Penalized regression and its applications to genetics and genomics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Erin Edward Austin

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Wei Pan, Ph.D

June, 2014

Acknowledgements

Please allow me to take a minute to thank my entire committee for their efforts: Dr. Wei Pan, Dr. Lin Yee Chen, Dr. Xiaotong Shen, Dr. Julian Wolfson, and Dr. Baolin Wu. I have been fortunate to work with them all, and I sincerely appreciate what they have done for me. It is important to note that all papers resulting from this dissertation work are coauthored by Dr. Pan and Dr. Shen. In mathematical terms Dr. Pan's guidance was the necessary condition for this dissertation to be completed. I hope to reward his patience and dedication with a career that makes him proud. Dr. Chen has given generously of his time, teaching me about collaborative research, academia, and balancing life. Both are tremendous models of professional and personal responsibility, and as they both have shown me, none of this matters without family. Thank you to my wife and son, Katie and Hugh. It would take another chapter to list the sacrifices they made for me.

The work in this dissertation was possible because of a number of other people and resources. This research was supported by NIH grants R01 HL65462, R01 HL105397 and R01 GM081535. Some of the dissertation work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute. The ongoing support of MSI staff is greatly appreciated. The work in Chapter 2 makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. The work in Chapters 3 and 4 makes use of Genetics Analysis Workshop 18 (GAW18) data. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18

were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575. I am appreciative of the valuable feedback from the GAW18 Machine Learning and Data Mining Group including Ashley Bonner, Hsin-Hsiung Huang, Dr. Stacey Knight, Dr. Ake Lu, and group leader Dr. Rita Cantor. I thank reviewers and editors for helpful and constructive comments. Thank you to Dr. Yen-Yi Ho for discussions on Modified Liquid Association.

Abstract

The quality of genetics-based personalized medicine is a direct function of the success of statistical genomics, defined here as the application of statistical methodologies to genome data. The following dissertation provides two new statistical tools and insights for three areas of interest within the statistical genomics field: (1) better disease outcome prediction using personal genomes, (2) describing the association between genome regions and an outcome, and (3) discovering previously unknown subpopulations within a population. With respect to each of the three problems, penalized regression, in particular regression utilizing the truncated L_1-penalty (TLP), is an essential element of the related methodology. Collectively, the dissertation reveals potential gains from using penalties better aligned with the data's structure and the research aim; for example, by syncing penalty features to underlying genetic architectures to improve prediction. Supported by both simulation and real data analysis, the work herein develops and demonstrates the promise of (1) a new global testing statistic for quantifying the association of a targeted genome region and a disease outcome and (2) a new group truncated L_1-penalty (gTLP) methodology akin to hierarchical clustering that in some settings is able to uncover previously unknown subpopulations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Genetics-based personalized medicine has the potential to positively impact preventive care, diagnosis, treatment, and quality of life. Better personalized medicine can be realized through advances in statistical genomics, by which I mean the application of statistical methodologies to genome data. The following dissertation offers insight and statistical tools for three emerging areas of interest for researchers wanting to incorporate genetic information. First, there is a strong demand for better disease outcome prediction using personal genomes. Second, researchers want new statistical tests to find associations between genome regions and an outcome. Third, it is becoming increasingly necessary to develop methodologies for uncovering previously unknown subpopulations within a population.

Penalized regression is a diverse and promising methodology useful for statistics-based work in these three areas. Penalized regression can use information about the disease structure to develop richer genotype-phenotype maps that can then be leveraged to improve disease prediction. For example, penalized regression has two possible features, variable selection and proportional coefficient shrinkage, that allow researchers to build models tailored to hypothesized characteristics of the genotype-phenotype map. In effect, the regression methodology boosts prediction by incorporating estimation components that leverage features of known or hypothesized networks of genetic variants linked to the disease outcome. As another example of its use, the coefficient sets estimated with penalized regression for a targeted genomic region can themselves be incorporated into

new testing statistics. The expansive range of penalties provides researchers great flexibility in how they apply penalized regression to problems. Penalizing differences between simultaneously estimated sample-level models represents a new application. The result is a new hierarchical clustering tool to partition a population into unknown subpopulations defined by unique models.

In Chapter 2 the predictive performance of penalized regression is investigated in the genome-wide association studies (GWAS) setting. Pioneering work in genetic statistical modeling focused on this setting, where the candidate mutations occurred with at least a certain frequency in the population (e.g. 5%). GWAS research commonly applied a stringent significance threshold to account for multiple testing. Therefore, the number of variants determined to be causative for a disease was usually small and consisted only of variants with a strong individual association with the disease outcome. The work described in Chapter 2 demonstrates how penalized regression can outperform maximum likelihood estimation. Penalized regression, in particular, can be used to overcome the issue of high-dimensionality; that is, the number of covariates exceeding the number of samples. Of equal importance, the methodology can enhance predictive and classification performance by considering variants that marginally fail to meet a GWAS-level significance threshold. Because of this it is reasonable to hypothesize that predictive performance gains are possible when penalized regression accounts for the simultaneous impact on a disease of groups of variants individually with only moderate association with a disease. The work used simulation to explore different underlying genetic architectures: sparse versus non-sparse, and strong versus moderate effect sizes. When the features of the penalty match the architecture of the disease, performance likewise improves. However, application to real data sets for Crohn's disease and bipolar disorder, show a need to incorporate more genetic variants and biological features of a studied disease into the penalized regression models.

The boom in available genome data through sequencing has provided adequate numbers of samples to facilitate dependable investigation of rarer variants; that is, mutations occurring at small frequencies (less than 5% here). The availability of whole genome sequence data means that statistical genetics models must evolve to meet the challenge of using both rare and common variants to link previously unidentified genome loci to disease related traits. In Chapter 3 I extend the penalized regression approach to include

rare variants. Using a real systolic blood pressure dataset provided by the Genetic Analysis Workshop 18 (GAW18), it is possible to confirm the value of penalized regression in predicting this hypertension related trait. The work presented in this chapter is a comparison of the performance of a spectrum of penalized regressions that used at first only common variants (mutation frequencies $\geq 5\%$) or only rare variants to predict systolic blood pressure (SBP). Next, combinations of common and rare variants were used to model SBP, and the impact on prediction was quantified. The study demonstrates how penalized regression can improve prediction for any combination of common and rare variants compared to maximum likelihood estimation. Models using both types of variants, though, provide better predictions of systolic blood pressure than those using only one variant type. The results in this chapter exhibit how penalties directed towards the interrelationships between these variants hold vast potential to improve prediction. Of note, this work provides evidence that penalized regression may be able to account for environmental covariates. Chapters 2 and 3 substantiate the use of penalized regression in a statistical genetics setting.

Natural genetic structures like genes may contain multiple variants that work as a group to determine a biologic outcome. The effect of rare variants are hypothesized to be explained best as groups collectively associated with a biologic function. Therefore, it is important to develop statistical tools such as powerful association tests to identify a true association between a group of variants and an outcome of interest. In Chapter 4 I delineate a novel penalized regression based global test for the association between sets of variants and a disease phenotype. Rare variant analysis is given particular emphasis to provide insight into potential gains in power when testing on sequence data. I again utilize GAW18 data, specifically 200 sets of simulated disease outcomes from the real genomes of nearly 150 unrelated individuals. The power values of the new statistic are calculated and compared to those obtained from five well-regarded global tests that do not use logistic regression (Score, Sum, SSU, SSUw, and UminP) and a set of tests using either the SSU or score statistics and LASSO logistic regression. The results in this chapter lend support for several conclusions upon which to build. First, the new penalized regression based association test may be more powerful in some settings for testing sets of rare variants or regions with both rare and common variants. Second, as shown with an analysis of the MAP4 gene, the new test may be able to better leverage

variants with non-overlapping disease information. In fact, there is some evidence that with further improvements the new global test may be able to be tailored to different rare variant architectures for further power gains. Third, the penalized regression approach provides meaningful, though not complete, information on associated variants in a group of interest. Other methods do not provide this information.

For some modeling problems a population may be better assessed as an aggregate of unknown subpopulations, each with a distinct relationship between a response and associated variables. The finite mixture of regressions (FMR) model, where an outcome is derived from one of a finite number of linear regression models, is a natural tool in this setting. In Chapter 5 I first propose a novel penalized regression approach, then demonstrate how it can, in some types of problems, better identify subpopulations and their corresponding models than a semiparametric FMR method. The new method fits models for each person via grouping pursuit, utilizing a new group truncated L_1-penalty (gTLP) that shrinks differences between estimated parameter vectors. The methodology causes the individuals' regression coefficients to cluster into a few common models, in turn revealing previously unknown subpopulations. In fact, by varying the penalty strength, the new method can reveal a hierarchical structure among the subpopulations that can be useful in exploratory analysis. Simulations using FMR models and real data analysis show the performance of the method is promising.

Chapter 6 provides conclusions for the aggregate work. Statistical genomics is a constantly expanding area providing biostatisticians and public health researchers numerous challenges, three of which are addressed herein. In this dissertation penalized regression is part of all the methodology used or derived. Its consistent use allows for some generalizable interpretations despite differences in the exact problem being addressed. Ideas for future work are included in the discussion.

Chapter 2

Penalized Regression and Risk Prediction in Genome-Wide Association Studies

2.1 Introduction

Genetic information has the potential to improve health outcomes by allowing an individual to tailor preventive care and treatment plans to his or her personalized medical needs. An important task in personalized medicine is using a person's genome to predict disease risk (and treatment response). A necessity for making accurate risk predictions based on individuals' genomes is obtaining data on their genetic variants and phenotypes. Genome-wide association studies (GWAS) provide such data to researchers. Now one critical question is how to best predict disease risk from a large number of genetic variants, such as single-nucleotide polymorphisms (SNPs). Penalized regression equipped with variable selection, such as LASSO (Tibshirani, 1996), is deemed to be promising in this setting. However, for some diseases the sparsity assumption used by penalized regression to facilitate variable selection may not hold, in which case it is not completely clear how to proceed: should we apply a penalized or unpenalized approach? how about other penalized methods that do not conduct variable selection, such as ridge regression

(Hoerl and Kennard, 1970)? To answer these questions, our current research investigated the performance of an unpenalized approach and several representative penalized regression approaches under various scenarios with sparse or non-sparse models.

GWAS identify risk SNPs by individually testing each SNP with a stringent significance level adjusting for multiple testing. Many SNPs discovered to be associated with disease have been validated (McCarthy et al., 2008). However, for many strongly heritable diseases, their risk cannot be adequately explained by only a small number of identified SNPs. For example, adding seven SNPs known to be associated with breast cancer to the National Cancer Institute's Breast Cancer Risk Assessment Tool increased the discriminatory accuracy of the tool by only a small amount as measured by the area under the receiver operating characteristic curve (AUC) (Gail, 2009). In related work Gail (2008) demonstrated that very large relative risks are needed for a single factor to meaningfully improve disease classification; therefore, estimation of the effect of many disease associated SNPs with small effects will require researchers to address the issue of candidate SNPs vastly outnumbering available case samples. Penalized regression with variable selection can address this issue. In another study the percent of phenotypic variance in the highly heritable trait height explained by SNPs increased from 5% to 45% when both genome-wide significant SNPs and many non-significant SNPs were considered simultaneously (Yang et al., 2010). Increasing the number of SNPs used may also impact risk prediction: the inclusion of many non-significant SNPs discriminated bipolar disorder, coronary heart disease, hypertension, and Crohn's disease to some degree better than when only fewer and more significant SNPs were included (Evans et al., 2009). Furthermore, there was evidence to support polygenic effects for many common diseases (Park et al., 2010). For example, the risk of schizophrenia seemed to be associated with hundreds to thousands of SNPs (The International Schizophrenia Consortium, 2009). It is now hypothesized that many common diseases are associated with many SNPs with small to moderate effects.

Two studies have confirmed the value in including up to thousands of SNPs when assessing disease risk (Kang et al., 2011; Wei et al., 2009). Importantly, both studies revealed that, while still noticeably better than random, logistic regression with maximum-likelihood estimation was suboptimal in utilizing large numbers of SNPs to classify disease status. A recent study concluded that utilizing penalized regression with

variable selection, specifically LASSO, on a large number of SNPs in addition to those reaching the genome-wide significance level could improve prediction of Crohn's disease (Kooperberg et al., 2010). This disease is a form of inflammatory bowel disease affecting as many as 1.4 million Americans (About Crohn's Disease, 2009; Crohn's Disease, 2010). Patients with Crohn's disease have a chronic inflammation of the gastrointestinal tract that causes mild to severe symptoms such as abdominal pain, fever, and fatigue (Crohn's Disease, 2010).

Gaya et al. (2006) presented evidence of the heritability of Crohn's disease. Two subsequent studies (WTCCC, 2007; Franke et al., 2010) identified six regions of chromosome 10 associated with Crohn's disease. To mimic real situations we use the real SNP data from chromosome 10 to generate simulated disease risks and disease phenotypes in order to assess the performance of various regression methods with respect to risk estimation and disease classification. Specifically, we consider four types of true models: (1) a sparse model with risk being determined by a small number of SNPs with large effect sizes, (2) a sparse model with a small number of SNPs with moderate effect sizes, (3) a non-sparse model with risk being determined by a large number of SNPs with moderate effects, and (4) a non-sparse model with an even larger number ($> 1/3$ of the sample size) of SNPs with small effect sizes. We consider both unpenalized and penalized regressions, the former based on maximum likelihood estimator (MLE) while the latter on (1) least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), (2) smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), (3) truncated L_1-penalty (TLP) (Shen et al., 2012), (4) ridge regression (Hoerl and Kennard, 1970) and (5) elastic net (Zou and Hastie, 2005). This study is a follow-up on Kooperberg et al. (2010), in that we consider several new penalized regression methods and contrast the performance of the methods between sparse and non-sparse true models.

We also study the discrimination capabilities of the regression methods on two real data sets, Crohn's disease and bipolar disorder provided by the Wellcome Trust Case Control Consortium (WTCCC) (2007). It was confirmed that the best performer was dependent on the number and effect sizes of causal SNPs in the true model, and the inclusion of SNPs failing to meet the genome-wide significance level impacted the prediction accuracy.

2.2 Methods

2.2.1 Data

We use the Crohn's disease and bipolar disorder case and control data provided by the WTCCC. The WTCCC has collected genotype data of about 500,000 SNPs for approximately 2,000 samples for each of seven diseases, such as type 1 diabetes, hypertension, bipolar disorder and Crohn's disease, and 3,000 controls (WTCCC, 2007). For simulations, we use the genotype data of 28501 SNPs on chromosome 10 for Crohn's disease cases and controls. For quality control purposes, per WTCCC recommendations, we remove some samples and retain 1748 Crohn's disease samples and 2938 control samples; we also exclude some SNPs as recommended. Next, we eliminate the SNPs with a minor allele frequency (MAF) less than 5%. Furthermore, to mimic practical situations while maintaining a reasonable size for repeated simulations, we test each SNP separately by a chi-squared test for its association with Crohn's disease, and remove those with p-values larger than 0.1. At the end, we have about 2300 SNPs left and use them throughout our simulations.

2.2.2 Model

Let $Y_i = 0$ or 1 be a binary disease indicator for subject $i = 1, ..., n$, and X_{ij} subject i's minor allele number (0,1, or 2) for SNP $j = 1, ..., m$. Our aim is to build a model to successfully estimate subject i's risk of disease, $P(Y_i = 1|x_i)$, based on his or her SNP data $x_i = (X_{i1}, ..., X_{im})^T$. As in standard practice for binary outcomes, we use a logistic regression model:

$$\text{logit}(P(Y_i = 1|x_i)) = \log(\frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)}) = \beta_0 + \sum_{k=1}^{p} X_{ik}\beta_k, \tag{2.1}$$

where β_0 and β_k are unknown regression coefficients to be estimated; $p \leq m$ indicates any user specified subset of the SNPs.

In *unpenalized* logistic regression with maximum-likelihood estimator (MLE), β_0 and $\beta = (\beta_1, ..., \beta_p)^T$ are estimated by maximizing the log-likelihood:

$$l(\beta_0, \beta) = \sum_{i=1}^{n} Y_j(\beta_0 + x_i^T\beta) - \log[1 + \exp(\beta_0 + x_i^T\beta)]. \tag{2.2}$$

The MLE is asymptotically unbiased with fixed p as $n \to \infty$, but it may not be for a large p. One possible remedy is to introduce regularization or penalization on regression coefficients. The use of certain penalties, such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), shrinks many regression coefficient estimates to be 0, effectively selecting a subset of SNPs to be used for prediction. *Penalized* logistic regression provides coefficient estimates for β_0 and β by maximizing a penalized log-likelihood (Friedman et al., 2008):

$$l(\beta_0, \beta) - \lambda P(\beta), \tag{2.3}$$

where $\lambda \geq 0$ is a tuning parameter controlling the extent of penalization imposed by penalty $P(\beta)$. LASSO regression uses

$$P(\beta) = \sum_{k=1}^{p} \mid \beta_k \mid, \tag{2.4}$$

which is convex and computationally convenient. However, LASSO estimates are biased and may not be consistent. To avoid these issues, Fan and Li (2001) proposed using the SCAD penalty $P(\beta, \lambda)$ replacing $\lambda P(\beta)$:

$$\frac{dP(\beta, \lambda)}{d\beta} = \sum_{k=1}^{p} \lambda \text{sign}(\beta_k)[I(|\beta_k| \leq \lambda) + \frac{(a\lambda - \mid \beta_k \mid)_+}{(a-1)\lambda} \cdot I(\mid \beta_k \mid > \lambda)] \tag{2.5}$$

for $a = 3.7$. While maintaining the capability of variable selection, the SCAD penalty does not introduce biased estimates for some larger coefficients. The truncated L_1-penalty (TLP) adaptively determines which larger coefficients will not be penalized by introducing a separate thresholding parameter $\tau > 0$ (Shen et al., 2012):

$$P(\beta) = \sum_{k=1}^{p} \min(|\beta_k|/\tau, 1). \tag{2.6}$$

If a coefficient $\beta_k > \tau$, it will not be further penalized. The TLP approaches the L_0-loss as $\tau \to 0^+$. A penalized method without the capability of variable selection is ridge regression (Hoerl and Kennard, 1970) with penalty

$$P(\beta) = \sum_{k=1}^{p} \beta_k^2. \tag{2.7}$$

Certain true models might be best estimated using a hybrid penalty that simultaneously performs variable selection and continuous shrinkage (Zou and Hastie, 2005). In these

settings elastic net penalized regression may be more suitable. Elastic net penalized regression has been shown to produce a sparse model with good prediction accuracy, possibly superior to LASSO, while simultaneously promoting the grouping of strongly correlated predictors (Zou and Hastie, 2005). Its penalty structure is a weighted combination of the LASSO and ridge penalties controlled by a user specified mixing parameter α, which is restricted to [0, 1]. The *naive* elastic net penalty (Zou and Hastie, 2005) is

$$P(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1, \tag{2.8}$$

where α is selected to match the desired balance of variable selection and coefficient shrinkage. Zou and Hastie (2005) suggested that further gains may be possible from using a rescaled version of the elastic net penalty. However, Friedman et al. (2008) used the *naive* version of the penalty in the R package `glmnet` they developed to perform elastic net penalized regression. Results presented here follow this convention and are not rescaled. For sparse true models (i.e. with few non-zero β_k's) with a large number of candidate predictors (i.e. a large p), variable selection is often beneficial. However, for non-sparse models with many small non-zero $|\beta_k|$'s, variable selection will be difficult and may not result in good performance. On the other hand, since the ridge penalty has the grouping function (Zou and Hastie, 2005), ridge regression performs like model averaging. It is known that neither model selection nor model averaging can dominate the other, and each performs better under different situations (Yuan and Yang, 2005; Shen and Huang, 2006). In the current context, especially with non-sparse true models, it is not clear how LASSO, SCAD, and TLP compare to the ridge penalty for risk prediction, or if the elastic net penalty is superior, which is one of our aims.

2.3 Simulations

2.3.1 Simulation Set-ups

We use the real SNP data of the WTCCC control cohort to generate disease probabilities, $\pi_i = P(Y_i = 1)$. First, we randomly select p_1 causal SNPs (i.e. with corresponding $\beta_k \neq 0$). The true correlations for any two SNPs range from -0.8371 to 1 and approximately fit a symmetric unimodal distribution centered at 0. Table 2.1 provides summary statistics for all pairwise correlations for example sets of size $p = 5, 10, 50, 100, 500, 1000$ randomly

selected SNPs. Table 2.1 demonstrates how the true models with various numbers of p SNPs contain a diverse range of minor, moderate or strong correlations among the SNPs.

p	**Minimum**	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
5	-0.096	-0.026	-0.003	0.002	0.019
10	-0.040	-0.013	-0.001	0.017	0.216
50	-0.136	-0.012	-0.001	0.011	0.994
100	-0.408	-0.013	0	0.012	0.999
500	-0.668	-0.013	0	0.013	1
1000	-0.835	-0.013	0	0.013	1

Table 2.1: Summary Statistics for All Pairwise Correlations among the top p SNPs.

We use $p_1 = 10$ for two sparse models, one with strong effects (i.e. large $|\beta_k|$'s) and the other with only moderate effects (i.e. smaller $|\beta_k|$'s); we also use $p_1 = 300$ and $p_1 = 900$ for two non-sparse models. Second, we set $\beta_0 = \log(0.05/0.95)$ to emulate diseases with low prevalence, and follow Wray et al. (2007) to create odds ratios (ORs, $\text{OR}_k = \exp(\beta_k)$) of having disease for the p_1 causal SNPs. Specifically, we set $\text{OR}_k = 1 + \epsilon(\text{OR}_0 - 1)$ with ϵ randomly generated from a standard exponential distribution $Exp(1)$ and OR_0 being the mean OR, which is 2.75 and 1.415 for the two sparse models and 1.17 and 1.125 for the two non-sparse models respectively. We also randomly choose the sign of each β_k to be positive or negative to reflect both risk and protective causal SNPs. Third, the disease probability π_i for each subject $i = 1, ..., 2938$ in the WTCCC control cohort, is generated according to logistic regression model (1) with only chosen causal SNPs.

Finally, we use each π_i sequentially to generate disease status $Y_i \sim Bin(\pi_i)$; this step is repeated until we have $n = 2000$ cases and $n = 2000$ controls (while the other cases or controls are ignored) for each simulated dataset. One hundred datasets were generated under each of the four true models.

For each simulated dataset a randomly selected half of both the cases and controls is used as training set for building regression models, while the remaining half is the

test set used for unbiased assessment of performance. The performance of each method is evaluated in two distinct settings. In the first setting we rank all SNPs by the p-values of their univariate association with disease. Starting with a few of the most significant SNPs, we fit and refit the logistic model for each method, sequentially adding more and more top ranked SNPs into the model (1) to be fit. The structure of this scenario informs when the inclusion of increasingly less significant SNPs improves or deteriorates the performance. Gail (2009) measured the impact of only seven SNPs on classification of one disease, breast cancer, finding a very minor effect. Although they were not directly studying prediction, Yang et al. (2010) identified one trait, height, whose heritability could be explained better with models that considered many non-significant SNPs. Our first modeling scenario generalizes this previous work to measure the impact of including more and more SNPs (by design including less significant SNPs) on a spectrum of models with less and less true sparsity. Thus, the results can inform about underlying genetic architectures for which penalized regression can use additional SNPs to improve risk classification. The results presented in the following section for the unpenalized regression are from the usual MLE, while those for LASSO, SCAD and ridge use the tuning parameter λ selected via 10-fold cross-validation to have the smallest prediction error for any given number of candidate SNPs.

As exhibited in equations (2.6) and (2.8), the elastic net penalty depends on an additional parameter, α, and the TLP penalty requires specification of τ. Elastic net estimates are generated for each of a sequence of penalties defined by a uniformly spaced sequence of values for the mixing parameter, α. The elastic net regression models are fit starting with $\alpha = 0$, corresponding to ridge regression, and then with α increased by units of 0.10 until α= 1, corresponding to LASSO regression. For the TLP we apply a range of τ values chosen to yield a series of models with minor to major coefficient shrinkage. To save computing time for tuning parameter selection for the simulated datasets, we use an independent tuning dataset of an equal size generated exactly like the training and test data sets. The idea is similar to the CV except that we only need to fit a model once with the training data, then use the tuning data to calculate the prediction error and thus select λ and τ.

