Marginal False Discovery Rates for Sparse Group Penalized Regression

Ryan Miller
Department of Mathematics
Xavier University

Patrick Breheny
Department of Biostatistics
University of Iowa

August 6, 2021

Abstract

Penalized regression methods represent a popular modeling approach due in part to their ability to simulataneously achieve variable selection and estimation. However, many of the most widely used penalization schemes, such as the lasso, can be unsatisfactory in applications involving either categorical predictors, or predictors that are likely to share a non-linear relationship with the outcome. These scenarios require single variables be expressed using multiple columns within the design matrix prior to model fitting, making grouped penalties, such as the group lasso, more useful in facilitating the indentification and interpretation of important variables. This work proposes methods of marginal false discovery rate control in the context of group penalized regression, placing a particular focus on the group lasso. We present methods for linear, logistic, and survival regression models subject to either the group lasso or group MCP penalties. In a variety of applications involving categorical predictors and non-linear relationships, we demonstrate via simulation study that our proposed methods control the false discovery rate while yielding more true positives than existing alternatives. We also illustrate the practical utility of these methods in a non-linear additive modeling application involving 22,215 genetic features obtained the bronchial epitheliums of smokers suspected of having lung cancer.

1 Introduction

Since its inception, lasso regression (Tibshirani, 1996) has quickly become a popular modeling approach due to its ability to simultaneously achieve variable selection and estimation via penalization. In the usual linear regression setting, the lasso models an n-dimensional vector of continuous outcomes, $\mathbf{y}$, as a linear combination of covariates contained in an n by p design matrix, $\mathbf{X}$, and a p-dimensional vector of regression coefficients, $\boldsymbol{\beta}$, where the size $\boldsymbol{\beta}$ is subject to an l_1 penalty. More precisely, the lasso estimator is defined as the minimizer, with respect to $\boldsymbol{\beta}$, of the objective function:

$$Q_{\text{lasso}}(\boldsymbol{\beta}) = \tfrac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_1$$

A desirable property of the lasso estimator is that some of its components are exactly zero for sufficiently large values of the penalty parameter, λ. This allows the lasso estimate to be *sparse*, with the variables corresponding to non-zero elements described as being *selected* by the lasso. Unfortunately, selections made by the lasso can be unsatisfactory in certain applications, with notable examples being nominal categorical predictors or applications where basis expansions are used in non-linear additive modeling. In the former scenario, the lasso will select individual dummy variables from the design matrix rather than the underlying categorical predictors, thereby

making its selections dependent upon the scheme used to encode the dummy variables and complicating efforts by the analyst to identify and interpret impactful categorical predictors. The later scenario faces a similar issue, as the overall selection status of an underlying predictor can be unclear if some columns of its basis expansion are estimated as exactly zero but others are not.

The group lasso (Yuan and Lin, 2006) alleviates these shortcomings by imposing an l_2 penalty on entire groups of coefficients, resulting in selections occuring at the group level. For linear regression applications, the group lasso estimator of $\boldsymbol{\beta}$ is the minimizer of:

$$Q(\boldsymbol{\beta}) = \tfrac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{J} \lambda_j ||B_j||$$

This penalization framework can be extended to generalized linear models, with logistic regression (Meier et al., 2008) serving as a prominent example, and other likelihood-based models, such as Cox regression (Wang et al., 2009). It can also be generalized to other penalty functions, such as non-convex penalties (Breheny and Huang, 2012), and penalties that are also sparse at the individual coefficient level (Simon et al., 2013). Huang et al. (2012) provides a selective review of these methods.

One potential drawback of the group lasso and related methods is the limited scope of inferential approaches that are currently available. The recently developed *selective inference* (Tibshirani et al., 2016) family of methods has been extended to group sparse settings, including the group lasso (Yang et al., 2016), but a software implementation is only available for forward stepwise selection. Similarly, the *knockoff filter* method of false discovery rate control (Barber and Candes, 2015; Candès et al., 2018) has also been extended to group sparse settings (Dai and Barber, 2016), but it too lacks an available software implementation for group lasso models. Meanwhile, computational inferential approaches with readily available software, such as the parametric bootstrap approach implemented in the `EAinference` R Package (Zhou and Min, 2017), focus on the uncertainty in the individual coefficient estimates of the group lasso, rather than the selections of entire groups, limiting their utility in regards to false discovery rate control.

The focus of this paper is on the reliability of group selections made by the group lasso and its extensions. More specifically, we propose methods for controlling the marginal false discovery rate of group selections in the context of group lasso regression. We generalize our methods to variants of the group lasso, including group penalized logistic regression, other generalized linear models, and group penalized Cox regression, as well as the group MCP penalization scheme. Our proposed methods provide accessible, computationally efficient alternatives to the limited set of existing inferential approaches currently available for the group lasso. We use several simulation studies to demonstrate the robustness of these methods across a variety of data structures, as well as their ability to achieve higher true positive rates than some existing inferential approaches.

2 Background

2.1 Group lasso regression

Consider data of the usual form, $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ records response values for $i \in \{1, \ldots, n\}$ independent observations, and $\mathbf{X}$ is an n by p design matrix of explanatory variables. We focus on situations where the columns of $\mathbf{X}$ can be naturally organized into J nonoverlapping groups, such that $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J\}$, where $\mathbf{X}_j$ denotes the n by K_j matrix containing the explanatory variables belonging to group j.

The explanatory variables in $\mathbf{X}$ can be related with $\mathbf{y}$ using a probability model involving a set of coefficients, $\boldsymbol{\beta}$. A well-known example is the linear regression model, which specifies the relationship:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{\epsilon}$ is a n-dimensional vector of errors that are independent and Normally distributed with mean 0 and variance σ^2.

Under the model described in Equation 2.1, the group lasso solution, $\hat{\boldsymbol{\beta}}$, is the minimizer, with respect to $\boldsymbol{\beta}$, of the following objective function, denoted $Q(\boldsymbol{\beta})$:

$$Q(\boldsymbol{\beta}) = \tfrac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{J} \lambda_j ||\boldsymbol{\beta}_j|| \tag{2.2}$$

where λ_j is a penalty imposed on the l_2-norm of the coefficients belonging group j.

It is possible to specify λ_j separately for each group, but the more common choice use $\lambda_j = \sqrt{K_j}\lambda$, which penalizes each group in accordance to its size and allows λ to be used universally across groups. To aid in the presentation of our methods, we will assume $\lambda_j = \sqrt{K_j}\lambda$; however, this choice is not critical in any of our derivations.

The underlying framework of the group lasso can be generalized to a wide range of loss functions and penalization schemes. These variants can be expressed using the following objective function:

$$Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) + \sum_{j=1}^{J} p_\lambda(\boldsymbol{\beta}_j) \tag{2.3}$$

where $p_\lambda(\cdot)$ is a penalty function applied to each group of coefficients, and $L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})$ is a loss function, which is based upon the log-likelihood in the case of generalized linear models, and the Cox partial likelihood in the case of Cox regression.

For sufficiently large λ_j, the elements of $\hat{\boldsymbol{\beta}}_j$, the group of estimated coefficients corresponding to the columns of $\mathbf{X}$ that belong to group j will be exactly zero. We refer to any groups whose coefficient estimates are non-zero as being "selected" by the group lasso. Additionally, the group lasso penalty shrinks coefficient estimates towards zero, which has the benefit of allowing identifiable estimation even when the dimensionality of $\mathbf{X}$ is such that $p > n$.

