Marginal false discovery rate approaches to inference on penalized regression models

Ryan Miller

University of Iowa

Copyright © 2018 Ryan Miller

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/6474

Recommended Citation
https://doi.org/10.17077/etd.n7cwkl17

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Biostatistics Commons
MARGINAL FALSE DISCOVERY RATE APPROACHES TO INFERENCE ON PENALIZED REGRESSION MODELS

by

Ryan Miller

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

August 2018

Thesis Supervisor: Patrick Breheny, Associate Professor of Biostatistics
This is to certify that the Ph.D. thesis of

Ryan Miller

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biostatistics at the August 2018 graduation.

Thesis Committee:  Patrick Breheny, Thesis Supervisor

Joe Cavanaugh

Mike Jones

Yuan Huang

Aixin Tan
ACKNOWLEDGMENTS

Foremost I’d like to thank my adviser Patrick Breheny for everything he has done for me over the past several years. Patrick has been foundational in my development as a statistician, and I am grateful for the mentorship he has provided me as a student, teacher, and researcher. During my first year I was lucky to work as a teaching assistant for Patrick’s Introduction to Biostatistics. Despite being pretty familiar with the material, sitting in on his lectures got me to really grasp the big picture ideas that central understanding statistics at a deeper level. During my third year I was able to take his High-Dimensional Data Analysis class, a truly excellent course that made the transition into my dissertation easy. Since then I’ve enjoyed being an advisee of his; Patrick is always accessible, helpful, and willing to provide advice on anything. Our regular meetings are thought provoking and insightful; and have made working on my dissertation enjoyable and fulfilling.

I’d also like to thank my committee members, Joe Cavanaugh, Mike Jones, Yuan Huang, and Aixin Tan for their time commitment and oversight of my dissertation. Their willingness to serve on my committee is something I greatly appreciate.

Additionally, I’m grateful for the support I’ve received as a teaching and research assistant, without the financial support provided by these positions it is likely that I would not have attended graduate school. I’m thankful to the Department of Biostatistics for its role in providing me these funding opportunities, and also thankful to Michelle Reyes, Betsy Momany, Natoshia Askelson, and Pete Damiano for their roles during my time as a research assistant.

I’d like to thank my parents, Carol and David Miller, for countless things they’ve done for me throughout my entire life. Without the wonderful upbringing they provided I would not have the confidence nor tenacity that are necessary to pursue a PhD.
I’d also like to thank “The Boys”, my fellow Biostatistics PhD candidates: Ryan Peterson, Brandon Butcher, and Javi Flores for making my time in the program enjoyable and full of wonderful memories.

Last, but not least, I’d like to thank my fiance Alissa for her unwavering support throughout my journey. I couldn’t be more thankful for having her in my life.
ABSTRACT

Penalized regression methods, most notably the lasso, are a popular analysis tool for data sets containing large numbers of variables due to their ability to naturally perform variable selection. However, performing statistical inference on lasso models is difficult, making the topic an active area of research. Much of the literature on inferential methods for penalized regression models is focused exclusively on the linear regression setting, with few options available for models with categorical or survival outcomes. This dissertation proposes novel false discovery rate methods for a broad class of penalized likelihood modeling approaches that includes not only linear regression, but also generalized linear models, such as logistic regression, and survival models, such as Cox regression. The methods we propose are designed to be suitable next steps after fitting a penalized likelihood model, requiring minimal added computation or inputs other than the fitted model.

The methods proposed in this dissertation deal with two broad categories of false discovery rates that exist in the large-scale hypothesis testing literature. In the first category are tail-area based approaches, which pertain to the overall false discovery rate for an entire set of rejected hypotheses. In the case of penalized regression, this is the entire set of variables active in the model. In the second category are local approaches, which refer to the specific probabilities of single variables being false discoveries, in the regression setting local approaches describe each active variable individually. Both types of false discovery rates are valuable tools depending upon the inferential question, and this dissertation proposes methods falling into both categories.

In addition to making these proposals and investigating their properties using simulation, we also look at applying the methods to several case studies involving
genetic data. These data sets are high-dimensional, with tens of thousands of genetic variables, but only a few hundred observations. While false discovery rate methods, such as those we propose, are particularly relevant in these applications, our methodology can be applied to any situation where penalized regression modeling is employed, regardless of the dimensionality. Currently the methods described in this dissertation are publicly available in the R package ncvreg using the summary and mfdr functions.
Data containing large number of variables is becoming increasingly more common and sparsity inducing penalized regression methods, such as the lasso, have become a popular analysis tool for these datasets due to their ability to naturally perform variable selection. However, quantifying the importance of the variables selected by these models is a difficult task. These difficulties are compounded by the tendency for the most predictive models, for example those which were chosen using procedures like cross-validation, to include substantial amounts of noise variables with no real relationship with the outcome. To address the task of performing inference on penalized regression models, this thesis proposes false discovery rate approaches for a broad class of penalized regression models. This work includes the development of an upper bound for the number of noise variables in a model, as well as local false discovery rate approaches that quantify the likelihood of each individual selection being a false discovery. These methods are applicable to a wide range of penalties, such as the lasso, elastic net, SCAD, and MCP; a wide range of models, including linear regression, generalized linear models, and Cox proportional hazards models; and are also extended to the group regression setting under the group lasso penalty. In addition to studying these methods using numerous simulation studies, the practical utility of these methods is demonstrated using real data from several high-dimensional genome wide association studies.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. ix
LIST OF FIGURES ............................................................... x
PREFACE ........................................................................... xiii

CHAPTER

1 Introduction ................................................................. 1
  1.1 Large-scale Testing .................................................. 4
    1.1.1 Family-wise Error Control ................................. 5
    1.1.2 False Discovery Rates ...................................... 6
    1.1.3 A Bayesian Derivation of Fdr ............................ 8
    1.1.4 Local False Discovery Rates .............................. 9
  1.2 Penalized Regression Models .................................... 11
    1.2.1 Ridge Regression ............................................ 12
    1.2.2 The lasso ...................................................... 13
    1.2.3 The Elastic Net .............................................. 16
    1.2.4 Non-convex Penalties ..................................... 18
    1.2.5 The Group lasso .............................................. 19
  1.3 Inference for Penalized Regression Models ................... 21
    1.3.1 Sample Splitting ............................................ 22
    1.3.2 Selective Inference ......................................... 23
    1.3.3 Low Dimensional Projection ............................ 25
    1.3.4 Other Approaches ......................................... 26

2 Marginal False Discovery Rates for Penalized Likelihood Models ... 28
  2.1 Abstract ................................................................. 28
  2.2 Introduction ............................................................. 28
  2.3 Marginal false discovery rates ................................... 30
    2.3.1 Penalized likelihood optimization ...................... 32
    2.3.2 Marginal false discovery rate bounds for penalized likelihood methods ............................................. 33
    2.3.3 Other penalty functions .................................. 37
    2.3.4 Logistic regression ......................................... 38
    2.3.5 Cox regression .............................................. 38
  2.4 Simulation studies .................................................. 40
    2.4.1 Accuracy, sample size, and correlation ................. 41
    2.4.2 Features associated with censoring ..................... 43
    2.4.3 Comparison with other methods ......................... 44
  2.5 Case studies ............................................................ 48
    2.5.1 Lung cancer survival and gene expression .............. 48
    2.5.2 Lung cancer status among smokers ...................... 52
  2.6 Discussion ............................................................... 54
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A summary of the potential outcomes of $h$ hypothesis tests for a given rejection rule. Here $h$ and $R$ are known while $h_0$ and $h_1$ are unobserved.</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>The ratio of Causal:Noise (A:C) feature selections for various values of the signal strength $b$.</td>
<td>48</td>
</tr>
<tr>
<td>2.2</td>
<td>Results for different choices of $\lambda$ in applying the lasso and MCP penalties to the Shedden and Spira data. mFDR = Marginal false discovery rate (%). CVE = Cross-validation error. MCE = Misclassification error (%).</td>
<td>50</td>
</tr>
<tr>
<td>3.1</td>
<td>Local false discovery rate accuracy results for the Hard Scenario. Features are binned based upon their estimated FDR and the observed proportion of noise variables in each bin is reported in the body of the table for each method.</td>
<td>74</td>
</tr>
<tr>
<td>3.2</td>
<td>Simulation results comparing the average number of selections of causal and correlated variables, as well as the proportion of noise variable selections, for various model-based false discovery rate control procedures. The “exact”, “spacing”, “mod-spacing”, and “covtest” methods are related tests performed by the <code>selectiveInference</code> package.</td>
<td>78</td>
</tr>
<tr>
<td>3.3</td>
<td>Local false discovery rate estimates of the top ten features, when performing univariate testing, for the Shedden survival data.</td>
<td>80</td>
</tr>
<tr>
<td>3.4</td>
<td>The top 10 selected genes from the univariate approach and their local false discovery rate estimates for each method.</td>
<td>83</td>
</tr>
<tr>
<td>4.1</td>
<td>The Scheetz case study results using several different analysis approaches that each control the false discovery rate at 10%. The different false discovery definitions of these approaches leads to the drastic differences in the number of significant findings.</td>
<td>109</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure

1.1 A 2-dimensional illustration of the lasso solution, demonstrating its tendency for sparse solutions. Here the constraint region defined by $||\beta||_1 \leq t = 1$ is shaded in yellow, This region first meets the contours of the likelihood when $\hat{\beta}_2 = 1$ and $\hat{\beta}_1 = 0$. \hspace*{1cm} 15

2.1 Finite-sample accuracy of the mFDR bound (2.3.4). \hspace*{1cm} 42

2.2 The number of true, “A”, variables selected by each lasso based method of false discovery rate control, averaged across 1,000 simulation iterations plotted as a function of $\beta$. \hspace*{1cm} 44

2.3 The average number of selections for each type of variable depicted in the causal diagram of Section 2.3, when cross-validation is used to select $\lambda$. \hspace*{1cm} 46

2.4 The average number of selections, for each type of variable depicted in the causal diagram of Section 2.3, is plotted as function of $\beta$ for each method which controls the false discovery rate. \hspace*{1cm} 47

2.5 Marginal false discovery estimates for the Shedden data. (left) Total number of genes selected and expected number of marginal false discoveries; (right) Estimated mFDR. \hspace*{1cm} 49

3.1 The rows of this figure correspond to two simulated datasets. The column on the left shows the usual lasso coefficient path, while the column on the right displays our method’s feature specific local false discovery rates (mfdr) along the lasso path. The triangles prior to the start of the mfdr path are the traditional local false discovery rate estimates resulting from large-scale testing. Along the mfdr path dashed lines indicate the portion of the path where a feature is inactive ($\hat{\beta} = 0$) in the model. The vertical dotted line shows the value of the penalty parameter, $\lambda$, chosen by cross validation. \hspace*{1cm} 58

3.2 Causal diagram showing three types of relationships between variables and the outcome. \hspace*{1cm} 71

3.3 In the easy scenario the expected proportion of false discoveries at a given estimated local false discovery rate is plotted using a linear model, the estimates produced by each of these methods are very accurate in this scenario. \hspace*{1cm} 73
3.4 The average number of features of each type with estimated local false discovery rates of less than 0.10 for each method in the hard and easy scenarios. 76

3.5 The mixture density estimates, \( \hat{f}(z) \), for the different methods applied to the Shedden data. We observe that the distribution of test statistics more closely resembles the null as more features are adjusted for by the model. 81

3.6 Causal diagram depicting possible gene relationships with BRCA1. 82

3.7 Smoothed calibration results for the Cox regression Easy Scenario. 86

3.8 Smoothed calibration results for the logistic regression Easy Scenario. 87

3.9 Smoothed calibration results for the linear regression Hard Scenario, ‘B’ variables are included as valid discoveries in this plot. 88

3.10 Smoothed calibration results for the linear regression Hard Scenario, ‘B’ variables are omitted from this plot. 89

3.11 Smoothed calibration results for the Cox regression Hard Scenario, ‘B’ variables are included as valid discoveries in this plot. 90

3.12 Smoothed calibration results for the logistic regression Hard Scenario, ‘B’ variables are included as valid discoveries in this plot. 90

3.13 The ROC and PR results for various methods of density estimation applied to the 200 replicated data sets from the easy scenario. 91

3.14 The ROC and PR results for various methods of density estimation applied to the 200 replicated data sets from the hard scenario, ‘B’ variables are omitted from these calculations. 91

4.1 The mean estimated false discovery rate, at the value of \( \lambda \) corresponding with an average false discovery rate of 15%, plotted as a function of sample size for various group sizes. 102

4.2 The estimated group local false discovery rates when \( K_j = 5 \) plotted versus each group’s null/non-null status. A linear model is used for smoothing such that each method’s calibration can be assessed by comparing the respective fitted line with the 45° dashed line. 103

4.3 The number of true groups found, while controlling the false discovery rate below 10%, by various methods for different effect sizes. The mFdr method performs well, while the mfdr method is comparable to large-scale testing using local false discovery rates. 105
4.4 The empirical null distribution of the statistic $\frac{1}{n\sigma^2}||\tilde{X}_j^T r_j||_2^2$ plotted versus its theoretical $\chi^2$ distribution for three choices of $\lambda$. $\lambda_{max}$ is the intercept only group lasso model, $\lambda_{cv}$ is the model chosen by cross validation, and $\lambda_{min}$ is the end of the $\lambda$ path - often the saturated model. In row 1, $J = 50$ and $K_j = 5$; in row 2, $J = 200$ and $K_j = 5$; in row 3, $J = 200$ and $K_j = 15$. 

107
PREFACE

This manuscript contains an introduction followed by three papers, each intended to make a novel contribution to the field of high-dimensional data analysis. These papers are intended to function as stand alone publications and therefore might contain some redundancies or overlaps in their content, particularly in their introductory sections. Additionally, space constraints limit the depth of the literature review and background material that can be covered in each paper, so as an accommodation the introductory chapter of this manuscript provides a comprehensive review, parts of which are summarized during the introduction section of each paper.

Broadly speaking, the first paper extends previous work done to bound the marginal false discovery rate of penalized linear regression models to a broader class of penalized likelihood models, with a specific focus on penalized logistic regression and penalized Cox regression. The second chapter adapts the methodology behind local false discovery rates to variable selections by lasso regression models. The third paper extends both of these methods to the group lasso setting. Collectively, the three papers seek to address the lack of inferential tools for various penalized regression models by providing methods of false discovery rate control and assessments of variable selection reliability.
CHAPTER 1
INTRODUCTION

Recent improvements in technologies related to data collection, storage, and processing have dramatically changed the nature of the data that is available for analysis. Modern datasets often contain very large numbers of features (variables), which can be challenging for the data analyst. The focus of this manuscript is on inferential tools that can be applied to situations where the data contains a large number of variables. Throughout the manuscript the phrase high-dimensional data is often used to refer to the specific setting where the number of features exceeds the number of observed data points. Generally speaking, all high-dimensional data has a large number of variables, however not all data with large numbers of variables is high-dimensional. The methods we propose are not exclusive to the high-dimensional data, but their usefulness in these settings is particularly appealing. High-dimensional data may also be referred in the context of the $p > n$ setting, where $p$ denotes the number of features, and $n$ denotes the number of observations. Some examples of high-dimensional data include: genetic microarrays, which allow for tens of thousands of gene expression measurements to be collected, proteomics and metabolomics technologies, which can collect data on thousands of metabolites, and imaging data, which can consist of thousands of pixels. Several case studies will be presented throughout this manuscript, each involve genetic microarray data sets with various types of outcome variables, however we reiterate that these methods can be applied to many applications that need not be high-dimensional.

Inference for high-dimensional data is a particularly challenging topic for reasons we will later describe. Broadly speaking, there are two major categories of approaches used to analyze high-dimensional data. The first are large-scale testing
approaches, which separately test each feature’s association with the outcome variable in a one at a time fashion and then use a procedure to intelligently combine the results of these many hypothesis tests. Well-established inferential methods are the primary appeal of large-scale testing. A second category of approaches are model-based, which center around creating a model that simultaneously considers all of the features and their relationships with the outcome variable. Model based approaches tend to be more recent developments in high-dimensional data analysis, and present numerous challenges with regards to inference.

To further narrow the scope of our work, this manuscript focuses specifically on false discovery rate approaches to inference, a class containing two related perspectives. The first are tail-area methods, which characterize the false discovery rate of an entire set of features/hypotheses; and the second are local methods, which characterize the individual false discovery rate of a single feature/hypothesis. Both tail-area and local methods have been extensively studied in the realm of large-scale testing, however they are either in their developmental stages, or completely absent, in the realm of high-dimensional regression modeling. The upcoming chapters propose new tail-area and local false discovery rate tools for penalized regression models. In doing so, we reconcile these new model-based approaches with those that currently exist in the large-scale testing literature.

The content of this dissertation is structured based upon the “three paper” format, where each chapter is designed to serve as a stand-alone publication. Chapter 1 is not one of these three papers; it instead provides background information at much greater level of detail than would be possible in space constrained publications. The contents of this introductory chapter begin with an overview of large-scale testing methods, placing a focus on both tail-area and local false discovery rate approaches and how the two approaches are connected. The second portion of the chapter
introduces a variety of popular penalized regression methods, including ridge regression, the lasso, non-convex penalties, and group penalties. The third portion provides a review of several existing inferential procedures for penalized regression. Much of the content in Chapter 1 will be briefly reiterated or summarized during the introduction sections of each subsequent chapter.

Chapter 2 is the first of the three papers that make up the body of this dissertation; it focuses on marginal false discovery rate control (mFdr) for penalized regression models. This chapter introduces the concept of mFdr, reviews existing work from the linear regression setting, and then extends the work to a more general class of likelihood based penalized regression models with a specific focus towards GLM and Cox proportional hazards models. The goal of this chapter is to develop a bound for the number of noise variable selections that are expected to be selected into a given penalized likelihood model by random chance. The method provides a convenient, computationally efficient assessment of the overall selection reliability of any fitted penalized likelihood model.

Chapter 3 develops a local approach to false discoveries in the context of penalized regression; it aims to provide more specific information about a model’s individual variable selections and their likelihood of being a false a discovery. Mirroring the “F”/“f” notation that is traditionally used to denote cumulative distribution functions (CDFs) and probability density functions (PDFs), a convention that is also used in the false discovery rate literature, we refer to this method as “mfdr”, or local marginal false discovery rates. The development of mfdr showcases several advantages of model based approaches for high-dimensional data analysis while also illustrating a connection the approach shares with large-scale testing.

The last of these three papers, which is presented in Chapter 4, extends the aforementioned mFdr and mfdr approaches to the grouped variable setting. Here features belong to predetermined groups and penalization/selection occur at the
group level. Some important examples include categorical variables which are represented by groups of indicator variables, polynomial or basis expansions of features presumed to have non-linear effects, or scientifically related features that share a common functional relationship. While this chapter focuses on group level selections, an assessment of the individual members of these groups is an interesting direction for future research.

1.1 Large-scale Testing

This manuscript centers around the analysis goal of discovering associations between an outcome variable, $Y$, and explanatory features, $X_1, X_2, \ldots, X_p$, while being mindful of making false discoveries. Building a reliable model capable of simultaneously assessing a large number of explanatory variables can be difficult, particularly in high-dimensions. In simple terms, when $p > n$, there isn’t enough data for most traditional methods, such as maximum likelihood or ordinary least squares, to uniquely estimate the effect of each individual variable. A straightforward and reliable alternative is to break up the single high-dimensional problem into a large number of low dimensional problems.

Let $X$ denote the $n$ by $p$ matrix of explanatory features such that the element $x_{ij}$ corresponds to the value of the $j^{th}$ variable for the $i^{th}$ observation, and let $y$ denote the vector of length $n$ of response values. Large-scale testing focuses on each feature’s univariate relationship with $Y$. For a continuous, normally distributed response variable the approach uses $p$ separate simple linear regression models:

$$ y_i = \alpha_j + \beta_j x_{ij} + \epsilon_{ij} $$

$$ \epsilon_{ij} \sim N(0, \sigma_j^2) $$

For other distributions of $Y$ the appropriate analogs (logistic regression, etc.) can be used. The major appeal of large-scale testing is that inferential tools are well-defined for these classical regression models. These tools can be used to obtain
a collection of \( p \) test statistics, \( \{t_1, \ldots, t_j, \ldots, t_p\} \), and \( p \)-values, \( \{p_1, \ldots, p_j, \ldots, p_p\} \), corresponding to each estimate \( \hat{\beta}_j \). The dimensionality of the data only manifests itself in the special care that needs to be paid to the fact that a large number of tests were conducted. However, the statistical concerns of testing multiple hypotheses are well-established, the following two subsections cover some of the most popular approaches to multiple testing.

1.1.1 Family-wise Error Control

When conducting multiple hypothesis tests the primary concern is often controlling the Type I error rate. A Type I error is defined to be the incorrect rejection of a true Null Hypothesis, or in other words a false positive. Methods of family-wise error rate (FWER) control consider a collection of multiple hypothesis tests and seek to control the probability of making at least one Type I error at level \( \alpha \).

The simplest approach to FWER control is the Bonferroni correction, which can be framed as an adjustment to the rejection threshold for each individual \( p \)-value such that a Null Hypothesis is rejected when \( p_j \leq \alpha / h \), where \( h \) is number of hypotheses that are tested. This more stringent rejection rule will control the FWER at level \( \alpha \). Other, more powerful, methods such as those of Sidak, Tukey, Holm, and Scheffe (Hsu, 1996) can also be used to control the FWER; because the focus of this manuscript is on false discovery rates and these other methods of FWER control will not be discussed. The Bonferroni method provides a simple illustration of FWER control, and is a suitable example for motivating the advantages of false discovery rate approaches. FWER, particularly in high-dimensions, is a very demanding condition to satisfy. An overall 5% chance of making a single Type I error when conducting tens of thousands of tests is highly restrictive; presumably some false discoveries could be tolerable when looking at many thousands of variables. These concerns are addressed by the less restrictive, false discovery rate (Fdr) approach to
managing the amount of Type I errors when conducting a large number of hypothesis tests.

1.1.2 False Discovery Rates

To clarify the difference between FWER and Fdr, let $h_0$ denote the number of true Null Hypotheses and $h_1$ denote the number of false Null Hypotheses for a set of $h = h_0 + h_1$ hypothesis tests. We can represent the outcomes of the these tests using a table (Table 1.1), where the columns represent the data analyst’s decision to reject or not to reject the Null Hypothesis, and the rows represent the true state of the hypothesis:

<table>
<thead>
<tr>
<th></th>
<th>Fail to Reject</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null is True</td>
<td>$h_0 - V$</td>
<td>$V$</td>
<td>$h_0$</td>
</tr>
<tr>
<td>Null is False</td>
<td>$h_1 - U$</td>
<td>$U$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$h - R$</td>
<td>$R$</td>
<td>$h$</td>
</tr>
</tbody>
</table>

Table 1.1: A summary of the potential outcomes of $h$ hypothesis tests for a given rejection rule. Here $h$ and $R$ are known while $h_0$ and $h_1$ are unobserved.

Traditionally the content of Table 1.1 has been approached “horizontally”, with Type I error being related to $V$ and power being based upon $U/h_1$. An attractive alternative is to approach the table “vertically” by considering $V/R$, which is the ratio of true Null Hypotheses that are deemed Non-Null relative to the total number of hypotheses deemed Non-Null. The expected value of this quantity, $E(V/R)$, is formally defined as the false discovery rate (Fdr). Throughout this manuscript we define the case where $R = 0$ to correspond with an Fdr of 0. This is
a nuance that is common to most, but not all, approaches to false discovery rates; it is often irrelevant in high-dimensional applications where rejecting exactly zero hypotheses is very unlikely.

The popularity surrounding Fdr originates from the work of Benjamini and Hochberg (1995), who first provided a simple Fdr control procedure in a seminal paper that remains one of the most highly cited publications in all of statistics. The Benjamini-Hochberg (BH) procedure is a two stage, step-up procedure that controls the Fdr at level $\alpha$. The first step is to order the $p$-values from smallest to largest; these ordered $p$-values are denoted by $p_{(1)}, p_{(2)}, \ldots, p_{(h)}$. The second is to step through the ordered $p$-values and find the largest index $k$ such that $p_{(k)} \leq \frac{k}{h} \alpha$. The procedure rejects the Null Hypotheses corresponding to $(1), \ldots, (k)$, and the entire set of rejected features is expected to contain no more than $\alpha k$ false discoveries. The BH procedure also suggests a simple correction to $p$-values: $p_{(k)} \frac{k}{h} \leq \alpha$, which can be used for Fdr control in a manner similar to the how the Bonferroni correction is used.