The second setting is designed to compare the performance of the methods with a

large number of the candidate SNPs. In penalized regression the regularization parameter λ is systematically varied to generate a solution path of the regression coefficients, from which we identify a global maximum of some performance measurement to represent the best ever performance of the corresponding method.

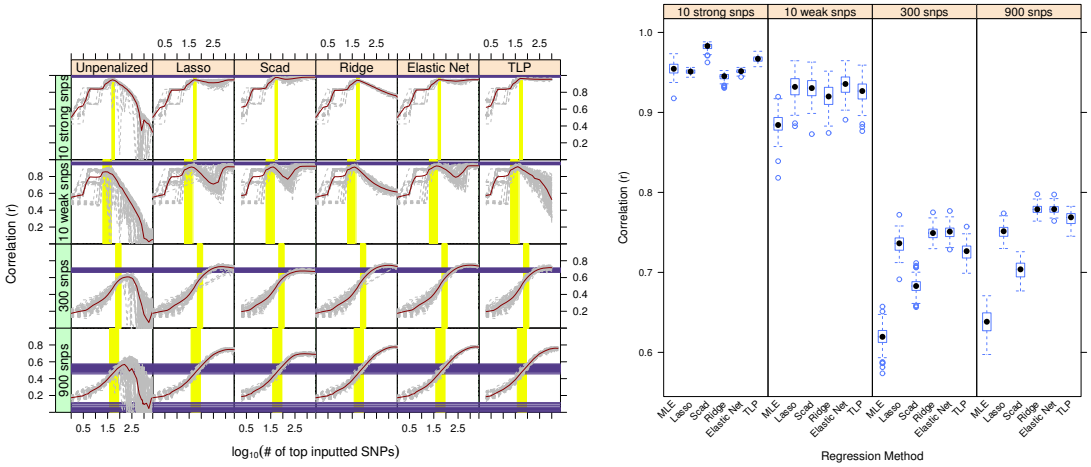
For each method, the estimated β_0 and β_k from a training set are applied to the corresponding test set to obtain risk estimates, $\hat{\pi}_i$. The correlation of the $\hat{\pi}_i$ and the true π_i for the test samples is computed and used to compare the predictive performance of the methods.

This metric has been used in risk and outcome prediction for GWAS data (Wray et al., 2007; Lee et al., 2008). In addition, we also utilize the area under a receiver operating curve (AUC) for test samples to assess the discriminatory capabilities of the regression methods. The AUC is the gold standard metric that has been most consistently used in the GWAS literature. The use of AUC also permits direct comparison to previous related work. R package `glmnet` was used to fit the LASSO, ridge and elastic net penalized regression models. SCAD models were fit using the R package `ncvreg`. TLP models for the simulated data sets were fit using Feature Grouping and Selection Over and Undirected Graph (FGSG) software implemented in Matlab (Yang et al., 2012) while those for the real data were fit using our own implemented R function. Computational time necessitated using the FGSG software, which was much faster in fitting penalized linear regression models. It is known that linear regression models perform well for binary traits with GWAS data (Wu et al., 2010). We also compared the results from penalized logistic regression models fitted by the R function with those from linear models by the FGSG software for the first ten simulated datasets; their differences were within 0.031 in the correlation metric and within 0.01 in the AUC metric.

2.3.2 Main Results

We first investigate the effect of using an increasing number of top SNPs for risk prediction. Figure 2.1(a) presents the correlation between true risk and predicted risk, $Corr(\pi_i, \hat{\pi}_i)$. For each of the four true models, $Corr(\pi_i, \hat{\pi}_i)$ for each method is plotted as a gray curve against the number of the top SNPs used in the candidate model (before penalization) for each of the 100 simulated datasets, and the mean correlation curve over all 100 simulations is plotted as a dark red curve. The elastic net and TLP results

are for the data-tuned values of α and τ respectively. In addition, vertical lines mark the number of the SNPs that would meet a Bonferroni adjusted genome-wide significance level at 0.05 when evaluated individually using a chi-squared test. Examination of the curves beyond the vertical lines reveals situations in which better estimates of the disease risk can be obtained by considering more SNPs, including those failing to meet the genome-wide significance level. The horizontal lines mark the correlations obtained from the MLE of the true model (with exactly all the causal SNPs).



(a) (b)

Figure 2.1: Correlation of the true π_i and the $\hat{\pi}_i$ estimated with various numbers of top SNPs. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum correlation obtained for each simulated dataset across the number of top SNPs.

In the sparse model with strong effect sizes, all penalized methods predict risk nearly as well or better than the unpenalized method, as shown in Figure 2.1(b) where only the maximum correlation across various numbers of top SNPs from each simulated dataset is plotted. For the sparse model with weaker effects and both non-sparse models, all penalized methods by far surpass the MLEs, even ones based on the true models. Among the penalized methods, LASSO, SCAD, TLP and elastic net outperform ridge regression for sparse models, but the trend reverses for non-sparse models for all but the elastic

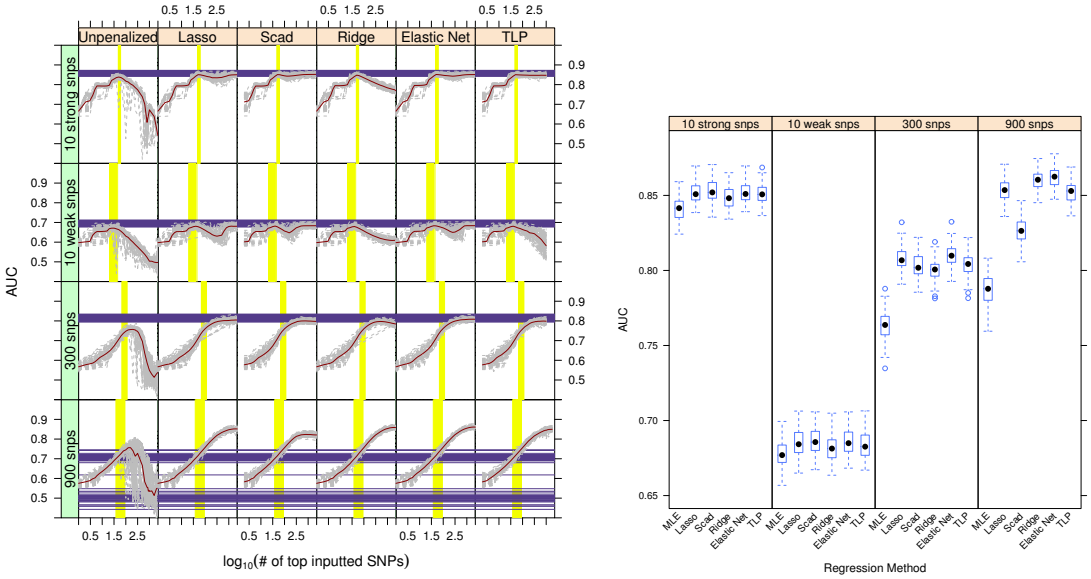
net. As the number of causal SNPs increases or strength of effect decreases, the relative performance of the elastic net and TLP penalties improves. In fact, the elastic net outperforms all methods for the $p_1 = 300$ case, which is to be expected as it is a model balanced between extreme sparse and non-sparse models. The best performing elastic net models are at least as good in the non-sparse $p_1 = 900$ case as those of ridge regression, the best overall performer of the non-mixture penalties. Table 2.2 provides the mean values of the maximum performance metrics of each regression method for the datasets. The table allows quick comparisons of the various methods in all modeling scenarios. These results reinforce the importance of using a suitable penalty for a given problem, depending on whether the model sparsity assumption holds.

Model/Data	**#SNPs**	**Metric**	**MLE**	**SCAD**	**LASSO**	**Elastic Net**	**Ridge**	**TLP**
10 Strong SNPs	Varying	Corr	0.954	0.982	0.951	0.951	0.944	0.966
		AUC	0.841	0.853	0.852	0.852	0.849	0.851
	Fixed	Corr	-	0.974	0.975	0.970	0.769	-
		AUC	-	0.850	0.851	0.850	0.775	-
10 Weak SNPs	Varying	Corr	0.885	0.931	0.931	0.935	0.920	0.925
		AUC	0.678	0.686	0.685	0.686	0.682	0.684
	Fixed	Corr	-	0.912	0.928	0.927	0.620	-
		AUC	-	0.682	0.684	0.684	0.607	-
300 SNPs	Varying	Corr	0.619	0.683	0.735	0.750	0.749	0.726
		AUC	0.763	0.803	0.808	0.810	0.800	0.804
	Fixed	Corr	-	0.659	0.716	0.720	0.725	-
		AUC	-	0.791	0.808	0.808	0.786	-
900 SNPs	Varying	Corr	0.638	0.702	0.751	0.779	0.779	0.767
		AUC	0.787	0.827	0.854	0.862	0.860	0.852
	Fixed	Corr	-	0.674	0.761	0.766	0.784	-
		AUC	-	0.815	0.854	0.856	0.860	-
Crohn's Disease	Varying	AUC	0.675	0.677	0.678	0.678	0.677	0.686
	Fixed	AUC	-	0.672	0.668	0.660	0.612	-
Bipolar Disorder	Varying	AUC	0.607	0.606	0.606	0.609	0.608	0.609
	Fixed	AUC	-	0.595	0.602	0.603	0.594	-

Table 2.2: Mean of the maximum correlations (Corr) and AUCs for each method under either varying or fixed numbers of input SNPs.

To quantify the impact of including more SNPs, we first examine the performance for the sparse models. The LASSO and SCAD, methods with a variable selection feature, are able to maintain near optimal performance even when the number of candidate SNPs far exceeds that of the true model. Further, the elastic net appears to improve on the LASSO. In contrast, both unpenalized and ridge regressions have their prediction accuracy worsened markedly with the inclusion of more SNPs. For non-sparse models containing many SNPs failing to meet the genome-wide significance level, LASSO, SCAD, and TLP are again able to deal with a large number of SNPs for better risk estimation than the MLE. TLP uses the additional SNPs noticeably better than LASSO and SCAD when the true number of causal SNPs grows. Ridge regression is able to surpass these three penalization methods. In all four models the elastic net performs comparably to the best of the other regressions. This is likely due to its being a hybrid of the sparse and non-sparse regression methods, and our method examined a range of α's corresponding to a range of models from those strongly favoring LASSO to those strongly favoring ridge regression. However, it is noteworthy that the elastic net was not bounded by the performances of LASSO and ridge regressions.

Next, the discriminatory abilities of the methods are assessed because correct classification of disease status is key to personalized medicine. The literature for the clinical application of disease assessments universally reported AUCs as the standard for comparing disease classification methods. Therefore, the current study will assess classification using this metric to enable comparisons to previous work. Figure 2.2 demonstrates the classification performance of the methods in terms of their AUCs. The main conclusions remain the same: SCAD, closely followed by LASSO , elastic net, and TLP, is the winner for the two sparse models, while elastic net and ridge regression beat other methods for the non-sparse model with $p_1 = 900$. However, for the non-sparse model with $p_1 = 300$, ridge regression performs worse than all other penalization methods. Elastic net performs best, followed by LASSO, TLP, and then SCAD. Overall, elastic net is either the top performer or close to the top for all true models,



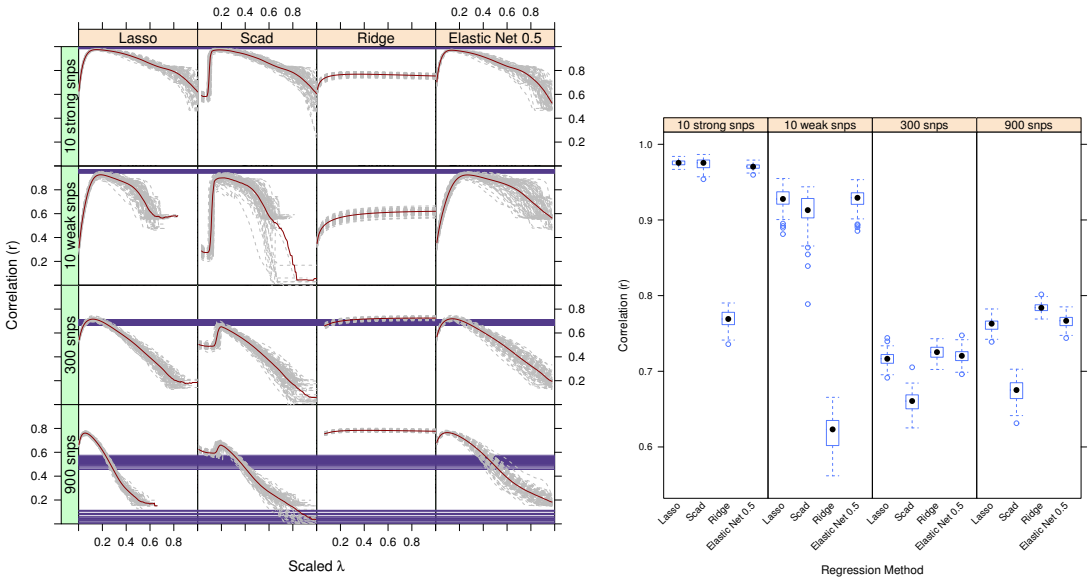
(a) (b)

Figure 2.2: AUC calculated for 100 simulated test datasets with various numbers of top SNPs. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum AUC obtained for each simulated dataset across the number of top SNPs

and every type of penalized regression always beats MLE.

The results presented in Figure 2.2 demonstrate the value of penalized regression in disease risk estimation and classification, especially in utilizing the information in less significant SNPs that may often go unused. A natural question is whether we can eliminate the need to rank SNPs marginally and examine all SNPs simultaneously. The below simulation results address this question. All the penalized methods start with a full model containing all available SNPs; by varying the tuning parameter λ monotonically, various models are fitted and their performance is assessed. Figure 2.3(a) provides curves for the correlations between true and predicted risk at any given value of λ for four of the penalized methods: LASSO, SCAD, ridge, and elastic net. Elastic net results for only models with $\alpha = 0.5$ are shown. Since one value of τ that provides a single intuitive interpretation across all four true models does not exist, TLP results as a regularization

path in terms of λ would have limited comparability to the results from the other models in this setting and are not presented here. As before, the result for each simulation is represented by a gray curve, and the mean curve across all simulations is plotted as a dark red curve. For comparison, the horizontal lines mark the correlations obtained from maximum likelihood estimation using exactly the true causal SNPs. To facilitate plotting, for each penalized method, the value of λ is scaled by its maximum so that it falls inside the interval $[0, 1]$.

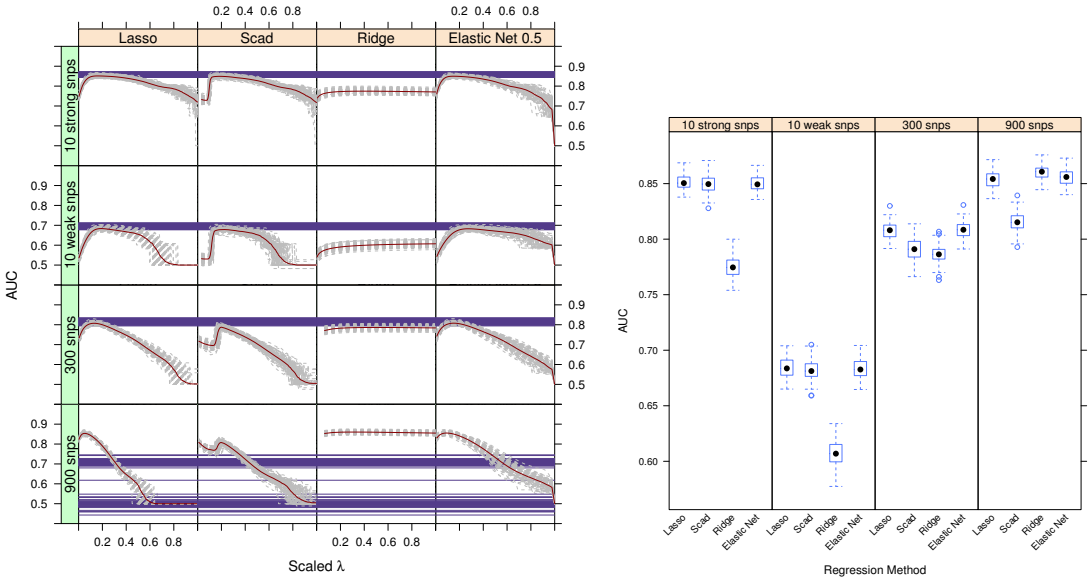


(a) (b)

Figure 2.3: Correlation of the true π_i and the $\hat{\pi}_i$ estimated from all SNPs with various values of regularization parameter λ. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum correlation obtained for each simulated dataset across the values of λ.

As before, SCAD, LASSO and elastic net with $\alpha = 0.5$ outperform ridge regression for sparse models, while for both non-sparse models ridge regression is the best when judged by their optimal performance shown in Figure 2.3(b). Interestingly, LASSO outperforms SCAD in all situations, suggesting the robustness of LASSO to a large number of input variables. The performance of the elastic net with $\alpha = 0.5$ is between

that of LASSO and ridge in all cases as expected. This elastic net's results are closer to the better of ridge and LASSO in all four models; however, the degree to which the best method outperforms the balanced elastic net ($\alpha = 0.5$) varies by true model. This provides strong evidence that matching the sparsity of the penalty to the model sparsity improves classification. Comparing with earlier results, we can conclude that simultaneous use of too many SNPs will deteriorate the performance of any penalized method, suggesting possible gain in performance by a preliminary screening of a large number of variables. Similar conclusions hold if AUC is used to measure the classification performance of the methods (Figure 2.4); however, LASSO, followed closely by the elastic net (α=0.5), is the overall winner, in particular it beats ridge regression even for the non-sparse model with $p_1 = 300$, indicating the necessity of variable selection for large p.



(a) (b)

Figure 2.4: AUC calculated for 100 simulated test datasets from all SNPs with various values of λ. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum AUC obtained for each simulated dataset across the values of λ.

2.3.3 Other Results

Two of the penalties, SCAD and TLP, are non-convex. Thus, there may be multiple local maxima with respect to their corresponding penalized log-likelihood functions, leading to possibly different estimates with different starting values. To examine this issue the authors refit some SCAD and TLP models for the first 20 data sets. The refit models considered the top ranked 1000 SNPs at a few fixed λ values (and a fixed $\tau = 0.1$ for TLP). Eight different sets of initial regression coefficient values were used as the starting values for SCAD and TLP: the estimated coefficient values with the true model, a vector of all zeros, and the coefficients estimated by LASSO at each of the six λ values: 0.01, 0.1, 1, 2.5, 5, and 10. This was done for the true models with 10 (strong) and 300 causal SNPs to represent one sparse true model and one non-sparse true model. The R package `SIS` was used to fit the SCAD models as it allowed user specified initial value sets. FGSG software was used to fit TLP models as before.

Figure 2.5 presents the findings. Each curve represents the average AUC at a given λ over the 20 data sets for each set of the starting values (with the solid one for the first set). The primary finding is that for the λ generating the best AUC given a set of initial coefficients, all eight sets yield comparable AUCs. Results for many of the SCAD models could not be obtained when λ exceeded 0.1 due to numerical problems in the R package; the partial curves are still provided. Many AUC values were the same or within 0.01 for the SCAD scenarios, thus, given the scale the curves appear to overlap in the plots. Not surprisingly the AUC is impacted by the starting values used to find regression coefficient estimates. Importantly, the impact appears to be small near tuning parameter values yielding the top performing SCAD or TLP models.

Below is a short summary on computing time needed to fit each type of penalized regression models. We calculated the average CPU time for one value or one set of tuning parameters for each penalized regression method with 1000 candidate SNPs for the true models with 10 (strong) and 300 causal SNPs. For the 10 strong SNP scenario, SCAD used approximately 30 seconds to fit a model, TLP used 20 seconds, and the model fitting using `glmnet` ranged from 1.5 seconds for ridge regression to 7 seconds for LASSO. For the 300 SNP scenario, SCAD used approximately 44 seconds per model-fitting, TLP used 20 seconds, and `glmnet` ranged from 1.2 seconds for ridge regression to 5 seconds with LASSO.

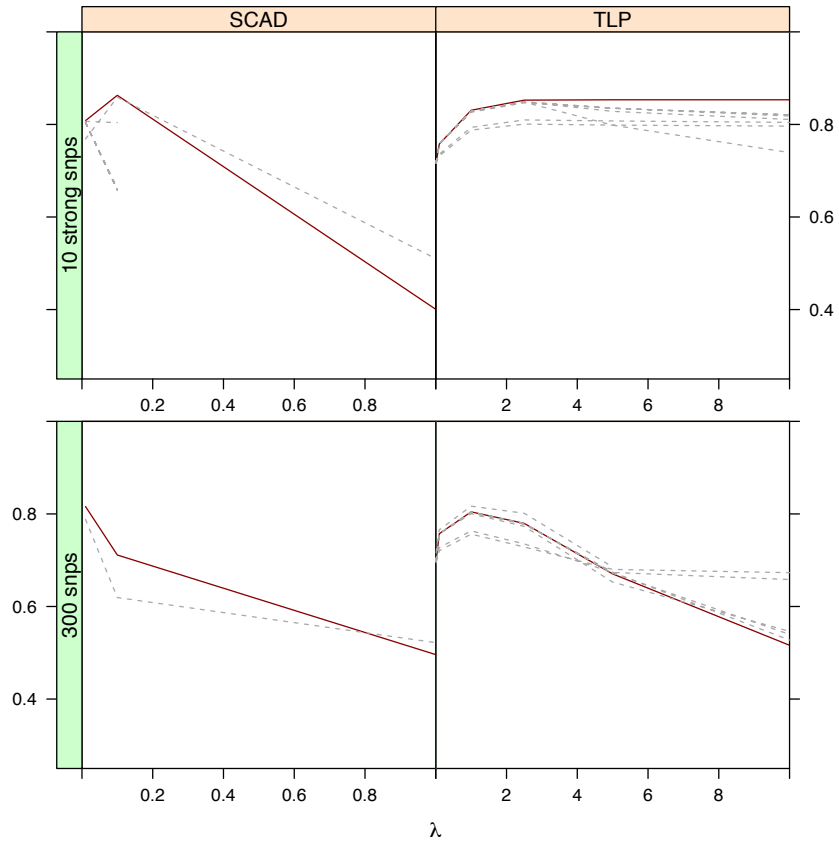


Figure 2.5: Results of SCAD and TLP with various starting values.

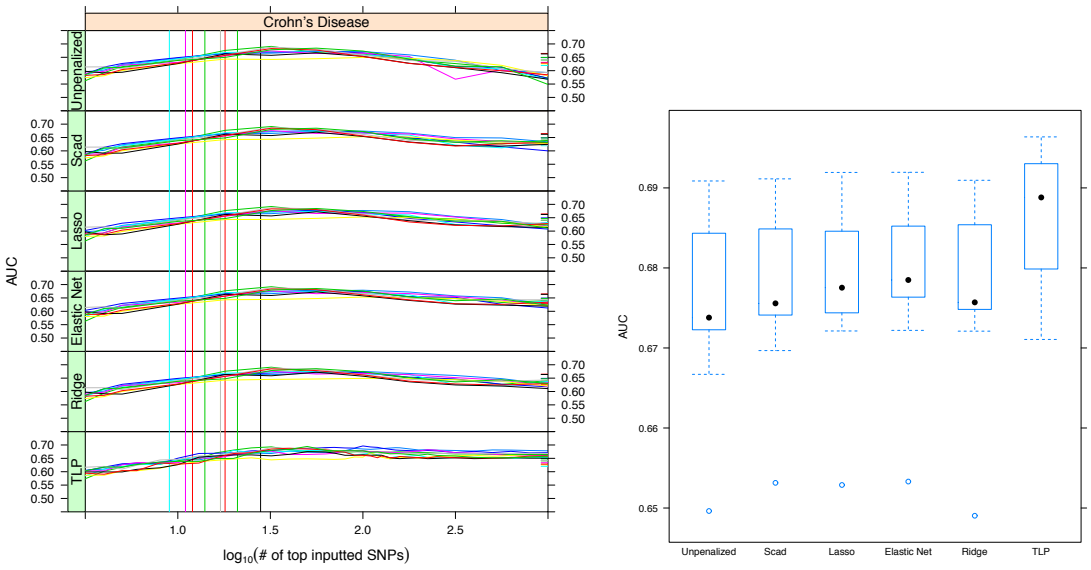
2.4 Examples

The final part of our study examines the classification accuracy of the six regression methods on two WTCCC datasets for Crohn's disease and bipolar disorder. The training and test data are created by randomly dividing the WTCCC disease (case) and WTCCC control samples into two (almost) equally sized sets, one for training and one for test. We consider the 5000 most significant SNPs from all chromosomes as determined by a univariate chi-squared test on each SNP. The whole process, including randomly dividing the true cases and controls into training and test sets, and identifying the 5000 most significant SNPs, is repeated ten times. The results for each of these ten datasets

are presented in the following plots. The number of the significant SNPs meeting the significance level of 0.05/373191 are plotted as vertical lines. Horizontal tick marks on the secondary y-axis represent the maximum AUC achieved by MLE with these significant SNPs.

