

---

Theses and Dissertations

---

Summer 2009

# Regularized methods for high-dimensional and bi-level variable selection

Patrick John Breheny  
*University of Iowa*

Copyright 2009 Patrick John Breheny

This dissertation is available at Iowa Research Online: <http://ir.uiowa.edu/etd/325>

---

## Recommended Citation

Breheny, Patrick John. "Regularized methods for high-dimensional and bi-level variable selection." PhD (Doctor of Philosophy) thesis, University of Iowa, 2009.  
<http://ir.uiowa.edu/etd/325>.

---

Follow this and additional works at: <http://ir.uiowa.edu/etd>



Part of the [Biostatistics Commons](#)

REGULARIZED METHODS FOR HIGH-DIMENSIONAL AND BI-LEVEL  
VARIABLE SELECTION

by

Patrick John Breheny

An Abstract

Of a thesis submitted in partial fulfillment of the  
requirements for the Doctor of Philosophy  
degree in Biostatistics in the  
Graduate College of The  
University of Iowa

July 2009

Thesis Supervisor: Professor Jian Huang

## ABSTRACT

Many traditional approaches to statistical analysis cease to be useful when the number of variables is large in comparison with the sample size. Penalized regression methods have proved to be an attractive approach, both theoretically and empirically, for dealing with these problems. This thesis focuses on the development of penalized regression methods for high-dimensional variable selection. The first part of this thesis deals with problems in which the covariates possess a grouping structure that can be incorporated into the analysis to select important groups as well as important members of those groups. I introduce a framework for grouped penalization that encompasses the previously proposed group lasso and group bridge methods, sheds light on the behavior of grouped penalties, and motivates the proposal of a new method, group MCP.

The second part of this thesis develops fast algorithms for fitting models with complicated penalty functions such as grouped penalization methods. These algorithms combine the idea of local approximation of penalty functions with recent research into coordinate descent algorithms to produce highly efficient numerical methods for fitting models with complicated penalties. Importantly, I show these algorithms to be both stable and linear in the dimension of the feature space, allowing them to be efficiently scaled up to very large problems.

In the third part of this thesis, I extend the idea of false discovery rates to penalized regression. I show how the Karush-Kuhn-Tucker conditions describing penalized regression estimates provide testable hypotheses involving partial residuals, thus connecting the previously disparate fields of multiple comparisons and penalized regression. I then propose two approaches to estimating false discovery rates for penalized regression methods and examine the accuracy of these approaches.

Finally, the methods from all three sections are studied in a number of simulations and applied to real data from microarray and genetic association studies.

Abstract Approved: \_\_\_\_\_  
Thesis Supervisor

\_\_\_\_\_  
Title and Department

\_\_\_\_\_  
Date

REGULARIZED METHODS FOR HIGH-DIMENSIONAL AND BI-LEVEL  
VARIABLE SELECTION

by

Patrick John Breheny

A thesis submitted in partial fulfillment of the  
requirements for the Doctor of Philosophy  
degree in Biostatistics in the  
Graduate College of The  
University of Iowa

July 2009

Thesis supervisor: Professor Jian Huang

Copyright by  
PATRICK JOHN BREHENY  
2009  
All Rights Reserved

Graduate College  
The University of Iowa  
Iowa City, Iowa

CERTIFICATE OF APPROVAL

---

PH.D. THESIS

---

This is to certify that the Ph.D. thesis of

Patrick John Breheny

has been approved by the Examining Committee  
for the thesis requirement for the Doctor of  
Philosophy degree in Biostatistics at the July 2009  
graduation.

Thesis Committee: \_\_\_\_\_  
Jian Huang, Thesis Supervisor

\_\_\_\_\_  
Joseph Cavanaugh

\_\_\_\_\_  
Michael Jones

\_\_\_\_\_  
Kai Wang

\_\_\_\_\_  
Dale Zimmerman

## ACKNOWLEDGMENTS

I would like to thank all of my professors in both the statistics and biostatistics departments at the University of Iowa for opening my eyes to the fascinating world of statistics, for their advice, and for providing me with an excellent education. Specifically, I would like to thank Luke Tierney for introducing me to computer-intensive statistics and my advisor Jian Huang for introducing me to the field of penalized regression and for two years of support, guidance, and thought-provoking conversation. His open-mindedness and intelligence have been an inspiration to me, and I couldn't ask for a better role model.

I would also like to thank Amy Andreotti, who taught me how to be a scientist, my mom and dad for encouraging my curiosity at a young age, my amazing wife Wyndee for making me happy, and my daughter Nyla for debugging all my code.



## ABSTRACT

Many traditional approaches to statistical analysis cease to be useful when the number of variables is large in comparison with the sample size. Penalized regression methods have proved to be an attractive approach, both theoretically and empirically, for dealing with these problems. This thesis focuses on the development of penalized regression methods for high-dimensional variable selection. The first part of this thesis deals with problems in which the covariates possess a grouping structure that can be incorporated into the analysis to select important groups as well as important members of those groups. I introduce a framework for grouped penalization that encompasses the previously proposed group lasso and group bridge methods, sheds light on the behavior of grouped penalties, and motivates the proposal of a new method, group MCP.

The second part of this thesis develops fast algorithms for fitting models with complicated penalty functions such as grouped penalization methods. These algorithms combine the idea of local approximation of penalty functions with recent research into coordinate descent algorithms to produce highly efficient numerical methods for fitting models with complicated penalties. Importantly, I show these algorithms to be both stable and linear in the dimension of the feature space, allowing them to be efficiently scaled up to very large problems.

In the third part of this thesis, I extend the idea of false discovery rates to penalized regression. I show how the Karush-Kuhn-Tucker conditions describing penalized regression estimates provide testable hypotheses involving partial residuals, thus connecting the previously disparate fields of multiple comparisons and penalized regression. I then propose two approaches to estimating false discovery rates for penalized regression methods and examine the accuracy of these approaches.

Finally, the methods from all three sections are studied in a number of simulations and applied to real data from microarray and genetic association studies.

# TABLE OF CONTENTS

|   |      |
|---|------|
| LIST OF TABLES . . . . .  | vii  |
| LIST OF FIGURES . . . . .   | viii |
| CHAPTER   |      |
| 1 INTRODUCTION . . . . .  | 1    |
| 1.1 Ridge regression and the lasso . . . . .  | 1    |
| 1.2 SCAD and MCP . . . . .  | 5    |
| 1.3 Overview of the thesis . . . . .  | 7    |
| 2 PENALIZED METHODS FOR BI-LEVEL VARIABLE SELECTION . . . . .                               | 8    |
| 2.1 A general framework for group penalization . . . . .                                    | 10   |
| 2.2 Group MCP . . . . .   | 11   |
| 2.3 Loss functions . . . . .  | 13   |
| 2.4 Simulations . . . . .   | 14   |
| 2.5 Application: Genetic association study of age-related macular<br>degeneration . . . . . | 18   |
| 3 LOCAL COORDINATE DESCENT ALGORITHMS . . . . .   | 21   |
| 3.1 Local coordinate descent . . . . .  | 21   |
| 3.2 Group MCP . . . . .   | 23   |
| 3.3 Group bridge . . . . .  | 24   |
| 3.4 Group lasso . . . . .   | 25   |
| 3.5 Convergence of the LCD algorithm . . . . .  | 26   |
| 3.6 Pathwise optimization and initial values . . . . .                                      | 28   |
| 3.7 Regularization parameter selection . . . . .  | 30   |
| 3.8 Adding an $L_2$ penalty . . . . .   | 31   |
| 3.9 Simulations . . . . .   | 32   |
| 3.9.1 Efficiency . . . . .  | 32   |
| 3.9.2 Regularization parameter selection . . . . .  | 34   |
| 4 FALSE DISCOVERY RATES FOR PENALIZED REGRESSION . . . . .                                  | 36   |
| 4.1 The false discovery rate of the lasso . . . . .   | 37   |
| 4.1.1 Linear regression . . . . .   | 39   |
| 4.1.2 Logistic regression . . . . .   | 41   |
| 4.1.3 Other penalties . . . . .   | 42   |
| 4.2 Heuristic approach . . . . .  | 43   |
| 4.3 Linear predictor approach . . . . .   | 44   |
| 4.3.1 Linear regression . . . . .   | 45   |
| 4.3.2 Logistic regression . . . . .   | 48   |
| 4.4 Simulations . . . . .   | 50   |

|       |  |    |
|-------|--|----|
| 4.4.1 | Accuracy . . . . .                           | 50 |
| 4.4.2 | Lasso FDR vs. competing approaches . . . . . | 54 |
| 4.5   | Applications . . . . .                       | 58 |
| 4.5.1 | Gene expression data . . . . .               | 58 |
| 4.5.2 | Genetic association study . . . . .          | 60 |
| 5     | SUMMARY . . . . .                            | 61 |
|       | REFERENCES . . . . .                         | 63 |

## LIST OF TABLES

Table

|     |  |    |
|-----|--|----|
| 2.1 | Simulation: Selection of variables and groups by group MCP, group lasso, and group bridge. . . . .   | 16 |
| 2.2 | Application of group penalization and a one-at-a-time methods to a genetic association study of age-related macular degeneration . . . .                                     | 19 |
| 3.1 | LCD algorithm efficiency: Linear regression, $n = 500$ and $p = 200$ . .   | 33 |
| 3.2 | LCD algorithm efficiency: Logistic regression, $n = 1000$ and $p = 200$ .  | 33 |
| 3.3 | LCD algorithm efficiency: Linear regression, $n = 500$ and $p = 2000$ .  | 34 |
| 4.1 | Schematic: Possible outcomes of feature selection . . . . .  | 38 |
| 4.2 | Simulation design: Number of causative, correlated, and spurious features for each setting . . . . .   | 55 |
| 4.3 | Features selected at various FDR levels by $t$ -test and lasso approaches for a gene expression study of leukemia. . . . .   | 59 |
| 4.4 | Number of features selected at various FDR levels by univariate trend test and lasso approaches for a genetic association study of age-related macular degeneration. . . . . | 60 |

## LIST OF FIGURES

Figure

|     |  |    |
|-----|--|----|
| 1.1 | Shapes of the bridge family of penalties . . . . .   | 4  |
| 1.2 | Shapes of SCAD and MCP penalties . . . . .   | 6  |
| 2.1 | Shapes of group penalties . . . . .  | 13 |
| 2.2 | Simulation results: Model error of group penalization methods . . .  | 16 |
| 2.3 | Simulation results: Prediction error of group penalization methods .   | 17 |
| 3.1 | Coefficient paths for group penalization methods . . . . .   | 28 |
| 3.2 | Simulation: Comparison of degree of freedom estimators for group<br>penalization methods . . . . .                       | 35 |
| 4.1 | Simulation: Accuracy of FDR estimation for the heuristic approach  | 52 |
| 4.2 | Simulation: Accuracy of FDR estimation for the linear predictor ap-<br>proach . . . . .                                  | 53 |
| 4.3 | Simulation: Causative and spurious features selected by various ap-<br>proaches (uncorrelated design) . . . . .          | 56 |
| 4.4 | Simulation: Causative, correlated, and spurious features selected by<br>various approaches (correlated design) . . . . . | 57 |

## CHAPTER 1 INTRODUCTION

Variable selection is an important issue in regression. Typically, measurements are obtained for a large number of potential predictors in order to avoid missing an important link between a predictive factor and the outcome. However, to reduce variability and obtain a more interpretable model, we are often interested in selecting a smaller number of important variables. The low cost and easy implementation of automated methods for data collection and storage has led to a recent abundance of problems for which the number of variables is large in comparison to the sample size.

For these high-dimensional problems, many traditional approaches to variable selection cease to be useful due to computational infeasibility, model nonidentifiability, or both. Incorporating additional information into such problems becomes a necessity. Penalized regression models are one attractive approach that has proven successful – both theoretically and empirically – for dealing with high-dimensional data.

This chapter introduces the concepts behind penalized regression and provides background on several previously proposed penalized regression methods. More specific introductions will be presented at the beginning of the relevant chapters, and overlap somewhat with the general introduction here.

### 1.1 Ridge regression and the lasso

Problems in which  $p$  is large in comparison to  $n$  present a problem for regression models. Solutions become unstable or no longer uniquely defined and models become difficult to interpret. Furthermore, searching through subsets of potential predictors for a good model is both unstable (Breiman, 1996) and computationally unfeasible even for moderately sized  $p$ . Penalized regression methods are much less

susceptible to these issues and have become a popular approach to dealing with these problems.

Suppose we have  $n$  observations indexed by  $i$ . Each observation contains measurements of an outcome  $y_i$  and  $p$  features  $\{x_{i1}, \dots, x_{ip}\}$  indexed by  $j$ . We assume without loss of generality that the features have been standardized such that  $\sum_{i=1}^n x_{ij} = 0$  and  $\frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1$ . Typically, an intercept term  $\beta_0$  is also included in the model, and left unpenalized. As we shall see, this ensures that the penalty is applied equally to all covariates in an equivariant manner. This is standard practice in regularized estimation; estimates are then transformed back to their original scale after the penalized models have been fit.

The problem of interest involves estimating a sparse vector of regression coefficients  $\boldsymbol{\beta}$ . Penalized regression methods accomplish this by minimizing an objective function  $Q$  that is composed of a loss function  $L$  plus a penalty function  $P$ :

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) + P_\lambda(\boldsymbol{\beta}), \quad (1.1)$$

where  $P$  is a function of the coefficients indexed by a parameter  $\lambda$  which controls the degree of penalization. Typically, the penalty function  $P$  has the following properties: it is symmetric about the origin,  $P(\mathbf{0}) = 0$ , and  $P$  is nondecreasing in  $\|\boldsymbol{\beta}\|$ .

This approach produces a spectrum of solutions depending on the value of  $\lambda$ ; such methods are often referred to as *regularization* methods, and  $\lambda$  is called the regularization parameter. The majority of the regularization literature concerns the least squares loss function, but least absolute deviation and negative log-likelihood loss functions are also common. In this chapter, we will leave the loss function unspecified.

Regularization naturally lends itself to a Bayesian interpretation in which  $P$  is the negative log-prior of the coefficients; the difference is that instead of sampling



from the posterior distribution, regularization methods find the posterior mode. Typically, penalties are chosen such that the standardized covariates are treated equally. More specific prior beliefs regarding the nature of the coefficients might motivate one to propose other penalties; this thesis, however, deals only with “objective” cases.

Regularization methods date back to the proposal of ridge regression by Hoerl and Kennard (1970). They proposed the objective function

$$Q_{\text{ridge}}(\boldsymbol{\beta}) = \frac{1}{2n}L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^J \beta_j^2. \quad (1.2)$$

Ridge regression shrinks coefficients towards 0 by imposing a penalty on their size. In an unpenalized, underdetermined model, a very large positive coefficient on one variable may be canceled by a very large negative coefficient on another variable; ridge regression prevents this from happening and yields unique solutions for all  $\lambda > 0$ , even when  $p > n$ .

Ridge regression produces very stable estimates and performs very well in certain settings, but it has two central flaws. First, ridge regression heavily penalizes large coefficients, leading to badly biased estimates when some of the coefficients are large. Second, ridge regression does not produce sparse solutions and thus fails to improve the interpretability of the model.

To remedy these flaws, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (lasso), which minimizes the objective function

$$Q_{\text{lasso}}(\boldsymbol{\beta}) = \frac{1}{2n}L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^J |\beta_j|. \quad (1.3)$$

Provided that  $\lambda$  is sufficiently large, a portion of the values that make up  $\hat{\boldsymbol{\beta}}$  will be exactly 0 for the lasso penalty function. Thus, the lasso provides a continuous subset selection procedure. Furthermore, the solutions to the lasso are less downwardly biased than those of ridge regression.

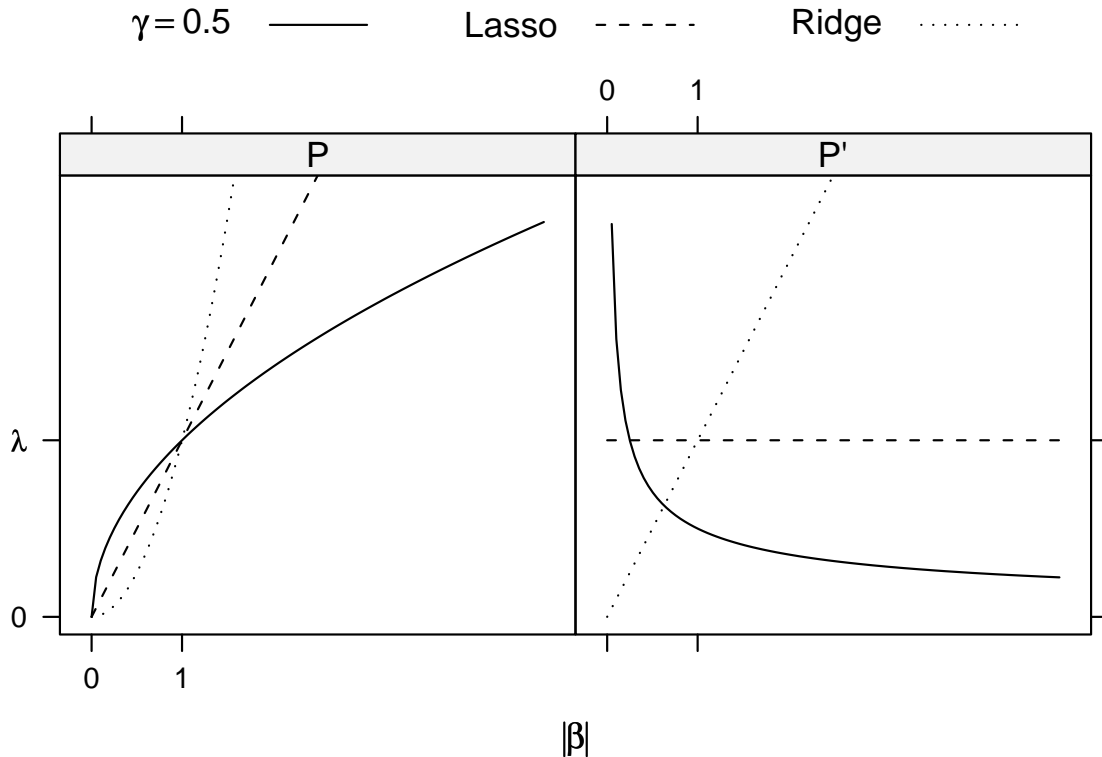


Figure 1.1: Shapes of the bridge family of penalties. On the horizontal axis is the absolute value of the regression coefficient. For the panel on the left, the penalty itself is on the vertical axis, while for the panel on the right, the derivative of the penalty is plotted on the vertical axis.