A useful approach to understanding the BH procedure is to view the observed $p$-values from large-scale testing as a two-component mixture:

$$g(p) = \pi_0 g_0(p) + (1 - \pi_0) g_1(p)$$

Here $g$ is the density of the $p$-value mixture, $g_0$ is the density under the Null Hypothesis, which by definition is $U(0, 1)$, and $\pi_0$ is the fraction of Null Hypotheses which are true. The mixture can be equivalently expressed using cumulative distribution functions (CDFs) instead of densities:

$$G(p) = \pi_0 G_0(p) + (1 - \pi_0) G_1(p) = \pi_0 p + (1 - \pi_0) G_1(p)$$

Using CDFs allows for a formal definition of tail area-based false discovery rates: $\text{Fdr}(p) = \pi_0 p / G(p)$, a quantity which is closely related to steps of
the BH procedure. To formalize the connection with the BH procedure, define order\((p_j)\) to be ordering position of the \(j^{th}\) feature’s \(p\)-value in the ascending sequence \(\{p_{(1)}, p_{(2)}, \ldots, p_{(h)}\}\). Using the empirical CDF, \(\tilde{G}(p_j) = \text{order}(p_j)/h\), we have:

\[
\tilde{F}_{\text{dr}}(p_j) = \pi_0 p_j h / \text{order}(p_j) \leq p_j h / \text{order}(p_j)
\]

The right hand side of this inequality is the exact same quantity that results from the BH “correction”. Thus indicating that controlling \(\tilde{F}_{\text{dr}}\) at level \(\alpha\) is equivalent to controlling \(F_{\text{dr}}\) using the BH rule. Motivated by this result, the next subsection provides a more detailed development of \(F_{\text{dr}}\) from the Bayesian perspective. This development will be based upon the test statistics themselves rather than \(p\)-values, which provides a smoother transition into how these concepts can be applied to penalized regression models.

1.1.3 A Bayesian Derivation of \(F_{\text{dr}}\)

In the years immediately following the initial development of BH procedure false discovery rate approaches to multiple testing rapidly gained in popularity. It was soon shown that \(F_{\text{dr}}\) could be characterized using a Bayesian perspective (Storey, 2002; Efron and Tibshirani, 2002) by using arguments similar to those made at the end of the previous subsection. To formalize this Bayesian approach, let \(z_j\) denote the normalized test statistic for the \(j^{th}\) feature. These normalized statistics can be obtained retroactively from a \(p\)-value or from an original test statistic following some other null distribution such that \(z_j = \Phi^{-1}(F_t(t_j))\), where \(\Phi\) is the standard normal CDF. We then define:

\[
\pi_0 = Pr(\text{Null}) \quad f_0(z) \text{ as the density of } z \text{ for null features}
\]
\[
\pi_1 = Pr(\text{Non-null}) \quad f_1(z) \text{ as the density of } z \text{ for non-null features}
\]
By construction, the density $f_0$ is the standard normal distribution and in most high-dimensional applications it is reasonable to presume $\pi_0$ is close to 1. It is worth mentioning here that the BH procedure implicitly uses $\pi_0 = 1$, which results in the procedure providing conservative control over the false discovery rate. Letting $\mathcal{Z}$ be any subset of the real line, we define:

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z) dz \quad F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z) dz$$

Consider $z_j \in \mathcal{Z}$ and suppose we are interested in whether or not this feature belongs to the null or non-null group. For the members of $\mathcal{Z}$, Bayes rule suggests:

$$\Pr(\text{Null} \mid z_j \in \mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{\pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z})} = F_{\text{dr}}(\mathcal{Z}) \quad (1.1.1)$$

The denominator of this expression is referred to as the mixture distribution and is denoted by $F$. In similar fashion we define the mixture density, denoted by $f$, to be: $\pi_0 f_0(z) + \pi_0 f_1(z)$. The expression in (1.1.1) provides a tail-area Fdr characterization of the collection of features with test statistics contained in the region $\mathcal{Z}$, thus enabling control over Fdr via the choice of $\mathcal{Z}$. This formulation of tail area-based Fdr gives rise to a localized approach for the false discovery rate of any particular Null Hypothesis and the corresponding statistic $z_j$.

### 1.1.4 Local False Discovery Rates

Suppose we take $\mathcal{Z}$ to be the single point $z_j$, it follows directly from the arguments of the previous section that:

$$\Pr(\text{Null} \mid z_j = z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + \pi_1 f_1(z)} = f_{\text{dr}}(z) \quad (1.1.2)$$

Here $f_{\text{dr}}(z)$ is defined as the local false discovery rate (Efron et al., 2001; Efron and Tibshirani, 2002), and refers to the specific probability that the $j$th feature is a false discovery. To summarize the distinction between Fdr and fdr, the Fdr result of Equation 1.1.1 can be used to quantify the reliability of the set of features.
determined by \( Z \), while the \( fdr \) result of Equation 1.1.2 can be used to quantify the reliability of a single feature. As one would expect, \( Fdr \) and \( fdr \) are linked in several ways, one of which is through the relationship:

\[
Fdr(Z_0) = E\left(fdr(z) | z \in Z_0\right) \text{where } Z_0 = [z_0, \infty)
\]

This result implies that the \( Fdr \) of the set of features with normalized test statistics exceeding \( z_0 \) is equal to the average \( fdr \) of features beyond that threshold, which ensures that selecting individual features using a threshold \( fdr(z) < \alpha \) also limits \( Fdr \) below \( \alpha \) for the entire set of features selected. This is an important relationship, and one that will also hold for the model-based methods we will propose in Chapters 2-4 of this manuscript.

One fortunate aspect of Equations 1.1.1 and 1.1.2 is that their numerators are known (provided we take \( \pi_0 = 1 \)); which means that we only require estimates of \( F \) or \( f \) in order to estimate \( Fdr \) or \( fdr \). When \( p \) is large, the empirical estimators, \( \hat{F} \) and \( \hat{f} \), are very good approximations, providing an empirical Bayes method for estimating \( Fdr \) and \( fdr \) in high-dimensional settings.

It is worth mentioning that, in practice, the theoretical null density, \( f_0 \), can be incorrectly specified; some examples include correlations amongst test statistics or unobserved covariates. These scenarios can lead to overly optimistic or pessimistic \( p \)-values and theoretical null densities that are too narrow or wide. There has been substantial amount of work to address this problem using empirical null modeling, which estimates \( f_0 \), typically as a normal density with its location and scale derived from the observed data using the assumption that a large fraction of hypotheses are null. Methods of empirical null modeling are employed automatically in the R packages “locfdr” (Efron et al., 2015) and “fdrtool” (Strimmer, 2008a) when violations of the theoretical null are detected. Empirical null modeling is not currently adopted in any of our work, however it might be an interesting area for future work.
1.2 Penalized Regression Models

As detailed in the previous section, large-scale testing approach to high-dimensional data analysis offer well-developed Fdr and fdr methods, however they come with the downside of using a separate model or test for each feature. Many regression models capable of handling high-dimensional data have been developed, often through the use of penalization. These penalized models provide the starting point for an alternative, and potentially more informative, approach for determining which features are related with the outcome variable.

Generally speaking, regression modeling simultaneously relates the set of explanatory features in data matrix, $X$, with the outcome, $y$, using a probability model involving coefficients $\beta$. Throughout the remainder of this manuscript we assume that the columns of $X$ have been standardized such that each variable has a mean of 0 and $\sum_i x_i^2 = n$, in the case of linear regression we also assume that $y$ is centered which removes the need for an intercept. These standardizations are necessary to ensure that penalization is distributed equally to each feature regardless of its units. After the model has been fit it is possible to unstandardize and obtain the regression coefficients, and the intercept, on their original scales; this is typically done automatically by most popular software packages.

The fit of a regression model can be summarized using the log-likelihood, denoted by $\ell(\beta|X,y)$. Often it is convenient to express the model in terms of the linear predictor $\eta = X\beta$, such that: $f(\eta|y) = \ell(\beta|X,y)$. It is important to recognize that these functions are equivalent, but their derivatives are not. Many of the derivations in upcoming the chapters will be done in the general likelihood setting, however the applications and numerical results will focus specifically on three common likelihood-based models:

1. **Linear Regression**: where $Y$ is normally distributed and $\ell(\beta|X,y) \propto (y - X\beta)^T(y - X\beta)$
2. Logistic Regression: where $Y$ follows a Bernoulli distribution and $\ell(\beta|X, y) \propto \sum_i (y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i))$ and $\pi_i = \exp(x_i^T\beta)/(1 + \exp(x_i^T\beta))$

3. The Cox Regression: where $Y$ is a time-to-event outcome, and the Cox partial likelihood is denoted by $\ell(\beta|X, y) = \sum_{i=1}^n (d_i x_i^T\beta - \sum_{k \in R_i} \exp(x_k^T\beta))$, in this setting $y$ and the rows $X$ have been arranged in order of ascending failure times, $d_i = 1$ indicates an observed event and $d_i = 0$ identifies right censoring, and $R_i$ denotes the set of observations at risk when observation $i$ occurs.

For each of these models, the classical approach to estimating $\beta$ is based upon maximizing $\ell(\beta|X, y)$. However, this approach is unstable when $p > n$ unless an appropriate penalty is imposed on the size of $\beta$. The instability lies in the attempt to uniquely estimate more parameters than there are available data points. By analogy, one could imagine trying to uniquely estimate the slope and intercept of a simple linear regression model where the data is a single point. Infinitely many lines will fit the point perfectly.

The remainder of this section provides the details for a few popular penalization schemes. The penalized regression literature is vast, and this section is not intended to be a comprehensive review, but rather a detailed introduction to a handful of important penalties that are either frequently used, historically significant, or are relevant to the work of future chapters.

1.2.1 Ridge Regression

One of the earliest, most well-studied, penalized regression methods is ridge regression. The approach originates in linear regression setting where the maximum likelihood estimate of $\beta$ is equivalent to the least squares estimate $(X^TX)^{-1}X^Ty$. It gets its name from its solution $\hat{\beta}_{\text{ridge}}^* = (X^TX + \lambda I)^{-1}X^Ty$, which is very similar to the ordinary least squares solution but with a “ridge” added to diagonal of $X^TX$. This ridge corresponds to imposing a penalty on the squared L2 norm of $\beta$,
denoted by \( ||\beta||^2 = \sum_j \beta_j^2 \), and the ridge solution can be defined by the optimization problem:

\[
Q(\beta | X, y) = -\frac{1}{n} \ell(\beta | X, y) + \lambda ||\beta||^2
\]

Here \( \lambda \) is a tuning parameter which dictates the degree of penalization. The L2 penalization in ridge regression leads to biased coefficient estimates which are shrunken towards zero, however it does not perform variable selection, and all features are included in the model with non-zero coefficient estimates. These tendencies may seem undesirable, but it is well-established that for at least one value of \( \lambda \) the ridge regression estimates of \( \beta \) will have smaller mean squared error (MSE) than the unpenalized estimates. Additionally, ridge regression provides stable estimates even when \( p > n \). These properties make ridge regression an attractive tool, and an approach worth mentioning when discussing high-dimensional regression modeling, however the primary focus of this manuscript will be on other forms of penalized regression which also perform variable selection.

### 1.2.2 The lasso

While ridge regression imposes a penalty on the squared L2 of \( \beta \), an alternative is to instead impose a penalty on the L1 norm of \( \beta \), denoted by \( ||\beta||_1 = \sum_j |\beta_j| \). This approach was first formalized by Robert Tibshirani in 1996 and named the “least absolute shrinkage and selection operator”, or the lasso (Tibshirani, 1996).

Generally speaking, lasso regression combines an L1 penalty with the log-likelihood such that the lasso estimator, denoted by \( \hat{\beta} \), is defined by:

\[
\hat{\beta} = \arg \min_{\beta} \left\{ Q(\beta) = -\frac{1}{n} \ell(\beta | X, y) + \lambda ||\beta||_1 \right\} \tag{1.2.1}
\]

As indicated by its name, the lasso penalty offers two major benefits, it provides both shrinkage and selection. Much like ridge regression, the lasso yields
biased coefficient estimates which are shrunken towards zero, and in doing so has
the ability to achieve smaller MSE than unpenalized methods. Unlike ridge
regression, the lasso naturally performs variable selection, yielding “sparse” solutions
where a subset of the regression coefficients are estimated to be exactly zero. The
degree of sparsity is governed by the tuning parameter \( \lambda \), with larger values of \( \lambda \)
resulting in more sparse models that contain fewer non-zero coefficient estimates.
Sparsity is particularly attractive in high-dimensions, and is commonly used as an
implicit assessment of which features are important.

The ability of the lasso to achieve sparse solutions is most easily understood
by formulating Equation 1.2.1 as the constrained optimization problem:

\[
\hat{\beta} = \arg \min_{\beta} \left\{ -\frac{1}{n} \ell(\beta | X, y) \right\} \text{ such that: } ||\beta||_1 \leq t
\]

When \( p = 2 \) the constraint region defined by \( t \) is a square, we can plot this
region against the contours of the likelihood in order to visualize the behavior of the
lasso solution. For a given value of \( t \) the lasso solution occurs at the first place where
the likelihood contours touch the constraint region. In the simulated example shown
in Figure 1.1 we see the tendency for this to occur at corners, which corresponds to
an exact zero estimate of \( \beta_1 \) for this example.

While sparsity is an alluring property, a practical challenge of the lasso is the
non-differentiability of the L1 penalty. In classical statistical theory the estimates
of \( \beta \) are typically found by differentiating the log-likelihood with respect to \( \beta \) to
obtain a score function, then setting that score function equal to 0 and solving. This
idea can be easily extended to penalized likelihoods by differentiating the objective
function \( Q(\beta) \), which includes both the log-likelihood and a penalty function, to
arrive at the set of penalized score equations. For ridge regression this is very
straightforward since the log-likelihood and the penalty term are both differentiable
everywhere. However, for the lasso, the penalty term is only semi-differentiable,
Figure 1.1: A 2-dimensional illustration of the lasso solution, demonstrating its tendency for sparse solutions. Here the constraint region defined by $||\beta||_1 \leq t = 1$ is shaded in yellow. This region first meets the contours of the likelihood when $\hat{\beta}_2 = 1$ and $\hat{\beta}_1 = 0$.

specifically it is non-differentiable at zero. Consequently differentiation of the lasso penalty with respect to $\beta$ must be characterized by a subdifferential. A simple example illustrating this concept is the subdifferential of $|a|$, which is defined:

$$
\partial |a| = \begin{cases} 
-1, & \text{if } a \text{ is less than 0} \\
[-1, 1], & \text{if } a \text{ equals 0} \\
1, & \text{if } a \text{ is greater than 0}
\end{cases}
$$

The necessity of subdifferentials for the lasso objective function requires that the penalized score equations of the lasso accommodate each of these cases. The resulting set of equations are known as the Karush-Kuhn-Tucker (KKT) conditions, and are necessary and sufficient in characterizing the solution of any convex optimization problem (such as the lasso). For the lasso penalty these conditions require
(Tibshirani, 2013):

\[
\frac{1}{n} u_j(\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \quad \text{if } \hat{\beta}_j \neq 0
\]

\[
\frac{1}{n} u_j(\hat{\beta}) \in [-\lambda, \lambda] \quad \text{if } \hat{\beta}_j = 0
\]

for j \in \{1, \ldots, p\}, where \( u_j \) is the \( j^{th} \) element of the unpenalized score function, \( u(\beta) = \nabla \ell(\beta|X,y) \).

Conceptually the lasso KKT conditions are most easily understood in the linear regression setting where \( u_j = x_j^T(y - X\hat{\beta}) \). Recognizing the standardization of \( X \), we see that the KKT conditions of linear regression mandate that the correlation between a predictor and the current residual must reach a certain threshold, defined by \( \lambda \), in order for that variable to become active in the model. When this correlation is less than \( \lambda \) the variable’s coefficient estimate must be exactly zero.

1.2.3 The Elastic Net

The lasso is one of the most significant developments in the realm of high-dimensional regression modeling, however it does have some limitations. One limitation is that in the \( p > n \) scenario the lasso can only select at most \( n \) variables before reaching saturation. Another limitation is that in situations where groups of variables have very high pairwise correlations the lasso tends to select only a single one of these variables from the group. The elastic net is capable of overcoming both of these limitations by combining the L2 penalty of ridge regression with the L1 penalty of the lasso. The combination provides shrinkage in way that encourages the selection of all members of correlated groups of variables, while still providing a sparse solution. This tendency led its creators to choose the name “elastic net”, as they describe: “it is like a stretchable fishing net that retains all the big fish” (Zou and Hastie, 2005).
Formally the elastic net solution is described by the optimization problem:

\[
\hat{\beta} = \arg \min_{\beta} \left\{ Q(\beta) = -\frac{1}{n} \ell(\beta|X, y) + \alpha \lambda \|eta\|_1 + (1 - \alpha) \frac{\lambda}{2} \|eta\|_2 \right\} \tag{1.2.3}
\]

Using this parametrization, \( \lambda \) represents the total amount of penalization, and \( \alpha \) controls the balance between the lasso and ridge penalties. It is easy to see that the lasso is a special case of the elastic net where \( \alpha = 1 \), and ridge regression is a special case of the elastic net when \( \alpha = 0 \). It is also common to parametrize the elastic using \( \lambda_1 \) and \( \lambda_2 \):

\[
Q(\beta) = -\frac{1}{n} \ell(\beta|X, y) + \lambda_1 \|eta\|_1 + \frac{\lambda_2}{2} \|eta\|_2
\]

Like the lasso, the L1 penalty of the elastic net necessitates the use of subdifferentials in characterizing its solution, the resulting KKT conditions for elastic net solution are:

\[
\frac{1}{n} \hat{u}_j(\hat{\beta}) - \lambda_2 \hat{\beta}_j = \alpha \lambda \text{ sign}(\hat{\beta}_j) \quad \text{if } \hat{\beta}_j \neq 0
\]

\[
\frac{1}{n} \hat{u}_j(\hat{\beta}) \in [-\lambda_1, \lambda_1] \quad \text{if } \hat{\beta}_j = 0 \tag{1.2.4}
\]

The condition for variable selection under the elastic net penalty appears to be very similar to that of the lasso, however the actual selections and their coefficient estimates can be quite different, particularly when there is high correlation between features. The ridge penalty in the elastic net induces a “grouping effect”, meaning that the coefficient estimates of highly correlated variables tend to nearly equal. More formally, suppose that two features, \( x_j \) and \( x_k \), have sample correlation \( \rho \) and are both active in an elastic net model with coefficient estimates \( \hat{\beta}_j, \hat{\beta}_k > 0 \). Define the standardized distance between these coefficients as:

\[
D_{\lambda, \alpha} = \frac{\sqrt{n} |\hat{\beta}_j - \hat{\beta}_k|}{||y||_2}
\]

This distance is bounded by:

\[
D_{\lambda, \alpha} \leq \frac{\sqrt{2(1 - \rho)}}{(1 - \alpha) \lambda}
\]
This result implies that the distance between the coefficient estimates of any two features is bounded by their pairwise correlation, in the extreme scenario where \( \rho = 1 \) both coefficient estimates will be identical. The lasso does not have a grouping property and instead has a tendency to only select a single representative from correlated groups of features. Furthermore, the representative can often be somewhat arbitrary and change frequently when the data is bootstrapped or subsetted (Zou and Hastie, 2005). These differences have the practical implication that the variable selections made by the lasso and the elastic net carry different interpretations.

1.2.4 Non-convex Penalties

The sparsity induced by the lasso or elastic net penalties can be highly desirable. However, each of these penalties yield a biased estimator with bias that does not disappear as \( n \to \infty \). Non-convex penalties are capable of achieving both variable selection and unbiased estimation by tapering off the degree of penalization applied to large coefficient estimates. Consider the general penalized likelihood problem:

\[
Q(\beta|X, y) = -\frac{1}{n} \ell(\beta|X, y) + \sum_j P(\beta_j|\lambda, \gamma)
\]

Here the penalty function, \( P \), depends on two tuning parameters \( \lambda \) and \( \gamma \). The focus of this section will be on the minimax concave penalty, or MCP, but there are many other non-convex penalties with the smoothly clipped absolute deviations (SCAD) penalty being one of the earliest and most influential. For a fixed value of \( \gamma > 1 \), the minimax concave penalty function is represented by:

\[
P(\beta|\lambda) = \begin{cases} 
\lambda|\beta| - \frac{\beta^2}{2\gamma}, & \text{if } |\beta| \leq \gamma\lambda \\
\frac{1}{2}\gamma\lambda^2, & \text{if } |\beta| > \gamma\lambda
\end{cases}
\]

The behavior of MCP is best understood by looking at its derivative, which
represents the rate of penalization as $x$ changes:
\[
\frac{\partial}{\partial \beta} P(\beta|\lambda) = \begin{cases} 
(\lambda - |\beta|/\gamma)\text{sign}(\beta), & \text{if } |\beta| \leq \gamma \lambda \\
0, & \text{if } |\beta| > \gamma \lambda
\end{cases}
\]

Consider the case where $\beta = 0$, in this situation MCP applies the exact same rate of penalization as the lasso, $\lambda$. As $|\beta|$ increases, the rate of penalization smoothly decreases until reaching zero when is large enough, $|\beta| > \gamma \lambda$. This relaxation allows for MCP to both select variables and achieve the unbiased estimation of sufficiently large effects, a result known as the *oracle property*. The parameter $\gamma$ controls how quickly the penalization rate goes to zero; as $\gamma \to \infty$, MCP approaches the lasso penalty. Conversely, when $\gamma$ is very small the coefficient estimates become unbiased very quickly, however these estimates also tend to be unstable due to the greater concavity of the penalty function pushing the entire objective function towards non-convexity.

In comparison with the lasso, non-convex penalties tend towards greater sparsity while having comparable predictive ability. Non-convex penalties are particularly effective when a small number of features have large signals. In these situations, methods like cross validation tend to favor lasso models with smaller values of $\lambda$ that allow for the larger effects; however these large effects often come at the cost of allowing more noise variables into the model (a frequent consequence of small values of $\lambda$). This tradeoff isn’t necessary with penalties like MCP where a smaller $\lambda$ isn’t necessary to accommodate large effects.

1.2.5 The Group lasso

In certain circumstances it is preferable to consider imposing sparsity on groups of variables rather than the individual variables themselves. For example categorical predictors are often expressed using groups of dummy indicator variables and it may be desirable for all of these indicators to be selected together since
the group represents a single underlying characteristic. Variables can belong to
groups based upon prior knowledge, for example genetic features can be grouped
using known gene pathways, and in some applications it might make sense to for
the regression model to select entire gene pathways rather than individual genes.
The group lasso behaves analogously to the ordinary lasso: it achieves sparsity at
the group level through its tendency to set entire groups of coefficients to exactly
zero.

To formally introduce the group lasso, let \( \{X_1, X_2, \ldots, X_J\} \) represent non-
overlapping portions of the design matrix corresponding to groups \( j \in \{1, \ldots, J\} \)
such that \( X_j \) is an \( n \) by \( K_j \) matrix whose columns contain the features in group \( j \).
The group lasso estimates are found by minimizing the objective function:

\[
Q(\beta|X, y) = -\frac{1}{n} \ell(\beta|X, y) + \sum_j \lambda_j ||\beta_j||_2
\]

(1.2.5)

The parameter \( \lambda_j \) controls the degree of penalization applied to each group of
coefficients, denoted by \( \beta_j \), and is allowed to vary in accordance to group size with a
common choice being \( \lambda_j = \sqrt{K_j} \lambda \). The group lasso exploits the non-differentiability
of \( ||\beta_j||_2 \) at \( \beta_j = 0 \) to achieve sparsity at the group level.

The group lasso was originally proposed under the assumption of orthonormal-
mality within each group (ie: \( 1/nX_j^TX_j = I \)). Groups that are non-orthonormal can
be orthonormalized such that \( X_j = \tilde{X}_jR_j \) and \( \frac{1}{n}\tilde{X}_j^T\tilde{X}_j = I \) using a singular value
decomposition of \( \frac{1}{n}X_j^TX_j \). Provided that that \( K_j < n \) for all \( j \), any optimization
performed on the orthonormal scale is equivalent to optimization on the original
scale. The solution on the orthonormal scale, which we denote by \( \hat{\theta}_j \), can easily be
converted back to the original scale by \( \hat{\beta}_j = R_j^{-1}\hat{\theta}_j \).