2.4.1 Crohn's Disease

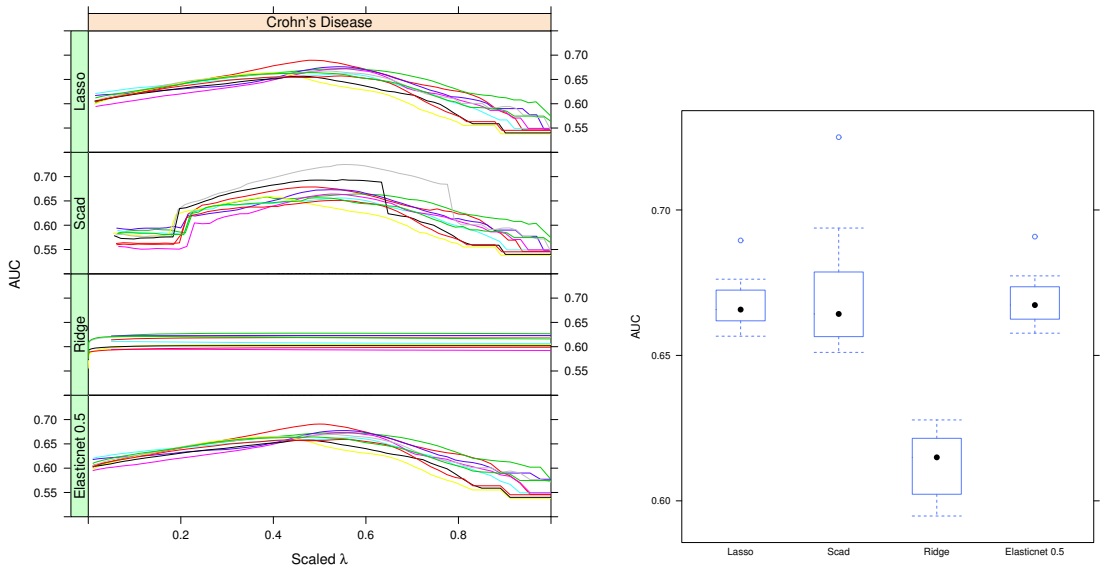
Current research has identified about 80 SNPs associated with Crohn's disease. Figure 2.6 shows that approximately only the top 50 SNPs are needed to obtain the best risk prediction for all the methods; however, this includes more than just those SNPs meeting the significance level of 0.05/373191. Interestingly, although TLP was the overall winner and all five penalized methods are better than the unpenalized one, the performance difference among the methods is small.



(a) (b)

Figure 2.6: AUC calculated for the Crohn's disease test datasets with various numbers of top SNPs. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each dataset across the number of top SNPs.

Figure 2.7 presents the results of the four penalized methods starting with all 5000 SNPs included in a candidate model. With such a large number of candidate SNPs, while the number of the truly predictive SNPs may be small, the ridge penalty is largely outperformed by LASSO and SCAD that are capable of variable selection. The ridge regression is similarly outperformed by an elastic net penalty that shifts part of the weight from the ridge penalty to the LASSO penalty.



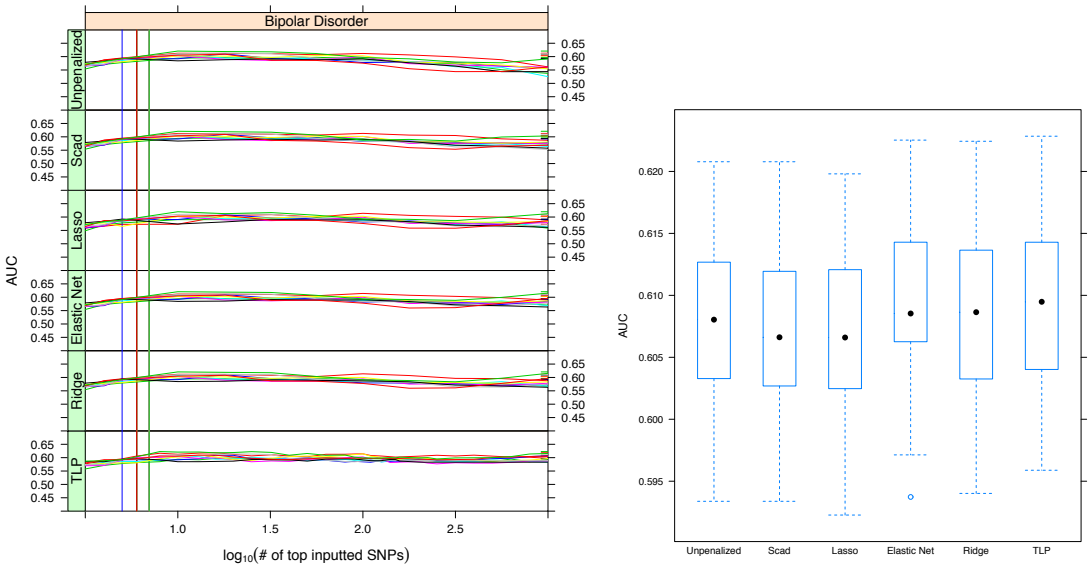
(a) (b)

Figure 2.7: AUC calculated for the Crohn's disease test datasets with all SNPs across the values of λ. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each dataset across the values of λ.

2.4.2 Bipolar Disorder

Bipolar disorder is a condition in which people go back and forth between mania periods of a very good or irritable mood and depression (Bipolar disorder, 2011). Figure 2.8 presents the AUC results as the number of candidate SNPs was increased. Unlike Crohn's disease, penalized regression does not always outperform MLE. Elastic net penalized regression and TLP perform best, though again the performance difference among the

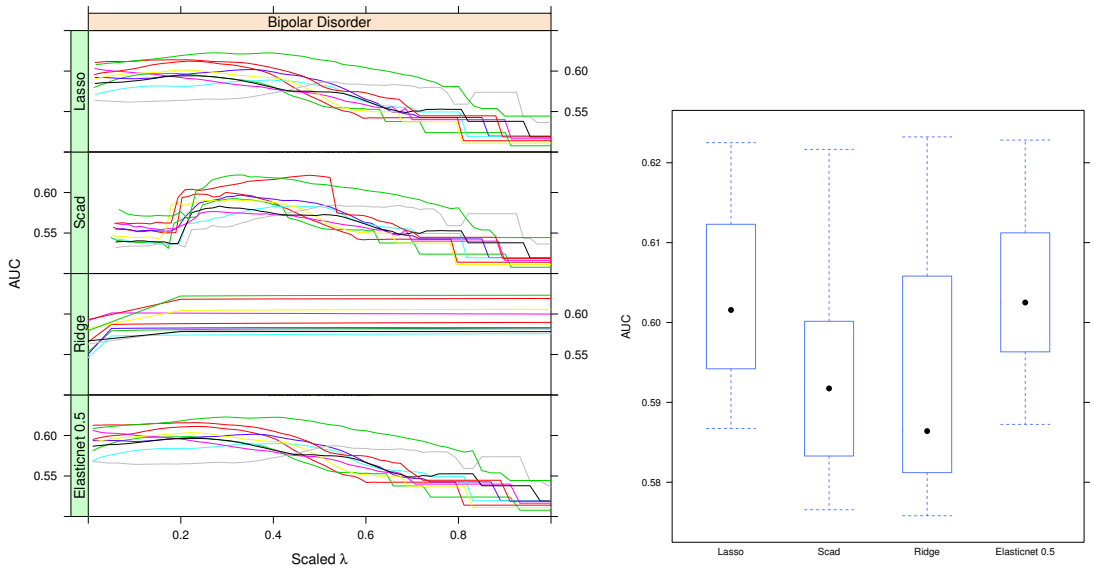
methods is small. As shown in 2.8(b), the inclusion of many SNPs failing to reach the genome-wide significance level does not diminish the discrimination strength of the penalized methods, and in fact ridge and elastic net regression and TLP better use these extra SNPs than both LASSO and SCAD to exceed or nearly exceed its performance achieved with only the few significant SNPs.



(a) (b)

Figure 2.8: AUC calculated for the bipolar disorder test datasets with various numbers of top SNPs. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each dataset across the number of top SNPs.

Next we include all SNPs in each penalized regression model and vary the tuning parameter λ (Figure 2.9). Again it seems that, with a large candidate model containing a large number of predictors, ridge regression performs less well than the other three penalized methods, perhaps due to the former's inability for variable selection. LASSO and elastic net with $\alpha = 0.5$ are the winners.



(a) (b)

Figure 2.9: AUC calculated for the bipolar disorder test datasets with all SNPs across various values of λ. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each data across the values of λ.

2.5 Discussion

The primary objective of our study was to provide insight into general categories of models for which penalized regression improved disease risk prediction and classification for GWAS data. More specifically, we investigated the performance of MLE, LASSO, SCAD, ridge, elastic net and TLP regression methods for four different true models. The four models were chosen to represent broad categories defined by sparsity and strength of SNPs associated with disease. Two sparse models were considered with strong or moderate association strengths of only 10 causal SNPs. Two non-sparse models included 300 and 900 causal SNPs with weak effects respectively. Overall, we confirmed the commonly held belief that penalized regressions based on the model sparsity assumption, such as LASSO, SCAD, TLP and elastic net weighted towards its LASSO component were most

suitable for sparse true models. This was true for both risk prediction and discrimination. However, we did discover that when effect sizes were strong in a sparse model, MLE performed as well. An interesting result was about how the penalized regressions used the information (or lack of information) when many SNPs were considered, in particular SNPs that would not meet a strict genome-wide significance level. As a rule, if a various number of top SNPs ranked by their marginal association significance are allowed to enter into a model, the LASSO and SCAD regressions were able to detect and thus ignore many unassociated SNPs in sparse model settings, while ridge regression was able to outperform LASSO and SCAD for non-sparse models with many SNPs with only weak associations. This may be important going forward as non-sparse and polygenic models may hold for many common diseases and complex traits. For sparse models the TLP's performance was comparable to LASSO and SCAD, but it outperformed LASSO and SCAD, but not ridge, when the true model was non-sparse with many weakly associated SNPs. The elastic net demonstrated the value in both variable selection and continuous shrinkage features of a penalty as it was able to adapt to the true underlying model and yield the best or nearly the best performance of all penalties. It is noteworthy, though, that the elastic net did not uniformly outperform either TLP or SCAD, in particular the TLP performed best on the real Crohn's disease and bipolar disorder data in the modeling scenario where the number of input SNPs was varied.

We have focused on penalized regression methods, but Bayesian approaches (Guan and Stephens, 2011) are also potentially useful and worth further investigation, which however is beyond the scope of this paper.

The current statistical research on high-dimensional data has largely focused on sparse models, yielding many important and insightful results. Nonetheless, non-sparse models are also useful, as manifested by polygenic models for complex and common diseases. There are few theoretical studies on non-sparse models; an exception is the work of Cook et al. (2012) on dimension reduction. The main message of our study, certainly not new, is that different penalized methods may be more suitable depending on the underlying architecture of the true model: for example if the model is sparse or non-sparse. Hopefully this will prompt more empirical and theoretical investigations for non-sparse models.

Chapter 3

Does the Inclusion of Rare Variants Improve Risk Prediction?

3.1 Introduction

The potential number of lives impacted by successful early identification of patients at high risk for hypertension has motivated researchers across a spectrum of fields. On the frontier of risk prediction is the identification of genetic variants linked to traits such as high blood pressure. Advancements in sequencing have fostered the identification of a growing number of loci related to blood pressure. One such study performed by The International Consortium for Blood Pressure Genome-Wide Association Studies identified 29 SNPs related to systolic blood pressure (The International Consortium for Blood Pressure Genome-Wide Association Studies, 2011). A second compelling study concluded that perhaps as many as hundreds of SNPs affect blood pressure; moreover, rare variants (variants with minor allele frequency less than 5%) in addition to novel common variants (minor allele frequencies greater than 5%) are necessary to explain the relationship between allelic variants and blood pressure (Levy et al., 2009).

One promising tool that may be able to simultaneously leverage risk information in both common and rare variants is penalized regression. The range of available penalties allows researchers to estimate models with a mixture of two desirable properties: variable selection and proportional shrinkage of regression coefficients. The following work systematically measured the advantages of the different types of penalized regression

methods in the prediction of SBP using only common variants, only rare variants, or combinations of the two types of variants.

3.2 Methods

3.2.1 Data

The primary source for genotypic, phenotypic, and covariate data was Genetic Analysis Workshop 18 (GAW18) data files. GAW18 data is provided for approximately 1000 Mexican American individuals comprising 20 pedigrees enriched for type 2 diabetes. The pedigrees contained between 21 and 76 individuals. The phenotype of interest was the systolic blood pressure (SBP) measure from the first time point. Genotype data for more than 8,000,000 genome locations was derived from sequencing data for all odd numbered chromosomes, representing all sequencing data made available by GAW18. Approximately one third of the variants were common. The analysis accounted for the covariates age, gender, smoking status, and antihypertensive medication.

The pairwise correlation structure resulting from either a family structure or a cryptic population structure was removed using an estimate of the variance-covariance matrix. We estimated the variance-covariance structure as a function of the Identity-By-State (IBS) matrix calculated from all available GWAS data. Efficient Mixed-Model Association eXpedited (EMMAX) software (Kang et al., 2010) was used to obtain our IBS matrix estimate. For IBS matrix convergence it was necessary to exclude individuals missing more than 10% of genotypes (pre-imputation). Therefore, the final sample size for this study was 759.

3.2.2 Model

Let Y_i be the SBP value at the first examination for subject $i = 1, ..., n$, and define X_{ij} as subject i's minor allele count (0,1, or 2) for SNP $j = 1, ..., p$. Covariate information for subject i is notated by $X_{i,age}$ for age, $X_{i,gen}$ for gender, $X_{i,smoke}$ for smoking status, and $X_{i,med}$ for antihypertensive medication use. The effect of antihypertensive medication on blood pressure is not consistent across samples; thus it is not ideal to include patients using this medication. However, removing patients who used treatment medication from a diabetes enriched sample would have excluded a significant part of the GAW18 data. The authors chose to incorporate use of antihypertensive medication as a covariate in

order to account for medication use while minimizing assumptions about its impact on SBP. We assumed the following model relates the genotypic data to the phenotype: $Y = X\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Sigma)$. Here, $Y = Y_{n\times 1}$, a vector of the phenotype measurement for the n samples, $X = X_{n\times(1+4+p)}$, the design matrix for the genotype and covariate data including a column of ones for β_0 estimation, and ϵ is a $n \times 1$ vector of random errors. The vector of predicted phenotypes, $\hat{Y}$, is then equal to $X\hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimate (MLE) of the coefficient vector, β. More specifically,

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y = [(\Sigma^{-1/2}X)'(\Sigma^{-1/2}X)]^{-1}(\Sigma^{-1/2}X)'\Sigma^{-1/2}Y = [(X^*)'(X^*)]^{-1}(X^*)'Y^*$$

where $Y^* = \Sigma^{-1/2}Y$ and $X^* = \Sigma^{-1/2}X$. Thus, we can decorrelate our samples by premultipling both Y and X by $\Sigma^{-1/2}$. Kang et al. (2010) demonstrated that the variance-covariance matrix, $\hat{\Sigma}$, can be estimated effectively as a function of the Identity-by-State (IBS) matrix. Kang et al. showed the effectiveness of their method on both seemingly unrelated samples and samples with a substantial population structure. For Kang's method $\hat{\Sigma} = \sigma_g^2 K + \sigma_r^2 I_n$ where $\sigma_g^2 =$ genetic variance parameter, $\sigma_r^2 =$ residual variance parameter, and $K = I\hat{B}S$. We decorrelated our samples using the $\hat{\Sigma}^{-1/2}$ derived with the Kang et al. method. During preparation of the final manuscript, work appeared by Rakitsch et al. (2013) using a similar method to correct for population structures in a penalized regression approach to multi-marker association mapping. The present investigation studied a model of the new vector of decorrelated phenotypes, Y^*, as a function of the new genotype and covariate matrix, X^*. To be clear the model utilized in the current study is $Y^* = X^*\beta + \epsilon^*$ where $\epsilon^* \sim \mathcal{N}(0, \sigma^2 I_n)$. Note, $\sigma^2 \approx 1$.

We first consider the unpenalized regression model. MLE is asymptotically unbiased with fixed p as $n \to \infty$, but it may not be for a large p. One possible remedy is to introduce regularization or penalization on regression coefficients. We obtained predictions of Y^* by first obtaining $\hat{\beta}$, then $\hat{Y^*} = X^*\hat{\beta}$. For penalized regression methods $\hat{\beta}$ is found by maximizing a penalized log-likelihood (Friedman et al., 2008): $l(\beta) - \lambda P(\beta)$.

Candidate penalties that perform variable selection are LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and the truncated L_1-penalty (TLP) (Shen et al., 2012). Perform LASSO regression by applying the penalty $P(\beta) = \sum_{k=1}^{p} \mid \beta_k \mid$. The SCAD penalty, $P(\beta, \lambda)$, replaces $\lambda P(\beta)$ with $dP(\beta, \lambda)/d\beta = \sum_{k=1}^{p} \lambda sign(\beta_k)[I(|\beta_k| \leq \lambda) + (a\lambda - \mid \beta_k \mid)_+/(a-1)\lambda \cdot I(\mid \beta_k \mid > \lambda)]$ for $a = 3.7$. TLP regression uses $P(\beta) = \sum_{k=1}^{p} \min(|\beta_k|/\tau, 1)$ where $\tau > 0$ is a thresholding parameter, beyond which there is no further penalty. Regressions using these penalties are three methods to shrink many regression coefficient

estimates to 0, effectively selecting a subset of SNPs to be used for prediction. The variable selection feature can be of particular value in genetics settings such as ours where the number of true causative variants is likely a small fraction of the considered SNPs. If instead of variable selection, it is advantageous to proportionally shrink all regression coefficients, a candidate penalized regression method is ridge regression (Hoerl and Kennard, 1970). Ridge regression employs the penalty $P(\beta) = \sum_{k=1}^{p} \beta_k^2$. Elastic net penalized regression (Zou and Hastie, 2005) is a hybrid of the two approaches, with a penalty structure that is a mixture of the LASSO and ridge penalties controlled by a user specified mixing parameter, α, which is restricted to [0,1]. The elastic net penalty is $P(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$ where α is selected to match the desired balance of variable selection and coefficient shrinkage.

3.2.3 Implementation

We restricted our study to the top 1000 common variant SNPs and top 1000 rare variant SNPs as identified by the marginal significance of a Kruskal-Wallis test of the minor allele counts and SBP values for the 759 samples. The real data observations were randomly divided into equally sized training, tuning, and testing sets ($n = 253$ for each), and a sequence of models was then fit on the training set. The sequence was defined by incremental increases in both the penalty and penalty specific parameters (e.g. α and τ). The sequence of penalty (and tuning parameter when applicable) values used to fit the models spanned a range comprehensive enough to allow identification of the values which optimized performance for SCAD, LASSO, elastic net, and ridge regression. The additional tuning parameter, τ, used in TLP penalized regression greatly increased the computational time; therefore, the number of λ and τ pairs considered was constrained. The TLP results presented here likely underestimate the true performance of this method. In all penalized regressions the optimal penalty value was the one minimizing prediction error in the estimated tuning phenotypes when applying the regression coefficients estimated from the training model based on that penalty value.

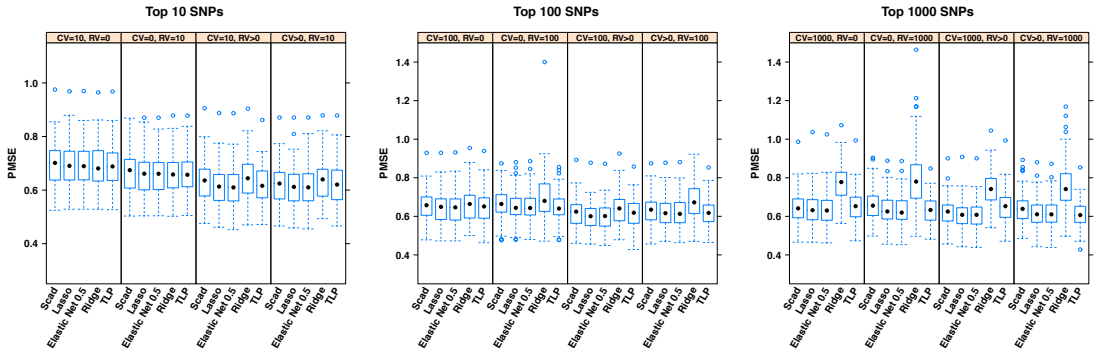
Models were fit in a directed way based on the number and type of variants. First, the authors examined only the top 10, 100, and 1000 most significant common variants. This examination was repeated using only the top 10, 100, and 1000 rare variants. Next, 1, 10, 100, and 1000 of the complimentary type of variant were added to the

model. For example, after fitting a model with only the top 10 common variants, four models were fit using these same 10 common variant SNPs *and* the top 1, then top 10, then top 100, and finally the top 1000 rare variants. The formal assessment of the regression methods was done by applying the training coefficients corresponding to the optimal penalty to the testing data. This process of randomly dividing the real data set into training, tuning, and testing sets, and then investigating the predictive performance of penalized regression methods was repeated 100 times, as a form of cross-validation. The regression approaches were compared using predictive mean squared error (PSME). Define PMSE= $\sum_{i=1}^{n}(\hat{Y}_i^* - Y_i^*)^2/n$. OLS, SCAD, LASSO, elastic net, and ridge regression estimates were generated using R packages `glmnet` (Friedman et al., 2008) and `ncvreg` (Breheny and Huang, 2011). TLP estimates were obtained using FGSG: Feature Grouping and Selection Over an Undirected Graph in Matlab (Yang et al., 2012).

3.3 Results

Descriptions of the predictive mean squared error (PMSE) of Y^* from the 100 randomly created testing data sets are presented in Figure 3.1 and Table 3.1. Figure 3.1 provides box plots for the predicted mean squared errors obtained using the different types of regression on the 100 data sets. The intent of Figure 3.1 is to provide an assessment of differences and reductions in PSME for different regression penalization methods within and between inputted SNP scenarios. Figure 3.1(a) presents results from models where fitting was based on the top 10 SNPs for each of the variant types. Figure 3.1(b) presents results where fitting was based on the top 100 SNPs for each of the variant types, and Figure 3.1(c) presents results where fitting was based on the top 1000 SNPs for each of the variant types. In each figure the first two columns represent models using only common variants (CVs) or only rare variants (RVs). The third column provides PMSEs of Y^* for the best model using the fixed number of common variants and either 1, 10, 100, or 1000 RVs. For example, the column labeled "CV=10,RV>0" gives the smallest PSME from the four models using exactly the top 10 common variants and either the top 1, 10, 100, or 1000 rare variants. Similarly, the fourth column describes the model with the smallest PMSE using the fixed number of rare variants and either 1, 10, 100, or

1000 CVs. Figures 3.1(a), 3.1(b), and 3.1(c) are plotted on the same scale to facilitate comparisons across them. Table 3.1 gives the median PMSE for the twelve modeling scenarios across the 100 data sets. Please note the OLS predicted mean squared errors are not presented in Figure 3.1 due to their relative size.