The form of the lasso and ridge penalties are very similar, the only difference being the exponent to which  $|\beta|$  is raised, which we will denote  $\gamma$ . Indeed, even prior to the publication of the lasso, Frank and Friedman (1993) proposed a general family of penalties in which  $\gamma$  is allowed to vary over all nonnegative values. The members of this family include both ridge and lasso and are known collectively as the “bridge” penalties. The shapes of three members of this family are illustrated in Figure 1.1, with the exponent  $\gamma$  equal to 2 (ridge), 1 (lasso), or 1/2.

The differences between the bridge penalties are apparent from the plots in Figure 1.1, in particular the plot of their derivatives. The rate of penalization for ridge regression increases with  $|\beta|$ , thus applying little to no penalization near 0 while strongly discouraging large coefficients. Meanwhile, the rate of penalization

for the lasso is constant – and, notably, greater than zero at  $|\beta| = 0$ , thereby producing sparse solutions. Finally, setting  $\gamma = 1/2$  results in a rate of penalization that is very high near 0 but steadily diminishes as  $\beta$  grows larger. This square-root penalty produces solutions that are even more sparse and less biased than the lasso. However, the rate of penalization goes to  $\infty$  as  $|\beta|$  goes to 0, which produces computational and analytic problems. Furthermore, the objective function for this penalty is no longer convex, and will therefore possess multiple minima.

## 1.2 SCAD and MCP

Another type of penalty outside of the bridge family was proposed by Fan and Li (2001). The penalty they propose, the smoothly clipped absolute deviation (SCAD) penalty, begins by applying the same rate of penalization as the lasso, but continuously relaxes that penalization until, when  $|\beta| \geq a\lambda$ , the rate of penalization drops to 0. The minimax concave penalty (MCP), proposed by (Zhang, 2007), behaves similarly, and the connection between the two methods are explored in detail by its author. The two penalties are plotted in Figure 1.2.

The goal of both penalties is to eliminate the unimportant variables from the model while leaving the important variables unpenalized. This would be equivalent to fitting an unpenalized model in which the truly nonzero variables are known in advance (the so-called “oracle” model). Both MCP and SCAD accomplish this asymptotically and are said to have the oracle property (Fan and Li, 2001; Zhang, 2007).

From Figure 1.2, we can observe that  $\lambda$  is the regularization parameter that determines the magnitude of penalization and  $a$  is a tuning parameter that affects the range over which the penalty is applied.

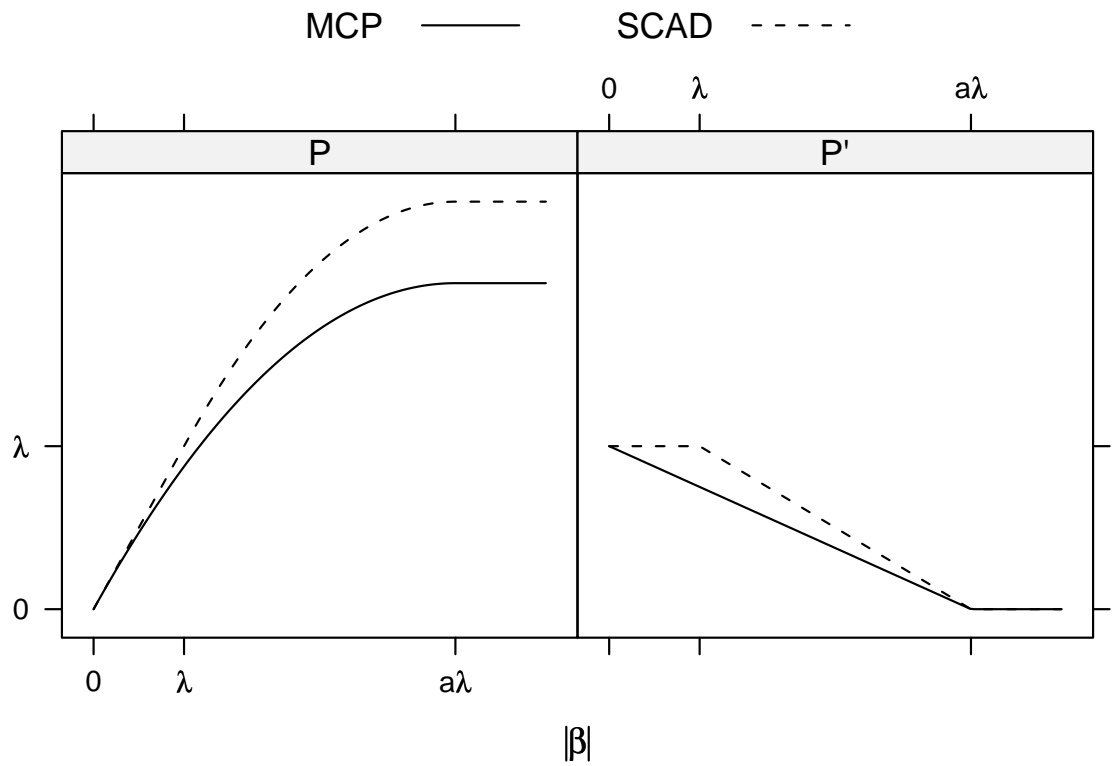


Figure 1.2: Shapes of SCAD and MCP penalties. On the horizontal axis is the absolute value of the regression coefficient. For the panel on the left, the penalty itself is on the vertical axis, while for the panel on the right, the derivative of the penalty is plotted on the vertical axis.

### 1.3 Overview of the thesis

In Chapter 2, I discuss penalized variable selection in the context of grouped covariates, in which the goal is to select important groups as well as important members of those groups – a problem of bi-level selection. This chapter introduces a framework which lends insight into previously proposed approaches and motivates a new method which I call group MCP.

In Chapter 3, I develop algorithms to fit penalized regression models with complicated penalties such as the group penalization methods discussed in Chapter 2. I demonstrate that these algorithms are stable and numerically efficient for high dimensional regression problems and can therefore be efficiently scaled up to very large problems.

The false discovery rate (FDR) is a statistically sound and intuitively appealing approach for assessing the number of false positives likely to arise when identifying important features from a large number of candidates. Chapter 4 extends the FDR idea to penalized regression and develops estimators for the FDR of penalized regression approaches.

Within these chapters, the empirical properties of the methods are investigated under a variety of simulations as well as applied to real data sets from two important high-dimensional problems in modern biomolecular research: gene expression and genetic association studies. Finally, the results of this thesis will be discussed and summarized in Chapter 5.

## CHAPTER 2

### PENALIZED METHODS FOR BI-LEVEL VARIABLE SELECTION

There is a large body of available literature on the topic of variable selection, but the majority of this work is focused on the selection of individual variables. In many regression problems, however, predictors are not distinct but arise from common underlying factors. Categorical factors are often represented by a group of indicator functions; likewise for continuous factors and basis functions. Groups of measurements may be taken in the hopes of capturing unobservable latent variables or of measuring different aspects of complex entities. Some specific examples include measurements of gene expression, which can be grouped by pathway, and genetic markers, which can be grouped by the gene or haplotype that they belong to. Methods for individual variable selection may perform inefficiently in these settings by ignoring the information present in the grouping structure, or even give rise to models that are not sensible.

In this chapter, I consider  $\mathbf{x}_i$  as being composed of an unpenalized intercept and  $J$  groups  $\mathbf{x}_{ij}$ , with  $K_j$  denoting the size of group  $j$ . The quantity  $x_{ijk}$  is therefore the  $k^{\text{th}}$  covariate of the  $j^{\text{th}}$  group for the  $i^{\text{th}}$  observation. The coefficient vector  $\boldsymbol{\beta}$  is composed likewise. Covariates that do not belong to a group may be thought of as a group of one.

Methods that take into account grouping information have recently begun to appear in the penalized regression literature. Yuan and Lin (2006) proposed the group lasso, in which  $\hat{\boldsymbol{\beta}}$  is defined to be the value that minimizes the objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^J \sqrt{K_j} \|\boldsymbol{\beta}_j\|, \quad (2.1)$$

where  $\|\cdot\|$  is the  $L_2$  norm. This penalty enforces sparsity at the group level, rather

than at the level of the individual covariates. Within a group, the covariates are either all equal to zero or else all nonzero.

The group lasso has some attractive qualities, such as the fact that its objective function is convex (*i.e.*, there are no local minima, only a single global minimum). However, the group lasso also has a number of drawbacks: it produces a strong bias towards zero, it tends to overselect the true number of groups, and it is incapable of selecting important elements within a group. To address these shortcomings, Huang et al. (2007) proposed the group bridge, whose estimate minimizes

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^J K_j^\gamma \|\boldsymbol{\beta}_j\|_1^\gamma, \quad (2.2)$$

where  $\|\cdot\|_1$  is the  $L_1$  norm. Throughout this paper, we take  $\gamma = 1/2$  for group bridge.

The group bridge produces sparse solutions both at the group level and at the level of the individual covariates within a group (we will refer to this as *bi-level selection*). Furthermore, its solutions tend to exhibit less bias than those of the group lasso and have been shown to be asymptotically consistent for group selection. Unlike the group lasso, however, the group bridge objective function is nonconvex and not differentiable at  $|\boldsymbol{\beta}_j| = \mathbf{0}$ , which in practice can lead to problems with model fitting.

Instead of considering these methods to be entirely distinct, in this chapter we will view them as part of a general framework for group penalization, which not only lends insight into their behavior but opens the door for further methodological development along these lines. We will take advantage of this general framework to develop and explore the behavior of a new method for bi-level selection, group

MCP.

## 2.1 A general framework for group penalization

As discussed in Chapter 1, the effect of a penalty upon the solution is determined by its derivative. To better understand the action of these penalties and to illuminate the development of new ones, we can consider grouped penalties to have a form in which an outer penalty  $f_O$  is applied to a sum of inner penalties  $f_I$ . The penalty applied to a group of covariates is

$$f_O \left( \sum_{k=1}^{K_j} f_I(|\beta_{jk}|) \right) \quad (2.3)$$

and the partial derivative with respect to the  $jk^{\text{th}}$  covariate is

$$f'_O \left( \sum_{k=1}^{K_j} f_I(|\beta_{jk}|) \right) f'_I(|\beta_{jk}|). \quad (2.4)$$

Note that both group lasso and group bridge fit into this framework with an outer bridge ( $\gamma = 1/2$ ) penalty; the former possesses an inner ridge penalty, while the latter has an inner lasso penalty. We have intentionally left the above framework general in the sense of not rigidly specifying the role of constants or tuning parameters such as  $\lambda$ ,  $\gamma$ , or  $\sqrt{K_j}$ . A more specific framework would obscure the main point as well as create the potential of excluding useful forms.

From (2.4), we can understand group penalization to be applying a rate of penalization to a covariate that consists of two terms: the first carrying information regarding the group; the second carrying information about the individual covariate. Variables can enter the model either by having a strong individual signal or by being a member of a group with a strong collective signal. Conversely, a variable with a strong individual signal can be excluded from a model through its association with a preponderance of weak group members.

However, one must be careful not to let it oversimplify the situation. Casually



combining penalties will not necessarily lead to reasonable results. For example, using the lasso as both inner and outer penalty is equivalent to the conventional lasso, and makes no use of grouping structure. Furthermore, properties may emerge from the combination that are more than the sum of their parts. The group lasso, for instance, possesses a convex penalty despite the fact that its outer bridge penalty is nonconvex. Nevertheless, the framework described above is a helpful lens through which to view the problem of group penalization which emphasizes the dominant feature of the method: the gradient of the penalty and how it varies over the feature space.

## 2.2 Group MCP

Zhang (2007) proposes a nonconvex penalty called MCP which possesses attractive theoretical properties. MCP and its derivative are defined on  $[0, \infty)$  by

$$f_{\lambda,a}(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2a} & \text{if } \theta \leq a\lambda \\ \frac{1}{2}a\lambda^2 & \text{if } \theta > a\lambda \end{cases} \quad f'_{\lambda,a}(\theta) = \begin{cases} \lambda - \frac{\theta}{a} & \text{if } \theta \leq a\lambda \\ 0 & \text{if } \theta > a\lambda \end{cases} \quad (2.5)$$

for  $\lambda \geq 0$ . The rationale behind the penalty can again be understood by considering its derivative: MCP begins by applying the same rate of penalization as the lasso, but continuously relaxes that penalization until, when  $\theta > a\lambda$ , the rate of penalization drops to 0. MCP is motivated by and rather similar to SCAD. The connections between MCP and SCAD are explored in detail by Zhang (2007). The derivatives of MCP and SCAD were plotted in Figure 1.2.

The goal of both penalties is to eliminate the unimportant variables from the model while leaving the important variables unpenalized. This would be equivalent to fitting an unpenalized model in which the truly nonzero variables are known in advance (the so-called ‘‘oracle’’ model). Both MCP and SCAD accomplish this asymptotically and are said to have the oracle property (Fan and Li, 2001; Zhang,

2007).

From Figure 1.2, we can observe that  $\lambda$  is the regularization parameter that determines the magnitude of penalization and  $a$  is a tuning parameter that affects the range over which the penalty is applied. When  $a$  is small, the region in which MCP is not constant is small; when  $a$  is large, MCP penalty has a broader influence. Generally speaking, small values of  $a$  are best at retaining the unbiasedness of the SCAD penalty for large coefficients, but they also run the risk of creating objective functions with problematic nonconvexity that are difficult to optimize and yield solutions that are discontinuous with respect to  $\lambda$ . It is therefore best to choose an  $a$  that is big enough to avoid problems but not too big. Zhang (2007) discusses the issue of choosing  $a$  in depth; for the results of sections 2.4 and 2.5, we use  $a = 3$  for penalized linear regression and  $a = 30$  for penalized logistic regression.

The group MCP estimate minimizes

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + \sum_{j=1}^J f_{\lambda,b} \left( \sum_{k=1}^{K_j} f_{\lambda,a}(|\beta_{jk}|) \right), \quad (2.6)$$

where  $b$ , the tuning parameter of the outer penalty, is chosen to be  $K_j a \lambda / 2$  in order to ensure that the group level penalty attains its maximum if and only if each of its components are at their maximum. In other words, the derivative of the outer penalty reaches 0 if and only if  $|\beta_{jk}| \geq a\lambda$  for all  $k \in \{1, \dots, K_j\}$ . The relationship between group lasso, group bridge, and group MCP is illustrated for a two-covariate group in Figure 2.1.

One can see from Figure 2.1 that the group MCP penalty is capped at both the individual covariate and group levels, while the group lasso and group bridge penalties are not. This illustrates the two rationales of group MCP: (1) to avoid overshrinkage by allowing covariates to grow large, and (2) to allow groups to remain sparse internally. Group bridge allows the presence of a single large predictor to continually lower the entry threshold of the other variables in its group. This

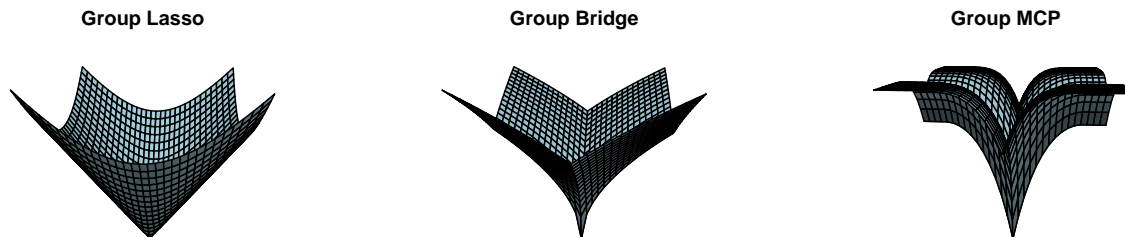


Figure 2.1: Shapes of group penalties. The penalty applied to a two-covariate group is plotted, with the two coefficients on the horizontal axes and the penalty on the vertical axis. The group lasso, group bridge, and group MCP penalties are illustrated. Note that where the penalty comes to a point or edge, there is the possibility that the solution will take on a sparse value; all penalties come to a point at  $\mathbf{0}$ , encouraging group-level sparsity, but only group bridge and group MCP allow for bi-level selection.

property, whereby a single strong predictor drags others into the model, prevents group bridge from achieving consistency for the selection of individual variables. Group MCP, on the other hand, limits the extent to which a single predictor can reduce the penalty applied to the other members of the group.

### 2.3 Loss functions

Much of the work in the field of penalized regression has focused on squared error loss:

$$\frac{1}{2n}L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^n (y_i - \mathbf{x}'_i\boldsymbol{\beta})^2.$$

In principle, however, a given penalty may be applied to any loss function, the most important class of which being likelihood-derived loss functions such as the (negative) log likelihood of a generalized linear model (McCullagh and Nelder, 1999) or the partial likelihood of a Cox proportional hazards regression model.