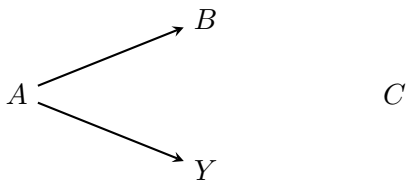
In their original proposal, Yuan and Lin (2006) assume have been orthonormalized within each group, such that $\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_j = \mathbf{I}$ for all $j \in \{1, \ldots, J\}$ with $\mathbf{I}$ representing the identity matrix, prior to model estimation. Although data are unlikely to naturally occur in this form, groups can be orthonormalized as a pre-processing step. Provided $K_j < n$ for all $j \in \{1, \ldots, J\}$, any optimization done using the orthonormalized data is equivalent to the original scale, thus group lasso solution found on the orthonormal scale can also be easily converted back to the original scale. Throughout the remainder of this paper we assume the data have been orthonormalized within in each group as a pre-processing step. See Simon and Tibshirani (2004) for further discussion of this topic.

2.2 Marginal false discovery rates

Following the seminal work of Benjamini and Hochberg (1995), false discovery rate (FDR) control has become one of the most adopted inferential paradigms in applications involving large numbers of simultaneous comparisons between variables. The literature on false discovery rate control is enormous, with many authors operating under slightly different definitions. One straightforward way to define the false discovery rate is as the expected number of false selections divided by the total number of selections, which corresponds to the fraction of significant features that are expected to be false positives. Under this definition, a procedure that controls the FDR at 10% conveys the expectation that no more than 10% of the comparisons it identifies as significant are expected to be false positives.

Much of the work done on false discovery rate control is in the context of *large-scale univariate testing*, or applications that involve aggregating the results of a large number of single variable hypothesis tests. Farcomeni (2008) and Strimmer (2008) provide a more detailed overviews of these methods. Our focus is on the regression modeling framework, where the notion of a false discovery can be complicated by the presence of relationships between predictors.

To better understand these complexities, consider the causal diagram shown below, which depicts a possible relationship between three explanatory variables, A, B, and C, and an outcome variable Y.



Variable A has a direct causal relationship with Y and should never be considered a false discovery, while variable C is independent of Y and should always be considered a false discovery. Whether or not variable B should be considered a false discovery less clear. In large scale univariate testing approaches, variable B is not be viewed as a false discovery because it is not *marginally independent* of variable Y, which is the criteria typically evaluated in a single variable hypothesis test. However, in the regression setting, the coefficient in the data generating model corresponding to variable B would be zero, suggesting variable B is a false discovery.

To clarity these distinctions, we introduce two contrasting false discovery rate perspectives: the *marginal* perspective, in which a variable, or group of variables, denoted X_j is a false discovery if it is declared significant despite being fully independent of the outcome: $\mathbf{X}_j \perp\!\!\!\perp Y$, and the *fully conditional* perspective, where X_j is a false discovery if it is deemed significant despite being independent of the outcome conditional upon all of the other variables in the data: $X_j \perp\!\!\!\perp Y | X_{k \neq j}$. Because penalized regression methods, including the group lasso, allow for only a subset of the available predictors to be active in a given model, a *pathwise conditional* perspective is also possible. This perspective focuses on the model where X_j first becomes active and conditions only on the other variables present in the model (a set we denote M_j) at that time when assessing whether or not variable j is a false discovery: $X_j \perp\!\!\!\perp Y | X_k$ for $k \in M_j$.

The methods developed in this paper use the less restrictive *marginal false discovery rate* definition. While the selection of variables like B in the causal diagram can be problematic in certain applications, it is often is impossible to untangle the true causal structure of the $A - B - Y$ relationship after the data have been collected. Additionally, it is always useful, regardless of application, to control the number of noise variables, like C, that are declared significant. For these reasons, in addition to the existing adoption of the marginal definition in the realm of large-scale univariate testing, we argue that the marginal false discovery rate is a valuable quantity to consider in applications involving the group lasso and its variants. Further, other marginal false discovery rate inferential approaches for regression models have gained traction in recent years, with Breheny (2019); Miller and Breheny (2019); Liang et al. (2021) developing methods for the ordinary lasso, penalized GLMs and survival models, and penalized transformation models respectively.

3 Marginal false discovery rates for the group lasso

3.1 Linear regression

Consider data generated by the usual linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\boldsymbol{\epsilon}$ is a n-dimensional vector of errors that are independent and Normally distributed with mean 0 and variance σ^2, and $\mathbf{X}$ is presumed to have a known grouping structure such that $\boldsymbol{\beta}_j$ is a vector of length K_j representing the regression coefficients associated with group $j \in \{1, \ldots, J\}$.

Our goal is to characterize the expected number and rate falsely selected groups for a given group lasso model. We begin with the group lasso solution, $\hat{\boldsymbol{\beta}}$, defined as the minimizer of the group lasso objective function, $Q(\boldsymbol{\beta})$:

$$Q(\boldsymbol{\beta}) = \tfrac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{J} \lambda_j ||B_j|| \tag{3.2}$$

Solving for $\hat{\boldsymbol{\beta}}$ involves the subdifferential of Q with respect to $\boldsymbol{\beta}_j$:

$$\begin{aligned} -\tfrac{1}{n}\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\boldsymbol{\beta}_{-j})+\boldsymbol{\beta}_j+\lambda_j\tfrac{\boldsymbol{\beta}_j}{||\boldsymbol{\beta}_j||} \quad &\text{if } \boldsymbol{\beta}_j \neq 0 \\ -\tfrac{1}{n}\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\boldsymbol{\beta}_{-j})+\lambda_j\mathbf{v} \quad &\text{if } \boldsymbol{\beta}_j = 0 \end{aligned} \tag{3.3}$$

Here, $\mathbf{v}$ is any vector satisfying $||\mathbf{v}|| < 1$, and the notation $\mathbf{X}_{-j}$ is used to denote the portion of $\mathbf{X}$ that remains after removal of the covariate group contained in $\mathbf{X}_j$, with $\boldsymbol{\beta}_{-j}$ respectively describing the associated model coefficients. In the convex optimization literature, the criteria described in Expression 3.3 are known as the KKT conditions.

From these conditions, it follows that if the j^{th} group is to be selected into the model with non-zero coefficients, it must be that the case that:

$$\tfrac{1}{n}\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})-\hat{\boldsymbol{\beta}}_j=\lambda_j\tfrac{\hat{\boldsymbol{\beta}}_j}{||\hat{\boldsymbol{\beta}}_j||} \tag{3.4}$$

Theorem 1. *For the group lasso solution $\hat{\boldsymbol{\beta}}$, the component $\hat{\boldsymbol{\beta}}_j \neq \mathbf{0}$ if and only if*

$$\tfrac{1}{n}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 > \lambda_j^2 \tag{3.5}$$

Further, given the group $\mathbf{X}_j$ is marginally independent of $\mathbf{y}$, if $\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_{-j}^T \xrightarrow{p} \mathbf{0}$ and λ is chosen such that $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is bounded in probability, then

$$\tfrac{1}{n\sigma^2}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 \xrightarrow{d} \chi^2_{K_j} \tag{3.6}$$

Proof. The first statement is a straightforward algebraic manipulation of Equation 3.4, and is a direct consequence of the KKT conditions.