(a) (b) (c)

Figure 3.1: Boxplots of the Median Predicted Mean Square Errors (PMSEs) for Y^* calculated from the 100 Randomly Generated Testing Sets.

Table 3.1 gives the median PMSE for the twelve modeling scenarios across the 100 data sets.

Regression Method	Top 10 SNPs				Top 100 SNPs				Top 1000 SNPs			
	CV Only	RV Only	CV=10,RV>0	CV>0,RV=10	CV Only	RV Only	CV=100,RV>0	CV>0 ,RV=100	CV Only	RV Only	CV=1000,RV>0	CV>0, RV=1000
OLS	3.723	0.719	1.856	0.722	107.775	24	98.109	21.748	370644.875	311336.868	37064.875	63274.28
SCAD	0.701	0.674	0.636	0.625	0.657	0.664	0.625	0.635	0.641	0.656	0.625	0.639
LASSO	0.691	0.661	0.613	0.612	0.649	0.644	0.601	0.616	0.632	0.625	0.608	0.611
Elastic Net ($\alpha = 0.5$)	0.689	0.661	0.610	0.610	0.646	0.643	0.601	0.613	0.630	0.619	0.608	0.610
Ridge	0.681	0.658	0.644	0.640	0.664	0.680	0.641	0.672	0.778	0.780	0.742	0.741
TLP	0.688	0.657	0.616	0.621	0.652	0.641	0.618	0.617	0.653	0.633	0.653	0.607

Table 3.1: Median Predicted Mean Square Errors (PMSEs) for Y^* calculated from the 100 Randomly Generated Testing Sets.

It is evident from Table 3.1 that penalized regression methods outperform OLS regardless of the number or type of candidate variants. Fixing the type of penalized regression and the number of top SNPs considered for the model allows us to uncover that rare variant only models usually outperformed common variant only models. The difference was small, though. The central question to be answered by this work was

whether adding rare variants to common variant models improved SBP prediction. We found that for penalized regression models the inclusion of at least 1 of the complementary type of variant improved or maintained the performance of the model. This was true whether we fixed 10, 100, or 1000 top SNPs, added common variants to rare variant only models, or added rare variants to common variant models. Again, the differences were small; however, small but perceptible shifts in the overall distributions as presented in Figure 3.1 support this conclusion.

Comparisons across models based on the top 10, top 100, and top 1000 SNPs revealed an interesting pattern. As the number of candidate SNPs increased, the sparse SCAD, LASSO, and TLP penalties were generally superior to the non-sparse ridge penalty. Differences were small, at most 0.1555 mmHg, and need confirmation on different SBP real data sets. The conclusion should also be corroborated with simulated SBP data sets generated from genetic models reflecting a comprehensive range of possible SBP genetic architectures. Further, while reductions in PMSE occurred within the same variant composition across the three top SNP groupings (e.g. comparing CV Only for the Top 10 to CV Only with the Top 100 SNPs), the gains were often less than those made by just adding the complementary type of variant to the model. Combined, these two results suggest that the true number of strong causative variants is at most moderate and includes both rare and common variants. Ridge regression was the best or nearly identical to the best penalty choice when only the top 10 common or rare variants were used, indicating that all of these top variants are integral in understanding the association between genotypes and SBP. TLP was a top performer with models using only the top 10 or top 100 rare variants. As more SNPs of any type were included, the elastic net equally weighted to LASSO and ridge was generally superior. That is, there was a need for a selection element to distinguish noise from true effect, and there was a need for a non-sparse penalty feature to still incorporate larger numbers of SNPs in the regression model. This perhaps indicates that beyond a small set of strong causative SNPs, there are many SNPs that are truly associated with the outcome, but the majority of them have small marginal effects sizes. This could prove important when considering that previous research has found at least 29 causative SNPs; thus, undiscovered variants associated with SBP may have at most moderate effect sizes.

3.4 Discussion

The strongest conclusion can be drawn about the effect of including rare variants in addition to common variants when predicting systolic blood pressure. The PMSE was reduced by up to 11.5%, and generally reduced between 4% and 9%, when rare variants were added to common variant only penalized regression models. This was true when any of 10, 100, or 1000 top SNPs were used. PMSE comparisons of single variant type models to combined variant type models revealed that both rare and common variants explain variance in SBP. Every penalty considered in the study improved SBP prediction over OLS. This was true whether estimation used only common variants, used only rare variants, or used both types of variants. The elastic net penalized regression was best at leveraging the information in the additional SNPs (rare or common), and produced the best overall models (again the absolute reduction in PMSE was too small to be statistically significant because of the variance in the PSME median distributions.). Caution when making conclusions about the TLP is needed because of the limited number of combinations of λ and τ studied due to time constraints. The results here likely understate the performance of TLP, thus the small gains from using TLP with the top 10 and top 100 rare variants warrant future analysis for possible confirmation. Work on the genotype-hypertension map should specifically consider rare and common variants. The interesting result that a hybrid penalty with both selection and proportional shrinkage components performed best hints at an underlying architecture where numerous SNPs with moderate main effects are interrelated in how they are associated with blood pressure. Overall, results presented here provide evidence that penalized regression, especially a hybrid of LASSO and ridge regression, can be used to improve SBP prediction.

Chapter 4

A Novel Statistic for Global Association Testing based on Penalized Regression

4.1 Introduction

Genetic variants naturally partition into meaningful structures, such as genes. As a result the variants within a specific gene may be a related set of genetic predictors that conjointly cause a biologic function such as disease. Consequently, it is essential to develop statistical tools such as association tests to identify a true link between a group of variants and disease outcomes. The challenge of manufacturing a powerful test is rich as its power can be impacted by genomic properties like the rarity of a variant's mutation, the size of the variant's effect, the number of null variants tested, and the linkage disequilibrium (LD) of the variants. In the following we first develop a new elastic net-based test statistic. We then use Genetic Analysis Workshop 18 (GAW 18) data to assess the power of the corresponding global association test to capture a relationship between an aggregated group of variants and hypertension. GAW18 provides 200 sets of simulated disease outcomes from the real genomes of nearly 1000 individuals (approximately 150 unrelated). We find that there are genetic settings determined by genome regions where the new test's power exceeds existing tests. More, the novel test

provides specific information about the relationship between individual variants and the outcome, something usually not available with existing methods.

Frame the exploration for powerful association tests by letting $Y_i = 0$ or 1 be a binary response variable for subject $i = 1, ..., n$, and denote the values of m ordinal genetic predictors for subject i with a $m \times 1$ vector $X_i = (X_{i1}, ..., X_{im})^T$. Describe the map from genetic predictor to disease with main effects logistic regression:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_j. \tag{4.1}$$

True associations in this framework can be quantified with a global test on the null hypothesis $H_0 : \beta = (\beta_i, ..., \beta_m)^T = 0$ versus a general alternative $H_1 : \beta \neq 0$ (Goeman et al., 2004). Here, the global test will assess the joint impact of a set of m predictors on the disease.

Because single variant tests have successfully identified many strong sources of heritability for common disease, an intuitive approach to the global testing of a set of variants is to combine or pool the results from the individual tests. One common approach is to use the minimum p-value from score statistic derived tests of individual variants. Seaman et al. (2005) effectively applied a direct simulation approach to minimum p-value calculations describing the associations between candidate genes and treatment response to anti depression drugs. The flexible combined statistic (CS) is another example allowing a test of no associated markers in a prespecified functionally related genome region (De la Cruz et al., 2009). Here, user defined truncation thresholds are applied to individual variant p-values that are then combined multiplicatively. The method is flexible and includes the Truncated Product method of Zaykin et al. (2002) which uses a fixed truncation threshold for all variants. The framework also allows for the Rank Truncated Product method of Dudbridge and Koeleman (2003) which uses a fixed number of p-values. The authors found the method increases power for testing a set of variants by allowing the inclusion of moderate signals, whereas the single variant tests find strong individual associations. Yu and colleagues (2009) developed the Adaptive Rank Truncated Product (ARTP) method as another way to combine the p-values from tests on individual variants like SNPs. The ARTP-based conclusions mirror those of De la Cruz et al., but the authors also discuss how power of a global null hypothesis may not increase if the marginal effects of the variants are weak while the interaction effect

is strong. The cohort allelic sums test (CAST) takes a different approach to aggregating single variant information, testing the difference in the sums of allelic mutation frequencies in a functional region between affected and unaffected populations (Morgenthaler et al, 2007). CAST was robust in identifying associations between common diseases and suspected risk genes. CAST, though, is skewed toward common variants; so, Madsen and Browning (2009) developed an alternative methodology comparing the sums of mutations in a group of variants. Their method weights the sums by the mutation frequencies in the unaffected samples. The method showed favorable power compared to existing similar methods, but the authors remark that their statistic's power gains are at the cost of generality and require correct determinations of the disease risk allele (Madsen and Browning, 2009). The numerous limitations of tests which depend on single variant statistics make it desirable to find alternative approaches. One such approach is to model the effect of each variant simultaneously and then test for a group level association with a disease outcome.

A uniformly most powerful unbiased (UMPU) test for H_0 is not possible; however, a scaled version of $\beta = \delta b/\sqrt{m}$ will permit a most powerful (MP) test (Cox and Hinkley, 1974). Here b is a fixed vector and δ is a scalar. The corresponding test statistic is $T_{MP} = b^T U$ where U is the score vector; thus, substituting $\hat{\beta}$ for b will provide estimates of the most powerful (EMP) test statistic, T_{EMP}. A number of existing test statistics have been used to estimate T_{MP} in the genetics context. For example, Schaid et al. (2002) used score statistics for haplotype analysis, and Pan (2009) developed two tests, SSU and SSUw, to overcome situations where the covariates have opposite causal effects on the trait.

Limited research, though, has attempted to find powerful global tests based on penalized regression; for example, using methodology based on the least absolute shrinkage and selection operator penalty (LASSO) (Tibshirani, 1996) or ridge regression (Hoerl and Kennard, 1970). The value of penalized regression features such as dimension reduction is shown in a setting with both primary and secondary variables Martinez et al. (2010). Reducing the dimension of secondary variables with Adaptive LASSO revealed the possibility of increasing the power of a score test on the coefficients of primary variables; however, the penalized regression applied only to variable selection and not association testing. In a genome-wide association setting, a selection procedure based

on LASSO was underpowered compared to marginal regression based methods (Alexander and Lange, 2011). The Gene set Ridge regression in Association Studies (GRASS) methodology generates a test statistic for the association between related sets of genes and a disease (Chen et al., 2010). GRASS regularizes with ridge regression between genes and LASSO penalized regression within genes. Using GWAS data (variants with minor allele frequencies $\geq 5\%$) the method is effective at identifying an association between a group of genes, e.g. those in a pathway, and a disease, especially if several of the genes are causal. There is evidence that statistical association tests using a subset of variants selected by LASSO penalized regression do not have more power than existing tests like the SSU and SSUw on the full set (Basu et al., 2011). The Basu et al. statistics used LASSO logistic regression to perform variant selection, and then the global association test used SSU and score statistics based on the components of the score vector corresponding to the selected variants. The present work uses penalized regression as in the Basu approach, but directly employs the model estimated by penalized logistic regression.

Our novel approach uses the elastic net penalty (Zou and Hastie, 2005). We assess the statistical significance of the estimated coefficients through the estimated $T_{EMP} = \hat{\beta}^T U$. In this paper we examine two genomic regions within the GAW18 data. We utilize coefficient estimates for variants in the these regions using a spectrum of elastic net penalties to show situations where gains in power are possible through the use of our penalized regression approach. Specifically, we quantify the power of global hypothesis tests using only a region's rare variants, using only a region's common variants, and considering both types of variants in a region. Rare variants are defined as those with minor allele frequency (MAF) less than or equal to 5%, and common variants are the complement (MAF>5%). The power values from our method are compared to power values obtained from a set of unpenalized logistic regression based test statistics and those using the Basu LASSO penalized regression testing approach. It is notable that our method has the added advantage of providing information about the association of individual variants. We provide information on the identification of true and false positive risk relationships, showing how the penalized regression methods can in fact provide insight into which variants are associated with the outcome.

4.2 Methods

4.2.1 A New Test Statistic Based on Penalized Regression

The primary goal of this article is to demonstrate that a penalized regression based association test can be more powerful (while controlling for Type I error) than either a maximum likelihood based approach or one of a representative set of well-established global tests. As describe previously, the most powerful test statistic is a function of the coefficient estimates of β.

To be complete in our description, we briefly describe how estimates are derived with penalized regression. Penalized regression based $\hat{\beta}$ are obtained by maximizing the penalized log-likelihood (Friedman et al., 2008) for a given set (α, λ):

$$l(\beta_0, \beta) - \lambda P_\alpha(\beta). \tag{4.2}$$

In our problem, the first term results from equation (4.1)

$$l(\beta_0, \beta) = \sum_{i=1}^{n} Y_j(\beta_0 + X_i^T\beta) - \log[1 + \exp(\beta_0 + X_i^T\beta)], \tag{4.3}$$

and the second term, $P_\alpha(\beta)$, denotes the elastic net penalty (Zou and Hastie, 2005)

$$P_\alpha(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1. \tag{4.4}$$

The elastic net penalty includes a mixing parameter $\alpha \in [0,1]$. The α is selected by the user to match the desired balance of simultaneous coefficient shrinkage ($\|\beta\|_2^2$) and variable selection ($\|\beta\|_1$). Ridge regression corresponds to $\alpha = 0$; therefore, more emphasis is placed on simultaneous coefficient shrinkage as $\alpha \to 0$. When $\alpha = 1$ the elastic net penalty becomes the LASSO penalty, meaning that the penalty features increased variable selection more as $\alpha \to 1$. We use the elastic net penalty in the present examination because it provides a framework to easily explore the relationship between power and different mixtures of its two desirable features by simply considering different values of α.

Evident from equations (4.2) and (4.4), finding the coefficient estimates necessary for calculating any T_{EMP} using elastic net penalized regression requires user-specification of the mixing parameter α and the tuning or penalty parameter λ. We wanted to study

the power and type I error for a range of penalties weighted towards ridge or LASSO regression to various degrees. Therefore, we examined and present results for a sequence of five evenly spaced α from [0,1]; that is, $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. An optimal λ is not known in advance. As will be shown shortly, our new statistic overcomes this issue by combining sets of coefficient estimates from models fit using different λ in a very specific way. That is, we do not need to try to specify a single best λ, but can instead specify a k-sized candidate set, $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$, whose range and refinement likely describe the overall performance of the penalty. The `R` package `glmnet` (Friedman et al., 2010) was used to obtain regression coefficient estimates. For each of the α we fit models for $k = 20$ λs covering a range determined automatically by the `glmnet()` function. The choice of $k = 20$ provided sufficiently refined candidate sets to capture a wide range of the performance possible given α. The result is that our method is to a noticeable extent robust to the choice of λ. To summarize up to this point in our process, we have only created $k = 20$ sets of $\hat{\beta}$s for a given data set and one of five user specified αs using the default settings of the highly regarded `glmnet` software. The key to our method is to take these sets, the result of limited user input, and aggregate them into a test statistic that is similarly free of the need for further user specification.

To see how we calculate the test statistic for our global test, it is helpful to first consider a single grid point (α, λ_1) representing the smallest λ used for one specified α. Using equation (4.2) the estimate of β is $\hat{\beta}_{(\alpha,\lambda_1)} = \arg\max_{\beta_0,\beta}\{l(\beta_0, \beta) - \lambda_1 P_\alpha(\beta)\}$ where $P_\alpha(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$. Extending the conclusions of Cox and Hinkley, the most powerful test statistic for this single grid point (α, λ_1) is

$$T_{(\alpha,\lambda_1)} = \left| \frac{\hat{\beta}_{(\alpha,\lambda_1)} U}{\|\hat{\beta}_{(\alpha,\lambda_1)}\|_2} \right| \tag{4.5}$$

where $U = \sum_{i=1}^{n}(Y_i - \bar{Y})X_i$.

It is important to remark that the desired statistical significance summaries of the penalized regression approach is for the specified α; that is, we want to summarize the performance models for all λ used with this penalty (determined by α). Descriptions at the α level require aggregation of the coefficients found with the k elements of $\Lambda = \{\lambda_1, ..., \lambda_k\}$. To our knowledge aggregations of the test statistics, notated generically as $T_{(\alpha,\Lambda)}$, do not have a parametric asymptotic distribution; therefore, a permutation based empirical p-value was derived to quantify power and Type I error rates. The calculation

and statistical significance of $T_{(\alpha,\Lambda=\{\lambda_1,...,\lambda_k\})}$ and the resulting decision can be found using B permutations in the following manner.

Step 1: Set α.

Step 2: Generate with the corresponding elastic net penalty k sets of coefficient estimates labeled $(\hat{\beta}_{(\alpha,\lambda_1)},...,\hat{\beta}_{(\alpha,\lambda_k)})$.

Step 3: Calculate the corresponding k *original* test statistics $T_{(\alpha,\lambda_1)}$ to $T_{(\alpha,\lambda_k)}$ with equation (4.5)

Step 4: Randomly permute B times the vector of the binary outcomes, Y. Label these permuted outcome vectors $Y^{(b)}$ with $b = 1,...,B$.

Step 5: Using the exact same $\Lambda = \{\lambda_1,...,\lambda_k\}$ and the same elastic net penalty, generate $\beta^{(b)} = (\beta^{(b)}_{(\alpha,\lambda_1)},...,\beta^{(b)}_{(\alpha,\lambda_k)})^T$ from $Y^{(b)}$ and the original X, the $n \times m$ covariate matrix.

Step 6: Calculate $T^{(b)}_{(\alpha,\lambda_1)}$ to $T^{(b)}_{(\alpha,\lambda_k)}$ for each $b = 1,...,B$ with equation (4.5)

Step 7: Calculate p-values for each of the k scalars $T_{(\alpha,\lambda_1)}$ to $T_{(\alpha,\lambda_k)}$ as

$$P_{(\alpha,\lambda_1)} = \frac{\sum_{b=1}^{B} I(T_{(\alpha,\lambda_1)} > T^{(b)}_{(\alpha,\lambda_1)})}{B} \text{ to } P_{(\alpha,\lambda_k)} = \frac{\sum_{b=1}^{B} I(T_{(\alpha,\lambda_k)} > T^{(b)}_{(\alpha,\lambda_k)})}{B}$$

and select minP=min($P_{(\alpha,\lambda_1)}$,...,$P_{(\alpha,\lambda_k)}$)

Step 8: For each permutation $b = 1,...,B$ calculate p-values for each of the k scalars $T^{(b)}_{(\alpha,\lambda_1)}$ to $T^{(b)}_{(\alpha,\lambda_k)}$ as

$$P^{(b)}_{(\alpha,\lambda_1)} = \frac{\sum_{h\neq b} I(T^{(b)}_{(\alpha,\lambda_1)} > T^{(h)}_{(\alpha,\lambda_1)})}{B-1} \text{ to } P^{(b)}_{(\alpha,\lambda_k)} = \frac{\sum_{h\neq b} I(T^{(b)}_{(\alpha,\lambda_k)} > T^{(h)}_{(\alpha,\lambda_k)})}{B-1}$$

and select $\text{minP}^{(b)}$ =min($P^{(b)}_1$,...,$P^{(b)}_k$)

Step 9: Calculate the p-value for $T_{(\alpha,\Lambda=\{\lambda_1,...,\lambda_k\})}$ as

$$P_{T(\alpha,\Lambda)} = \frac{\sum_{b=1}^{B} I(\text{minP} < \text{minP}^{(b)})}{B}.$$

Step 10: If $P_{T(\alpha,\Lambda)} \leq 0.05$ the decision is to reject H_0.

The new test statistic's decisions will be compared to those from existing global tests detailed in the following section. Even when these tests have an asymptotic distribution, we compare to the permutation p-value for each of these tests to maintain the same context. Here, the permutation p-value is the percentile of the p-value calculated from the original data set compared to those from the B permuted data sets.

4.2.2 Other Global Tests

In this section we will provide background on five unpenalized logistic regression global tests and a set of LASSO based global test of $H_0 : \beta = (\beta_i, ..., \beta_m)^T = 0$ versus $H_1 : \beta \neq 0$.

First to describe the five unpenalized tests: the score test, SSU, SSUw, Sum, and UminP. The tests are a function of model (1)'s score vector. For reference, under H_0 the score vector, denoted U, and its covariance matrix, denoted V, equal

$$\begin{aligned} U &= \sum_{i=1}^{n}(Y_i - \bar{Y})X_i, \\ V &= Cov(U) = \bar{Y}(1-\bar{Y})\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \end{aligned}$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ and $\bar{X} = \sum_{i=1}^{n} X_i/n$.

Score test

The multivariate score test statistic is

$$T_{score} = U^T V^{-1} U,$$

where V^{-1} is the inverse of the variance matrix of the score vector (generalized inverse if it is not of full rank). Under H_0, $T_{Score} \sim \chi^2(rank(V))$.

It is common practice with case-control studies to use a likelihood-ratio test statistic to assess the association between a set of variants and disease status. The multivariate score statistic is asymptotically equivalent without requiring potentially time consuming computation of the maximum-likelihood estimates of β (Schaid et al., 2002).

UminP

In high-dimensional data the score test may lose power or it may be difficult to estimate the covariance matrix V. Therefore, it may be desirable to test each covariate univariately and aggregate the results from these marginal tests. One such way is to use the UminP test which equals the minimum of the m p-values resulting from a univariate test of each predictor. The UminP test statistic is

$$T_{UminP} = max_{j=1,2,...,m} U_j^2/v_j,$$

where U_j is the j^{th} element of U and v_j is the $(j,j)^{th}$ diagonal element of V. Conneely and Boehnke (2007) showed that the asymptotic p-value could be calculated by integration of the multivariate normal density because a vector of score statistics has asymptotic normality.

Sum test

First, make a working assumption that $\beta_1 = \cdots \beta_m \equiv \beta_c$ such that equation (4.1) can be written:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_j = \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_c.$$

Then test $H_0 : \beta_c = 0$ with

$$T_{Sum} = \hat{\beta}_c^2 / V_c$$

where $\hat{\beta}_c = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 \hat{\beta}_{M,j}}{\sum_{i=1}^{n}(\sum_{j=1}^{m} X_{ij})^2}$ and $\hat{\beta}_M$'s are the MLEs of the β_M's of the marginal linear regression models. Under H_0, $T_{Sum} \sim \chi^2(1)$. The sum test has asymptotically correct test size when H_0 is true. More, β_C is likely non-zero when H_1 is false, leading to good power (Pan, 2009).

SSU and SSUw

A problem with the sum test is that power is reduced when the direction of the effect size varies; that is, when the signs of the β_j are both positive and negative. Two tests called SSU and SSUw overcome this issue by utilizing the sum of the squares of the marginal score statistics (Pan 2009). The SSU test statistic is

$$T_{SSU} = U^T U.$$

SSUw is based on $V_d = Diag(V)$ and is defined

$$T_{SSUw} = U^T V_d^{-1} U.$$

Under H_0, T_{SSU} and T_{SSUw} have quadratic forms and approximately follow a mixture of χ_1^2, which can be approximated by a scaled and shifted chi-sqared distribution (Pan 2009).

Tests based on LASSO logistic regression

Basu et al. (2011) created LASSO penalized regression based score and SSU statistics, hypothesizing that dimension reduction would increase power (2011). First, with a user specified penalty parameter, λ, generate the LASSO estimate, $\hat{\beta}_L(\lambda)$, using:

$$\hat{\beta}_L(\lambda) = \arg\max_{\beta}\{\log L(\beta) - \lambda\|\beta\|_1\}.$$

Next, let $U(\lambda)$ be the component of the score vector corresponding to the nonzero components of $\hat{\beta}_L(\lambda)$. As will be explained next, the Basu et al. test statistics are created using $U(\lambda)$ and $V(\lambda)$, the respective submatrix of V, through

$$T_{SSU}(\lambda) = U(\lambda)^T U(\lambda) \text{ and } T_{Sco}(\lambda) = U(\lambda)^T V^{-1}(\lambda) U(\lambda).$$

Three SSU (Score) type statistics were created from a k-sized candidate set of penalty parameters, $\Lambda = \{\lambda_1, \cdots, \lambda_k\}$. Calculate and standardize $T_{SSU}(\lambda_i)$ for $i = 1, \cdots, k$. Permutation based p-values for all λ_i in Λ were combined or averaged with three methods: minimum (Min) p-value, Fisher's method (1932), and the truncated product method or TPM (Zaykin et al., 2002). Specifically,

$$\begin{aligned} T_{Ave,SSU,Min} &= \max_{\lambda_i \in \Lambda} P_{SSU}(\lambda_i) \\ T_{Ave,SSU,Fisher} &= \prod_{\lambda_i \in \Lambda} P_{SSU}(\lambda_i) \\ T_{Ave,SSU,TPM} &= \prod_{\lambda_i \in \Lambda} P_{SSU}(\lambda_i)^{I_{(P_{SSU}(\lambda_i) \le \alpha_0)}} \text{ with } \alpha_0 = 0.05 \end{aligned}$$

Label the corresponding Score versions of the three statistics: $T_{Ave,Score,Min}$; $T_{Ave,Score,Fisher}$; and $T_{Ave,Score,TPM}$.