The advantage of squared error loss is computational tractability: algorithms are generally easy to implement and efficient. However, for many loss functions, we can make a quadratic approximation to the loss function using the current estimate

of the linear predictors  $\boldsymbol{\eta}^{(m)}$ , and update coefficients using an iteratively reweighted least squares algorithm:

$$L(\boldsymbol{\eta}) \approx L(\boldsymbol{\eta}^{(m)}) + (\boldsymbol{\eta} - \boldsymbol{\eta}^{(m)})' \mathbf{v} + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(m)})' \mathbf{W}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(m)}),$$

where  $\mathbf{v}$  and  $\mathbf{W}$  are the first and second derivatives of  $L(\boldsymbol{\eta})$  with respect to  $\boldsymbol{\eta}$ , evaluated at  $\boldsymbol{\eta}^{(m)}$ . Now, letting  $\mathbf{z} = \boldsymbol{\eta}^{(m)} - \mathbf{W}^{-1}\mathbf{v}$  and dropping terms that are constant with respect to  $\boldsymbol{\beta}$ , we can complete the square to obtain

$$L(\boldsymbol{\beta}) \approx \frac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}). \quad (2.7)$$

For generalized linear models,  $\mathbf{W}$  is a diagonal matrix, and the quadratic approximation renders the loss function equivalent to squared error loss in which the observations are weighted by  $\mathbf{w} = \text{diag}(\mathbf{W})$ . This allows algorithms developed for squared error loss to be easily adapted to generalized linear models, with slight modifications for the iterative reweighting. Chapter 3 will discuss algorithms for fitting penalized regression models in greater detail.

## 2.4 Simulations

In this section, we will compare the performance of the group lasso, group bridge, and group MCP methods across a variety of independently generated data sets. Here, we use BIC as the model selection criterion; simulations I have conducted for logistic regression and using AIC and GCV all illustrate the same basic trends.

Data were simulated from the generating model

$$y_i = \mathbf{x}'_{i1}\boldsymbol{\beta}_1^{(0)} + \dots + \mathbf{x}'_{i10}\boldsymbol{\beta}_{10}^{(0)} + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} \text{N}(0, 1), \quad (2.8)$$

with 100 observations and 10 groups, each of which contained 10 members ( $n = p = 100$ ). The sparsity of the underlying models varied over a range of true nonzero groups  $J_0 \in 2, 3, 4, 5$  and over a range of nonzero members within a group  $K_0 \in 2, 3, \dots, 10$ . Furthermore, the magnitude of the coefficients was determined

according to

$$\beta_{jk}^{(0)} = a j k I(j \leq J_0) I(k \leq K_0),$$

where  $a$  was chosen such that the signal to noise ratio (SNR) of the model was approximately one (actual range from 0.84 to 1.45). This specification ensures that each model covers a spectrum of groups ranging from those with with small effects to those with large effects, and that each group contains large and small contributors. Model error (ME), mean squared prediction error (MSPE), and SNR are defined as follows:

$$\begin{aligned} \text{ME} &= (\hat{\beta} - \beta^{(0)})' \text{E}(\mathbf{xx}') (\hat{\beta} - \beta^{(0)}), \\ \text{MSPE} &= \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\beta})' (\mathbf{y} - \mathbf{X}\hat{\beta}), \end{aligned}$$

and

$$\text{SNR} = \frac{1}{\sigma^2} \beta^{(0)'} \text{E}(\mathbf{xx}') \beta^{(0)}.$$

For each combination of  $J_0$  and  $K_0$ , 500 independent data sets were generated. We note the average number of groups and coefficients selected by the approaches for two representative cases in Table 2.1, plot model errors in Figure 2.2, and plot mean squared prediction error in Figure 2.3.

The most striking difference between the methods is the extent to which the form of the penalty enforces grouping: group lasso forces complete grouping, group MCP encourages grouping to a rather slight extent, and group bridge is somewhere in between. This is seen most clearly by observing the average number of variables selected per group for the cases listed in Table 2.1. For group lasso, of course, this number is always 10. For group MCP, approximately two or three variables were selected per group, while group bridge selected four or five per group.

Table 2.1: Simulation: Selection of variables and groups by group MCP, group lasso, and group bridge.

|                  | Variables | Groups                          |     |     | Variables |      |      |
|------------------|-----------|---------------------------------|-----|-----|-----------|------|------|
|                  | / group   | Selected                        | FP  | FN  | Selected  | FP   | FN   |
| Generating model |           | 3 groups, 3 variables per group |     |     |           |      |      |
| Group lasso      | 10.0      | 2.9                             | 0.3 | 0.4 | 28.5      | 20.7 | 1.2  |
| Group bridge     | 4.2       | 2.5                             | 0.3 | 0.8 | 9.9       | 5.2  | 4.3  |
| Group MCP        | 2.2       | 5.9                             | 3.0 | 0.1 | 12.6      | 7.5  | 3.9  |
| Generating model |           | 3 groups, 8 variables per group |     |     |           |      |      |
| Group lasso      | 10.0      | 2.9                             | 0.2 | 0.3 | 28.9      | 7.3  | 2.4  |
| Group bridge     | 5.0       | 2.5                             | 0.3 | 0.8 | 11.8      | 2.1  | 14.3 |
| Group MCP        | 2.7       | 5.6                             | 2.6 | 0.0 | 14.4      | 4.7  | 14.3 |

FP=False positive; FN=False negative

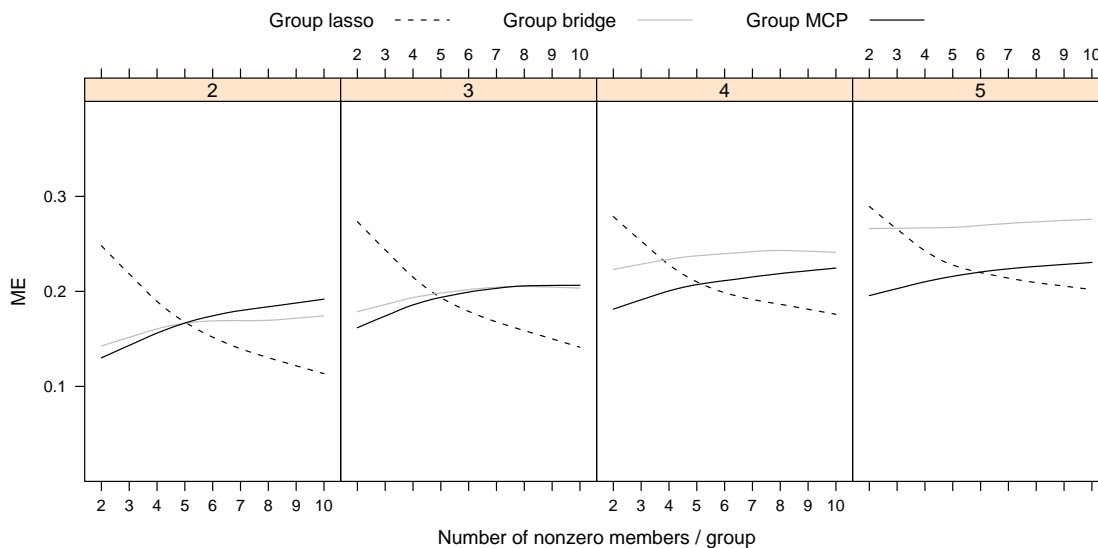


Figure 2.2: Simulation results: Model error of group penalization methods. In each panel, the number of nonzero groups is indicated in the strip at the top. The  $x$ -axis represents the number of nonzero elements per group. At each tick mark, 500 data sets were generated. A lowess curve has been fit to the points and plotted.

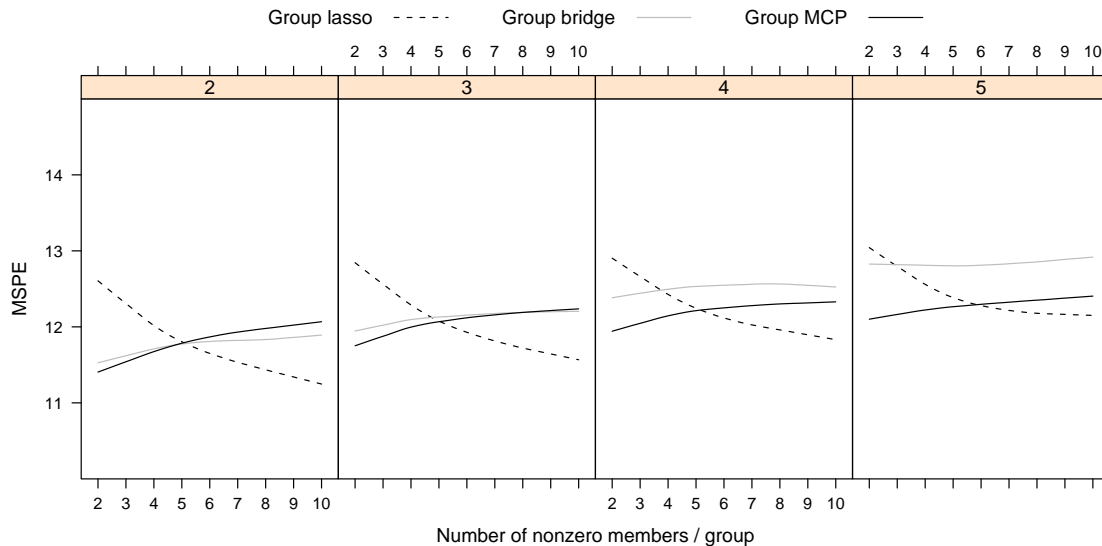


Figure 2.3: Simulation results: Prediction error of group penalization methods. In each panel, the number of nonzero groups is indicated in the strip at the top. The  $x$ -axis represents the number of nonzero elements per group. At each tick mark, 500 data sets were generated. A lowess curve has been fit to the points and plotted.

This number varies little with the true number of variables per group. Although this may seem surprising, it is important to keep in mind that the study design in question contains variables with heterogeneous effects within groups. The main difference between the methods is the extent to which the selection threshold is lowered for the group members given that the group has already been selected. As remarked upon in section 2.2, the extent to which this occurs is lower for group MCP than for group bridge. For group lasso of course, the threshold is lowered to 0 given that a group is selected, and thus all members are selected.

Based on these selection properties, we would expect group MCP to perform best in settings where the underlying model is relatively sparse. Indeed, that is exactly what is seen in Figures 2.2 and 2.3. Based on this simulation study, I would put forward the following advice regarding the choice between group penalization methods: if one expects that the proportion of nonzero group members to be greater

than one-half, use group lasso; otherwise, use group MCP. If one expects this proportion to be close to one-half, one may wish to use group bridge. However, as we will see in Chapter 3, the fact that the gradient of the group bridge penalty goes to  $\infty$  as  $\beta_j$  goes to zero raises concerns about the model fitting process.

## 2.5 Application: Genetic association study of age-related macular degeneration

Genetic association studies are an increasingly important tool for detecting links between genetic markers and diseases. The example that we will consider here involves data from a case-control study of age-related macular degeneration consisting of 400 cases and 400 controls. We confine our analysis to 30 genes that previous biological studies have suggested may be related to the disease. These genes contained 532 markers with acceptably low rates of missing data ( $< 20\%$  no call rate) and high minor allele frequency ( $> 10\%$ ).

We analyzed the data with the group lasso, group bridge, and group MCP methods by considering markers to be grouped by the gene they belong to. Logistic regression models were fit assuming an additive effect for all markers (homozygous dominant = 2, heterozygous = 1, homozygous recessive = 0). Missing (“no call”) data was imputed from the nearest non-missing marker for that subject. In addition to the group penalization methods, we analyzed these data using a traditional one-at-a-time approach, in which univariate logistic regression models were fit and marker effects tested using a  $p < .05$  cutoff. For group lasso and group bridge, using BIC to select  $\lambda$  resulted in the selection of the intercept-only model. Thus, more liberal model selection criteria were used for those methods: AIC for group lasso and GCV for group bridge.

To assess the performance of these methods, we computed 10-fold cross-validation error rates for the methods. For the one-at-a-time approach, predictions



Table 2.2: Application of group penalization and a one-at-a-time methods to a genetic association study of age-related macular degeneration.

|               | # of<br>groups | # of<br>covariates | Error<br>rate | CV error<br>rate |
|---------------|----------------|--------------------|---------------|------------------|
| One-at-a-time | 19             | 47                 | .302          | .450             |
| Group lasso   | 7              | 139                | .318          | .429             |
| Group bridge  | 2              | 11                 | .372          | .414             |
| Group MCP     | 10             | 15                 | .354          | .408             |

were made from an unpenalized logistic regression model fit to the training data using all the markers selected by individual testing. The results are presented in Table 2.2.

Table 2.2 strongly suggests the benefits of using group penalized models as opposed to one-at-a-time approaches: the three group penalization methods achieve lower test error rates and do so while selecting fewer groups. Although the fact that the error rates exceed 0.4 indicate that these 30 genes likely do not include SNPs that exert an overwhelming effect on an individual's chances of developing age-related macular degeneration, the fact that they are well below 0.5 demonstrates that these genes do contain SNPs related to the disease. In particular, bi-level selection methods seem to perform quite well for these data. Group bridge identifies 3 promising genes out of 30 candidates, and group MCP achieves a similarly low test error rate while identifying 10 promising SNPs out of 532.

There are a number of important practical issues that arise in genetic association studies that are beyond the scope of this paper to address. Nearby genetic markers are linked; indeed, this is the impetus for addressing these problems using grouped penalization methods. However, genetic linkage also results in highly

correlated predictors. We have observed that the choice of  $\lambda_2$  for group bridge and group MCP has a noticeable impact on the SNPs selected. Furthermore, most genetic association studies are conducted on much larger scales than we have indicated here: moving from hundreds of SNPs to hundreds of thousands of SNPs presents a new challenge to both the computation and the assigning of group labels. The handling of missing data, the search for interactions, and the incorporation of non-genetic covariates are also important issues. In spite of these complications, the fact that markers are known to be grouped in genetic association studies is a strong motivation for the further development of bi-level selection methods.

## CHAPTER 3

### LOCAL COORDINATE DESCENT ALGORITHMS

The algorithms that have been proposed thus far to fit models with grouped penalties are either (a) inefficient for models with large numbers of predictors, or (b) limited to linear regression models, models in which the members of a group are orthogonal to each other, or both. We combine the ideas of coordinate descent optimization and local approximation of penalty functions to introduce a new, general algorithm for fitting models with grouped penalties. The resulting algorithm is stable and very fast even when the number of variables is much larger than the sample size. We apply the algorithm to models with grouped penalties, but note that the idea may be applied to other penalized regression problems in which the penalties are complicated but not necessarily grouped. These algorithms are provided as an R package, `grpreg` (available at <http://cran.r-project.org>).

#### 3.1 Local coordinate descent

The approach that we describe for minimizing  $Q(\boldsymbol{\beta})$  relies on obtaining a first-order Taylor series approximation of the penalty. This approach requires continuous differentiability. Here, we treat penalties as functions of  $|\beta|$ ; from this perspective, penalties like the lasso are continuously differentiable, with domain  $[0, \infty)$ .

Coordinate descent algorithms optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached. The idea is simple but efficient – each pass over the parameters requires only  $O(np)$  operations. Since the number of iterations is typically much smaller than  $p$ , the solution is reached faster even than the  $np^2$  operations required to solve a linear regression problem by QR decomposition. Furthermore, since the computational burden increases only linearly with  $p$ , coordinate descent algorithms can be applied to very high-dimensional problems. Only recently has the power

of coordinate descent algorithms for optimizing penalized regression problems been fully appreciated; see Friedman et al. (2007) and Wu and Lange (2008) for additional history and a more extensive treatment.

Coordinate descent algorithms are ideal for problems like the lasso where deriving the solution is simple in one dimension. The group penalties discussed in this paper do not have this feature; however, one may approximate these penalties to obtain a locally accurate representation that does. The idea of obtaining approximations to penalties in order to simplify optimization of penalized likelihoods is not new. Fan and Li (2001) propose a local quadratic approximation (LQA), while Zou and Li (2008) describe a local linear approximation (LLA). The LQA and LLA algorithms can also be used to fit these models, but as we will see in section 3.9.1, the LCD algorithm is much more efficient.

Letting  $\tilde{\boldsymbol{\beta}}$  represent the current estimate of  $\boldsymbol{\beta}$ , the overall structure of the local group coordinate descent (LCD) algorithm is as follows:

- (1) Choose an initial estimate  $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(0)}$
- (2) Approximate loss function, if necessary
- (3) Update covariates:
  - (a) Update  $\tilde{\beta}_0$ , if necessary
  - (b) For  $j \in \{1, \dots, J\}$ , update  $\tilde{\boldsymbol{\beta}}_j$
- (4) Repeat steps 2 and 3 until convergence

First, let us consider the updating of the intercept in step (3)(a). For squared error loss, this step is unnecessary since  $\hat{\beta}_0$  will always equal the mean of  $\mathbf{y}$ . Nevertheless, going through the procedure is a helpful introduction to the coordinate descent idea. The partial residual for updating  $\tilde{\beta}_0$  is  $\tilde{\mathbf{r}}_0 = \mathbf{y} - \mathbf{X}_{-0}\tilde{\boldsymbol{\beta}}_{-0}$ , where the  $-0$  subscript refers to what remains of  $\mathbf{X}$  or  $\tilde{\boldsymbol{\beta}}$  after the  $0^{\text{th}}$  column or element has

been removed, respectively. The updated value of  $\tilde{\beta}_0$  is therefore the simple linear regression solution:

$$\tilde{\beta}_0 \leftarrow \frac{\mathbf{x}'_0 \tilde{\mathbf{r}}_0}{\mathbf{x}'_0 \mathbf{x}_0} = \frac{1}{n} \mathbf{x}'_0 \tilde{\mathbf{r}}_0.$$

Performing the above calculation directly is somewhat wasteful, however; a more efficient way of updating  $\tilde{\beta}_0$  is to take advantage of the current residuals  $\tilde{\mathbf{r}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$  (Friedman et al., 2008). Here, we note that  $\tilde{\mathbf{r}}_0 = \tilde{\mathbf{r}} + \mathbf{x}_0 \tilde{\beta}_0$ ; thus

$$\tilde{\beta}_0 \leftarrow \frac{1}{n} \mathbf{x}'_0 \tilde{\mathbf{r}} + \tilde{\beta}_0. \quad (3.1)$$

Updating  $\tilde{\beta}_0$  in this way costs only  $2n$  operations:  $n$  operations to calculate  $\mathbf{x}'_0 \tilde{\mathbf{r}}$  and  $n$  operations to update  $\tilde{\mathbf{r}}$ . In contrast, obtaining  $\tilde{\mathbf{r}}_0$  requires  $n(p-1)$  operations. Meanwhile, for iteratively reweighted optimization, the updating step is

$$\tilde{\beta}_0 \leftarrow \mathbf{x}'_0 \mathbf{W} \tilde{\mathbf{r}} / \mathbf{x}'_0 \mathbf{W} \mathbf{x}_0 + \tilde{\beta}_0, \quad (3.2)$$

requiring  $3n$  operations.