Then, expanding the left side of Expression 3.5 yields:

$$\begin{aligned} \tfrac{1}{n}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 &= \tfrac{1}{n}||\mathbf{X}_j^T(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}-\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 \\ &= \tfrac{1}{n}||\mathbf{X}_j^T\boldsymbol{\epsilon}-\mathbf{X}_j^T\mathbf{X}_{-j}(\boldsymbol{\beta}_{-j}-\hat{\boldsymbol{\beta}}_{-j})||^2 \end{aligned} \tag{3.7}$$

Noting $\frac{1}{\sqrt{n}}\mathbf{X}_j^T\mathbf{X}_{-j}(\boldsymbol{\beta}_{-j}-\hat{\boldsymbol{\beta}}_{-j}) \xrightarrow{p} \mathbf{0}$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0},\sigma^2\mathbf{I})$, we have $\frac{1}{n\sigma^2}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 \xrightarrow{d} \chi^2_{K_j}$. □

The first condition of Theorem 1, $\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_{-j}^T \xrightarrow{p} \mathbf{0}$, is satisfied when pairwise correlations between the columns of $\mathbf{X}_j$ and the other columns of $\mathbf{X}$ become negligible as n increases. While the second condition, that $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is bounded in probability, is met for suitable choices of λ (Liu and Zhang, 2009).

The first condition is unlikely to be met for many real data applications, instead it represents a valuable worst-case scenario regarding false discoveries. The heuristic explanation is that in penalized regression applications where two groups are related to each other, the selection of one group reduces the chances that the other is also selected. Thus, treating groups as independent will yield a conservative estimate of the marginal false discovery rate. Quantifying the exact degree of conservatism introduced by these relationships is significantly less mathematically tractable, but we extensively explore the issue via simulation study in Section 4. Breheny (2019) presents a more detailed theoretical discussion of this topic in the context of the ordinary lasso.

Corollary 1. *Under the conditions outlined in Theorem 1, for the group lasso model characterized by λ, the expected number of false discoveries and the expected rate of marginal false discoveries are respectively bounded by:*

$$\begin{aligned} FD &= \sum_{j=1}^{J} \Pr\left(\chi^2_{K_j} > \frac{n\lambda_j^2}{\sigma^2}\right) \\ mFDR &= \frac{FD}{S} \end{aligned} \tag{3.8}$$

where S denotes the total number of selected groups in the model characterized by λ.

Theorem 1 implies the probability that the j^{th} group is selected, given the group is marginally independent of the outcome, corresponds to the probability that a $\chi^2_{K_j}$ random variable is larger than $\frac{n\lambda_j^2}{\sigma^2}$. In principle, we'd then sum over all groups that are marginally independent of the outcome to determine the expected number of falsely selected groups; however, the identity of such groups is unknown in practice, so summing overall J groups provides a conservative alternative. In many applications of the group lasso, the number of groups that are truly related to the outcome is small relative to the total number of groups, making this effect relatively small.

For a given value of λ, the process of calculating mFDR is summarized in Algorithm 1.

Algorithm 1 Calculating the mFDR upper bound

procedure

Estimate σ^2 as $\hat{\sigma}^2$

for $j \in \{1, \ldots, J\}$ **do**

$\widehat{\text{FD}}_{j,\lambda} = \Pr\left(\chi^2_{K_j} > \frac{n\lambda_j^2}{\hat{\sigma}^2}\right)$ by the result of Theorem 1

$\widehat{\text{FD}}_\lambda = \sum_{j=1}^{J} \widehat{\text{FD}}_{j,\lambda}$

$\widehat{\text{mFDR}}_\lambda = \min\left(\frac{\widehat{\text{FD}}_\lambda}{S_\lambda}, 1\right)$

return $\widehat{\text{mFDR}}_\lambda$

We point out that the initial step of Algorithm 1 requires estimating σ^2, either by dividing the residual sum of squares by its degrees of freedom or via the cross-validated error of the model under consideration.

3.2 Generalized linear models

The group lasso penalty can be extended to other likelihood-based models, such as GLMs and Cox regression. Solving for $\hat{\boldsymbol{\beta}}$ in these applications typically relies upon a quadratic approximation of the log-likelihood function, L:

$$L(\boldsymbol{\eta}) = L(\tilde{\boldsymbol{\eta}}) + (\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}})^T \mathbf{v} + \tfrac{1}{2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}})^T \mathbf{W} (\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) \tag{3.9}$$

where $\eta = \mathbf{X}\boldsymbol{\beta}$, with $\tilde{\boldsymbol{\eta}}$ denoting the current estimate, and $\mathbf{v}$ and $\mathbf{W}$ respectively denote the first and second derivatives of $L(\boldsymbol{\eta})$ evaluated at $\tilde{\boldsymbol{\eta}}$. It is worthwhile noting that $\mathbf{W}$ is a diagonal matrix in popular GLMs such as logistic regression.

Letting $\mathbf{z} = \tilde{\boldsymbol{\eta}} - \mathbf{W}^{-1}\mathbf{v}$, and dropping terms that are constant with respect to $\boldsymbol{\beta}$, the approximation in Equation 3.9 results in a loss function that is equivalent to weighted squared error loss:

$$L(\boldsymbol{\beta}) \approx \tfrac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) \tag{3.10}$$

Thus, the optimization algorithms that solve for the group lasso solution, as well as the corresponding KKT conditions, can be adapted to likelihood-based models by the minor addition of a weight matrix, $\mathbf{W}$.

Consequently, groups selections made by these models are characterized by the following condition:

$$\tfrac{1}{n}||\mathbf{X}_j^T \mathbf{W}(\mathbf{z} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 > \lambda_j^2 \tag{3.11}$$

Applying the same general steps previously described in the linear regression setting, we can work towards characterizing the marginal false discovery rate of these models using the left-hand side of Expression **??**:

$$\tfrac{1}{n}||\mathbf{X}_j^T \mathbf{W}(\mathbf{z} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 = \tfrac{1}{n}||\mathbf{X}_j^T \mathbf{v} - \mathbf{X}_j^T \mathbf{W} \mathbf{X}_{-j}(\boldsymbol{\beta}_{-j} - \hat{\boldsymbol{\beta}}_{-j})||^2 \tag{3.12}$$

Proposition 1. *Provided* $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1/2}\mathbf{X}^T\mathbf{v} \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$*, where* $\mathbf{I}$ *is the p x p identity matrix, if* $\frac{1}{n}\mathbf{X}_j^T\mathbf{W}\mathbf{X}_{-j}^T \xrightarrow{p} \mathbf{0}$*, and* $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *is bounded in probability, then the number and rate of marginal false discoveries of a likelihood-*

based model subjective to the group lasso penalty characterized by λ can be estimated by:

$$\begin{aligned} FD &= \sum_{j=1}^{J} \Pr\left(\chi^2_{K_j} > \frac{n^2\lambda_j^2}{\mathrm{Tr}(\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j)}\right) \\ mFDR &= \frac{FD}{S} \end{aligned} \tag{3.13}$$

The first condition involved in Proposition 1 is a standard result of classical likelihood theory, which can be shown for many types of generalized linear models and Cox regression. The remaining conditions are direct analogs of those described in Theorem 1 for the linear regression setting. As a result, the derivations of the estimates in Equations 3.13 are analogous to those of Theorem 1 and Corollary 1, with the primary distinction being the variance used in normalization.

