td>
</tr>
</tbody>
</table>

Comparison of ranked gene and SNP lists

A number of factors complicate the comparison of ranked gene and SNP lists across both datasets. Firstly, sets of mapped SNPs and genes differ slightly between the two datasets (see Table 4.2). Secondly, even if we consider only those variables mapped in both datasets, different, though overlapping sets of variables are ranked in each. Thirdly, ranked variables are not independent (Jurman et al., 2008). For example, genes may be grouped into pathways, so that a reordering of genes within a pathway might be considered less significant than a reordering of genes mapping to different pathways. Similarly a reordering of SNPs mapping to a single gene might be considered less significant than a reordering of SNPs mapping to different genes.

In order to compute a distance measure between pairs or ranked lists, we therefore make two simplifying assumptions. First, we consider only variables ranked in one or both
datasets. This seems reasonable, since we can necessarily only compile a distance measure from variables that are ranked in one or both datasets. Second, we assume that variables are independent. This makes our distance measure conservative, in the sense that it will treat all reordering of SNPs or genes equally, irrespective of any potential functional relationship between them.

With these assumptions in mind, we begin by denoting the set of all $p^*$ variables (genes or SNPs) that are ranked in \textit{either} dataset by $\mathcal{L} = \{1, \ldots, p^*\}$. We further denote the corresponding sets of ranked variables for SP2 and SiMES datasets by $\mathcal{L}_\tau$ and $\mathcal{L}_\sigma$ respectively. We then have the following set relations: $\mathcal{L}_\tau, \mathcal{L}_\sigma \subset \mathcal{L}$; $\mathcal{L}_\tau \neq \mathcal{L}_\sigma$; and $|\mathcal{L}_\tau| \neq |\mathcal{L}_\sigma|$.

We now extend the previous Canberra distance measure to encompass the above set relations. We begin, as before, by defining two ranked lists corresponding to the rankings of all the variables in $\mathcal{L}$ for each dataset, although this time we must account for the fact that not all variables in $\mathcal{L}$ are ranked in both. We denote SP2 rankings by $\tau(i), i = 1, \ldots, p^*$, where $\tau(i)$ is the rank of variable $i$ if $i \in \mathcal{L}_\tau$, and $\tau(i) = p^*$ otherwise. SiMES rankings are defined in the same way, and denoted by $\sigma(i), i = 1, \ldots, p^*$.

Applying this revised ranking scheme, we can then define a top-$k$ normalised Canberra distance (4.24) as
\begin{equation}
Ca^*(k; \tau, \sigma) = \frac{Ca(k; \tau, \sigma)}{E[Ca(k, p^*)]},
\end{equation}
for any $k \leq \min\{|\mathcal{L}_\tau|, |\mathcal{L}_\sigma|\}$. The restriction on $k$ follows from the fact that we cannot distinguish between top-$k$ rankings for all $k > \min\{|\mathcal{L}_\tau|, |\mathcal{L}_\sigma|\}$.

\textit{i. Gene rankings}

Information summarising the relationship between the two ranked lists of genes is given in Table 4.12.

We consider normalised Canberra distances, $Ca^*(k; \tau, \sigma)$, for $k = 1, \ldots, 500$ only, and plot these in Figure 4.15 (left, blue curve), along with $Ca^*_\pi(k; \tau, \sigma)$ (4.26) for $Z = 10,000$ permutations of the SiMES pathway rankings, $\sigma^\pi, \pi = 1, \ldots, 10,000$ (green curve). Once again this latter curve confirms that the expected value, $E[Ca(k, p^*)]$, is indeed a good measure of $Ca$ in the random case where there is no agreement between rankings. We also plot FDR $q$-values using the same procedure as described previously for pathways. FDR is
Table 4.12: Summary of genes analysed and ranked in SP2 and SiMES datasets.

<table>
<thead>
<tr>
<th></th>
<th>SP2</th>
<th>SiMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of genes mapped to pathways</td>
<td>4,734</td>
<td>4,751</td>
</tr>
<tr>
<td>number of genes mapping to both datasets</td>
<td></td>
<td>4,726</td>
</tr>
<tr>
<td>number of ranked genes ((</td>
<td>\mathcal{L}_\tau</td>
<td>,</td>
</tr>
<tr>
<td>number of genes ranked in either dataset ((\rho^*))</td>
<td></td>
<td>3,913</td>
</tr>
<tr>
<td>number of genes ranked in both datasets ((</td>
<td>\mathcal{L}_\tau</td>
<td>\cap</td>
</tr>
</tbody>
</table>

controlled at a nominal 5% level for all \(k > 13\) in the region tested \((1 \leq k \leq 500)\). The distance \(Ca^*\) between SP2 and SiMES pathway rankings attains its minimum value when \(k = 244\), so that on this measure, the two gene rankings are in closest agreement when we consider the top 244 pathways in each ranked list only.

Figure 4.15: Comparison of top-\(k\) SP2 and SiMES gene rankings, for \(k = 1, \ldots, 500\). Left hand plot: Variation of normalised Canberra distance, \(Ca^*\) with \(k\) (4.27) (blue curve), and corresponding mean values over 10,000 permutations of SiMES rankings (4.26) (green curve). Right hand plot: FDR \(q\)-values (blue curve). Dotted green line shows the threshold for FDR control at the 5% level.

Following the same strategy as implemented for pathways, we then form the consensus set, \(\Psi_{244}^{\text{gene}}\), and average rankings \(\psi_{244}^{\text{gene}}\). The consensus set contains 84 genes, and we list the top 30 genes ordered by their average rank in the two datasets, in Table 4.13.
Table 4.13: Top 30 consensus genes ordered by their average rank, $\psi_{244}^{\text{gene}}$.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene</th>
<th>Average rank ($\psi_{244}^{\text{gene}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LAMA2</td>
<td>9.0</td>
</tr>
<tr>
<td>2</td>
<td>ADCY2</td>
<td>11.0</td>
</tr>
<tr>
<td>3</td>
<td>CACNA1C</td>
<td>11.5</td>
</tr>
<tr>
<td>4</td>
<td>PRKCB</td>
<td>11.5</td>
</tr>
<tr>
<td>5</td>
<td>PRKCA</td>
<td>21.0</td>
</tr>
<tr>
<td>6</td>
<td>EGFR</td>
<td>21.5</td>
</tr>
<tr>
<td>7</td>
<td>ITGA1</td>
<td>24.5</td>
</tr>
<tr>
<td>8</td>
<td>CACNA2D3</td>
<td>25.5</td>
</tr>
<tr>
<td>9</td>
<td>RYR2</td>
<td>26.5</td>
</tr>
<tr>
<td>10</td>
<td>IGF1R</td>
<td>30.5</td>
</tr>
<tr>
<td>11</td>
<td>PAK7</td>
<td>36.5</td>
</tr>
<tr>
<td>12</td>
<td>ADCY8</td>
<td>37.5</td>
</tr>
<tr>
<td>13</td>
<td>VAV2</td>
<td>41.0</td>
</tr>
<tr>
<td>14</td>
<td>SLC8A1</td>
<td>41.5</td>
</tr>
<tr>
<td>15</td>
<td>CACNB2</td>
<td>42.5</td>
</tr>
<tr>
<td>16</td>
<td>CACNA2D1</td>
<td>43.0</td>
</tr>
<tr>
<td>17</td>
<td>ITGA9</td>
<td>44.0</td>
</tr>
<tr>
<td>18</td>
<td>KRAS</td>
<td>47.5</td>
</tr>
<tr>
<td>19</td>
<td>MAPK10</td>
<td>50.5</td>
</tr>
<tr>
<td>20</td>
<td>CACNA1S</td>
<td>51.0</td>
</tr>
<tr>
<td>21</td>
<td>VAV3</td>
<td>54.0</td>
</tr>
<tr>
<td>22</td>
<td>PLCG2</td>
<td>55.5</td>
</tr>
<tr>
<td>23</td>
<td>BCL2</td>
<td>57.0</td>
</tr>
<tr>
<td>24</td>
<td>CD80</td>
<td>60.0</td>
</tr>
<tr>
<td>25</td>
<td>ITGA11</td>
<td>60.5</td>
</tr>
<tr>
<td>26</td>
<td>CTNNA2</td>
<td>61.0</td>
</tr>
<tr>
<td>27</td>
<td>ALDH1B1</td>
<td>61.5</td>
</tr>
<tr>
<td>28</td>
<td>MGST3</td>
<td>63.0</td>
</tr>
<tr>
<td>29</td>
<td>NEDD4L</td>
<td>63.0</td>
</tr>
<tr>
<td>30</td>
<td>PRKAG2</td>
<td>66.0</td>
</tr>
</tbody>
</table>

**ii. SNP rankings**

Information summarising the relationship between the two ranked lists of SNPs is given in Table 4.14. In contrast to both pathway and gene rankings, it is apparent that relatively few
ranked SNPs overlap both datasets – 8,151 out of 41,452 SNPs that are ranked in either dataset. This results in values for $Ca^*(k)$ that are close to 1, corresponding to the random list case, over a wide range of possible values for $k$ (data not shown).

For this reason, we compute a simple summary measure

$$\psi^{SNP} = \left\{ \frac{T(j) + \sigma(j)}{2} : j \in L_\tau \cap L_\sigma \right\}$$  \hspace{1cm} (4.28)

and report only the top ranking SNPs using this measure in Table 4.15.

Table 4.14: Summary of SNPs analysed and ranked in SP2 and SiMES datasets.

<table>
<thead>
<tr>
<th></th>
<th>SP2</th>
<th>SiMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of SNPs mapped to pathways</td>
<td>75,389</td>
<td>78,933</td>
</tr>
<tr>
<td>number of SNPs mapping to both datasets</td>
<td>74,864</td>
<td></td>
</tr>
<tr>
<td>number of ranked SNPs ($</td>
<td>L_\tau</td>
<td>,</td>
</tr>
<tr>
<td>number of SNPs ranked in either dataset ($p^*$)</td>
<td>41,452</td>
<td></td>
</tr>
<tr>
<td>number of SNPs ranked in both datasets ($</td>
<td>L_\tau</td>
<td>\cap</td>
</tr>
</tbody>
</table>

4.4 Discussion

We began this chapter by outlining a strategy for pathways-driven SNP selection, using the sparse group lasso, that aims to maximise the power to select causal SNPs, possibly of low effect size, that might otherwise be missed if pathways information is ignored. In a simulation study we were then able to demonstrate that where causal SNPs are enriched within a single causal pathway, SGL does indeed have greater SNP selection power, compared to an alternative sparse regression model, the lasso, that disregards pathways information. These results mirror previous findings that support the intuition that a sparse selection penalty that promotes dual-level sparsity is better able to recover the true model in these circumstances (Friedman, Hastie, and Tibshirani, 2010; Simon et al., 2012).

We then argued from a theoretical standpoint that where SNPs can map to multiple pathways, a modification (SGL-CGD) of the SGL-BCGD estimation algorithm that treats pathways as independent, may offer greater sensitivity for the detection of causal SNPs and
Table 4.15: Top 30 SNPs ranked in both SP2 and SiMES datasets, ranked in order of mean ranking, $\psi^{SNP}$ (4.28).