Updating  $\tilde{\boldsymbol{\beta}}_j$  in step (3)(b) depends on the penalty. We discuss the updating step separately for group MCP, group bridge, and group lasso.

### 3.2 Group MCP

Group MCP has the most straightforward updating step. We begin by reviewing the univariate solution to the lasso. When the penalty being applied to a single parameter is  $\lambda|\beta|$ , the solution to the lasso (Tibshirani, 1996) is

$$\beta = \frac{S(\frac{1}{n} \mathbf{x}' \mathbf{y}, \lambda)}{\frac{1}{n} \mathbf{x}' \mathbf{x}} = S(\frac{1}{n} \mathbf{x}' \mathbf{y}, \lambda),$$

where  $S(z, c)$  is the soft-thresholding operator (Donoho and Johnstone, 1994) defined for positive  $c$  by

$$S(z, c) = \begin{cases} z - c & \text{if } z > c \\ 0 & \text{if } |z| \leq c \\ z + c & \text{if } z < -c. \end{cases}$$

Group MCP does not have a similarly convenient form for updating individual parameters. However, by taking the first order Taylor series approximation about  $\tilde{\boldsymbol{\beta}}_j$ , the penalty as a function of  $\beta_{jk}$  is approximately proportional to  $\tilde{\lambda}_{jk}|\beta_{jk}|$ , where

$$\tilde{\lambda}_{jk} = f'_{\lambda,b} \left( \sum_{m=1}^{K_j} f_{\lambda,a}(|\tilde{\beta}_{jm}|) \right) f'_{\lambda,a}(|\tilde{\beta}_{jk}|) \quad (3.3)$$

and  $f, f'$  were defined in equation (2.5). Thus, in the local region where the penalty is well-approximated by a linear function, step (3)(b) consists of simple updating steps based on the soft-thresholding cutoff  $\tilde{\lambda}_{jk}$ : for  $k \in \{1, \dots, K_j\}$ ,

$$\tilde{\beta}_{jk} \leftarrow S \left( \frac{1}{n} \mathbf{x}'_{jk} \tilde{\mathbf{r}} + \tilde{\beta}_{jk}, \tilde{\lambda}_{jk} \right) \quad (3.4)$$

or, when weights are present,

$$\tilde{\beta}_{jk} \leftarrow \frac{S(\frac{1}{n} \mathbf{x}'_{jk} \mathbf{W} \tilde{\mathbf{r}} + \frac{1}{n} \mathbf{x}'_{jk} \mathbf{W} \mathbf{x}_{jk} \tilde{\beta}_{jk}, \tilde{\lambda}_{jk})}{\frac{1}{n} \mathbf{x}'_{jk} \mathbf{W} \mathbf{x}_{jk}}. \quad (3.5)$$

### 3.3 Group bridge

The local coordinate descent algorithm for group bridge is rather similar to that for group MCP, only with

$$\tilde{\lambda}_{jk} = \lambda \gamma K_j^\gamma \|\tilde{\boldsymbol{\beta}}_j\|_1^{\gamma-1}. \quad (3.6)$$

The difficulty posed by group bridge is that, because the bridge penalty is not everywhere differentiable,  $\tilde{\lambda}_{jk}$  is undefined at  $\tilde{\boldsymbol{\beta}}_j = \mathbf{0}$  for  $\gamma < 1$ . This is not a problem with the algorithm;  $\mathbf{0}$  presents a fundamental issue with the penalty itself. For any positive value of  $\lambda$ ,  $\mathbf{0}$  is a local minimum of the group bridge penalty. Clearly,

this complicates optimization. Our approach is to begin with an initial value away from  $\mathbf{0}$  and, if  $\tilde{\boldsymbol{\beta}}_j$  reaches  $\mathbf{0}$  at any point during the iteration, to restrain  $\tilde{\boldsymbol{\beta}}_j$  at  $\mathbf{0}$  thereafter. Obviously, this incurs the potential drawback of dropping groups that would prove to be nonzero when the solution converges. Essentially, this approach screens groups from further consideration if they contain no members that show significant correlation with the outcome given the current model parameters.

### 3.4 Group lasso

Updating is more complicated in the group lasso because of its sparsity properties: group members go to 0 all at once or not at all. Thus, we must update  $\tilde{\boldsymbol{\beta}}_j$  at step (3)(b) in two steps: first, check whether  $\tilde{\boldsymbol{\beta}}_j = \mathbf{0}$  and second, if  $\tilde{\boldsymbol{\beta}}_j \neq \mathbf{0}$ , update  $\tilde{\beta}_{jk}$  for  $k \in \{1, \dots, K_j\}$ .

The first step is performed by noting that  $\tilde{\boldsymbol{\beta}}_j \neq \mathbf{0}$  if and only if

$$\frac{1}{n} \|\mathbf{X}'_j \tilde{\mathbf{r}} + \mathbf{X}'_j \mathbf{X}_j \tilde{\boldsymbol{\beta}}_j\| > \sqrt{K_j} \lambda. \quad (3.7)$$

The logic behind this condition is that if  $\boldsymbol{\beta}_j$  cannot move in any direction away from  $\mathbf{0}$  without increasing the penalty more than the movement improves the fit, then  $\mathbf{0}$  is a local minimum; since the group lasso penalty is convex,  $\mathbf{0}$  is also the unique global minimum. The conditions defined by (3.7) are in fact the Karush-Kuhn-Tucker conditions for this problem, and were first pointed out by Yuan and Lin (2006).

If this condition does not hold, then we can set  $\tilde{\boldsymbol{\beta}}_j = \mathbf{0}$  and move on. Otherwise, we once again make a local approximation to the penalty and update the members of group  $j$ . However, instead of approximating the penalty as a function of  $|\beta_{jk}|$ , for group lasso we can obtain a better approximation by considering the penalty as a function of  $\beta_{jk}^2$ . Now, the penalty applied to  $\beta_{jk}$  may be approximated

by  $\tilde{\lambda}_{jk}\beta_{jk}^2/2$ , where

$$\tilde{\lambda}_{jk} = \frac{\lambda\sqrt{K_j}}{\|\tilde{\boldsymbol{\beta}}_j\|}. \quad (3.8)$$

This approach yields a shrinkage updating step instead of a soft-thresholding step:

$$\tilde{\beta}_{jk} \leftarrow \frac{\frac{1}{n}\mathbf{x}'_{jk}\tilde{\mathbf{r}} + \tilde{\beta}_{jk}}{1 + \tilde{\lambda}_{jk}} \quad (3.9)$$

or, for weighted optimization,

$$\tilde{\beta}_{jk} \leftarrow \frac{\frac{1}{n}\mathbf{x}'_{jk}\mathbf{W}\tilde{\mathbf{r}} + \tilde{\beta}_{jk}}{\frac{1}{n}\mathbf{x}'_{jk}\mathbf{W}\mathbf{x}_{jk} + \tilde{\lambda}_{jk}}. \quad (3.10)$$

Note that, like (3.6), (3.8) is undefined at  $\mathbf{0}$ . Unlike group bridge, however, this is merely a minor algorithmic inconvenience. The penalty is differentiable; its partial derivatives simply have a different form at  $\mathbf{0}$ . This issue can be avoided by adding a small positive quantity  $\delta$  to the denominator in equation (3.8).

### 3.5 Convergence of the LCD algorithm

Let  $\boldsymbol{\beta}^{(m)}$  denote the value of the coefficients at a given step of the algorithm, and let  $\boldsymbol{\beta}^{(m+1)}$  be the value after the next updating step has occurred. With the exception of the sparsity check during the first stage of the group lasso algorithm,  $\boldsymbol{\beta}^{(m+1)}$  and  $\boldsymbol{\beta}^{(m)}$  will differ by, at most, one element.

We now prove that the proposed algorithms for squared error loss decrease the objective function with every step. For other loss functions, making a quadratic approximation to the loss function will not in general decrease the objective function. However, it is still the case that with every step that updates  $\boldsymbol{\beta}$ , the approximated objective function will be decreased.

**Proposition 1.** *At every step of the algorithms described in sections 3.2-3.4,*

$$Q(\boldsymbol{\beta}^{(m+1)}) \leq Q(\boldsymbol{\beta}^{(m)}) \quad (3.11)$$

*Thus, all three algorithms decrease the objective function at every step and therefore*



are guaranteed to converge.

*Proof.* This result follows from the general theory of MM (majorization-minimization) algorithms (Lange et al., 2000). A function  $h$  is said to majorize a function  $g$  if  $h(x) \geq g(x) \forall x$  and there exists a point  $x^*$  such that  $h(x^*) = g(x^*)$ .

Then, at a given updating step  $i$ , let  $g$  denote the objective function,  $h^{(i)}$  denote the approximation being made at the current step,  $\beta^{(i)}$  the current value of  $\beta$ , and  $\beta^{(i+1)}$  the value that minimizes  $h^{(i)}$ ,

$$\begin{aligned} g(\beta^{(i+1)}) &\leq h^{(i)}(\beta^{(i+1)}) && (h^{(i)} \text{ majorizes } g) \\ &\leq h^{(i)}(\beta^{(i)}) && (\beta^{(i+1)} \text{ minimizes } h^{(i)}) \\ &= g(\beta^{(i)}) && (\text{expansion is made about } \beta^{(i)}) \end{aligned}$$

Because the loss function is unchanged between  $g$  and  $h$ , all that remains to prove the theorem is to show that the approximations referred to by (3.3), (3.6), and (3.8) majorize their respective penalty functions. This is straightforward for group bridge and group MCP, as both penalties are concave on  $[0, \infty)$ . They are therefore majorized by any tangent line. For group lasso, we can demonstrate majorization through inspection of second derivatives by observing that  $\{h(\beta_{jk}) - g(\beta_{jk})\}'' \geq 0$  for all  $\beta_{jk} \in (0, \infty)$ .  $\square$

The LCD algorithm is therefore stable and guaranteed to converge, although not necessarily to the global minimum of the objective function. The group bridge and group MCP penalty functions are nonconvex; group bridge always contains local minima and group MCP may have them as well. Furthermore, coordinate descent algorithms for penalized squared error loss functions are guaranteed to converge to minima only when the penalties are separable. Group penalties are separable between groups, but not within them. Convergence to a minimum cannot be guaranteed, then, for the one-at-a-time updates that we propose here. Nevertheless, we have not observed this to be a significant problem in practice. Comparing the

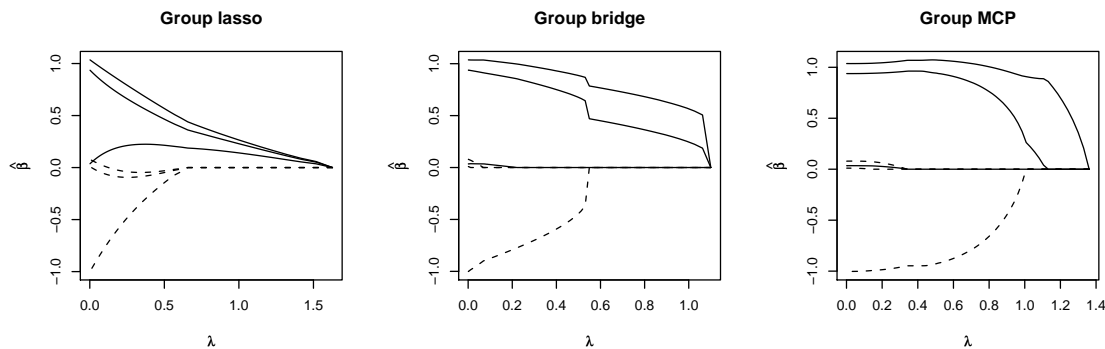


Figure 3.1: Coefficient paths for group penalization methods. Paths for group lasso, group bridge, and group MCP are illustrated, with  $\lambda$  varying from 0 to  $\lambda_{\max}$ . The simulated data set features two groups, each with three covariates. In the underlying model, the solid group has two covariates equal to 1 and the other equal to 0; the dashed group has two coefficients equal to 0 and the other equal to -1.

convergence of the LCD algorithms to LQA/LLA algorithms (which update all parameters simultaneously) for the same data, the algorithms rarely converge to different values, and when they do, the differences are quite small.

### 3.6 Pathwise optimization and initial values

The local coordinate descent algorithm requires an initial value  $\beta^{(0)}$ . Usually, we are interested in obtaining  $\hat{\beta}$  not just for a single value of  $\lambda$ , but for a range of values and then applying some criterion to choose an optimal  $\lambda$ .

Usually, the range of  $\lambda$  values one is interested in extends from a maximum value  $\lambda_{\max}$  for which all penalized coefficients are 0 down to  $\lambda = 0$  or to a minimum value  $\lambda_{\min}$  at which the model becomes excessively large or ceases to be identifiable. The estimated coefficients vary continuously with  $\lambda$  and produce a path of solutions regularized by  $\lambda$ . Example coefficient paths for group lasso, group bridge, and group MCP over a fine grid of  $\lambda$  values are presented in Figure 3.1; inspecting the path of solutions produced by a penalized regression method is often a very good way to gain insight into the methodology.

Figure 3.1 depicts a toy example, yet reveals much about the behavior of

grouped penalties. In the example, there are two groups, each of which containing three members. In the solid group, two of the members have nonzero coefficients, while in the dashed group, member has a nonzero coefficient. Even though each of the nonzero coefficients is of the same magnitude, the coefficients from the solid group enter the model much more easily than the lone nonzero coefficient from the dashed group. Note also, however, that this assumption is less pronounced for group MCP. Finally, notice the extent to which solutions are shrunken toward zero. The effect is quite strong for group lasso, much less so for group MCP, and in between for group bridge. Indeed, for group MCP at  $\lambda \approx 0.4$ , all of the variables with true zero coefficients have been eliminated while the remaining coefficients are unpenalized. In this region, the group MCP approach is equivalent to the oracle model.

Because the paths are continuous, a reasonable approach to choosing initial values is to start at one extreme of the path and use the estimate  $\hat{\boldsymbol{\beta}}$  from the previous value of  $\lambda$  as the initial value for the next value of  $\lambda$ .

For group MCP and group lasso (and in general for any penalty function that is differentiable at  $\mathbf{0}$ ), we can easily determine  $\lambda_{\max}$ , the smallest value for which all penalized coefficients are 0. From (3.7), it is clear that

$$\lambda_{\max} = \max_j \frac{\|\mathbf{X}'_j \mathbf{W} \tilde{\mathbf{r}}\|}{n\sqrt{K_j}},$$

where the current residuals and weights are obtained using a regression fit to the intercept-only model. For group MCP,

$$\lambda_{\max} = \max_{j,k} \sqrt{\frac{|\mathbf{x}'_{jk} \mathbf{W} \tilde{\mathbf{r}}|}{n}}.$$

For these methods, we can start at  $\lambda_{\max}$  using  $\boldsymbol{\beta}^{(0)} = \mathbf{0}$  and proceed towards  $\lambda_{\min}$ .

This approach does not work for group bridge, however, because  $\tilde{\boldsymbol{\beta}}$  must be initialized away from 0. We must therefore start at  $\lambda_{\min}$  and proceed toward  $\lambda_{\max}$  (*i.e.*, work in the opposite direction as group MCP and group lasso). For the initial

value at  $\lambda_{\min}$ , we suggest using the unpenalized univariate regression coefficients.

For all the numerical results in this paper, we follow the approach of Friedman et al. (2008) and compute solutions along a grid of 100  $\lambda$  values that are equally spaced on the log scale.

### 3.7 Regularization parameter selection

Once a regularization path has been fit, we are typically interested in selecting an optimal point along the path. Three widely used criteria are:

$$AIC(\lambda) = 2L_\lambda + 2df_\lambda, \quad (3.12)$$

$$BIC(\lambda) = 2L_\lambda + \log(n)df_\lambda, \quad (3.13)$$

and

$$GCV(\lambda) = \frac{2L_\lambda}{[1 - (df_\lambda/n)]^2}, \quad (3.14)$$

where  $df_\lambda$  is the effective number of parameters. The optimal value of  $\lambda$  is chosen to be the one that minimizes the criterion.