Under group orthonormalization, the KKT which characterize the group lasso
solution are:

$$\frac{1}{n} u_j(\hat{\beta}) = \lambda_j s_j, \quad \text{where: } s_j = \begin{cases} \frac{\hat{\beta}_j}{\|\hat{\beta}_j\|_2}, & \text{if } \hat{\beta}_j \neq 0 \\ v : \|v\|_2 \leq 1, & \text{if } \hat{\beta}_j = 0 \end{cases}$$  \hspace{1cm} (1.2.6)

Where $u_j$ denotes the vector of unpenalized score components, in terms of $\hat{\theta}$ and $\tilde{X}$, corresponding to the features in group $j$.

The KKT conditions under group orthonormalization for the linear regression setting reveal an interesting connection between the group lasso and the universally most powerful invariant (UMPI) test for testing a group of variables in classical linear model theory. Classically this test rejects the null hypothesis that $\beta_j = 0$, at significance level $\alpha$, if:

$$\|P_{X_j}r_0\|_2^2 \geq \sigma^2 \chi^2_{K_j,1-\alpha}$$  \hspace{1cm} (1.2.7)

Where $P_{X_j}$ is the orthogonal projection operator onto the column space of $X_j$, and $r_0$ is the vector of residuals calculated without variables contained in $\beta_j$.

For orthonormal groups, $P_{X_j} \propto \tilde{X}\tilde{X}^T$ and $\|\tilde{X}\tilde{X}^T r_0\|_2^2 \propto \|\tilde{X}^T r_0\|_2^2$, thus making the group lasso selection condition, $\|\tilde{X}^T(y - \tilde{X}\hat{\theta})\|_2^2 = \lambda_j$, expressed in the KKT conditions in Equation 1.2.6 essentially equivalent to the UMPI $\chi^2$ test provided $\lambda_j$ accounts for group size using $\sqrt{K_j}$.

Numerous variants of the group lasso exist, including the overlapping group lasso (Jacob et al., 2009) where variables are allowed to be members of multiple groups, and the sparse group lasso (Simon et al., 2013) where an additional L1 penalty is used to impose sparsity at the individual feature level for selected groups.

1.3 Inference for Penalized Regression Models

The usual regression modeling inference on $\hat{\beta}$ is complicated for penalized regression models due to their coefficient estimates being biased towards zero, and their sampling distributions containing point masses at zero. Several approaches
to inference have been proposed, some attempt to obtain de-biased estimates using the penalized regression results as a starting point, while others perform inference directly on the penalized model. Many of these approaches are computationally intensive, even for small data sets, and are particularly burdensome to carry out if several different models are being used. This section will focus on two particular approaches, sample splitting and selective inference, which will be compared with our methods in future chapters. It will also provide a cursory overview of several other proposed methods whose results do not easily translate to conventional false discovery rates.

1.3.1 Sample Splitting

A conceptually simple method of obtaining regression based $p$-values for sparse high-dimensional data is to use a sample splitting approach which randomly separates the data into two disjoint parts, using the first part to perform variable selection, and then using the second part to perform inference. This procedure avoids the over-optimism induced by using the data to select variables and then ignoring variable selection step when performing inference (Wasserman and Roeder, 2009). While statistically valid, carrying out such a procedure is very sensitive to the specific split, with different splits yielding potentially different results. Repeating the procedure many times will reduce this uncertainty; however, it introduces a new challenge regarding how to handle variables that are selected in some splits but not others, thereby motivating the multi-split approach. Formally, the approach is defined (Meinshausen et al., 2009; Dezeure et al., 2015):

1. Randomly split the data into two sets $I_1$ and $I_2$ such that $I_1 \cup I_2 = \emptyset$. Typically these sets are of equal size.

2. Fit a penalized regression model using $I_1$ to select a subset of the available variables: $\hat{S} \subset \{1, \ldots, p\}$. 
3. Fit an unpenalized regression model using $I_2$ with only the variables in $\hat{S}$.

4. Compute the unadjusted $p$-values, $p_j^{raw}$, using the classical regression based inferential methods for variables in $\hat{S}$; for $j \notin \hat{S}$ assign $p_j^{raw} = 1$.

5. Correct $p_j^{raw}$ for multiple testing.

6. Repeat steps 1-5 $B$ times to obtain a collection of $p$-values for each variable: $p_j^{(1)}, \ldots, p_j^{(B)}$ for $j \in \{1, \ldots, p\}$.

7. Aggregate the collections in step 6 by taking an appropriate sample quantile.

More repetitions (a larger $B$) adds stability and reproducibility to the multi-split method. A potential downside of this approach is that it relies on either an underlying sparsity structure with the number of non-zero effects being fewer than $n/2$, or alternatively a beta-min assumption which defines the set $S_0$:

$$\min_{j \in S_0} |\beta_j| \geq \beta_{\min}$$

These screening assumptions are necessary in order for the variable selection procedure of step 2 to provide to asymptotically consistent selection of the true active set. The main appeal of the multi-split method is that it is extremely general and can easily be applied to GLM and Cox Regression models. In contrast, many other methods that have been developed in the high-dimensional linear regression framework do not easily extend to more general settings. Consequently, the multi-split method will frequently be used as a competitor to our methods in later sections.

1.3.2 Selective Inference

While sample splitting approaches use separate portions of data to first select variables and then conduct inference, selective inference methods (Tibshirani et al., 2016) use the same data for both selection and inference. Selective inference approaches are able to avoid overly optimistic results by considering the relevant Null
Hypotheses to be random. Formally, the goal of selective inference is, after having selected a model $M$ based upon the data, to test the hypothesis $\hat{H}_0$. The proper rejection region for this test needs to condition upon $M$ and $\hat{H}_0$.

Suppose $y \sim N(\mu, \Sigma)$, as is the case in the linear regression setting. For the lasso it can be shown that any variable selection event can be expressed in terms of:

$$\{y : Ay \leq b\}$$

Where $A$ and $b$ do not depend upon $y$.

While selective inference methods can be applied to the lasso, they are most easily understood using forward stepwise regression, a familiar variable procedure is related to lasso regression. After running a forward stepwise selection procedure for $k$ steps, $\{y : Ay \leq b\}$ is the set of possible $y$ vectors that would yield an identical set of selected predictors and signs at each step of the procedure. In forward selection, each of these steps can be thought of as a competition between each of candidate predictors involving their inner product: $\langle x_j, y \rangle$. If the first predictor “wins” this competition, it is known that $|\langle x_1, y \rangle| > |\langle x_j, y \rangle|$ for all $j \neq 1$. When conducting inference regarding $\beta_1$, the selective inference approach conditions on the set of $y$ vectors that satisfy: $|\langle x_1, y \rangle| > |\langle x_j, y \rangle|$ for all $j \neq 1$. It turns out that this set can expressed in the form $Ay \leq b$. For any vector $\eta$:

$$\eta^T y | \{Ay \leq b\} \sim \text{Truncated Normal} \quad (1.3.1)$$

Unconditionally, $\eta^T y$ would ordinarily follow a normal distribution, however conditioning on the selection process that lead to the current model changes this to a truncated normal distribution. This result allows for hypothesis testing, and confidence interval construction, that properly account for the uncertainty introduced by model selection. As previously mentioned, a lasso model at a given value
of $\lambda$ can be expressed as $\{y : Ay \leq b\}$, which enables the use of selective inference methods for hypothesis testing.

A natural question with selective inference is “when to stop adding variables?”. G’Sell et al. (2016); Taylor and Tibshirani (2015) describe stopping rules for sequential hypothesis tests which can be used to control the FWER or Fdr in the context of selective inference. Focusing on Fdr, the rule is similar in spirit to the BH procedure but the $p$-values are not reordered:

$$k_{\text{stop}, \alpha} = \max\{k : \frac{1}{k} \sum_{i=1}^{k} i = 1k\log(1 - p_i) \leq \alpha\}$$

Conceptually the rule leads to stopping at the last point when the average complementary log $p$-value, up to that point, is below the threshold $\alpha$. The set of variables contained in the final model obtained after $k_{\text{stop}, \alpha}$ steps has an Fdr of at most $\alpha$. These results illustrate how selective inference can used for Fdr based inference on lasso regression models.

1.3.3 Low Dimensional Projection

One recent approach to inference involves a de-biasing procedure introduced by Zhang and Zhang (2014) which can provide low-dimensional confidence intervals and $p$-values for individual parameters in high-dimensional penalized regression models. This approach is most easily motivated in the low-dimensional, $n > p$, linear regression setting, where the $j^{th}$ component of the ordinary least squares (OLS) estimator of $\beta$ can be obtained by the following two-step procedure:

1. Fit an ordinary least squares regression model on $x_j$ using the remaining variables, $X_{-j}$, as predictors

2. Then $\hat{\beta}_j^{\text{OLS}} = \frac{y^T \gamma_j}{x_j^T \gamma_j}$, where $\gamma_j$ is the vector of residuals from the regression model in step 1

When $p > n$ the first step inherently leads to residual vector of exactly $0$,
however an alternative would be to use something other than OLS, for example a penalized regression model, in step 1 to obtain $\gamma_j$. In general this will result in step two yielding the quantity:

$$\frac{y^T \gamma_j}{x_j^T \gamma_j} = \beta_j + \frac{e^T \gamma_j}{x_j^T \gamma_j} + \sum_{k \neq j} \frac{x_k^T \gamma_j}{x_j^T \gamma_j} \beta_k$$  \hspace{1cm} (1.3.2)

Notice that when OLS is used in step 1 that $x_k$ is orthogonal to $\gamma_j$ for $\{k \neq j\}$ and consequently the third term in Equation 1.3.2 will be exactly zero. When using penalized regression in step 1 this term is non-zero, however it implies the bias corrected estimate: $\beta_{j}^{LDPE} = \frac{y^T \gamma_j}{x_j^T \gamma_j} - \sum_{k \neq j} \frac{x_k^T \gamma_j}{x_j^T \gamma_j} \hat{\beta}_k$, where $\hat{\beta}_k$ is estimate from the penalized regression of $y$ versus $X$. When using the lasso in step 1, it has been shown that this de-biasing approach results in $\beta_{j}^{LDPE} \sim N(\beta_j, s_j^2)$ where $s_j = \sigma z_j^T z_j / |z_j^T x_j|$, a result which allows for hypothesis tests or confidence intervals. The validity of these results relies upon a consistent estimator of $\sigma$, and the lasso estimator, $\hat{\beta}$, being close to the true $\beta$, with the formal requirement being $||\hat{\beta} - \beta||_1 = o_P(1/\sqrt{\log(p)})$ (Zhang and Zhang, 2014; van de Geer et al., 2014). From a practical perspective, this method has a steep computational burden, requiring the fitting of $p + 1$ lasso regression models and then performing cross validation on each one in order to select a suitable value of $\lambda$ for that step (Dezeure et al., 2015).

1.3.4 Other Approaches

Other approaches to inference using penalized regression models have been proposed, many which rely on resampling or subsampling techniques, with one example being the stability selection method proposed by (Meinshausen and Bühlmann, 2010). To summarize, stability selection involves taking random subsets of the data and tracking which variables are selected in a high proportion of these subsamples. Variables are deemed “stable” if they are selected more often than some threshold proportion: $\pi_{\text{thr}}$. The authors show this approach can be used to bound the number
of false discoveries present in the set of stable variables (for $\pi_{\text{thr}} > 0.5$) such that:

$$\mathbb{E}(V) \leq \frac{1}{2\pi_{\text{thr}}} \frac{q_{\Lambda}^2}{p}$$  \hspace{1cm} (1.3.3)

where $q_{\Lambda}$ is the average number of selected variables or a given penalty parameter $\lambda$, across the collection of subsamples.

In practice, it can be challenging to apply stability selection due to the computational burden of repeated subsampling, where each subsample requires the fitting of its own penalized regression model. There is also the difficulty of choosing values of $\lambda$ and $\pi_{\text{thr}}$ that achieve the Fdr control at the desired level. Additionally, the bound in Equation 1.3.3 tends to be conservative when applied to real data.

Stability selection is conceptually related to bootstrapping. However, generally speaking, standard bootstrapping methods are unable to provide consistent estimates of lasso parameters, even in situations where the lasso has $\sqrt{n}$-consistency (Chatterjee and Lahiri, 2010). A great deal of work has been done to develop modified bootstrapping procedures (Chatterjee and Lahiri, 2011; Dezeure et al., 2017), however the details of these procedures are beyond the scope of this dissertation and are fundamentally different approach that is at odds with our goal of developing computationally efficient methods that can be easily applied to a fitted model.
CHAPTER 2
MARGINAL FALSE DISCOVERY RATES FOR PENALIZED LIKELIHOOD MODELS

2.1 Abstract

The popularity of penalized regression in high-dimensional data analysis has led to a demand for new inferential tools for these models. False discovery rate control is widely used in high-dimensional hypothesis testing, but has only recently been considered in the context of penalized regression. Almost all of this work, however, has focused on lasso-penalized linear regression. In this paper, we derive a general method for controlling the marginal false discovery rate that can be applied to any penalized likelihood-based model, such as logistic regression and Cox regression. Our approach is fast, flexible and can be used with a variety of penalty functions including lasso, elastic net, MCP, and MNet. We derive theoretical results under which the proposed method is valid, and use simulation studies to demonstrate that the approach is reasonably robust, albeit slightly conservative, when these assumptions are violated. Despite being conservative, we show that our method often offers more power to select causally important features than existing approaches. Finally, the practical utility of the method is demonstrated on gene expression data sets with binary and time-to-event outcomes.

2.2 Introduction

High-dimensional data poses a challenge to traditional likelihood-based modeling approaches. Penalized regression, which can provide sparse models in which only a subset of the available features have non-zero coefficients, is an increasingly popular approach that is well suited to handle high-dimensional data. Inferential
methods for controlling the error rates of variable selection for these methods, however, have been limited, especially outside the least-squares setting.

In this manuscript, we build upon the recently proposed idea of marginal false discovery rates and extend them to the more general class of likelihood-based models, which includes generalized linear models such as logistic regression as well as Cox proportional hazards models. Our presentation focuses mainly on the most popular penalized regression method, the least absolute shrinkage and selection operator, or lasso (Tibshirani, 1996), but the methods we develop apply to many other penalties, including the minimax concave penalty (MCP; Zhang, 2010), the smoothly clipped absolute deviations (SCAD; Fan and Li, 2001), and the elastic net (Zou and Hastie, 2005).

There is an enormous literature on false discovery rate control as it applies to separately testing each individual feature one-at-a-time, but the issue is much more complex in the regression setting. A fundamental challenge of inference under variable selection is how to account for using the same data to both select features as well as to fit the model. Nevertheless, several approaches to address this challenge have been proposed.

One approach, which we refer to as sample splitting, is based upon dividing the data into two parts, using the first part for variable selection and the second part for inference. Wasserman and Roeder (2009) first proposed this approach using single split, and Meinshausen et al. (2009) extended it by considering multiple random splits. More recently, Lockhart et al. (2014) and Tibshirani et al. (2016) have proposed a family of methods known as selective inference, which test the significance of each variable selection along the lasso solution path as \( \lambda \) is decreased, conditional upon the other variables already active in the model. Based on this sequence of tests, formal stopping rules can be derived in order to control the false discovery rate at a specified level (G’Sell et al., 2016).
These approaches, however, suffer from two primary drawbacks. First, they are very restrictive, in the sense that they are often unable to select more than one or two features, even at high false discovery rates. This is especially true in high dimensions. Second, they are quite computationally intensive. For both approaches, it typically takes several orders of magnitude longer to estimate the false discovery rate than to fit the model in first place.

To address both of these drawbacks, Breheny (2018) proposed controlling the marginal false discovery rate (mFDR), and showed that focusing only on this weaker definition of false discovery yields a method that is less restrictive and far less computationally intensive than approaches for controlling conditional FDRs. Breheny (2018) considered only the case of linear regression models. Here, we extend that work to the more general class of regression models based on likelihood; in particular, this class includes generalized linear models and Cox proportional hazards models. We compare these newly developed methods to sample splitting and selective inference using simulated data and apply our approach to two case studies involving high-dimensional data, one with a time-to-event outcome and the other with a binary outcome, to demonstrate the practical utility of our method on real data.

2.3 Marginal false discovery rates

False discoveries are straightforward to define when conducting single variable hypothesis tests; a false discovery occurs when a feature $X_j$ is declared to be associated with an outcome $Y$ even though the feature is actually independent of the outcome: $X_j \perp \!\!\!\!\!\!\perp Y$. In the regression framework, where many variables are being considered simultaneously, the idea of a false discovery is more complicated. The most common approach, which we refer to as the fully conditional perspective, is to consider a feature $X_j$ to be a false discovery if it is independent of the outcome
conditional upon all other features: $X_j \perp \perp Y | X_{k \neq j}$. Penalized likelihood methods typically result in only a subset of the available variables being active in the model, thereby motivating the *pathwise conditional* perspective. This perspective focuses on the model where $X_j$ first becomes active and conditions only on the other variables present in the model (denote this set $M_j$) at that time when assessing whether or not variable $j$ is a false discovery: $X_j \perp \perp Y | X_k$ for $k \in M_j$.

In this paper we derive our method under the less restrictive *marginal false discovery* definition (Breheny, 2018), which we illustrate using the causal diagram depicted above. In this diagram variable $A$ has a direct causal relationship with the outcome variable $Y$ and should never be considered a false discovery. Variable $C$ is independent of variable $Y$ regardless of any variables we adjust for and should always be considered a false discovery.

Variable $B$, on the other hand, is a more subtle case: $B$ is correlated with $Y$, but after adjusting for $A$, $B$ and $Y$ are independent. Depending upon the perspective taken, $B$ might or might not be considered a false discovery. In a fully conditional approach, $B$ is considered a false discovery. In a pathwise approach, whether $B$ is a false discovery or not depends on whether $A$ is active in the model.

In practice, however, both of these approaches suffer from the difficulty of determining the $A - B - Y$ relationship: is $A$ driving changes in $Y$ and $B$ merely correlated, or vice versa? Avoiding these complications is one of the primary motivations behind the marginal perspective, which is only concerned with false discoveries arising from variables like $C$. Depending on the application, selecting variables like $B$ may or may not be problematic, but it is almost always the case that a pure noise variable like $C$ is the worst kind of feature to select. This point, along with
the fact that the marginal false discovery rate is easy to interpret, makes the mFDR broadly useful and informative, although certainly there are scenarios in which the conditional FDRs are useful as well.

2.3.1 Penalized likelihood optimization

Consider data of the usual form \((y, X)\), where \(y\) denotes the response for \(i = \{1, \ldots, n\}\) independent observations, and \(X\) is a matrix containing the values of \(j = \{1, \ldots, p\}\) explanatory variables such that entry \(x_{i,j}\) corresponds to the value of the \(j^{th}\) variable for the \(i^{th}\) observation. We assume the columns of \(X\) are standardized such that each variable has a mean of 0 and \(\sum_i x_{i,j}^2 = n\).

The explanatory variables in \(X\) are related to \(y\) through a probability model involving coefficients \(\beta\). The fit of the model to the data can be summarized using the log-likelihood, which we denote \(\ell(\beta|X, y)\). In the classical setting, \(\beta\) is estimated by maximizing \(l(\beta|X, y)\). However, this approach is unstable in high dimensions unless an appropriate penalty, denoted \(P_\lambda(\beta)\), is imposed on the size of \(\beta\). In this case, \(\hat{\beta}\) is found by minimizing the objective function

\[
Q(\beta|X, y) = -\frac{1}{n} \ell(\beta|X, y) + P_\lambda(\beta).
\] (2.3.1)

In the classical setting, the maximum likelihood estimate is found by setting the score, \(u(\beta) = \nabla \ell(\beta|X, y)\), equal to zero. The penalized maximum likelihood estimate, \(\hat{\beta}\), is found similarly, although allowances must be made for the fact that the penalty function is typically not differentiable. These penalized score equations are known as the Karush-Kuhn-Tucker (KKT) conditions in the convex optimization literature, and are both necessary and sufficient for a solution \(\hat{\beta}\) to minimize \(Q(\beta|X, y)\).

In a likelihood-based regression model, the likelihood depends on \(X\) and \(\beta\) through a linear predictor \(\eta = X\beta\); in other words, we can equivalently express
the likelihood in terms of a loss function \( f(\eta|y) = -\ell(\beta|X, y) \). In what follows, we assume that the loss function is strictly convex with respect to the linear predictors \( \eta \); note that this does not imply strict convexity with respect to \( \beta \). Under these conditions, any solution \( \hat{\beta} \) that minimizes (2.3.1) with the lasso penalty \( P_\lambda(\beta) = \lambda \|\beta\|_1 \) must satisfy (Tibshirani, 2013):

\[
\frac{1}{n} u_j(\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \quad \text{if} \quad \hat{\beta}_j \neq 0
\]

\[
\frac{1}{n} u_j(\hat{\beta}) \in [-\lambda, \lambda] \quad \text{if} \quad \hat{\beta}_j = 0
\]

(2.3.2)

for \( j \in \{1, \ldots, p\} \).

2.3.2 Marginal false discovery rate bounds for penalized likelihood methods

In this section, we use classical distributional properties of the score function along with the KKT conditions given above to derive an upper bound for the number of marginal false discoveries in the lasso model. The basic intuition behind the derivation is that, given certain regularity conditions, if feature \( j \) is a marginally independent of \( y \), then \( Pr(\hat{\beta}_j \neq 0) \) is approximately equal to \( Pr(\frac{1}{n} |u_j(\beta)| > \lambda) \), where the classical score function \( u \) is evaluated at the true value of \( \beta \) provided that the log-likelihood is correctly specified (i.e., that the model assumptions hold). Given this result, the asymptotic normality of the score allows us to estimate this tail probability, and with it, the expected number of marginal false discoveries at a given value of \( \lambda \).

Three regularity conditions are required for these theoretical results to hold. These are given below, where we let \( W = \nabla^2 f \) denote the \( n \times n \) matrix of second derivatives of the loss function with respect to \( \eta \), such that the classical Hessian matrix \( \nabla^2 \ell(\beta) = -X^T W X \). We use \( W \) to denote this matrix evaluated at the true value of \( \beta \) and \( \hat{W} \) if evaluated at the lasso estimate. In addition, we let \( v_j = x_j^T W x_j \), with \( \hat{v}_j \) defined similarly.
(A1) Asymptotic normality of the score function: $(X^TWX)^{-1/2}u(\beta) \overset{d}{\to} N(0, I)$, where $I$ denotes the $p \times p$ identity matrix.

(A2) Vanishing correlation: $\frac{1}{n}x_j^TWx_{-j} \overset{p}{\to} 0$.

(A3) Estimation consistency: $\sqrt{n}(\hat{\beta} - \beta)$ is bounded in probability.

(A1) is a standard result of classical likelihood theory and can be shown for many types of models. (A3) is not a trivial condition, but has been studied and shown to hold for various models and various types of penalties under certain conditions. (A2), on the other hand, is unlikely to be truly satisfied by most features in practice. Our theoretical results illustrate what is required for the proposal to work perfectly in the sense of providing a consistent estimate for the mFDR. In practice, (A2) serves as a worst-case scenario for the correlation structure in the sense that other correlation structures will, on average, lead to fewer false discoveries. Viewed as an estimator, this means that the mFDR equation we will derive (2.3.4) is inherently conservative, in the sense that it will overestimate the true mFDR. Viewed instead as a control procedure, this means that (2.3.4) provides a probabilistic upper bound on the false discovery rate, and that if we use this equation to limit the mFDR to, say, 10%, we can be confident that the true mFDR is even smaller.

The fundamental reason for this is that if a pool of noise features are uncorrelated, a variable selection method may select several of them, whereas if they are correlated, the method will tend to select just one. More explicit results are shown in Section 2.4, which shows that when (A2) is violated, the mFDR bound is less tight, as one would expect. Assumptions (A1)-(A3) are further discussed in the specific cases of logistic and Cox regression later in this section.

We now formally state our main theoretical result. In interpreting this result, it is important to keep in mind that only features that are marginally independent of the outcome will satisfy both assumption (A2) and $\beta_j = 0$. In other words, the
Theorem applies to variables like $C$ in the causal diagram of Section 2.3, but not variable $B$: although the regression coefficient for $B$ is zero, its correlation with $A$ violates (A2).