4.3 GAW18

The Genetic Analysis Workshop 18 (GAW18) collected and provide genotypic, hypertension phenotypic, and covariate data for approximately 1000 Mexican American individuals comprising 20 pedigrees enriched for type 2 diabetes. Genetic variant data for more than 8,000,000 genome locations was derived from sequencing data for all odd numbered chromosomes. The ratio of rare to common variates is approximately 2:1. Phenotypic

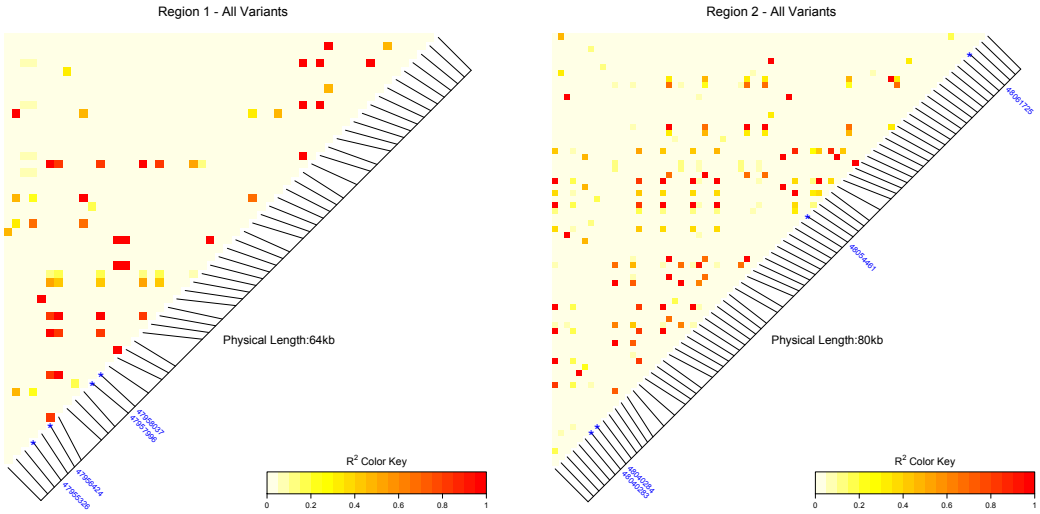
variables included systolic blood pressure (SBP), diastolic blood pressure (DBP), and a binary hypertension (HT) classification, all reported for up to four exams over a 30 year period.

Additionally, GAW18 provides 200 sets of simulated hypertension outcomes from the true genome and covariate data. The outcomes were generated with reported effect sizes from a known subset of nearly 1500 variants. MAP4 on chromosome 3 was the gene given the strongest effect on the phenotype with 15 selected variants (of 894) explaining about 7 percent of the variance. Therefore, we examined the global tests for quantifying the significance of association tests between regions on this gene and the samples' first simulated hypertension outcome. Of import to the following analysis, GAW18 identified nearly 150 individuals that are unrelated. The results described below are for only these samples.

Our investigation focused on two regions chosen because they contain both rare and common variants causal variants as well as a range of correlation strengths among variants in the region. More, the minor allele frequencies of the variants and correlation within and between causal and null variants exhibit interesting patterns and differences across the regions. Each region contains four causal variants, but the number of rare variants is different. There are two rare variants in Region 1, and three variants are rare in Region 2. Region 1 contains 62 null variants (58 rare) while Region 2 has 77 null variants (65 rare). Each region had exactly two of the 55 variants most affecting on SBP. To be clear the selected regions are intended to provide examples and insight about when the various global tests could leverage the information in multiple variants (and rare variants in particular) to increase power.

The correlation structures for the full sample are presented in Figure 4.1. The left panel shows the pairwise LD in r^2 for Region 1, and the right panel shows the LD plot for Region 2. Please note that case and control correlation structures are omitted because the disease statuses varied by simulation. The causal variants are identified with stars and labeled along the center diagonal in each plot. The causal variants in Region 1 exhibit a range of correlation strength from mild to strong with both the null and other causal variants. In moderate contrast the casual variants in Region 2 appear to have minimal correlation with both the null variants and with each other. Studying these two regions will thereby provide some insight into how power is affected when the causal

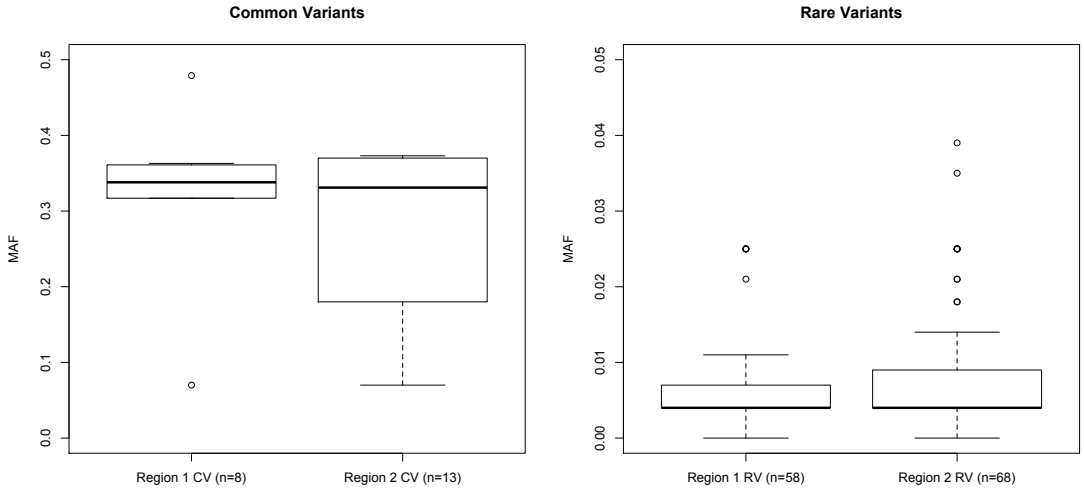
variants are providing various degrees of the same information.



(a) (b)

Figure 4.1: Pairwise LD in r^2 with notated true causal variants for (a) MAP4 Gene Region 1 (b) MAP4 Gene Region 2

The minor allele frequencies (MAFs) of the four variants in Region 1 were 0.7%, 35.9%, 2.1%, and 31.7%. Region 2 MAFs equaled 2.5%, 2.1%, 7.4%, and 0.4%. Note that the common causal variant's MAF is close to the 5% threshold. Figure 4.2 (a) provides a box plot of the MAFs for each region's common variants and (b) rare variants. Interestingly, Region 2 has a more rare variants with MAF in the 1% to 5% than Region 1 and more common variants with MAF between 5% and 30%.



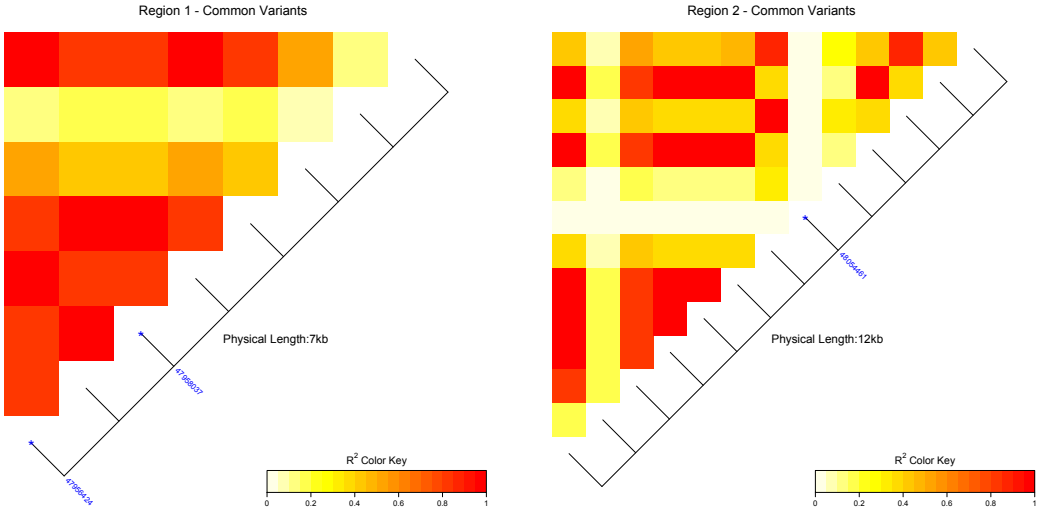
(a) (b)

Figure 4.2: Minor Allele Frequency (MAF) boxplots by region for MAP4 Gene (a) *common* variants and (b) *rare* variants

When considering the MAFs and LD structures, Region 1 has relatively stronger correlation associated with the causal variants and more with MAFs further from 5%. Region 2 shows much less correlation associated with causal SNPs but more variants of each type closer to the 5% MAF boundary. With the literature stressing a need for insight specific to variant type, we begin by partitioning the regions into rare and common variants. Next, we investigated the power of association tests for subsets of each variant type: only Region 1 common variants (section 4.3.1), only Region 2 common variants (section 4.3.1), only Region 1 rare variants (section 4.3.2), and only Region 2 rare variants (section 4.3.2). To conclude we report power results at a 5% level for the full regions: Region 1 combined common and rare variants (section 4.3.3), and Region 2 combined common and rare variants (section 4.3.3). For the full Region analyses, we show summaries of the true and false positive rates calculated using the variants selected via the penalized regression element of our new association test.

4.3.1 Common variants

The LD structures are noticeably different among the common variants in the two MAP4 regions (Figure 4.3). The correlation between the markers in Region 1 is noticeably stronger (Figure 4.3(a)). The two causal variants appear to be in a subset of strongly overlapping markers. While Region 2 does have variants in strong LD, there are also less correlated variants (Figure 4.3(b)). The most striking difference is how the causal variant in Region 2 has limited correlation with all other variants.



(a) (b)

Figure 4.3: Pairwise LD in r^2 with notated true causal *common* variants for (a) MAP4 Gene Region 1 (b) MAP4 Gene Region 2

Table 4.1 provides the power and median p-values [interquartile range] by region for the various global association tests at a 5% level for the 200 GAW18 data sets of common variants. The first five rows present results for our new penalized regression based testing approach. The next six give results for the association tests based on LASSO logistic regression developed by Basu et al. The final set of results, on the last five rows, shows the breakdown for the five tests that do not use penalized regression.

Test Category	Test Statistic	Region 1		Region 2	
		Power	Median P-value [Q1,Q3]	Power	Median P-value [Q1,Q3]
New Penalized Regression Based Test	$T_{(Ridge)}$	0.055	0.294 [0.168,0.451]	0.040	0.314 [0.184,0.502]
	$T_{(\alpha=0.25)}$	0.110	0.212 [0.116,0.406]	0.120	0.264 [0.131,0.471]
	$T_{(\alpha=0.50)}$	0.125	0.198 [0.106,0.386]	0.125	0.266 [0.117,0.462]
	$T_{(\alpha=0.75)}$	0.140	0.190 [0.101,0.379]	0.115	0.274 [0.116,0.452]
	$T_{(LASSO)}$	0.145	0.196 [0.088,0.36]	0.075	0.278 [0.143,0.492]
Tests Based on LASSO Logistic Regression	$T_{Ave,SSU,Min}$	0.120	0.300 [0.108,0.556]	0.085	0.396 [0.178,0.648]
	$T_{Ave,SSU,Fisher}$	0.030	0.349 [0.186,0.540]	0.030	0.040 [0.252,0.617]
	$T_{Ave,SSU,TPM}$	0.030	0.138 [0.122,0.152]	0.030	0.224 [0.196,0.250]
	$T_{Ave,Score,Min}$	0.215	0.189 [0.066,0.414]	0.115	0.312 [0.116,0.565]
	$T_{Ave,Score,Fisher}$	0.220	0.194 [0.062,0.434]	0.100	0.278 [0.138,0.536]
	$T_{Ave,Score,TPM}$	0.240	0.152 [0.058,0.174]	0.100	0.264 [0.154,0.292]
Tests without Penalized Regression	T_{Score}	0.215	0.181 [0.067,0.391]	0.080	0.260 [0.120,0.470]
	T_{SSU}	0.035	0.384 [0.213,0.519]	0.025	0.396 [0.241,0.564]
	T_{SSUw}	0.035	0.336 [0.200,0.455]	0.035	0.365 [0.225,0.525]
	T_{Sum}	0.060	0.426 [0.187,0.681]	0.025	0.538 [0.313,0.727]
	T_{UminP}	0.120	0.265 [0.104,0.467]	0.090	0.336 [0.172,0.571]

Table 4.1: Summary of Global Test P-values for 200 sets using only *common* variants

In Region 1 the tests incorporating the score statistic are most powerful. The top performer was the score statistic truncated product LASSO based test, but the Fisher and minimum P versions of the LASSO based score test, as well as the score test itself, all have similar power. All tests using a SSU statistics performed poorer than those using the score test in Region 1. If our elastic net has positive weight on the selection feature, then our method has power like tests using the minimum P. In these cases power was roughly half of the score based tests. As the selection feature is given more weight, then the power of our method increases, but is still less that the score test related values. Interestingly, though, the median p-value for our method was comparable to all score statistic based tests.

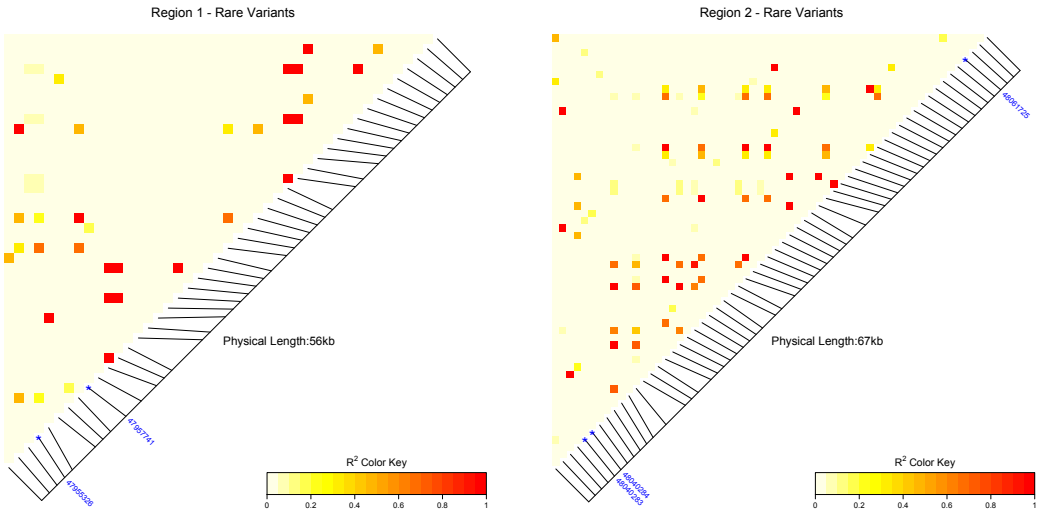
The number of causal variants decreases and the number of null variants increases in Region 2. Consequently, the penalized regression with a selection feature may be more effective relative to other tests. Table 4.1 shows how in Region 2 our method is now the most powerful when the elastic net includes both ridge and LASSO components ($\alpha \in (0, 1)$). The LASSO based score association tests, one of the other penalized

regression style tests, are the next best and have only slightly less power. Besides the score test, the tests without a penalized regression feature are less powerful, and even T_{score} performs worse than the best performer in each of the penalized regression groups. With and without penalized regression, the minimum P and score tests are more valuable than the SSU test. Overall, the main finding is that the more direct the use of penalized regression the better the power.

The global testing of common variants in the two Regions did not have good power in general. Our method shows the most promise in harnessing information in uncorrelated sets of variants. This possibly means that our test does a better job of aggregating or leveraging the information from multiple variants with limited overlap. Even in the scenario with strong LD between causal variants and null variants, our method had average performance across all methods.

4.3.2 Rare variants

Next, association testing on variants with $MAF \leq 5\%$ was investigated. Begin with the two region specific LD plots in Figure 4.4.



(a) (b)

Figure 4.4: Pairwise LD in r^2 with notated true causal *rare* variants for (a) MAP4 Gene Region 1 (b) MAP4 Gene Region 2

There is much less dependence among the rare variants than the common variants. This is true for both regions and the GAW18 selected causal markers in particular. The significance testing results in Table 4.2 show the promise of our new method when assessing a group of rare variants.

Test Category	Test Statistic	Region 1		Region 2	
		Power	Median P-value [Q1,Q3]	Power	Median P-value [Q1,Q3]
New Penalized Regression Based Test	$T_{(Ridge)}$	0.365	0.084 [0.022,0.228]	0.340	0.103 [0.032,0.281]
	$T_{(\alpha=0.25)}$	0.390	0.082 [0.018,0.211]	0.355	0.091 [0.029,0.229]
	$T_{(\alpha=0.50)}$	0.410	0.075 [0.017,0.194]	0.350	0.094 [0.025,0.224]
	$T_{(\alpha=0.75)}$	0.430	0.070 [0.017,0.183]	0.355	0.091 [0.023,0.218]
	$T_{(LASSO)}$	0.310	0.110 [0.035,0.248]	0.390	0.081 [0.024,0.232]
Tests Based on LASSO Logistic Regression	$T_{Ave,SSU,Min}$	0.060	0.456 [0.182,0.690]	0.085	0.352 [0.155,0.636]
	$T_{Ave,SSU,Fisher}$	0.407	0.086 [0.013,0.263]	0.281	0.160 [0.038,0.347]
	$T_{Ave,SSU,TPM}$	0.412	0.226 [0.012,0.566]	0.312	0.248 [0.033,0.428]
	$T_{Ave,Score,Min}$	0.166	0.230 [0.087,0.49]	0.136	0.260 [0.119,0.458]
	$T_{Ave,Score,Fisher}$	0.241	0.166 [0.058,0.322]	0.171	0.206 [0.089,0.384]
	$T_{Ave,Score,TPM}$	0.226	0.248 [0.062,0.487]	0.156	0.324 [0.121,0.500]
Tests without Penalized Regression	T_{Score}	0.030	0.285 [0.156,0.411]	0.015	0.284 [0.181,0.408]
	T_{SSU}	0.390	0.096 [0.020,0.262]	0.290	0.139 [0.039,0.358]
	T_{SSUw}	0.260	0.140 [0.048,0.260]	0.155	0.231 [0.085,0.402]
	T_{Sum}	0.185	0.272 [0.074,0.505]	0.100	0.384 [0.122,0.697]
	T_{UminP}	0.015	0.524 [0.375,0.662]	0.010	0.619 [0.376,0.725]

Table 4.2: Summary of Global Test P-values for 200 sets using only *rare* variants

In both regions our method provided the single most powerful test of all the tests ($\alpha = 0.75$ in Region 1 and LASSO in Region 2). Examining only Region 1 our hybrid versions of ridge and LASSO, the original SSU, and the Lasso based SSU tests using Fisher's method or the TPM essentially do as well. Interestingly, our LASSO test has less power than these tests, but it still outperforms all score, sum, and minimum p-value based tests. Besides the score test and any test using minimum p-values, the global tests have vastly more power to find associated rare variants in Region 1 than what was found testing only common variants.

The results in Region 2 indicated an increase in power from the most powerful of the five tests without logistic regression to the most powerful of the LASSO based tests. There is another increase for the most powerful of our new tests, the overall top performer is in this group (LASSO). The median [Q1,Q3] values provide supporting evidence of a shift in performance of the three categories of tests. Using our methodology with LASSO was best, but unlike the common variant results, using our approach with ridge did not cause a dramatic loss of power. Tests using the score or minimum p-value statistics

underperformed comparatively. It is also noteworthy that in this larger set of candidate variants, any test with penalized regression has at least the power and usual more than the power of test on the smaller common variant candidate sets. This was not true for the unpenalized association tests.

Comparing the two regions' power results, several observations are noteworthy. First, the relative gain in power with our method increases as more rare variants are in the region. The rare variant LD structure in both regions is more similar to that of the common variant LD in Region 2 than Region 1. This is significant as it adds further evidence that when there is an absence of strong correlation between causal and null variants our test may be the better choice. Finally, the relative power (rank of performance) within the second two groups of tests does not change by region, but it does with our test. It may be possible that our method could make further power gains once it is better understood how to match a balance of the elastic net features to a genetic architecture.

4.3.3 Combined Common and Rare variants

The penalized regression based test in general showed at least some potential for testing common variants and possible much more potential with rare variants. This prompted us to wonder if our test could leverage both rare and common variants simultaneously. We combined and tested all variants in each region and a summary of the results for the 200 GAW18 data sets is given in Table 4.3.

Test Category	Test Statistic	Region 1		Region 2	
		Power	Median P-value [Q1,Q3]	Power	Median P-value [Q1,Q3]
New Penalized Regression Based Test	$T_{(Ridge)}$	0.350	0.086 [0.025,0.251]	0.285	0.126 [0.043,0.239]
	$T_{(\alpha=0.25)}$	0.440	0.078 [0.016,0.209]	0.320	0.104 [0.034,0.246]
	$T_{(\alpha=0.50)}$	0.455	0.067 [0.017,0.200]	0.295	0.105 [0.037,0.244]
	$T_{(\alpha=0.75)}$	0.470	0.062 [0.017,0.194]	0.315	0.100 [0.038,0.235]
	$T_{(LASSO)}$	0.375	0.104 [0.026,0.259]	0.300	0.095 [0.038,0.230]
Tests Based on LASSO Logistic Regression	$T_{Ave,SSU,Min}$	0.105	0.381 [0.148,0.628]	0.095	0.370 [0.145,0.602]
	$T_{Ave,SSU,Fisher}$	0.045	0.251 [0.140,0.420]	0.045	0.303 [0.176,0.476]
	$T_{Ave,SSU,TPM}$	0.055	0.284 [0.128,0.616]	0.055	0.350 [0.158,0.578]
	$T_{Ave,Score,Min}$	0.155	0.248 [0.090,0.510]	0.135	0.259 [0.104,0.472]
	$T_{Ave,Score,Fisher}$	0.200	0.198 [0.062,0.420]	0.135	0.254 [0.102,0.429]
	$T_{Ave,Score,TPM}$	0.225	0.282 [0.058,0.520]	0.160	0.319 [0.118,0.555]
Tests without Penalized Regression	T_{Score}	0.040	0.294 [0.180,0.478]	0.010	0.319 [0.206,0.450]
	T_{SSU}	0.050	0.305 [0.179,0.447]	0.025	0.349 [0.218,0.482]
	T_{SSUw}	0.200	0.171 [0.071,0.316]	0.105	0.259 [0.124,0.402]
	T_{Sum}	0.015	0.521 [0.307,0.78]	0.030	0.553 [0.300,0.765]
	T_{UminP}	0.020	0.517 [0.280,0.646]	0.015	0.579 [0.302,0.728]

Table 4.3: Summary of Global Test P-values for 200 sets using only *all* variants

The primary result is that the simultaneous testing of common and rare variants is noticeably more powerful with our method than any other method. Quite interestingly, when comparing these results to the rare variant results, common variants increased the power of the rare variants in any meaningful way for only our test and $T_{Ave,SSU,Min}$ and only in Region 1. The power based on our approach was reduced from rare variant only testing in Region 2 as well, but did not diminish on average as much as other methods. This was slightly unexpected as the common variant's MAF was only 7.5%. The LASSO based tests were better when based on the score test, but the weighted SSU was the best

among the tests without logistic regression. For all tests but weighted SSU in these two groups of tests, the power seems nearly capped by the power obtained using only common variants. Considering that there is likely nonzero LD between rare and common variants in practice, our method holds even more promise when testing the full region.