We propose the following estimator for  $df_\lambda$ . Let  $\hat{\beta}_{jk}$  denote the fitted value of  $\beta_{jk}$  and  $\hat{\beta}_{jk}^*$  denote the unpenalized fit to the partial residual:  $\hat{\beta}_{jk}^* = \mathbf{x}'_{jk}\tilde{\mathbf{r}}_{jk}/n$ . Then

$$\hat{df}_\lambda = \sum_{j=1}^J \sum_{k=1}^{K_j} \frac{\hat{\beta}_{jk}}{\hat{\beta}_{jk}^*}. \quad (3.15)$$

This estimator is attractive for a number of reasons. For linear fitting methods such that  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ , there are several justifications for choosing  $\hat{df} = \text{trace}(\mathbf{S})$  (Hastie et al., 2001). Ridge regression is an example of a linear fitting method in which  $\mathbf{S} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'$ . For the special case of an orthonormal design, (3.15) is equal to the trace of  $\mathbf{S}$ . The estimator also has an intuitive justification, in that it makes a smooth transition from an unpenalized coefficient with  $df = 1$  to a coefficient that has been eliminated with  $df = 0$ . Another attractive feature is

convenience: the estimator is obtained as a byproduct of the coordinate descent algorithm with no additional calculation.

Yuan and Lin (2006) propose an estimator for the effective number of parameters of the group lasso, but it involves the ordinary least squares estimator, which is undefined in high dimensions, so we do not consider it here. Another common approach is to set  $\hat{df}$  equal to the number of nonzero elements of  $\hat{\beta}$  (Efron et al., 2004; Zou et al., 2007). However, this has two drawbacks. One is that the estimator (and, hence, the model selection criterion) is not a continuous function of  $\lambda$ . The other is that this approach is inappropriate for methods that perform a heavy amount of coefficient shrinkage like the group lasso. We examine the performance of this estimator and estimator (3.15) using simulation studies in section 3.9.2.

### 3.8 Adding an $L_2$ penalty

Zou and Hastie (2005) have suggested that incorporating an additional, small  $L_2$  penalty can improve the performance of penalized regression methods such as the lasso, especially when the number of predictors is larger than the number of observations or when large correlation exists between the predictors. This does not pose a complication to the above algorithms. When minimizing the previously defined objective functions plus  $\lambda_2 \sum_{j,k} \beta_{jk}^2/2$ , the updating step (3.4) becomes

$$\tilde{\beta}_{jk} \leftarrow \frac{S(\frac{1}{n} \mathbf{x}'_{jk} \tilde{\mathbf{r}} + \tilde{\beta}_{jk}, \tilde{\lambda}_{jk})}{1 + \lambda_2}$$

for group MCP and group bridge and the updating step (3.9) becomes

$$\tilde{\beta}_{jk} \leftarrow \frac{\frac{1}{n} \mathbf{x}'_{jk} \tilde{\mathbf{r}} + \tilde{\beta}_{jk}}{1 + \tilde{\lambda}_{jk} + \lambda_2}$$

for group lasso. We use  $\lambda_2 = .001\lambda$  for the numerical results presented in the remainder of this chapter, as well as for the results in Chapter 2.

### 3.9 Simulations

#### 3.9.1 Efficiency

We will examine the efficiency of the LCD algorithm by measuring the average time to fit the entire path of solutions for group lasso, group bridge, and group MCP, as well as the lasso as a benchmark. Besides LCD, we consider the following algorithms: `lars` (Efron et al., 2004), the most widely used algorithm for fitting lasso paths as of this writing; `glmnet` (Friedman et al., 2008), a very efficient coordinate descent algorithm for computing lasso paths; `glmpath` (Park and Hastie, 2007), an approach to fitting lasso paths for GLMs not based on coordinate descent; and the LQA (Fan and Li, 2001) and LLA (Zou and Li, 2008) algorithms mentioned in section 3.1.

We will consider three situations:

- Linear regression with  $n = 500$ ,  $p = 200$
- Logistic regression with  $n = 1000$ ,  $p = 200$
- Linear regression with  $n = 500$ ,  $p = 2000$

For the data sets with  $n > p$ , paths were computed down to  $\lambda = 0$ ; for the  $p > n$  data sets, paths were computed down to 5% of  $\lambda_{\max}$ .

The results of these efficiency trials are presented in Tables 3.1, 3.2, and 3.3. All entries are the average time in number of seconds, averaged over 100 randomly generated data sets.

These timings dramatically verify the efficiency of coordinate descent algorithms for high-dimensional penalized regression. The LCD algorithm is not only much faster than LLA/LQA for small  $p$ , its computational burden increases in a manner that is roughly linear with  $p$  as opposed to the polynomial increase suffered by LLA/LQA. Indeed, the LCD algorithms are, for large  $p$ , even faster than

Table 3.1: LCD algorithm efficiency: Linear regression,  $n = 500$  and  $p = 200$ .

| Penalty      | Algorithm           | Average Time (s) |
|--------------|---------------------|------------------|
| Lasso        | <code>glmnet</code> | .03              |
| Lasso        | <code>lars</code>   | .43              |
| Group lasso  | LQA                 | 3.54             |
| Group bridge | LLA                 | 7.02             |
| Group MCP    | LLA                 | 5.13             |
| Group lasso  | LCD                 | .63              |
| Group bridge | LCD                 | .11              |
| Group MCP    | LCD                 | .10              |

Table 3.2: LCD algorithm efficiency: Logistic regression,  $n = 1000$  and  $p = 200$ .

| Penalty      | Algorithm           | Average Time (s) |
|--------------|---------------------|------------------|
| Lasso        | <code>glmnet</code> | 0.24             |
| Lasso        | <code>glmnet</code> | 13.77            |
| Group lasso  | LQA                 | 21.78            |
| Group bridge | LLA                 | 29.77            |
| Group MCP    | LLA                 | 15.08            |
| Group lasso  | LCD                 | 1.80             |
| Group bridge | LCD                 | 0.67             |
| Group MCP    | LCD                 | 0.47             |

Table 3.3: LCD algorithm efficiency: Linear regression,  $n = 500$  and  $p = 2000$ .

| Penalty      | Algorithm           | Average Time (s) |
|--------------|---------------------|------------------|
| Lasso        | <code>glmnet</code> | 1.60             |
| Lasso        | <code>lars</code>   | 22.69            |
| Group lasso  | LQA                 | 1900.49*         |
| Group bridge | LLA                 | 1985.19*         |
| Group MCP    | LLA                 | 1823.32*         |
| Group lasso  | LCD                 | 23.00            |
| Group bridge | LCD                 | 1.46             |
| Group MCP    | LCD                 | 3.47             |

\*Only one replication

the LARS algorithm, a somewhat remarkable fact considering that the latter takes explicit advantage of special piecewise linearity properties of linear regression lasso paths.

Among the grouped penalties, group lasso is the slowest due to its two-step updating procedure. Group bridge was timed here to be the fastest, although this is potentially misleading. Group bridge saves time by not updating groups that reach  $\mathbf{0}$  with no guarantee of converging to the true minimum. This is a weakness of the method, not a strength, although it does result in shorter computing times.

### 3.9.2 Regularization parameter selection

In this section, we will conduct a simulation study to compare the performance of our proposed estimator of the number of effective model parameters versus using the number of nonzero covariates as an estimator. As in section 2.4, we will look at simulations for penalized linear regression using BIC as the model selection criterion;



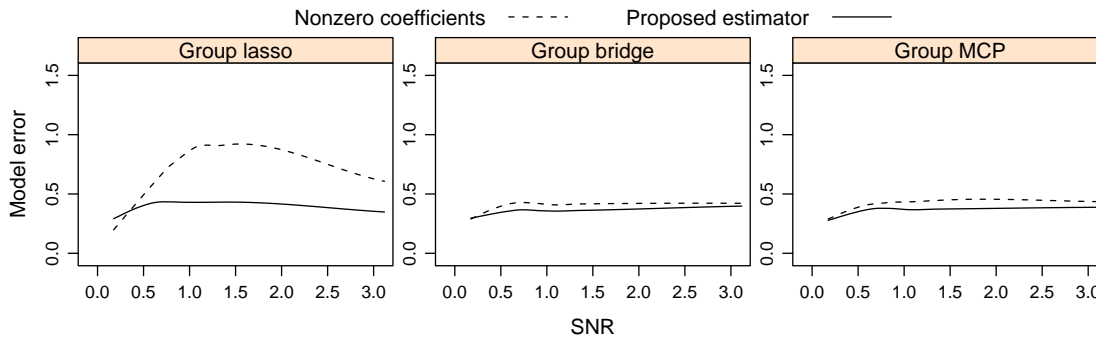


Figure 3.2: Simulation: Comparison of degree of freedom estimators for group penalization methods. The model error for each method is plotted, after selecting  $\lambda$  with BIC using one of two estimators for the effective number of model parameters. Dashed line: Estimator (3.15). Solid line: Using number of nonzero elements of  $\beta$ .

simulations for logistic regression and using AIC and GCV illustrate the same trend.

As before, data were simulated from model (2.8). Here,  $J_0 = 3$  and the elements of  $\beta_1$  through  $\beta_3$  were randomly generated in such a way as to have the models span SNR ratios over the range  $(0.5, 3)$  in a roughly uniform manner. Data sets were generated independently 500 times. Model error was chosen as the outcome; lowess curves were fit to the results and plotted in Figure 3.2.

As Figure 3.2 illustrates, the performance of estimator (3.15) is similar to (although slightly better than) that of counting the nonzero elements of  $\beta$  for group bridge and group MCP, but much better for the more ridge-like penalty group lasso. We consider this sufficient justification for the use of (3.15) throughout the remainder of this article; however, further study of this approach to estimating model degrees of freedom is warranted.

## CHAPTER 4

### FALSE DISCOVERY RATES FOR PENALIZED REGRESSION

In this chapter, I consider high-dimensional regression problems in which the goal is to select a small number of covariates that contribute to the outcome from a large number of potential features. In contrast to existing approaches, the methods introduced here select variables while limiting the expected fraction of falsely selected features.

Feature selection is an important issue in high-dimensional data analysis. Many contemporary studies collect information on a large number of features that are potentially related to an outcome of interest with the expectation that only a small number of those features will exhibit meaningful effects. Examples include gene expression and proteomics studies, genetic association studies, signal processing, image analysis, and financial applications (Donoho, 2000; Fan and Li, 2006).

Many of the successful approaches that have been proposed for prediction and feature selection for high-dimensional data analysis fall into the general framework of penalized regression models. These approaches seek to minimize an objective function consisting of a loss function plus a penalty term. The loss function characterizes the accuracy of the model's fit to the data, while the penalty term encourages both shrinkage and sparsity of the solution by penalizing large regression coefficients. The balance between the loss function and penalty is controlled by a regularization parameter. A well-chosen regularization parameter allows the model to explain the data while limiting the amount of overfitting.

Despite the success of penalized regression methods, their practical utility for feature selection has been limited by a lack of inferential results for these models. Regularization parameters for these models are usually selected on the basis of information criteria or cross-validation. These approaches may yield accurate prediction

methods, but they provide little information regarding the significance of the selected features. For univariate hypothesis testing, in comparison, the false discovery rate (Benjamini and Hochberg, 1995) has proved to be an extremely valuable and intuitive measure of feature significance in light of large-scale multiple comparison.

This chapter applies the idea of false discovery rates to penalized regression models. In doing so, we obtain an easily interpreted measure of the significance of the features selected by these models. The problem is defined explicitly in section 4.1. Exact distributional results for penalized regression methods are difficult to obtain, and the methods introduced here are approximations to the true false discovery rate. We consider two approaches to this approximation. The first is based a heuristic line of reasoning (section 4.2), while the second is based on approximating the penalized regression model with a linear predictor (section 4.3). Section 4.4 examines the accuracy of the resulting estimators and compares the methodology to competing approaches for feature selection, while section 4.5 applies this approach to a gene expression study of leukemia and a genetic association study of age-related macular degeneration. We will concentrate primarily on the simplest and most popular of these models, the lasso, but the idea may be applied to other penalized regression methods as well.

#### 4.1 The false discovery rate of the lasso

The following notation will be used throughout: suppose we have  $n$  observations indexed by  $i$ . Each observation contains measurements of an outcome  $y_i$  and  $p$  features  $\{x_{i1}, \dots, x_{ip}\}$  indexed by  $j$ . We assume without loss of generality that the features have been standardized such that  $\sum_{i=1}^n x_{ij} = 0$  and  $\frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1$  for all  $j$ .

We will consider penalized linear and logistic regression. In both cases, the mean of the outcome is assumed to depend on the covariates through the linear function  $\eta_i = \beta_0 + \sum_j x_{ij}\beta_j$ . In this context, we define a feature  $\mathbf{x}_j$  to be a *null*

Table 4.1: Schematic: Possible outcomes of feature selection.

|                | Selected | Not selected | Total |
|----------------|----------|--------------|-------|
| $\beta = 0$    | $F$      | $p_0 - F$    | $p_0$ |
| $\beta \neq 0$ | $T$      | $p_1 - T$    | $p_1$ |
| Total          | $S$      | $p - S$      | $p$   |

feature if  $\beta_j = 0$ . If a null feature is selected by a given procedure, it becomes a *false discovery*. Table 4.1 lists the outcomes of a feature selection procedure:  $p_0$  denotes the number of features whose coefficients are truly 0, and  $p_1$  denotes the the number of features whose coefficients are not equal to 0. A given procedure selects  $S$  features, of which  $T$  are correctly chosen and  $F$  are false discoveries.

The quantity  $F/S$  is of interest, as it is the proportion of selected features which are false discoveries. However,  $F$  is an unknown quantity. A widely used approach in univariate hypothesis testing is to estimate the expected number of false discoveries  $E(F)$  and then estimate the *false discovery rate* (FDR) by

$$\widehat{\text{FDR}} = \frac{E(F)}{S}. \quad (4.1)$$

This is the general form that our FDR estimates will take.

Our goal is to select features which contribute to the outcome while limiting the false discovery rate. Even though penalized regression methods such as the lasso are not conventionally thought of as performing multiple comparisons, we will see that for each feature, a statistic derived from that feature is compared to a threshold. The central idea behind our FDR approach is to estimate the probability that a null feature will exceed that threshold. Letting  $\alpha_j$  denote the probability that feature  $j$  will exceed this threshold given that  $\beta_j = 0$ , we will estimate the expected number

of false discoveries with

$$E(F) = \sum_{j=1}^p \alpha_j. \quad (4.2)$$

Note that this is an overestimate of  $E(F)$  in that the sum should be taken only over the null features rather than all features. However, since we do not know which of the features are null, a conservative approach is to treat all features as null. As established by Benjamini and Hochberg (1995), this approach guarantees that the resulting FDR estimate will be greater than or equal to the true FDR under all configurations of null vs. nonnull features.

For univariate hypothesis testing, the probability that a feature will be selected given that the null hypothesis is true is prespecified as  $\alpha$  and thus  $E(F)$  is simply  $p\alpha$ . For the lasso, the features are not necessarily tested at the same significance level, and thus feature-specific subscripts are required in equation (4.2).

#### 4.1.1 Linear regression

The lasso (Tibshirani, 1996) is a popular regression-based approach to feature selection for sparse problems. For linear regression, the lasso estimate  $\hat{\boldsymbol{\beta}}$  is defined as the value of the coefficient vector that minimizes the objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^n (y_i - \eta_i)^2 + \lambda \sum_{j=1}^p |\beta_j|. \quad (4.3)$$

Note that we have eliminated the intercept; for linear regression with standardized covariates, the intercept will equal the mean of the outcome. Without loss of generality, then, we can consider the outcome to have been centered prior to fitting and ignore the intercept. Typically, this minimization is performed over a range of values for  $\lambda$  and some criterion is used to choose an optimal value.

Algorithmic approaches to fitting lasso models have advanced drastically since Tibshirani's original proposal (Efron et al., 2004; Friedman et al., 2007; Wu and

Lange, 2008; Friedman et al., 2008), making it feasible to fit lasso models to sparse problems involving hundreds of thousands of variables. Because of this, the lasso is now commonly used for problems in which the primary goal is to select a small number of important variables from a large pool of potential predictors. It is desirable in these problems to obtain a measure of the accuracy of this feature selection, which is what we propose here with false discovery rates.

Let  $\mathbf{r}$  denote the  $n$ -dimensional residual vector with elements  $y_i - \eta_i$ . Taking the partial derivatives of (4.3), we can see that

$$\frac{1}{n} \mathbf{x}'_j \mathbf{r} = \lambda \text{sign}(\hat{\beta}_j) \quad \forall \hat{\beta}_j \neq 0, \quad (4.4a)$$

$$\frac{1}{n} |\mathbf{x}'_j \mathbf{r}| \leq \lambda \quad \forall \hat{\beta}_j = 0. \quad (4.4b)$$

For convex objective functions such as that of the lasso, these conditions are both necessary and sufficient for any solution  $\hat{\boldsymbol{\beta}}$ . In the convex optimization literature, they are known as the Karush-Kuhn-Tucker (KKT) conditions.

Introducing the notation  $-j$  to refer to the portion of  $\mathbf{X}$  or  $\hat{\boldsymbol{\beta}}$  that remains after the  $j^{\text{th}}$  column or element has been removed and defining  $\mathbf{r}_{-j} = \mathbf{y} - \mathbf{X}_{-j} \hat{\boldsymbol{\beta}}_{-j}$ , the above conditions imply that

$$\frac{1}{n} |\mathbf{x}'_j \mathbf{r}_{-j}| > \lambda \quad \forall \hat{\beta}_j \neq 0, \quad (4.5a)$$

$$\frac{1}{n} |\mathbf{x}'_j \mathbf{r}_{-j}| \leq \lambda \quad \forall \hat{\beta}_j = 0. \quad (4.5b)$$

Thus, lasso feature selection is based on a series of multiple comparisons involving the correlation of feature  $j$  with its partial residual vector  $\mathbf{r}_{-j}$ , and

$$\begin{aligned} \alpha_j &= \Pr(\hat{\beta}_j \neq 0 | \beta_j = 0) \\ &= \Pr\left(\frac{1}{n} |\mathbf{x}'_j \mathbf{r}_{-j}| > \lambda | \beta_j = 0\right) \end{aligned} \quad (4.6)$$

In regression problems, the covariates are generally treated as fixed. Thus, the probability that  $n^{-1} |\mathbf{x}'_j \mathbf{r}_{-j}| > \lambda$  given that  $\beta_j = 0$  is determined by the distribution

of  $\mathbf{r}_{-j}$ . In general,  $\mathbf{r}_{-j}$  has a rather complicated distribution. Its elements are not independent, homoskedastic, or normally distributed, which renders derivation of the exact probability in equation (4.6) difficult. Sections 4.2 and 4.3 present two approaches to approximating the above probability.