**Theorem 1.** For any solution $\hat{\beta}$ of the lasso-penalized objective (2.3.1), we have $\hat{\beta}_j \neq 0$ if and only if

$$\frac{1}{n} \left| u_j(\hat{\beta}) + v_j \hat{\beta}_j \right| > \lambda. \quad (2.3.3)$$

Furthermore, provided that feature $j$ satisfies (A1)-(A3) and $\beta_j = 0$,

$$\frac{u_j(\hat{\beta}) + v_j \hat{\beta}_j}{\sqrt{v_j}} \xrightarrow{d} N(0,1).$$

**Proof.** The first remark follows directly from the KKT conditions (2.3.2) and the fact that, if the loss function $f(\eta|y)$ is strictly convex, then $W$ is positive definite and $v_j$ is positive. The asymptotic normality of the score function (A1) implies that the following Taylor series expansion holds:

$$u(\hat{\beta}) = u(\beta) - X^T WX(\hat{\beta} - \beta) + o_p(n)(\hat{\beta} - \beta).$$

Since $\beta_j = 0$, we then have

$$u_j(\hat{\beta}) = u_j(\beta) - x_j^T WX_{-j}(\hat{\beta}_{-j} - \beta_{-j}) - x_j^T WX_j \hat{\beta}_j + o_p(n) \hat{\beta}_j,$$

or

$$\frac{u_j(\hat{\beta}) + v_j \hat{\beta}_j}{\sqrt{v_j}} = \frac{u_j(\beta)}{\sqrt{v_j}} - \sqrt{\frac{n}{v_j}} \left[ x_j^T WX_{-j} [\sqrt{n}(\hat{\beta}_{-j} - \beta_{-j})] + o_p(1) \sqrt{\frac{n}{v_j}} \sqrt{n} \hat{\beta}_j. \right]$$

Noting that $v_j/n = O(1)$, the first term on the right side of the equation converges to $N(0, 1)$ by (A1); the second term goes to zero by conditions (A2) and (A3); and the third term also goes to zero by (A3).

Theorem 1 therefore implies that the probability that feature $j$ is selected,
given that it is marginally independent of the outcome, is approximately the probability that a random variable following a $N(0, v_j/n^2)$ distribution exceeds $\lambda$ in absolute value. In principle, the expected number of marginal false selections could be obtained by summing this probability over the set of marginally independent noise variables; in practice, since the identity of this set is unknown, a conservative alternative is to sum over all $p$ variables. This leads to the following upper bounds for the number and rate of marginal false discoveries:

$$\hat{FD} = 2 \sum_{j=1}^{p} \Phi \left( \frac{-n\lambda}{\sqrt{v_j}} \right)$$

$$\hat{mFdr} = \frac{\hat{FD}}{|S|}.$$  

where $S$ is the set of selected variables and $|S|$ its size. Note that because $p$ is used as an upper bound for the total number of noise features, $\hat{FD}$ and $\hat{mFdr}$ will be somewhat conservative even when (A1)-(A3) are fully satisfied. However, in the scenario where most features are null, as is often presumed in high dimensional applications, this effect will be relatively minor. For a given value of $\lambda$ the process of calculating $\hat{mFdr}$ is encapsulated by Algorithm 1.

---

**Algorithm 1** Calculating the mFDR upper bound.

```plaintext
procedure
    Estimate $\hat{W} \leftarrow \nabla^2 f(\hat{\eta})$
    for $j \in \{1, \ldots, p\}$ do
        $\hat{v}_j \leftarrow x_j^T \hat{W} x_j$
        $\hat{FD}_{j,\lambda} = 2 \Phi \left( \frac{-n\lambda}{\sqrt{\hat{v}_j}} \right)$ by the distributional result of Theorem 1
    $\hat{FD}_\lambda = \sum_{j=1}^{p} \hat{FD}_{j,\lambda}$
    $\hat{mFdr}_\lambda = \min \left( \frac{\hat{FD}_\lambda}{|S|}, 1 \right)$
return $\hat{mFdr}_\lambda$
```
2.3.3 Other penalty functions

The form of the mFDR calculation (2.3.4) is determined by the KKT conditions (2.3.2). Although the theorem in previous Section is specific to the lasso, many other penalties proposed in the literature have very similar KKT conditions. This in turn allows our results to be easily extended to other penalized methods.

For example, consider the elastic net (Zou and Hastie, 2005), which utilizes two penalty parameters: $\lambda_1$ and $\lambda_2$, with $\lambda_1, \lambda_2 > 0$. The elastic net solution is found by minimizing $-\frac{1}{n} \ell(\beta | X, y) + \lambda_1 ||\beta||_1 + \frac{\lambda_2}{2} ||\beta||_2^2$. The resulting KKT conditions dictate that $\hat{\beta}_j \neq 0$ if and only if

$$\frac{1}{n} \left| \frac{u_j(\hat{\beta})}{v_j} \right| + \frac{v_j \hat{\beta}_j}{\lambda_1} > 1.$$

Compared with the corresponding equation for the lasso (2.3.3), the only change is that the right hand side of the selection condition has changed from $\lambda$ to $\lambda_1$. Thus, equation (2.3.4) applies to the elastic net as well, with a similarly trivial change (replacing $\lambda$ with $\lambda_1$). Note that the actual estimates $\hat{\beta}$ (and as a result $\tilde{W}$, $\hat{v}$, and $|S|$) may certainly change a great deal, and thus the resulting inferences may be very different, but the form of the mFDR upper bound is essentially identical.

Furthermore, in some cases, such as for MCP and SCAD, the form of the upper bound is exactly the same. In other words, equation (2.3.3) holds for these penalties as well as the lasso, and therefore the mFDR upper bound is unchanged from (2.3.4) – although again, the actual estimates $\hat{\beta}$ and any quantities based on them will be very different.

It is worth noting here that MCP and SCAD differ from the lasso by relaxing the degree of penalization on variables with large effects. This leads to solutions with greater sparsity and reduced bias. In particular, both theoretical analysis and simulation studies have demonstrated that convergence is typically faster for MCP and SCAD than for the lasso (Zhang, 2010; Fan and Li, 2001;
Breheny and Huang, 2011). Thus, as we will see later in the paper, the accuracy of regularity condition (A3) is typically better at finite sample sizes for these estimators than for the lasso and as a consequence, the resulting mFDR bound is tighter (i.e., less conservative).

2.3.4 Logistic regression

So far we have presented our results in the general setting of penalized likelihood optimization. One specific type of model to which Section 2.3.2 applies is penalized logistic regression. Suppose $y_i$ follows a Bernoulli distribution with success probability $\pi_i$. In logistic regression the logit of $\pi_i$ is modeled as a function of $\eta = X\beta$. This results in a likelihood consisting of the product of $n$ independent Bernoulli distributions with success probabilities $\pi_i = \exp(x_i^T\beta)/(1 + \exp(x_i^T\beta))$ and $W$ a diagonal matrix whose entries are given by $\pi_i(1 - \pi_i)$.

The asymptotic normality of the score function for logistic regression (A1) can be found in, e.g., McCullagh and Nelder (1989), while necessary conditions for $\sqrt{n}$-consistency of the lasso estimates for logistic regression (A3) are established in Fan and Li (2001).

2.3.5 Cox regression

The results of Section 2.3.2 also apply to penalized Cox proportional hazards regression. Here the outcome of interest contains two components: a time, $y_i$, along with an accompanying indicator variable, $d_i$, where $d_i = 1$ indicates $y_i$ is an observed event time and $d_i = 0$ indicates $y_i$ is a right censoring time.

Let $t_1 < t_2 < \ldots < t_m$ be an increasing list of unique failure times indexed by $j$. The Cox model assumes a semi-parametric form of the hazard such that $h_i(t) = h_0(t)e^{x_i^T\beta}$, where $h_i(t)$ is the hazard for observation $i$ at time $t$ and $h_0(t)$ is a common baseline hazard. Cox regression is based upon the partial likelihood (Cox,
where $R_j$ denotes the set of observations still at risk at time $t_j$, known as the risk set.

Letting $\pi_{ij} = \exp(\eta_i) / \sum_{k \in R_j} \exp(\eta_k)$, the $i$th diagonal element of $W$ is given by $\sum_j d_j \pi_{ij} (1 - \pi_{ij})$, while the $i, k$th off-diagonal element is $- \sum_j d_j \pi_{ij} (1 - \pi_{kj})$. As discussed in Simon (2011), however, the off-diagonal elements of $W$ are typically negligible except for very small sample sizes. Thus, for the simulations we present in Section 2.4, we took the off-diagonal elements of $W$ to be zero in order to speed up the calculations. We found no meaningful difference in the calculation of $\hat{mFdr}$ when using the diagonal approximation of $W$ in place of the full matrix.

The asymptotic normality of the score function for Cox regression (A1) is established in Andersen and Gill (1982), while necessary conditions for $\sqrt{n}$-consistency of the lasso estimates for Cox regression (A3) can be found in Fan (2002). There are, however, some additional considerations considering censoring with respect to assumption (A2) in Theorem 1. As discussed earlier, (A2) holds for features that are marginally independent of all other features as well as the outcome. For (A2) to hold for Cox regression, however, the feature must also be independent of the censoring mechanism. This additional requirement is needed because when a variable is related to the censoring mechanism, its distribution will drift over time as certain values are disproportionately removed from the risk set, which can induce correlations between variables that would otherwise be uncorrelated. The impact of this additional concern on $\hat{mFdr}$ is further assessed via simulation in the next section.
2.4 Simulation studies

In this section we study the behavior of our proposed method for controlling the mFDR via several simulation studies. In each study we generate \( j \in \{1, \ldots, p\} \) features from standard normal distributions for \( i = 1, \ldots, n \) subjects. We focus on penalized logistic and Cox regression models, although we also carry out some simulations involving normally distributed outcomes in order to compare our results with those in Breheny (2018).

For logistic regression scenarios, binary outcomes are generated from independent Bernoulli distributions with parameter \( \pi_i = \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta)) \). For Cox regression scenarios, survival outcomes are generated from independent exponential distributions with rate parameter \( \theta_i = \exp(x_i^T \beta) \). And for linear regression scenarios, outcomes are generated from the model \( y_i = x_i^T \beta + N(0, \sigma^2) \). Factors such as \( p, n, \beta \), the correlation between features, and censoring are varied throughout these simulations.

In this section and the next, we compare our proposed mFDR control method with sample splitting and covariance testing. The covariance testing approach was implemented using the `covTest` package (Lockhart et al., 2014) and the `forwardStop` function of the `SelectiveInference` package (Tibshirani et al., 2016); however, the current versions of these packages caution that this approach is considered “developmental” for logistic regression and is not currently implemented for Cox regression. Although software does exist for sample splitting for linear regression via the `hdial` package Dezeure et al. (2015), the current version of the package does not offer methods for logistic or Cox regression models, so we manually implemented our own sample splitting approach.
2.4.1 Accuracy, sample size, and correlation

In Section 2.3.2, the mFDR bound (2.3.4) was shown to be tight in an asymptotic sense provided that condition (A2) was met. To see how useful this result is in practice, we wish to observe how tight this bound is at finite sample sizes as well as when condition (A2) is violated – i.e., when noise features are correlated. We evaluate the accuracy of the mFDR bound by comparing \( \hat{mFdr} \) with the true mFDR – i.e., the empirical proportion of noise features selected – at each value along a fixed \( \lambda \) sequence, averaged across 1,000 simulated data sets. In this simulation, time-to-event outcomes were uncensored; the effect of censoring on \( \hat{mFdr} \) is considered in the next section.

We generate our data using \( p = 40 \) with \( \beta_{1:4} = 10/\sqrt{n} \) and \( \beta_{5:40} = 0 \) while varying \( n \) from 100 to 1,000 in increments of 100. We take \( \beta \) to be function of \( n \) in order to maintain the difficulty of the variable selection problem as \( n \) increases. Without this provision, the feature selection problem becomes too easy at large sample sizes and the “interesting” region where the mFDR is not close to 0 or to 1 is very small.

In total we assess twelve different scenarios consisting of each possible combination of:

- Two different correlation structures: independent noise variables and correlated noise variables
- Three different regression methods: linear regression, logistic regression, and Cox regression
- Two different penalties: lasso and MCP.

For our correlated setting noise variables are given an autoregressive correlation structure based upon their index such that \( \text{Cor}(x_j, x_k) = 0.8^{|j-k|} \). We display the
mean value of $\hat{mFdr}$ at the $\lambda$ value where the average observed proportion of noise variables is 10%.

![Finite-sample accuracy of the mFDR bound (2.3.4).](image)

Focusing first on the case of independent noise features, we see that unlike in the linear regression case, where $\hat{mFdr}$ is essentially perfect at all sample sizes considered, for logistic and Cox regression $\hat{mFdr}$ tends to provide a conservative upper bound at small sample sizes, although this effect does diminish and the bound becomes tighter as $n$ increases. Furthermore, the MCP penalty leads to faster convergence and tighter bounds than the lasso penalty; this is particularly noticeable in the logistic and Cox regression settings.

When noise features are correlated, the $\hat{mFdr}$ bound is always conservative, at all sample sizes. In terms of Theorem 1, when noise features are correlated, (A2) is violated because the term $\frac{1}{n}x_j^TWX_{-j}$ does not converge to zero, leaving behind a remainder term that is unaccounted for in (2.3.4). Nevertheless, while this does mean that the bound is less tight, it does not prevent the use of the method for the purposes of mFDR control. Furthermore, in this case, the effect is relatively slight: for the most part, $\hat{mFdr}$ is approximately 20% when the true mFDR is 10%.
2.4.2 Features associated with censoring

As discussed in Section 2.3.5, the presence of censoring in Cox regression means that condition (A2) will also be violated if noise features are associated with the censoring mechanism. Here, we assess the impact such an association has on the $\hat{mFDR}$.

Consider two scenarios, A and B. In each scenario censoring times are generated from exponential distributions with $\theta_i = \exp(x_i^T \gamma)$, with

$$\gamma_A = (0, 0, 0, 0.45, -0.45, 0.45, -0.45, 0, \ldots, 0)$$

for scenario A, while in scenario B all variables are independent of censoring (i.e., $\gamma_B = 0$). By design, each of these scenarios results, on average, in a censoring rate of 50%. The same set of $n = 500$ true failure times is used for both scenarios and are generated from independent exponential distributions with rate parameter $\theta_i = \exp(x_i^T \beta)$, where $\beta_{1:4} = 0.45$ and $\beta_{5:40} = 0$. The value 0.45, used in $\gamma$ and $\beta$, is approximately $10/\sqrt{n}$ for this scenario, so that the results of this simulation are comparable with those displayed in Figure 2.1.

In scenario A, several noise features are associated with the censoring mechanism. Here, when $\hat{mFdr}$ is 10%, the observed false discovery rate averaged across 1,000 simulations is 4.03% using the lasso penalty and 6.31% using MCP. In scenario B, where all censoring assumptions are met, the observed false discovery rates are 5.01% and 6.85% for the lasso and MCP, respectively. These results demonstrate that noise variable associated with the censoring mechanism will lead to a more conservative upper bound; however, the effect is negligible compared to the considerations described in Section 2.4.1.
2.4.3 Comparison with other methods

In this section, we compare our proposed mFDR approach with other methods of variable selection used in the high-dimensional setting. We generate our data based upon the structure depicted in the causal diagram of Section 2.3, using $n = 400$ and $p = 1000$ with ten variables causally related to the outcome (like “A”), such that $\beta_{1:10} = b$ and vary $b$ throughout the study. Corresponding to each of these variables, we generate nine correlated ($\rho = 0.5$) variables (akin to “B”). The remaining 900 variables are noise (akin to “C”), however they are correlated with each other such that $\text{Cor}(x_j, x_k) = 0.8^{|j-k|}$, thereby creating a situation where the mFDR upper bound will be conservative.

![Figure 2.2: The number of true, “A”, variables selected by each lasso based method of false discovery rate control, averaged across 1,000 simulation iterations plotted as a function of $\beta$.](image)

For each of the following methods, we assess the average number of selections for each variable type:

- Our mFDR approach, where variables are selected by the lasso using the smallest $\lambda$ with $\widehat{mFdr} \leq .1$.

- Sample splitting (Wasserman and Roeder, 2009), where we first fit a lasso regression model on half of the data to select the top 20 variables. We then
use the remaining data to fit an unpenalized regression model on the variables selected in the first stage. With the unpenalized model we conduct traditional Wald hypothesis tests on the regression coefficients and apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate at 10%. Here, we limit the first-stage selections to 20 variables so that in the second stage, the model contains 10 events per variable (Peduzzi et al., 1995).

- Covariance testing (Lockhart et al., 2014), which we use in conjunction with the forward stopping rules proposed by G’Sell et al. (2016) to control the pathwise false discovery rate at 10%.

- Univariate testing, where we fit separate, unpenalized regression models to each variable individually, and then adjust the resulting p-values using the Benjamini-Hochberg procedure to control the false discovery rate at 10%.

- Cross-validation (CV), where 10-fold CV is used to choose $\lambda$ for the lasso model; note that this approach makes no attempt to control the false discovery rate.

The comparison of FDR control procedures for the lasso (mFDR, sample splitting, and covariance testing) is shown in Figure 2.2. Compared to other false discovery rate control methods for penalized regression, the mFDR approach is more powerful – at the same nominal FDR, the marginal approach selects more causally important variables at each signal strength. This is not a surprising result: a primary motivation of the marginal approach is that, by adopting a less restrictive definition of false discovery, we could improve power. The covariance testing approach in particular is particularly conservative in the logistic regression setting, selecting far fewer causal features than either mFDR or sample splitting. As of the time of this writing, the covariance test has not been extended to Cox regression.
Figure 2.3: The average number of selections for each type of variable depicted in the causal diagram of Section 2.3, when cross-validation is used to select \( \lambda \).

Simulation results for cross validation are shown in Figure 2.3. As the figure shows, while cross-validation has excellent sensitivity, selecting nearly all the causal ("A") variables even at low signal strength, it does not control the number of false selections. In fact, at all levels of \( b \), the majority of the features in the CV-selected model are noise features. This problem becomes increasingly worse as the signal strength increases. For example, at \( b = 0.8 \), over 70% of the features in the CV model are noise features. In contrast, as may be seen in Figure 2.4, the vast majority of the features selected by the mFDR control approach are causally related to the outcome.

Figure 2.4 compares the mFDR and univariate approaches, both of which are designed to control the proportion of marginal false discoveries without making any claims regarding features indirectly associated with the outcome. The top panels show that the two methods are comparable in terms of their ability to select the true causal variables. The advantage of using a regression-based approach over univariate testing, however, is clearly seen when looking at the number of other selections. Lasso with mFDR control greatly reduces the number of correlated
(“B”) selections compared to univariate testing: whereas univariate testing tends to identify dozens of indirectly associated features as discoveries, lasso with mFDR control tends to select at most two.

The mFDR control approach also selects far fewer noise (“C”) variables than univariate testing. The reason for this is that the large number of “B” features selected by univariate testing allow it to select a much larger number of features while maintaining the overall FDR at 10%.

![Figure 2.4: The average number of selections, for each type of variable depicted in the causal diagram of Section 2.3, is plotted as function of $\beta$ for each method which controls the false discovery rate.](image)

For this reason, Figure 2.4 is potentially somewhat misleading in terms of comparing mFDR and univariate testing in terms of their power to identify type “A” features. Table 2.1 presents an alternative metric for comparing the univariate and mFDR approaches. Given the large difference between the two approaches with respect to number of indirectly associated (“B”) features selected, the table presents the ratio of the number of causally associated features to the number of
noise features – i.e., the A:C ratio – for various values of $\beta$. Whereas Figure 2.4 suggests that the two approaches were approximately equally effective at identifying the truly important variables, Table 2.1 shows that the lasso-mFDR approach is far more powerful in terms of the number of true discoveries per noise feature selected.

<table>
<thead>
<tr>
<th>$b$</th>
<th>Cox</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate</td>
<td>Lasso-mFDR</td>
</tr>
<tr>
<td>0.25</td>
<td>3.5</td>
<td>17.7</td>
</tr>
<tr>
<td>0.35</td>
<td>3.0</td>
<td>36.2</td>
</tr>
<tr>
<td>0.45</td>
<td>2.9</td>
<td>102.2</td>
</tr>
<tr>
<td>0.65</td>
<td>2.4</td>
<td>333.3</td>
</tr>
<tr>
<td>0.75</td>
<td>1.9</td>
<td>750.0</td>
</tr>
</tbody>
</table>

Table 2.1: The ratio of Causal:Noise (A:C) feature selections for various values of the signal strength $b$.

Indeed, for the univariate approach, the rate of true variable selections per noise variable selection is not only low, but in fact decreases as the signal strength increases. In contrast, with the mFDR approach, the causal:noise selection ratio increases with signal strength, as one would hope.

2.5 Case studies

2.5.1 Lung cancer survival and gene expression

Shedden et al. (2008) studied the survival of 442 early-stage lung cancer subjects. Researchers collected high-dimensional gene expression data on 22,283 genes and additional clinical covariates of age, race, gender, smoking history, cancer grade, and whether or not the subject received adjuvant chemotherapy. The retrospective study used time to death as its outcome, which was observed in 236 (53.3%) of the
In our analysis we aim to select important genes while controlling the marginal false discovery rate. We use a semi-penalized Cox proportional hazards regression model that allows all of the clinical covariates to enter the model unpenalized while using the sparsity introduced by penalization to screen for additional genes related to survival. We aim to limit the mFDR to 10% and compare our results to those of other methods. Note that in this example, the number of features $p$ in equation (2.3.4) is the number of penalized features (i.e., the number of genes), since the clinical covariates will always be in the model.

Figure 2.5 illustrates the mFDR estimates for this study. On the left, we see a gap between the number of selected genes and the expected number of false selections at $\log(\lambda) \approx -2$. This indicates that for these values of $\lambda$, many of the genes selected by the lasso are likely to be truly related to survival, as it would be unlikely for so many noise features to be selected merely by random chance. On the right, these expected number of false discoveries is plotted as a fraction of the selected features (i.e., the mFDR).

Table 2.2 displays results for three potential approaches to choosing $\lambda$: the
<table>
<thead>
<tr>
<th></th>
<th>λ</th>
<th>EF</th>
<th>S</th>
<th>mFDR</th>
<th>CVE</th>
<th>MCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shedden</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lasso CV</td>
<td>0.121</td>
<td>9.16</td>
<td>13</td>
<td>70.4</td>
<td>1308</td>
<td>-</td>
</tr>
<tr>
<td>CV(1se)</td>
<td>0.180</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>1320</td>
<td>-</td>
</tr>
<tr>
<td>mFDR</td>
<td>0.146</td>
<td>0.52</td>
<td>8</td>
<td>6.6</td>
<td>1313</td>
<td>-</td>
</tr>
<tr>
<td>MCP CV</td>
<td>0.133</td>
<td>2.53</td>
<td>8</td>
<td>31.7</td>
<td>1310</td>
<td>-</td>
</tr>
<tr>
<td>CV(1se)</td>
<td>0.180</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>1320</td>
<td>-</td>
</tr>
<tr>
<td>mFDR</td>
<td>0.155</td>
<td>0.16</td>
<td>2</td>
<td>8.2</td>
<td>1315</td>
<td>-</td>
</tr>
<tr>
<td><strong>Spira</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lasso CV</td>
<td>0.029</td>
<td>49</td>
<td>49</td>
<td>100</td>
<td>1.00</td>
<td>24.5</td>
</tr>
<tr>
<td>CV(1se)</td>
<td>0.063</td>
<td>32</td>
<td>32</td>
<td>100</td>
<td>1.07</td>
<td>25.0</td>
</tr>
<tr>
<td>mFDR</td>
<td>0.146</td>
<td>0.78</td>
<td>10</td>
<td>7.8</td>
<td>1.29</td>
<td>30.7</td>
</tr>
<tr>
<td>MCP CV</td>
<td>0.077</td>
<td>13</td>
<td>13</td>
<td>100</td>
<td>1.11</td>
<td>30.2</td>
</tr>
<tr>
<td>CV(1se)</td>
<td>0.093</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>1.18</td>
<td>33.3</td>
</tr>
<tr>
<td>mFDR</td>
<td>0.151</td>
<td>0.42</td>
<td>5</td>
<td>8.4</td>
<td>1.31</td>
<td>33.9</td>
</tr>
</tbody>
</table>

Table 2.2: Results for different choices of $\lambda$ in applying the lasso and MCP penalties to the Shedden and Spira data. mFDR = Marginal false discovery rate (%). CVE = Cross-validation error. MCE = Misclassification error (%).

value that minimizes cross-validation error (CV), the value that comes within 1 standard error (SE) of minimizing CV (1se), and the smallest value of $\lambda$ satisfying mFdr < 10%. From the table, we can see that there appears to be an inherent tradeoff between the optimal prediction and restricting a model to containing few noise features. In this case, we see that the lasso model with lowest prediction error has an estimated mFDR of 70.4%, indicating that despite its prediction accuracy, a substantial fraction of the 13 features selected by the model may be noise. Conversely, if we aim to control the mFDR at 10%, we select only 8 features. Although
we can be more confident that these 8 features are truly related to survival, the model is somewhat less accurate from a prediction standpoint (CV error of 1313 compared to 1308). The CV (1se) choice λ aims to select a parsimonious model whose accuracy is comparable with the best model (Hastie et al., 2009). However, in this example the CV (1se) method selects the null model, giving the impression that no genes can be reliably selected. Note that the mFdr < 10% approach is also within 1 SE of the best model, and therefore may also be thought of as striking an appropriate balance between prediction and parsimony in this example.