Despite power gains or top performance, the penalized regression holds one more major advantage. Our method can incorporate variant selection, allowing our method to provide variant specific association information. Table 4.4 gives true positive and false positive rates for the 200 simulated sets.

Test	Region 1		Region 2	
Statistic	TP	FP	TP	FP
Ridge	3.180 (1.584)	38.43 (18.978)	2.86 (1.793)	49.315 (30.849)
$\alpha = 0.25$	1.535 (1.407)	16.585 (14.959)	1.490 (1.446)	15.255 (15.692)
$\alpha = 0.50$	1.225 (1.242)	13.015 (13.552)	1.200 (1.364)	11.585 (13.372)
$\alpha = 0.75$	1.185 (1.203)	13.495 (13.733)	1.155 (1.400)	10.375 (12.597)
LASSO	1.04 (1.04)	10.93 (11.731)	0.995 (1.270)	8.57 (11.192)

Table 4.4: Mean (SD) of True Positive (TP) and False Positive (FP) for 200 sets using *all* variants

Here we see that for even strong selection there is possibly actionable information on associated variables. That is, the TP rates capture at least 1 of the 4 variants even though the null variants are more than 90% of each region. This indicates that beyond power information, our method provides insight into variants deserving of possible further investigation.

4.4 Discussion

The results in this article offer support for several conclusions that might contribute to future research. First, our penalized regression based association test, one that uses the full coefficient sets resulting from penalization, may be more powerful when global testing sets of rare variants or regions with both variants. This was true even when compared to other penalized regression approaches that use coefficient estimates only for variable selection. Second, as shown with the MAP4 gene analysis, our new statistic may be able to better leverage variants with non-overlapping disease information. In

fact, there is some evidence that with further improvements our new method may be able to be tailored to different rare variant architectures for further power gains. Third, the penalized regression approach provides meaningful, though not complete, information on associated variants in a group of interest. Other methods do provide this information.

Chapter 5

A New Semiparametric Approach to Finite Mixture of Regressions using Penalized Regression via Fusion

5.1 Introduction

A traditional way to assess the association between candidate variables and an outcome of interest is to generate model estimates at a population level. However, it is often reasonable to hypothesize that for different, unknown subpopulations, an outcome results from different sets of variables (or possibly from different sized effects of the same variables). For example, a disease outcome may be a function of different sets of genetic variants for different groups of individuals within a population. Modeling approaches that don't account for subpopulation induced heterogeneity and the possibility of subpopulation specific effect sizes could easily fail to identify factors associated with a response for only some of the subpopulations.

Statistically, modeling outcomes for a population may in fact require the assumption of a distinct relationship for distinct but unknown subpopulations. One modeling framework useful for this strategy is the finite mixture of regressions (FMR) model. Here, an individual's outcome is predicted from one regression model (known as a component) out of a set of possible regression models. Because the actual component is unknown

for any given observation, a natural choice for fitting FMR models is the Expectation-Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977). Methods based on the EM algorithm yield density estimates and component level regression coefficient estimates depending on the likelihood assumptions used when fitting the model. Wedel and DeSarbo (1995) showed how the algorithm could successfully estimate regression parameters for mixtures of common distributions such as normal or binomial. An EM-like algorithm was developed by Benaglia, Chauveau, and Hunter (2009) to allow for more generality in the error term. Their algorithm was able to lower error rates when compared to current best methods. However, in this algorithm it is unclear what objective function is being maximized and whether successive iterations guarantee an increase in the objective function. A maximum smoothed likelihood algorithm was created by Levine, Hunter, and Chauveau (2011) to remedy the Benaglia shortcomings without decreasing its success in estimating FMR models. The algorithm's advantages, though, did not hold when using the Benaglia, Chauveau, and Hunter (2009) approach to updating bandwidths; the authors remarked how effective bandwidth choice remains a problem with their algorithm. Subsequently, Hunter and Young (2012) developed a semiparametric EM-like algorithm removing the parametric assumptions on the components. The authors showed the method was successful when the initialization was directed towards true values. EM or EM-like algorithms have been successful in FMR problems where it is possible to both dependably specify the mixture distribution including the number of components and initialize the algorithm well.

The EM algorithm has served as the main statistical tool for another category of approaches to subpopulation estimation. For these approaches the focus is on clustering subject-specific regression models. In earlier work, DeSarbo and Cron (1988) used the EM algorithm for clusterwise linear regression. The methodology estimated sets of linear regression parameters assuming normal densities and a given number of clusters of individuals, each cluster with its own regression function. Individuals could then be assigned to a cluster using the estimated posterior probabilities. Interested in the model-based clustering of cyclone tracks or curves, Gaffney and Smyth (2003) used a maximum a posteriori (MAP) EM algorithm for random effects regression mixtures. Specifically, the authors defined a hierarchical model structure with a mixture of parameters at the top level and a simple cyclone-specific regression model at the data level. Again, the

cyclone-specific regression parameters were estimated under the assumption that they were from one of k prespecified subpopulations that follow a normal density. Xu and Hedeker (2001) were also successful using the random-effects mixture model framework with normal density assumptions when classifying subjects in longitudinal clinical trials. Their method revealed if patients receiving either the treatment or placebo comprised distinct subpopulations. As part of the methodology, either one or two components were specified and the respective models were compared. While still dependent on the density and components assumptions, the work of these authors demonstrated the potential for the clustering of subject-specific models.

In settings where the number of subpopulations is unknown or the error distribution cannot be reasonably assumed, alternatives to or enhancements of the EM algorithm must be considered. Penalized regression has shown promise as one such improvement. Specific to the goal of variable selection, some investigators have integrated in a limited way penalized regression into common Markov chain Monte Carlo (MCMC) approaches. An EM algorithm developed by Khalili and Chen (2007) for a penalized mixture model was applied to the FMR setting for the purpose of variable selection, but estimation was based on a parametric likelihood assumption. The method had computational superiority and equal performance to BIC for variable selection in some simulation scenarios. The same authors, Khalili, Chen, and Lin, followed-up in 2011, commenting that despite the many recent advances on the variable selection problem for linear and generalized linear models, methods that gear toward finite-mixture models are still very limited. In their 2011 work an EM approach, again using penalized likelihood for variable selection, was effective in simulations at selecting important covariates, but this was after applying a screening method. To our knowledge the most successful approaches to date for estimating FMR models depend on methodology using some form or approximation of the EM algorithm, and thus depend on making successful likelihood assumptions or successful density estimations. Previous research has also, though, demonstrated potential value in using penalized regression as part of successful FMR estimation. We will go further and exclusively approach the clustered models problem with a penalized regression-based methodology that does not use EM techniques.

In the following work we take a novel approach towards identifying unknown subgroups and their corresponding regression models via grouping pursuit (fusion). Our

approach does not depend on any likelihood assumptions or component density estimations. The key to our methodology is the application of a new type of penalized regression to simultaneous fitting of *separate* regression models for *each* subject. If there exist unknown subpopulations, then the individual fitted models should be the same within the same subpopulation but different across the subpopulations. Specifically, the subjects within a subpopulation share a common model, but the common models differ by subpopulation. Thus, a logical methodological step is the inclusion of a grouping feature to penalize differences in the estimated covariate coefficients *across* individuals. As we will elaborate on shortly, we develop just such a penalty that enables us to force the individuals' models to cluster into a few common models, corresponding to different subpopulations. The methodology can be used as an exploratory data analysis tool akin to hierarchical clustering versus model-based clustering or k-means clustering where the number of clusters is specified.

Penalized regression has been researched to specifically assess its ability to identify and/or leverage groups of variables associated with an outcome. Yuan and Lin (2006) demonstrated that when groups of variables appeared (or disappeared) together in a model, using a group LASSO (Least Absolute Shrinkage and Selection Operator) penalty to select groups of variables or factors (group LASSO) resulted in better performance than the standard LASSO. Another penalized regression approach, the fused LASSO from Tibshirani et al. (2005), added an additional penalty to LASSO specifically for differences in successive regression coefficients. In situations where the features had a natural order, the additional grouping penalty showed promise for both regression and classification. Again using the EM algorithm and a penalized likelihood, Chen and Khalili (2008) applied a smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) to differences in parameter values for adjacent components in a one covariate FMR model. That is, when it could be assumed that the parameter values could be ordered, their penalized likelihood approach (MSCAD) was as effective as existing methods at finding the number of components of the FMR model. A correlation-based penalty applied to all pairs of coefficients used by Tutz and Ulbricht (2009) outperformed an elastic net penalized regression approach that did not use a grouping penalty. Recently, Shen, Huang, and Pan (2012) developed a penalized regression method for simultaneous supervised clustering and feature selection over a given undirected graph that utilized a

truncated-L_1 penalty (TLP) for grouping pursuit. Successful identification and estimation of unknown homogenous groups of effects were possible with their approach. The method used a single linear regression model for a single response, but assumed that the full coefficient vector could be partitioned into subsets of homogeneous coefficients. The new method improved parameter estimation and group identification by penalizing differences within these smaller vectors. In related work, Pan, Shen, and Liu (2013) developed a penalized regression-based clustering (PRclust) method where the TLP penalty was applied to differences in the centroids of data points. PRclust performed well in situations such as non-convex clusters where other more common methods did not. Pivotal to the current work, the success of PRclust demonstrated the potential for comparisons across subjects with a grouping penalty. While using penalized regression to exploit the structure of networks or underlying factors is promising, it still ignores the possibility that any set of variables might only affect a subset of the population. That is, penalized regression to our knowledge has not been used to directly seek or identify subpopulation structures via multiple regression models without explicit use of FMR. The following work incorporates a grouping pursuit framework to shrink differences between subject-specific models for problems similar to FMR. Our approach to penalized regression uses grouping pursuit when simultaneously fitting *separate* models for *each* subject. To be explicit, we penalize only the differences in corresponding parameter estimates between each pair of subject-specific regression models. We study both the LASSO penalty developed by Tibshirani (1996) and the TLP invented by Shen, Pan, and Zhu (2012) in two ways. First, we penalize without using a group feature by applying the penalty to the individual coefficient differences. In a sense we are grouping the subjects for each coefficient separately. This approach shrinks differences in the subjects' models parameter by parameter and does not explicitly shrink differences between the full models. Therefore, we next apply two group penalties based on LASSO and TLP to the differences in the estimated parameter vectors for each pair of samples' regression model. The resulting estimates are compared to the very successful Hunter and Young semiparametric FMR which uses an EM-like approach.

When applied, it is our hypothesis that we will see a hierarchical clustering of individual models depending on the magnitude of the penalty and thresholding parameters. In turn we reveal a partition of the population into subpopulations; although, we do not

focus here on the choice of the number of subpopulations. The following discussion uses simulated FMR models to permit comparison to previous methods and is followed by application to two real data settings. The intent of the following is to show the new penalized regression-based method can handle FMR models and the clustering of subject-level regression models. Because of this we establish its efficacy in the cornerstone case of one covariate problems, a necessary step before building to higher dimensions in subsequent work.

5.2 Methods

In this section we first detail the FMR model. A second section delineates our penalized regression approach and its computation.

5.2.1 Finite Mixture of Regressions Model (FMR)

To motivate and contrast with our new method, we briefly review the Finite Mixture of Regressions (FMR) Model. Using the language of McLachlan and Peel (2000) and notation of Khalili and Chen (2007), suppose Y_i represents the value of a continuous random variable, or response, for subject $i = 1, ..., n$. Let X_{ij} equal subject i's value for covariate $j = 1, ..., p$; therefore, $X_i = (x_{1i}, x_{2i}, ..., x_{pi})$ is the vector of covariates for subject i. Next, let $f(y; \theta_k(x), \phi_k)$ for $k = 1, \ldots, K$ represent K conditional parametric densities of y given x as a function of a canonical parameter, θ_k, and a dispersion parameter, ϕ_k. Utilize the identity link function $g(\mu) = \mu$ such that $\theta = X\beta = \mu$, and (x, Y) follows a FMR model of order K where the conditional density function of Y given x has the form:

$$f(y; x, \Psi) = \sum_{k=1}^{K} \pi_k f(y; \theta_k(x), \phi_k). \tag{5.1}$$

The FMR model has order $K < \infty$ as it is a mixture of K densities (known as component densities). In this equation the unknown parameters are $\Psi = (\beta_1, \beta_2, ..., \beta_K, \phi, \pi)$, where $\beta_k = (\beta_{1k}, \beta_{2k}, ..., \beta_{pk})^T$, $\phi = (\phi_1, \phi_2, ..., \phi_K)^T$, $\pi = (\pi_1, \pi_2, ..., \pi_{K-1})^T$ such that both $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

Parametric approaches by specifying a parametric form of $f(\theta, \phi)$ and estimating

$f(\hat{\theta}, \hat{\phi})$ are most common. As described in the introduction, though, parametric approaches can be too restrictive; therefore, we compare our penalized regression approach to a semiparametric method developed by Hunter and Young. Their method estimates each of the component densities by a nonparametric kernel estimate $\hat{f}(\cdot)$ and provides component level regression coefficients based on a specific K. The Hunter and Young method generates K sets of regression coefficient estimates, partly depending on the specified and estimated likelihoods in an EM-like algorithm. As is described in the next section, our method starts with over-specified n sets of regression coefficients and uses grouping pursuit with group penalties to find a hierarchical clustering of the individual regression models without specifying or estimating a parametric model or likelihood.

5.2.2 A Novel Semiparametric Approach Based on Penalized Regression

Model

We begin by hypothesizing that the parameters of the underlying model for a response can vary by subpopulation. To capture this we estimate a model for each subject in the study using penalized regression with a group feature intended to reveal subpopulations via clustering among these models.

As before suppose Y_i represents the value of a continuous response for subject $i = 1, \ldots, n$. Again, let $X_i = (x_{1i}, \ldots, x_{pi})$ be the vector of p covariates for subject i. Assume for each subject i there is a subject-specific linear model:

$$Y_i | X_i = \beta_{0i} + X_i \beta_i + \epsilon_i \tag{5.2}$$

where $\beta_i = (\beta_{1i}, \ldots, \beta_{pi})^T$ and $\mathrm{E}(\epsilon_i) = 0$. Please note how we initially allow for a sample-dependent (β_{0i}, β_i^T) for each subject, and we at no time specify or estimate a density function for ϵ_i. Our method is semiparametric as we specify the linear form of the relationship, but we do not use $f(\cdot)$ in the FMR model.

Observe from our model how the covariates associated with an outcome would have non-zero values in β_i, but we do not assume the set of non-zero coefficients are identical for all i. For example, a set of covariates might affect the responses of only a subset of the populations (affect only a subpopulation). Even in cases where the same set of covariates

affect multiple subpopulations, the magnitude and/or direction of effect can vary. That is, a set of covariates might affect the outcome of interest for several subpopulations, but affect each differently. In each of these scenarios there is one overarching principle: if multiple subjects' outcomes are affected by the group of covariates in the same functional way, then the (β_{0i}, β_i^T)'s for this subset of the population should be identical. In this way we can partition our population into groups defined by identical (β_{0i}, β_i^T)'s. Our method provides estimates for β_{0i} and β_i by minimizing

$$(1/2) \parallel Y - X\beta \parallel_2^2 + \lambda P(\beta)$$

$$\text{with} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & 1 & X_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & & \mathbf{0} & \cdots & \mathbf{0} & 1 \quad X_n \end{bmatrix} \quad \text{with} \quad \mathbf{0}^T = \left.\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} p+1 \quad , \text{ and } \quad \beta = \begin{bmatrix} \beta_{01} \\ \beta_1 \\ \beta_{02} \\ \beta_2 \\ \vdots \\ \beta_{0n} \\ \beta_n \end{bmatrix}.$$

The penalty parameter λ is applied to a specified penalty, $P(\beta)$. We consider two penalty forms and require $\lambda > 0$ for identifiability: Tibshirani's convex LASSO penalty (1996) and Shen, Pan, and Zhu's non-convex truncated L_1- penalty (TLP) (2012). For our two approaches with respect to the LASSO penalty and grouping pursuit:

1. $P_L(\beta) := LASSO(\beta) := \sum_{i<j} \|\beta_{0i} - \beta_{0j}\|_1 + \sum_{m=1}^{p} \sum_{i<j} \|\beta_{mi} - \beta_{mj}\|_1$
2. $P_{gL}(\beta) := gLASSO(\beta) := \sum_{i<j} \| \binom{\beta_{0i}}{\beta_i} - \binom{\beta_{0j}}{\beta_j} \|_2$

where $\| \cdot \|_1$ is the L_1 norm and $\| \cdot \|_2$ is the L_2 norm. The nongroup version, $P_L(\beta)$, bases selection on the between sample differences in individual coefficient estimates. Depending on the size of λ, the nongroup version chooses the nonzero differences between final estimated sample models by comparing corresponding parameters separately. In contrast, the group version, $P_{gL}(\beta)$, will shrink differences between the full estimated parameter sets and more likely have $(\beta_{0i}, \beta_i^T) = (\beta_{0j}, \beta_j^T)$.

The LASSO penalty will shrink all coefficient differences. However, if there are in fact multiple groups, then group LASSO will encourage shrinkage between and not just within groups. To better maintain between group while reducing within group differences, one strategy is to truncate the penalty for large coefficient differences. Potentially, this

could lessen the between group shrinkage, thus maintaining between group differences for better clustering or subpopulation identification. The TLP does exactly this by implementing a thresholding parameter, $\tau > 0$. For our two approaches with respect to TLP:

1. $P_{TLP}(\beta) := TLP(\beta) := \sum_{i<j} \min(\|\beta_{0i} - \beta_{0j}\|_1/\tau, 1)$
   $+ \sum_{m=1}^{p} \sum_{i<j} \min(\|\beta_{mi} - \beta_{mj}\|_1/\tau, 1)$
2. $P_{gTLP}(\beta) := gTLP(\beta) := \sum_{i<j} \min(\|\binom{\beta_{0i}}{\beta_i} - \binom{\beta_{0j}}{\beta_j}\|_2/\tau, 1)$

In comparing the LASSO and TLP versions, there is no further penalty for differences greater than τ for the TLP version, but there is with LASSO. Overall, LASSO yields biased parameter estimates that the, per Shen, Pan, and Zhu (2012), TLP penalty corrects through adaptive shrinkage that combines shrinkage and thresholding.

Computation

Given λ and τ (TLP only), estimates using the nongroup penalties P_L and P_{TLP} were obtained from slight modifications of the `gflasso` and `ncTLF` functions in FGSG: Feature Grouping and Selection Over an Undirected Graph in Matlab engineered by Yang et al. (2012).

We develop an alternating direction method of multipliers (ADMM) to fit the models when using group penalties. The ADMM form introduces another variable, Z, reflecting how the objective function can be separated and subsequently solved in parallel. In ADMM the problem with respect to group LASSO (gLASSO) is stated as:

$$\begin{aligned} &\text{minimize} \;\; f(\beta) = (1/2)\|Y - X\beta\|_2^2 + \lambda P_{gL}(Z) \\ &\text{subject to} \;\; F\beta - Z = 0 \end{aligned}$$

where F is the linear transformation matrix comparing vectors of coefficients for all pairs of samples ($1 \leq i < j \leq n$). That is, $F = \left[F_{1,2}^T, \quad F_{1,3}^T, \quad \cdots \quad , F_{n-1,n}^T\right]^T$ where each $F_{i,j}$ is a $(p+1) \times n(p+1)$ matrix

$$F_{i,j} = \begin{array}{c} \begin{matrix} & & & & & & \overset{\text{(i(p+1)-1)}^{th}}{\underset{\text{column}}{\downarrow}} & & & \overset{\text{(j(p+1)-1)}^{th}}{\underset{\text{column}}{\downarrow}} & & \end{matrix} \\ \begin{bmatrix} \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & \cdots \\ \cdots & \cdots & & & & & & & & \\ \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots \\ \cdots & \cdots & & & & & & & & \\ \cdots & \cdots & \cdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \cdots & \cdots & & & & & & & & \\ \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & \cdots & & & & & & & & \end{bmatrix} \end{array}$$

The corresponding gLASSO objective function, derived as in the method of multipliers from an augmented Lagrangian with u the scaled dual variable, is

$$L_\rho(\beta, z, u) = (1/2)\|Y - X\beta\|_2^2 + \lambda P_{gL}(Z) + (\rho/2)\|F\beta - z + u\|_2^2.$$

Boyd et al. (2011) showed the ADMM algorithm then iterates three steps until converging to coefficient estimates:

1. $\beta^{(h+1)} = (X^T X + \rho F^T F)^{-1}\big(X^T Y + \rho F^T (z^{(h)} - u^{(h)})\big)$
2. $z^{(h+1)} = \begin{bmatrix} S_{\lambda/\rho}\big(F_{1,2}\beta^{(h+1)} + u_{1,2}^{(h)}\big) \\ \vdots \\ S_{\lambda/\rho}\big(F_{n-1,n}\beta^{(h+1)} + u_{n-1,n}^{(h)}\big) \end{bmatrix}$
3. $u^{(h+1)} = u^{(h)} + F\beta^{(h+1)} - z^{(h+1)}$.

In the above, the notation "(h)" is for the h^{th} iteration. S is the vector *soft thresholding operator*: $S_\kappa(a) = \big(1 - \kappa/\|a\|_2\big)_+ a$, and a_+ is equal to the positive part of a. Remark how the $S_\kappa(a)$ can shrink a whole vector to 0 if the coefficient vectors being compared are the same, which is in contrast to the individual soft thresholding used in $LASSO(\beta)$. Finally, u is partitioned corresponding to the pairwise differences in coefficient vectors; thus, $u_{i,j}$ represents the subvector of u corresponding to the comparison made with $F_{i,j}$. For our estimation we set $\rho = 1$.

The group TLP (gTLP) penalty is not convex, an important distinction from gLASSO; therefore, we use a difference convex method to facilitate computation. First, define the objective function:

$$S(\beta) = (1/2)\left\|Y - X\beta\right\|_2^2 + \lambda \sum_{i<j} \min\left(\left\|\begin{pmatrix}\beta_{0i} \\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j} \\ \beta_j\end{pmatrix}\right\|_2 / \tau, 1\right).$$

Similar to Shen, Huang, and Pan (2012), $S(\beta)$ can be written as a difference of two convex functions $S_1(\beta) - S_2(\beta)$ with

$$S_1(\beta) = (1/2)\Big\|Y - X\beta\Big\|_2^2 + (\lambda/\tau)\sum_{i<j}\Big\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\Big\|_2$$
$$S_2(\beta) = (\lambda/\tau)\sum_{i<j}\Big(\Big\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\Big\|_2 - \tau\Big)_+.$$

As demonstrated by those authors, a sequence of upper approximations can be constructed iteratively by replacing $S_2(\beta)$ at the iteration $h+1$ by its piecewise affine minimization

$$S_2(\beta)^{(h)} = S_2(\hat{\beta}^{(h)}) + (\lambda/\tau)\sum_{i<j} I\Big(\Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2 \geq \tau\Big)\ \mathrm{x}$$
$$\Big(\Big\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\Big\|_2 - \Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta_{0j}}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2\Big)$$

at iteration h, yielding an upper convex approximation for $S(\beta)$ at iteration $h+1$:

$$S^{(h+1)}(\beta) = (1/2)\Big\|Y - X\beta\Big\|_2^2 + (\lambda/\tau)\sum_{i<j}\Big(\Big\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\Big\|_2\Big)\times$$
$$I\Big(\Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2 < \tau\Big).$$

Because of this we can use ADMM for gTLP by replacing step two of the gLASSO algorithm with

$$z^{(h+1)} = \begin{bmatrix} S_{\lambda_h/\rho}\big(F_{1,2}\beta^{(h+1)} + u_{1,2}^{(h)}\big) \\ \vdots \\ S_{\lambda_h/\rho}\big(F_{n-1,n}\beta^{(h+1)} + u_{n-1,n}^{(h)}\big)\end{bmatrix}$$

where we calculate for each comparison $i < j$

$$\lambda_{h/\rho} = \lambda(\rho\tau)^{-1}\,\mathrm{I}\Big(\Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2 < \tau\Big).$$

Our method is distinct from the competing FMR estimation methods, which are intended to find estimates at the component level. In particular, it is semiparametric.