<table>
<thead>
<tr>
<th>rank</th>
<th>SNP</th>
<th>$\psi^{SNP}$</th>
<th>mapped gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs897799</td>
<td>133.0</td>
<td>COX6B2 IL11</td>
</tr>
<tr>
<td>2</td>
<td>rs2126953</td>
<td>203.0</td>
<td>ITGA1</td>
</tr>
<tr>
<td>3</td>
<td>rs7714110</td>
<td>213.0</td>
<td>ADCY2</td>
</tr>
<tr>
<td>4</td>
<td>rs6924886</td>
<td>274.5</td>
<td>PDSS2</td>
</tr>
<tr>
<td>5</td>
<td>rs2447867</td>
<td>275.5</td>
<td>ITGA1</td>
</tr>
<tr>
<td>6</td>
<td>rs9386622</td>
<td>283.0</td>
<td>PDSS2</td>
</tr>
<tr>
<td>7</td>
<td>rs6568474</td>
<td>283.5</td>
<td>PDSS2</td>
</tr>
<tr>
<td>8</td>
<td>rs10446497</td>
<td>349.5</td>
<td>PAK2</td>
</tr>
<tr>
<td>9</td>
<td>rs6583177</td>
<td>385.5</td>
<td>PAK2</td>
</tr>
<tr>
<td>10</td>
<td>rs4765961</td>
<td>429.0</td>
<td>CACNA1C</td>
</tr>
<tr>
<td>11</td>
<td>rs759440</td>
<td>457.0</td>
<td>PDSS2</td>
</tr>
<tr>
<td>12</td>
<td>rs10457161</td>
<td>465.0</td>
<td>PDSS2</td>
</tr>
<tr>
<td>13</td>
<td>rs10462842</td>
<td>479.5</td>
<td>ADCY2</td>
</tr>
<tr>
<td>14</td>
<td>rs9373932</td>
<td>529.0</td>
<td>PDSS2</td>
</tr>
<tr>
<td>15</td>
<td>rs12206487</td>
<td>532.0</td>
<td>LAMA2</td>
</tr>
<tr>
<td>16</td>
<td>rs743567</td>
<td>543.0</td>
<td>MYH7</td>
</tr>
<tr>
<td>17</td>
<td>rs12472674</td>
<td>543.0</td>
<td>CTNNA2</td>
</tr>
<tr>
<td>18</td>
<td>rs11759792</td>
<td>557.5</td>
<td>PDSS2</td>
</tr>
<tr>
<td>19</td>
<td>rs9373924</td>
<td>566.0</td>
<td>PDSS2</td>
</tr>
<tr>
<td>20</td>
<td>rs319070</td>
<td>623.0</td>
<td>PDSS2</td>
</tr>
<tr>
<td>21</td>
<td>rs751877</td>
<td>630.0</td>
<td>ADCY4 LTB4R RIPK3</td>
</tr>
<tr>
<td>22</td>
<td>rs2047698</td>
<td>714.0</td>
<td>PDGFD</td>
</tr>
<tr>
<td>23</td>
<td>rs4804505</td>
<td>727.5</td>
<td>PDE4A KEAP1</td>
</tr>
<tr>
<td>24</td>
<td>rs12672417</td>
<td>764.0</td>
<td>SMURF1</td>
</tr>
<tr>
<td>25</td>
<td>rs7766689</td>
<td>800.5</td>
<td>LAMA2</td>
</tr>
<tr>
<td>26</td>
<td>rs157694</td>
<td>860.0</td>
<td>MAP3K7</td>
</tr>
<tr>
<td>27</td>
<td>rs554192</td>
<td>878.0</td>
<td>NEDD4L</td>
</tr>
<tr>
<td>28</td>
<td>rs2746543</td>
<td>896.5</td>
<td>SDHB</td>
</tr>
<tr>
<td>29</td>
<td>rs742257</td>
<td>942.0</td>
<td>LAMB3</td>
</tr>
<tr>
<td>30</td>
<td>rs1798619</td>
<td>944.5</td>
<td>PAK2</td>
</tr>
</tbody>
</table>
pathways. A potential concern is that this gain in power may be accompanied by an inflated number of false positives. However, in a simulation study with overlapping pathways we found that SGL-CGD demonstrated relative gains in both sensitivity and specificity, compared to SGL-BCGD. This gain in specificity was unexpected, and appears to arise directly from treating pathways as independent in the model estimation process. As with GL, the ability of SGL to recover the true model is likely to be affected by the complexity of the pathway overlap structure (Percival, 2012), although we expect that the gains in power and sensitivity achieved with the independence assumption will also be apparent with real data.

As with the methods introduced in preceding chapters, P-SGLAW combines the SGL model and SGL-CGD estimation algorithm with a weight-tuning algorithm to reduce selection bias, and a resampling technique designed to provide a robust measure of SNP, gene and pathway importance. As such, the latter is expected to confer advantages, in terms of the down ranking of unimportant predictors, previously observed for the lasso (Meinshausen and Bühlmann, 2010; Chatterjee and Lahiri, 2011). Once again it would be interesting to explore this further using simulations derived from real pathway and genotype data.

We do however develop a heuristic measure of ranking performance in our P-SGLAW application study identifying SNPs and pathways associated with high-density lipoprotein cholesterol levels (Section 4.3). By comparing empirical pathway and SNP rankings with null rankings obtained by permuting phenotype labels, we gain some confidence that pathway and SNP signals captured in the top rankings can be distinguished from those arising from noise or spurious associations. Interestingly, when the same comparison between empirical and null rankings is made with a reduced value for the regularisation parameter, $\alpha$, there is evidence of selection bias, in the sense that pathways and SNPs tend to be highly ranked both empirically and under the null. Since a smaller $\alpha$ corresponds to a greater number of SNPs being selected at each subsample, this would seem to suggest that too many SNPs are being selected. In this case, pathway and SNP rankings may in part reflect spurious associations, with a bias towards SNPs overlapping multiple pathways.

Turning to the study results, we conduct two separate analyses on independent discovery and replication datasets. Since subjects from both datasets are genotyped on the same platform, the large majority of SNPs mapping to pathways in one dataset do so also in the
other dataset. Thus 99.3% of SNPs mapping to pathways in the SP2 dataset are similarly
mapped in the SiMES dataset. For the SiMES dataset, the corresponding figure is 94.8%.
As expected, the concordance of gene coverage is even greater. Thus 99.8% of mapped
genes in the SP2 dataset are also mapped in the SiMES dataset, and 99.5% of mapped
genes in the SiMES dataset are also mapped in SP2.

On the assumption that similar patterns of genetic variation are likely to impact HDLC
levels in both cohorts, this gives us an opportunity to validate results from each dataset
by comparing rankings. Where there are common mechanisms affecting phenotypes in
both cohorts however, we would expect to observe the most concordance between the two
studies at the pathway level, followed by genes, and lastly SNPs. Indeed this increased
heterogeneity at the SNP, and to a lesser extent at the gene level is one motivation for
adopting a pathways approach in the first place (Holmans et al., 2009; Hirschhorn, 2009;
Cantor, Lange, and Sinsheimer, 2010).

We obtain consensus pathway and gene rankings by considering only the top $k$ ranks
in each dataset, with $k$ obtained as the value that minimises the distance between the two
rankings. We additionally derive a significance measure for each top-$k$ distance by compar-
ing empirical distances against a null distribution obtained by permuting ranks in one list.
We note that this can only be an approximation of the true null, since in reality rankings
for both datasets may be influenced by the extent to which genes and SNPs overlap multi-
ple pathways. However, some support for the reasonableness of this approximation can be
gained from our earlier analysis, showing that the correlation between empirical and null
pathway and SNP rankings is low, so that rankings under the null are indeed approximately
random.

Considering the consensus pathway rankings in Table 4.11, three out of the seven con-
sensus pathways (ranked 1, 2 and 5), are related to cardiomyopathy. These three pathways
are the only cardiomyopathy-related pathways amongst the 185 KEGG pathways used in
our analysis, so it is noteworthy that all three fall within the consensus pathway rankings.
The link between HDLC levels and cardiomyopathy is already well established (Ansell
et al., 2005; Gordon et al., 1989; Toth, 2005; Freitas et al., 2009; Gaddam et al., 2011).
Furthermore, numerous references in the literature also describe the links between lipid
metabolism and T cell receptor (consensus pathway ranking 3) and PPAR signaling (rank 7)
(Janes et al., 2000; Calder and Yaqoob, 2007; Staels et al., 1998; Bensinger and Tontonoz, 2008).

Turning to a consideration of the top 30 consensus genes and SNPs presented in Tables 4.13 and 4.15 (and see also pathway ranking tables 4.6, 4.9 and 4.11, and extended results in supplementary information). We found that many are enriched in one of several gene families:

1. L-type calcium channel genes, including \textit{CACNA1C}, \textit{CACNA1S}, \textit{CACNA2D1}, \textit{CACNA2D3} and \textit{CACNB2}

2. Adenylate cyclase genes, including \textit{ADCY2}, \textit{ADCY4} and \textit{ADCY8}

3. Integrin and laminin genes, including \textit{ITGA1}, \textit{ITGA9}, \textit{ITGA11}, \textit{LAMA2}, and \textit{LAMA3}

4. MAPK signaling pathway genes, including \textit{MAPK10} and \textit{MAP3K7}

5. Immunological pathway genes, including \textit{PAK2}, \textit{PAK7}, \textit{PRKCA}, \textit{PRKCB}, \textit{VAV2} and \textit{VAV3}

These genes are highly enriched in several high ranking pathways from both datasets. Notably, the focal adhesion pathway alone has 12 gene hits, as does the dilated cardiomyopathy pathway. Cardiomyopathy pathways as a whole have 30 genes hits (several of the genes overlap more than one cardiomyopathy pathway). 10 of these genes feature in the MAPK signaling pathway, while GnRH (8 genes), T and B cell receptor (8), calcium (7), ErbB (5), and Wnt signaling (4) pathways also contain several genes in the list. To elucidate the biological relevance of these gene families and the connections between them, we investigated their known functional links with cardiovascular phenotypes (not restricted to HDLC) by referencing the KEGG and Genetic Association (http://geneticassociationdb.nih.gov) databases.

\textit{Voltage dependent L-type calcium channel gene family}

The genes in this family encode the subunits of the human voltage dependent L-type calcium channel (CaV1). The $\alpha$-1 subunit (encoded by \textit{CACNA1C}, \textit{A1S}, \textit{A2D1} and \textit{A2D3} in our study) determines channel function in various tissues. CaV1 function has significant
4.4 Discussion

Impact on the activity of heart cells and smooth muscles. For example, patients with malfunctioning CaV1 develop arrhythmias and shortened QT interval (Splawski et al., 2004; Antzelevitch et al., 2007; Templin et al., 2011). Furthermore, CACNA1C polymorphisms have been associated with variation in blood pressure in Caucasian and East Asian populations by pharmacogenetic analysis. In 120 Caucasians, 3 SNPs in this gene were significantly associated with the response to a widely applied antihypertensive CaV1 blocker (Bremer et al., 2006). Kamide et al. (2009) also found that polymorphisms in CACNA1C were associated with sensitivity to an antihypertensive in 161 Japanese patients. The CaV1 β subunit encoding CACNB2 has also been associated with blood pressure (Levy et al., 2009).

This gene family was mapped to several pathways in our study, with the KEGG dilated cardiomyopathy pathway achieving highest rank both within individual datasets, and in the consensus pathway rankings. Dilated cardiomyopathy is the most common form of cardiomyopathy, and features enlarged and weakened heart muscles. Although high levels of serum HDLC lowers the risk of heart disease (Castelli, 1988; Toth, 2005), there is still no direct evidence that CaV1 is involved in HDLC metabolism.

Adenylate cyclase gene family

Three adenylate cyclases genes, ADCY2, ADCY4 and ADCY8 were highly ranked in our study. Currently, there are no reported associations of these genes with cardiovascular disease or lipid levels. Adenylate cyclase genes catalyse the formation of cyclic adenosine monophosphate (cAMP) from adenosine triphosphate (ATP), while cAMP serves as the second messenger in cell signal transduction. Note that ADCY2 is insensitive to calcium concentration, suggesting that any association of this gene family with HDLC levels may not be due to any interactions with the CaV1 gene family.

Among high ranking pathways, ADCY2 and ADCY8 feature in the dilated cardiomyopathy pathway. Although ADCY4 was not in the top 30 consensus genes, rs751877 in this gene was among the top 30 consensus SNPs.

Integrin and laminin gene families

We found 3 genes encoding integrin subunits in our study. Integrins hook to the extracellu-
lar matrix (ECM) from the cell surface, and are also important signal transduction receptors which communicate aspects of the cell’s physical and chemical environment (Nermut et al., 1988). Interestingly, laminins are the major component of the ECM, and are relevant to the shape and migration of almost every type of tissue. Both of these two families of genes are therefore highly relevant to the survival and shape of heart muscles. A recent GWAS conducted in a Japanese population confirmed a previous association between ITGA9 and blood pressure in European populations (Takeuchi et al., 2010).

