Marginal False Discovery Rates for Group Lasso Regression

Ryan Miller
Department of Mathematics
Xavier University

Patrick Breheny
Department of Biostatistics
University of Iowa

July 23, 2021

Abstract

Placeholder

1 Introduction

Since its proposal, the lasso (Tibshirani, 1996) has become an immensely popular modeling approach due to its ability to simulataneously 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, $\boldsymbol{\beta}$, with an l_1 penalty imposed on the size of $\boldsymbol{\beta}$. More precisely, the lasso estimator of $\boldsymbol{\beta}$ is defined as the minimizer of:

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

Perhaps the most desirable property of the lasso is that some components of $\hat{\boldsymbol{\beta}}$ are estimated to be exactly 0 for sufficiently large values of the penalty parameter, λ, implying the remaining non-zero elements have been *selected* by the lasso. Unfortunately, selections made by the lasso can be unsatisfactory in applications involving categorical predictors. In these circumstances, the lasso willl select amongst the individual dummy variables rather than the underlying predictors themselves, thereby making selections dependent upon the scheme used to encode the dummy variables and complicating efforts to identify and interpret impactful categorical predictors. The lasso faces a similar barrier in applications where basis expanasions are used to represent non-linear additive effects of predictors, as the overall selection status of an underlying predictor can be unclear if some columns of the basis expansion are selected but others are not.

The group lasso (Yuan and Lin, 2006) is capable of addressing these shortcomings by imposing an l_2 penalty upon groups of coefficients, rather than individual coefficients, thus resulting in selections occuring at the group level. For linear regression applications, the group lasso estimator of $\boldsymbol{\beta}$ is defined:

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

The penalization scheme employed by the group lasso has since been extended to many generalized linear models, with logistic regression (Meier et al., 2008) serving as a notable example among many others.

A potential drawback of the group lasso is the limited scope of inferential methods available after model fitting. While 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), 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 found using the group lasso, rather than the selections of entire groups.

The focus of this paper is on the reliability of group level selections made by the group lasso and its variants. More specifically, we propose methods for controlling the marginal false discovery rate of group selections in the context of group lasso regression. Our proposed methods provide computationally efficient alternatives to the limited set of existing inferential approaches for the group lasso. We demonstrate the robustness of these methods across a variety of data structures, as well as their ability to achieve higher true positive rates than existing inferential approaches. Further, we generalize our methods to extensions of the group lasso, including logistic regression and other generalized linear models, as well as non-convex group penalization schemes (Breheny and Huang, 2012).

2 Background

2.1 Group lasso regression

Consider data of the usual form, $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ records the response values of $i \in \{1, \ldots, n\}$ independent observations, and $\mathbf{X}$ is an n by p design matrix of explanatory variables. We focus on situations in which the columns of $\mathbf{X}$ can be naturally placed 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 regeression model, which specifies the relationship:

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

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

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

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

where λ_j is a penalty applied to l_2-norm of the coefficients in group j. While it is possible to specify λ_j individually for each group, the more common choice is to penalize in accordance to group size by setting $\lambda_j = \sqrt{K_j}\lambda$, where λ is universal across groups.

The underlying framework of the group lasso can be generalized to a wide range of loss functions and penalization schemes whose solutions minimize 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 upon the log-likelihood in the case of generalized linear models. See Huang et al. (2012) for selective review of other grouped penalization schemes.

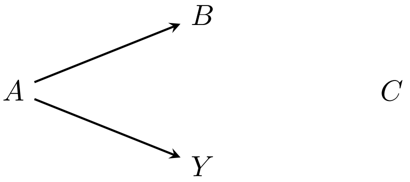
For sufficiently large λ_j, the entire group of coefficients belonging group j will all be estimated as exactly zero, making it possible to use the group lasso to perform variable selection at the group level. Throughout the remainder of this paper, we refer to groups whose group lasso coefficient estimates are non-zero as being "selected" by the group lasso. The group lasso penalty has the added benefit of shrinking coefficient estimates towards zero, therefore allowing identifiable estimation even when the dimensionality of $\mathbf{X}$ is such that $p > n$.

