Marginal false discovery rate control for likelihood-based penalized regression models

Ryan Miller
Department of Biostatistics
University of Iowa

Patrick Breheny
Department of Biostatistics
University of Iowa

November 20, 2018

Abstract

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

1 Introduction

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

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

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

One approach, which we refer to as *sample splitting*, is based upon dividing the data into two parts, using the first part for variable selection and the second part for inference. **?** first proposed this approach using single split, and **?** extended it by considering multiple random splits. More recently, **?** and **?** have proposed a family of methods known as *selective inference*, which test the significance of each variable selection along the lasso solution path as λ is decreased, conditional upon the other variables already active in the model. Based on this sequence of tests, formal stopping rules can be derived in order to control the false discovery rate at a specified level (**?**). Several other approaches exist – e.g., procedures based on the bootstrap (**?**) or the idea of de-biasing (**?**). A comprehensive review is beyond the scope of this manuscript, so we pay particular attention to sample splitting and selective inference as representative of two different frameworks, as described in Section 2.

All of these approaches, however, suffer from two primary drawbacks. First, they are very restrictive, in the sense that they are often unable to select more than one or two features, even at high false discovery rates. This is especially true in high dimensions. Second, they are quite computationally intensive. For both approaches, it typically takes several orders of magnitude longer to estimate the false discovery rate than to fit the model in first place.

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

2   Marginal false discovery rates

False discoveries are straightforward to define when conducting single variable hypothesis tests; a false discovery occurs when a feature X_j is declared to be associated with an outcome Y even though the feature is actually independent of the outcome: $X_j \perp\!\!\!\perp Y$. In the regression framework, where many variables are being considered simultaneously, the idea of a false discovery is more complicated. The most common approach, which we refer to as the *fully conditional* perspective, is to consider a feature X_j to be a false discovery if it is independent of the outcome conditional upon all other features: $X_j \perp\!\!\!\perp Y|X_{k\neq j}$. Penalized likelihood methods typically result in only a subset of the available variables being active in the model, thereby motivating the *pathwise conditional* perspective. This perspective focuses on the model where X_j first becomes active and conditions only on the other variables present in the model (denote this set M_j) at that time when assessing whether or not variable j is a false discovery: $X_j \perp\!\!\!\perp Y|X_k$ for $k \in M_j$.
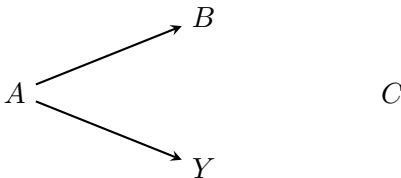
In this paper we derive our method under the less restrictive *marginal false discovery* definition (**?**), which we illustrate using the causal diagram depicted above. In this diagram variable A has a direct causal relationship with the outcome variable Y and should never be considered a false discovery. Variable C is independent of variable Y regardless of any variables we adjust for and should always be considered a false discovery.

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

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

2.1   Penalized likelihood optimization

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

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

$$Q(\boldsymbol{\beta}|X, \mathbf{y}) = -\frac{1}{n}\ell(\boldsymbol{\beta}|X, \mathbf{y}) + P_\lambda(\boldsymbol{\beta}). \tag{2.1}$$

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

In a likelihood-based regression model, the likelihood depends on $\mathbf{X}$ and $\boldsymbol{\beta}$ through a linear predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$; in other words, we can equivalently express the likelihood in terms of a loss function $f(\boldsymbol{\eta}|\mathbf{y}) = -\ell(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})$. In what follows, we assume that the loss function is strictly convex with respect to the linear predictors $\boldsymbol{\eta}$; note that this does not imply strict convexity with respect to $\boldsymbol{\beta}$. Under these conditions, any solution $\hat{\boldsymbol{\beta}}$ that minimizes (2.1) with the lasso penalty $P_\lambda(\boldsymbol{\beta}) = \lambda ||\boldsymbol{\beta}||_1$ must satisfy (**?**):

$$\begin{aligned} \frac{1}{n}u_j(\hat{\boldsymbol{\beta}}) &= \lambda \, \text{sign}(\hat{\beta}_j) && \text{if } \hat{\beta}_j \neq 0 \\ \frac{1}{n}u_j(\hat{\boldsymbol{\beta}}) &\in [-\lambda, \lambda] && \text{if } \hat{\beta}_j = 0 \end{aligned} \tag{2.2}$$

for $j \in \{1, \ldots, p\}$.

2.2   Marginal false discovery rate bounds for penalized likelihood methods

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

Three regularity conditions are required for these theoretical results to hold. These are given below, where we let $\mathbf{W} = \nabla^2 f$ denote the $n \times n$ matrix of second derivatives of the loss function with respect to $\boldsymbol{\eta}$, such that the classical Hessian matrix $\nabla^2 \ell(\boldsymbol{\beta}) = -\mathbf{X}^T \mathbf{W} \mathbf{X}$. We use $\mathbf{W}$ to denote this matrix evaluated at the true value of $\boldsymbol{\beta}$ and $\widehat{\mathbf{W}}$ if evaluated at the lasso estimate. In addition, we let $v_j = \mathbf{x}_j^T \mathbf{W} \mathbf{x}_j$, with $\hat{v}_j$ defined similarly.