Integrin family genes and LAMA2 were selected primarily within high-ranking cardiomyopathy, focal adhesion and ECM receptor signaling pathways, with once again the dilated cardiomyopathy pathway achieving the highest ranks. However, evidence for LAMA3 association is weaker, since it was not in the top 30 consensus genes, although a SNP from the LAMB3 (laminin β-3) gene was ranked 29 in the consensus SNP list.

**MAPK signaling pathway**

TAK1 (MAP3K7) and JNK3 (MAPK10) are kinases which regulate cell cycling. They activate or depress downstream transcription factors which mediate cell proliferation, differentiation and inflammation.

JNK activity has been associated with obesity in a mouse model, where the absence of JNK1 (MAPK8), a protein in the same family as MAPK10, protects against the obesity-induced insulin resistance (Hirosumi et al., 2002). The negative correlation between HDLC level and obesity has been well accepted (Howard, Ruotolo, and Robbins, 2003).

**Immunological pathways**

PAK (PAK2 and PAK7) genes feature in the high ranking T cell signaling pathway in both SP2 and SiMES datasets. PRKC (including PRKCA and PRKCB), along with VAV (VAV2 and VAV3) genes also feature in various high ranking immunological pathways including T cell signaling, Pathogenic Escherichia Coli Infection and Natural Killer Cell Mediated Cytotoxicity. Genes from all 3 of these families are frequently top ranked in these pathways.

PAK and VAV are activated by antigens, and regulate the T cell cytoskeleton, indicating a possible impact on T cell shape and mobility. In a candidate gene association analysis, PRKCA was reported to be associated with HDLC at a nominally significant level, but was
not significant after adjusting for multiple testing (Lu et al., 2008).

In summary, genes enriched in the above gene clusters and pathways may be relevant to heart muscle cell signal transduction, shape and migration, and may thus have functional relevance to the onset of cardiovascular diseases. Many highly ranked genes in our study are also involved in neurological pathways. For example polymorphisms in CACNA1C have been associated with bipolar disorder, schizophrenia and major depression (Ferreira et al., 2008; Moskvina et al., 2009; Green et al., 2010). This points to an interesting hypothesis that serum HDLC levels might be regulated not only by metabolism but also by neurological pathways, although the elucidation of any putative biological mechanism underlying such an association obviously exceeds the scope of this study.

Besides the gene families and associated pathways discussed above, a notable feature of the top 30 consensus SNP ranking results presented in Table 4.15 is the inclusion of 9 SNPs mapping to the PDSS2 gene. PDSS2 achieves its high ranking in the consensus SNP list, through its inclusion in the highly ranked terpenoid backbone biosynthesis pathway in SiMES, and this gene is in fact the second highest consensus ranking gene with this dataset. In contrast, this gene is ranked low (216) in the SP2 dataset, which explains why it fails to make the top 30 consensus gene rankings. PDSS2 encodes subunit 2 of prenyl diphosphate synthase, which determines the length of the isoprenoid chain of coenzyme Q10 (CoQ10) (Saiki et al., 2005). A deficiency in biosynthesis of CoQ10 has previously been associated with delayed motor development and abnormal renal function, with excess serum lipids (Sobreira et al., 1997).

Despite the well established links between lipid metabolism and PPAR signaling noted above, no genes in this high-ranked pathway fall in the top 30 gene rankings for either dataset (see Tables 4.6 and 4.15). This could be because the association signal in this pathway is more widely distributed, compared to other high ranking pathways, perhaps indicating heterogeneity in genetic causal factors within our sample, so that different genes and SNPs are highlighted in different subsamples. This would result in reduced gene selection frequencies. Also, genes that overlap multiple putative causal pathways are more likely to be selected in a given subsample, meaning that associated genes mapping to pathways with relatively few overlaps may have lower selection frequencies. This may be the case with genes in the PPAR signaling pathway, whose 63 genes map to an average $2.7 \pm 1.8$ path-
ways. As a comparison, the 84 genes in the top-ranked dilated cardiomyopathy pathway map to an average $7.2 \pm 3.8$ pathways.

Our study failed to highlight HDLC-associated SNPs identified in previous GWAS (see for example www.genome.gov/gwastudies for an up to date list). A primary reason for this is that the large majority of SNPs identified in previous studies do not map to pathways in our study, either because they fall in intergenic regions, or because they do not feature on the Illumina arrays used here. In addition our method is designed to highlight distributed, small SNP effects that accumulate across gene pathways, and so will likely fail to identify those SNPs with significant marginal effects targeted by GWAS.

As noted above, for those SNPs that are ranked in our study, we expect there to be considerably less concordance between SNP ranking results for the two datasets, compared with pathway and gene rankings, and this is indeed the case. This reduced concordance may be due to increased heterogeneity of genetic risk factors at the SNP level between the two datasets. Another important source of variation in SNP selection frequencies is LD between SNPs. The within-pathway lasso penalty will tend to select one of a group of highly correlated SNPs at random, reducing SNP selection frequencies within LD blocks harbouring causal SNPs. An alternative approach would be to consider a different penalty within selected pathways, for example the elastic net (Zou and Hastie, 2005), which selects groups of correlated variables jointly, although this comes at the cost of introducing a further regularisation parameter to be tuned.
Chapter 5

Sparse regression methods for pathway and SNP selection: Conclusions and further work

In the preceding chapters we have considered a range of strategies and models for the detection of pathways, genes and SNPs associated with univariate and multivariate quantitative traits.

For the modelling of a univariate quantitative trait, we developed a method, P-GLAW, that combines all SNPs and pathways together in a single penalised regression model, and imposes a group-sparsity constraint to drive pathway selection. We demonstrated through simulations over a wide range of scenarios, that this multivariate approach was better able to identify implicated pathways, compared to an alternative method based on combining univariate SNP statistics. We then extended this method to the case of a multivariate quantitative trait by applying a group lasso penalty in a sparse reduced-rank regression (sRRR) model previously applied to the problem of identifying SNPs. We used the resulting PsRRR method in an imaging genetic analysis modelling a high-dimensional multivariate phenotype characteristic of structural changes in the brains of patients with AD. Our study identified a number of pathways previously implicated in the disease, and in a follow up analysis we performed sRRR with a lasso penalty to identify important SNPs and genes within previously selected pathways. Here we were able to identify a number of SNPs and genes
Conclusions and further work

previously implicated in AD, alongside other putative candidates. Finally, returning to the case of a univariate trait, we explored a method, P-SGLAW, that enforces sparsity at both the pathway and SNP level, and is therefore able to perform simultaneous pathway and SNP selection. We demonstrated by simulation that where pathways information is important, this dual-level sparse selection model is better able to recover underlying true sparsity, compared to the simple lasso that disregards such information. We then argued that in the case of overlapping pathways, a modification to the estimation algorithm that assumes that pathways are independent may improve model selection performance, and again demonstrated this in a simulation study. Finally we applied our method to a study of pathways, SNPs and genes associated with high-density lipoprotein cholesterol levels.

We have highlighted various theoretical and practical issues and problems surrounding the development and implementation of each of these methods in previous chapters. Here we take a broader view, and focus on points that are relevant to the methods as a whole, as well as considering possible future areas of investigation.

Beginning with the models themselves, each can be broadly classified as enforcing either group-wise sparse (P-GLAW and PsRRR) or dual-level sparse (P-SGLAW) solutions to the problem of modelling quantitative traits with genetic predictors. The former makes the assumption that all SNPs within an implicated pathway may influence the phenotype, whereas the latter assumes that this influence is restricted to only a subset of all SNPs within a pathway. As such, for pathways selection the former might be expected to be more sensitive to widely distributed pathway signals than the latter. A further interesting contrast can be drawn between a competitive approach to the modelling of pathways (P-GLAW), as against a non-competitive approach (P-SGLAW) that treats each pathway as independent in the estimation algorithm. This distinction arises primarily because of the fact that pathways overlap. Again pathway selection performance with each approach is expected to vary, depending on the nature of the pathways signal, and in particular on the distributions of overlapping causal genes. We have argued that the non-competitive approach is best-suited to the detection of important SNPs. However this may be at the expense of pathway selection performance. A compromise may be to take the ‘two-step’ approach to SNP selection described in the PsRRR AD application study. Ultimately the choice of model may depend on which assumptions seem most appropriate to the data, as
well as the particular research question being investigated.

For the modelling of a multivariate trait, at least two outstanding issues with our proposed PsRRR method warrant further work. Firstly, we make the simplifying assumption that phenotypes are uncorrelated by setting the covariance matrix for the phenotype, \( \Gamma = I_Q \). This has previously been shown to be a reasonable assumption when working with high dimensional imaging genetic data (Vounou, Nichols, and Montana, 2010). However, while we describe a simulation study with a lasso based penalty on the phenotype in Section 3.4.3, a natural extension to the current model would be to impose a suitable regularisation penalty to account for structure in the phenotype, for example by using the elastic net, a ROI-based sparse group lasso, or other more complex hierarchical penalty (Jenatton et al., 2011). Secondly, in this work we have only considered the rank-1 solution for our PsRRR model. We therefore capture only the first latent factor summarising the relationship between genotypes and phenotypes. Vounou, Nichols, and Montana (2010) describe a procedure for capturing further latent factors by regressing out the effects of important predictors from each successive rank estimation. The situation here is complicated by a number of factors including the presence of overlapping SNPs, the typically large number of SNPs mapping to selected pathways, and the need to recalibrate weights. It would nonetheless be interesting to explore the possibility of estimating further ranks by regressing out the effects of pathways selected above a nominal threshold. Finally we note interesting work in imaging genetics has recently been done on combining imaging modalities to detect genetic associations with multi-modal phenotypes (Wang et al., 2012). This mirrors parallel theoretical and applied work on multi task learning using sparse representations (Kim and Xing, 2012; Yang, Kim, and Xing, 2009; Puniyani, Kim, and Xing, 2010), and offers a number of interesting avenues for the extension of the current model.

All the methods developed here are subject to potential bias, in the sense that where there is no association between genotypes and phenotypes, pathways may not be selected with an equal probability. Such biases reflect differences in the statistical properties of each pathway, for example variations in pathway size, number of genes, patterns of LD between SNPs, and correlations between pathways. A related problem discussed widely in the sparse modelling literature is the issue of consistency in variable selection, that is the ability to recover the true underlying sparsity. A number of methods have been proposed
to improve variable selection consistency for both the lasso and the group lasso, including
the use of adapted weights that reflect certain statistical properties of the predictor vari-
ables (Zou, 2006; Wang and Leng, 2008), and for the lasso, a method that applies small
perturbations to the regularisation parameter $\lambda$ (Meinshausen and Bühlmann, 2010). Our
bias correction strategy uses adapted weights, but is data-driven, on the grounds that it is
not clear \textit{a priori} which particular properties may be most relevant to the problem of bias.
We demonstrate through simulations that this strategy gives clear benefits in terms of group
selection consistency. Our iterative weight tuning procedure is however quite computa-
tionally intensive, and it would be interesting to compare our results with some of the proposed
alternatives, which may be easier to implement.