Like the linear regression setting, we can summarize the calculation of mFDR for given value of λ in a group lasso penalized likelihood-based model using the following algorithm:

Algorithm 2 Calculating the mFDR upper bound (GLMs

procedure

Estimate $\mathbf{W} \leftarrow \nabla^2 f(\hat{\boldsymbol{\eta}})$

for $j \in \{1, \ldots, J\}$ **do**

$\widehat{\mathrm{FD}}_{j,\lambda} = \Pr\left(\chi^2_{K_j} > \frac{n^2\lambda_j^2}{\mathrm{Tr}(\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j)}\right)$ by Proposition 1

$\widehat{\mathrm{FD}}_{\lambda} = \sum_{j=1}^{J} \widehat{\mathrm{FD}}_{j,\lambda}$

$\widehat{\mathrm{mFDR}}_{\lambda} = \min\left(\frac{\widehat{\mathrm{FD}}_{\lambda}}{S_\lambda}, 1\right)$

return $\widehat{\mathrm{mFDR}}_{\lambda}$

As was the case in the linear regression setting, these estimates will be conservative in the presence of pairwise correlations across columns of $\mathbf{X}$ corresponding to different groups, and in applications were J is substantially exceeds the number of groups that are marginally independent of the outcome.

Additionally, the estimates produced by Algorithm 2 are subject to further uncertainty arising from the use of the average diagonal element of $\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j$ in normalization. Section 4 explores the reliability of these estimates via simulation study for group penalized logistic and Cox regression models under several different design matrices and correlation structures.

3.3 Other penalty functions

The general form of the mFDR estimators described in prior sections are directly applicable to several other penalization schemes related to the group lasso. One example is the group minimax concave penalty, or group MCP. In non-grouped applications, the MCP penalty function, $f_{\lambda,a}(\theta)$, is defined as:

$$\begin{aligned} \lambda\theta - \tfrac{\theta^2}{2a} \quad &\text{if } \theta \leq a\lambda \\ \tfrac{1}{2}a\lambda^2 \quad &\text{if } \theta > a\lambda \end{aligned} \tag{3.14}$$

for values of $\lambda \geq 0$. Here, a is a tuning parameter, with $a = 3$ being the typical value used in most software implementations. MCP is a concave penalty where the degree of penalization is diminished, eventually becoming zero, for large coefficients.

In grouped applications, the composite group MCP estimate (Breheny and Huang, 2012; Huang et al., 2012) is defined as the minimizer of:

$$Q(\boldsymbol{\beta}) = \tfrac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{J} f_{\lambda,b}\Big(\sum_{k=1}^{K_j} f_{\lambda,a}(|\beta_{jk}|)\Big) \tag{3.15}$$

where the tuning parameter b is typically chosen to be $K_j a\lambda/2$ to ensure that the group level penalty attains its maximum if and only if each of its components are at their maximum.

Although the coefficient estimates arising from group MCP regression tend to differ from those of the group lasso, the optimization conditions that define the entry criteria for a group becoming active in the model are identical. Consequently, the same estimators proposed in **??** and 3.8 can also be used to control the marginal false discovery rate in group MCP regression. In Section 4 we present results for both the group lasso and group MCP penalties in several of the simulation experiments used to explore various aspects our proposed mFDR estimators.

4 Simulation experiments

4.1 Data generation

In the following sections we present results from a variety of simulation experiments exploring the validity and efficacy of our proposed mFDR methods. In all of these experiments, the covariates recorded in $\mathbf{X}$, are derived from numeric values randomly generated from a multivariate normal distribution with a mean of $\mathbf{0}$ and a covariance matrix, Σ_X. The off-diagonal entries of Σ_X are used to invoke specific correlation structures, we focus on two particular cases: Independence - $cor(\mathbf{x}_a\mathbf{x}_b) = 0$, and Autoregressive - $cor(\mathbf{x}_a, \mathbf{x}_b) = \rho^{|a-b|}$, for all $a, b \in \{i, \ldots, J\}$. After these numeric values are generated, they are then used to construct design matrices for the two primary applications we are focused on, models containing nominal categorical predictors and non-parametric additive models.

In applications focused on categorical predictors, the final design matrix is constructed by binning the underlying numeric values into k equally sized groups, which are then expressed in the design matrix using k dummy variables. Outcomes are then generated from the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where ϵ_i is independently drawn from a $N(0, 1)$ distribution, in linear regression applications, from a Bernoulli distribution with $\Pr(y_i = 1) = 1/(1 + exp(-\mathbf{x}_i^T\boldsymbol{\beta}))$ in logistic regression applications, or from an exponential distribution with a rate parameter equal to $exp(-\mathbf{x}_i^T\boldsymbol{\beta})$ in Cox regression applications.

In applications focused on non-parametric additive models, outcomes are generated directly from the underlying numeric values. For linear regression scenarios, this is done such that $\mathbf{y} = \sum_{j=1}^{t} f(\mathbf{x}_j) + \boldsymbol{\epsilon}$, where the function, f, defines a non-linear relationship between active features and the outcome, t is a pre-specified number of non-noise features, and ϵ_i is independently drawn from a $N(0, 1)$ distribution. For logistic regression scenarios, the outcome is drawn from a Bernoulli distribution with $\Pr(y_i = 1) = 1/(1 + exp(-\sum_{j=1}^{t} f(\mathbf{x}_{ij})))$, and for Cox regression scenarios the outcome is drawn from an exponential distribution with a rate parameter equal to $exp(\sum_{j=1}^{t} f(\mathbf{x}_{ij}))$. We consider two different forms of f: Quadratic - $f(X) = \beta X^2$, Piecewise Linear - $f(X) = \beta X$ if $X > 0$ and $f(X) = 0$ otherwise. Prior to model fitting, each feature undergoes a basis expansion with K_j degrees of freedom, thereby yielding a design matrix containing J groups of size K_j.

4.2 False discovery rate control

Standard software packages will fit a series group lasso models along a decreasing sequence of λ values. The estimators described in Equations 3.8 and 3.13 can be used to estimate the number of false discoveries that are expected in each model of this sequence. The curves displayed in Figure 1 display the mean estimated number of marginal false group selections and the mean empirical number of false group selections averaged over 200 simulation repetitions along a fixed, decreasing sequence of λ values in the piecewise linear data generation scenario for group lasso penalized linear, logistic, and Cox regression with $n = 1000$, $J = 100$, and $t = 10$ under the independence correlation structure.