#### 4.1.2 Logistic regression

The false discovery conditions are quite similar for lasso-penalized logistic regression. Here, the covariates are assumed to have a linear relationship on the log-odds of a binary response occurring. Specifically,

$$\begin{aligned}\pi_i &= \Pr(y_i = 1 | \eta_i) \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.\end{aligned}$$

The objective function of this model is then

$$Q(\beta_0, \boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^n \{y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)\} + \lambda \sum_{j=1}^p |\beta_j|.$$

This model leads to the same conditions as linear regression:

$$\frac{1}{n} \mathbf{x}'_j \mathbf{r} = \lambda \text{sign}(\hat{\beta}_j) \quad \forall \hat{\beta}_j \neq 0, \quad (4.7a)$$

$$\frac{1}{n} |\mathbf{x}'_j \mathbf{r}| \leq \lambda \quad \forall \hat{\beta}_j = 0. \quad (4.7b)$$

and

$$\frac{1}{n} |\mathbf{x}'_j \mathbf{r}_{-j}| > \lambda \quad \forall \hat{\beta}_j \neq 0, \quad (4.8a)$$

$$\frac{1}{n} |\mathbf{x}'_j \mathbf{r}_{-j}| \leq \lambda \quad \forall \hat{\beta}_j = 0, \quad (4.8b)$$

where  $\mathbf{r} = \mathbf{y} - \boldsymbol{\pi}$  and the  $-j$  subscript again refers to quantities calculated by leaving the contribution from the  $j^{\text{th}}$  coefficient out of the model. Once again, sections 4.2 and 4.3 present two approaches to approximating the resulting probability.

### 4.1.3 Other penalties

Note that the FDR conditions (4.5) presented above are not necessarily specific to the lasso. The form of the penalty is only present in the sense that  $\lambda$  is the correlation threshold set by the lasso. Specifically,  $\lambda$  is the derivative of the lasso penalty as  $|\beta| \rightarrow 0$ . The above conditions for the false discovery rates of penalized regression models can therefore be applied to any penalized regression method that is differentiable at 0 with respect to  $|\beta|$ . Examples of such methods include the elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), and MCP (Zhang, 2007).

Conveniently, all of the above methods are usually parameterized in such a way that  $\lambda$  (or  $\lambda_1$  for the elastic net) represents the derivative of the penalty function as  $|\beta| \rightarrow 0$ ; thus, the above equations apply equally well for those methods as for the lasso. One notable exception, however, is bridge regression (Frank and Friedman, 1993) with  $\gamma < 1$ , for which the limit of the derivative as  $\beta \rightarrow 0$  goes to infinity.

Of course, the probability of  $n^{-1}|\mathbf{x}'_j \mathbf{r}_{-j}|$  exceeding  $\lambda$  will certainly depend on the penalty. Of the two approaches to estimating  $\alpha_j$  that will be presented in sections 4.2 and 4.3, the linear predictor approach makes specific use of the form of the lasso whereas the heuristic estimator does not. Thus, the heuristic estimator could be applied with no modifications to FDR estimation for penalties such as SCAD or the elastic net, whereas the estimate based on forming a linear predictor would require further derivations specific to the penalty under investigation.



## 4.2 Heuristic approach

In this section, I present a simple estimator of  $\alpha_j$  based on the observation that residuals in regression problems tend to follow an approximately normal distribution. Thus, the following approximation may prove reasonable:

$$\mathbf{r}_{-j} \stackrel{\text{approx}}{\sim} N(\mathbf{0}, \tau_j^2 \mathbf{I}),$$

with the variance  $\tau_j^2$  to be estimated from the data.

In unpenalized regression, the expectation of  $\mathbf{r}_{-j}$  is exactly 0 when  $\beta_j = 0$ . However, in penalized regression,

$$\begin{aligned} E(\mathbf{r}_{-j}) &= E(\mathbf{y} - \mathbf{X}_{-j} \hat{\boldsymbol{\beta}}_{-j}) \\ &= \mathbf{X}_{-j} \{\boldsymbol{\beta}_{-j} - E(\hat{\boldsymbol{\beta}}_{-j})\} \end{aligned}$$

when  $\beta_j = 0$ . Thus,  $\mathbf{r}_{-j}$  will have nonzero mean proportional to the bias of the estimator  $\hat{\boldsymbol{\beta}}$ . We will ignore this bias, however, and assume  $E(\mathbf{r}_{-j}) = 0$  for null features. Note that the lasso estimator  $\hat{\boldsymbol{\beta}}$  is a consistent estimator of  $\boldsymbol{\beta}$ ; thus, this assumption is justified asymptotically.

A reasonable estimate of  $\tau_j^2$  given that feature  $j$  is a false discovery is the observed variance of the residuals,

$$\hat{\tau}_j^2 = \frac{\mathbf{r}'\mathbf{r}}{n}. \tag{4.9}$$

The elements of  $\mathbf{r}$  always sum to 0, so there is no need to adjust for the mean when estimating this variance. Also note that we are interested in the variance of the fitted (as opposed to the true) residual; therefore, adjusting for the degrees of freedom of the fit in the denominator is inappropriate. Finally, it is tempting to replace  $\mathbf{r}$  with  $\mathbf{r}_{-j}$  in the above equation, but this estimate fails to condition on  $j$  being a false discovery and leads to the unattractive result that nonnull features will contribute more to the FDR than null features.

Applying the approximations described above, we have

$$\begin{aligned}\alpha_j &\approx 2\Phi\left(\frac{-\sqrt{n}\lambda}{\tau_j}\right) \\ &\approx 2\Phi\left(-\frac{n\lambda}{\sqrt{\mathbf{r}'\mathbf{r}}}\right),\end{aligned}$$

where  $\Phi(\cdot)$  is the cdf of the standard normal distribution. Thus,

$$E(F) \approx 2p\Phi\left(-\frac{n\lambda}{\sqrt{\mathbf{r}'\mathbf{r}}}\right).$$

Substituting this expression into equation (4.1) yields the FDR estimate

$$\widehat{\text{FDR}} = \frac{2p}{S}\Phi\left(-\frac{n\lambda}{\sqrt{\mathbf{r}'\mathbf{r}}}\right) \quad (4.10)$$

Certainly, the above line of reasoning makes some rather strong assumptions that may not hold true in all settings. However, as we shall see in section 4.4, the above estimator performs quite well in practice.

Furthermore, in extending FDR estimation to other models and penalties, no modification of the above estimator is necessary. The assumptions are made directly upon the distribution of  $\mathbf{r}_{-j}$ . Certainly, these approximations will be more accurate in certain settings than in others. For example, the approximation that  $E(\mathbf{r}_{-j}) = 0$  for null features may be inaccurate for a penalty with heavy shrinkage (such as the elastic net with a large  $L_2$  component), and may be exact for penalties that attempt to eliminate the shrinkage of nonnull features (such as SCAD and MCP). We restrict our attention here to the lasso, but further study would also be of interest.

### 4.3 Linear predictor approach

In this section, we take a different approach to approximating the distribution of  $n^{-1}\mathbf{x}'_j\mathbf{r}_{-j}$  based on representing the lasso predictor as  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ , where  $\mathbf{H}$  is akin to the familiar “hat” matrix from linear regression. However, this approach is only approximate in the sense that  $\mathbf{H}$  depends on  $\mathbf{y}$ . Thus, treating  $\mathbf{H}$  as fixed (as

we shall do) will not yield exact results. Furthermore, for logistic regression,  $\hat{\mathbf{y}}$  can only approximately be represented as  $\mathbf{H}\mathbf{y}$ . Unlike the previous section, the estimators developed below rely on forms specific to the lasso, and therefore are not as readily adaptable to other penalties. Nevertheless, although new derivations would be required, the basic approach of forming a linear predictor from the KKT conditions can certainly be extended to methods other than the lasso.

#### 4.3.1 Linear regression

The approach presented below in deriving a linear approximation to the lasso estimates is similar to that of Osborne et al. (2000). Let  $A = \{j : \hat{\beta}_j \neq 0\}$ , and let  $\mathbf{X}_A$  and  $\hat{\boldsymbol{\beta}}_A$  denote the portions of  $\mathbf{X}$  and  $\hat{\boldsymbol{\beta}}$  that belong to the set  $A$ . Then (4.4a) implies that

$$\mathbf{X}'_A(\mathbf{y} - \mathbf{X}_A\hat{\boldsymbol{\beta}}_A) = n\lambda\mathbf{s}_A,$$

where  $\mathbf{s}_A$  is a vector with elements  $\{\text{sign}(\hat{\beta}_j) : j \in A\}$ . Thus,

$$\begin{aligned} \mathbf{X}'_A\mathbf{y} &= \mathbf{X}'_A\mathbf{X}_A\hat{\boldsymbol{\beta}}_A + n\lambda\mathbf{s}_A \\ &= \mathbf{X}'_A\mathbf{X}_A\hat{\boldsymbol{\beta}}_A + n\lambda\mathbf{s}_A\mathbf{s}'_A\|\hat{\boldsymbol{\beta}}_A\|_1^{-1}, \end{aligned}$$

because  $\mathbf{s}'_A\hat{\boldsymbol{\beta}}_A = \|\hat{\boldsymbol{\beta}}_A\|_1$ , where  $\|\cdot\|_1$  denotes the  $L1$  norm. Thus,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A &= \mathbf{M}^{-1}\mathbf{X}'_A\mathbf{y} \\ &= \mathbf{S}\mathbf{y}, \end{aligned}$$

where  $\mathbf{M} = \mathbf{X}'_A\mathbf{X}_A + n\lambda\mathbf{s}_A\mathbf{s}'_A\|\hat{\boldsymbol{\beta}}_A\|_1^{-1}$  and  $\mathbf{S} = \mathbf{M}^{-1}\mathbf{X}'_A$ .

We now have

$$\begin{aligned} \frac{1}{n}\mathbf{x}'_j\mathbf{r}_{-j} &= \frac{1}{n}\mathbf{x}'_j(\mathbf{y} - \mathbf{H}_{-j}\mathbf{y}) \\ &= \frac{1}{n}\mathbf{x}'_j(\mathbf{I} - \mathbf{H}_{-j})\mathbf{y}, \end{aligned} \tag{4.11}$$

where  $\mathbf{H}_{-j} = (\mathbf{X}_A)_{-j}\mathbf{S}_{-j}$  and  $\mathbf{I}$  is the identity matrix. Here, the  $-j$  subscript refers to the removal of the portion of  $\mathbf{X}_A$  or  $\mathbf{S}$  that corresponds to the  $j^{\text{th}}$  feature. Note that this will not usually be the  $j^{\text{th}}$  row or column of the matrix, and that if  $\hat{\beta}_j = 0$ , the  $j^{\text{th}}$  portion will not be present and  $(\mathbf{X}_A)_{-j} = \mathbf{X}_A$ .

In the above,  $\mathbf{H}_{-j}$  is akin to the familiar “hat” matrix from linear regression in the sense that  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ , but there are two important differences:  $\mathbf{H}$  is neither idempotent nor fixed. Instead, it contains the random elements  $\mathbf{s}_A$  and  $\|\hat{\boldsymbol{\beta}}_A\|_1$ . Furthermore, it is important to note that  $\mathbf{H}_{-j}$  is not symmetric. Nevertheless, if we assume, as is common in linear regression, that

$$y_i = \mathbf{x}'_i\boldsymbol{\beta} + \epsilon_i \quad (4.12)$$

with

$$\epsilon_i \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma^2), \quad (4.13)$$

then the representation contained in equation (4.11) suggests that  $n^{-1}\mathbf{x}'_j\mathbf{r}_{-j}$  follows an approximate normal distribution. Thus,

$$\alpha_j = \Phi\left(\frac{\mu_j - \lambda}{\tau_j}\right) + \Phi\left(\frac{-\mu_j - \lambda}{\tau_j}\right), \quad (4.14)$$

where  $\mu_j$  and  $\tau_j^2$  are the mean and variance of  $n^{-1}\mathbf{x}'_j\mathbf{r}_{-j}$  under the null hypothesis that  $\beta_j = 0$ .

Based on (4.11),  $\mu_j$  and  $\tau_j^2$  can be reasonably estimated by

$$\hat{\mu}_j = \frac{1}{n}\mathbf{x}'_j(\mathbf{I} - \mathbf{H}_{-j})\widehat{\mathbf{E}}_0(\mathbf{y}) \quad (4.15)$$

$$\hat{\tau}_j^2 = \frac{\hat{\sigma}^2}{n^2}\mathbf{x}'_j(\mathbf{I} - \mathbf{H}_{-j})(\mathbf{I} - \mathbf{H}_{-j})'\mathbf{x}_j, \quad (4.16)$$

where  $\hat{\sigma}^2$  is an estimate of  $\sigma^2$  and  $\widehat{\mathbf{E}}_0(\mathbf{y})$  is an estimate of the expected value of  $\mathbf{y}$  under the null hypothesis  $\beta_j = 0$ .

Together, equations (4.15), and (4.16) allow us to estimate the probability of

false selection  $\alpha_j$  needed to estimate the false discovery rate using equations (4.1) and (4.6).

Before moving on, let us consider two interesting limiting cases. The first is the case when  $x'_j x_{j'} \rightarrow 0 \forall j \neq j'$ ; *i.e.*, as the correlation between predictors tends towards 0 and the design matrix becomes orthogonal. In this case,  $x'_j \mathbf{H}_{-j} \rightarrow \mathbf{0}$ , causing  $\hat{\mu}_j \rightarrow 0$  and  $\hat{\tau}_j^2 \rightarrow \hat{\sigma}^2/n \forall j$ . Therefore, for the case of orthogonal design, the above false discovery rate calculations are exact.

Furthermore, note that, as  $\lambda \rightarrow 0$ , the random contributions to  $\mathbf{H}_{-j}$  disappear and the derivation is exactly that which would arise from a standard linear models approach. Thus, in two limiting cases for which exact solutions are available, the linear approximation is in agreement with the exact results.

However, there are several issues with implementing the above approach in practice:

- *Estimating  $\sigma^2$* : Little work has been done regarding estimation of  $\sigma^2$  in high-dimensional regression. One reasonable approach is to choose a value of  $\lambda$  intended to achieve high predictive accuracy and then estimate  $\sigma^2$  as

$$\hat{\sigma}^2 = \frac{\mathbf{r}'_{\lambda} \mathbf{r}_{\lambda}}{n - df_{\lambda}}, \quad (4.17)$$

where  $df_{\lambda}$  is the degrees of freedom of the fit. For the lasso, this is commonly taken to be the number of nonzero coefficients in the model (Zou et al., 2007). We implement this approach for the results in section 4.4 and 4.5, using AIC as the model selection criterion.

- *Estimating  $\widehat{E_0(\mathbf{y})}$* : In a traditional regression setup, coefficient estimation is unbiased, so  $\mathbf{X}_{-j} \hat{\boldsymbol{\beta}}_{-j}$  is an unbiased estimator of  $E_0(\mathbf{y})$ . However, in penalized regression, coefficient estimation is biased towards zero, so  $\mathbf{X}_{-j} \hat{\boldsymbol{\beta}}_{-j}$  will underestimate  $E_0(\mathbf{y})$ . Little research has been published regarding the bias of the lasso, however, so the extent of this underestimation remains unclear. It

is also unclear how to correct for it; therefore, we use  $\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}$  as an estimator in section 4.4 and 4.5, with  $\hat{\boldsymbol{\beta}}$  again chosen using AIC.

- *The ramifications of treating  $\mathbf{H}_{-j}$  as fixed:* Treating  $\mathbf{H}_{-j}$  as fixed even though it contains the random elements  $\mathbf{s}_A$  and  $\|\hat{\boldsymbol{\beta}}_A\|_1$  will affect the accuracy of the estimator. However, we note that the random elements pertain only to the size of  $\hat{\boldsymbol{\beta}}$  (the number of its nonzero elements and its  $L_1$  norm), not the coefficient estimates themselves. The size of  $\hat{\boldsymbol{\beta}}$  is controlled largely by  $\lambda$ , which is taken to be fixed. It is possible, therefore, that the treatment of  $\mathbf{H}_{-j}$  as fixed will not present major problems.

We will examine the impact of these issues upon FDR estimation in section 4.4.

### 4.3.2 Logistic regression

For logistic regression,  $\hat{\boldsymbol{\pi}}$  is not a linear function of  $\mathbf{y}$ . In order to apply the approach here, we will therefore need to construct an approximate linear predictor. Consider a Taylor series approximation of  $\mathbf{y} - \hat{\boldsymbol{\pi}}$  about  $\boldsymbol{\beta}$ , the true value of the regression coefficients:

$$\mathbf{y} - \hat{\boldsymbol{\pi}} \approx \mathbf{y} - \boldsymbol{\pi} - \mathbf{W}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad (4.18)$$

where  $\boldsymbol{\pi}$  is the true value of  $E(\mathbf{y})$  and  $\mathbf{W}$  is a diagonal matrix with elements  $\{\pi_i(1 - \pi_i)\}$ . It is worth mentioning that for logistic regression, the intercept cannot be eliminated from the objective functions as it can in linear regression; hence, in our derivations of logistic regression, it is important to remember that  $\mathbf{X}$  and  $\boldsymbol{\beta}$  contain entries for the intercept and therefore have  $p + 1$  elements.