Another means of potentially improving consistency in model selection is through the
use of bootstrapping or resampling techniques, and these have been widely demonstrated
to provide a robust measure of variable importance in finite samples (Wang et al., 2010;
Chatterjee and Lahiri, 2011). This type of approach can also provide a solution to the dif-
ficult problem of choosing the correct regularisation, $\lambda$, where it has been demonstrated to
give superior model selection performance when compared to alternative methods such as
cross validation, for determining the correct amount of model sparsity (Meinshausen and
Bühlmann, 2010). We use a subsampling approach to rank variables, and demonstrate its
effectiveness for ranking pathways using P-GLAW in a simulation study. We expect these
advantages to extend to the case of dual level sparsity, although it would be interesting to
investigate this further using real data. We do not explore the issue of determining a selec-
tion frequency threshold for the control of false positives here. This has been investigated
in the context of lasso selection (Meinshausen and Bühlmann, 2010), but the direct applica-
tion of existing theory to the present case is not feasible, since overlapping pathways make
clear distinctions between causal and noise variables problematic. In the P-SGLAW applica-
tion study described in Section 4.3, we do however compare empirical pathway and SNP
selection frequency distributions against those obtained using permuted phenotype labels.
These comparisons give some indication of which, amongst the top-$k$ ranked variables, are
likely to be important. There may be scope for formalising this comparison by for example
comparing the shapes of empirical and null distributions or deriving an exact permutation
test.
There are other potentially interesting areas to explore with regard to the subsampling method used here. For example, standard approaches consider only the set of variables selected at each subsample, and ignore potentially relevant information captured in the coefficient estimates themselves. The use of this additional information would result in a set of ranked lists, one for each subsample, and the joint consideration of these lists has the potential to provide a more robust measure of variable importance, by taking account of the relative importance of each variable for each subsample. Various techniques for aggregating ranked, partial lists (including those that we used for comparing variable rankings between datasets in the study described in Section 4.3) have been proposed (Sculley, 2007; Kolde et al., 2012; Jurman et al., 2012), and we have begun to investigate these in the context of variable selection. For imaging genetic analysis using PsRRR with sparse selection of genotypes and phenotypes, it would also be interesting to consider joint selection probabilities for voxels and pathway selection. This has the potential to boost power by observing patterns of covariation between voxels (or voxel clusters) and pathways.

As with all PGAS techniques, our methods rest on the use of prior information on the functional interaction of genes within pathways. Our methods can however be easily adapted to accommodate other ways of grouping SNP data, for example using protein interaction networks (Wu, Feng, and Stein, 2010), or GO and other ontologies (Jensen and Bork, 2010).

Despite great efforts, pathway assembly is still in its infancy, and the relative sparsity of gene-pathway annotations reflects the fact that our understanding of how the majority of genes functionally interact is at an early stage. As a consequence, annotations from different pathways databases often vary (Soh et al., 2010), so that the choice of pathways database can impact PGAS results (Elbers et al., 2009; Cantor, Lange, and Sinsheimer, 2010). Results are also subject to bias resulting from SNP to gene mapping strategies, so that for example SNP to gene mapping distances will affect the number of unmapped SNPs falling within gene ‘deserts’ (Eleftherohorinou et al., 2009); SNPs may map to relatively large numbers of genes in gene rich areas of the genome; and the mapping of a SNP to its closest gene may obscure a true functional relationships with a more distant gene (Wang et al., 2009b). Indeed recent research from the ENCODE project indicates that functional elements may in fact be densely distributed throughout the genome (Bernstein et al., 2012;
Sanyal et al., 2012), and this information has the potential to radically alter future pathways analysis. These issues, together with the fact that PGAS methods are by construction designed to highlight distributed, moderate to small SNP effects, emphasise the point that pathways analysis should be seen as complementary to studies searching for single markers (Wang, Li, and Hakonarson, 2010). This observation is further reinforced in our PsRRR application study, where a number of genes previously implicated in AD do not map to known pathways in our study. These genes and their associated SNPs, many of which are well validated, are therefore excluded, so that a significant part of the known AD-associated genetic signal is missing.
Chapter 6

False positives in neuroimaging genetics using cluster size inference with Voxel-based morphometry data

While our focus in preceding chapters has been on multivariate methods, mass univariate modelling, whereby each genetic predictor is tested against each voxel, is still the prevailing paradigm in imaging genetics (Potkin et al., 2009a). As previously discussed, a major issue with this approach is the need to apply a severe correction for multiple testing, owing to the very large number of tests performed, with a resulting detrimental affect on power. One approach to reducing the scale of this correction is simply to reduce the number of tests. However, this generally requires assumptions to be made concerning candidate genes or brain regions, or involves the discarding of potentially informative data (see Section 1.3). An alternative strategy is to recognise that structure present in both imaging and genetic data means that the tests are not strictly independent. This means that the effective number of tests is reduced, rendering a Bonferroni-corrected threshold overly conservative.

In the genetic domain, multiple testing corrections that account for SNP-SNP dependencies are well established (Johnson et al., 2010). An important source of structure in the imaging domain arises from spatial dependencies between voxels, leading to localised patterns of widespread correlation. These correlations are introduced by the image acquisition process itself, by physiological signal not included in the model, by image resampling dur-
Cluster size inference

ing re-alignment, or by explicit smoothing applied in image pre-processing (Frackowiak et al., 2003). Random field theory (RFT) is a widely used multiple testing method for determining corrected significances while accounting for these spatial dependencies between voxels (Worsley et al., 1996b). RFT however rests on a number of assumptions regarding the underlying imaging data (Ashburner et al., 2006), and these correction techniques must therefore be tested on representative empirical data before their efficacy can be firmly established. In a recent study Meyer-Lindenberg et al. (2008) measured rejection (i.e. false positive) rates in an imaging genetic study using voxel-wise inference against a pre-selected set of ‘null’ SNPs considered to have no effect on brain structure or function. Subjects included patients with schizophrenia, as well as healthy controls. Gene effects on both brain structure (derived from structural MRI scans using voxel-based morphometry), and brain function (using fMRI response to cognitive tests) were considered. The study looked at rejection rates both across the brain as a whole, and for specific regions of interest. Significance thresholds were adjusted for comparisons across multiple voxels using both family-wise error (FWER) and false discovery rate (FDR) corrections. At a nominal significance level of 0.05, they found empirical rejection rates ranging from 0.2–4.1%, suggesting that for the methods studied, false positive rates are well controlled, and that inferences are if anything conservative.

Voxel-wise and cluster size tests

Within the MULM framework, a variety of different approaches have been used to identify significant signals (or ‘activation’) in the brain. In voxel-wise tests, differences in activation are assessed at each individual voxel, so that locations where there is a strong association between voxel intensity and a disease phenotype for example, are labelled as significant. In cluster size tests on the other hand, an arbitrary cluster-forming threshold is applied to define connected components, with significance then being assessed on the basis of the size of clusters of contiguous super-threshold voxels. Cluster size tests can be relatively more sensitive than voxel-wise tests for spatially extended signals (Friston et al., 1996; Poline et al., 1997; Moorhead et al., 2005), since they make use of the spatial nature of the signal and require a less severe multiple testing correction (there are always fewer clusters than voxels). This necessarily comes at the cost of reduced localising power, as rejecting the cluster
null hypothesis just implies that one or more voxels within the cluster are significant.

As with voxel-wise tests, cluster size tests must account for the fact that neighbouring voxels are correlated, even without any experimentally-induced activation. RFT is used to assign p-values to each cluster in the statistic image accounting for smoothness and search volume. However, RFT rests on a number of assumptions, and simulation studies have shown that the performance of this technique depends on the choice of cluster-forming threshold, and on the use of sufficiently smooth images (Hayasaka and Nichols, 2003).

Inferences using cluster size are also subject to confounding effects arising from ‘non-stationarity’—i.e., from local variations in noise smoothness. Under non-stationarity, even when there is no signal present clusters will be larger in ‘smoother’ regions, and smaller in ‘rougher’ ones. Methods that fail to take such local non-stationarity into account will provide unreliable inferences, with areas of extended smoothness producing large clusters and increased incidence of false positives, and conversely rough areas demonstrating greater incidence of false negatives (Hayasaka et al., 2004). One approach to tackling this problem is to adjust cluster sizes according to local smoothness using non-stationary RFT methods (Worsley, 2002).

Hayasaka et al. (2004) compared stationary and non-stationary RFT cluster size inference methods in the identification of activated areas using simulated and PET data sets. They found that the stationary RFT method was anticonservative\(^1\) under non-stationarity, but that the non-stationary RFT test performed well only for smooth images under high degrees of freedom. While that work suggested the use of corresponding nonparametric nonstationary cluster size permutation tests, here we are motivated to use parametric RFT in the imaging genetics context, as testing massive numbers of SNPs may make permutation tests impractical.

In this chapter we investigate rejection (type I error) rates for voxel-wise and cluster size neuroimaging genetic inference on a dataset comprising 181 MRI images and associated genotype information from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) online database. We follow a similar experimental design to that of Meyer-Lindenberg et al. (2008), although we restrict our analysis to genotypic effects on whole-brain structure,

\(^1\)Anticonservative tests produce p-values that are too small, giving rise to rejection (false positive) rates that are higher than the nominal (expected) rate for the test.
using voxel-based morphometry (VBM). We supplement our real data with simulated data evaluations to aid in the interpretation of the real VBM data results.

VBM is a technique used to derive quantitative information on the distribution of particular types of brain tissue, for example grey and white matter, from structural MRI scans. Voxel-wise images are first processed using a series of techniques, required for example to extract (‘segment’) the required feature, for example grey matter concentration, from the raw image; to correct for subject movement; and to correct for differences in large-scale brain morphology, thus enabling images taken from different subjects to be compared. Additional smoothing is also generally applied at the image pre-processing stage to improve the signal to noise ratio, and to decrease sensitivity to any misalignment between subjects (Ashburner and Friston, 2000; Frackowiak et al., 2003).

### 6.1 Imaging data

#### 6.1.1 ADNI images

181 T1-weighted 3D structural MRI scans from subjects with mild cognitive impairment (MCI) were obtained from the ADNI database (see Jack et al. (2008) for image acquisition details). SPM5’s unified segmentation and normalisation (http://www.fil.ion.ucl.ac.uk/spm/software/spm5) was used to obtain gray matter (GM) images in standard space, modulated to account for volume changes in the warping to the MNI atlas. Modulated GM images were smoothed with 6mm and 12mm Gaussian kernels. 12mm smoothing is the de-facto standard in VBM studies (Ashburner and Friston, 2000), and was used by Meyer-Lindenberg et al. (2008). A second set of images with 6mm smoothing enables the performance of RFT at relatively low smoothness to be assessed. A grey matter analysis mask was constructed by thresholding the mean grey matter image at 0.025. All pre-processing and smoothing was carried out using SPM5.

---

2Smoothing with a Gaussian kernel has the effect that each voxels intensity is a weighted average of surrounding local voxels, with the weighting following a Gaussian function
6.2 Genotype data

6.1.2 Simulated images

Stationary and non-stationary random images were generated using FSL (http://www.fmrib.ox.ac.uk/fsl/). 3D simulated images had the same voxel size (2mm³) as MNI-warped ADNI images, and were also masked with the real data analysis mask.

Stationary realisations were generated using white noise images convolved with 6mm and 12mm 3D Gaussian smoothing kernels. Non-stationary realisations were generated from white noise images smoothed with 3 different Gaussian kernels extending over distinct, adjacent 3D regions of the image volume. “6mm” non-stationary images were composed of a central region smoothed with a 9mm kernel, with intermediate and outer regions smoothed with 6 and 4mm kernels respectively. “12mm” images were made up of regions smoothed with 12, 8, and 18mm kernels (see Figure 6.1). Final images were smoothed with a 1.5mm kernel to eliminate discontinuities at the boundaries between different regions, resulting in final non-stationary smoothnesses of 4.3, 6.2 and 9.1mm FWHM for “6mm”, and 8.1, 12.1 and 18.1 for “12mm”. All images were truncated from a larger initial volume to avoid edge artefacts. Finally, to match the real data under consideration, only voxels within the real-image grey matter mask were used.

6.2 Genotype data

In their study of genetic effects on brain structure, Meyer-Lindenberg et al. (2008) selected 720 ‘null’ SNPs, found to have no significant association with disease phenotype (at the 5% level) in separate case-control and family-based analyses. The results of the subsequent neuroimaging genetic analysis were considered to set an upper bound on null rejection rates, since individual SNPs may still influence brain structure after all. To establish a lower bound, the authors repeated their analysis, but with the genotype-phenotype relationship removed by permuting genotypes across subjects. 4 such permutations were analysed. In the present study, 700 selected ‘null’ SNPs are used, with 10 subsequent permutations.