Figure 1 demonstrates that the expected number of false selections as estimated via Equations 3.8 and 3.13 very closely resembles the empirical number of false selections along the entire sequence of λ values, with only a very slight degree of over-estimation induced by incidental correlations between the columns of $\mathbf{X}$. In the linear regression scenario, the estimated and empirical curves are nearly indistinguishable. In the logistic regression and Cox regression scenarios, the additional sources of variability discussed in Section 3.2 can be observed, though the

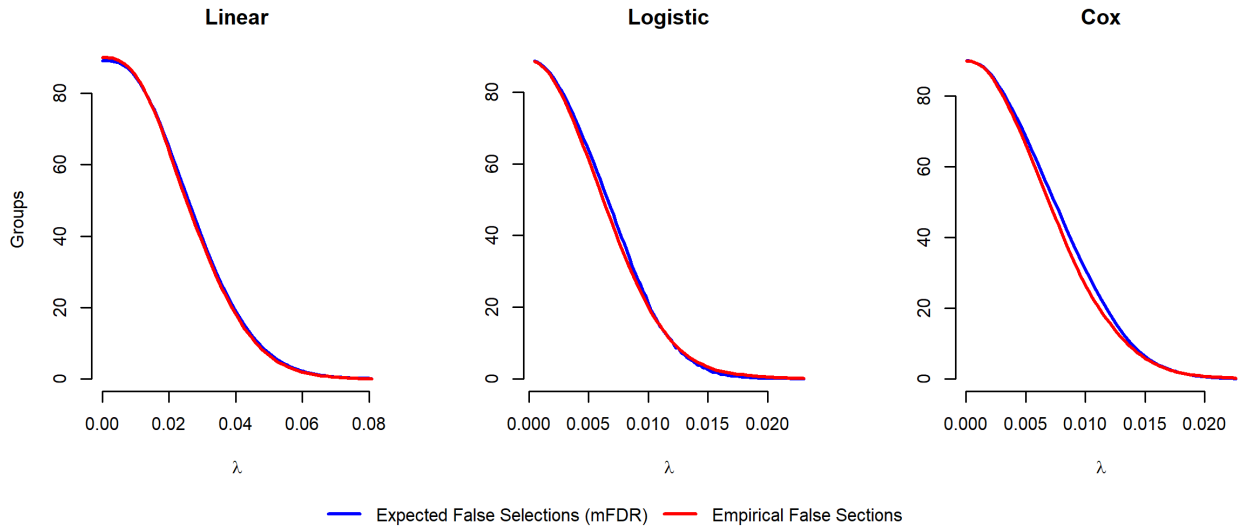


Figure 1: Comparison of the average expected and empirical number of marginal false discoveries along a sequence of λ values for various likelihood-based models when covariates are generated independently.

practical implications appear minimal.

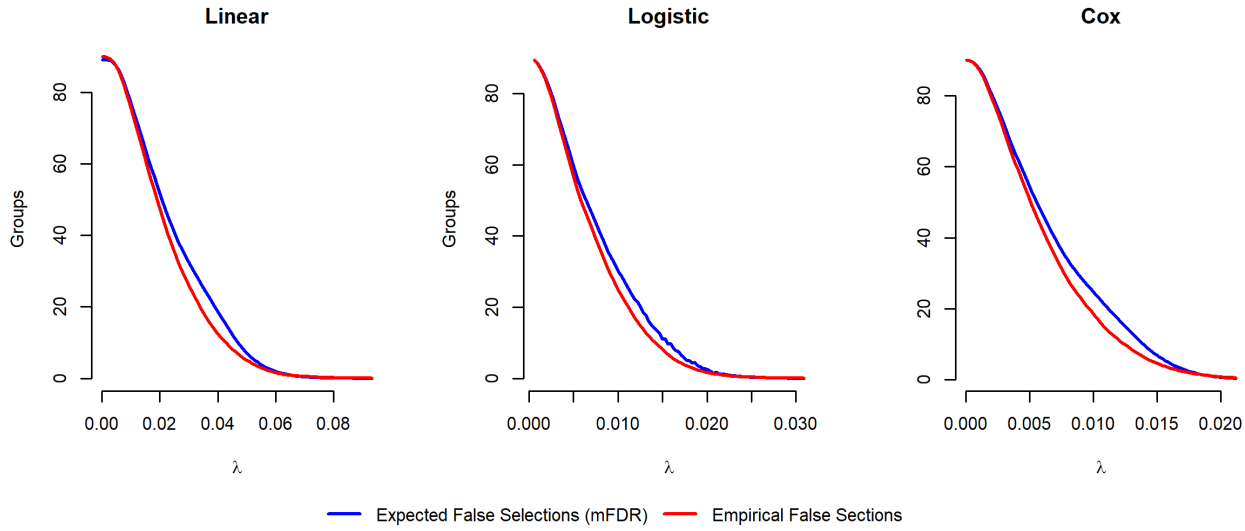


Figure 2: Comparison of the average expected and empirical number of marginal false discoveries along a sequence of λ values for various likelihood-based models when covariates are generated using an autoregressive structure with $\rho = 0.9$.

Figure refFig:lseq2 displays the same set of curves when noise features are generated under the autoregressive correlation structure with $\rho = 0.9$. For the reasons discussed in Section 3, the expected number of false selections estimated via Equations 3.8 and 3.13 tends to be larger than the empirical average at most values of λ, illustrating the conservative nature of these estimators in the presence of correlations between the columns of $\mathbf{X}$ belonging to different groups. However, in the portions of the λ sequence that are likely of greatest interest, those where the estimated mFDR is between 10% and 20%, the discrepancy is relatively small in absolute terms.

As an example, at $\lambda = 0.064$ in the linear regression setting, the average estimated marginal false discovery rate is 11.2%, while the average empirical false discovery rate is only 8.1%. Furthermore, even at the values of λ

that exhibit the greatest discrepancies, the estimated and empirical marginal false discovery rate tend to be close enough to lead to similar conclusions. An example of this is at $\lambda = 0.039$ in the linear regression setting, where the average estimated number of false discoveries is 19.27, which suggests the analyst should conclude that many false discoveries are to be expected among the groups selected by this model. The observed empirical average number of false discoveries at this value of λ is 12.94, which supports the conclusion suggested by the mFDR approach.

Focusing further on the linear regression scenario under an independence correlation structure, we conduct a second set of simulation experiments that vary $n \in \{150, 300, 500, 750, 1000, 1300, 1600\}$ to evaluate the role of sample size in the accuracy of our proposed mFDR estimates. We record the mean difference between the estimated and empirical marginal false discovery rate, averaged across 500 simulation repetitions, when selecting the smallest value of λ with $\widehat{\text{Mfdr}} \leq 0.15$ along a fixed sequence of decreasing λ values. Aside from the sample size, all other aspects of this simulation are the same as those previously described in this section, with the exception that we also include results for nominal categorical predictors with groups of size $K_j = 3$ and $K_j = 6$, as well as for the group MCP penalty.