In their original proposal, Yuan and Lin (2006) assume the data 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. Further, the group lasso solution of the orthonormal scale can also be easily converted back to the orginal scale. Throughout the remainder of this paper we assume the data have already 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

The work of Benjamini and Hochberg (1995) has led to false discovery rate (FDR) control becoming perhaps the most widely utilized inferential paradigm in applications involving large numbers of simulataneous comparisons between variables. The literature on false discovery rate control is enormous, with many authors operating under slightly different definitions. One straightfoward way to charactertize the false discovery rate is as the expected number of false selections divided by the total number of selections, or as the fraction of significant features that are false positives. In this regard, a selection procedure that controls the FDR at 10% corresponds to the expectation that no more than 10% of the comparisons it identifies as siginificant are expected to be false positives.

Much of the work in the realm of false discovery rate control pertains to *large scale univariate testing*, or applications that entail the aggregation of results of a large number of single variable hypothesis tests, Farcomeni (2008) and Strimmer (2008) provide a more detailed overview of these methods. In the regression modeling framework, the notion of a false discovery can be considerably more complicated in the presence of relationships between predictors. To better understand these complexities, consider the causal diagram shown below, which illustrates a possible relationship between three explanatory variables, A, B, and C, and an outcome variable Y.



In this scenario, 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. According to the paradigm used in large scale univariate testing, variable B would not be seen as a false discovery because it is not *marginally independent* of variable Y, the criteria that is typically considered in each single variable hypothesis test. Alternatively, in the regression setting, the coefficient in the data generating model corresponding to variable B would be zero, suggesting that variable B should be viewed as a false discovery. In the context of group selections, the underlying idea conveyed by this causal structure can be generalized such that A, B, and C represent groups of variables acting in concert.

To further clarity these distinctions, we introduce two contrasting false discovery rate perspectives: the *marginal* perspective, in which the variable, or group of variables, denoted X_j is a false discovery if it is declared significant despite being fully indpendent 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 as 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 utilize the less restrictive *marginal false discovery rate* definition. While the selection of variables like B in the aforementioned causal diagram can be problematic in certain applications, in most applications it is impossible to untangle the true causal structure of the $A - B - Y$ relationship, and it is universally useful, regardless of application, to control the number of variables like C that are deemed significant. For these reasons, as well as the existing adoption of the marginal definition within the realm of large scale univariate testing, we argue the marginal false discovery rate is a valuable quantity to consider in applications involving the group lasso. Furthermore, marginal false discovery rate inferential approaches 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 arising from the usual linear model:

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

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ 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 $j \in \{1, \dots, J\}$. 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, $\hat{\boldsymbol{\beta}}$, which is 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 $\hat{\boldsymbol{\beta}}$ requires the subdifferential (cite Bertsekas 1999) 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 $\boldsymbol{\beta}_{-j}$ describing the associated model coefficients. In the optimzation literature, the criteria described in 3.3 are known as the KKT conditions.

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

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

Thereby prompting the following theorem.

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 straightfoward algebraic manipulation of **??**, which is a direct consequence of the KKT conditions in 3.3. 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, $\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_{-j}^T \xrightarrow{p} \mathbf{0}$, can be satisified when correlations between the columns belonging to group j and the others in $\mathbf{X}$ become neglible 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 (CITE http://proceedings.mlr.press/v5/liu09a/liu09a.pdf).