An alternative to lasso-penalized regression is to use the MCP penalty, which relaxes the degree of shrinkage for large coefficients, thereby allowing a smaller number of features to account for the observed signal. Our approach indicates that the MCP model minimizing CV error is likely to have an mFDR no higher than 31.7%; compared to the lasso model that minimized CV error, we can be considerably more confident in the smaller number of features selected by MCP. The prediction accuracy of MCP was similar to lasso in this case (CV error 1310 vs. 1308), with the lasso being slightly more accurate and MCP slightly more parsimonious.

MCP strikes an attractive balance in this case between prediction accuracy and false discovery rate. One reason for this is that two genes, ZC2HC1A, and FAM117A, have very large effects, exactly the scenario MCP is designed to perform well in. In the lasso model with λ selected by cross validation these variables have coefficients of -0.170 and -0.118 respectively, while in the MCP model these variables have coefficients of -0.220, -0.175 despite cross-validation selecting a larger λ (more penalization) for MCP. This illustrates the fact that lasso models can only accommodate large effects by lowering λ, which comes at the cost of allowing additional noise variables into the model.

For comparison we also analyzed the data using repeated sample splitting
with 100 random splits, as advocated by Meinshausen et al. (2009). However, this approach was unable to select any genes at a false discovery rate of 10%, or even at a more liberal rate of 50%. As mentioned earlier, the covTest package does not (currently) accommodate survival data.

We also applied a large scale univariate testing approach to the Shedden data, fitting a series of Cox regression models adjusting for all clinical covariates and containing a single gene. After using the Benjamini-Hochberg procedure to control the false discovery rate at 10%, this approaches selects 803 genes, far more than any of the regression-based approaches. As discussed in Section 2.4.3, the primary difference between univariate and regression-based methods with respect to false discoveries is that univariate approaches tend to select large numbers of features that are indirectly correlated with the outcome. Our simulation results would suggest that for these data, univariate testing yields a large number of genes, most of which are only indirectly associated with survival, while the mFDR approach yields a smaller number of genes, most of which are directly associated with survival.

2.5.2 Lung cancer status among smokers

Spira et al. (2007) collected RNA expression data for 22,215 genes from histologically normal bronchial epitheliums of \( n = 192 \) smokers, of which 102 had developed lung cancer and 90 had not developed lung cancer. The goal of the study was to identify genes that are indicative of whether or not a smoker has lung cancer. For our analysis we fit penalized logistic regression models, and compare the results using mFDR, cross validation, and the covariance test to choose \( \lambda \), as well as the genes selected via sample splitting. We also compare these model-based approaches to the traditional univariate approach with false discovery rate control.

Table 2.2 shows results for the Spira data at various \( \lambda \) values for the lasso and MCP penalty functions. In contrast with the Shedden case study, for both
penalties in this example the mFDR bound is 100% at the value of \( \lambda \) selected by cross validation. In other words, we are unable to provide an upper bound for the mFDR at the value of \( \lambda \), and one should be cautious about interpreting the selected features as significant. For example, at \( \lambda = 0.029 \), there may very well be some true discoveries among the 49 selected features, but a lasso-penalized logistic regression model could easily select 49 features just by chance at that value of \( \lambda \) even if none of them were related to the outcome.

In order to arrive at a set of features with a low false discovery rate, we must choose a much smaller model (for the lasso model, 10 features instead of 49). As in the earlier case study, however, the price of an FDR restriction is a decrease in prediction accuracy. The prediction error is considerably higher in the mFdr < 10% lasso and MCP models than for the models selected by cross-validation. In the Spira example, no genes have particularly large signals; as a result, the MCP approach is less successful than at finding a model that is attractive from both the mFDR and prediction perspectives than it was in the Shedden example.

As in the Shedden case study, neither the sample splitting nor covariance testing approaches were able to select any genes at a 10% FDR, nor could any genes be selected by either approach at the considerably more liberal cutoff of 50%. Finally, we tested for significant features in a univariate approach, fitting a separate logistic regression model for each gene and controlling the false discovery rate at 10% using the Benjamini-Hochberg procedure. This approach identifies 2,833 genes as significant (single-gene tests based on \( t \)-tests yielded similar results).

Depending on the goals of the analysis and whether indirect associations are of interest, either the univariate or mFDR approach might prove useful here. However, as in the earlier case study, fully conditional and pathwise conditional approaches to controlling the false discovery rate tend to be so restrictive that they are unable to identify any features in high-dimensional studies.
2.6 Discussion

Controlling the marginal false discovery rate of a penalized likelihood model is a useful way of assessing the reliability of a selected set of features. Unlike other approaches that have a similar goal, such as sample splitting or the covariance test, the mFDR approach uses a less strict definition of a false discovery which does not require conditioning on any other variables and consequently only limits the selection of variables that are noise in an unconditional – i.e., marginal – sense. The simulations in this paper demonstrate that while mFDR is based upon a weaker false discovery definition, when used in combination with a penalized regression model it tends to perform very well at limiting the number of indirect (non-causal) feature selections.

Furthermore, mFDR is far more convenient from the standpoint of computational burden than other approaches. For the Spira data, where \( n = 192 \) and \( p = 22215 \), fitting a penalized logistic regression model and then estimating the mFDR for the entire \( \lambda \) sequence takes only 1.4 seconds. This is over 400 times faster than sample splitting and the covariance test, each of which took over 9 minutes, and is nearly 30 times faster than even univariate testing, which took over 3 minutes. Results are similar for the Shedden survival analysis. The computational efficiency of mFDR makes it a particularly appealing tool during the early stages of an analysis when a number of candidate models are being considered.

The mFDR control procedure proposed here is currently implemented in the R package ncvreg \((\text{Breheny and Huang, 2011})\) using the mFDR function. The function accepts an ncvreg fitted model and calculates the expected number of false discoveries, as well as the mFDR for each value of the \( \lambda \) sequence used in fitting the model. The package also provides a plotting method, which produces plots like those in Figure 2.5. The ncvreg package accommodates lasso, MCP, SCAD, elastic net, and MNet \((\text{Huang et al., 2016})\) penalties and provides a convenient model
summary measure for penalized linear, logistic, and Cox regression models.

In summary, this manuscript has demonstrated how to extend control over the marginal false discovery rate for penalized regression models to any general likelihood-based loss function. Controlling the mFDR is much faster, considerably less restrictive, and is conveniently available in an open-source software package. The marginal false discovery rate is an easily interpreted and broadly useful approach to inference concerning the reliability of selected features in a penalized regression model, capable of being generalized to a wide variety of penalty functions and modeling frameworks.
3.1 Abstract

Penalized regression methods, most notably the lasso, are a popular approach to analyzing high-dimensional data. An attractive property of the lasso is that it naturally performs variable selection. An important area of concern, however, is the reliability of these variable selections. Motivated by local false discovery rate methodology from the large-scale hypothesis testing literature, we propose a method for calculating a local false discovery rate for each variable under consideration by the lasso model. These rates can be used to assess the reliability of an individual feature, or to estimate the model’s overall false discovery rate. The method can be used for all values of the lasso penalty parameter, $\lambda$. This is particularly useful for models with a few highly significant features but a high overall Fdr, which are a relatively common occurrence when using cross validation to select $\lambda$. It is also flexible enough to be applied to many varieties of penalized likelihoods including GLM and Cox models, and a variety of penalties, including MCP and SCAD. We demonstrate the validity of this approach and contrast it with other inferential methods for penalized regression as well as with local false discovery rates for univariate hypothesis tests. Finally, we show the practical utility of our method by applying it to two case studies involving high dimensional genetic data.

3.2 Introduction

In recent years, data involving large numbers of features has become increasingly prevalent. There are many approaches to analyzing these high-dimensional
data sets. Broadly speaking, these approaches can be divided into two major categories, \textit{large-scale testing} approaches and \textit{model-based} approaches. The former entails conducting separate tests for each feature, while the later considers many features simultaneously using a single model. Depending on the research question and the experimental design, one of these approaches might be preferred; however in many applications both are reasonable and afford their own advantages. Large-scale testing offers well-established methods, such as \(q\)-values and local false discovery rates, that provide an individual assessment of feature importance while controlling the false discovery rate. However, the information loss when only considering univariate relationships can be unsettling. Sparsity inducing penalized regression methods, such as the lasso, provide an interpretable framework for model-based approaches, but determining feature importance or achieving false discovery rate control when using these models can be difficult. Before progressing into the technical details, we present two simulated examples which motive our \textit{mfdr} method and illustrate the relationship it shares with both of these two seemingly disjoint approaches to high-dimensional data analysis.

Figure 3.1 shows two different simulated datasets, each containing two features that are causally related to the outcome, two features that are correlated with the causal variables, and 96 features that are purely noise. The panels on the left show the usual lasso coefficient paths that are returned by standard software, while the panels on the right display the results of our \textit{mfdr} method along the lasso path. In the upper left panel, the model at \(\lambda_{CV}\)\footnote{The \(\lambda_{CV}\) value which minimizes cross validation error}, contains several noise variables. However, it is quite easy to visually distinguish the two causal features from the noise simply by inspecting their coefficient paths. The \textit{mfdr} plot in the upper right panel confirms this visual assessment, showing clear separation between features in the region around \(\lambda_{CV}\). The dataset in
Figure 3.1: The rows of this figure correspond to two simulated datasets. The column on the left shows the usual lasso coefficient path, while the column on the right displays our method’s feature specific local false discovery rates (mfdr) along the lasso path. The triangles prior to the start of the mfdr path are the traditional local false discovery rate estimates resulting from large-scale testing. Along the mfdr path dashed lines indicate the portion of the path where a feature is inactive ($\hat{\beta} = 0$) in the model. The vertical dotted line shows the value of the penalty parameter, $\lambda$, chosen by cross validation.

The second row demonstrates a more challenging case where determining the important features using the coefficient path is difficult. The mfdr plot in the lower right panel lends clarity to this situation, showing a relatively clear separation between feature types near $\lambda_{CV}$.

The two mfdr paths of Figure 3.1 also illustrate the connection between the mfdr approach and traditional large-scale testing approach to local false discovery rates. These traditional local false discovery rates are denoted by the triangles prior to the start of the mfdr path, and are equivalent to the mfdr estimates at the
beginning of the mfdr path when no features are active in the model. Initially, each method identifies both causal features, along with some of the correlated features, as important. However as $\lambda$ decreases, and the causal features become active in the model, the mfdr method reveals the correlated features are more likely to be false discoveries.

Having presented an initial case for the utility of mfdr, and an illustration of the connections it shares with both traditional local false discovery rates and lasso regression, we structure the remainder of the paper as follows: Section 3.3 gives a more formal introduction to false discovery rates approaches in the context of both large-scale testing and model based approaches to high dimensional data. Section 3.4.1 introduces our lasso based mfdr estimator in the linear regression setting; and then Section 3.4.2 does so for a more general class penalized likelihood models, putting a focus on penalized logistic and Cox regression models. Section 3.5 studies of the mfdr approach using simulation, comparing it to existing methods commonly used in high-dimensional analysis, and Section 3.6 explores two real data case studies where the method proves to be useful.

### 3.3 Background

In the context of both large-scale testing and model-based approaches, this paper will focus on false discovery rates, a common approach to inference in high-dimensional data analysis. There are two main types of false discovery rates, tail-area approaches, which describe the expected rate of false discoveries for all features beyond a given threshold, and local approaches, which describe the density of false discoveries at a specific point. We adopt the general convention throughout, used by many other authors, of using Fdr to refer to tail-area approaches, and fdr to refer to local approaches, paralleling the traditional use of $F$ and $f$ to refer to distribution and density functions. Both Fdr and fdr have been well studied in the realm of
large-scale testing. The seminal Fdr procedure proposed Benjamini and Hochberg (1995) remains a widely popular approach to Fdr and has led many others to study
tail-area false discovery rate control (Storey et al., 2004; Genovese and Wasserman,
2004). Using an empirical Bayes framework, Efron et al. (2001) extended this idea
and proposed local false discovery rates, a proposal which has also been extended
in many ways; for additional background see the reviews of Strimmer (2008b) and
Farcomeni (2008).

Recently, false discovery rates have been considered in the realm of high-
dimensional modeling as well, although this research has focused exclusively on tail-
area approaches. The methods of this paper focus on lasso regression (Tibshirani,
1996), a popular modeling approach which naturally performs variable selection by
using L1 regularization. Among the variable selections made by the lasso, false
discoveries can be quite prevalent (Su et al., 2017).

The false discovery rate control provided by large-scale testing approaches is
marginal in the sense that a feature $X_j$ is considered a false discovery only if that
feature is marginally independent of the outcome $Y$: $X_j \perp \perp Y$. In regression, where
many features are being considered simultaneously, the issue is more complicated
and can involve various kinds of conditional independence. For example, we can
adopt a fully conditional perspective, which considers a feature $X_j$ to be a false
discovery if it is independent of the outcome conditional upon all other features:
$X_j \perp \perp Y \mid X_{k \neq j}$. For lasso models, sample splitting (Wasserman and Roeder, 2009;
Meinshausen et al., 2009) can be used to control the fully conditional false discovery
rate.

An alternative definition that can be used for penalized models is the pathwise
conditional perspective. Pathwise approaches focus on the point in the regulariza-
tion path at which feature $j$ first becomes active and condition only on the other
variables present in the model (denote this set $M_j$) when assessing whether or not
variable $j$ is a false discovery: $X_j \perp Y|X_k$ for $k \in M_j$. The methods of Lockhart et al. (2014) and Tibshirani et al. (2016) used in conjunction with the sequential stopping rule of G’Sell et al. (2016) allow for control over the pathwise conditional Fdr.

Weaker approaches to false discovery rates for penalized regression models have also been proposed, Breheny (2018) developed an analytic method which bounds the marginal false discovery rate (mFdr) of penalized linear regression models. Miller and Breheny (2018) extended the method to more general class of penalized likelihood models, while Huang (2017) addressed a similar question using a Monte Carlo approach.

In this paper we adopt the marginal perspective on false discoveries and develop a local approach which can be applied to a general class of penalized likelihood models. Our proposed method is related to the tail area mFdr approach of Breheny (2018) and Miller and Breheny (2018), and also local false discovery rates in large-scale testing. Despite being marginal in development, in practice our approach tends to capitalize on the conditional adjustments that naturally occur in regression modeling.

3.3.1 Large-Scale Testing, Fdr, and fdr

Consider data of the usual form $(y, X)$, where $y$ denotes the response for $i = \{1, \ldots, n\}$ independent observations and $X$ is an $n \times p$ matrix containing the values of $j = \{1, \ldots, p\}$ explanatory features. We presume that only a small subset of the available features have an important relationship with the outcome, and the goal of our analysis is to correctly identify those important features.

Large-scale univariate testing considers $p$ separate null hypotheses, each corresponding to a single feature, and conducts a univariate test on each of those
hypotheses, resulting in a collection of test statistics \( \{t_1, t_2, \ldots, t_p\} \) and corresponding p-values \( \{p_1, p_2, \ldots, p_p\} \). To meaningfully aggregate the results of these many hypothesis tests, Benjamini and Hochberg (1995) developed a procedure which limits \( V \), the number of falsely rejected null hypotheses, such that the tail-area Fdr, \( E(V/R) \), where \( R \) is the total number of rejections, is less than a pre-specified threshold \( \alpha \) when \( R > 0 \) and is defined to be 0 when \( R = 0 \).

Alternatively, Fdr can also be approached from a Bayesian perspective (Storey, 2002; Strimmer, 2008b). To outline this approach, let \( z_j \) denote the normalized test statistic for the \( j^{th} \) feature, specifically \( z_j = \Phi^{-1}(F_i(t_j)) \) where \( \Phi \) is the standard normal CDF, then define:

\[
\pi_0 = Pr(\text{Null}) \quad f_0(z) = \text{the density of null test statistics} \\
\pi_1 = Pr(\text{Non-null}) \quad f_1(z) = \text{the density of non-null test statistics}
\]

By construction \( f_0 \) is the standard normal distribution, and in most high-dimensional applications \( \pi_0 \) is presumed close to 1. Letting \( Z \) denote any subset of the real line, we define:

\[
F_0(Z) = \int_Z f_0(z)dz \\
F_1(Z) = \int_Z f_1(z)dz
\]

Suppose we observe \( z_j \in Z \) and we are interested in whether or not this feature belongs to the null or non-null group. For the members of \( Z \), Bayes rule suggests:

\[
Pr(\text{Null}|z_j \in Z) = \frac{\pi_0 F_0(Z)}{\pi_0 F_0(Z) + \pi_1 F_1(Z)} = \text{Fdr}(Z) \quad (3.3.1)
\]

The denominator of this expression is referred to as the mixture distribution and is denoted by \( F \). In similar fashion we define the mixture density, denoted by \( f \), to be: \( \pi_0 f_0(z) + \pi_0 f_1(z) \). The expression in (3.3.1) provides a tail-area Fdr characterization of the collection of features with test statistics contained in the region \( Z \), thus enabling control over Fdr via the choice of \( Z \). Alternatively, we may
take $Z$ to be the single point $z_j$:

$$\Pr(\text{Null}|z_j = z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + \pi_1 f_1(z)} = \text{fdr}(z)$$  \hspace{1cm} (3.3.2)

Here fdr($z$) is referred to as the local false discovery rate, the specific probability that the $j^{th}$ feature is a false discovery. In summary, 3.3.1 can be used to quantify the reliability of the set of features determined by $Z$, while 3.3.2 can be used to quantify the reliability of a single feature. As one would expect, Fdr and fdr are linked in several ways, one of these is through the relationship:

$$\text{Fdr}(Z_0) = E\left(\text{fdr}(z) | z \in [z_0, \infty)\right)$$

The Fdr of the set of features with normalized test statistics exceeding $z_0$ is equal to the average fdr of features with test statistics beyond that threshold, which ensures that selecting individual features using a threshold $\text{fdr}(z) < \alpha$ also limits Fdr below $\alpha$ for the entire set of features selected. Further exposition of the numerous links between Fdr and fdr can be found in the works of Efron (2005) and Strimmer (2008b).

An important aspect of Equations (3.3.1) and (3.3.2) is that their numerators are known (provided we take $\pi_0 = 1$); this means that we only require estimates of $F$ or $f$ in order to estimate Fdr or fdr. When $p$ is large, the empirical estimators, $\hat{F}$ and $\hat{f}$, are very good approximations, providing an empirical Bayes method for estimating Fdr and fdr in high-dimensional settings.

3.3.2 Penalized Regression and mFdr

In contrast with the univariate nature of large-scale testing, regression models simultaneously relate the explanatory features in $X$ with $y$ using a probability model involving coefficients $\beta$. In what follows we assume the columns of $X$ are standardized such that each variable has a mean of 0 and $\sum_j x_j^2 = n$. The fit
of a regression model can be summarized using the log-likelihood, which we denote \( \ell(\beta|X,y) \). In the classical setting, \( \beta \) is estimated by maximizing \( l(\beta|X,y) \). However, this approach is unstable when \( p > n \) unless an appropriate penalty is imposed on the size of \( \beta \). In the case of the lasso penalty, estimates of \( \beta \) are found by minimizing the objective function:

\[
Q(\beta|X,y) = -\frac{1}{n}\ell(\beta|X,y) + \lambda||\beta||_1
\]  
(3.3.3)

In the classical setting, the maximum likelihood estimate is found by setting the score, \( u(\beta) = \nabla \ell(\beta|X,y) \), equal to zero. The lasso estimate, \( \hat{\beta} \), is found similarly, although allowances must be made for the fact that the penalty function is typically not differentiable. These penalized score equations are known as the Karush-Kuhn-Tucker (KKT) conditions in the convex optimization literature, and are both necessary and sufficient for a solution \( \hat{\beta} \) to minimize \( Q(\beta|X,y) \).

An important property induced by the lasso penalty is sparsity, meaning that \( \hat{\beta}_j = 0 \) for a large number of features, and only a subset of the available features being selected into the model to be non-zero. The penalty parameter, \( \lambda \), governs the degree of sparsity with smaller values of \( \lambda \) leading to more variables having non-zero coefficients. The KKT conditions mathematically characterize feature selection at a given value of \( \lambda \).

Miller and Breheny (2018) use the KKT conditions to develop an upper bound for the number of features that could be expected in lasso model by random chance. Broadly speaking, if feature \( j \) is marginally independent of \( y \), then \( Pr(\hat{\beta}_j \neq 0) \) is approximately equal to \( Pr(\frac{1}{n}|u_j(\beta)| > \lambda) \), where the classical score function \( u \) is evaluated at the true value of \( \beta \). The asymptotic normality of the score can be used to estimate this tail probability, and provide a bound over the expected number of marginal false discoveries at a given value of \( \lambda \).

This approach provides an assessment for the overall set of selections made
by a lasso regression model; however it does not make any distinctions among these selections. It would be unusual to expect all selected features to each have the same likelihood of being a false discovery, which motivates the need for a local approach. Additionally, we mention that this approach can be used to select a model with low false discovery rate, but doing so often requires choosing a model with suboptimal predictive performance. A local approach offers the potential to reliably select features while using the most predictive model.

3.4 Estimating mfdr

3.4.1 Linear Regression

Consider the linear regression setting:

$$y = X\beta + \epsilon$$

$$\epsilon_i \sim N(0, \sigma^2)$$

As mentioned in Section 3.3.2, the lasso solution, $$\hat{\beta}$$, is mathematically characterized by the KKT conditions. For linear regression, these conditions require (Tibshirani, 2013):

$$\frac{1}{n} x_j^T (y - X\hat{\beta}) = \lambda \text{sign} (\hat{\beta}_j)$$

for all $$\hat{\beta}_j \neq 0$$

$$\frac{1}{n} x_j^T (y - X\hat{\beta}) \leq \lambda$$

for all $$\hat{\beta}_j = 0$$

We define the partial residual as $$r_j = y - X_{-j}\hat{\beta}_{-j}$$ where the subscript $$-j$$ indicates the removal of the $$j^{th}$$ feature. Using this definition it follows directly from the KKT conditions that:

$$\frac{1}{n} |x_j^T r_j| > \lambda$$

for all $$\hat{\beta}_j \neq 0$$

$$\frac{1}{n} |x_j^T r_j| \leq \lambda$$

for all $$\hat{\beta}_j = 0$$

The quantity $$\frac{1}{n} x_j^T r_j$$ governs the selection of the $$j^{th}$$ feature into the lasso
model and is analogous to a test statistic for selection in the sense that if its absolute value is large enough, relative to \( \lambda \), then that feature is selected into the model. Furthermore, in the special case of orthonormal design where \( \frac{1}{n}X^TX = I \), it straightforward to show that \( \frac{1}{n}x_j^Tr_j \sim N(\beta_j, \sigma^2/n) \) (Breheny, 2018). Under the null hypothesis that \( \beta_j = 0 \), this distributional result can be used to construct a normalized test statistic and an associated local false discovery rate estimator:

\[
z_j = \frac{\frac{1}{n}x_j^Tr_j}{\hat{\sigma}/\sqrt{n}}
\]

\[
\text{mfdr}(z_j) = \frac{\phi(z_j)}{\hat{f}(z_j)}
\]

(3.4.1)

Here \( \phi() \) denotes the standard normal density function, \( \hat{f} \) is an empirically estimated mixture density of the normalized test statistics, and \( \hat{\sigma} \) is an estimate of \( \sigma \). Numerous choices can be made regarding the estimation of \( f, \sigma, \) and \( \pi_0 \); for now we take a simple approach, by conservatively assuming \( \pi_0 = 1 \), using kernel density estimation to obtain \( \hat{f} \), and using the residual sum of squares divided by the model degrees of freedom as an estimate of \( \sigma^2 \) (Zou et al., 2007). Section 3.5.4 uses simulation to explore how sensitive this estimator is to different methods of density estimation.