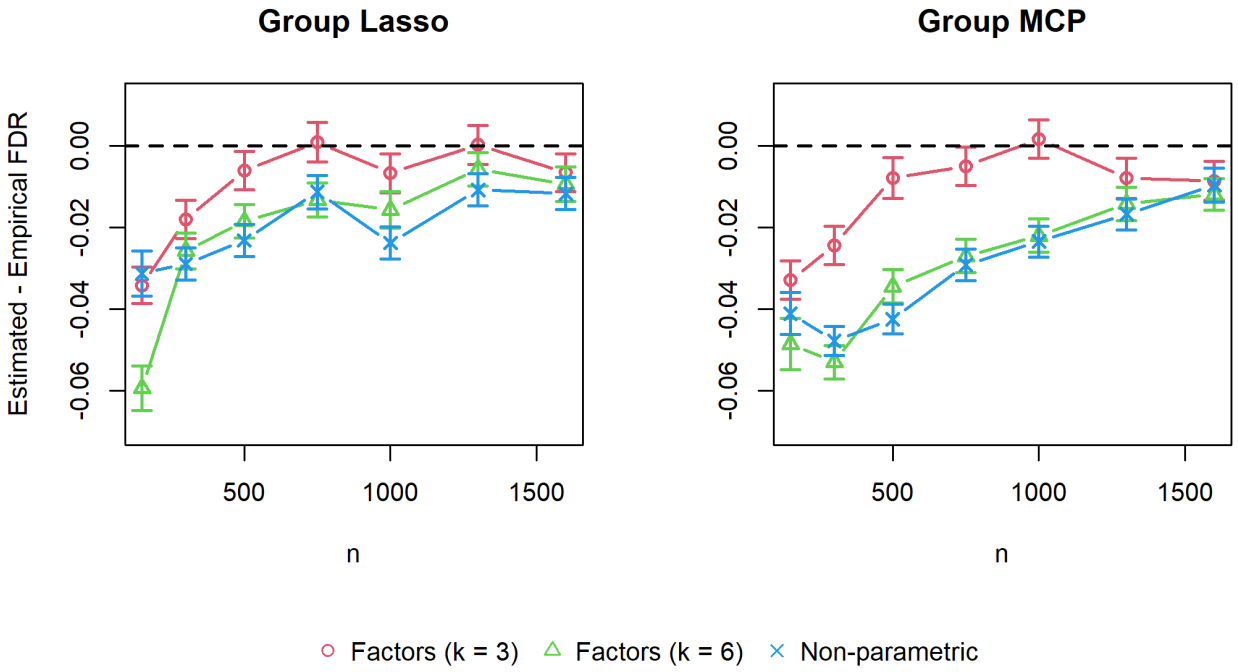


Figure 3: Tightness of the mFDR upper bound as n increases.

Figure 3 shows that as the sample size increases, the tightness of the mFDR bound improves. The mFDR bound tends to be less conservative for smaller sized groups, and less conservative in applications involving categorical predictors than in non-linear additive modeling applications. In absolute terms, the degree of conservatism is small for most sample sizes, with the estimates exceeding the truth by an average of less than 0.02 in many instances.

Additionally, Figure 3 demonstrates the applicability of our proposed mFDR method with the group MCP penalty. While the group selections and coefficient estimate under this penalization scheme tend to differ from those of the group lasso, the accuracy of our mFDR estimates tend to be comparable.

4.3 Comparison with other regression-based methods

Our third simulation experiment explores the ability our proposed mFDR methods to correctly identify true positives in comparison to other existing methods capable of controlling the false discover rate. We generate data using $n = 200$, consisting of $J = 100$ groups with only $t = 10$ of those groups sharing a real relationship with the outcome. We focus on the linear regression setting under the independence correlation structure, and we present

the average number of true positive selections across 50 simulation repetitions for range of different signal strengths (as determined by β) while controlling the false discovery rate at 10%. In our comparisons, consider the following methods:

- *Forward selection with selective inference* where the selective inference approaches implemented in the `groupfsInf` function contained in the `selectiveInference` R package (Tibshirani et al., 2016) are used to obtain p-values for sequentially added groups. The forward stopping rules developed by G'Sell et al. (2016) are then applied to these results in order to control the false discovery rate.

- *Large-scale testing* where single group hypothesis tests are performed separately, and the results are aggregated to control the false discovery rate using the methods of Benjamini and Hochberg (1995) as implemented in the `p.adjust` function contained in base R.

- *Data splitting* where half of the available data is used to select groups via the group lasso with λ chosen to minimize the 5-fold cross-validation error, and the other half is used to fit a classical least squares regression model to the selected groups and perform analysis of variance tests. The results of these tests are aggregated using the methods of Benjamini and Hochberg (1995).

- *Selective inference via the ordinary lasso* where the `selectiveInference` function from the`selectiveInference` R package is applied to an ordinary lasso regression model fit to the original data (before undergoing basis expansion) and the false discovery rate is controlled using forward stopping rules. This approach is only considered in the non-parametric additive modeling scenarios (those previously referred to as "piecewise linear" and "quadratic").

- *Knockoff filter via the ordinary lasso* where the `knockoff.filter` function contained in the `knockoff` R package (Barber and Candes, 2015) is applied to an ordinary lasso regression model fit to the original data (before undergoing basis expansion). This approach is also only considered in the scenarios previously referred to as "piecewise linear" and "quadratic".

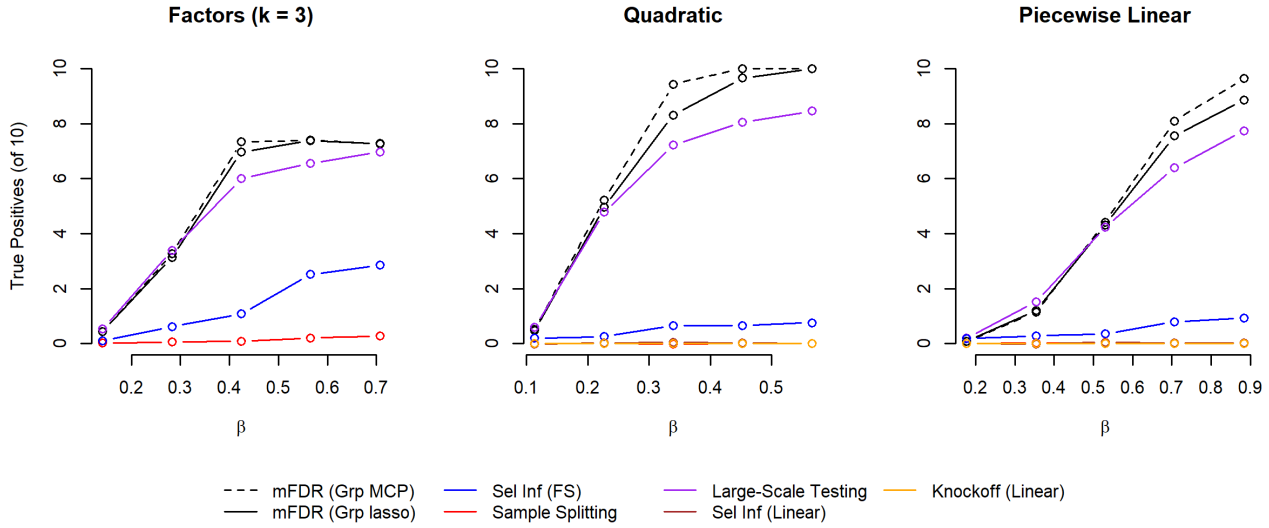


Figure 4: The average number of true positives in response to β for various methods of false discovery rate control applied to different linear regression applications involving $n = 200$, $J = 100$, $t = 10$ and the independence correlation structure.

Figure 4 demonstrates the ability of our proposed mFDR methods to on average identify more true positives than all of the aforementioned methods. We also find that when the method is slighter more powerful when

applied to models using the group MCP penalty, despite previously discussed simulations showing it to offer more conservative false discovery rate control. Large-scale testing, which also adopts a marginal perspective on false discoveries, exhibits comparable performance when the signal size is relatively small, but falls short of mFDR for larger values of β.

Forward selection with selective inference, which adopts a stronger pathwise perspective on false discoveries, is the only approach that on average selects at least one true positive for the values of β we consider. The other methods struggle to identify any of the variables that share a relationship with the outcome. These results highlight the advantages of the mFDR method and its marginal perspective on false discoveries when it comes to limiting the number of noise selections without being overly restrictive.