Substituting this approximation into (4.7), we arrive at a set of approximate

KKT conditions:

$$\frac{1}{n}\mathbf{x}'_j\mathbf{W}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \lambda\text{sign}(\hat{\beta}_j) \quad \forall \hat{\beta}_j \neq 0, \quad (4.19a)$$

$$\frac{1}{n}|\mathbf{x}'_j\mathbf{W}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})| \leq \lambda \quad \forall \hat{\beta}_j = 0. \quad (4.19b)$$

where  $\mathbf{z} = \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\pi}) + \mathbf{X}\boldsymbol{\beta}$ . Note that  $\mathbf{z}$  here has the same form as the adjusted response in the traditional quadratic approximation to model fitting and asymptotic inference for logistic regression models. Letting  $\mathbf{r}_{-j} = \mathbf{z} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}$ , we have the familiar

$$\frac{1}{n}|\mathbf{x}'_j\mathbf{W}\mathbf{r}_{-j}| > \lambda \quad \forall \hat{\beta}_j \neq 0, \quad (4.20)$$

$$\frac{1}{n}|\mathbf{x}'_j\mathbf{W}\mathbf{r}_{-j}| \leq \lambda \quad \forall \hat{\beta}_j = 0, \quad (4.21)$$

Starting from (4.19), we may follow the same approach as in section 4.3.1 to obtain the relation  $\hat{\boldsymbol{\beta}} = \mathbf{S}\mathbf{z}$  as before, except with

$$\mathbf{S} = \mathbf{M}^{-1}\mathbf{X}'_A\mathbf{W}$$

and

$$\mathbf{M} = \mathbf{X}'_A\mathbf{W}\mathbf{X}_A + n\lambda\mathbf{s}_A\mathbf{s}'_A\|\hat{\boldsymbol{\beta}}_A\|_1^{-1},$$

where the first entry of  $\mathbf{s}_A$  here is zero, resulting from the unpenalized intercept.

If, as is done in traditional logistic regression, we treat  $\mathbf{z}$  as following an approximately normal distribution, and we again treat  $\mathbf{H}_{-j} = (\mathbf{X}_A)_{-j}\mathbf{S}_{-j}$  as fixed, then  $n^{-1}\mathbf{x}'_j\mathbf{W}\mathbf{r}_{-j}$  will follow an approximately normal distribution, with mean and variance as follows:

$$\hat{\mu}_j = \frac{1}{n}\mathbf{x}'_j\mathbf{W}(\mathbf{I} - \mathbf{H}_{-j})\widehat{\mathbf{E}}_0(\mathbf{y}) \quad (4.22)$$

$$\hat{\tau}_j^2 = n^{-2}\mathbf{x}'_j\mathbf{W}(\mathbf{I} - \mathbf{H}_{-j})\mathbf{W}^{-1}(\mathbf{I} - \mathbf{H}_{-j})'\mathbf{W}\mathbf{x}_j, \quad (4.23)$$

since  $\mathbf{E}(\mathbf{z}) = \mathbf{X}\boldsymbol{\beta}$  and  $\text{Var}(\mathbf{z}) = \mathbf{W}^{-1}$ . The same issues will be present with this

estimator as with the linear regression approach:

- *Estimating  $\widehat{E_0(\mathbf{y})}$*
- *Estimating  $\mathbf{W}$*
- *The ramifications of treating  $\mathbf{H}_{-j}$  as fixed*

However, with logistic regression, there is also another concern:

- *The accuracy of approximation (4.18)*

The estimation of  $\widehat{E_0(\mathbf{y})}$  and  $\mathbf{W}$  will again derive from the AIC-selected model for the numeric results of section 4.4, in which the accuracy of the proposed approach will be assessed.

## 4.4 Simulations

### 4.4.1 Accuracy

We begin by examining whether the estimators proposed in sections 4.2 and 4.3 provide accurate estimates of the true false discovery rate. We will test this by simulating data sets from the following models:

$$y_i = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i, \quad \text{where } \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1) \quad (\text{Gaussian}) \quad (4.24)$$

$$\text{logit}(y_i) = \sum_{j=1}^p x_{ij}\beta_j \quad (\text{binomial}) \quad (4.25)$$

The elements of the design matrix were generated from a standard normal distribution:  $x_{ij} \stackrel{\text{iid}}{\sim} N(0, 1)$ , while

$$\beta_j = \begin{cases} 1 & \text{if } j \in \{1, \dots, \frac{p_1}{2}\} \\ -1 & \text{if } j \in \{\frac{p_1}{2} + 1, \dots, p_1\} \\ 0 & \text{otherwise,} \end{cases} \quad (4.26)$$

where  $p_1$  denotes the number of features contributing to the outcome (*i.e.*, the number of nonzero coefficients in the generating model). For the simulations presented



below,  $p_1 = 6$ ,  $n = 100$  for the Gaussian response and  $n = 200$  for the binomial response. Two values for  $p$  were considered ( $p = 50$  and  $p = 500$ ) in order to examine behavior in both low dimensional and high dimensional settings. Data sets were independently generated 500 times and the estimated false discovery rates as well as the true false discovery rate (proportion of features selected for which  $\beta_j = 0$ ) as a function of  $\lambda$  were recorded. The smoothed average of these rates is plotted in Figures 4.1 and 4.2.

As Figure 4.1 indicates, the heuristic approach seems to estimate the true false discovery rate quite well for lasso-penalized linear and logistic regression in both high- and low-dimensional settings. The estimated and true curves are rather similar, and when discrepancy arises, the heuristic estimate is slightly conservative.

For the linear predictor approach, however, the agreement with the true FDR is not as good. In order to separate the effect of estimating nuisance parameters from the fundamental accuracy of the approach, we examine two forms of the estimator. The first (“linear”) involves estimating  $\sigma^2$  and  $E_0(\mathbf{y})$  as described earlier. The other (“linear-true”) involves using replacing those quantities with their known values. Clearly, the linear-true approach is not available in practice. Figure 4.2 illustrates the performance of these approaches in estimating the FDR for the same data sets as in Figure 4.1.

From the figure, it appears that while the linear approach performs well in the low-dimensional Gaussian case, its accuracy in other settings is not as good as the heuristic approach. For logistic regression cases, the linear predictor approach produces very conservative FDR estimates, while for the high-dimensional Gaussian case, it results in liberal estimates of the FDR at the key  $\lambda$  values where the true FDR is close to 10%.

The linear-true approach is certainly more accurate than the linear approach,

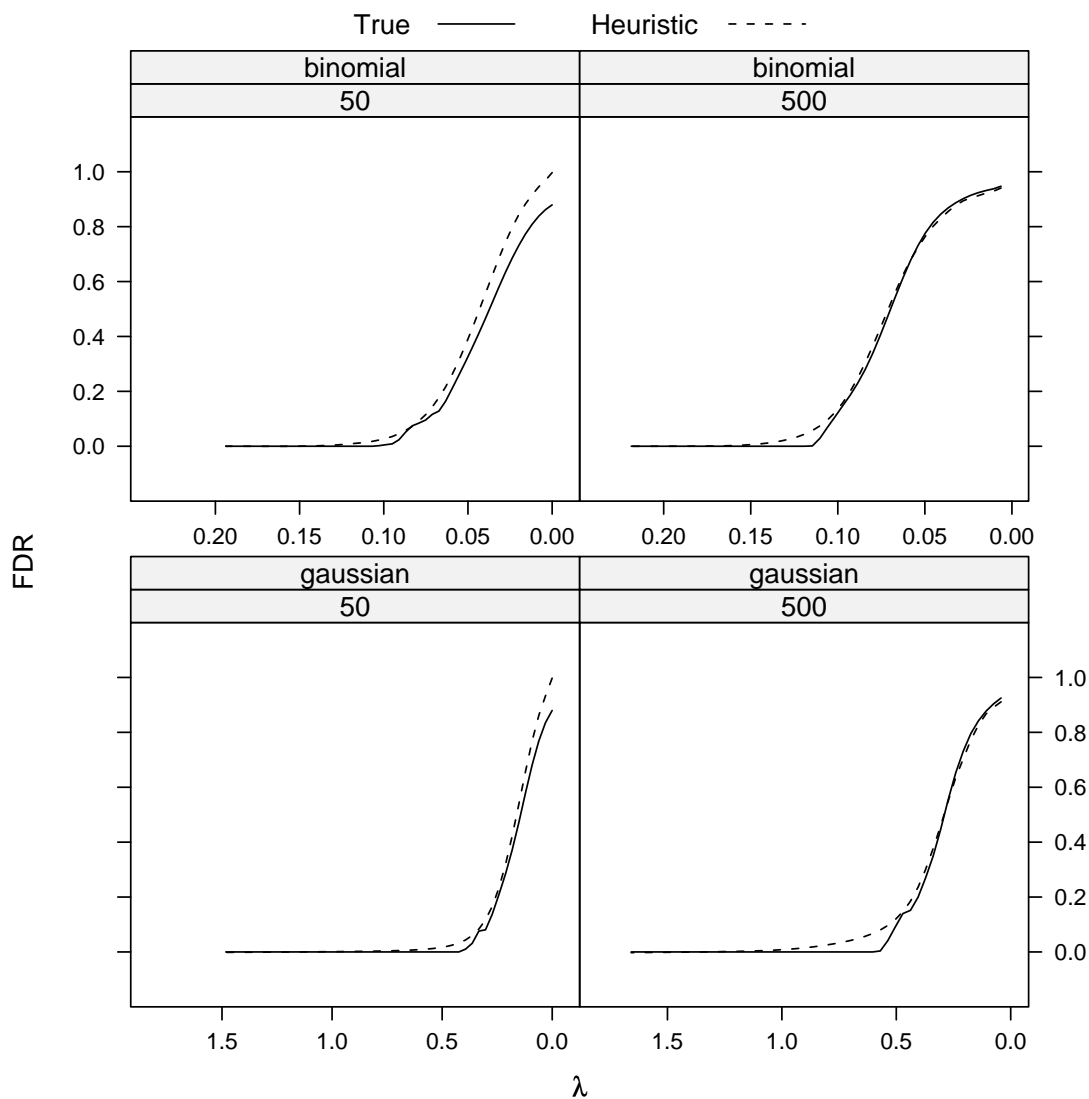


Figure 4.1: Simulation: Accuracy of FDR estimation for the heuristic approach. In each panel, the response distribution and number of features are indicated in strips at the top. The regularization parameter  $\lambda$  is on the horizontal axis, while the true/estimated FDR is on the vertical axis. For each of the 500 independently generated data sets, FDR estimation was carried out for 50 values of lambda. A loess curve was then fit to all the points and plotted.

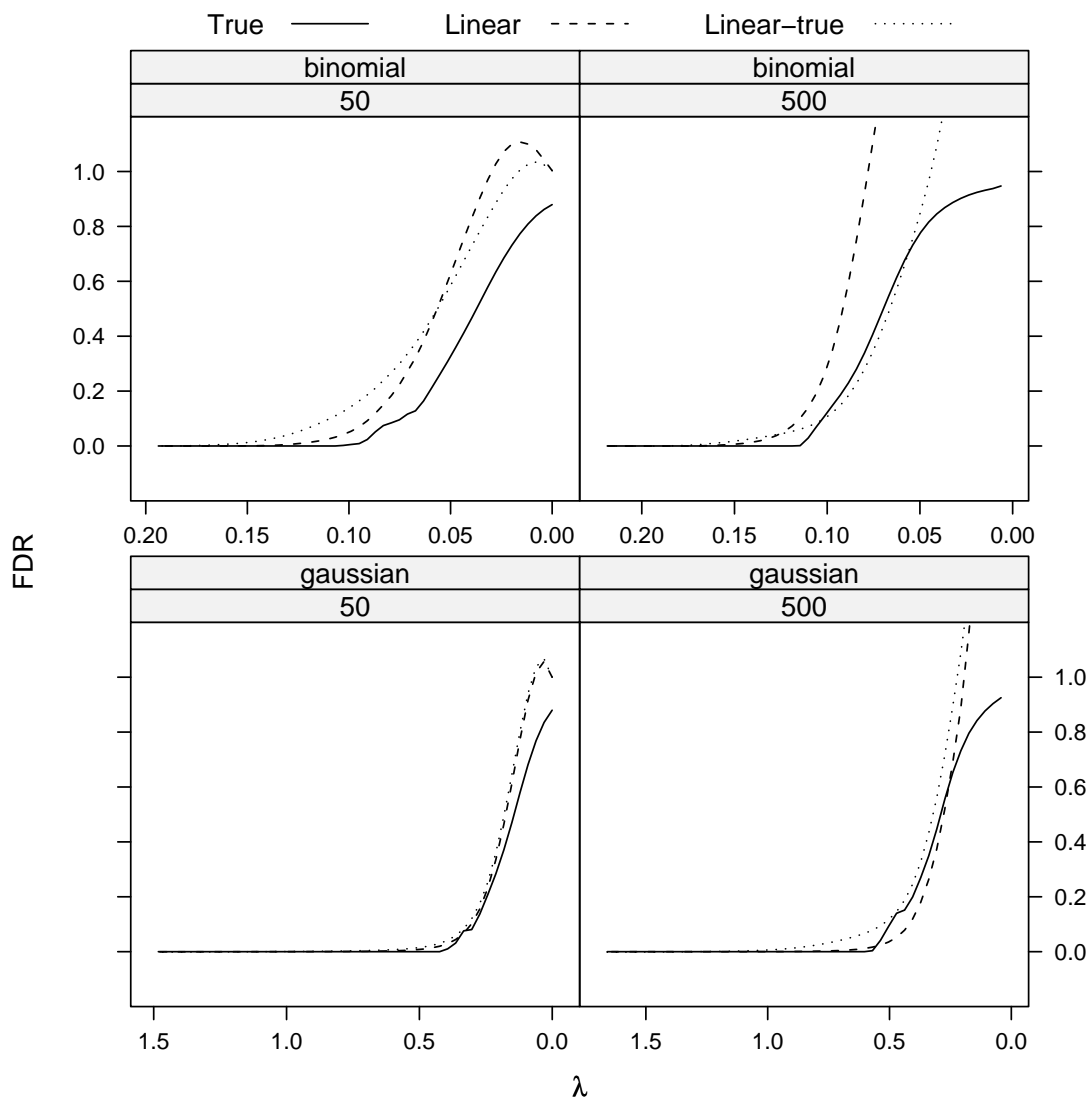


Figure 4.2: Simulation: Accuracy of FDR estimation for the linear predictor approach. In each panel, the response distribution and number of features are indicated in strips at the top. The regularization parameter  $\lambda$  is on the horizontal axis, while the true/estimated FDR is on the vertical axis. For each of the 500 independently generated data sets, FDR estimation was carried out for 50 values of lambda. A loess curve was then fit to all the points and plotted. The data sets here are the same as those in 4.1, and thus the “true” line is the same in both plots.

particularly for the high dimensional settings. However, the linear-true FDR estimates are rather inaccurate for the low dimensional binomial setting, indicating that either (4.18) or the approximate normality of the adjusted response  $\mathbf{z}$  is poor in certain cases. Furthermore, it is the failure to estimate  $E_0(\mathbf{y})$  accurately, not  $\sigma^2$ , that is the cause of this failure, as we can verify by replacing the estimates with their known values separately (simulations not shown). Finally, we see that the treatment of  $\mathbf{H}_{-j}$  as fixed leads to problems for heavily overfit models, as the linear-true FDR estimates are far from the actual FDR for models in which  $\lambda$  is small. Fortunately, however, the FDR estimates are conservative in these cases.

Although the linear predictor approach has merit, the inability to estimate  $E_0(\mathbf{y})$  well limits our ability to apply this approach in practice. Therefore, we will consider only the heuristic estimator in the remainder of the simulations and applications.

#### 4.4.2 Lasso FDR vs. competing approaches

We now turn our attention to the merits of a lasso FDR-based approach to feature selection in comparison with competing approaches. We consider lasso-based approaches using AIC (Akaike, 1973) or BIC (Schwarz, 1978) to select the tuning parameter  $\lambda$  as well as univariate approaches with the FDR controlled using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

As before, we will consider the generating model to be (4.24) for the Gaussian response and (4.25) for the binomial response. However, we will now consider correlated design matrices as well. In the correlated design, each feature for which  $\beta_j \neq 0$  will be correlated with  $m$  features for which  $\beta_k = 0$  in the following manner:

$$x_k = x_j + \sqrt{3}Z,$$

where  $Z \sim N(0, 1)$ . This results in a correlation of 0.5 between the two features.

Table 4.2: Simulation design: Number of causative, correlated, and spurious features for each setting.

| Dimension | Features     | Causative | Correlated | Spurious | Total |
|-----------|--------------|-----------|------------|----------|-------|
| Low       | Uncorrelated | 6         | 0          | 44       | 50    |
| Low       | Correlated   | 6         | 12         | 32       | 50    |
| High      | Uncorrelated | 6         | 0          | 494      | 500   |
| High      | Correlated   | 6         | 54         | 440      | 500   |

There will therefore be  $p_1$  “causative” features that contribute to the outcome,  $mp_1$  “correlated” features that are correlated with the outcome but only through their mutual correlation with a causative feature, and  $p - (m + 1)p_1$  “spurious” features whose only association with the outcome is due to chance. For the low-dimensional case,  $m = 2$ , while for the high-dimensional case,  $m = 9$ . The number of variables in each category for the various simulation designs are summarized in Table 4.2.