(A1) Asymptotic normality of the score function: $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1/2}\mathbf{u}(\boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ denotes the $p \times p$ identity matrix.

(A2) Vanishing correlation: $\frac{1}{n}\mathbf{x}_j^T\mathbf{W}\mathbf{X}_{-j} \xrightarrow{p} \mathbf{0}$.

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

(A1) is a standard result of classical likelihood theory and can be shown for many types of models. (A3) is not a trivial condition, but has been studied and shown to hold for various models and various types of penalties under certain conditions. (A2), on the other hand, is unlikely to be truly satisfied by most features in practice. Our theoretical results illustrate what is required for the proposal to work perfectly in the sense of providing a consistent estimate for the mFDR. In practice, (A2) serves as a worst-case scenario for the correlation structure in the sense that other correlation structures will, on average, lead to fewer false discoveries. Viewed as an estimator, this means that the mFDR equation we will derive (2.4) is inherently conservative, in the sense that it will overestimate the true mFDR. Viewed instead as a control procedure, this means that (2.4) provides a probabilistic upper bound on the false discovery rate, and that if we use this equation to limit the mFDR to, say, 10%, we can be confident that the true mFDR is even smaller.

The fundamental reason for this is that if a pool of noise features are uncorrelated, a variable selection method may select several of them, whereas if they are correlated, the method will tend to select just one. More explicit results are shown in Section 3, which shows that when (A2) is violated, the mFDR bound is less tight, as one would expect. Assumptions (A1)-(A3) are further discussed in the specific cases of logistic and Cox regression later in this section.

We now formally state our main theoretical result. In interpreting this result, it is important to keep in mind that only features that are marginally independent of the outcome will satisfy both assumption (A2) and $\beta_j = 0$. In other words, the theorem applies to variables like C in the causal diagram of Section 2, but not variable B: although the regression coefficient for B is zero, its correlation with A violates (A2).

Theorem 1. *For any solution $\hat{\boldsymbol{\beta}}$ of the lasso-penalized objective (2.1), we have $\hat{\beta}_j \neq 0$ if and only if*

$$\frac{1}{n}\left|u_j(\hat{\boldsymbol{\beta}}) + v_j\hat{\beta}_j\right| > \lambda. \tag{2.3}$$

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

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

Proof. The first remark follows directly from the KKT conditions (2.2) and the fact that, if the loss function $f(\boldsymbol{\eta}|\mathbf{y})$ is strictly convex, then $\mathbf{W}$ is positive definite and v_j is positive. The asymptotic normality of the score function (A1) implies that the following Taylor series expansion holds:

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

Since $\beta_j = 0$, we then have

$$u_j(\hat{\boldsymbol{\beta}}) = u_j(\boldsymbol{\beta}) - \mathbf{x}_j^T\mathbf{W}\mathbf{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{\beta}_{-j}) - \mathbf{x}_j^T\mathbf{W}\mathbf{x}_j\hat{\beta}_j + o_p(n)\hat{\beta}_j,$$

or

$$\frac{u_j(\hat{\boldsymbol{\beta}})+v_j\hat{\beta}_j}{\sqrt{v_j}} = \frac{u_j(\boldsymbol{\beta})}{\sqrt{v_j}} - \sqrt{\frac{n}{v_j}}[\frac{1}{n}\mathbf{x}_j^T\mathbf{W}\mathbf{X}_{-j}][\sqrt{n}(\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{\beta}_{-j})] + o_p(1)\sqrt{\frac{n}{v_j}}\sqrt{n}\hat{\beta}_j.$$

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

4

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

$$\widehat{\text{FD}} = 2\sum_{j=1}^{p} \Phi\left(\frac{-n\lambda}{\sqrt{\hat{v}_j}}\right)$$
$$\widehat{\text{mFDR}} = \frac{\widehat{\text{FD}}}{|S|},$$

(2.4)

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

Algorithm 1 Calculating the mFDR upper bound

procedure
    Estimate $\widehat{W} \leftarrow \nabla^2 f(\hat{\boldsymbol{\eta}})$
    for $j \in \{1, \ldots, p\}$ **do**
        $\hat{v}_j \leftarrow \mathbf{x}_j^T \widehat{W} \mathbf{x}_j$
        $\widehat{\text{FD}}_{j,\lambda} = 2\Phi\left(\frac{-n\lambda}{\sqrt{\hat{v}_j}}\right)$ by the distributional result of Theorem 1
    $\widehat{\text{FD}}_\lambda = \sum_{j=1}^{p} \widehat{\text{FD}}_{j,\lambda}$
    $\widehat{\text{mFDR}}_\lambda = \min\left(\frac{\widehat{\text{FD}}_\lambda}{|S_\lambda|}, 1\right)$
    return $\widehat{\text{mFDR}}_\lambda$

It is worth noting that the version of Theorem 1 presented in this paper takes p to be fixed. It is of interest to extend Theorem 1 to allow $p \to \infty$ as $n \to \infty$ and determine how this affects convergence; we leave this as a topic to be pursued in a future manuscript.

2.3 Other penalty functions

The form of the mFDR calculation (2.4) is determined by the KKT conditions (2.2). Although the theorem in previous Section is specific to the lasso, many other penalties proposed in the literature have very similar KKT conditions. This in turn allows our results to be easily extended to other penalized methods.