Because the majority of lasso applications do not have an orthonormal design, we investigate the estimator in (3.4.1) by further studying the quantity \( \frac{1}{n}x_j^Tr_j \):

\[
\frac{1}{n}x_j^Tr_j = \frac{1}{n}x_j^T(X\beta + \epsilon - X_{-j}\hat{\beta}_{-j})
\]

\[
= \frac{1}{n}x_j^T\epsilon + \beta_j + \frac{1}{n}x_j^TX_{-j}(\beta_{-j} - \hat{\beta}_{-j})
\]

The component \( \frac{1}{n}x_j^T\epsilon + \beta_j \) is unaffected by the structure of \( \frac{1}{n}X^TX \); thus the estimator in (3.4.1) will be accurate in situations where the final term, \( \frac{1}{n}x_j^TX_{-j}(\beta_{-j} - \hat{\beta}_{-j}) \), is negligible. Under orthonormal design \( \frac{1}{n}x_j^TX_{-j} \) is exactly zero, making the entire term zero. If feature \( j \) is independent of all other features then \( \frac{1}{n}x_j^TX_{-j} \) will converge to zero as \( n \) increases, making the term asymptotically negligible provided \( \sqrt{n}(\beta_{-j} - \hat{\beta}_{-j}) \) is bounded in probability, which occurs when \( \lambda = O(n^{-1/2}) \).
When correlation exists between features, \( \frac{1}{n}x_j^T X_{-j} \) will not converge to zero and \( z_j \) will follow a distribution different than \( N(0, 1) \). The empirically estimated mixture density in the denominator of (3.4.1) effectively accounts for this discrepancy, however the numerator does not, resulting in potentially inaccurate estimates. This topic is addressed further in Section 3.5, where we show via simulation that even in the presence of correlation our approach performs well in comparison to both traditional large-scale testing fdr approaches and model based Fdr approaches.

In the ideal setting where features are orthonormal, the mfdr estimator of (3.4.1) shares an important relationship with the mFdr estimator proposed by Bre-heny (2018) which is captured in the following theorem:

**Theorem 2.** Suppose \( z_j \sim N(0, \sigma/\sqrt{n}) \), and define \( Z_- = (-\infty, -\lambda] \), \( Z_+ = [\lambda, \infty) \), and \( \mathcal{M}_\lambda = \{ Z_- \cup Z_+ \} \), Then:

\[
mFdr(\mathcal{M}_\lambda) = \mathbb{E}(mfdr(z)|z \in \mathcal{M}_\lambda)
\]

Theorem 2 states that, on average, the marginal false discovery rate of a model is the average local false discovery rate of its selections. Alternatively, this result implies that the expected number of false discoveries in a model can be decomposed into the sum of each individual selection’s mfdr. Proof of Theorem 2 can be found in the appendix of this chapter.

### 3.4.2 GLM and Cox Models

Now consider the more general case where \( y \) need not be normally distributed. Specifically we focus on the cases of a binary outcome (logistic regression) and survival outcome (Cox regression), although this approached can generally be applied to various other likelihood based models.

Similar to the linear regression setting, we can develop a local false discovery rate estimator by studying minimization of the objective function, \( Q(\beta|X, y) \), as defined in (2.3.1). When \( y \) is not normally distributed, \( Q(\beta|X, y) \) can be minimized
using an algorithm similar to the standard Newton-Raphson algorithm, with the difference being that at each iteration the algorithm solves a penalized re-weighted least squares problem rather than a general least squares problem. This algorithm relies on a two term Taylor Series expansion of $\ell(\beta | X, y)$ centered at a working estimate $\tilde{\beta}$. We also note that $X$ and $\beta$ are related through a linear predictor $\eta = X\beta$; thus we can equivalently express the likelihood in terms of $\eta$ with: $f(\eta | y) = \ell(\beta | X, y)$.

$$\ell(\beta | X, y) \approx l(\beta) + (\beta - \tilde{\beta})^T l'(\tilde{\beta}) + \frac{1}{2} (\beta - \tilde{\beta})^T l''(\tilde{\beta})(\beta - \tilde{\beta})$$

$$\approx \frac{1}{2} (\tilde{y} - \eta)^T f''(\tilde{\eta})(\tilde{y} - \eta) + \text{const.}$$

Here $\tilde{y} = \tilde{\eta} - f''(\tilde{\eta})^{-1} f'(\tilde{\eta})$ serves as a pseudo-response in the weighted least squares expression. Additional details regarding the algorithm and its convergence to the correct solution are given by Simon (2011). We instead direct our focus on the KKT conditions that arise when using this approach, which are very similar to those of the linear regression setting, differing only by the inclusion of a weight matrix $W = f''(\tilde{\eta})$ and the use a pseudo-response $\tilde{y}$:

$$\frac{1}{n} x_j^T W (\tilde{y} - X\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \quad \text{for all } \hat{\beta}_j \neq 0$$

$$\frac{1}{n} x_j^T W (\tilde{y} - X\hat{\beta}) \leq \lambda \quad \text{for all } \hat{\beta}_j = 0$$

From here we are able to follow an argument analogous to that of Section 3.4.1, defining the partial pseudo-residual $r_j = \tilde{y} - X_{-j} \hat{\beta}_{-j}$, which from the KKT conditions implies:

$$\frac{1}{n} |x_j^T W r_j| > \lambda \quad \text{for all } \hat{\beta}_j \neq 0$$

$$\frac{1}{n} |x_j^T W r_j| \leq \lambda \quad \text{for all } \hat{\beta}_j = 0$$

As was the case in Section 3.4.1, we have identified a quantity, $\frac{1}{n} x_j^T W r_j$, that is analogous to a test statistic on selection. Studying this quantity leads to
an estimator of mfdr. To begin we expand the selection condition for $\hat{\beta}_j \neq 0$ by evaluating the pseudo-response at the converged solution, $\hat{\beta}$, under the null hypothesis where $\beta_j = 0$:

$$\frac{1}{n}x_j^TWr_j = \frac{1}{n}x_j^T(WX\hat{\beta} - W^{-1}f'(\hat{\eta}) - X_{-j}\hat{\beta}_{-j})$$

$$= \frac{1}{n}x_j^Tf'(\hat{\eta}) + x_j^TWx_j\hat{\beta}_j$$ (3.4.2)

This implies that selection occurs when $\frac{1}{n}|x_j^Tf'(\hat{\eta}) + x_j^TWx_j\hat{\beta}_j| > \lambda$. Recognizing that $x_j^Tf'(\hat{\eta}) = u_j(\beta)$, where $u_j$ is the $j^{th}$ element of the unpenalized score function, we apply the results of Theorem 1 provided by Miller and Breheny (2018), which shows that under appropriate regularity conditions:

$$\frac{u_j(\hat{\beta}) + v_j\hat{\beta}_j}{\sqrt{v_j}} \xrightarrow{d} N(0, 1)$$

where $v_j = x_j^TWx_j$.

This distributional result suggests the following local false discovery rate estimator:

$$z_j = \frac{u_j(\hat{\beta}) + v_j\hat{\beta}_j}{\sqrt{v_j}}$$

$$\widehat{\text{mfdr}}(z_j) = \frac{\phi(z_j)}{f(z_j)}$$ (3.4.3)

At this point we mention that the estimator in Equation 3.4.3 relies on a vanishing correlation assumption made by Miller and Breheny (2018); consequently it will be accurate in the case of independent features but may be conservative when features are correlated. While independent features are unlikely to occur in practice, simulation results presented in Section 3.5 suggest that the estimator remains useful for analyzing data with correlated features.

### 3.5 Simulation Studies

In this section we conduct a series of simulations studying the behavior of the mfdr estimators in (3.4.3) and (3.4.1), and comparing their performance with
other methods, focusing on comparisons with the traditional univariate local false discovery rates.

We define the traditional fdr procedure to consist of fitting a univariate regression model to each of the \( j \in \{1, \ldots, p\} \) features, extracting the test statistic, \( t_j \), corresponding to the test on the hypothesis that \( \beta_j = 0 \), then normalizing these test statistics such that \( z_j = \Phi^{-1}(Pr(T < t_j)) \), and inputting the normalized statistics into the \texttt{locfdr} function in the R package \texttt{locfdr} (Efron et al., 2015) in order to obtain local false discovery rate estimates.

For our mfdr approach we include results for two different values of \( \lambda \). The first, \( \lambda_{\text{mFdr}} \), characterizes the largest model where the estimated mFdr is less than 10\%. The second, \( \lambda_{\text{CV}} \), characterizes the model with the minimum cross-validated error. These two choices of \( \lambda \) can be viewed as representative of the trade-off that usually occurs between limiting the number of false discoveries and having high predictive power. We look at the cases of linear regression, logistic regression, and Cox regression; and we present results for two data generating scenarios referred to as the “Easy Scenario” , which is designed for \texttt{mfdr} to be highly accurate, and the “Hard Scenario”, which is designed to replicate the challenges of real data.

The Easy Scenario:

- We use a large sample size of \( n = 1000 \)
- Covariate values, \( x_{ij} \), independently generated from the standard normal distribution.
- Response variables are generated by:
  - Linear regression, where \( y = X\beta + \epsilon \) where \( \epsilon_i \sim N(0, \sigma^2) \), \( \beta_{1:60} = 4 \), and \( \beta_{61:600} = 0 \), and \( \sigma = \sqrt{n} \)
  - Logistic regression, where \( y_i \sim \text{bin} \left( 1, \pi_i = \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta)} \right) \), \( \beta_{1:60} = .1 \), and \( \beta_{61:600} = 0 \)
– Cox regression, where \( y_i \sim \exp(\theta_i = \exp(x_i^T \beta)) \), \( \beta_{1:60} = .1 \), and \( \beta_{61:600} = 0 \), and no censoring

**The Hard Scenario:**

In the Hard Scenario we impose a dependence structure motivated by the causal diagram of Figure 3.2. In this diagram variable A has a direct causal relationship with the outcome variable Y, variable B is correlated with Y through its relationship with A, but is not causally related, and variable C is unrelated to all of the other variables and the outcome. The marginal perspective on false discoveries makes no explicit distinction between ‘A’ and ‘B’, however very different results can be observed for variables like ‘B’ depending upon the method used.

![Figure 3.2: Causal diagram showing three types of relationships between variables and the outcome.](image)

With the structure of Figure 3.2 in mind, the Hard Scenario consists of:

- A sample size of \( n = 200 \)

- Covariates generated with the following dependence structure:
  - 6 causative features, which are independent of each other
  - 54 correlated features, grouped such that 9 are related to each causative feature with \( \rho = 0.5 \)
  - 540 noise features, which are correlated with each other by an autoregressive correlation structure where \( \text{cor}(x_j, x_k) = 0.8^{|j-k|} \)
• Response variables are generated from the same distributions as in the Easy Scenario; however, $\beta$ differs to reflect the change in sample size:

- Linear regression, $\beta_{1:6} = (5, -5, 4.5, -4.5, 4, -4)$, and $\beta_{7:600} = 0$, and $\sigma = \sqrt{n}$
- Logistic regression, $\beta_{1:6} = (1.25, -1.25, 1, -1, .75, -.75)$, and $\beta_{7:600} = 0$
- Cox regression, $\beta_{1:6} = (1, -1, .75, -.75, .5, -.5)$, and $\beta_{7:600} = 0$, and no censoring

In summary, in the terms of Figure 3.2, the Hard Scenario consists of 6 features akin to variable A, 54 features akin to variable B, and 540 features akin to variable C. The Easy Scenario consists of 60 features akin to variable A, 0 features akin to variable B, and 540 features akin to variable C. The Hard Scenario, relative to the Easy Scenario, also has a smaller sample size and noise variables with an autoregressive correlation structure.

3.5.1 Calibration

We assess the calibration of $\hat{\text{mfdr}}$ by gauging how well the estimates track with the observed proportion of noise features with that estimate. For example, the estimator is well-calibrated at 0.2 if 20% of features with $\hat{\text{mfdr}} = .2$ are observed to be false discoveries.

Figure 3.3 displays results for the linear regression model in the Easy Scenario, similar results are available in the appendix for other response variables and show no major differences from what is observed in the linear regression setting. We determine calibration by fitting a simple linear regression model using each feature’s estimated mfdr to predict noise variable status, and then comparing the fitted line with a 45-degree line. Our mfdr estimates appear very accurate at both values of $\lambda$, with fitted lines which are nearly indistinguishable from perfect 1-1 correspondence.
Figure 3.3: In the easy scenario the expected proportion of false discoveries at a given estimated local false discovery rate is plotted using a linear model, the estimates produced by each of these methods are very accurate in this scenario.

The traditional univariate method also appears to be very well calibrated, showing only a very slight departure at small estimated fdr values.

Despite being well-calibrated, the three approaches of Figure 3.3 show markedly different feature classification potential. At $\lambda_{CV}$, the estimated mfdr of noise features tends to cluster tightly near 1, while the estimated mfdr of causal features tends to cluster very tightly near 0. In contrast, the traditional univariate method more often provides intermediate estimates for variables of both types. The results at $\lambda_{mFdr}$ fall somewhere in between the results at $\lambda_{CV}$ and results of the univariate fdr approach.

The treatment of correlated ‘B’ variables in the Hard Scenario makes assessing calibration difficult. Strictly speaking, our method is based upon a marginal false discovery definition, meaning that ‘B’ variables are not false discoveries. However our estimator is based upon the correlation between a variable and the partial
residual, so when the corresponding causal ‘A’ variable is active in the model this correlation will be diminished for the ‘B’ variable.

Table 3.1: Local false discovery rate accuracy results for the Hard Scenario. Features are binned based upon their estimated fdr and the observed proportion of noise variables in each bin is reported in the body of the table for each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bin 1</th>
<th>Bin 2</th>
<th>Bin 3</th>
<th>Bin 4</th>
<th>Bin 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>(0, 0.2]</td>
<td>(0.2, 0.4]</td>
<td>(0.4, 0.6]</td>
<td>(0.6, 0.8]</td>
<td>(0.8, 1]</td>
</tr>
<tr>
<td>Univariate fdr</td>
<td>0.12</td>
<td>0.43</td>
<td>0.68</td>
<td>0.84</td>
<td>0.95</td>
</tr>
<tr>
<td>mfdr at $\lambda_{mFdr}$</td>
<td>0.12</td>
<td>0.49</td>
<td>0.66</td>
<td>0.80</td>
<td>0.93</td>
</tr>
<tr>
<td>mfdr at $\lambda_{CV}$</td>
<td>0.05</td>
<td>0.32</td>
<td>0.79</td>
<td>0.91</td>
<td>0.91</td>
</tr>
</tbody>
</table>

| Logisitic | (0, 0.2] | (0.2, 0.4] | (0.4, 0.6] | (0.6, 0.8] | (0.8, 1] |
| Univariate fdr | 0.11 | 0.41 | 0.67 | 0.84 | 0.95 |
| mfdr at $\lambda_{mFdr}$ | 0.06 | 0.36 | 0.57 | 0.77 | 0.93 |
| mfdr at $\lambda_{CV}$ | 0.00 | 0.18 | 0.91 | 0.92 | 0.89 |

| Cox      | (0, 0.2] | (0.2, 0.4] | (0.4, 0.6] | (0.6, 0.8] | (0.8, 1] |
| Univariate fdr | 0.09 | 0.42 | 0.67 | 0.85 | 0.96 |
| mfdr at $\lambda_{mFdr}$ | 0.02 | 0.23 | 0.39 | 0.79 | 0.92 |
| mfdr at $\lambda_{CV}$ | 0.01 | 0.36 | 0.92 | 0.93 | 0.89 |

Table 3.1 displays accuracy results for the hard scenario. Here features are sorted into five equally spaced bins based upon their estimated local false discovery rate under a given procedure. Within in each bin we calculate the proportion of features which are completed unrelated to the outcome, ie: “C” variables. An accurate estimator should result in the proportion of noise features within a bin inside the range defining the bin. The Hard Scenario, represents a situation where we expect the numerator of our estimator be inaccurate, which clearly impacts its calibration. However, this inaccuracy is not unique to our method, the univariate
approach shows similar miscalibration in the intermediate bins. Nevertheless, the estimator does a good job in differentiating important features from noise, which is demonstrated by very few noise features end up in the leftmost bin.

We also see that the choice of $\lambda$ can have a noticeable influence on the accuracy of the estimator. On average the mFdr method chooses larger values of $\lambda$ resulting in models with greater sparsity and making it unsurprising to see results that are similar to those of the traditional univariate procedure. Cross validation, which selects smaller values of $\lambda$, leads to larger models and tends to be accurate in the low and high bins but not in the middle bins. This is in part due to the mfdr approach rarely resulting in intermediate estimates when a relatively large number of variables are active in the model, which leads to small cell counts in the middle bins; but also due to the tendency of mfdr to estimate higher local false discovery rates for “B” variables, an aspect that we will discuss further in the next section.

3.5.2 Comparison with Univariate fdr

In this section we compare the number of feature selections of each type, “A”, “B”, and “C”, for the Easy and Hard Scenarios. As depicted in Figure 3.4, using a local false discovery rate selection threshold of 0.1, leads to the selection of more causally important, “A”, variables in mfdr approach in both scenarios for both choices of $\lambda$. In fact, when using cross validation to select $\lambda$, the mfdr approach selects nearly twice as many “A” variables as the univariate approach in the easy scenario. In the hard scenario mfdr is still more powerful than the univariate approach, selecting on average 16% more “A” variables (4.08 vs. 3.53) than the univariate approach at the $\lambda$ value chosen by cross validation.

In addition to selecting more causally important variables, the mfdr approach drastically reduces the amount of correlated, non-causal features with low local false discovery rate estimates. This is most notable at $\lambda_{CV}$, where the number of “B”
variables with \( \text{fdr} < 0.1 \) is nearly ten times lower than with the univariate approach. Because of the marginal perspective these methods take, the high number of “B” variable selections made by the univariate approach comes with the unfortunate consequence allowing for more noise variable selections. Thus, relative to mfdr, we observe traditional local false discovery rate approaches resulting in fewer “A” variables and more “C” variables.

The results displayed in Figure 3.4 are based upon a somewhat arbitrary threshold of 0.1. While they demonstrate the ability of mfdr to outperform univariate fdr at this threshold, it remains to be shown that mfdr does a better job discriminating between important and unimportant features in a more general sense.

To address this issue we look at the number of false positives, defined as noise features classified as important at a given threshold, and false negatives, defined as important features classified as noise at a given threshold, using the area under the
ROC curve (AUC) to assess each method’s ability to discriminant between important variables and noise variables for various fdr thresholds. Because it is unclear whether the correlated “B” variables in the Hard Scenario should be considered false discoveries, we omit them for the purposes of this comparison.

At the cross validation choice of $\lambda$, the mfdr approach results in average AUC values of 0.943 and 0.935 respectively for the Easy and Hard Scenarios. This is an improvement over the average AUC values of 0.927 and 0.921 that result from the traditional univariate procedure, which demonstrates the general effectiveness of mfdr at a variety of thresholds.

3.5.3 Comparison with Selective Inference and Sample Splitting

As shown in Theorem 2, mFdr and mfdr are linked in a way which allows mfdr to be used to control mFdr, and enables a comparison between the mfdr method and existing methods of Fdr control for lasso models. In this section we use mfdr to select features with $\hat{\text{mfdr}} < 0.1$, which will conservatively control Fdr below 10%, and compare these results with the selective inference approach of Tibshirani et al. (2016) using the ForwardStop rules of G’Sell et al. (2016), which controls the pathwise-wise Fdr at 10%; we also compare our method with the repeated sample splitting method implemented by the hdi package (Dezeure et al., 2015), which is used to control the fully conditional Fdr at 10%.

The results displayed in Table 3.2 demonstrate the conservative nature each of these conditional Fdr control approaches, with the mfdr method selecting far more causal features than any other approach. However, due to their stronger false discovery definitions the other approaches do end up selecting fewer correlated, non-causal features relative to the mfdr approach. This difference is relatively minor, and comes at the expense of only being able to select half as many causal features.
Table 3.2: Simulation results comparing the average number of selections of causal and correlated variables, as well as the proportion of noise variable selections, for various model-based false discovery rate control procedures. The “exact”, “spacing”, “mod-spacing”, and “covtest” methods are related tests performed by the selectiveInference package.

3.5.4 Density Estimation

As mentioned in Sections 3.4.1 and 3.4.2, the density estimation method used to obtain \( \hat{f} \) can influence the estimates of \( \hat{\text{mfdr}}(z) \). In the simulation results shown in this paper we used the default kernel density estimation method of base R, however other density estimation methods have the potential to produce better results. The locfdr package estimates the mixture density by binning \( z \), and then fitting either a natural spline, or a polynomial Poisson regression model, to the counts within each bin. Additional details regarding this estimation procedure are available in Efron (2005).

In additional simulations we explored several different choices of density estimation motivated by the recommendations made by Deng and Wickham (2011). We assessed each of these methods using AUC and found that they all produced relatively similar results, with mean AUC values within roughly 0.02 of each other.
regardless of the density estimation method that was used. Based upon minor improvements in AUC, we recommend the average shifted histogram method (ASH) as implemented by the \texttt{ash} package in R (Scott, 2009), or the binned kernel density estimation using the linear binning method as implemented by the function \texttt{bkde} in the \texttt{KernSmooth} package (Wand and Ripley, 2013) to optimize the performance of \( \hat{\text{mfdr}} \). The complete results for our density estimation comparisons can be seen in the supplemental material.

3.6 Case Studies

3.6.1 Lung Cancer Survival

Shedden et al. (2008) studied the survival of 442 early-stage lung cancer subjects, 236 experienced the outcome event of death during the study. Researchers collected high-dimensional gene expression data consisting of 22,283 genetic features as well as additional clinical covariates of age, race, gender, smoking history, cancer grade, and whether or not the subject received adjuvent chemotherapy. The goal of our analysis is to identify genetic features that are associated with survival after adjusting for the clinical covariates.

We first analyze the data using the traditional univariate fdr approach, which is based upon the test statistics from 22,283 separate Cox regression models, where each of these models contains a single genetic feature in addition to the clinical covariates. While each of these models contains more than one variable, we will refer to this method as a univariate approach to remain consistent with our prior terminology. We compare these results with those obtained using the mfdr approach where clinical covariates are included into the model unpenalized. We consider both cross validation and marginal false discovery rates as methods of selecting \( \lambda \).

Cross validation selects a \( \lambda \) value of 0.119, while controlling the marginal false discovery rate below 10% results in a \( \lambda \) value of 0.157. This corresponds with 16
and 2 genetic features being selected into each respective regression model.

<table>
<thead>
<tr>
<th>Feature</th>
<th>ZC2HC1A</th>
<th>FAM117A</th>
<th>CSRP1</th>
<th>PDPK1</th>
<th>SCGB1D2</th>
<th>BSDC1</th>
<th>ARHGEF2</th>
<th>AFFX-M27830</th>
<th>ETV5</th>
<th>Hs28SrRNA-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate fdr</td>
<td>0.00075</td>
<td>0.00111</td>
<td>0.00461</td>
<td>0.00641</td>
<td>0.00647</td>
<td>0.00658</td>
<td>0.00667</td>
<td>0.00699</td>
<td>0.00715</td>
<td>0.00787</td>
</tr>
<tr>
<td>mFdr at $\lambda_{mFdr}$</td>
<td>0.00424</td>
<td>0.02221</td>
<td>0.02632</td>
<td>0.03827</td>
<td>0.03867</td>
<td>0.03876</td>
<td>0.03910</td>
<td>0.04356</td>
<td>0.04364</td>
<td>0.04532</td>
</tr>
<tr>
<td>$\hat{\beta}_j \neq 0$</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>0.00789</td>
<td>0.03732</td>
<td>0.05094</td>
<td>0.16775</td>
<td>0.16784</td>
<td>0.17916</td>
<td>0.18398</td>
<td>0.18827</td>
<td>0.23666</td>
<td>0.24182</td>
</tr>
</tbody>
</table>

Table 3.3: Local false discovery rate estimates of the top ten features, when performing univariate testing, for the Shedden survival data.

The ten features with the lowest local false discovery rates for each approach are displayed in Table 3.3. We observe that the features ZC2HC1A, SCGB1D2, FAM117A, and BSDC1 appear in the top ten for all approaches, albeit with different estimates. The estimates for the univariate fdr and the mFdr approach at $\lambda_{mFdr}$ tend to be smaller than those at $\lambda_{CV}$, a result that is likely attributed to how each approach behaves in the presence of correlated variables. As demonstrated in the simulation results of Section 3.5, the mFdr approach at $\lambda_{CV}$ tends to prevent features from achieving low fdr estimates through their correlation with other variables.