Therefore, for comparison we present results from application of a semiparametric FMR methodology of Hunter and Young (2012). The Hunter and Young method estimates β_{0k} and β_k for $k = 1, ..., K$ (refer to equation (5.1)); that is, an estimate of β_0 and β for each *component* k. The semiparametric models were fitted with the default settings of the `spregmix` function in the `R` package `mixtools` from Benaglia et al. (2009).

For both penalty types models were fit with a large decreasing sequence of λ in order to show a wide range of degree of selection. When fitting models for a data set we started with the largest value of penalty. The resulting parameter estimates were used to initialize the subsequent model's estimation for the same data set (the model fit using the next smallest candidate in the sequence). We repeated this process until the fitting of the model with the smallest λ was initialized with the estimates found with the second smallest λ. For the TLP models we considered a range of small to large candidates for the tuning parameter τ's in order to show results from situations where nearly all differences exceeded the threshold to situations with performance similar to LASSO.

The threshold and penalty parameters used for the presented results were determined with generalized cross-validation (GCV). Golub, Heath, and Wahba (1979) showed GCV's viability in selecting the parameter in ridge regression, and Pan, Shen, and Liu (2013) used GCV successfully to choose the threshold parameter when applying their TLP based PRclust clustering algorithm. When calculating the GCV in our setting, first allow $\hat{\mu}_i = \hat{\beta}_{0i} + X_i\hat{\beta}_i$. Following Golub, Heath, and Wahba (1979) generalized cross-validation can be defined as

$$GCV(df) = \frac{RSS}{(n - df)^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2}{(n - df)^2}.$$

Here, the notation shows how the GCV statistic is a function of df, equal to the degrees of freedom used when generating the μ_i. Pan, Shen, and Liu (2013) found estimates could be improved by using generalized degrees of freedom (GDF) instead of the usual $df = p$. Ye (1998) provided the calculation for GDF, which in our problem is

$$GDF = \sum_{i=1}^{n} \lim_{\delta \to 0} E_\mu \Big[\frac{\hat{\mu}_i(Y_i + \delta e_i) - \hat{\mu}_i(Y_i)}{\delta} \Big]$$

where e_i is the i^{th} column of the $n \times n$ identity matrix. Correspondingly, Ye (1998)

provided the following Monte Carlo algorithm to estimate GDF (adapted to our setting) when applying one of our four penalties:

1. Repeat steps 2 and 3 for $b = 1, \ldots, B$. In the following we set $B = 100$
2. Generate $\Delta_b = (\delta_{b,1}, \ldots, \delta_{b,n})$ with $\delta_{b,i}$ iid $\mathcal{N}(0, \nu)$. For our problems $\nu \approx .5\sigma_Y$
3. Compute $\hat{\mu}(Y + \Delta_b)$ with the penalty-specific algorithm using data $Y + \Delta_b$
4. Calculate $\hat{h}_i$ as the regression slope from $\hat{\mu}_i(Y + \Delta_b) = \alpha + \hat{h}_i \delta_{b,i}$ for $b = 1, \ldots, B$
5. Use $GDF = \sum_{i=1}^{n} \hat{h}_i$ when calculating GCV for the $\hat{\beta}$ found with a specified λ and τ (TLP only).

The parameter values for the following results are those with the smallest $GCV(GDF)$ statistic among the candidates considered.

5.3 Simulations

We explored two related settings using a single continuous response generated from a standard linear regression model with one continuous covariate ($p = 1$) and an intercept for $n = 200$ subjects. The responses were generated from a FMR model with $K = 2$ components; that is, the responses were generated using different regression models for $k = 1$ and $k = 2$.

5.3.1 Simulation Design

The component for sample i was simulated from a Bernoulli distribution with mean equal to 0.5; that is, equal probability of either component generating the true response. Resulting from the use of the Bernoulli distribution to randomly assign group, 100 subjects' responses were created with each component. The simulated response was generated as

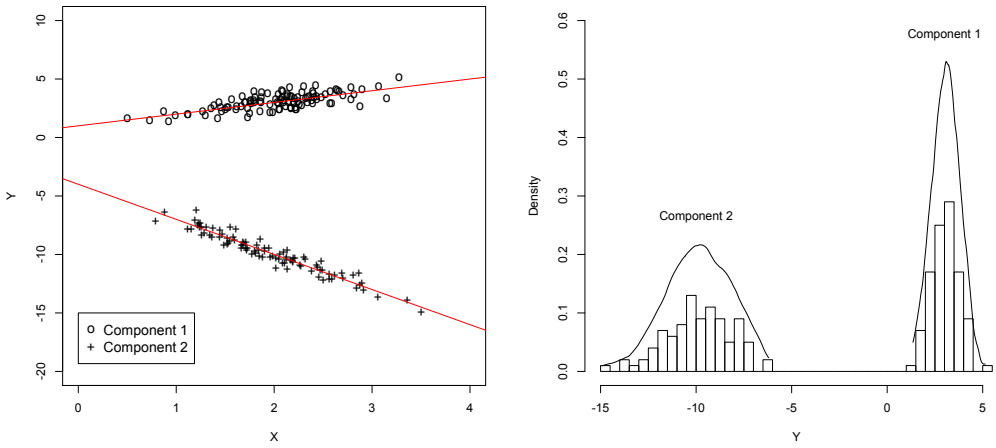
$$Y_i | X_i, k = \beta_{0k} + X_i \beta_{1k} + \epsilon_i, \tag{5.3}$$

where $k \in \{1, 2\}$ indicates the component generating Y_i and $(\beta_{0k}, \beta_{1k})^T$ are the intercept and regression coefficient for the k^{th} regression component.

The first stage of the simulation is the generation of the covariate value. Let X_i represent a continuous covariate. Specifically, X_i is generated from a normal distribution with mean 2 and standard deviation 0.5. Next, we let $\epsilon_i \sim \mathcal{N}(0, 0.5)$. In the following two simulations we consider two different $(\beta_{01}, \beta_{11})^T$ and $(\beta_{02}, \beta_{12})^T$ combinations and generate Y_i subsequently from the respective regression components using equation (5.3).

5.3.2 Simulation Results

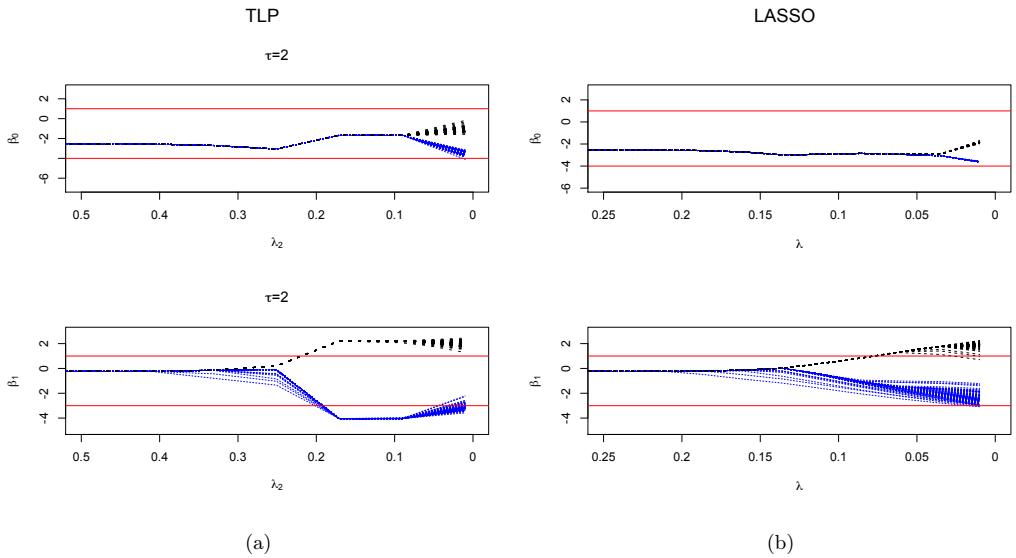
The first simulation evaluates a scenario with strong separation between responses generated with different components. Set $\beta_{01} = 1$ and $\beta_{11} = 1$ for component one and $\beta_{02} = -4$ and $\beta_{12} = -3$ for component two. The (X_i, Y_i) pairs are plotted in Figure 5.1(a). Subjects from the first component are plotted with circles and subjects from the second component are plotted with pluses. Additionally, the true regression lines for the two components are plotted with solid lines. The resulting Y_i are plotted in Figure 5.1(b) and include component specific Epanechnikov kernel density estimates to provide a sense of their distribution. For completeness, we remark that the Epanechnikov kernel was chosen because it minimizes the asymptotic mean integrated squared error, a well-established performance measure for kernel density estimation. The curve was generated using the `density` function with its default bandwidth provided in the `R` base `stats` package version 2.15.1.



(a) (b)

Figure 5.1: (a) Y_i and X_i scatterplot with true regression lines (b) Y_i distribution

The first set of results examine the performance of penalized regression with the nongroup penalties: LASSO and TLP. In Figures 5.2(a) and (b) the individual λ regularization paths for each subject i are plotted for β_{0i} (top row) and β_{1i} (bottom row). In our usage, a regularization path is the curve connecting the estimates obtained for person i when using each value of λ in sequential order (λ value given on the horizontal axis). From left to right the value of the penalty parameter is decreasing to allow any natural hierarchical structure to be exhibited. For TLP the plot is based on $\tau = 2$, the value with the lowest combined GCV statistics across the candidate penalty parameters. Specifically, (a) provides the TLP version with penalized pairs of individual coefficients. The respective LASSO results are given in (b). True values are given as horizontal lines, and the regularization paths for subjects from the first component are darker than those from the second.



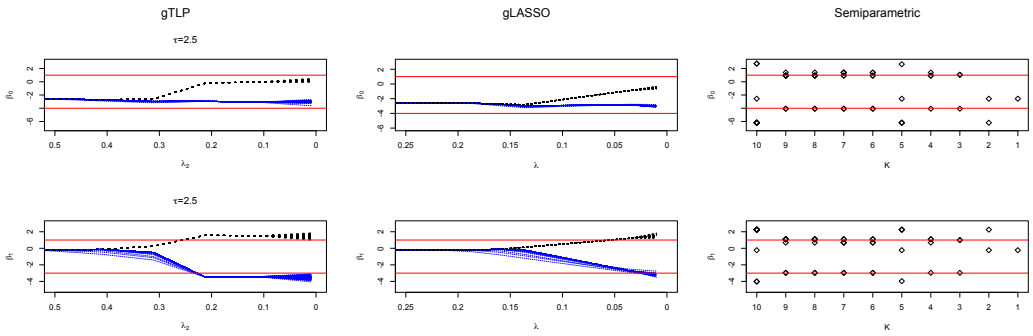
(a) (b)

Figure 5.2: β_0 (row 1) and β_1 (row 2) estimates using (a) TLP and (b) LASSO

Subjects from the two components can be distinguished for both TLP and LASSO for small enough λ. Not unexpectedly, the divergence in parameter estimates for subjects in the same component generally increases with both the TLP and LASSO methods as the penalty decreases. This becomes significant because the λ at which the groups separate is different for the β_{0i}'s and the β_{1i}'s. TLP does outperform LASSO in terms of providing closer estimates of the true β_i as λ decreases, but there is still no λ range for either method at which both components' β_0 or β_1 estimates are simultaneously within even one unit for all n subjects (using a course metric for illustrative purposes). These two deficiencies prompted an investigation of the effect of a group penalty applied to the distance between the samples' coefficient vectors.

Figure 5.3(a) reveals the success of our group TLP (gTLP) method at overcoming these issues. The individual λ regularization path for each sample i are plotted for $\tau = 2.5$ (lowest total λ path GCV). As before the hierarchical structure can be seen in both the β_{0i} and β_{1i} plots, where the two distinct groups become more apparent as the penalty is reduced. The estimates themselves show increased β_0 and β_1 accuracy for both components simultaneously unlike in the TLP or LASSO versions (closer to the true values for small λ). gTLP definitely exhibits this property more than that of the

group LASSO plots in Figure 5.3(b). We see the gLASSO is effective at identifying two distinct components, but shows less accuracy (distance between the true and estimated values) than the gTLP approach in at least one parameter. Comparing the group and nongroup approaches, the largest penalty parameter value which induces separation between components is the same for both the slope and coefficient. For both types of penalties the group versions improve parameter estimates relative to their nongroup counterparts.



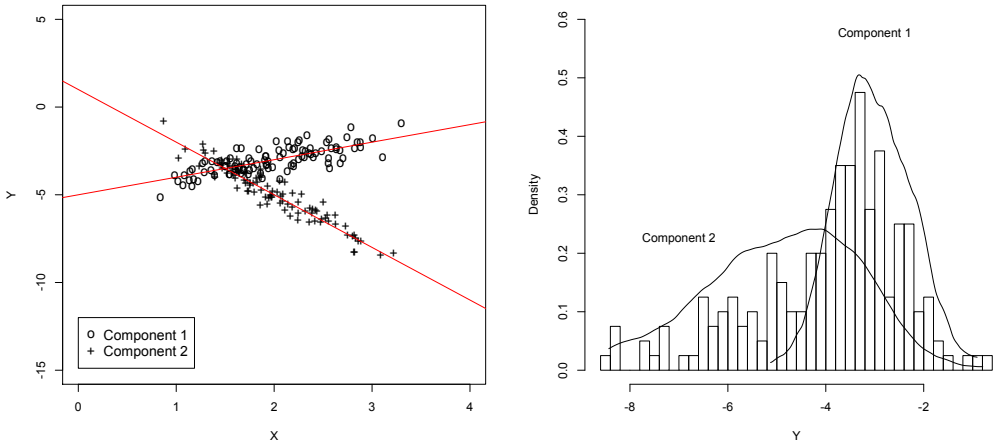
(a) (b) (c)

Figure 5.3: β_0 (row 1) and β_1 (row 2) estimates using (a) gTLP, (b) gLASSO, and (c) SP

Semiparametric (abbreviated SP) FMR models were fit with $K = 1, ..., 10$ specified components (x-axis), and the parameter estimates are plotted in the third panel of the figure. Figure 5.3(c) reports β_{0k} (top row) and β_{1k} (bottom row) for $k = 1, ..., K$ components. Please note the descending order of the axis. The figures reveal that for $K = 2$, semiparametric estimation is overall not successful, seeming to provide estimates centered around one of the two true component parameter values for both β_0 and β_1. Interestingly, when specifying $K = 3, 4, 6, 7, 8$, or 9 the SP method finds essentially two groups with good parameter estimation (note that multiple component estimates overlap significantly in the plot). However, when K is specified as two components, the method essentially centers the two β_0 estimates around component two's value of -4 and centers the two β_1 estimates around component ones's value of 1. That is, one component drives the estimates for the intercept and the other component drives the estimation for the

regression coefficient, something that it appears the group feature prevents.

The first simulation for outcomes that are distinct with respect to component yielded evidence that gTLP can outperform the other methods and modeling approaches. Our second simulation considered the complementary context, that of overlapping responses for both components. With $\beta_{01} = -5$ and $\beta_{11} = 1$ for component one and $\beta_{02} = 1$ and $\beta_{12} = -3$ for component two, the regression lines intersect within our X range. The (X_i, Y_i) pairs and the Y_i distribution are presented in Figures 5.4(a) and (b). Unlike scenario one there is considerable overlap in responses generated from the different components (compare Figure 5.4b to Figure 5.1b).

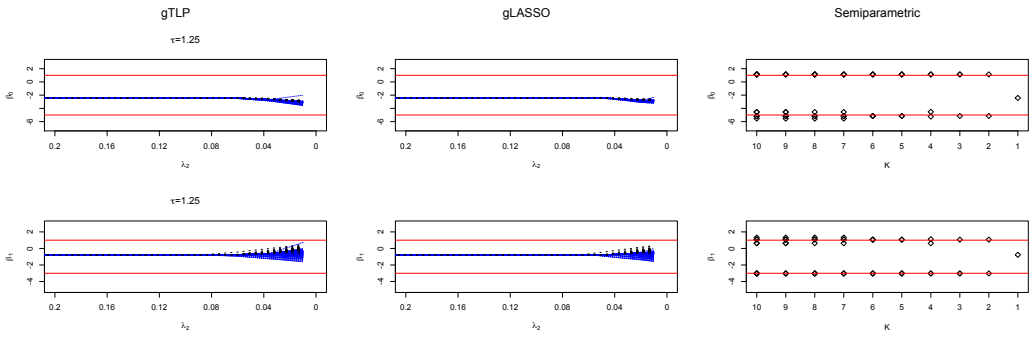


(a) (b)

Figure 5.4: (a) Y_i and X_i scatterplot with true regression lines (b) Y_i distribution

Here, the semiparametric method excels and the two penalized regression based methods are unsuccessful as can be seen in Figure 5.5. The gTLP (a) and gLASSO (b) methods are not able to distinguish any subpopulations; thus, they essentially provide estimates centered around a population mean (the $K = 1$ value in the SP method in (c)). Even with small λ the penalized regression approaches do not provide solid parameter estimates for any samples for any reasonable distance metric. Figure 5.5(c) show how the SP method provides estimates with little bias for the true number of components

$K = 2$. TLP and LASSO penalized regression results were less successful and are not presented.



(a) (b) (c)

Figure 5.5: β_0 (row 1) and β_1 (row 2) estimates using (a) gTLP, (b) gLASSO, and (c) SP

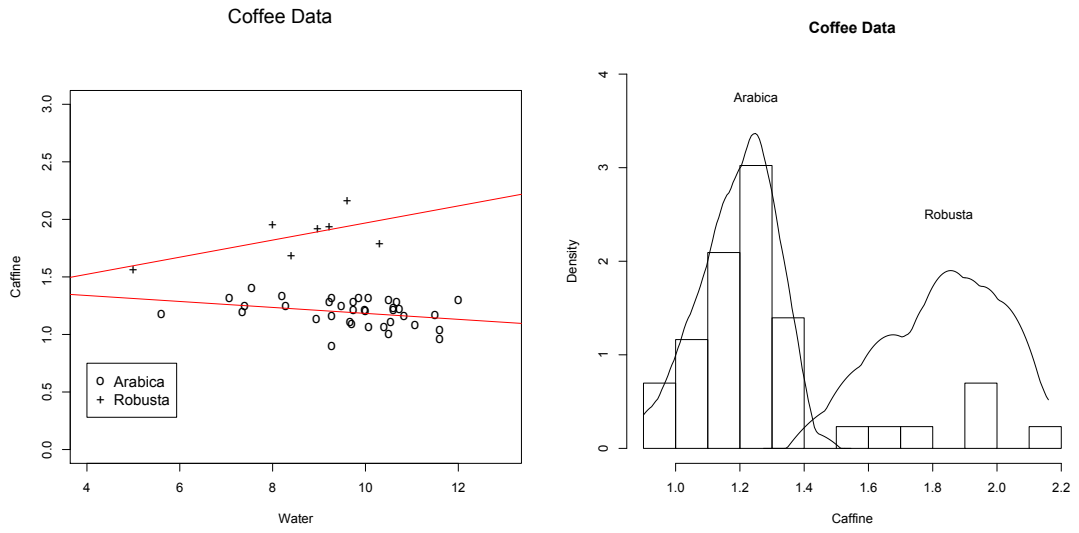
An early possible insight from the two simulations is that gTLP and to a lesser degree gLASSO are best when responses are generally distinguishable between components. The additional thresholding parameter when using TLP may be advantageous when the distance between component coefficient vectors is dominated by one parameter. Similarly, it may be valuable to truncate penalization in order to reduce the effect of penalizing samples that are truly in different subpopulations. However, the SP method was the only method that could distinguish components in the scenario where the Y_i's overlapped for the two components.

5.4 Examples

The final data section shows examples from two nonsimulated data sets. In the first the number of components is known, allowing the method to be tested on a real data set. The second data set exploration represents an extension of our methodology to the investigation of how a third factor may influence the relationship between two other factors (and possibly allow for population partitioning based on this influence).

5.4.1 Coffee Data

In 1973 Streuli presented coffee data collected from 43 samples of one of two varieties: Arabica or Robusta. The data is available in the `pgmm` R package built by McNicholas et al. (2011) and consists of measurements for chemical properties such as "water" and "caffine" content, the variables used in our analysis. The seven Robusta samples are plotted with pluses on the top of plot Figure 5.6(a), and their fitted regression line with a positive effect for water is provided. The caffine and water relationship for the remaining thirty-six Arabica samples are represented by circles surrounding their decreasing fitted regression line.



(a) (b)

Figure 5.6: (a) Caffine and Water scatterplot with fitted regression lines (b) Caffine distribution

Compare the coffee plots in Figure 5.6 to those of our first simulation in Figure 5.1 and note the (slightly less) distinct groupings of outcomes by variety of coffee. In the simulation setting the gTLP method was superior to the gLASSO, which was in turn superior to the semiparametric method. In particular, the semiparametric method produced parameter estimates of the regression coefficients further from the true values.

The results are even more extreme for the real data example. The coefficient estimates for all three methods are presented in Figure 5.7, starting with SP for comparison. Because the true variety is known, lighter colored lines are used for the seven Robusta samples.

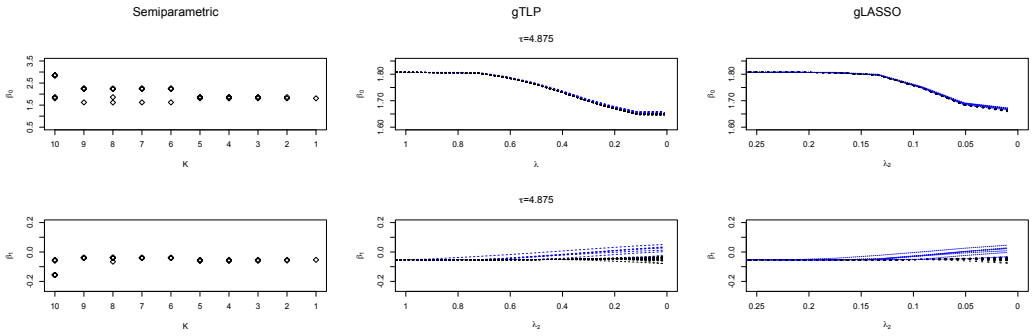
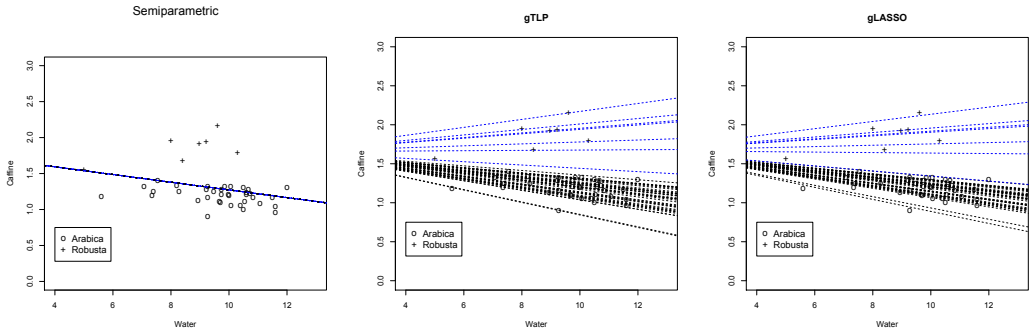


Figure 5.7: β_0 (row 1) and β_1 (row 2) estimates using (a) SP (b) gTLP and (c) gLASSO

The semiparametric approach does not distinguish between the two varieties. In fact, for $K \leq 5$ the method sees the 43 samples as essentially one population. This result held for 10 different random starting values. Both the gTLP and gLASSO approaches show evidence of two groups, particularly in the β_1 parameters (second row), as the penalty is decreased. There is only small separation in the intercept parameters especially across varieties (more in gTLP than gLASSO), but this is not unexpected after examining fitted lines in Figure 5.6(a). Perhaps the biggest difference between the two varieties when examining Figure 5.6(a) is the likely direction of the β_{1i}'s. Importantly, the gTLP method estimates more positive β_{1i}'s for the Robusta samples than gLASSO as λ decreases. The effect is noticeably better fits at a sample level (Figure 5.8(b) versus (c)) when applying the λ associated with the smallest individual GCV value.