For example, consider the elastic net (**?**), which utilizes two penalty parameters: λ_1 and λ_2, with $\lambda_1, \lambda_2 > 0$. The elastic net solution is found by minimizing $-\frac{1}{n}\ell(\boldsymbol{\beta}|X, \mathbf{y}) + \lambda_1||\boldsymbol{\beta}||_1 + \frac{\lambda_2}{2}||\boldsymbol{\beta}||_2^2$. The resulting KKT conditions dictate that $\hat{\beta}_j \neq 0$ if and only if

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

Compared with the corresponding equation for the lasso (2.3), the only change is that the right hand side of the selection condition has changed from λ to λ_1. Thus, equation (2.4) applies to the elastic net as well, with a similarly trivial change (replacing λ with λ_1). Note that the actual estimates $\hat{\boldsymbol{\beta}}$ (and as a result $\widehat{\mathbf{W}}$, $\hat{v}$, and $|S|$) may certainly change a great deal, and thus the resulting inferences may be very different, but the form of the mFDR upper bound is essentially identical.

Furthermore, in some cases, such as for MCP and SCAD, the form of the upper bound is exactly the same. In other words, equation (2.3) holds for these penalties as well as the lasso, and therefore the mFDR upper bound is unchanged from (2.4) – although again, the actual estimates $\hat{\boldsymbol{\beta}}$ and any quantities based on them will be very different.

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

2.4 Logistic regression

So far we have presented our results in the general setting of penalized likelihood optimization. One specific type of model to which Section 2.2 applies is penalized logistic regression. Suppose y_i follows a Bernoulli distribution with success probability π_i. In logistic regression the logit of π_i is modeled as a function of $\boldsymbol{\eta} = X\boldsymbol{\beta}$. This results in a likelihood consisting of the product of n independent Bernoulli distributions with success probabilities $\pi_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta})\}$ and $\mathbf{W}$ a diagonal matrix whose entries are given by $\pi_i(1 - \pi_i)$.

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

2.5 Cox regression

The results of Section 2.2 also apply to penalized Cox proportional hazards regression. Here the outcome of interest contains two components: a time, y_i, along with an accompanying indicator variable, d_i, where $d_i = 1$ indicates y_i is an observed event time and $d_i = 0$ indicates y_i is a right censoring time.

Let $t_1 < t_2 < \ldots < t_m$ be an increasing list of unique failure times indexed by j. The Cox model assumes a semi-parametric form of the hazard such that $h_i(t) = h_0(t)e^{\mathbf{x}_i^T\boldsymbol{\beta}}$, where $h_i(t)$ is the hazard for observation i at time t and $h_0(t)$ is a common baseline hazard. Cox regression is based upon the partial likelihood (**?**)

$$L(\boldsymbol{\eta}) = \prod_{j=1}^{m} \frac{\exp(\eta_j)}{\sum_{i \in R_j} \exp(\eta_i)},$$

where R_j denotes the set of observations still at risk at time t_j, known as the risk set.

Letting $\pi_{ij} = \exp(\eta_i)/\sum_{k \in R_j}\exp(\eta_k)$, the ith diagonal element of $\mathbf{W}$ is $\sum_j d_j\pi_{ij}(1 - \pi_{ij})$, while its i, kth off-diagonal element is given by $-\sum_j d_j\pi_{ij}(1 - \pi_{kj})$. As discussed in **?**, however, the off-diagonal elements of $\mathbf{W}$ are typically negligible except for very small sample sizes. Thus, for the simulations we present in Section 3, we took the off-diagonal elements of $\mathbf{W}$ to be zero in order to speed up the calculations. We found no meaningful difference in the calculation of $\widehat{\mathrm{mFDR}}$ when using the diagonal approximation of $\mathbf{W}$ in place of the full matrix.

The asymptotic normality of the score function for Cox regression (A1) is established in **?**, while necessary conditions for $\sqrt{n}$-consistency of the lasso estimates for Cox regression (A3) can be found in **?**. There are, however, some additional considerations considering censoring with respect to assumption (A2) in Theorem 1. As discussed earlier, (A2) holds for features that are marginally independent of all other features as well as the outcome. For (A2) to hold for Cox regression, however, the feature must also be independent of the censoring mechanism. This additional requirement is needed because when a variable is related to the censoring mechanism, its distribution will drift over time as certain values are disproportionately removed from the risk set, which can induce correlations between variables that would otherwise be uncorrelated. The impact of this additional concern on $\widehat{\mathrm{mFDR}}$ is further assessed via simulation in Section 3.3.

3 Simulation studies

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

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

In this section and the next, we compare our proposed mFDR control method with sample splitting and covariance testing. The covariance testing approach was implemented using the `covTest` package (**?**) and the forwardStop function of the SelectiveInference package (**?**); however, the current versions of these packages caution that this approach is considered "developmental" for logistic regression and is not currently implemented for Cox regression. Although software does exist for sample splitting for linear regression via the `hdi` package **?**, the current version of the package does not offer methods for logistic or Cox regression models, so we manually implemented our own sample splitting approach.