As before, 500 data sets were generated independently for each combination of response distribution, number of features, and design (correlated/uncorrelated). For each data set, a lasso model was fit to the data and AIC/BIC/FDR used to select  $\lambda$ . In addition, a (Wald) hypothesis testing procedure was carried out for each feature individually in a one covariate linear/logistic regression model, with FDR used to decide the significance level. The numbers of features of each type (“causative”, “correlated”, and “spurious”) selected by the procedure were recorded. Note that the uncorrelated design has no features in the “correlated” category. The results of this simulation are shown in Figures 4.3 and 4.4.

We can draw a number of conclusions from Figures 4.3 and 4.4. First, the FDR approaches are successful at limiting the fraction of selected features that arise from spurious association. The fraction of the bar colored white is always small for the FDR approaches, whereas a large portion of the features selected by AIC and

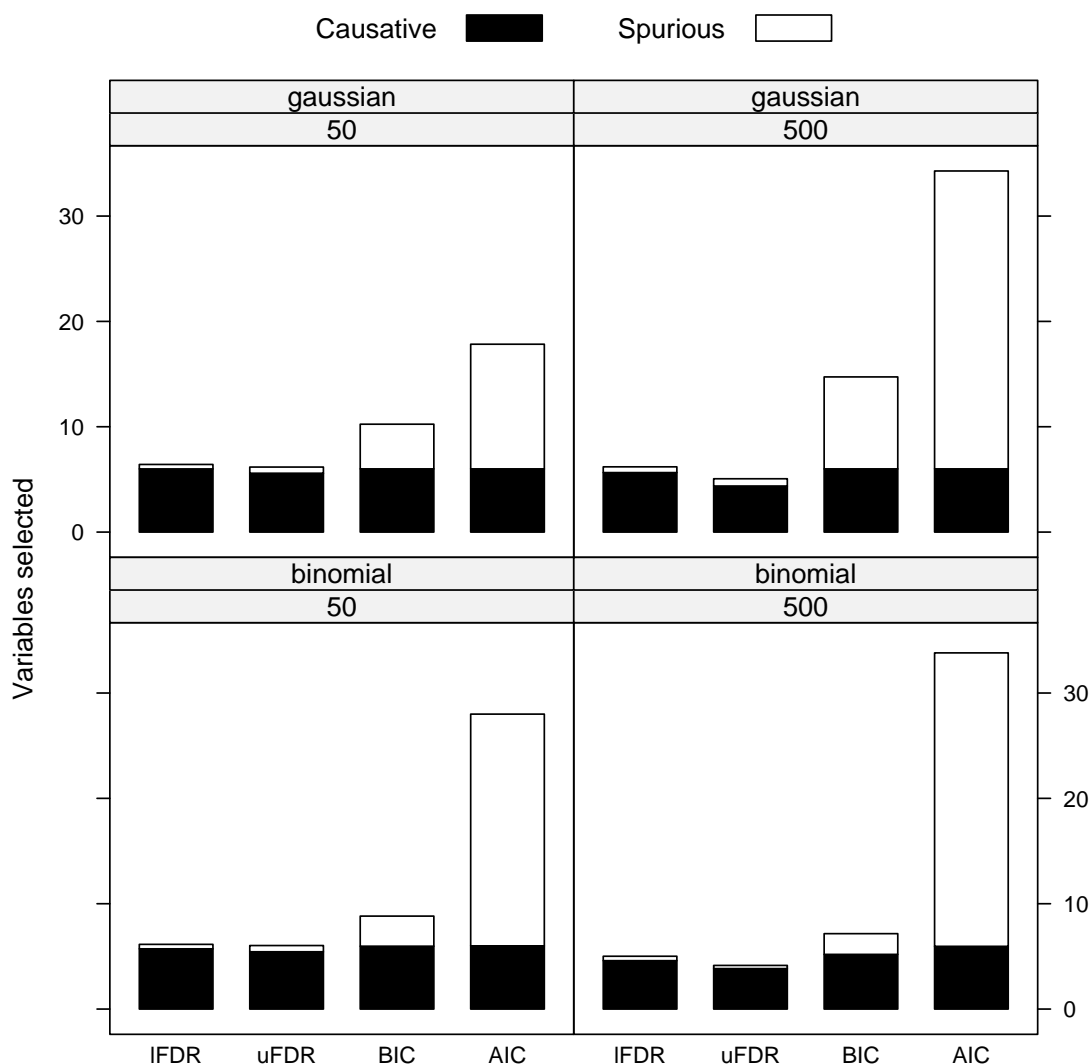


Figure 4.3: Simulation: Causative and spurious features selected by various approaches (uncorrelated design). In addition to the lasso FDR approach (lFDR) and FDR-controlled univariate hypothesis tests (uFDR), features were selected using AIC and BIC to choose the lasso regularization parameter. For lFDR and uFDR, the FDR threshold was set to 10%. In each panel, the response distribution and number of features are indicated in strips at the top. The number of variables of each type are stacked, with bars shaded by the type of feature selected: causative features in black, spurious features in white. Each panel contains the averaged results of 500 independently generated data sets. Note that the total height of each bar is the average number of features selected by each approach.

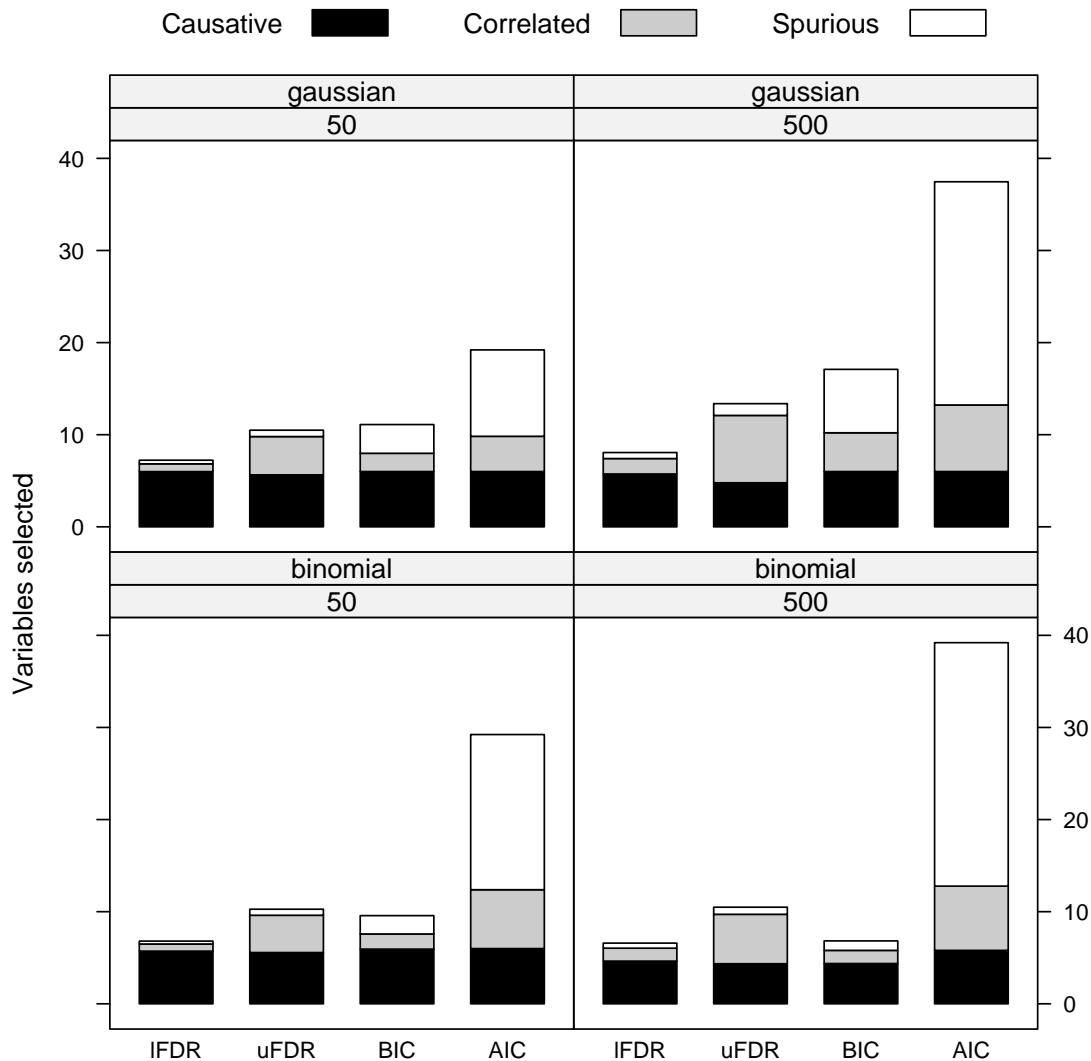


Figure 4.4: Simulation: Causative, correlated, and spurious features selected by various approaches (correlated design). In addition to the lasso FDR approach (IFDR) and FDR-controlled univariate hypothesis tests (uFDR), features were selected using AIC and BIC to choose the lasso regularization parameter. For IFDR and uFDR, the FDR threshold was set to 10%. In each panel, the response distribution and number of features are indicated in strips at the top. The number of variables of each type are stacked, with bars shaded by the type of feature selected: causative features in black, correlated features in gray, and spurious features in white. Each panel contains the averaged results of 500 independently generated data sets. Note that the total height of each bar is the average number of features selected by each approach.

BIC may be spurious. Second, univariate approaches are unable to distinguish between causative and correlated features, and thus select a large number of correlated features. Third, in the settings considered here, an FDR of 10% is a more stringent cutoff than BIC or AIC. Finally, the lasso FDR approach seems to be slightly more powerful than the univariate FDR approach at selecting causative features, even when the univariate FDR approach selects a much larger number of features.

## 4.5 Applications

### 4.5.1 Gene expression data

We now apply the FDR-regularized lasso methodology to applications in high-dimensional biomedical research. First, we examine the gene expression study of leukemia patients presented in Golub et al. (1999). In the study, the expression levels of 7129 genes were recorded for 27 patients with acute lymphoblastic leukemia (ALL) and 11 patients with acute myeloid leukemia (AML) (in the study, additional samples were collected to form a testing set for prediction methods, but we will not deal with the additional samples here).

Features were selected from the leukemia data set using  $t$ -tests to select differentially expressed genes as well as using the lasso with  $\lambda$  chosen by FDR. For both methods, logarithms of the expression levels were taken prior to model fitting. The number of features selected by each method at three different FDR levels, as well as the number of features in common, are presented in Table 4.3. No assumption of equal variance was made for the  $t$ -tests.

From table 4.3, we can draw several conclusions. First, association with the outcome conditional on the other variables in the model is clearly a much more stringent criterion than marginal association with the outcome. Thus, the multivariate, lasso-based approach selects far fewer features than the univariate,  $t$ -test approach. However, the features selected by the lasso are contributing independently to the



Table 4.3: Features selected at various FDR levels by  $t$ -test and lasso approaches for a gene expression study of leukemia.

| FDR | $t$ -test | Lasso | In common |
|-----|-----------|-------|-----------|
| .01 | 165       | 4     | 4         |
| .05 | 554       | 13    | 11        |
| .10 | 898       | 13    | 13        |

outcome, whereas the features selected by the  $t$ -tests may be providing largely redundant information. Second, as the FDR threshold is relaxed, the univariate approach selects ever greater numbers of features. However, the lasso can never select more than  $n$  features without the model ceasing to be identifiable. As a result, when  $p \gg n$ , as the number of selected features becomes appreciable in size compared to  $n$ , the false discovery rate increases rapidly while the number of selected features remains relatively static. This sharp increase in the false discovery rate was also observed in Figure 4.1. Lastly, the  $t$ -test approach fails to identify features chosen by the lasso despite the  $t$ -test approach selecting hundreds of additional features. In particular, at a false discovery rate of 5%, two of the 13 lasso-selected features are not among the 554 most significant univariate results. Furthermore, of the 13 genes selected by the lasso at a 5% FDR, only 2 are among the 25 most significant as selected by the  $t$ -test.

If the results of this gene expression study were used to select features, perhaps to further investigate as biomarkers, obtaining the small set of features that contribute independently to the outcome from the lasso FDR approach may be preferable to the large set of possibly redundant features obtained from univariate screening.

Table 4.4: Number of features selected at various FDR levels by univariate trend test and lasso approaches for a genetic association study of age-related macular degeneration.

| FDR | Trend test | Lasso | In common |
|-----|------------|-------|-----------|
| .01 | 2          | 1     | 1         |
| .05 | 5          | 3     | 3         |
| .10 | 5          | 3     | 3         |

#### 4.5.2 Genetic association study

We also apply the lasso FDR approach to the genetic association study presented in Chapter 2. The data here is the same as that analyzed in section 2.5. Again we compare the lasso-based approach with a univariate approach, here a Cochran-Armitage linear trend test to reflect the categorical nature of the genetic covariates. As in the previous section, the number of features selected by each method at three at three different FDR levels, as well as the number of features in common, are presented in table 4.4.

In sharp contrast to the leukemia gene expression study, in the genetic association study, the lasso-based and multivariate approaches are largely in agreement. The numbers of genetic markers selected by the two methods are similar, and all the markers selected by the lasso approach were also identified by the univariate approach. From these two examples, then, we see that depending on the situation, FDR-guided feature selection may differ markedly between univariate and multivariate approaches, or the two approaches may be in close agreement.

## CHAPTER 5

### SUMMARY

As the automated collection and storage of data becomes cheaper to obtain and easier to implement, high-dimensional problems are becoming increasingly common. For these problems, traditional approaches to regression break down and new methods are needed. Introducing additional information in the form of a penalized objective function is an elegant, flexible, and practical approach for dealing with these problems. However, penalized methods have been lacking in several areas, three of which are addressed in this thesis.

First, there has been little work on incorporating specific prior covariate structure into penalized regression models. This is a particular problem in systems biology, in which much is often known about relationships involving genes and genetic markers prior to the study in question. Chapter 2 of this thesis deals with the incorporation of a specific type of structure: grouped covariates. I introduce a framework that sheds light on the behavior of grouped penalization methods and apply these methods to an important study design in modern genetics – genetic association studies. In addition, I develop a novel group penalty, group MCP, demonstrate that its grouping assumptions are less severe than those of group lasso and group bridge, and show that it performs better than those competing methods in situations with substantial ( $> 50\%$ ) within-group sparsity.

Second, it is of crucial importance for high dimensional problems that algorithms for fitting penalized regression models be as efficient as possible. In Chapter 3, I develop fast, stable algorithms for fitting models with complicated penalties, such as group penalties. Importantly, I show these algorithms to require  $O(np)$  computations and to decrease the objective function with every step, as well as demonstrating their numerical efficiency compared with competing algorithms in a

variety of simulations.

Third, and perhaps most importantly, meaningful inference has been limited. Chapter 4 addresses this problem by applying the idea of false discovery rates to penalized regression models. Through inspection of the Karush-Kuhn-Tucker conditions, I demonstrate that penalized regression models involve a series of multiple comparisons that can be used to define a false discovery rate. I propose two approaches to estimating these false discovery rates, investigate their accuracy, and compare FDR-regularized model selection with competing approaches, demonstrating multiple attractive features of the proposed methodology.

As always, there is the possibility for much future work along these lines. As illustrated in Chapter 2, group bridge, group MCP, and group lasso all have benefits and drawbacks, with each method performing well in certain situations and not in others. It would be desirable to develop a more robust method that performs well across a variety of situations. Investigating false discovery rate estimation for penalized regression models other than the lasso – in particular, for grouped penalization methods – also warrants further study. Finally, comprehensive studies of the application of penalized regression methods to specific applications would be extremely valuable. One application of particular interest is genome-wide association studies, where much study is needed regarding the impact of issues in genetics (such as linkage, penetrance, and the genetic basis of the disease) upon the performance of penalized regression and other approaches to analyzing these data.

In conclusion, the research described in this thesis, in addition to being valuable in and of itself, lays the foundation for future analyses of high-dimensional data in which we can propose complex penalties that incorporate prior understanding and specific covariate structure, fit such models efficiently, and obtain useful and intuitive inferential measures.

## REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pp. 267–281.
- Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Methodological* 57, 289–300.
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 24(6), 2350–2383.
- Donoho, D. (2000, August). High-dimensional data analysis: The curses and blessings of dimensionality. Lecture delivered at the conference "Math Challenges of the 21st Century" held by the American Math. Society organised in Los Angeles, August 6-11.
- Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–455.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics* 32(2), 407–499.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- Fan, J. and R. Li (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians*, pp. 595–622. European Mathematical Society.
- Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools (Disc: P136-148). *Technometrics* 35, 109–135.
- Friedman, J., T. Hastie, H. Hofling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2), 302–332.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Regularization paths for generalized linear models via coordinate descent. <http://www-stat.stanford.edu/~hastie/Papers/glmnet.pdf>.
- Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.
- Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag Inc.
- Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

- Huang, J., S. Ma, H. Xie, and C.-H. Zhang (2007). A group bridge approach for variable selection. Technical Report #376, Department of Statistics and Actuarial Science, University of Iowa.
- Lange, K., D. R. Hunter, and I. Yang (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics* 9(1), 1–59.
- McCullagh, P. and J. A. Nelder (1999). *Generalized Linear Models*. Chapman & Hall Ltd.
- Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics* 9(2), 319–337.
- Park, M. Y. and T. Hastie (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 69(4), 659–677.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological* 58, 267–288.
- Wu, T. T. and K. Lange (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* 2(1), 224–244.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 68(1), 49–67.
- Zhang, C.-H. (2007). Penalized linear unbiased selection. Technical Report #2007-003, Department of Statistics and Biostatistics, Rutgers University.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 67(2), 301–320.
- Zou, H., T. Hastie, and R. Tibshirani (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics* 35(5), 2173–2192.
- Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* 36(4), 1509–1533.