4.4 Comparison with cross-validation

Cross-validation is a resampling procedure that is commonly used to determine the optimal values of the penalty parameter(s) in penalized regression applications. Most standard penalized regression software will perform k-fold cross-validation, which splits the data into k subsets such that model evaluation is done on data that are different from those used in model fitting. Despite its widespread use and its ability to provide an estimate of out-of-sample model performance, cross-validation will not necessarily limit the number or rate of false discoveries.

In our final simulation experiment, we compare the number of true positives and the false discovery rate of various group lasso and group MCP linear regression models for the factor ($K_j = 3$), piecewise linear, and quadratic scenarios with coefficient vectors of varying signal strength. In each scenario, we consider three different approaches for choosing λ. The first chooses the smallest value which controls the marginal false discovery rate at 10%, the second chooses the value that minimizes the model's 5-fold cross-validation error, and the third chooses the largest value of λ where the cross-validation error is within 1 standard error of the minimum (referred to as the "1se" method). All other simulation parameters are held at the same values used in Section 4.3.

Table 1: The average number of true positive selections and the average empirical false discovery rate (shown in parentheses) for various methods of choosing λ. mFDR = the smallest λ where the estimated marginal false discovery rate is less than 10%, CV = the value of λ that minimizes the 5-fold cross-validation error, 1se = the largest value of λ where the cross-validation error is within 1 standard error of the minimum.

Penalty	Scenario	β	mFDR	1se	CV
lasso	Factors ($K_j = 3$)	0.2	1.3 (5.8%)	1.4 (7.4%)	5.2 (57.4%)
		0.3	4.2 (7.5%)	5.5 (23.3%)	7.5 (65.7%)
		0.4	6.3 (6.7%)	7.4 (30.8%)	8.0 (68.2%)
	Piecewise Linear	0.25	0.2 (4.4%)	0.3 (2.2%)	3.2 (45.9%)
		0.5	3.7 (9.0%)	6.8 (27.9%)	9.2 (63.0%)
		0.75	7.7 (4.8%)	9.8 (34.2%)	10.0 (67.2%)
	Quadratic	0.15	1.7 (4.3%)	2.2 (8.7%)	6.7 (57.3%)
		0.2	3.9 (7.3%)	6.3 (24.8%)	9.2 (63.6%)
		0.25	6.0 (6.7%)	9.1 (33.1%)	9.9 (66.9%)
MCP	Factors ($K_j = 3$)	0.2	1.3 (6.3%)	0.8 (4.3%)	3.9 (41.3%)
		0.3	4.3 (7.2%)	4.8 (12.5%)	6.7 (44.4%)
		0.4	6.8 (7.7%)	7.3 (14.5%)	7.7 (42.9%)
	Piecewise Linear	0.25	0.2 (4.5%)	0.2 (1.4%)	2.0 (28.6%)
		0.5	3.8 (8.9%)	5.2 (12.3%)	8.0 (35.8%)
		0.75	8.4 (4.2%)	9.6 (12.9%)	9.9 (35.8%)
	Quadratic	0.15	1.6 (3.8%)	1.2 (2.5%)	4.6 (34.0%)
		0.2	4.1 (6.4%)	5.1 (12.5%)	8.2 (39.4%)
		0.25	6.4 (6.0%)	8.0 (14.6%)	9.5 (36.9%)

The results displayed in Table **??** demonstrate the trade-off that typically occurs when choosing between a

highly predictive model and a model where the analyst can be confident in importance of the selected groups. In many scenarios, cross-validation leads to the selection of models where somewhere between 50%-70% of the selected groups are false discoveries. For smaller values of β, cross-validation leads to a substantially higher number of true positives than the mFDR approach, but this diminishes as β increases. Additionally, the "1se method appears to offer a middle ground; however, it does not provide any quantification of what the false discovery rate might be, and we observe real false discovery rates as high as 33% in some scenarios when using this approach.

5 Real data case study

5.1 Data

Lung cancer is among the leading causes of death in the United States and the world, with a high mortality rate that is in part attributable to a lack of effective diagnostic tools while the disease is still in its early stages. Spira et al. (2007) studied the histologically normal bronchial epitheliums of smokers, collecting RNA expression data for $p = 22,215$ genetic features using Affymetrix HG-U133A microarrays. Among the $n = 192$ participants, 102 were cases who had already developed lung cancer and 90 were controls who had not developed lung cancer. The goal of the study was to determine whether gene expression data obtained at bronchoscopy from smokers with suspicion of lung cancer could be used as a lung cancer biomarker.

5.2 Methods

In our analysis we consider several different approaches using the high-dimensional genetic data to predict the binary outcome of case-control status. Our primary approach applies a basis expansion with 4 degrees of freedom to each genetic feature, creating a new design matrix, $\tilde{\mathbf{X}}$, that contains 88,860 columns, corresponding to 22,215 groups of size $K_j = 4$. We then use the mFDR estimator proposed in 3.13 to select a group lasso model that controls the marginal false discovery rate at 10%.

For comparison, we conduct a large-scale univariate testing approach where separate logistic regression models are fit corresponding to each genetic feature and these models are then summarized by a likelihood-ratio test comparing their fit to that of an intercept-only model. The Benjamini-Hochberg procedure is applied to the resulting set of p-values to control the false discovery rate at 10%. This approach is done separately for both the original design matrix, $\mathbf{X}$, and the expanded design matrix, $\tilde{\mathbf{X}}$. We also model the data without considering non-linear associations by fitting a lasso regression model using the binomial likelihood and applying the methods of Miller and Breheny (2019) to control the marginal false discovery rate. Finally, we also include results for each of the penalized regressions models favored by 5-fold cross-validation.

5.3 Results

Table 2: A summary of various analysis approaches applied to the Spira data. S = number of selections. mFDR = estimated marginal false discovery rate (%). MCE = cross-validated misclassification error (%).

Method	Design Matrix	λ	S	mFDR	MCE
lasso	$\mathbf{X}$	CV	55	100%	24.5%
lasso	$\mathbf{X}$	mFDR	10	7.8%	31.8%
group lasso	$\tilde{\mathbf{X}}$	CV	45	53.4%	25.5%
group lasso	$\tilde{\mathbf{X}}$	mFDR	21	5.6%	27.1%
large-scale testing	$\tilde{\mathbf{X}}$	-	12,902	10.0%	-
large-scale testing	$\mathbf{X}$	-	2,426	10.0%	-

Table 2 summarizes the results of each analysis of the Spira dataset. Among penalized regression approaches, the lowest cross-validated misclassification error is observed when the ordinary lasso used on the unexpanded

design matrix; however, the estimated mFDR of this model is 100%. Since this mFDR estimate is inherently conservative, a value of 100% doesn't necessarily indicate that all these selections are noise; however, it does suggest that we cannot be confident in the reliability of these selections. For the ordinary lasso, controlling the marginal false discovery rate at 10% limits the number of genetic features selected to only 10, while at the same time substantially increasing the cross-validated misclassification error. In contrast, when the group lasso is applied to the expanded design matrix, 21 genetic features can be selected while controlling the marginal false discovery rate at 10% and achieving a misclassification error that is much closer to the minimum achieved by the ordinary lasso.