An interesting aspect of mFdr for GLM/Cox Regression is that $z_j$ isn’t monotonically related to $\hat{\beta}_j$ across features due to estimator’s denominator, $\sqrt{x_j^TWx_j}$. This enables features that were not selected by the lasso to potentially have smaller local false discovery rate estimates than some features which were selected. The feature SCGB1D2 is an example of this phenomenon, it has the second smallest local marginal false discovery rate estimate at $\lambda_{mFdr}$ despite not being active in the lasso model. Interestingly it is active at $\lambda_{CV}$, and is one of three features in that model with local false discovery rates below 10%. It also appears among the top
univariate feature selections where it has the fifth smallest univariate fdr estimate.

Figure 3.5: The mixture density estimates, $\hat{f}(z)$, for the different methods applied to the Shedden data. We observe that the distribution of test statistics more closely resembles the null as more features are adjusted for by the model.

To further investigate these differences we apply a local false discovery rate threshold of $\alpha = 0.10$. This leads to the selection of 293 features using the univariate approach, 31 features using the mfdr approach at $\lambda_{mFdr}$, and 3 features using the mfdr approach at $\lambda_{CV}$. These differences demonstrate the usefulness of mfdr’s ability to limit the number of correlated, non-causal features that are deemed significant.

Figure 3.5 shows the estimated mixture density, $\hat{f}$, for each method. With the univariate approach, which does not account for any correlations between features, we see that the distribution of univariate test statistics is most different from the null distribution. In the lasso model where $\lambda$ is selected by mFdr, the genetic features ZC2HC1A and FAM117A are adjusted for by the model, and consequently the mixture distribution narrows relative to that of the univariate approach. When cross validation is used to select $\lambda$ there are 16 genetic features that are adjusted for
by the model, and the mixture distribution narrows even further to the point where it closely resembles the null. These observations are in line with the simulation results of Section 3.5, which suggest that cross validation should be used to choose \( \lambda \) if the goal is to eliminate correlated features.

### 3.6.2 BRCA1 Gene Expression

Our second case study looks at the gene expression of breast cancer patients from The Cancer Genome Atlas (TCGA) project. The data set is publicly available at [http://cancergenome.nih.gov](http://cancergenome.nih.gov), and consists of 17,814 gene expression measures for 536 subjects. One of these genes, BRCA1, is a tumor suppressor that is widely regarded as playing a causal role in the development of breast cancer. When BRCA1 is under-expressed the risk of breast cancer is significantly increased, which makes genes that are related to BRCA1 expression interesting candidates for future research. As is the case in many genetic data scenarios, we expect a large number of genes to have non-causal, correlated relationships with BRCA1. This makes univariate approaches potentially undesirable due to their tendency to select all of these correlated genes. Unsurprisingly, when using an fdr threshold of 0.10, the univariate fdr approach selects 8,431 genes, too many features to look at individually. Additionally, most of these selections are likely to have correlated, non-causal relationships with BRCA1.

![Causal Diagram](Figure 3.6: Causal diagram depicting possible gene relationships with BRCA1.)

An alternative approach would be to use a lasso regression model to jointly...
model the relationship between BRCA1 and the 17,813 remaining genes. However, cross validation selects a model which contains 116 features and a very high false discovery rate, our method estimates the average mfdr of these features to be 0.827. Using mFdr to choose $\lambda$ is one way of obtaining a more reliable set of features, but this model adjusts for genes and is considerably less predictive. The mfdr approach offers another alternative that can be directly applied to the most predictive model, resulting in 16 significant genes with local false discovery rates under 10%.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Univariate $\hat{\text{fdr}}$</th>
<th>$\hat{\text{mfdr}}$ at $\lambda_{\text{mFdr}}$</th>
<th>$\hat{\text{mfdr}}$ at $\lambda_{\text{CV}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C17orf53</td>
<td>1.98E-37</td>
<td>0.000697</td>
<td>1</td>
</tr>
<tr>
<td>TUBG1</td>
<td>9.34E-33</td>
<td>0.204282</td>
<td>1</td>
</tr>
<tr>
<td>DTL</td>
<td>1.52E-31</td>
<td>2.41E-10</td>
<td>0.000082</td>
</tr>
<tr>
<td>VPS25</td>
<td>1.32E-30</td>
<td>0.000014</td>
<td>0.025972</td>
</tr>
<tr>
<td>TOP2A</td>
<td>2.12E-30</td>
<td>0.000030</td>
<td>0.021379</td>
</tr>
<tr>
<td>PSME3</td>
<td>1.05E-29</td>
<td>0.000022</td>
<td>0.002878</td>
</tr>
<tr>
<td>TUBG2</td>
<td>6.07E-29</td>
<td>0.314260</td>
<td>1</td>
</tr>
<tr>
<td>TIMELESS</td>
<td>4.30E-27</td>
<td>0.139503</td>
<td>1</td>
</tr>
<tr>
<td>NBR2</td>
<td>1.43E-26</td>
<td>8.65E-42</td>
<td>4.26E-44</td>
</tr>
<tr>
<td>CCDC43</td>
<td>8.18E-26</td>
<td>0.146387</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.4: The top 10 selected genes from the univariate approach and their local false discovery rate estimates for each method.

Table 3.4 shows the 10 genes with the lowest univariate local false discovery rates, along with their mfdr estimates at $\lambda_{\text{mFdr}}$ and $\lambda_{\text{CV}}$. Many of the top selections made by the univariate approach with the lowest fdr estimates have biological roles that make them unlikely to be causally related with BRCA1. For example TUBG1 and TUBG2 both have high mfdr estimates in the lasso based approaches but have
extremely low fdr estimates in the univariate approach. Biologically, their role is to mediate microtubule formation and progression of the cell cycle. However, they are located on chromosome 17, the location of BRCA1, and are highly correlated with many other genes on chromosome 17, including: NBR2, VPS25, TOP2A, PSME3, and CCDC43. Given their functions and locations, it seems somewhat unlikely for TUBG1 and TUBG2 to be actively driving BRCA1 expression, which is suggested by the mfdr results but not the univariate testing results.

Other selections that have low local false discovery rates in both the univariate and lasso approaches have very plausible relationships with BRCA1. For example PSME3 encodes a protein that is known to interact with p53, a protein that is widely regarded as playing a crucial role in cancer formation (Zhang and Zhang, 2008). Another example is DTL, which interacts with p21, a protein that is also known to have a role in cancer formation (Abbas et al., 2008). These results demonstrate the potential of the mfdr approach to identify more scientifically relevant relationships by reducing the number of correlated, non-causal feature selections.

3.7 Discussion

The local approach to false discoveries for lasso regression models provides useful way of quantifying the reliability of individual feature selections after a model is fit. The estimator can be quickly computed, even in high dimensions, using only quantities that can be easily obtained from a fitted lasso model. This makes it a natural next step when there is an interest in the reliability of the variables selected by lasso model. The method is currently implemented in the R package ncvreg (Breheny and Huang, 2011) using the summary function. By default this will report local false discovery rates for all variables selected at a given value of $\lambda$, but includes options to report only variables which meet a specified mfdr threshold or report only a specified number of selections ordered by lowest mfdr.
While the estimate works well in most situations where regression would typically be employed, one setting where it does not work well is when analyzing small sample case-control genetic studies such as the HIV study that was used to motivate local false discovery rates in Efron’s 2005 paper. However, with a moderate sample size our approach offers advantages over the univariate approach, including increased power and a potential ability to reduce the selection rates of correlated, non-causal features, although we make the caveat that, as with many methods, there is an inherent identifiability issue in distinguishing causal from correlated non-causal features.

3.8 Appendix

3.8.1 Proof of Theorem 2

Proof. The tail areas defined by $Z_-$ and $Z_+$ are disjoint such that:

$$
E(\text{mfdr}(z) | z \in M_\lambda) = E(\text{mfdr}(z) | z \in Z_-) + E(\text{mfdr}(z) | z \in Z_+)
$$

Beginning with the left tail characterized by $Z_-:

$$
E(\text{mfdr}(z) | z \in Z_-) = \int_{Z_-} \text{mfdr}(z) Pr(\text{mfdr}(z) | z \in Z_-) \, dz
$$

$$
= \int_{Z_-} \frac{\pi_0 f_0(z)}{f(z)} \left( \frac{f(z)}{\int_{Z_-} f(z) \, dz} \right) \, dz
$$

$$
= \frac{\pi_0 F_0(-\lambda)}{F(-\lambda)}
$$

Define $|S_-|$ as $E(\# z_j < -\lambda)$, and recognize that empirically $F(-\lambda) = \frac{|S_-|}{p}$ and $F_0$ is CDF of $N(0, \sigma^2/n)$, thus:

$$
E(\text{mfdr}(z) | z \in Z_-) = \frac{\pi_0 p \phi(-\sqrt{n}\lambda/\sigma)}{|S_-|}
$$

Similarly, define $|S_+|$ as $E(\# z_j > \lambda)$, and recognize $|S| = |S_-| + |S_+|$. By following an analogous derivation for the right tail, $Z_+$, and using the symmetry of
the standard normal CDF we arrive at:

$$E(\text{mfdr}(z)|z \in \mathcal{M}_\lambda) = \frac{2\pi_0 p \Phi(-\sqrt{\lambda}/\sigma)}{|S|}$$

which, when $\pi_0 = 1$, is the definition of $\text{mFdr}(\mathcal{M}_\lambda)$.

3.9 Supplemental Results

3.9.1 Calibration

![Smoothed calibration results for the Cox regression Easy Scenario.](image)

Figure 3.7: Smoothed calibration results for the Cox regression Easy Scenario.

In the Easy Scenario, the calibration of $\widehat{\text{mfdr}}$ for logistic and Cox regression is essentially identical to what is seen in the linear regression setting. The estimates are well calibrated for both choices of $\lambda$, as well as for the univariate approach.

As previously mentioned, the treatment of correlated ‘B’ variables in the Hard Scenario makes assessing calibration difficult. Strictly speaking, our method is based
Figure 3.8: Smoothed calibration results for the logistic regression Easy Scenario.

upon a marginal false discovery definition, meaning that ‘B’ variables are not false discoveries. However our estimator is based upon the correlation between a variable and the partial residual, so when the corresponding causal ‘A’ variable is active in the model this correlation will be diminished for the ‘B’ variable. This behavior is markedly displayed when contrasting Figure 3.9 with Figure 3.12.

Figures 3.12, 3.12, 3.12, and 3.12 show calibration results for the Hard Scenario for logistic and Cox regression. Relative to linear regression, calibration appears to be much better for Cox and logistic regression when ‘B’ variables are included. We suspect that this is attributable to the relative lack of features with intermediate mfd estimates in the linear regression setting, which leads to the generalized additive model (GAM) smoothing method being heavily influenced by the large number of estimates near 0 and 1. Because there are so few points in the middle of the plot, there is little consequence for the GAM smoother not accurately capturing trends.
Figure 3.9: Smoothed calibration results for the linear regression Hard Scenario, ‘B’ variables are included as valid discoveries in this plot.

3.9.2 Density Estimation

Deng and Wickham (2011) evaluated numerous methods of density estimation that are currently available in R. The authors describe three broad approaches of density estimation: histogram-based, kernel-based, and penalization-based. Motivated by their assessments we consider the performance of $\hat{mfdr}$ using four different methods to estimate the mixture density: the kernel-based approach implemented in the default `density` function in base R, the average shifted histogram approach implemented by the `ash` package, the kernel-based approach implemented by the `KernSmooth` package, and the penalized likelihood approach implemented by the `logspline` package. We evaluate the performance of these methods based upon area under the ROC and precision-recall (PR) curves using data generated under
both the Easy and Hard scenarios. For the Hard Scenario, ‘B’ variables are omitted from the calculations of ROC and PR. While these comparisons are not an exhaustive examination of all available means of density estimation, they represent a variety of commonly recommended methods which encompass differing approaches to density estimation.

Figures 3.13 and 3.14 display ROC and PR results over 200 simulated data sets for the easy and hard scenarios. Regardless of the density estimation method the mean AUC tends to differ by less than a few percentage points for the ROC and PR curves in each scenario. We note that the ash and bkde methods tend perform slightly better and we recommend these methods if optimal performance is desired; however, for practical purposes, the default density function in R appears adequate.
Figure 3.11: Smoothed calibration results for the Cox regression Hard Scenario, ‘B’ variables are included as valid discoveries in this plot.

Figure 3.12: Smoothed calibration results for the logistic regression Hard Scenario, ‘B’ variables are included as valid discoveries in this plot.
Figure 3.13: The ROC and PR results for various methods of density estimation applied to the 200 replicated data sets from the easy scenario.

Figure 3.14: The ROC and PR results for various methods of density estimation applied to the 200 replicated data sets from the hard scenario, 'B' variables are omitted from these calculations.
4.1 Abstract

In many modeling applications the predictor variables are not unrelated entities, but instead are members of structured groups. The group lasso is a powerful tool for fitting regression models that preserve group structures and are sparse at the group level. This paper focuses on false discovery rate methods for grouped data. Specifically, we propose marginal false discovery rate methods for the group lasso which can be easily implemented, at little added computational cost, after fitting a group lasso model. Our methods adopt a weaker false discovery definition than some regression based approaches, however we argue that they provide convenient and informative next steps for performing inference after fitting a group lasso model. We demonstrate these advantages using several simulation studies and a real data case study involving the identification of non-linear genetic effects in Bardet-Biedl syndrome.

4.2 Introduction

Frequently in modeling applications the predictor variables are not unrelated entities, but are instead members of structured groups. Some common examples include indicator variables representing the levels of a categorical variable, or basis expansions representing the non-linear effects of continuous variable. Grouping can also be used to incorporate relevant external information into the model, providing enhanced prediction or interpretation; for example genetic features can be grouped based upon scientifically known gene pathways. In many of these applications the number of available predictors is large and variable selection is necessary. Methods
such as the lasso (Tibshirani, 1996) are an attractive tool for selecting individual predictors using penalization, and the group lasso (Yuan and Lin, 2006) extends these ideas to perform selection at the group level by penalizing entire groups rather than individual predictors. Group lasso models are sparse at the group level, with all members of selected groups having non-zero effects. Variants such as the sparse group lasso (Simon et al., 2013) allow for another layer of sparsity within selected groups; however our focus will be only on group level selection.

The amount of sparsity in a group lasso model depends upon a tuning parameter $\lambda$. Most popular methods for choosing $\lambda$, such as cross validation, make no attempt to limit the false discovery rate and can result in models containing a substantial number of unimportant groups. In order to address this concern, this work extends that of Breheny (2018); Miller and Breheny (2018) by developing marginal false discovery rate bounds for group lasso models (mFdr), which can be used to assess the number of false discoveries expected to occur in a model by random chance.

We also develop local marginal false discovery rate estimates for individual group lasso selections (mfdr), which can be used to assess the probability that a given group selection is a false discovery. These mfdr estimates have the advantage of applying to any model, which we later demonstrate is particularly valuable given the tension that can exist between the choosing the most predictive model and choosing a model with a reasonable mFdr.

The mFdr and mfdr approaches parallel existing tail-area based false discovery rates (Fdr) and local false discovery rates (fdr) proposals in the large-scale testing literature, a vast and actively growing area of research. The seminal proposal of Benjamini and Hochberg (1995) sparked considerable interest in the area and has since lead many others to study tail-area false discovery rates. Efron and Tibshirani (2002) made an important extension, using an empirical Bayes framework to develop local false discoveries rates, thereby allowing false discovery rate inference
on individual hypotheses rather than entire sets. Storey et al. (2004) and Strimmer (2008b) build off of these proposals, establishing several connections between tail-area and local approaches. Several others have also made contributions in this area; Farcomeni (2008) provides a detailed review. While much of this research has been based upon single variable tests, the methods can easily be applied to single group tests.

It is only recently that false discovery rates have been considered in the realm of penalized regression modeling, with nearly all of the research studying ungrouped models. This research is complicated by the question of how to define a false discovery in the regression setting where many variables or groups of variables are considered simultaneously. One option, which mirrors the definition most commonly used in the large-scale testing literature, is to adopt marginal perspective where a selected group is considered a false discovery if the group is marginally unrelated with \( Y: X_j \perp \perp Y \) (Breheny, 2018). Our proposed mFdr and mfdr methods adopt this marginal perspective.

More stringent false discovery definitions are possible when considering conditional relationships between variables. Data splitting approaches, such as the sample-splitting method of Wasserman and Roeder (2009) and the multiple-split extension by Meinshausen et al. (2009), adopt a stronger, fully-conditional definition where a group is considered a false discovery if it unrelated to the outcome, conditional upon all other groups, ie: \( X_j \perp \perp Y|X_k \neq j \). Data splitting approaches can easily extend to the group lasso, although we are not aware of any formal applications that focus on this setting. Another approach is to adopt a partially conditional definition which conditions upon the selected model such that false discoveries are defined as: \( X_j \perp \perp Y|X_k \) for \( k \in M_j \), where \( M_j \) is the model under consideration. This definition is adopted by the selective inference approach of Lockhart et al. (2014); Tibshirani et al. (2016) which can be used with the stopping rules of G’Sell
et al. (2016) to achieve pathwise conditional false discovery rate control for forward stepwise regression and the ungrouped lasso models. Selective inference has been extended by Loftus and Taylor (2015) to groups of variables, however the work focuses primarily on grouped selections in forward stepwise regression and not the group lasso.

Although our methods adopt the same false discovery rate definition used in large-scale testing, the focus of this paper is not on one-at-a-time testing but rather on modeling using the group lasso. Specifically we seek to address two questions that arise after fitting a group lasso model. The first being “how many false discoveries are expected in the a model by random chance?”, and the second being “how likely is selected group to be a false discovery?”. After presenting our methods, we use simulation to assess their accuracy and to compare them with traditional large-scaling testing approaches, data splitting approaches, and selective inference. However, despite making comparisons with these methods, we reiterate that many aspects of our methods are fundamentally different from those approaches. Our goal is not to show superiority of our methods, but rather to illustrate the additional information which can be gleaned from a fitted group lasso model with relatively little additional computation.

4.2.1 The Group lasso

Consider data of the form \((X, y)\), where \(X\) is an \(n\) by \(p\) design matrix and \(y\) is a vector of outcomes. We assume the outcomes arise from the linear model \(y = X\beta + \epsilon\), where the errors, \(\epsilon\), are independent identically normally distributed with mean 0 and variance \(\sigma^2\), and \(\beta\) is the vector of true regression coefficients. We assume the columns of \(X\) have been standardized such that each variable has a mean of 0 and \(\sum_i x_j^2 = n\). Let \(\{X_1, X_2, \ldots, X_J\}\) represent sub-matrices of \(X\) corresponding to non-overlapping groups \(j \in \{1, \ldots, J\}\), where \(X_j\) is an \(n\) by \(K_j\)
matrix whose columns contain the features in group $j$. Similarly let $\beta_j$ represent the $K_j$-dimensional sub-vector of regression coefficients corresponding to the features in group $j$. We focus on the group lasso solution, which is defined as the minimizer, with respect to $\beta$, of the objective function:

$$Q(\beta|X, y) = \frac{1}{n}||y - X\beta||_2^2 + \sum_j \lambda_j||\beta_j||_2$$  \hspace{1cm} (4.2.1)

The tuning parameter $\lambda_j$ is allowed to vary in accordance to group size. Throughout the remainder of this paper we use $\lambda_j = \sqrt{K_j}\lambda$; however, other choices are possible.

In solving the group lasso it is convenient to orthonormalize each group such that $X_j = \tilde{X}_j R_j$ and $\frac{1}{n}\tilde{X}_j^T \tilde{X}_j = I$. This can be done using the singular value decomposition: $\frac{1}{n}X_j^T X_j = U_j \Lambda_j U_j^T$, which corresponds to $R_j = U_j \Lambda_j^{-1/2}$. Orthonormalizing the groups implies the following objective function:

$$Q(\theta|\tilde{X}, y) = \frac{1}{n}||y - \tilde{X}\theta||_2^2 + \sum_j \lambda_j||\theta_j||_2$$  \hspace{1cm} (4.2.2)

Where $\theta$ represents the coefficient vector on the orthonormal scale, and can easily be used to obtain the coefficients on the original scale using $\beta_j = R_j^{-1}\theta_j$. We mention here that the minimization problem defined by Equation 4.2.2 differs from that of Equation 4.2.1, however both result in identical solutions provided $K_j < n$ for all $j \in \{1, \ldots, J\}$. Simon and Tibshirani (2004) provides a detailed description of how these two group lasso formulations are related.

The group lasso achieves group level sparsity because its penalty term is not strictly differentiable at 0. This requires the use of subdifferentials to characterize the coefficient estimates which minimize Equation 4.2.2, and the resulting sets of equations are known as the Karush-Kuhn-Tucker (KKT) conditions in the convex optimization literature. The KKT conditions are necessary and sufficient for $\hat{\theta}$ to
minimize (4.2.2); they state (Friedman et al., 2010; Breheny and Huang, 2015):

\[
\frac{1}{n} \tilde{X}_j^T (y - \tilde{X}\hat{\theta}) = \lambda_j s_j, \quad \text{where: } s_j = \begin{cases} 
\frac{\theta_j}{||\theta_j||_2}, & \text{if } \theta_j \neq 0 \\
v : ||v||_2 \leq 1, & \text{if } \theta_j = 0
\end{cases}
\]  

(4.2.3)

The KKT conditions define the entry criteria for active groups in the group lasso model. To illustrate, define the partial residual, for group \( j \), to be: \( r_j = y - \tilde{X}_j\hat{\theta}_j \). In order for group \( j \) to be selected, the KKT conditions imply that

\[
\frac{1}{n} \tilde{X}_j^T r_j - \hat{\theta}_j = \lambda_j \frac{\hat{\theta}_j}{||\hat{\theta}_j||_2}.
\]

After some minor algebraic manipulation this leads to the selection condition:

\[
\frac{1}{n} ||\tilde{X}_j^T r_j|| > \lambda_j^2 \tag{4.2.4}
\]

For the standardized group lasso the selection condition in Equation 4.2.4 shares an interesting connection with the uniformly most powerful invariant (UMPI) test used for testing the significance of a group of variables in unpenalized ordinary least squares regression (Simon and Tibshirani, 2004). Let \( \hat{y}_{-j} \) denote the fitted values after omitting group \( j \) from the model, The UMPI test declares significance for group \( j \), at level \( \alpha \), if:

\[
||y - \hat{y}_{-j}||_2^2 \geq \sigma^2 \chi^2_{K_j, 1-\alpha}
\]

The squared difference, \( ||y - \hat{y}_{-j}||_2^2 \), can be expressed as \( ||P_{\text{col}(x_j)} r_j||_2^2 \), where \( P_{\text{col}(x_j)} \) denotes orthogonal projection onto the column space of \( X_j \). In the case of orthonormalized groups \( P_{\text{col}(\tilde{x}_j)} = \tilde{X}\tilde{X}^T \) and \( ||\tilde{X}\tilde{X}^T r_j||_2^2 \propto ||\tilde{X}^T r_j||_2^2 \), thus the expression in Equation 4.2.4 corresponds directly with the UMPI test. This connection provides a promising starting point for assessing the role of random chance in group lasso selections.
4.3 Group lasso mFdr

In this section we use the KKT conditions to derive an upper bound for the number and rate of marginal false discoveries in a group lasso model. While we refer to our results as “upper bounds” they are actually unbiased under ideal conditions where there are no correlations between groups, however in overwhelming majority of real data applications they will be overestimates.

The basic intuition behind the approach is that under the null hypothesis that group $j$ is marginally independent of $y$ the probability of a group being selected into the group lasso model can be expressed as $Pr(\frac{1}{n}||\tilde{X}_j^T(y - \tilde{X}_{-j}\hat{\theta}_{-j})||_2 > \lambda_j^2)$, and under certain regularity conditions we are able to determine this probability. This probability will be identical for groups of the same size, thereby allowing for an upper bound on the number of false group selections and consequently the marginal false discovery rate that requires little computational burden.

To proceed in our development of the mFdr upper bound, we assume the following conditions:

(A1) Vanishing correlation between groups: $\lim_{n \to \infty} \frac{1}{n} \tilde{X}_j^T \tilde{X}_{-j} = 0$.

(A2) Estimation consistency: $\sqrt{n}(\hat{\theta} - \tilde{\theta})$ is bounded in probability.