(a) (b) (c)

Figure 5.8: Scatterplot coffee sample level regression estimates using (a) SP (b) gTLP and (c) gLASSO

From Figure 5.8(a) it is clear that the SP does not find evidence of multiple subpopulations. In this plot the posterior weighted regression lines ($K = 2$) for each sample are plotted. In contrast both penalized regression approaches do see two varieties, but the gTLP outperforms gLASSO in capturing the key difference in varieties: likely regression coefficient direction. The results provide evidence that the success of penalized regression with a group feature (gTLP in particular) demonstrated by simulation can also hold in real data settings.

5.4.2 Saccharomyces Cerevisiae Cell-cycle Data

The first simulation and real coffee data support the promise of our gTLP methodology. Natural questions emerge in our setting that we will start to address in this section. First, what factors lead to the underlying subgroups? In particular, how do the various methodologies perform when the relationship between our independent and outcome variables is dependent on the value of a third variable? Finally, what if this third variable is continuous? That is, what if the subpopulation definition is really a continuum and not a strict ordinal classification?

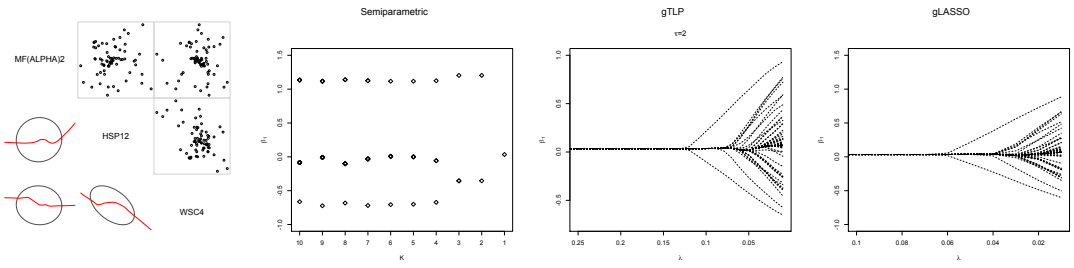
In a genetics context this situation was considered by Li (2002) in the development of the liquid association (LA) statistic. LA is the quantification of the dependency of the coexpression (as correlation) of two genes on a third gene. Ho et al. (2011) extended

the statistic to variable triplets with more complex codependencies. Specifically, Ho and colleagues created the modified liquid association (MLA) statistic and demonstrated its value on the Saccharomyces cerevisiae cell-cycle data set from Spellman et al. (1998). Ho et al. (2011) defined modified liquid association (MLA) as the expected value of change in X_1 and X_2's conditional correlation with a standard normal X_3. Specifically,

$$MLA(X_1, X_2|X_3) = E\{h^{'}(X_3)\} = E\{h(X_3)X_3\} \text{ where } h(X_3) = \rho(X_1, X_2|X_3).$$

This data is available in the R package `LiquidAssociation` from Ho (2009). The RNA abundance measures of eleven gene triplets in the Saccharomyces cerevisiae met the criterion used by Ho et al. (2011) for large MLA, indicating strong evidence that the correlation of two of the genes depends on a third. Using this data we can begin studying our new methods in a setting where Ho et al. documented evidence of a linear relationship between two continuous variables impacted by a known third continuous variable. The results begin to provide some insight into the questions posed above.

We studied the 69 samples with RNA values for all variables in the triplet consisting of the MF(ALPHA)2, HSP12, and WSC4 genes. Figure 5.9(a) shows the pairwise scatterplots among the three genes. The lower diagonal panels provide pairwise 68% concentration ellipses with a loess smoothed curve.



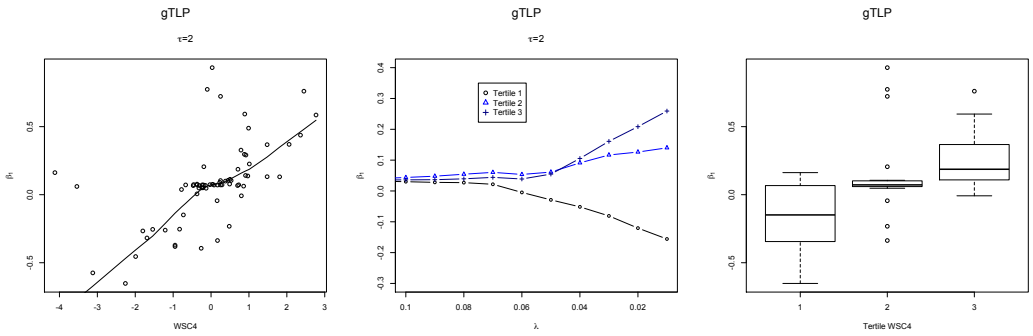
(a) (b) (c) (d)

Figure 5.9: (a) MF(ALPHA)2 and HSP12 scatterplot (b) SP (c) gTLP (d) gLASSO

Ho et al. (2011) found good evidence that the relationship between the MF(ALPHA)2 and HSP12 genes was dependent on the WSC4 gene. Thus, we regressed the RNA abundance of MF(ALPHA)2 onto the corresponding abundances of HSP12 by the SP,

gTLP,and gLASSO methods. We then explored basic ways to capture the dependance on the WSC4 gene by comparing our subject-specific regression slope estimates and their respective WSC4 abundances.

Figure 5.9(b), (c), and (d) focus on the regression coefficient of HSP12 which we label β_1. The SP results in plot (b) seem to indicate two or three groups. However, the pseudo likelihoods provided by the `spregmix` function favor 5 or 9 groups and not 2 or 3, preventing strong conclusions about a subpopulation structure from being drawn. By design the penalized regression approaches permit divergence in the sample level $\hat{\beta}_{1i}$ (if it exists). In this sense the fan patterns shown in Figures 5.9(c) and (d) are not unexpected. Moreover, the thresholding parameter of the gTLP does seem to allow some loose clustering for sufficiently small λ that is not apparent with gLASSO. Comparing the gTLP to the SP, we see the penalized regression approach can provide better insight into the degree of clustering. The example here illuminates that gTLP is more flexible than SP in that it will not force aggregation that may not exist. This property could prove valuable when the relationship between two variables is fluid and not easily categorized by clustering methods but still can be described meaningfully in the context of a third factor. For example, if the third gene (WSC4) truly affects the relationship of MF(ALPHA)2 and HSP12, then consideration of each sample's WSC4 abundance should yield insight into the pattern shown in the gTLP results. Consequently, there should be some predictable structure in the fan pattern dependent on the samples' corresponding WSC4_i RNA abundances. The final set of Figures (5.10) begins to describe this dependence by revealing a simple but distinct pattern when the samples are classified by tertile of WSC4.



(a) (b) (c)

Figure 5.10: (a) WSC4$_i$ and β_{1i} scatterplot with Loess curve (b) mean gTLP by tertile (c) distribution of β_1's by tertile for fixed $\lambda = 0.01$ and $\tau = 2$

In Figure 5.10 (a) a Loess curve is fit to a scatterplot of the WSC4$_i$ RNA abundances and the β_{1i}'s from the subject-specific regression models estimated with $\lambda = 0.01$ and $\tau = 2$. The curve shows a distinct positive linear trend for the smaller and larger WSC4 abundances and a flattened area for the average values. Therefore, we divided the WSC4 abundances into tertiles as a form of partioning the population. The mean coefficient estimate by tertile for each value of λ is plotted Figure 5.10 (b). The first tertile is plotted with circles, the second with triangles, and the largest third of WSC4 abundances with plus signs. We see a modest pattern based on the tertiles. The coefficients within higher WSC4 tertiles are in general larger as λ decreases. Even with the basic exploratory analysis used here, the gTLP can provide some insight into how association between two genes depends on a third genetic covariate. This can be seen better in panel (c) where the $\hat{\beta}_{1i}$ distributions by tertile are plotted for the (λ,τ) value with the lowest GCV value.

The MLA value was 0.522 for our triplet as reported in Ho et al. (2011), a value which indicated a significant dependency on WSC4 of the correlation between MF(ALPHA)2 and HSP12. Our method examines an association between the third variable, WSC4, and the regression coefficients for predicting each sample's MF(ALPHA)2 value from its HSP12 measure; specifically, the samples' coefficients for MF(ALPHA)2 regressed onto HSP12. Using even a simple metric like correlation we find a strong association. In this example the correlation between the $\hat{\beta}_{i1}$ and the WSC4 abundances for the 69 samples

was 0.55.

The relatively large correlation value shows the potential for using comparisons between the gTLP coefficient estimates and the actual values of a third gene as a complement to MLA for capturing the dependency of the relationship between two factors on a third. The MLA calculation requires specification of the third variable, but the gTLP approach does not when estimating regression models. Therefore, the gTLP derived estimates can be compared to multiple candidate variables or sets of candidate variables. The novel gTLP holds promise as an complement to MLA by providing sample-level regression estimates that can be integrated into potentially quick exploratory analyses.

5.5 Discussion

The article has provided evidence using real data supported by simulation that our new grouping pursuit gTLP method, and to a lesser extent a grouping pursuit gLASSO, handles certain types of problems for which previous methods such as Hunter and Young's semiparametric approach were not successful. Our novel gTLP approach was successful in scenarios using FMR when responses generated by different component regression models were distinguishable. The gTLP method, which applies group penalization to differences between coefficient vectors, was able to correctly classify subpopulations and provide good subject level estimates of regression coefficients. While warranting further investigation, the truncation threshold parameter (τ) used by the gTLP may improve on gLASSO methods by weighting the penalty more towards within component differences. If the responses from different component regression models are well separated, the gTLP may be better than gLASSO at maintaining between component/subpopulation separation in the coefficients while reducing within component differences. In addition, this work confirms that group penalties, such as gTLP and gLASSO, can improve component identification and regression model estimation over their corresponding coefficient specific penalties, TLP and LASSO.

Importantly, our new method focuses on the estimation (and then clustering) of individual regression models. This holds great promise for application to personalized medicine. In the present work we have only begun to show how a different grouping approach to penalized regression may be able to overcome some of the limitations of

current approaches. The simulations were basic and do not cover a large range of possible combinations of component models, but they do provide support for gTLP's value in the essential setting (single variable) needed for analysis of more complicated scenarios. Future work will need to apply the method to more scenarios to further define the class of problems for which gTLP shows strong promise. A particular problem of interest occurs when a variant has a true effect for only one of several subsets of the population. Also, future work must include scenarios with more covariates and variable selection features in addition to grouping features. The work to date employed the squared loss function only, but the method could be modified to accommodate different loss functions that might better serve a problem. For example, it could be interesting to look at an L_1 function in data with outliers. The authors thought it advantageous to show how the penalty magnitude could uncover a hierarchal structure; thereby, showing the potential for different partitions of the population as for the MLA example. Finally, in our examples the GCV was successful at choosing a single set of coefficient estimates among those generated by different threshold and penalty values, but it will be beneficial to revisit this issue and potentially develop a better criterion for selecting optimal tuning parameters and the number of components (if indeed they exist). Our main goal here is to demonstrate the feasibility and promise of our proposed penalized regression approach as a proof of concept.

Chapter 6

Conclusion and Discussion

The work presented above provides evidence that penalized regression can be valuable in translating genetic information. Analysis using simulated and real data show its potential to advance risk prediction, even in the presence of environmental covariates. A new statistic has potential as a useful tool for assessing the relationship between the genetic information in a genome region and a disease outcome. Finally, the initial success of a new semiparametric group penalized regression approach reveals its potential as a clustering methodolgy that can partition populations into risk groups.

The individual chapters give context-specific insight on the strengths and limitations of the three approaches. Taken collectively there is strong evidence to advance the Truncated L_1-Penalty (TLP) as a candidate on par with the LASSO or ridge regression. First, it outperformed other penalties in some of the prediction scenarios in the second and third chapters. Second, the group version (gTLP) was demonstrably superior to group LASSO in the FMR setting. While its additional thresholding parameter can result in improved function, it can also complicate computation and cross-validation (relating to parameter choice). Future work will need to address these issues beyond what is shown in this dissertation.

A broad theme in the second through fourth chapters is penalty choice. More specifically, how the choice of penalty influenced the strength of conclusions with respect to prediction, classification, power analysis, and true associated variant identification. This was in no way surprising, but the results reinforce the value in gaining more understanding of why the different penalty features have such an impact in different genetic settings.

If possible, future work will need to provide more insight into how to choose a penalty to best leverage a genetic architecture.

The work that I am most excited about going forward was established at the proof-of-concept level in the fifth chapter. The gTLP clustering methodology can be taken in many directions. In fact, some of this has already begun. The technical strength of the method can be improved by better addressing the issue of identifiability. A natural extension needed to impact genetics questions is to increase dimensionality, eventually applying to the $p >> n$ setting. One possible enhancement is to introduce additional penalties; for example, address the identifiability with an additional ridge penalty and address high dimensionality with an added variable selection penalty. A major goal is to create a general hierarchical clustering tool; thus, it will be necessary to find criteria for cases where more than two subgroups can be found, even if the boundaries are fluid. The examples provided in Chapter 5 begin to show the potential of the tool, but they also reveal concerns that must be addressed.

This dissertation work in its entirety can be describe as solving initial problems on the path to personalized medicine. The collective work shows how penalized regression can be used in more and more settings, and the reported results substantiate the potential of our new penalized-regression based statistical tools. In each of the three settings, though, limitations were found that must be addressed before the new methodologies can make more meaningful contributions to the field. As this dissertation is the cornerstone of my career, future work will focus (at least initially) on adapting and evolving the methodologies in order to solve increasingly more challenging problems.

References

About Crohn's Disease (2009). Retrieved from http://www.ccfa.org/info/about/crohns on 8/21/2011.

Alexander DH, Lange K (2011). Stability Selection for Genome-Wide Association. *Genetic Epidemiology* 35: 722-728.

Basu S, Pan W, Shen X, Oetting W (2011). Multilocus Association Testing With Penalized Regression. *Genetic Epidemiology* 35: 755-765.

Benaglia T, Chauveau D, Hunter DR (2009). An EM-Like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures. *Journal of Computational and Graphical Statistics* 18(2): 505-526.

Benaglia T, Chauveau D, Hunter DR (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software* 32(6): 1-29.

Bipolar disorder (2011). Retrieved from http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001924/ on 8/21/2011.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* 3(1): 1-122.

Breheny P, Huang J (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Statist* 5: 232-253.

Chen J, Khalili A (2008). Order Selection in Finite Mixture Models With A Nonsmooth Penalty. *Journal of the American Statistical Association* 103(484): 1674-1683.

Chen L, Hutter C, Potter J, Liu Y, Prentice R, Peters U, Hsu L (2010). Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. *The American Journal of Human Genetics* 86: 860-871.

Cook RD, Forzani L, Rothman AJ (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Annals of Statistics* 40: 352-384.

Cox DR, Hinkley DV (1974). *Theoretical Statistics*. London: Chapman & Hall.

Crohn's Disease (2010). Retrieved from http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001295/ on 8/21/2011.

De la Cruz O, Wen X, Ke B, Song M, Nicolae D (2010). Gene, Region and Pathway Level Analyses in Whole-Genome Studies. *Genetic Epidemiology* 34: 222-231.

Dempster AP, Laird NM, Rubin DB (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1): 1-38.

DeSarbo WS, Cron WL (1988). Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification* 5(2): 249-282.

Dudbridge F, Koeleman BP (2003). Rank truncated product of P-values, with application to genomewide association scans. *Genetic Epidemiology* 25(4): 360-366.

Evans D, Visscher P, Wray N (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics* 18: 3525-3531.

Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association* 96: 1348-1360.

Fisher RA (1932). *Statistical Methods for Research Workers*, 4th edition. London: Oliver & Boyd.

Franke A, McGovern D, Barrett J, Wang K, Radford-Smith G, Ahmad T, Lees C, Balschun T, Lee J, Roberts R, Anderson C, Bis J, Bumpstead S, Ellinghaus D, Festen E, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew C, Montgomery G, Prescott N, Raychaudhuri S, Rotter J, Schumm P, Sharma Y, Simms L, Taylor K, Whiteman D, Wijmenga C, Baldassano R, Colombel J, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery S, Halfvarson J, Verspaget H, Hugot J, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor D, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart A, Stokkers P, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan S, Brant S, Rioux J, D'Amato M, Weersma R, Kugathasan S, Griffiths A, Mansfield J, Vermeire S, Duerr R, Silverberg M, Satsangi J, Schreiber S, Cho J, Annese V, Hakonarson H, Daly M, Parkes M (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* 42: 1118-1125.

Friedman J, Hastie T, Tibshirani R (2008). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1): 1-22.

Gaffney SJ, Smyth P (2003). Curve clustering with random effects regression mixtures. *Proceedings of the ninth international workshop on artificial intelligence and statistics*. Key West, FL.

Gail M (2008). Discriminatory Accuracy From Single-Nucleotide Polymorphisms in Models to Predict Breast Cancer Risk. *Journal Natl Cancer Inst* 100: 1037-1041.

Gail M (2009). Value of Adding Single-Nucleotide Polymorphism Genotypes to a Breast Cancer Risk Model. *Journal Natl Cancer Inst* 101: 959-963.

Gaya D, Russell R, Nimmo E, Satsangi J (2006). New genes in inflammatory bowel disease: lessons for complex diseases? *Lancet* 367: 1271-1284.

Goeman J, van de Geer S, de Kort F, van Houwelingen H (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20(1): 93-99.

Golub GH, Heath M, Wahba G (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2): 215-223.

Guan Y, Stephens M (2011). Bayesian Variable Selection Regression for Genome-wide Association Studies, and other Large-Scale Problems. *Annals of Applied Statistics* 5(3): 1780-1815.

Ho Y (2009). LiquidAssociation: LiquidAssociation. R package version 1.12.0.

Ho Y, Parmigiani G, Louis TA, Cope LM (2011). Modeling Liquid Association. *Biometrics* 67(1): 133-141.

Hoerl A, Kennard R (1970). Ridge regression: Biased estimation for non-orthogonal problem. *Technometrics* 12: 55-67.

Hunter DR, Young DS (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics* 24(1): 19-38.

Hunter D, Wang S, Hettmansperger Rl (2007). Inference for Mixtures and Symmetric Distributions. *The Annals of Statistics* 33(1): 224-251.

Kang H, Sul J, Service S, Zaitlen N, Kong SY, Freimer N, Sabatti C, Eskin E (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42(4): 348-354.

Kang J, Kugathasan S, Georges M, Zhao H, Cho J.H.; the NIDDK IBD Genetics Consortium (2011) Improved risk prediction for Crohn's disease with a multi-locus approach. *Human Molecular Genetics* 20: 2435-2442.

Khalili A, Chen J (2007). Variables Selection in Finite Mixture of Regression Models *Journal of the American Statistical Association* 102(479): 1025-1038.

Khalili A, Chen J, Lin S (2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics* 12(1): 156-172.

Kooperberg C, LeBlanc M, Obenchain V (2010). Risk Prediction Using Genome-Wide Association Studies. *Genetic Epidemiology* 34: 643-652.

Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008). Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. *PLoS Genet* 4(10): e1000231.

Levine M, Hunter DR, Chauveau D (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* 98(2): 403-416.

Levy D, Ehret G, Rice K, Verwoert G, Launer L, Dehghan A, Glazer N, Morrison A, Johnson A. Aspelund T, Aulchenko Y, Lumley T, Köttgen A, Vasan R, Rivadeneira F, Eiriksdottir G, Guo X, Arking D, Mitchell G, Mattace-Raso F, Smith A, Taylor K, Scharpf R, Hwang SJ, Sijbrads E, Bis J, Harris T, Ganesh S, O'Donnell C, Hofman A, Rotter J, Coresh J, Benjamin E, Uitterlinden A, Heiss G, Fox C, Witteman J, Boerwinkle E, Wang T, Gudnason V, Larson M, Chakravarti A, Psaty B, van Duijn C (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics* 41(6): 666-676.

Li KC (2002). Genome-wide coexpression dynamics: Theory and application. *Proceedings of the National Academy of Sciences of the United States of America* 99(26): 16875-16880.

Madsen B, Browning S (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics* 5(2): 1-11.

Martinez J, Carroll R, Muller S, Sampson J, Chatterjee N (2010). A Note on the Effect on Power of Score Tests via Dimension Reduction by Penalized Regression under the Null. *The International Journal of Biostatistics* 6(1-Article 12): 1-12.

McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, Ioannidiis J, Hirschhorn J (2008). Genome-wide Significance for Dense SNP and Resequencing Data. *Nature Genetics* 9: 356-369.

McLachlan G, Peel D (2000). *Finite Mixture Models.* New York: John Wiley & Sons, Inc.

McNicholas P, Jampani R, McDaid A, Murphy B, Banks L, McNicholas MP (2011). pgmm: Parsimonious Gaussian Mixture Models. R package version 1.0.

Morgenthaler S, Thilly WG (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615(1):28-56.

Pan W (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* 33: 497-507.

Pan W, Shen X, Liu B (2013). Cluster Analysis: Unsupervised Learning via Supervised Learning with a Non-convex Penalty. *Journal of Machine Learning Research* 14(1): 1865-1889.

Park J, Wacholder S, Gail M, Peters U, Jacobs K, Chanock S, Chatterjee N (2010). Estimating effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics* 42: 570-575.

Rakitsch B, Lippert C, Stegle O, Borgwardt K (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics* 29(2): 206-214.

Schaid D, Rowland C, Tines D, Jacobson R, Poland G (2002). Score Tests for Association between Traits and Haplotypes when Linkage Phase is Ambiguous *Am J Hum Genet* 70: 425-434.

Seaman SR, Müller-Myhsok B (2005). Rapid Simulation of P Values for Product Methods and Multiple-Testing Adjustment in Association Studies. *The American Journal of Human Genetics*, 76:399-408.

Shen X, Huang H-C (2006). Optimal model assessment, selection, and combination. *JASA*, 101: 554-568.

Shen X, Huang HH, Pan W (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika* 99(4): 899-914.

Shen X, Pan W, Zhu Y (2012). Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association* 107(497): 223-232.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* 9(12): 3273-3297.

Streuli, H (1973). Der heutige stand der kaffeechemie. *International Colloquium on Coffee Chemisrty* 9: 61-72. Bogatá: Association Scientifique International du Café.

The International Consortium for Blood Pressure Genome-Wide Association Studies (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478: 103-109.

The International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460: 748-752.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678.

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

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* 67(1) 91-108.

Tutz G, Ulbricht J (2009). Penalized regression with correlation-based penalty. *Statistical Computing* 19: 239-253.

Wedel MG, DeSarbo W (1995). A Mixture Likelihood Approach for Generalized Linear Models. *Journal of Classification* 12(1): 21-55.

Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SF, Polychronakos C, Hakonarson H (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics* 2009 5(10):e1000678. Epub 2009 Oct 9.

Wray N, Goddard M, Visscher P (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* 17: 1520-1528.

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

Xu W, Hedeker D (2001). A Random-Effects Mixture Model for Classifying Treatment Response in Longitudinal Clinical Trials. *Journal of Biopharmaceutical Statistics* 11(4): 253–273.

Yang J, Benyamin B, McEvoy B, Gordon S, Henders A, Nyholt D, Madden P, Heath A, Martin N, Montgomery G, Goddard M, Visscher P (2010). Commom SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42: 565-569.

Yang S, Yuan L, Lai Y, Shen X, Wonka P, Ye J (2012). Feature grouping and selection over an undirected graph. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (*KDD*), 922-930, New York: ACM.

Ye J (1998). On measuring and correcting the effects of data mining and model selection. *Journal of American Statistical Association* 93(441): 120-131.

Yu K, Li Q, Bergen A, Pfeiffer R, Rosenberg P, Caporaso N, Kraft P, Chatterjee N (2009). Pathway Analysis by Adaptive Combination of P-Values. *Genetic Epidemiology* 33: 700-709.

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

Yuan Z, Yang Y (2005). Combining linear regression models: when and how? *JASA* 100: 1202-1214.

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002). Truncated product method for combining P?values. *Genetic Epidemiology* 22(2):170-185.

Zou H, Hastie T (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B* 76: 301-320.