The first condition is not a trivial assumption; however, it represents a valuable worst-case scenario in regards to false discoveries. Heuristically, in penalized regression applications where two groups are related the selection of one group decreases the chances that the other is also selected, thereby leading to fewer chance selections than would be expected had there been no relationship across groups. Characterizing the precise degree of conservatism induced in these scenarios is siginificantly less mathematically tractable, but we extensively explore the issue via simulation study in section (REF sim). (CITE BREHENY 2019) provides a more detailed theoretical discussion 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 defined 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 indpendent of the outcome to determine the expected number of falsely selected groups; however, the identity of such groups is unknown in practice, so summing over all J groups provides a conservative alternative. In many applications of the group lasso, the number of groups that are truly related to the outcome tends to be 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 $\sigma^2$ as either $\hat{\sigma}^2 = \frac{\mathbf{r}^T\mathbf{r}}{n-df}$ or $\hat{\sigma} = \text{cve}_\lambda$
    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 taking it to be the cross-validated error of the model under consideration.

3.2 Generalized linear models

The penalization scheme of the group lasso can be extended to other likelihood-based models, notably those corresponding to various generalized linear models (GLMs). In these settings, a quadratic approximation of the log-likelihood function, L, is typically used in solving for the group lasso coefficients:

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

where $\eta = \mathbf{X}\boldsymbol{\beta}$, with $\tilde{\boldsymbol{\eta}}$ denoting the current estimate, and $\mathbf{v}$ and $\mathbf{W}$ respectively represent 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 $\boldsymbol{\beta}$, this approximation yields 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 their corresponding KKT conditions, can be adapted to generalized linear models with the minor addition of a weight matrix, $\mathbf{W}$. Consequently, groups selections within 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 used 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, $\frac{1}{n}\mathbf{X}_j^T\mathbf{W}\mathbf{X}_{-j}^T \xrightarrow{p} \mathbf{0}$, and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is bounded in probability. The number and rate of marginal false discoveries in likelihood-based models subjective to the group lasso penalty can be characterized as follows:*

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

The first condition involved in 1 is a standard result of classical likelihood theory which can be shown for many types of generalized linear models, while the other two conditions are direct analogs of those in 1 for the linear regression setting. Consequently, the underlying derivation giving rise to the estimates in 3.13 is analgous to that of Theorem 1, with the primary difference being that the variance used in normalization. Like before, we can summarize the procedure involvied in calculating mFDR for given value of λ using the following algorithm:

Algorithm 2 Calculating the mFDR upper bound (GLMs

```
procedure
    Estimate $\widehat{W} \leftarrow \nabla^2 f(\hat{\boldsymbol{\eta}})$
    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, the estimates arising from 2 will be conservative in the presence of correlations between explanatory variables, or in applications were J is substantially greater than the number of groups that are marginally independent of the outcome. Additionally, these estimates are subject to added uncertainty from the use of the average diagonal element of $\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j$ in normalization. In Section 4 we demonstrate the reliability of these estimates when applied to group lasso logistic regression.

3.3 Other penalty functions

The general form of the Mfdr estimator in 3.8 is directly applicable to a number of other penalizations schemes that are related to the group lasso. One example is the minimax concave penalty, or MCP, whose penalty function,

$f_{\lambda,a}(\theta)$, is defined by

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

for values of $\lambda \geq 0$, where a is a tuning parameter, with $a = 3$ being typical in software implementations. MCP is a nonconvex penalty where the degree of penalization is diminished, eventually becoming zero, for large coefficients. In the grouped setting, the composite group MCP estimate (Breheny and Huang, 2012; Huang et al., 2012) is the minimizer of

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

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.

While the coefficient estimates that arise from group MCP regression may be somewhat different than those of the group lasso, the optimization conditions that describe the entry criteria for groups to become active in the model remains the same. Consequently, the same estimators proposed in **??** and 3.8 can also be used to control the marginal false discovery rate in group MCP regression, even though the resulting group selections may differ. In Section 4 we consider both the group lasso and group MCP penalization schemes when presenting numerical results on simulated data.

4 Simulation experiments

4.1 Data generation

In all experiments, the covariates stored in the design matrix, $\mathbf{X}$, are derived from numeric values randomly generated from a multivariate normal distribution, $N(0, \Sigma_X)$. The off-diagonal entries of Σ_X are used to invoke specific correlation structures between features. More specifically, we focus on two different correlation structures: Independence - $cor(\mathbf{x}_a\mathbf{x}_b) = 0$, and Autoregressive - $cor(\mathbf{x}_a, \mathbf{x}_b) = \rho^{|a-b|}$, for all $a, b \in \{i, \ldots, J\}$. These numeric values are used to construct design matrices for the two primary applications of the group lasso that we choose to focus on, models which involve nominal categorical predictors and non-parametric additive models.