3.1 Accuracy, sample size, and correlation

In Section 2.2, the mFDR bound (2.4) was shown to be tight in an asymptotic sense provided that condition (A2) was met. To see how useful this result is in practice, we wish to observe how tight this bound is at finite sample sizes as well as when condition (A2) is violated – i.e., when noise features are correlated. We evaluate the accuracy of the mFDR bound by comparing $\widehat{\text{mFDR}}$ with the true mFDR – i.e., the empirical proportion of noise features selected – at each value along a fixed λ sequence, averaged across 1,000 simulated data sets. In this simulation, time-to-event outcomes were uncensored; the effect of censoring on $\widehat{\text{mFDR}}$ is considered in the next section.

We generate our data using $p = 100$ with $\beta_{1:4} = 10/\sqrt{n}$ and $\beta_{5:40} = 0$ while varying n from 200 to 1,600 in increments of 200. We take $\boldsymbol{\beta}$ to be function of n in order to maintain the difficulty of the variable selection problem as n increases. Without this provision, the feature selection problem becomes too easy at large sample sizes and the "interesting" region where the mFDR is not close to 0 or to 1 is very small.

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

- Two different correlation structures: independent noise variables and correlated noise variables

- Three different regression methods: linear regression, logistic regression, and Cox regression

- Two different penalties: lasso and MCP.

For our correlated setting noise variables are given an autoregressive correlation structure based upon their index such that $\text{Cor}(\mathbf{x}_j, \mathbf{x}_k) = 0.8^{|j-k|}$. We display the mean value of $\widehat{\text{mFDR}}$ at the λ value where the average observed proportion of noise variables is 20%.

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

When noise features are correlated, the $\widehat{\text{mFDR}}$ bound is always conservative, at all sample sizes. In terms of Theorem 1, when noise features are correlated, (A2) is violated because the term $\frac{1}{n}\mathbf{x}_j^T \mathbf{W} \mathbf{X}_{-j}$ does not converge to zero, leaving behind a remainder term that is unaccounted for in (2.4). Nevertheless, while this does mean that the bound is less tight, it does not prevent the use of the method for the purposes of mFDR control. Furthermore, in this case, the effect is relatively slight: for the most part, $\widehat{\text{mFDR}}$ is between 35% and 40% when the true mFDR is 20%.
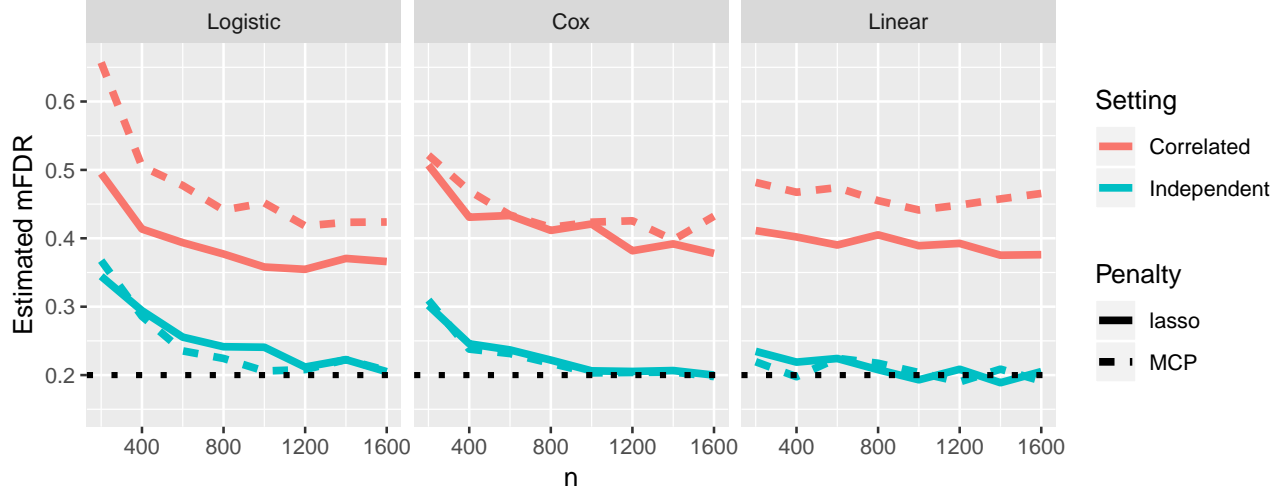
Figure 1: Finite-sample accuracy of the mFDR bound (2.4).

3.2 Features associated with censoring

As discussed in Section 2.5, the presence of censoring in Cox regression means that condition (A2) will also be violated if noise features are associated with the censoring mechanism. Here, we assess the impact such an association has on the $\widehat{mFDR}$.

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

$$\boldsymbol{\gamma}_A = (0, 0, 0, 0, .45, -.45, .45, -.45, 0, \ldots, 0)$$

for scenario A, while in scenario B all variables are independent of censoring (i.e., $\boldsymbol{\gamma}_B = \mathbf{0}$). By design, each of these scenarios results, on average, in a censoring rate of 50%. The same set of $n = 500$ true failure times is used for both scenarios and are generated from independent exponential distributions with rate parameter $\theta_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, where $\beta_{1:4} = .45$ and $\beta_{5:40} = 0$. The value 0.45, used in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, is approximately $10/\sqrt{n}$ for this scenario, so that the results of this simulation are comparable with those displayed in Figure 1.