ADNI subjects’ genotype information, assayed using the Illumina 610-Quad BeadChip microarray, was obtained from the ADNI website. Each genotype file contains information pertaining to 620,901 SNPs and copy number variations (CNVs). 700 ‘null’ SNPs
Figure 6.1: Non-stationary image simulation. (a) Schematic illustrating extent of 3 different smoothness regions. (b) as (a) with ADNI image brain mask applied. (c) Realisation of non-stationary image with outer, middle and inner regions smoothed with 8, 12 and 18mm FWHM Gaussian smoothing kernels. (d) as (c) with final 1.5mm smoothing kernel and ADNI mask applied.
were selected as follows. Firstly, all CNVs were excluded and only SNPs from chromosome 3 were considered. Chromosome 3 was chosen since none of 4 prime candidate AD-associated genes (APOE, PSEN1, PSEN2 and SORL1 at the time of the study) are located on this chromosome. Of the remaining 39,928 SNPs, those with a minor allele frequency of less than 5% were excluded, as were any SNPs with a Bonferroni-corrected Hardy Weinberg equilibrium p-value < 0.05/700. From the remaining 18,285 SNPs, 700 uniformly-spaced (by rank order in position) SNPs were selected, in order to minimise any possible linkage disequilibrium effects. Finally, as per standard practice, an adjustment was made to those SNPs (310 in total) with low numbers (< 10) of homozygous alleles, merging the rare homozygous and heterozygous groups. This is to minimise any potential biasing effects in the regression, and is equivalent to fitting a dominant or recessive model at the SNPs in question.

6.3 Statistical inference

Voxel-wise and cluster size tests for association between genotype and grey matter intensity were performed under the General Linear Model (GLM) (Friston et al., 1995) using SPM5. Non-stationary tests were carried out using Hayasaka’s non-stationary toolbox for SPM (http://fmri.wfubmc.edu/cms/NS-General). The non-stationary toolbox corrects for expected variation in cluster size in non-stationary images under the null, using RFT (Hayasaka et al., 2004; Worsley et al., 1999). Note that we did not compare our results with standard permutation tests (Hayasaka et al., 2004). Permutation methods are guaranteed to be valid under the null hypothesis, and our prime motivation in this study was to address the accuracy of RFT methods in a large data (i.e. imaging genetics) setting where permutation might not be practical computationally.

For the real (ADNI) image dataset, genotype effects were measured by modelling modulated grey matter intensity as a response to SNP allele frequency, with subject age and sex as nuisance covariates. Each SNP was analysed separately, with SNP significance determined from t and F-tests, corrected for multiple comparisons using family-wise error and false discovery rate. For cluster size tests, cluster-forming thresholds, $\alpha_c$, of 0.001, 0.01 and 0.05 under both stationary and non-stationary assumptions were considered. Overall
rejection rates express the proportion of the 700 SNPs found to cause any significant activation. All tests were repeated a total of 10 times with genotype-phenotype labels permuted to remove any possible remaining association.

Equivalent tests on simulated images were conducted with the same SNP and covariate (age, sex) data, so that degrees of freedom for all tests were the same as those on the ADNI dataset. These tests were performed without permutation since there can be no association between genotype and phenotype with random images.

6.3.1 Non-stationary cluster size inference

While the RFT non-stationary cluster size test is described in detail elsewhere (Hayasaka et al., 2004), we review it again here to facilitate later discussion. Under the GLM, the intensity $y(v)$ at voxel location $v$ is expressed as a linear combination of regressors

$$y(v) = X\beta(v) + \epsilon(v)$$

(6.1)

where, in a study with $N$ subjects and $P$ regressors, $X$ is an $(N \times P)$ design matrix, $\beta(v)$ is a $P$-dimensional vector of parameters to be estimated, and $\epsilon(v)$ is an $N$-vector of error terms, assumed to be independent and normally distributed.

With cluster size tests, significant clusters are formed from contiguous voxels whose $t$ or $F$-statistic exceed a fixed cluster-forming threshold, $u_c$ (or equivalently, an uncorrected significance level $\alpha_c$ that uniquely determines $u_c$). Briefly, the non-stationary toolbox corrects for image non-stationarity by measuring the ‘smoothness’ at each voxel, a quantity that is related to the variance of the spatial partial derivatives of the model errors, $\epsilon$ in (6.1). From this a measure of image smoothness, measured in FWHM is obtained. FWHM refers to the ‘full-width at half-maximum’ of a Gaussian kernel required to smooth a random (white noise) image into equivalent smoothness of the data at hand; note that isotropy is not assumed, and FWHM is fully specified by ($\text{FWHM}_x$, $\text{FWHM}_y$, $\text{FWHM}_z$). A related quantity is the ‘resolution element’ or RESEL (Worsley, 2002), a ‘virtual’ voxel of size $\text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z$. The RESEL count $N_{\text{res}}$ is the number of RESELS that fit
into the search volume,

\[ N_{\text{res}} = \frac{V}{\text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z} \]  \hspace{1cm} (6.2)

where \( V \) is the number of voxels in the image. When stationarity (i.e. uniform smoothness across the image) is assumed, FWHM is calculated by pooling FWHM across the entire image volume\(^3\). Under non-stationarity, FWHM is estimated at each voxel \( v \), giving a RESEL measure as well. The size of this local RESEL, \( 1/[\text{FWHM}_x(v) \times \text{FWHM}_y(v) \times \text{FWHM}_z(v)] \), is denoted \( \text{RPV}(v) \) for RESELS per voxel at voxel \( v \). In this way a voxel’s effective volume, relative to image smoothness, is obtained. The next step is to calculate the smoothness-adjusted cluster size, \( S' \), by summing effective voxel volumes over a cluster:

\[ S' = \sum_{v \in C} \text{RPV}(v) \]

where \( C \) denotes the set of voxel indices in the cluster. This procedure is equivalent to measuring cluster size in a distorted image, where space has been warped in such a way so as to ensure that stationarity holds (Worsley et al., 1999).

Finally, the probability of obtaining clusters of a given size \( S' \) under the null is calculated, corrected for multiple comparisons. This probability is derived from the image’s Euler Characteristic, \( \rho(\alpha_c) \), a topological property which approximates the expected number of clusters or ‘blobs’ in a thresholded image of given smoothness. In the stationary case, the expected cluster size under the null is

\[ E(S) = \frac{E(N_v)}{E(C)} \]  \hspace{1cm} (6.3)

where \( E(N_v) \) is the expected number of suprathreshold voxels (= \( V\alpha_c \)), and \( E(C) \) is the expected number of clusters (= \( V\rho(\alpha_c) \)). This expression also holds for \( S' \) but suprathreshold voxels must be measured in RESELS, i.e. \( E(N_v) = N_{\text{res}}\alpha_c \). The expected cluster size is then used to estimate the null distribution of \( S \) (or \( S' \)) and obtain uncorrected p-values.

---

\(^3\)Precisely the roughness measure is pooled and then converted to FWHM in order to minimise bias; see (Worsley et al., 1999) for details.
which are then converted to either FWER-corrected p-values or FDR-corrected p-values (Chumbley and Friston, 2009) that account for searching the brain for significant clusters.

The use of RFT in cluster size tests rests on a number of assumptions (Hayasaka and Nichols, 2003). These include:

- Lattice approximation - images are assumed to be derived from a smooth (continuous) random field sampled at regular points on a lattice; sampling is assumed to be fine enough to capture the local features of the field

- Image smoothness - images are smooth at the voxel scale; equivalently, total search volume is small compared to the size of a resel

- Uniform smoothness (for stationary tests only)

- High cluster-forming thresholds - RFT’s estimate of cluster size distribution under the null is derived asymptotically, under the assumption that the cluster-forming threshold $\alpha_c$ is sufficiently high

These assumptions present particular practical difficulties for those using cluster size tests, since low thresholds with as little smoothing as possible - the very conditions under which RFT performs worst - tend to maximise sensitivity and localising power (Hayasaka and Nichols, 2003).

### 6.4 Results

#### 6.4.1 ADNI images

Full cluster and voxel-wise results with FWER correction are presented in Table 6.1. Results with FDR correction are presented in Table 6.2. Relevant whole brain, voxel-wise rejection rates reported by Meyer-Lindenberg et al. (2008) are also included for comparison.

Results from tests with permuted genotype-phenotype labels (FWER corrected results only) were broadly similar to those with observed, unpermuted labels, indicating that for
### 6.4 Results

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Permuted&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Observed</th>
<th>Permuted&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Observed</th>
<th>Permuted&lt;sup&gt;3&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>t-tests</strong></td>
<td></td>
<td>6mm smoothing</td>
<td>12mm smoothing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td></td>
<td>10.7</td>
<td>23.3</td>
<td>47.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.2±1.1</td>
<td>24.8±2.2</td>
<td>46.7±1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.4</td>
<td>9.4</td>
<td>20.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.6±0.7</td>
<td>9.4±1.5</td>
<td>18.6±1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-stationary</td>
<td></td>
<td>0.001</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0</td>
<td>19.9</td>
<td>45.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.1±1.0</td>
<td>21.2±1.6</td>
<td>43.2±2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.9</td>
<td>9.1</td>
<td>20.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.8±0.8</td>
<td>9.3±1.5</td>
<td>18.4±1.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voxel-wise</td>
<td></td>
<td>-</td>
<td>3.4</td>
<td>2.7±0.6</td>
<td>1.9</td>
<td>1.1±0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.7±0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0</td>
<td>2.7±0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F-tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td></td>
<td>13.0</td>
<td>30.9</td>
<td>60.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.3±1.6</td>
<td>29.6±2.1</td>
<td>60.6±1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.4</td>
<td>11.9</td>
<td>26.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.8±1.1</td>
<td>10.7±1.7</td>
<td>24.5±2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-stationary</td>
<td></td>
<td>0.001</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.6</td>
<td>25.6</td>
<td>57.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.1±1.2</td>
<td>25.1±2.1</td>
<td>55.6±2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.7</td>
<td>11.4</td>
<td>25.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5±1.2</td>
<td>10.9±1.7</td>
<td>23.9±2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voxel-wise</td>
<td></td>
<td>-</td>
<td>3.6</td>
<td>2.6±0.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0</td>
<td>2.9±0.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>Mean rejection rate±SD across 10 permutations

<sup>2</sup>Results reported in Meyer-Lindenberg et al. (2008). These refer to whole-brain t-tests using structural (VBM) data with 12mm smoothing only

<sup>3</sup>Mean rejection rate±SD for 4 permutations

Table 6.1: FWER-corrected results - real (ADNI) images
Cluster size inference

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_c$</th>
<th>6mm smoothing</th>
<th>12mm smoothing</th>
<th>M-L et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$t$-tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cluster size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
<td>12.7</td>
<td>2.7</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>31.4</td>
<td>8.0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>51.6</td>
<td>17.7</td>
<td>-</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
<td>10.7</td>
<td>2.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>26.9</td>
<td>7.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>48.1</td>
<td>16.0</td>
<td>-</td>
</tr>
<tr>
<td>voxel-wise</td>
<td>-</td>
<td>3.3</td>
<td>1.9</td>
<td>1.8</td>
</tr>
<tr>
<td><strong>$F$-tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cluster size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
<td>15.9</td>
<td>2.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>41.1</td>
<td>11.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>75.7</td>
<td>25.4</td>
<td>-</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
<td>13.6</td>
<td>3.3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>35.9</td>
<td>11.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>70.3</td>
<td>24.4</td>
<td>-</td>
</tr>
<tr>
<td>voxel-wise</td>
<td>-</td>
<td>2.9</td>
<td>1.6</td>
<td>-</td>
</tr>
</tbody>
</table>

1Rejection rates for unpermuted data only were considered for FDR-corrected tests

Table 6.2: FDR-corrected results - real (ADNI) images
the purposes of the present study, chromosome 3 SNP effects on brain structure were negligible.

The key finding was that rejection rates were poorly controlled for all cluster size tests, except for those performed on 12mm smoothed images with the highest (most stringent) cluster-forming threshold, $\alpha_c = 0.001$. In this latter instance, FWER-corrected rejection rates approached the desired nominal 5% level, with a $3.8 \pm 0.8 \%$ rejection rate for a $t$-test with non-stationary correction under permutation, and $4.5 \pm 1.2 \%$ under the corresponding $F$-test. FDR-corrected results were broadly similar to FWER-corrected results for $\alpha_c = 0.01$ and $0.001$.