In the former application, each variable's underlying numeric values are binned into k equally sized groups, which are then expressed in the design matrix via k dummy variables. Following discretization, outcomes are generated either 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, or from a Bernouli distribution with $\Pr(y_i = 1) = 1/(1 + exp(-\mathbf{x}_i^T\boldsymbol{\beta}))$ in logistic regression applications.

In the later application, for linear regression scenarios, outcomes are generated directly from the underlying numeric values according to the model $\mathbf{y} = \sum_{j=1}^{t} f(\mathbf{x}_j) + \boldsymbol{\epsilon}$, 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. In logistic regression scenarios, the outcome is drawn from a Bernoulli distribution with $\Pr(y_i = 1) = 1/(1 + exp(-\sum_{j=1}^{t} f(\mathbf{x}_{ij})))$. We consider three forms of f: Quadratic - $f(X) = X^2$, Piecewise Linear - $f(X) = \frac{0.7}{\sqrt{n}}X$ if $X > 0$ and $f(X) = 0$ otherwise, and Periodic - $f(X) = sin(X)$. Prior to modeling, each feature undergoes a basis expansion with 3 degrees of freedom, thereby yielding a design matrix, $\mathbf{X}$, consisting of J groups of size $k = 3$.

4.2 False discovery rate control

The estimators described in Equations 3.8 and 3.13 allow for the expected number of false discoveries to be calculated for every model along a the decreasing sequence of λ values that is typically returned by software that fits group lasso models. The curves in (FIG1) 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, non-parametric additive model scenario for both linear and logistic regression with $n = 1000$, $J = 100$, and $t = 10$ for three different correlation structures, autoregressive with $\rho \in \{0.5, 0.9\}$, and independence.

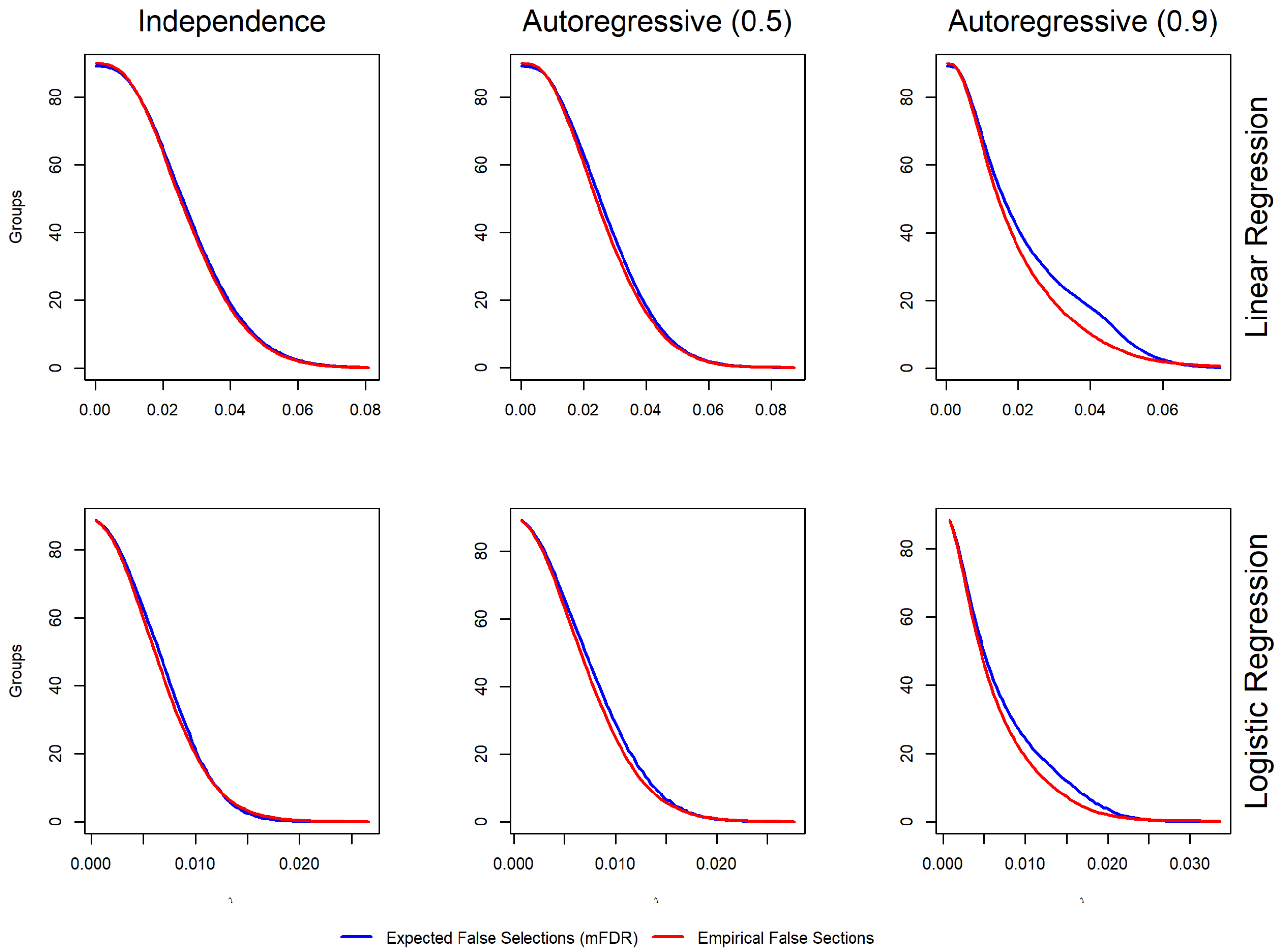


Figure 1: Comparison of the average expected and empirical number of marginal false discoveries along a sequence of λ values.

In the linear regression setting, under independence, Figure 1 demonstrates that the expected number of false selections very tightly resembles the empirical number, with only a very slight degree of over-estimation induced by incidental correlations between the columns of $\mathbf{X}$. As groups become more correlated, the estimates provided by Equation 3.8 become more conservative. However, even for $\rho = 0.9$ at the value of λ exhbititng the greatest discrepency Equation 3.8 on average suggests only YYY more false discoveries than tend to actually occur, a diference of ZZZ percent. Results are similar in the logisitic regression setting.