In scenario A, several noise features are associated with the censoring mechanism. Here, when $\widehat{mFDR}$ is 10%, the observed false discovery rate averaged across 1,000 simulations is 4.03% using the lasso penalty and 6.31% using MCP. In scenario B, where all censoring assumptions are met, the observed false discovery rates are 5.01% and 6.85% for the lasso and MCP, respectively. These results demonstrate that noise variable associated with the censoring mechanism will lead to a more conservative upper bound; however, the effect is negligible compared to the considerations described in Section 3.1.

3.3 Comparison with other methods

In this section, we compare our proposed mFDR approach with other methods of variable selection used in the high-dimensional setting. We generate our data based upon the structure depicted in the causal diagram of Section 2, using $n = 400$ and $p = 1000$ with ten variables causally related to the outcome (like "A"), such that $\beta_{1:10} = b$ and vary b throughout the study. Corresponding to each of these variables, we generate nine correlated ($\rho = 0.5$) variables (akin to "B"). The remaining 900 variables are noise (akin to "C"), however they are correlated with each other such that $\text{Cor}(\mathbf{x}_j, \mathbf{x}_k) = 0.8^{|j-k|}$, thereby creating a situation where the mFDR upper bound will be conservative.

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

- Our mFDR approach, where variables are selected by the lasso using the smallest λ with $\widehat{mFDR} \leq .1$.
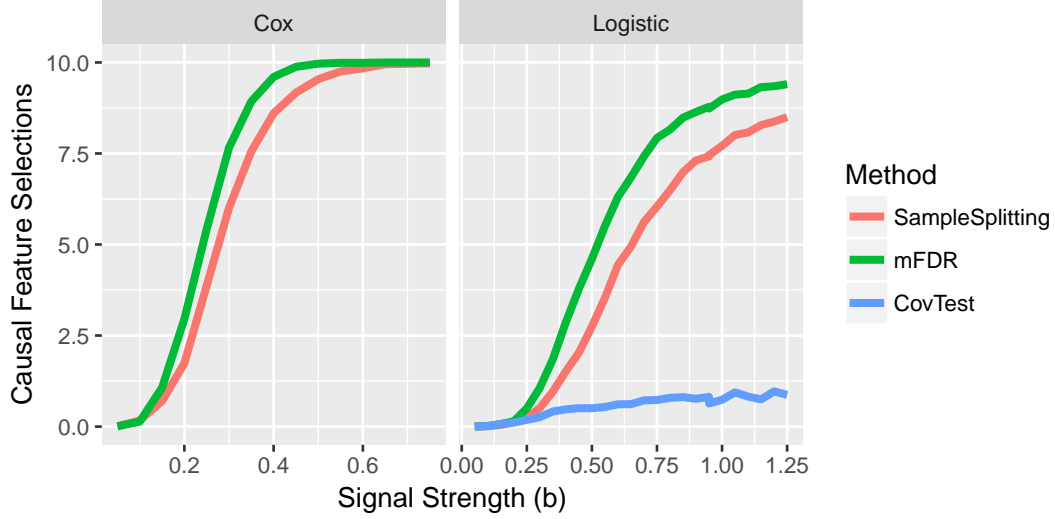
Figure 2: The number of true, "A", variables selected by each lasso based method of false discovery rate control, averaged across 1,000 simulation iterations plotted as a function of β.

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

- Covariance testing (**?**), which we use in conjunction with the forward stopping rules proposed by **?** to control the pathwise false discovery rate at 10%.

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

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

The comparison of FDR control procedures for the lasso (mFDR, sample splitting, and covariance testing) is shown in Figure 2. Compared to other false discovery rate control methods for penalized regression, the mFDR approach is more powerful – at the same nominal FDR, the marginal approach selects more causally important variables at each signal strength. This is not a surprising result: a primary motivation of the marginal approach is that, by adopting a less restrictive definition of false discovery, we could improve power. The covariance testing approach in particular is particularly conservative in the logistic regression setting, selecting far fewer causal features than either mFDR or sample splitting. As of the time of this writing, the covariance test has not been extended to Cox regression and is therefore not shown in the left panel of Figure 2.

Simulation results for cross validation are shown in Figure 3. As the figure shows, while cross-validation has excellent sensitivity, selecting nearly all the causal ("A") variables even at low signal strength, it does not control the number of false selections. In fact, at all levels of b, the majority of the features in the CV-selected model are noise features. This problem becomes increasingly worse as the signal strength increases. For example, at $b = 0.8$, over 70% of the features in the CV model are noise features. In contrast, as may be seen in Figure 4, the vast majority of the features selected by the mFDR control approach are causally related to the outcome.