FWER and FDR-corrected voxel-wise tests were conservative for both 6 and 12mm smoothed images, in agreement with results for FWER-corrected voxel-wise $t$-tests on 12mm smoothed images reported by Meyer-Lindenberg et al. (2008).

In general, cluster size tests became more anticonservative at lower thresholds (decreasing $u_c$, increasing $\alpha_c$), and this effect was exacerbated for low smoothness images. Image smoothness had a pronounced effect on all results, with tests performed on 6mm smoothed images having substantially higher rejection rates than those performed on images with 12mm smoothing. The degree of smoothing, however, showed little effect on voxel-wise rejection rates.

Cluster size tests corrected for image non-stationarity were generally closer to nominal than those assuming stationarity. Finally, $F$-tests were generally more anticonservative than equivalent $t$-tests.

### 6.4.2 Simulated images

Rejection rates for tests on simulated, random Gaussian images are presented in Table 6.3. For stationary (constant smoothness) 6 and 12mm FWHM Gaussian images, both stationary and non-stationary cluster size $t$-tests are highly conservative at higher thresholds ($\alpha_c = 0.001, 0.01$), but are anti-conservative at the lowest threshold ($\alpha_c = 0.05$). $F$-tests are conservative at all thresholds. As with stationary images, non-stationary cluster size $t$-tests are conservative at $\alpha_c = 0.001, 0.01$, and anti-conservative at $\alpha_c = 0.05$, whereas $F$-tests are conservative at all thresholds. As might be expected, stationary cluster size $t$
% Rejection Rates

<table>
<thead>
<tr>
<th></th>
<th>6mm smoothing</th>
<th></th>
<th>12mm smoothing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_c$</td>
<td>Stationary</td>
<td>Non-stationary</td>
<td>Stationary</td>
</tr>
<tr>
<td><strong>t-tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cluster size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
<td>0.5</td>
<td>5.5</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>1.0</td>
<td>11.7</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>13.0</td>
<td>35.9</td>
<td>6.9</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
<td>0.6</td>
<td>1.1</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.9</td>
<td>0.3</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>9.3</td>
<td>10.7</td>
<td>6.4</td>
</tr>
<tr>
<td>voxel-wise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FWE</td>
<td>3.7</td>
<td>3.6</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>4.7</td>
<td>4.0</td>
<td>2.9</td>
</tr>
<tr>
<td><strong>F-tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cluster size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
<td>0.3</td>
<td>5.6</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.4</td>
<td>10.7</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>2.1</td>
<td>20.3</td>
<td>2.4</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
<td>0.4</td>
<td>0.7</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.4</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1.1</td>
<td>1.4</td>
<td>2.2</td>
</tr>
<tr>
<td>voxel-wise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FWE</td>
<td>2.6</td>
<td>2.7</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>2.0</td>
<td>3.4</td>
<td>1.4</td>
</tr>
</tbody>
</table>

1 Images constructed from concentric regions smoothed with 4, 6 and 9mm Gaussian kernels
2 As above with 8, 12 and 18mm Gaussian kernels

Table 6.3: Results - simulated images
and $F$-tests on both “6mm” and “12mm” FWHM non-stationary images perform poorly.

Voxel-wise tests are generally conservative, or close to nominal for both stationary and non-stationary images at 6 and 12mm.

6.5 Discussion

This study provides the first analysis of false positive rates in an imaging genetics study of VBM data using cluster size inference. Images from a group of 181 subjects with mild cognitive impairment were tested against a set of 700 ‘null’ SNPs. The analysis presented here suggests that rejection rates under both stationary and non-stationary assumptions are poorly controlled at low cluster-forming thresholds or for images with low smoothness.

Null SNPs were selected from chromosome 3, with the simple rationale that none of the genes reported to have the strongest link with AD at the time of this study were present on this chromosome. Since this is a somewhat crude measure for selecting SNPs with no effect on grey matter distribution, we tested this assumption by comparing our results with those obtained with permuted data, so that any possible SNP effects were removed through breaking the association between genotype and phenotype. We found that rejection rates obtained using permuted SNPs were not significantly different from those obtained without permutation (considering a 95% confidence interval at $\pm 2$ sd), indicating that, for the purposes of this study, SNP effects on brain structure are indeed negligible.

We begin by considering the results obtained with the ADNI image dataset.

*Effect of cluster-forming threshold, $\alpha_c$*

The choice of cluster-forming threshold, $\alpha_c$ was found to have a significant effect on cluster size inference rejection rates. For images smoothed with a 12mm Gaussian kernel, both stationary and non-stationary tests were found to be well-controlled or conservative at the most stringent threshold ($\alpha_c = 0.001$). However, tests became increasingly anticonservative at lower thresholds $u_c$ (higher $\alpha_c$) for both 12mm and 6mm smoothed images.
Figure 6.2: Accuracy of estimation of \( c \), the theoretical number of clusters under RFT. Histograms show the empirical distribution of \( c \) across all 700 SNPs at three different cluster-forming thresholds (left to right), and with two different smoothing kernels (top and bottom). The theoretical (RFT) and empirical mean number of clusters, \( E(C) \), are shown as dashed and dotted lines respectively. The amount by which RFT overestimates \( E(C) \) increases as the cluster-forming threshold \( u_c \) is lowered, and with images of lower smoothness. (Note that the x-axis for 6mm smoothed images has a larger range, reflecting the fact that many more clusters are observed).
A possible explanation for the poor performance at low $u_c$ is bias in RFT’s estimate of the expected number of clusters, $E(C)$ (Figure 6.2). If $E(C)$ is over-estimated, expected cluster size is under-estimated (see Eqn. (6.3)), meaning that more clusters of a given size are labelled as significant. This over-estimation of $E(C)$ may reflect the inability of the Euler Characteristic, $\rho(\alpha_c)$, to accurately estimate the number of clusters at low thresholds, where clusters are more numerous and tend to coalesce to form topologically complex patterns (Taylor and Worsley, 2008).

12mm vs. 6mm smoothing kernels

The application of a wide range of Gaussian smoothing kernels in VBM is evident in the literature — e.g. 4mm (Schwartz et al., 2010), 8mm (Folley et al., 2010) and 10mm (Shen et al., 2010), as well as the ‘standard’ 12mm (Rosen et al., 2010; Ueda et al., 2010). However guidelines on the particular choice of smoothing kernel have been described as ‘vague’ (Hayasaka and Nichols, 2003), and there is a suggestion that kernel widths should be determined empirically (Worsley et al., 1996a). Notably, with the use of high-dimensional warping methods like DARTEL (Ashburner, 2007), there appears to be trend towards lower smoothing kernels. Improved intersubject alignment means there is a reduced need for smoothing to ‘blur out’ warping errors. For example, Bergouignan et al. (2009) use 12mm smoothing with SPM’s standard normalisation and 8mm with DARTEL.

In the present study, differing amounts of smoothing were found to have a pronounced effect on rejection rates. Tests on images smoothed with a 6mm Gaussian kernel were highly anticonservative at all thresholds including the highest ($\alpha_c = 0.001$), and were consistently more anticonservative when compared with 12mm smoothing results.

Poor performance for low smoothness images is in fact to be expected under the lattice assumption of random field theory (Hayasaka and Nichols, 2003). As image smoothness decreases, this lattice approximation breaks down, since the underlying variation is poorly-captured by discrete, voxel-wise sampling. This means that continuous RFT results are modelling unobserved, large intensity changes between sampled voxels. While previous reports have suggested 3 voxel FWHM smoothing (i.e. 6mm FWHM smoothing for the 2mm voxels considered here) is sufficient (Nichols and Hayasaka, 2003), for the ADNI data this is insufficient, as we find an over-estimation of the expected number of clusters,
with the gap between expected and observed values, $E(C) - C$, generally greater at 6mm than at 12mm (see Figure 6.2).

**Stationary vs. non-stationary tests**

Non-stationary cluster-wise rejection rates were generally similar, or slightly better-controlled than those assuming stationarity, suggesting that there is at least some non-stationarity present in the images. For non-stationary images, stationary tests would also be expected to perform worse at lower thresholds where clusters are larger and more likely to encompass extra-smooth regions, and this is indeed the case. A heuristic measure of image non-stationarity was obtained by plotting the distribution of voxel-wise FWHM, obtained from the RPV image produced by SPM (FWHM=RPV$^{-1/3}$). A completely stationary image would be expected to have constant FWHM across the entire image volume. Any pronounced departure from this suggests non-stationarity. An analysis of 12mm and 6mm FWHM images (see Figure 6.3) finds a spread of around 7mm to 17mm and SD of 2.6mm for 12mm smoothed images, (4mm to 8mm and SD of 1.0mm for 6mm images).

![Figure 6.3: Distribution of voxel-wise FWHM for ADNI images smoothed with 6mm (left) and 12mm (right) Gaussian smoothing kernels. Voxel-wise FWHM gives an indication of local smoothness and corresponds to the ‘full-width at half-maximum’ of a Gaussian kernel required to produce a random (white noise) image of equivalent smoothness. A perfectly stationary image would have constant FWHM at all voxels. In contrast, a highly non-stationary image would have a large spread in FWHM, as is seen here.](image)
While this spread of FWHM could be attributed to sampling variation, the theoretical SD of the FWHM estimator can be computed by simulation (see Appendix B of Hayasaka et al. (2004)). We find theoretical SDs of 0.696mm for 12mm smoothed stationary images, and 0.348mm for 6mm images, which are much smaller than our observed values. While these theoretical SDs under stationarity again depend on the accuracy of the RFT results, they provide further evidence of substantial image non-stationarity.

**T vs. F image results**

The $t$ and $F$ image cluster size results cannot be directly compared. While the single degree-of-freedom $F$ test we used is exactly equal to the square of the $t$ test used, the set of clusters generated will be different for two reasons. First, the one-sided $\alpha$ level used to define a $t$ statistic threshold will not equal the square root of the $F$ statistic threshold of the same $\alpha$ level (an $F$’s level corresponds to the $t$’s two-sided alpha level). Further, the $F$ image has the clusters arising from negative $t$ values. Thus there will be both more and different clusters in the $F$ images for the same data and $\alpha_c$.

These caveats aside, the rejection rates on the real data were largely similar for the same $\alpha_c$’s, with valid performance found only for 12mm smoothed data with $\alpha_c = 0.001$.

**Simulated images**

In marked contrast to tests performed on the ADNI image dataset, non-stationary cluster size tests on simulated stationary and non-stationary random images were found to be valid (conservative) at both high and moderate cluster-forming thresholds ($\alpha_c = 0.001, 0.01$), irrespective of image smoothness.

Other studies using simulated images produced from stationary and non-stationary, Gaussian random fields have also considered the effect of varying both the cluster-forming threshold and image smoothing kernel. With stationary simulated images, Hayasaka and Nichols (2003, Figure 2) found that cluster size tests were conservative over the same range of image smoothness with $\alpha_c = 0.001, 0.01$, in agreement with our results. Using similar non-stationary simulated data, Hayasaka et al. (2004) also found that non-stationary cluster size tests were conservative with images of low smoothness (comparable to our 6mm non-stationary images), and with 20 subjects, but only considered $\alpha_c = 0.01$. 
The large discrepancy in cluster size inference rejection rates between real and simulated image data over a range of thresholds and smoothing kernels suggests that there are features of the real VBM data that may be incompatible with the RFT method. This may for example be due to the inherent non-normality of VBM data, or to patterns of non-stationarity in real images that are more complex than those simulated here. Non-normality of VBM data has been reported before, but only when considering the accuracy of voxel-wise significance (Viviani et al., 2007; Salmond et al., 2002). This other work found that imbalanced group comparisons required 12mm FWHM smoothing to accurately control voxel-wise false-positives, though balanced group comparisons were accurate with smaller kernel sizes. As genotypes are rarely equally frequent, the imbalanced results are most relevant to this setting.