These conditions are not trivial. (A2) has been studied and shown to hold for the group lasso under certain conditions. (A1) is unlikely to be fully satisfied in most real data applications, however it represents somewhat of a worst case scenario for false discoveries. Correlation reduces the unique information available contained across the groups, on average leading to fewer false discoveries. This leads to our proposal overestimating the true false discovery rate in the presence of correlation.

To begin, recall that $r_j = y - \tilde{X}_{-j}\hat{\theta}_{-j}$, denotes the partial residual that results from the removal of group $j$. It follows directly from the KKT conditions that:

$$Pr(\hat{\theta}_j \neq 0) = Pr(\frac{1}{n}||\tilde{X}_j^T r_j||_2 > \lambda_j^2)$$
When $\theta_j = 0$, as is the case when $\tilde{X}_j$ is marginally independent of $y$:

$$||\tilde{X}_j^T r_j||_2^2 = ||\tilde{X}_j^T \epsilon + \tilde{X}_j^T \tilde{X}_{-j}(\theta_{-j} - \hat{\theta}_{-j})||_2^2$$

Under (A1) and (A2), $\frac{1}{\sqrt{n}} \tilde{X}_j^T \tilde{X}_{-j}(\theta_{-j} - \hat{\theta}_{-j}) \to 0$. Additionally, $\frac{1}{\sqrt{n}} \tilde{X}_j \epsilon \to N(0, \sigma^2 I_{K_j})$. Together these imply $\frac{1}{n\sigma^2}||\tilde{X}_j^T \epsilon||_2^2 \to \chi^2_{K_j}$, which allows for the probability of a null group $j$ being selected by the group lasso as a marginal false discovery to expressed:

$$Pr(\chi^2_{K_j} > \frac{n\lambda_j^2}{\sigma^2})$$

This distributional result suggests the following upper bounds on the number marginal false discoveries, and corresponding the marginal false discovery rate:

$$\hat{FD} = 2 \sum_{j=1}^{J} Pr(\chi^2_{K_j} > \frac{n\lambda_j^2}{\sigma^2})$$

$\hat{mFdr} = \frac{\hat{FD}}{|S|}$, (4.3.1)

Where $S$ is the set of selected groups and $|S|$ its size. Because $J$ is an upper bound for the number of null groups, $\hat{FD}$ and $\hat{mFdr}$ will be somewhat conservative even if (A1) and (A2) are fully satisfied. An appeal of $\hat{mFdr}$ is that it can easily computed for any fitted group lasso model, as it only requires $\lambda$, the size of each group, and an estimate of $\sigma^2$.

### 4.4 Group lasso mFdr

The upper bounds of (4.3.1) provide convenient summary measures for a particular group lasso model; however they are not informative at the individual group level. Phrased differently, it is unlikely that each selected group has the same chance of being a false discovery. To address this question we turn to local false discovery rates.

To begin we return to the KKT conditions, letting $t_j = \frac{1}{n\sigma^2}||\tilde{X}_j^T r_j||_2^2$. The
statistic $t_j$ governs the selection of group $j$ into the group lasso model, and follows a $\chi^2_{K_j}$ distribution under the null hypothesis when conditions (A1) and (A2) of Section 4.3 are satisfied. Next $t_j$ is mapped to the normal scale such that $z_j = \Phi^{-1}\left(\frac{1}{2}(1 - F(t_j, K_j))\right)$, where $\Phi$ represents the standard normal cumulative density function (CDF), and $F(t_j, K_j)$ represents the $\chi^2_{K_j}$ CDF evaluated at $t_j$.

Applying the empirical Bayes arguments of Efron et al. (2001) to $z_j | j \in \{1, \ldots, J\}$, the collection of normalized statistics, suggests the following group level local false discovery rate estimates:

$$\hat{mfdr}(z_j) = \hat{\pi}_0 \phi(z_j) \hat{f}(z_j) \quad (4.4.1)$$

Where $\hat{\pi}_0$ is an estimate of the fraction of truly null groups. It is often convenient to use a conservative estimate $\hat{\pi}_0 = 1$, which we will use throughout the remainder of the paper. $\hat{f}(z_j)$ is an empirical estimate of the density of $z_j$, a mixture of null and non-null densities. Density estimation is a broad area of statistics with direct implications on (4.4.1). For convenience we estimate $\hat{f}$ using kernel density estimation using the default density function in R, however we acknowledge that other methods of density estimation might be capable of producing better results.

The appeal of (4.4.1) is the group specific information it provides. The basic intuition behind its formulation is that $z$ follows a standard normal distribution for null groups. So if the empirically observed mixture density deviates from standard normal it is an indication that non-null groups are present. The density ratio at a particular value, $z_j$, quantifies how likely it is that the corresponding group could arise from the null density component of what was empirically observed. Efron (2012) provides a more technical explanation of local false discovery rates.
4.5 Simulations

To illustrate the results of Sections 4.3 and 4.4 we carry out several simulation studies, focusing on two common applications of the group lasso. The first is the categorical variable scenario, where groups are comprised of multiple indicator variables. The second is the external information scenario, where variables are continuous but are scientifically grouped based upon suspected similar relationships with the outcome. In the majority of these scenarios we hold number of groups constant at 50, and the number of non-null groups constant at 5. We generate our outcomes from the model \( Y \sim N(X\beta, \sigma^2) \), where for the non-null groups, \( j \in \{1, \ldots, 5\} \), \( \beta_j = (b/\sqrt{n}, \ldots, b/\sqrt{n}) \), where \( b \) is a constant, and for the null groups, \( j \in \{6, \ldots, 50\} \), \( \beta_j = (0, \ldots, 0) \). Throughout our simulations we vary aspects including group size, sample size, and effect size \( (b) \) in order to investigate the properties of our proposed mFdr and mfdr methods.

4.5.1 mFdr Convergence

The mFdr bound of Section 4.3 is an asymptotic result which is based upon assumptions of vanishing correlation between groups and estimation consistency of \( \hat{\beta} \). In finite sample situations spurious correlations between independent groups are common, and will cause the mFdr bound to be conservative. Additionally, larger group sizes increase the potential for these chance correlations, potentially exacerbating the problem. To investigate the finite sample behavior of the mFdr bound we simulate data and fit a group lasso model using a fixed \( \lambda \) sequence. For each value of \( \lambda \), we calculate the average observed false discovery rate and the corresponding average mFdr bound, repeating this process for a range of sample sizes.

Figure 4.1 shows the impact of sample size on the mFdr bound in both the categorical variable and external information scenarios. In each of these scenarios
mFdr exhibits similar trends; namely that as the sample size increases the bound approaches the true marginal false discovery rate, and the bound tends to be tighter for smaller group sizes, with the difference in group size becoming less influential as \( n \) increases. We will further discuss sample size, group size, and additional reasons for the mFdr bound being conservative in Section 4.5.4.

### 4.5.2 mfdr Calibration

The mfdr estimator of Section 4.4 provides very useful interpretations of individual group selections. However, the validity of these interpretations depends upon the estimator being well-calibrated. In other words, it is important to establish that roughly 20% of groups with mfdr estimates of around 0.2 end up being false discoveries. The calibration plots in Figure 4.2 display the estimated mfdr's of null and non-null groups and the simple linear regression line associating these estimates with the group’s true status. We include these estimates for two choices of \( \lambda \): the value selected by cross validation, and the smallest \( \lambda \) such that the model’s
estimated mFdr is below 20%. In order to establish a baseline for comparison, we include a large-scale testing approach which fits a separate least squares model to each group in a one-at-a-time fashion and then applies local false discovery rate methodology to the classical F-test of these model’s fit.

![Figure 4.2](image)

**Figure 4.2:** The estimated group local false discovery rates when $K_j = 5$ plotted versus each group’s null/non-null status. A linear model is used for smoothing such that each method’s calibration can be assessed by comparing the respective fitted line with the 45° dashed line.

The simulation results depicted in Figure 4.2 indicate fairly good calibration for each method. For groups having very low local false discovery rate estimates there is a slight liberal bias, an effect which is most pronounced in the large-scale testing approach. Nevertheless, the results in Figure 4.2 are reassuring in that they indicate that our mFdr estimates tend to accurately reflect the likelihood that a group is a false discovery. Additionally, relative to large-scale testing, the mFdr approach does a slightly better job assigning high false discovery rate estimates to noise groups. This can be difficult to see in Figure 4.2 where 93.6% of noise variables have false discovery rate estimates larger than 0.75 using the univariate approach, compared to 94.6% and 95.5% using the mFdr approach.
4.5.3 Power Comparisons

In this simulation study we investigate the power of mFdr and mfdr to identify important groups while controlling the marginal false discovery rate. We compare the selection results for several different methods that can be applied to group level selection, we define each of these methods as follows:

1. **large-scale testing**: For each group we fit a separate linear model using ordinary least squares, and then test for significance using the classical F-test against the intercept only model. The resulting test statistics are normalized and used as inputs for the traditional local false discovery rate methodology, which is used to identify significant groups with local false discovery rate estimates less than 10%.

2. **selective inference**: Using the entire collection of groups we apply forward stepwise regression using the `groupfs` function in the `selectiveInference` package (Tibshirani et al., 2016), followed by conditional testing using `groupfsInf`. The resulting sequence of $p$-values is input into the `forwardStop` function, which determines a stopping point that controls the pathwise conditional false discovery rate at 10%. All groups selected prior to this stopping point are declared significant.

3. **data-splitting**: The data is randomly split into two sets, each containing $n/2$ observations. The first set is used to select groups using a group lasso model with cross validation being used to $\lambda$. To conduct inference on these selections, an ordinary least squares regression model using only the selected groups from the prior step is fit using the second half of the data, and traditional significance tests and local false discovery rate methods are applied to determine the final set of selected groups.

4. **mFdr**: A group lasso model is fit using the `grpreg` package (Breheny and
Huang, 2015). At each step of the $\lambda$ sequence we calculate the mFdr upper bound, choosing the smallest value of $\lambda$ where the marginal false discovery rate is below 10% and declaring all active groups in this model significant.

5. $mfdr$: A group lasso model is fit using grpreg and cross validation is used to select the optimal model characterized by $\lambda_{cv}$. Using this model we estimate $mfdr$ for each group and declare those with $\hat{mfdr} < .1$ to be significant.

![Figure 4.3: The number of true groups found, while controlling the false discovery rate below 10%, by various methods for different effect sizes. The mFdr method performs well, while the mfdr method is comparable to large-scale testing using local false discovery rates.](image)

Figure 4.3: The number of true groups found, while controlling the false discovery rate below 10%, by various methods for different effect sizes. The mFdr method performs well, while the mfdr method is comparable to large-scale testing using local false discovery rates.

Figure 4.3 compares the power of several methods to select groups with non-zero effects while controlling the false discovery rate at a threshold of 10%. The mFdr approach is able to select the most true groups for every effect size ($\beta_j = b/\sqrt{n}$). The mfdr and large-scale testing approaches, both of which are based upon local false discovery rates, behave similarly and select slightly fewer true groups than mFdr. It is worth noting that using a local threshold of 10% will conservatively control the tail-area false discovery rate below 10%. A more precise approach could search for a stopping point such that the average local false discovery rate was 10%. 
Using a simple threshold rather than searching for a stopping point makes these two approaches appear conservative relative to the tail-area based mFdr. For all effect sizes, data-splitting and selective inference select the fewest groups; however, they also rely on stricter false discovery definitions. Despite all groups being independently generated in this simulation, the lower power exhibited by data-splitting and selective inference demonstrates that there is a price to be paid in order to guarantee stronger, conditional, false discovery rate control.

4.5.4 Estimating $\sigma^2$

As alluded to in previous sections, estimating $\sigma^2$ can be difficult, particularly in settings where $p > n$ and $K_j$ is large. This difficulty has a direct impact on estimating mFdr and mfdr. Some natural approaches to estimating $\sigma^2$ are to use the squared sum of the residuals divided by their degrees of freedom, or to use cross validation. The first approach is complicated by degrees of freedom; when $p$ is large it is possible for the number of non-zero regression effects in a group lasso model to exceed $n$, so clearly the usual lasso degrees of freedom cannot apply. Cross validation avoids this dilemma, and usually not an inconvenience as it is typically used to select $\lambda$ in most applications. The previous simulations we presented have all used cross validation to estimate $\sigma^2$.

However, even when a valid estimate of $\sigma^2$ is available, for example through the use of cross validation, that estimate comes with its own uncertainty. Typically this uncertainty is accounted for by using an $F$ distribution, rather than a $\chi^2$ distribution. However this again leads to an issue regarding the correct residual degrees of freedom. We’ve found the $\chi^2$ distribution to be reasonably accurate in cases where estimating $\sigma^2$ is relatively easy, such as when $K_j$ is small or when $n > p$. However, in cases where estimating $\sigma^2$ is difficult we’ve found that using a $\chi^2$ distribution can be woefully inaccurate.
Figure 4.4: The empirical null distribution of the statistic $\frac{1}{n\sigma^2}||\hat{X}_j^T r_j||_2^2$ plotted versus its theoretical $\chi^2$ distribution for three choices of $\lambda$. $\lambda_{\text{max}}$ is the intercept only group lasso model, $\lambda_{\text{cv}}$ is the model chosen by cross validation, and $\lambda_{\text{min}}$ is the end of the $\lambda$ path - often the saturated model. In row 1, $J = 50$ and $K_j = 5$; in row 2, $J = 200$ and $K_j = 5$; in row 3, $J = 200$ and $K_j = 15$.

Figure 4.4 shows the empirical null distribution of $\frac{1}{n\sigma^2}||\hat{X}_j^T r_j||_2^2$, relative to the theoretical $\chi^2$ distribution, for various dimensions and group sizes, and choices of $\lambda$, in the external information scenario. In the first row, $n = 500$ and there are a total of $J = 50$ groups of size $K_j = 5$, with 5 (10%) of the groups having non-null effects. For each $\lambda$ choice, indicated by the columns of Figure 4.4, the theoretical $\chi^2$ distribution accurately reflects the empirically observed distribution for the null groups. In the second row, $n = 500$ and there are a total of $J = 200$ groups of size $K_j = 5$, with 20 (10%) of the groups having non-null effects. Here the theoretical $\chi^2$ distribution is accurate at $\lambda_{\text{max}}$ but becomes slightly inaccurate for models where the number of active variables is large. In the third row, $n = 500$ and there are
a total of \( J = 200 \) groups of size \( K_j = 15 \), again with 20 (10\%) of the groups having non-null effects. In this situation, where \( p > n \) and \( K_j \) is large, we see severe departure from the theoretical null distribution, indicating the substantial impact of ignoring the uncertainty when estimating \( \sigma^2 \) in these difficult situations. Improved estimation of \( \sigma^2 \) and a better understanding of the effective degrees of freedom involved in its estimation could improve our mFdr and mfdr methods; this is an area of future research.

4.6 Case Study: Gene expression in Bardet-Biedl syndrome

In this case study we analyze the data of Scheetz et al. (2006), which consists of microarray gene expression measurements from the eye tissue of 120 male rats. Our outcome variable is the expression of TRIM32, a gene which has been shown to cause Bardet-Biedl syndrome, a genetic disease impacting multiple organ systems including the retina. Of the 31,042 probesets in the dataset we restrict our analysis to the 5,000 that exhibit the largest variance in their expression (on the log scale). In order to capture potentially non-linear relationships between these 5,000 genes and TRIM32, we use a natural cubic spline basis expansion, which results in 5,000 groups of size 3.

We apply the group lasso to the expanded data, identifying significant genes in two ways: first by using mFdr to select a model with a marginal false discovery rate below 10\%, and second by using cross validation to select the model and then choosing genes with \( \hat{mfdr} \leq .1 \) to control the marginal false discovery rate below 10\%. For comparison we consider several alternative approaches: fitting an ungrouped lasso regression to the unexpanded set of 5,000 genes and using the previously described mFdr and mfdr approaches; large-scale testing on the expanded data using traditional local false discovery rates; selective inference on the expanded data using
group forward stepwise regression with stopping rules to control the conditional false discovery rate below 10%; and a repeated data splitting approach where a group lasso model first is applied to random half of the data to select groups and inference is done using an ordinary least squares model fit to the remaining half of the data, using only that groups were selected. Because each half of the data contains only 60 observations, in the first stage we select all active groups at the cross-validated choice of $\lambda$ if there are 15 or fewer groups active in this model, or we select the first 15 groups if there are more than 15 active groups in the cross-validated model. We repeat the data splitting procedure 100 times and aggregate the results using the approach of Meinshausen et al. (2009).

<table>
<thead>
<tr>
<th>Method</th>
<th>CV Error</th>
<th>Std. Error</th>
<th>Selected</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>grouped mFdr</td>
<td>0.0084</td>
<td>0.0013</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>grouped mfdr</td>
<td>0.0076</td>
<td>0.0010</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>ungrouped mFdr</td>
<td>0.0104</td>
<td>0.0035</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>ungrouped mfdr</td>
<td>0.0099</td>
<td>0.0028</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>large-scale testing (grouped)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2434</td>
</tr>
<tr>
<td>selective inference (grouped)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>repeated data splitting (grouped)</td>
<td>NA</td>
<td>NA</td>
<td>9.54</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.1: The Scheetz case study results using several different analysis approaches that each control the false discovery rate at 10%. The different false discovery definitions of these approaches leads to the drastic differences in the number of significant findings.

Table 4.1 displays the cross validation error (if applicable), the number of genes that are active in the model (if applicable), and the number of genes that can be declared significant while controlling the false discovery rate at 10%. The group lasso is able to achieve smaller cross validation errors than the ungrouped lasso, indicating
that allowing for non-linear effects yields a better fitting model. Additionally, the
group lasso approaches are able to identify at least as many significant genes as
their ungrouped counterparts.

In comparison with other methods, large-scale testing suggests an extremely
large number of significant genes, a result which can likely be attributed to the
approach providing no protection against over-fitting in the single group regression
models, as well as the abundant correlations between genes. These correlations often
lead to redundancy in the results of large-scale testing, while regression approaches
tend to instead favor a single representative from each correlated group. Selective
inference cannot select any genes while limiting the conditional false discovery rate
at 10%, relaxing the acceptable false discovery rate to 25% leads to the selection of
29 genes, surprisingly only two of these genes are also selected by the group mfdr
approach and none are selected by the ungrouped mfdr approach. Data splitting
also is unable to identify any significant groups using a 10% false discovery rate
threshold, but unlike selective inference more liberal thresholds do not yield any
significant findings. After excluding the large-scale testing results, and using the
relaxed 25% cutoff for selective inference, a total of 16 genes are selected by more
than one approach, with four genes being selected by three approaches and two
genes, NR3C2 and Thumpd3-as1, being selected by four different approaches. If
we restrict our attention to only group-based approaches, a single gene, NR3C2, is
selected by the mfdr, mFdr, and selective inference methods. NR3C2 encodes the
mineralocorticoid receptor, which mediates aldosterone actions on salt and water
balance. Although this link between NR3C2 and TRIM32 that we have identified
appears to be novel, it is biologically plausible, as NR3C2 has been previously linked
to a different tripartite motif protein, TRIM24 (Zennaro et al., 2001)
4.7 Discussion

The tail-area and local marginal false discovery rate approaches proposed in this article provide straightforward and computationally inexpensive means of assessing the group selections made by the group lasso. Compared with other regression based approaches, such as data-splitting or selective inference, mFdr uses a weaker false discovery definition that allows for increased power to detect important groups while still limiting the number of selections that are purely noise. The marginal false discovery perspective is the same definition used in one-at-a-time large-scale testing approaches, making it a reasonable choice in numerous settings. The straightforward interpretations of marginal false discovery rate methods, and their low computational cost, make inference using mFdr and mfdr a sensible next steps after fitting a group lasso model. However, as is true with any measure, it is important to be aware of limitations. The proposed mFdr and mfdr methods tend to be overly conservative in situations where $\sigma^2$ is difficult to estimate, such as when groups sizes are large in $p > n$ scenarios. Additionally, conditional false discovery rates might be the more relevant quantity to consider in certain situations depending upon the goals of the analysis.
5.1 Accomplishments

Penalized regression models have become a popular tool for modeling high-dimensional data, yet performing valid statistical inference on predictor variables using these models is a difficult task. The chapters of this dissertation contribute new false discovery rate based inferential tools that can be applied to various penalized regression models. A particular goal of this work was to develop methods for high-dimensional logistic and Cox regression models, where, in spite of their widespread use, very few inferential methods have been developed.

Chapter 2 addresses the development of an upper bound for the number and rate of marginal false discovery in any sparse penalized likelihood model. This bound is tight under feature independence, with correlations causing it to be conservative, but still useful in comparison with existing methods. The major strengths of this method are its speed and flexibility; mFdr can be calculated in just a few seconds for the entire solution path of very large penalized regression models. While we focused on the lasso throughout this dissertation, mFdr and can be applied to a wide variety of penalty functions, including the lasso, elastic net, SCAD, MCP, and MNet. The mFdr method is currently implemented in the R package ncvreg using the mfdr function, an example using an MCP penalized Cox regression model can be seen below:

```r
> library(ncvreg)
> data("Lung")
> fit <- ncvsurv(Lung$X, Lung$y)
> mfdr(fit)
```

```
EF S mFDR
```
Chapter 3 develops a local marginal false discovery rate estimate for any variable under consideration by a penalized likelihood model at a given value of $\lambda$. The method is informative at the individual feature level and can be used to make inferences on the most predictive model, a model which will almost certainly contain some noise variables. The method is related to the mFdr approach, the expected mfdr of the active variables in a model is equal to the model’s mFdr. Like mFdr, the mfdr method is fast and flexible, and can be applied to lasso, elastic net, MCP, and MNet models for a variety of likelihoods using the `summary` function in `ncvreg`.

An example is shown below:

```r
> library(ncvreg)
> data("Lung")
> fit <- ncvsurv(Lung$X, Lung$y)
> summary(fit, lambda = 0.08)

MCP-penalized Cox regression with n=137, p=9

At lambda=0.08:
```

```r
0.44806  0.000000e+00  0  0.000000e+00
0.41786  9.926598e-06  1  9.926598e-06
0.38970  3.996605e-05  1  3.996605e-05
0.36344  1.392150e-04  1  1.392150e-04
0.33894  4.272559e-04  1  4.272559e-04
0.31610  1.173937e-03  1  1.173937e-03
0.29479  2.925856e-03  1  2.925856e-03
0.27493  6.681342e-03  1  6.681342e-03
0.25640  1.406510e-02  2  7.032551e-03
0.23912  2.735765e-02  2  1.367882e-02
...```
Nonzero coefficients: 5

Expected nonzero coefficients: 2.84

mFDR: 0.568

(local) Expected nonzero coefficients: 2.24

(local) Overall mfdr (5 features) : 0.449

<table>
<thead>
<tr>
<th>Estimate</th>
<th>z</th>
<th>mFDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>karno</td>
<td>-0.03121</td>
<td>-6.212</td>
</tr>
<tr>
<td>squamous</td>
<td>-0.78609</td>
<td>-3.516</td>
</tr>
<tr>
<td>large</td>
<td>-0.37310</td>
<td>-2.062</td>
</tr>
<tr>
<td>adeno</td>
<td>0.21578</td>
<td>1.617</td>
</tr>
<tr>
<td>trt</td>
<td>0.13616</td>
<td>1.482</td>
</tr>
</tbody>
</table>

Chapter 4 extends the mFdr and mfdr methods to the grouped variable setting under the group lasso penalty. In this work inference is performed at the group level, which is particularly useful for assessing categorical variables, non-linear effects, or variables that can grouped scientifically, such as genes within regulatory pathways. These methods tend to be conservative in cases where the residual variance, $\sigma^2$, is difficult to estimate. However, they are competitive with existing approaches in simulations and appear to be useful when applied to real data.

5.2 Future Work

The methods of Chapter 4 work well in settings where $\sigma^2$ can be accurately estimated; but there many scenarios where this is not the case that can arise in high-dimensional data analysis. Future work involves addressing this problem by determining the appropriate null distribution for these settings. The work in Chapter 4 also suggests numerous extensions to more general settings. One natural next
step is to extend the mFdr and mfdr methods to general likelihood models under the group lasso penalty. This is extension is currently in progress; it relies on approximating the matrix $W$ using $cI$, where $I$ is the identity matrix and $c$ is a constant. In the case of logistic regression $W$ is already a diagonal, and in the case of Cox regression there is some justification for ignoring the off diagonal elements of $W$ (Simon, 2011). However, it is unclear the best way of determining $c$, and while some early results look promising, more work still needs to be done. A final piece of future work is to implement these methods into the grpreg package and in doing so explore their applicability to other group penalty functions, including those which induce sparsity at both the group and individual variable levels.