Figure 4 compares the mFDR and univariate approaches, both of which are designed to control the proportion of marginal false discoveries without making any claims regarding features indirectly associated with the outcome.
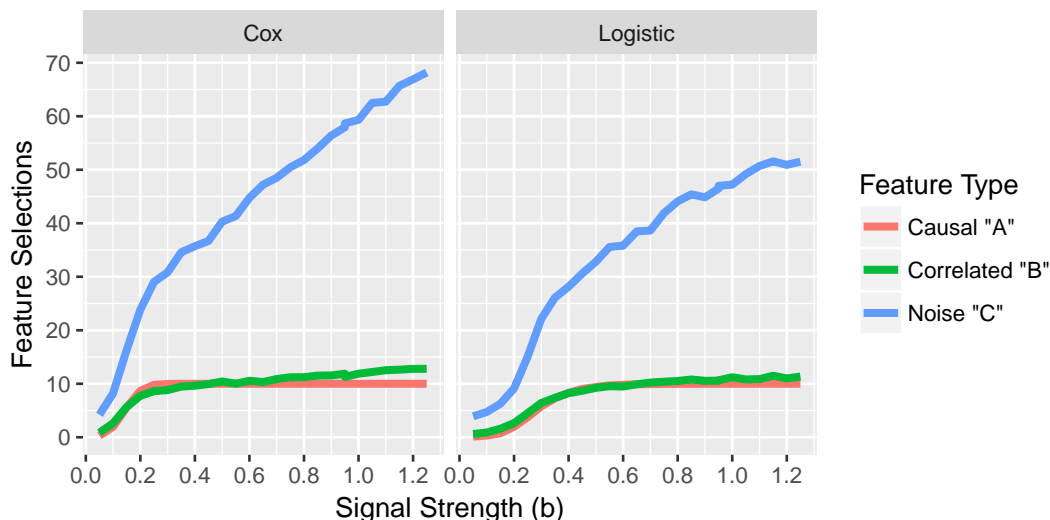
Figure 3: The average number of selections for each type of variable depicted in the causal diagram of Section 2, when cross-validation is used to select λ.

The top panels show that the two methods are comparable in terms of their ability to select the true causal variables. The advantage of using a regression-based approach over univariate testing, however, is clearly seen when looking at the number of other selections. Lasso with mFDR control greatly reduces the number of correlated ("B") selections compared to univariate testing: whereas univariate testing tends to identify dozens of indirectly associated features as discoveries, lasso with mFDR control tends to select at most two.

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

For this reason, Figure 4 is potentially somewhat misleading in terms of comparing mFDR and univariate testing in terms of their power to identify type "A" features. Table 1 presents an alternative metric for comparing the univariate and mFDR approaches. Given the large difference between the two approaches with respect to number of indirectly associated ("B") features selected, the table presents the ratio of the number of causally associated features to the number of noise features – i.e., the A:C ratio – for various values of β. Whereas Figure 4 suggests that the two approaches were approximately equally effective at identifying the truly important variables, Table 1 shows that the lasso-mFDR approach is far more powerful in terms of the number of true discoveries per noise feature selected.

Table 1: The ratio of Causal:Noise (A:C) feature selections for various values of the signal strength b.

	Cox		Logistic	
b	Univariate	Lasso-mFDR	Univariate	Lasso-mFDR
0.25	3.5	17.7	2.3	3.3
0.35	3.0	36.2	4.9	11.1
0.45	2.9	102.2	5.3	15.3
0.65	2.4	333.3	4.6	32.6
0.75	1.9	750.0	4.1	47.6

Indeed, for the univariate approach, the rate of true variable selections per noise variable selection is not only low, but in fact decreases as the signal strength increases. In contrast, with the mFDR approach, the causal:noise selection ratio increases with signal strength, as one would hope.
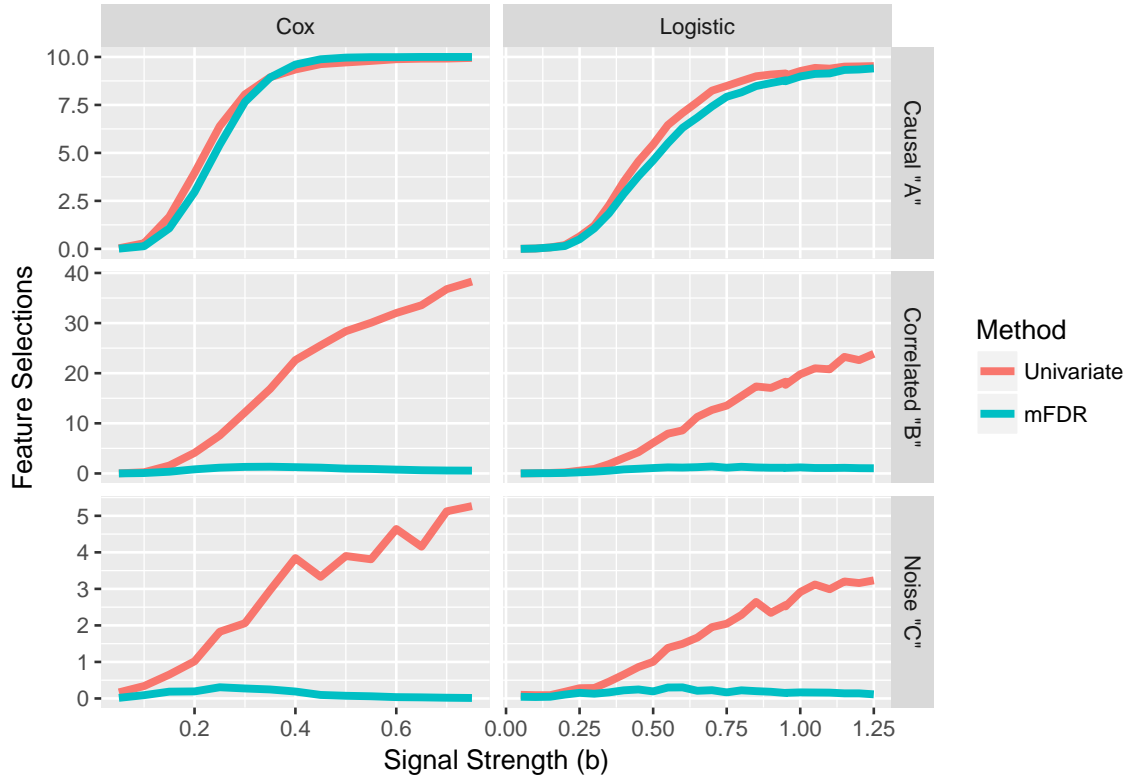
Figure 4: The average number of selections, for each type of variable depicted in the causal diagram of Section 2, is plotted as function of β for each method which controls the false discovery rate.