We performed a number of additional simulations in order to investigate the role of non-normality in cluster size inference. VBM data is hard bounded between 0 and 1, and modulated VBM nearly so. A Shapiro-Wilks test for normality at each voxel, using the spmd5beta diagnostic toolbox (http://www.sph.umich.edu/$\sim$sim$nichols/SPMd/) reveals that both 6mm and 12mm smoothed images are indeed highly non-normal. This deviation is particularly marked for 6mm images, with around 45% of voxels exceeding a nominal 5% Shapiro-Wilks threshold. In contrast, the stationary Gaussian noise-derived simulated images describe above show no significant deviation from normality. To test the effect of introducing non-normality to our simulations, we generated a set of images by first thresholding Gaussian noise images smoothed with 6mm and 8mm kernels, to produce ‘patchy’, binary images. These were then smoothed with 6mm and 12mm kernels to produce images with a range of deviations from non-normality that mimicked or exceeded the deviations from normality exhibited by the real VBM data, as measured with a Shapiro-Wilks test. Regression of these images against all 700 SNPs produced similar results to those described above, with conservative results at high and moderate cluster-forming thresholds with both 6mm and 12mm smoothing.

To test the effect of more complex patterns of non-stationarity, we segmented FWHM images derived from 6mm and 12mm smoothed ADNI images to produce a set of topologically complex masks corresponding to regions of high, medium and low ‘smoothness’. Non-stationary simulated images were then generated by filling each masked region
with differently smoothed Gaussian noise, as described in section 6.1.2. Once again, a full analysis produced conservative results, with rejection rates below a nominal 5% for $\alpha_c = 0.001, 0.01$ for both 6mm and 12mm smoothed images.

One final set of simulated images was produced by again generating complex, non-stationary FWHM-segmented masks, this time filled with non-normal, Gaussian noise-derived data, as described above. Rejection rates were again well controlled, in marked contrast to results obtained using real ADNI data.

### 6.6 Conclusion

We found that RFT nonstationary cluster size tests on real VBM data perform poorly at low cluster-forming thresholds and for images with low smoothness. In a second analysis with synthetic image data generated using Monte Carlo simulations, we found performance was instead excellent, if conservative. The contradictory results indicate there are features of the real VBM data that are incompatible with the RFT method.

We suggest two possible reasons for this difference in performance. First, as grey matter segmented data is hard bounded between 0 and 1, and modulated VBM data nearly so, the data may exhibit non-normality, violating a foundational assumption of the RFT method. Second, while we simulated nonstationarity, the pattern of nonstationarity observed in real VBM is substantially more complex (Figure 6.4). However, further tests using simulated images with both significant deviations from normality, and with more complex patterns of non-stationarity still produced conservative results, so that we were unable to find evidence that either aspect of real VBM data is responsible for the poor performance observed with real image data.

There are many ways to characterise deviations from normality in image data, and it may be that the VBM data deviates from normality in ways which we have been unable to capture in our simulations. The same is true of our attempts to model the true complexity of nonstationarity. Additionally, while RFT assumes that images can be warped to approximate stationarity, for VBM these hypothetical warps could be so convoluted so as to render the constituent approximations inaccurate.

Fortunately, an alternative to parametric, RFT-based cluster size inference is available
Figure 6.4: 12mm VBM image non-stationarity. The figure illustrates the variation in image smoothness measured in FWHM, derived from the SPM RPV image. There is a wide variation, ranging from 3.8 to 27.7mm.

— a nonparametric permutation test where the data itself is used to derive an empirical cluster size distribution under the null (Hayasaka et al., 2004). While this approach carries a greater computational burden, the false positive rates are exact (Hayasaka and Nichols, 2003), and the permutation approach should be reasonable for studies examining only a small number of SNPs.
Appendix A

Review of Existing Pathways Methods

In the following review we briefly describe a selection of existing PGAS analysis methods. We divide these into univariate-based methods that rely on SNP-wise association tests, and multivariate-based methods that depend on multivariate or multilocus modelling. The number of methods is now large and growing, particularly in the univariate category (see Wang, Li, and Hakonarson (2010); Fridley and Biernacka (2011); Khatri, Sirota, and Butte (2012) for reviews). We therefore select a small number of methods to give a sense of some of the issues involved. A short summary of key features for each method is given at the end in Table A.1.

A.1 Univariate-based methods

The majority of existing PGAS methods begin with multiple univariate tests of association, in which each SNP in the study is scored, for example using $\chi^2$ statistics in case-control studies, or $t$-statistics or p-values from regression-based tests in the case of quantitative traits. A wide variety of techniques are then used to combine univariate statistics into pathway scores. All methods employ strategies for dealing with the problem of differing correlation structures (LD) within genes and pathways, and also with variations in gene and pathway size. Without a correction for LD, signals within pathways containing large groups of highly correlated SNPs, for example residing within large genes, will have inflated numbers of high-scoring SNPs, and so bias pathway scores. To account for this, most
methods compare observed pathway scores against null distributions formed from multiple permutations of phenotype labels, which has the added advantage of also correcting for variation in gene and pathway size.

For the combination of univariate SNP statistics, a group of methods build on the Gene-set Enrichment Analysis (GSEA) technique, first proposed for the pathway analysis of gene expression data (Subramanian et al., 2005). GSEA methods attempt to determine the degree to which high ranking SNPs or genes are over-represented in a given gene set or pathway, in comparison with a null dataset formed from multiple permutations of phenotype labels.

Consider for example the situation where $H$ genes are partitioned into $L$ groups or pathways, $G_l$, ($l = 1, \ldots, L$), each containing $H_l$ genes. The GenGen method (Wang, Li, and Bucan, 2007) first assigns a score to gene $W_k$, ($k = 1, \ldots, H$), by taking the highest statistic value (e.g. $\chi^2$) of all SNPs mapped to that gene. Gene scores are then sorted into a ranked list, $r(1), \ldots, r(H)$, where $r(1)$ is the gene score corresponding to the gene showing greatest association with the phenotype. An enrichment score, $ES(l)$ for pathway $l$, is then given by

$$ES(l) = \max_{1 \leq t \leq H} \left\{ \sum_{W_{k^*} \in G_l, k^* \leq t} \frac{|r(k^*)|^\gamma}{H_R} - \sum_{W_{k^*} \notin G_l, k^* \leq t} \frac{1}{H - H_l} \right\}$$  \hspace{1cm} \text{(A.1)}$$

where $H_R = \sum_{W_{k^*} \in G_l} |r(k^*)|^\gamma$. When $\gamma = 0$, the enrichment score depends only on gene rank, whereas scores are weighted by gene score when $\gamma = 1$. (A.1) is a Kolmogorov-Smirniv-type running sum that measures the extent to which high ranking genes are enriched in $G_l$, compared to a randomly selected set of genes. $ES(l)$ will be high when a large part of the association signal is to be found in genes in $G_l$.

As a next step, a correction for the influence of both LD and the number of SNPs mapped to each gene on pathway enrichment scores is obtained as follows. Calculation of the pathway enrichment score is repeated with phenotype labels permuted, to give a distribution of enrichment scores, $ES(l, \pi)$, ($\pi = 1, \ldots, N_\pi$), generated from $N_\pi$ permutations. From this, a normalised enrichment score $NES(l)$ is calculated as a Z score,

$$NES(l) = \frac{ES(l) - \text{mean}(ES(l, \pi))}{\text{SD}(ES(l, \pi))}$$

When multiple pathways are tested, a measure of pathway significance corrected for mul-
Multiple testing is obtained using an FDR procedure. The FDR for pathway $l^*$ is calculated as

$$p^{FDR}(l^*) = \frac{\% \text{ of all } (l, \pi) \text{ with } NES(l, \pi) \geq NES(l^*)}{\% \text{ of observed } l \text{ with } NES(l) \geq NES(l^*)}$$

where the numerator is the proportion of false discoveries, obtained by measuring the number of permuted datasets having an NES greater than that observed for pathway $l^*$; and the denominator is the proportion of results declared positive, i.e. the proportion of pathways having an NES greater than or equal to that observed for pathway $l^*$.

**GSEA-SNP** (Holden et al., 2008) follows a similar strategy to GenGen, this time using individual SNP scores, rather than gene scores.

A different approach is used in the **PLINK set-based test**, part of the popular PLINK genome data analysis toolset (Purcell et al., 2007). Here pathway scores are obtained by taking the mean, marginal p-value of a pre-determined maximum number of ‘independent’ SNPs within the pathway, defined as those whose pairwise LD is below a certain threshold value. Consider $P$ mapped SNPs, $j = 1, \ldots, P$, once again partitioned into $L$ groups or pathways, $G_l$, ($l = 1, \ldots, L$), each containing $P_l$ SNPs. A measure of significance for pathway $l$ is then obtained using the following algorithm:

1. rank SNP p-values, obtained from multiple univariate QTTs (see Section 1.1), in order of decreasing significance, $s_{(1)}, \ldots, s_{(P)}$, where $s_{(1)}$ is the SNP p-value corresponding to the SNP in pathway $l$ showing greatest association with the phenotype (i.e. having the smallest p-value)

2. form the set $G_l^* \subset G_l$, containing $P_l^*$ independent SNPs as follows:
   
   (a) let $D = G_l$ and $G_l^* = \emptyset$

   (b) add the top-ranked SNP, $j^*$ from $D$, to $G_l^*$. Denote the number of SNPs in $G_l^*$ by $N_l^*$

   (c) remove from $D$, SNP $j^*$, and all SNPs, $k$ whose pairwise LD, $r^2(j^*, k)$ passes a threshold $R$. Here $r^2(j^*, k)$ is the $r^2$ coefficient measuring correlation between SNPs $j^*$ and $k$ (Hedrick and Kumar, 2001)

   (d) repeat steps (b) and (c) until $N_l^* = P_l^*$
3. form the pathway statistic,
\[ u(l) = \frac{1}{F_l} \sum_{j^* \in G^*_l} s(j^*) \]

4. repeat steps 1. to 3. \(N_\pi\) times, each time permuting phenotype labels, to obtain a null distribution for the pathway statistic \(u(l, \pi), (\pi = 1, \ldots, N_\pi)\)

5. the empirical pathway p-value is then calculated as the proportion of all permutations with \(u(l, \pi) \leq u(l)\)

6. as a final step, where more than one pathway is tested, a correction for multiple testing can be applied

Other strategies for combining univariate SNP scores include the rank truncated product method proposed by Yu et al. (2009), which uses the truncated product of top-ranking univariate p-values to determine pathway scores; the SNP ratio test (O’Dushlaine et al., 2009), where the proportion of SNPs passing a predefined significance threshold within a given pathway is used to determine pathway significance, and the Aligator method (Holmans et al., 2009), which also sets a threshold for SNP significance, but defines as significant any gene containing at least one significant SNP. Eleftherohorinou et al. (2009) assess pathway significance by summing univariate SNP statistics across a pathway and comparing this with a parametised null distribution obtained through permutation. In a separate second stage, SNPs driving the pathway effect in significant pathways are selected using lasso logistic regression.

One interesting way to distinguish PGAS methods is by focusing on the type of null hypothesis being tested. With the PLINK set-based test for example, pathway significance is determined with reference to a null distribution formed from SNPs within the pathway under consideration. PGAS methods testing null hypotheses of this type can be classified as self-contained (Wang, Li, and Hakonarson, 2010), and are clearly useful in studies focusing on candidate genes and pathways. In contrast, with the GenGen method, \(ES(l)\) in (A.1) is calculated by comparing enrichment for genes within pathway \(l\), to other mapped genes not in \(l\). Such methods have been described as competitive, and this approach might be expected to give better control of false positives, although at the cost of lower power,
since the self-contained null hypothesis is more restrictive (Goeman and Buhlmann, 2007; Wang, Li, and Hakonarson, 2010).

Univariate-based PGAS studies have identified numerous pathways, previously implicated in a wide range of diseases, as well as providing new candidate pathways for future consideration. However, despite their obvious advantages for detecting small, coordinated signals in PGAS, such methods have a number of disadvantages and potential weaknesses. First of all, many methods require the generation of null distributions of univariate test statistics through permutation, making them computationally intensive for genome wide datasets. Also, permutation methods assume that the permuted units - subjects, SNPs or genes - are independent and identically distributed. SNP-SNP and gene-gene interactions, together with uncorrected relatedness between subjects may therefore introduce bias, depending on the techniques being used (Wang, Li, and Hakonarson, 2010). Each of the methods considered here uses different strategies for combining univariate statistics to assess pathways significance, so that each method may highlight different pathways depending on the distribution and effect sizes of causal SNPs within genes and pathways. Power may also suffer when testing multiple pathways due to the application of stringent multiple testing corrections.