Focusing on the independence correlation structure and varying $n \in \{200, 400, 600, 800, 1000\}$, we can evaluate the role of sample size in the accuracy mFDR estimates from Equation 3.8. Figure 2 displays the mean difference between the estimated and empical false discovery rate, which is defined as $\frac{\text{\#Noise selections}}{\text{\#Total selections}}$, when selecting the smallest value of λ with $\widehat{\text{Mfdr}} \leq 0.10$ along a fixed sequence across 400 simulation iterations. As the sample size increases, the discrepency between mFDR and the nominal rate diminishes towards zero.

Figure 2 also demonstrates the compatibility of the mFDR method with the group MCP penalty. While the group selections and coefficient estimates under this penalization scheme tend to differ from those of the group lasso, the performance of the mFDR estimator remains consistent, if not slightly improved due to more accurate estimation of $\boldsymbol{\beta}$.

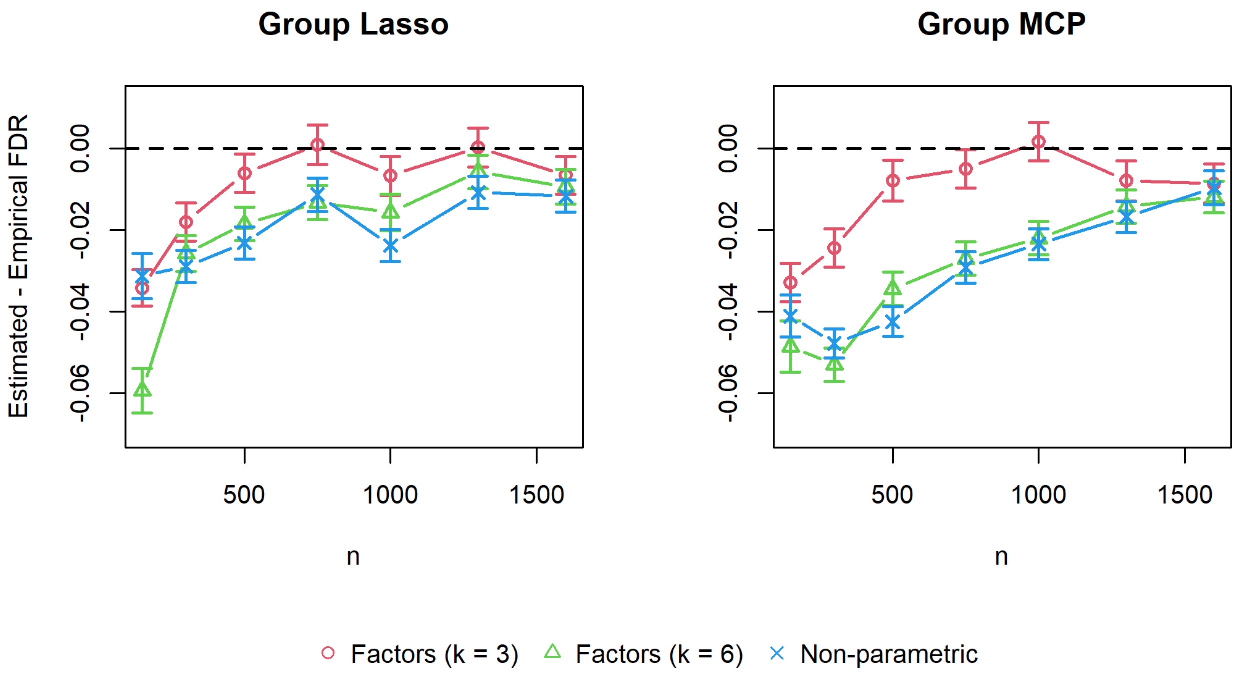


Figure 2: Tightness of the mFDR upper bound as n increases.

4.3 Comparison with other regression-based methods

Our third simulation experiment explores the true positive rate of Mfdr approach in comparison with other existing methods for controlling the false discovery rate in the context of group-level selection. We consider the following competing approaches:

- Forward selection with selective inference
- Data-splitting

4.4 Comparison with cross-validation

4.5 Comparison with univariate methods

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 due 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. Of 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 analysis approaches using the genetic data to predict case-control status. Our primary approach applies a basis expansion with 4 degrees of freedom to each genetic feature, thereby creating a new desing matrix, $\tilde{X}$, containing 88,860 columns which correspond to 22,215 groups of size $K_j = 4$. The mFDR estimator proposed in 3.13 is then used to select a group lasso model that controls the marginal false discovery rate.

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

5.3 Results

Table 1: 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	$\tilde{X}$	CV	45	53.4%	25.5%
group lasso	$\tilde{X}$	mFDR	21	5.6%	27.1%
large-scale testing	$\tilde{X}$	-	12,902	10.0%	-
large-scale testing	$\mathbf{X}$	-	2,426	10.0%	-