The large-scale testing results displayed in Table 2 highlight an important practical advantage of the mFDR approach. Although these methods are all based the same marginal perspective on false discoveries, the regression-based methods tend to naturally limit the number of highly correlated features that are deemed significant. That is, in situations where many features are strongly related with both each other and the outcome, penalized regression tends to select only a single representative from the group, while large-scale testing will select all of them. In applications like this one, large-scale testing can result in an overwhelming amount of "leads" that researchers must then filter, group, or assess manually. In contrast, regression-based methods like mFDR tend to yield a more manageable set of features that likely contain less redundancy.

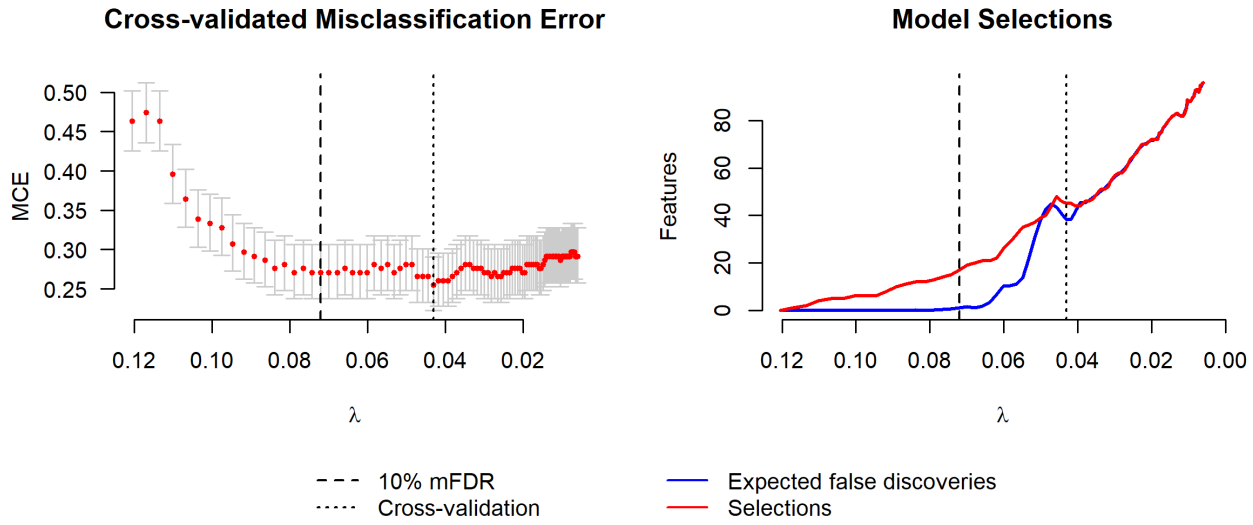


Figure 5: Group lasso modeling results for the Spira case study. The left panel displays the cross-validated misclassification error for the models corresponding to various values of a decreasing sequence of λ, while the right panel shows the total number of selections alongside the expected number of false discoveries for these models. The model favored by cross-validation marked by a dotted vertical line, while the most inclusive model with an estimated marginal false discovery rate less than 10% marked by a dashed vertical line.

Figure 5 provides a more detailed look at the results of the group lasso modeling approach. The left panel displays the cross-validated misclassification error (± 1 standard error) for models along decreasing sequence of λ values, demonstrating a wide range of models that achieve statistically similar levels of accuracy. The right panel shows the expected number of false discoveries, calculated using the estimators in Equation 3.13, and the total number of selections for each of these models. Together these plots can help the analyst determine a model with a suitable tradeoff between model accuracy and false discovery rate control. In this application, the improvements in misclassification error when the penalty parameter is decreased below 0.07 are relatively small and come at the cost of a substantial increase in the expected number of false discoveries present in the corresponding models.

6 Discussion

The ability to estimate the marginal false discovery rate of a group lasso model, or a related variant, is a useful tool in assessing the reliability of selected groups. This is particularly true when considering that many popular methods for deciding upon a group lasso model, such as cross-validation or other goodness of fit criteria, may result in undesirable numbers of false discoveries. While the mFDR approach is intended to control the marginal false discovery rate, which uses a more relaxed definition than some other methods with similar goals, it also tends to limit the number of indirect (non-causal) feature selections to much greater extent than other approaches, such as large-scale univariate testing, that adopt the same marginal perspective on false discoveries. This makes mFDR an attractive and versatile tool in a variety of settings.

Additionally, there are very few methods of false discovery rate control that currently have readily available software implementations. The mFDR procedures described in this paper, namely those outlined in Algorithm's 1 and 2, are implemented in the `grpmfdr` function that is currently available in a forked version of the 'R' package `grpreg` (Breheny and Huang, 2009) which can installed directly from https://github.com/remiller1450/grpreg. Implementations are available for the linear, logistic, and Cox regression models under either the group lasso or group MCP penalty.

From the standpoint of computational burden, the time it takes to incorporate mFDR as part of a broader analysis involving the group lasso is minimal. In the Spira case study, where the expanded design matrix contained 88,860 columns corresponding to 22,215 groups, it takes only a few minutes to estimate the number and rate of marginal false discoveries across an entire sequence of 100 λ values.

Supporting information

Data and source code to reproduce all results and figures are available at https://github.com/remiller1450/grp_mfdr_paper.

References

Barber, R. F. and Candes, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43** 2055–2085.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57** 289–300.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Stat Interface*, **2**.

Breheny, P. and Huang, J. (2012). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, **25**.

Breheny, P. J. (2019). Marginal false discovery rates for penalized regression models. *Biostatistics*, **20** 299–314.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: model-x knockoffs for high dimensional controlled variable selection. *J. Roy. Stat. Soc. B*, **80** 551–577.

Dai, R. and Barber, R. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*. PMLR, New York, New York, USA, 1851–1859. URL http://proceedings.mlr.press/v48/daia16.html.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.*, **17** 347–388.

G'Sell, M. G., Wager, S., Chouldechova, A. and Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *J. Roy. Stat. Soc. B*, **78** 423–444.

Huang, J., Breheny, P. and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, **27** 481 – 499.

Liang, W., Ma, S. and Lin, C. (2021). Marginal false discovery rate for a penalized transformation survival model. *Computational Statistics and Data Analysis*, **160** 107232.

Liu, H. and Zhang, J. (2009). Estimation consistency of the group lasso and its applications. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.), vol. 5 of *Proceedings of Machine Learning Research*. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 376–383. URL http://proceedings.mlr.press/v5/liu09a.html.

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

Miller, R. E. and Breheny, P. (2019). Marginal false discovery rate control for likelihoodbased penalized regression models. *Biometrical Journal*.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22** 231–245.

Simon, N. and Tibshirani, R. (2004). Standardization and the group lasso penalty. *Stat. Sinica*, **22** 983–1001.

Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E. and Brody, J. S. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.*, **13** 361–366.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, **9** 303.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, **58** 267–288.

Tibshirani, R., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Stat. Assoc.*, **111** 600–620.

Wang, S., Nan, B., Zhu, N. and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika*, **96** 307–322.

Yang, F., Foygel Barber, R., Jain, P. and Lafferty, J. (2016). Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds.), vol. 29. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2016/file/7c82fab8c8f89124e2ce92984e04fb40-Paper.pdf.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, **68** 49–67.

Zhou, Q. and Min, S. (2017). Estimator augmentation with applications in high-dimensional group inference. *Electronic Journal of Statistics*, **11** 3039 – 3080.