A.2 Multivariate methods

In contrast to univariate-based methods, multivariate PGAS methods in which SNPs are modelled jointly have only recently begun to emerge. Potential benefits of a multivariate approach include the capacity to model covariates, and the prospect of increased power and better control of false positives, since small SNP effects may be revealed when other, stronger effects have been accounted for, and similarly false signals can be weakened by the inclusion of stronger signals from true causal associations (Hoggart et al., 2008).

Pathways of Distinction Analysis (Braun and Buetow, 2011) uses distance metrics between SNPs within pathways to assess whether cases are more similar to other cases than to controls, and vice versa. This multilocus approach directs attention away from SNPs exhibiting independent main effects towards a focus on epistatic interactions, whereby joint interactions play a much larger part in driving pathway significance.
Appendix A. Review of Existing Pathways Methods

The use of sparse regression models in PGAS is a relatively recent development. As with univariate-based tests, these must take variations in gene and pathway size, and LD when establishing pathway significance. Once again permutation methods are often employed to correct for this, along with various techniques to reduce correlation within genes or pathways prior to model fitting. Here we review a few examples, focusing on the regression models used, rather than on other aspects such as the different permutation methods employed for establishing significance.

Gene set ridge regression in association studies or GRASS (Chen et al., 2010) uses a novel combination of methods including penalised regression to determine pathway significance in case control studies. They begin by mapping groups of SNPs to genes, \( W_k \) \((k = 1, \ldots, H_l)\), in a single pathway of interest, \( G_l \). The dimensionality of each gene is then reduced by characterising within-gene variation as a set of orthogonal ‘eigenSNPs’, obtained by PCA. This is justified on a number of grounds including the desirability of reducing redundant information in a gene due to LD, and of reducing undesirable effects due to correlation that can arise when using lasso-based models (Zou and Hastie, 2005). Individual eigenSNP coefficients are obtained through the following objective function

\[
\hat{\beta} = \arg\min_\beta \left\{ L(\beta) + \lambda \sum_{k=1}^{H_l} (||\beta_k||_1)^2 \right\}
\]

(A.2)

where \( \beta \) is the vector of eigenSNP coefficients for all SNPs in \( G_l \) and \( \beta_k = (\beta_{k1}, \ldots, \beta_{kS_k}) \) is the parameter vector corresponding to eigenSNPs in \( W_k \), and \( L(\beta) \) is the logistic loss function (2.13). Here, the penalty promotes sparsity amongst eigenSNPs by applying a lasso penalty within each gene, while retaining all genes in the model through the imposition of a ridge penalty among genes. \( \lambda \) is determined using Akaike’s information criterion (AIC) (Akaike, 1974). Gene level association is estimated from \( ||\hat{\beta}_k||_2 \), from which a standardised value

\[
\tilde{\beta}_k = \frac{||\hat{\beta}_k||_2 - \mu_{\hat{\beta}_k}}{\sigma_{\hat{\beta}_k}}
\]

by permuting phenotype labels. Finally a pathway score is obtained as

\[
T = ||\tilde{\beta}||_2 = \sqrt{\tilde{\beta}_1 + \ldots + \tilde{\beta}_{H_l}}
\]
from which a pathway p-value is obtained as the proportion of permuted datasets with a $T$ score greater than the observed value. The optimisation (A.2) is described as a form of ‘group ridge’ regression, in that it affords smooth shrinkage of grouped eigenSNP predictors. Note that the assumption is that all genes in a pathway will contribute to the pathway association signal.

In contrast Zhao et al. (2011) impose an additional layer of sparsity at the gene level, on the assumption that this is not the case. Once again this method is implemented for case-control data with a logistic loss function. In a similar way to GRASS, the method begins by summarising genetic variation in each gene using PCA, but this time only the first principal component (FPC) is used for the subsequent analysis. Gene selection then proceeds by solving the optimisation

$$
\hat{\beta} = \arg \min_\beta \left\{ L(\beta) + \lambda \sum_{k=1}^{H_l} ||\beta_k||_1 \right\}
$$

where the $\beta_k, k = 1, \ldots, H_l$ are coefficients corresponding to the FPCs for each gene in pathway $l$. For a given value of $\lambda$, this produces a set, $S(\lambda) = \{k : \beta_k \neq 0\}$ of genes selected by the model. The method proceeds by selecting multiple such sets, each with a different value of $\lambda$, with the ‘best’ subset $S^*$ selected using the AIC criterion. A pathway likelihood ratio (LR) statistic is then obtained by comparing a logistic likelihood function with $\beta = \hat{\beta}$ against the null with $\hat{\beta} = 0$. Finally a pathway p-value is obtained by comparing the pathway LR statistic against those obtained from multiple permuted datasets.

We conclude this section by describing the penalised regression method proposed by (Zhou et al., 2010). While not strictly a PGAS method, it provides an interesting example of penalised regression since it combines $\ell_2$ and $\ell_1$ penalties, with SNPs grouped into genes, to drive simultaneous SNP and gene selection, with a particular emphasis on the identification of rare variants. They apply their method in a small application study comprising 148 SNPs and 17 genes, and this is thus an early example of the application of the sparse group lasso model (see Chapter 4) to SNP data. SNPs are mapped to $H$ genes, $\mathcal{W}_k (k = 1, \ldots, H)$, each containing $S_k$ SNPs, with the parameter vector corresponding to SNPs in $\mathcal{W}_k$, denoted by $\beta_k = (\beta_{k1}, \ldots, \beta_{kS_k})$. SNPs not mapped to genes are however still retained in the model.
The objective function is then given by

\[
\hat{\beta} = \arg \min_{\beta} \left\{ L(\beta) + \lambda_g \sum_{k=1}^{H} ||\beta_k||_2 + \lambda_s \sum_{j=1}^{P} ||\beta_j||_1 \right\}
\]

Sparsity at the SNP level is encouraged through the use of the $\ell_1$ (lasso) penalty. The $\ell_2$ (group lasso) penalty encourages sparsity at the gene level, and aids the detection of rare variants residing within selected genes, since the ridge penalty relaxes the constraint on SNP coefficients within those genes. The regularisation penalties, $\lambda_g$ and $\lambda_s$ are adjusted to ensure that each SNP experiences the same overall penalty, irrespective of whether it is mapped to a gene or not. Thus if a SNP $j$ that is mapped to a gene experiences a penalty $\lambda_g ||\beta_k||_2 + \lambda_s |\beta_j|$, an unmapped SNP experiences an equal penalty $\lambda |\beta_j|$, where $\lambda = \lambda_g + \lambda_s$. 

Table A.1: Key features of reviewed pathways-based genetic association methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Univariate/multivariate</th>
<th>Procedure for determining pathway significance</th>
<th>Null hypothesis tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligator</td>
<td>univariate</td>
<td>significant genes contain at least one SNP passing predefined p-value threshold; count significant genes in pathway</td>
<td>competitive</td>
</tr>
<tr>
<td>Yu et al. (2009)</td>
<td>univariate</td>
<td>rank truncated product of SNP p-values</td>
<td>self-contained</td>
</tr>
<tr>
<td>Eleftherohorinou et al. (2009)</td>
<td>univariate</td>
<td>sum of Cochrane Armitage Trend SNP statistics, followed by lasso logistic regression to identify SNPs</td>
<td>self-contained</td>
</tr>
<tr>
<td>GenGen</td>
<td>univariate</td>
<td>most significant mapped SNP determines gene score; test for over-representation of high ranking genes</td>
<td>competitive</td>
</tr>
<tr>
<td>GRASS</td>
<td>multivariate</td>
<td>lasso regression selects eigenSNPs for each gene; grouped ridge regression followed by permutation establishes pathway significance</td>
<td>self-contained</td>
</tr>
<tr>
<td>GSEA-SNP</td>
<td>univariate</td>
<td>test for over-representation of high ranking SNPs in each pathway</td>
<td>competitive</td>
</tr>
<tr>
<td>PLINK Set-test</td>
<td>univariate</td>
<td>average, p-value of ‘independent’ SNPs passing pre-defined threshold</td>
<td>self-contained</td>
</tr>
<tr>
<td>PoDA</td>
<td>multivariate</td>
<td>distance-based metric to establish SNP similarity across pathways in case-control study</td>
<td>self-contained</td>
</tr>
<tr>
<td>SNP ratio test</td>
<td>univariate</td>
<td>proportion of SNPs passing predetermined significance threshold in a given pathway</td>
<td>competitive</td>
</tr>
<tr>
<td>Zhao et al. (2011)</td>
<td>multivariate</td>
<td>gene score established using first principal component eigenSNP; reduced gene set obtained using lasso penalised regression; LR test for pathway significance</td>
<td>self-contained</td>
</tr>
</tbody>
</table>
Bibliography


convex functions: a survey and extensions”. Mathematical Methods of Operations Re-

Green, E. K. et al. (2010). “The bipolar disorder risk allele at CACNA1C also confers
risk of recurrent major depression and of schizophrenia.” Molecular psychiatry 15.10,
pp. 1016–22.

Green, E. D. et al. (2011). “Charting a course for genomic medicine from base pairs to

Gunter, J. L. et al. (2006). “Validation testing of the MRI calibration phantom for the
Alzheimer’s Disease Neuroimaging Initiative Study”. ISMRM 14th Scientific Meeting
and Exhibition.

Guzzetta, G., G. Jurman, and C. Furlanello (2010). “A machine learning pipeline for quanti-
tative phenotype prediction from genotype data.” BMC bioinformatics 11 Suppl 8.Suppl 8,
S3.


Hariri, A. R. et al. (2002). “Serotonin Transporter Genetic Variation and the Response of
the Human Amygdala”. Science 297.5580, pp. 400–403.

Hastie, T., R. Tibshirani, and J. Friedman (2008). The Elements of Statistical Learning:
Data Mining, Inference and Prediction. 2nd. Springer, New York.


Hayasaka, S. et al. (2004). “Nonstationary cluster-size inference with random field and

Hedrick, P and S Kumar (2001). “Mutation and linkage disequilibrium in human mtDNA.”

Surveys 2, pp. 61–93.

Genetics 2.October, pp. 1–11.


BIBLIOGRAPHY


Skarman, A. et al. (2012). “A Bayesian variable selection procedure to rank overlapping

tomatic correction of intensity nonuniformity in MRI data”. *IEEE transactions on med-
ical imaging* 17.1, pp. 87–97.

Sleegers, K. et al. (2010). “The pursuit of susceptibility genes for Alzheimer’s disease:

Sobreira, C et al. (1997). “Mitochondrial encephalomyopathy with coenzyme Q10 defi-
ciency.” *Neurology* 48.5, pp. 1238–43.


Splawski, I. et al. (2004). “Ca(V)1.2 calcium channel dysfunction causes a multisystem

*Circulation* 98.19, pp. 2088–2093.

Steen, E. et al. (2005). “Impaired insulin and insulin-like growth factor expression and
signaling mechanisms in Alzheimer’s disease - is this type 3 diabetes”. *Journal of

lobe structure with relevance to neurodegeneration in Alzheimer’s disease.” *NeuroImage*
51.2, pp. 542–54.

Stein, J. L. et al. (2010b). “Voxelwise genome-wide association study (vGWAS).” *Neu-
roImage* 53.3, pp. 1160–74.

Stein, J. L. et al. (2012). “Identification of common variants associated with human hip-
focampal and intracranial volumes”. *Nature Genetics* 44.5.

Subramanian, A. et al. (2005). “Gene set enrichment analysis: a knowledge-based ap-
proach for interpreting genome-wide expression profiles.” *Proceedings of the National
Academy of Sciences of the United States of America* 102.43, pp. 15545–50.

Sun, Y. V. et al. (2009). “Identification of correlated genetic variants jointly associated with


BIBLIOGRAPHY