4 Case studies

4.1 Lung cancer survival and gene expression

? studied the survival of 442 early-stage lung cancer subjects. Researchers collected high-dimensional gene expression data on 22,283 genes and additional clinical covariates of age, race, gender, smoking history, cancer grade, and whether or not the subject received adjuvant chemotherapy. The retrospective study used time to death as its outcome, which was observed in 236 (53.3%) of the subjects, the remainder being right censored.

In our analysis we aim to select important genes while controlling the marginal false discovery rate. We use a semi-penalized Cox proportional hazards regression model that allows all of the clinical covariates to enter the model unpenalized while using the sparsity introduced by penalization to screen for additional genes related to survival. We aim to limit the mFDR to 10% and compare our results to those of other methods. Note that in this example, the number of features p in equation (2.4) is the number of penalized features (i.e., the number of genes), since the clinical covariates will always be in the model.

Figure 5 illustrates the mFDR estimates for this study. On the left, we see a gap between the number of selected genes and the expected number of false selections at $\log(\lambda) \approx -2$. This indicates that for these values of λ, many of the genes selected by the lasso are likely to be truly related to survival, as it would be unlikely for so many noise features to be selected merely by random chance. On the right, these expected number of false discoveries is plotted as a fraction of the selected features (i.e., the mFDR).

Table 2 displays results for three potential approaches to choosing λ: the value that minimizes cross-validation error (CV), the value that comes within 1 standard error (SE) of minimizing CV (1se), and the smallest value of λ satisfying mFDR < 10%. From the table, we can see that there appears to be an inherent tradeoff between the optimal prediction and restricting a model to containing few noise features. In this case, we see that the lasso model with lowest prediction error has an estimated mFDR of 84.6%, indicating that despite its prediction
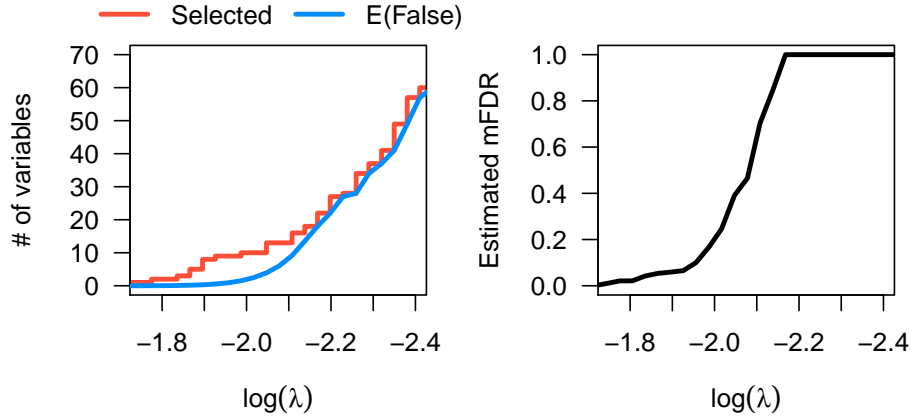
Figure 5: Marginal false discovery estimates for the Shedden data. (left) Total number of genes selected and expected number of marginal false discoveries; (right) Estimated mFDR.

Table 2: Results for different choices of λ in applying the lasso and MCP penalties to the Shedden and Spira data. mFDR = Marginal false discovery rate (%). CVE = Cross-validation error. MCE = Misclassification error (%).

		Shedden case study					Spira case study					
		λ	EF	S	mFDR	CVE	λ	EF	S	mFDR	CVE	MCE
lasso	CV	0.118	13.53	16	84.6	11.05	0.029	71	71	100	1.00	24.5
	CV(1se)	0.180	0	0	-	11.17	0.063	32	32	100	1.07	25.0
	mFDR	0.146	0.52	8	6.6	11.10	0.146	0.78	10	7.8	1.29	30.7
MCP	CV	0.155	0.16	8	8.2	11.21	0.067	14	14	100	1.13	29.7
	CV(1se)	0.180	0	0	-	11.22	0.093	10	10	100	1.19	31.8
	mFDR	0.155	0.16	2	8.2	11.21	0.151	0.42	5	8.4	1.29	33.9

accuracy, a substantial fraction of the 16 features selected by the model may be noise. Conversely, if we aim to control the mFDR at 10%, we select only 8 features. Although we can be more confident that these 8 features are truly related to survival, the model is somewhat less accurate from a prediction standpoint (CV error of 1313 compared to 1308). The CV (1se) choice λ aims to select a parsimonious model whose accuracy is comparable with the best model (**?**). However, in this example the CV (1se) method selects the null model, giving the impression that no genes can be reliably selected. Note that the mFDR < 10% approach is also within 1 SE of the best model, and therefore may also be thought of as striking an appropriate balance between prediction and parsimony in this example.