Table 1 summarizes the outcome each analysis approach applied to the Spira dataset. Among penalized regression approaches, the lowest cross-validated misclassification error is obtained by applying the ordinary lasso to the unexpanded design matrix; however, the estimated mFDR of this model is 100%. Because the mFDR bound is inherently conservative, an estimate of 100% doesn't necessarily indicate that all of 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% would limit the number of genetic features selected to only 10 while at the same time substantially increasing the cross-validated misclassification error. In contrast, when applying the group lasso to the expanded design matrix, 21 features can be selected while controlling the marginal false discovery rate below 10% while achieving a misclassification error that is much closer to the minimum achieved by the ordinary lasso.

Comparing the number of selections made by the regression-based methods in Table 1 to the results of large-scale testing highlights a key advantage of the mFDR. Although all of these methods are based the same marginal perspective on false discoveries, regression-based methods tend to naturally limit the number of highly correlated features that are deemed significant. That is, in situations where a large number of 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 such as this one, large-scale testing can result in an overwhelming amount of "leads" that researchers must then filter, group, or assess manually, while regression-based approaches mFDR tend to yield a more managable set of features with less redundancy.

Figure 3 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 the models corresponding to various values of a decreasing λ sequence, demonstrating a wide range of models that achieve statistically similar levels

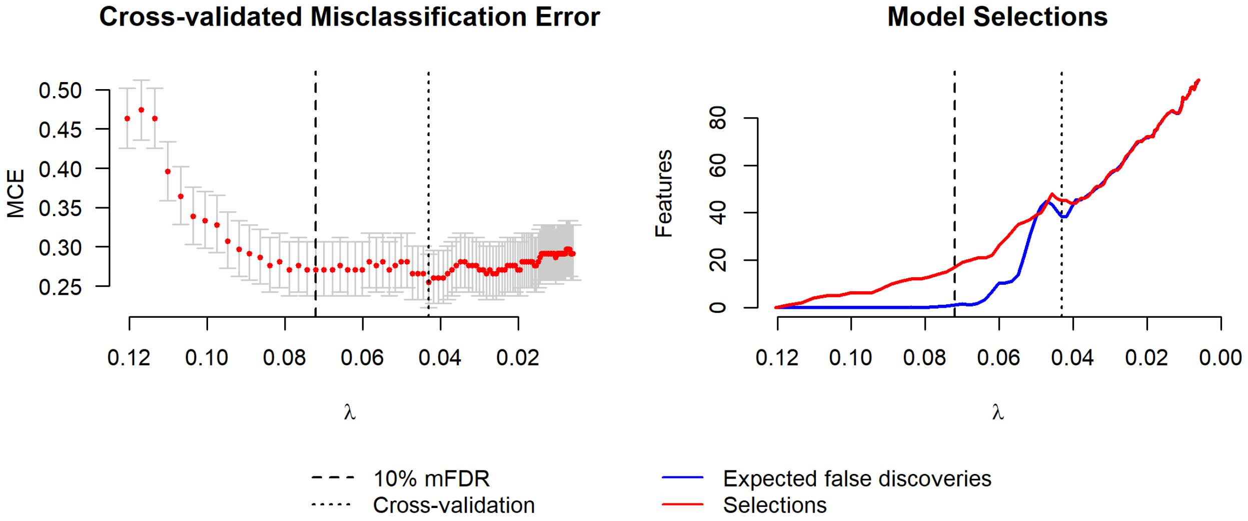


Figure 3: Group lasso modeling results for the Spira case study. The left panel displays the cross-validated misclassification error for the models corresponding to various values of a decreasing sequence of λ, while the right panel shows the total number of selections alongside the expected number of 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.

of accuracy. The right panel shows the expected number of false discoveries, calculated using the estimators in 3.13, and the total number of selections for each of these models. Together these plots can be used to selection 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

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. (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.

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.

Meier, L., Van De Geer, S. and Bühlmann, 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 likelihood-based penalized regression models. *Biometrical Journal*.

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.

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.