An alternative to lasso-penalized regression is to use the MCP penalty, which relaxes the degree of shrinkage for large coefficients, thereby allowing a smaller number of features to account for the observed signal. Our approach indicates that the MCP model minimizing CV error is likely to have an mFDR no higher than 8.2%; compared to the lasso model that minimized CV error, we can be considerably more confident in the smaller number of features selected by MCP.

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

For comparison we also analyzed the data using repeated sample splitting with 100 random splits, as advocated

by **?**. However, this approach was unable to select any genes at a false discovery rate of 10%, or even at a more liberal rate of 50%. As mentioned earlier, the `covTest` package does not (currently) accommodate survival data.

We also applied a large scale univariate testing approach to the Shedden data, fitting a series of Cox regression models adjusting for all clinical covariates and containing a single gene. After using the Benjamini-Hochberg procedure to control the false discovery rate at 10%, this approaches selects 803 genes, far more than any of the regression-based approaches. As discussed in Section 3.3, the primary difference between univariate and regression-based methods with respect to false discoveries is that univariate approaches tend to select large numbers of features that are indirectly correlated with the outcome. Our simulation results would suggest that for these data, univariate testing yields a large number of genes, most of which are only indirectly associated with survival, while the mFDR approach yields a smaller number of genes, most of which are directly associated with survival.

4.2 Lung cancer status among smokers

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

Table 2 shows results for the Spira data at various λ values for the lasso and MCP penalty functions. In contrast with the Shedden case study, for both penalties in this example the mFDR bound is 100% at the value of λ selected by cross validation. In other words, we are unable to provide an upper bound for the mFDR at the value of λ, and one should be cautious about interpreting the selected features as significant. For example, at $\lambda = 0.029$, there may very well be some true discoveries among the 71 selected features, but a lasso-penalized logistic regression model could easily select 71 features just by chance at that value of λ even if none of them were related to the outcome.

In order to arrive at a set of features with a low false discovery rate, we much choose a much smaller model (for the lasso model, 10 features instead of 71). As in the earlier case study, however, the price of an FDR restriction is a decrease in prediction accuracy. The prediction error is considerably higher in the mFDR $< 10\%$ lasso and MCP models than for the models selected by cross-validation. In the Spira example, no genes have particularly large signals; as a result, the MCP approach is less successful than at finding a model that is attractive from both the mFDR and prediction perspectives than it was in the Shedden example.

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

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

5 Discussion

Controlling the marginal false discovery rate of a penalized likelihood model is a useful way of assessing the reliability of a selected set of features. Unlike other approaches that have a similar goal, such as sample splitting or the covariance test, the mFDR approach uses a less strict definition of a false discovery which does not require conditioning on any other variables and consequently only limits the selection of variables that are noise in an unconditional – i.e., marginal – sense. The simulations in this paper demonstrate that while mFDR is based upon

a weaker false discovery definition, when used in combination with a penalized regression model it tends to perform very well at limiting the number of indirect (non-causal) feature selections.

Furthermore, mFDR is far more convenient from the standpoint of computational burden than other approaches. For the Spira data, where $n = 192$ and $p = 22,215$, fitting a penalized logistic regression model and then estimating the mFDR for the entire λ sequence takes only 1.4 seconds. This is over 400 times faster than sample splitting and the covariance test, each of which took over 9 minutes, and is nearly 30 times faster than even univariate testing, which took over 3 minutes. Results are similar for the Shedden survival analysis. The computational efficiency of mFDR makes it a particularly appealing tool during the early stages of an analysis when a number of candidate models are being considered.

A natural question is whether marginal independence is a realistic assumption. In many applications, complete independence is unlikely to be literally true. However this is, in a sense, the point: if one selects 20 features and one would expect to select 16 of them even if all of those features were completely independent of the outcome, this calls into question the confidence with which we can claim associations between any given feature and the outcome. A relevant analogy is with the testing of a point null hypothesis – an effect of precisely zero may be unrealistic, if that hypothesis cannot be rejected, this is cause for skepticism.

The mFDR control procedure proposed here is currently implemented in the R package `ncvreg` (**?**) using the `mFDR` function. The function accepts an ncvreg fitted model and calculates the expected number of false discoveries, as well as the mFDR for each value of the λ sequence used in fitting the model. The package also provides a plotting method, which produces plots like those in Figure 5. The `ncvreg` package accommodates lasso, MCP, SCAD, elastic net, and MNet (**?**) penalties and provides a convenient model summary measure for penalized linear, logistic, and Cox regression models.

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