Marginal false discovery rates for penalized regression models

PATRICK J. BREHENY[*]

Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

patrick-breheny@uiowa.edu

SUMMARY

Penalized regression methods are an attractive tool for high-dimensional data analysis, but their widespread adoption has been hampered by the difficulty of applying inferential tools. In particular, the question "How reliable is the selection of those features?" has proved difficult to address. In part, this difficulty arises from defining false discoveries in the classical, fully conditional sense, which is possible in low dimensions but does not scale well to high-dimensional settings. Here, we consider the analysis of marginal false discovery rates (mFDRs) for penalized regression methods. Restricting attention to the mFDR permits straightforward estimation of the number of selections that would likely have occurred by chance alone, and therefore provides a useful summary of selection reliability. Theoretical analysis and simulation studies demonstrate that this approach is quite accurate when the correlation among predictors is mild, and only slightly conservative when the correlation is stronger. Finally, the practical utility of the proposed method and its considerable advantages over other approaches are illustrated using gene expression data from The Cancer Genome Atlas and genome-wide association study data from the Myocardial Applied Genomics Network.

Keywords: Penalized regression; False discovery rate; Lasso; High-dimensional data analysis.

1. INTRODUCTION

Penalized regression is an attractive methodology for dealing with high-dimensional data where classical likelihood approaches to modeling break down. However, its widespread adoption has been hindered by a lack of inferential tools. In particular, penalized regression is very useful for variable selection, but how confident one may be about those selections has proven difficult to quantify. As we will see, this difficulty is partially due to the complexity of defining a "false discovery" in the penalized regression setting. In this article, I will focus mainly on the lasso for linear regression models (Tibshirani, 1996), although the idea is very general and can be extended to a variety of other regression models and penalty functions. In particular, suppose we use the lasso to select variables from a pool of potentially important features. This article addresses the question, "How many of those selections would likely have occurred by chance alone?"

There has been a fair amount of recent work on the idea of hypothesis testing in the high-dimensional penalized regression setting. A comprehensive review is beyond the scope of this article, but it is worth

---

[*]To whom correspondence should be addressed.

introducing two approaches to which I will compare the method proposed in this article. One approach is to split the sample into two parts, using the first part for variable selection and the second part for hypothesis testing. This idea was first introduced in Wasserman and Roeder (2009), who studied the problem using a single split of the dataset. Meinshausen *and others* (2009) extended this approach by considering multiple random splits and combining the results. Dezeure *and others* (2015) provide a comprehensive review of this approach and the details involved, along with procedures for limiting the overall false discovery rate through this form of testing.

An alternative approach is to test the significance of adding a variable along the solution path as the degree of penalization is relaxed, conditional on the other variables already included in the model. Several tests fall into this category, including the covariance test (Lockhart *and others*, 2014), the spacing test, and an exact test (Tibshirani *and others*, 2016, which also introduces the spacing test). The details of the tests differ, but all involve modifying classical tests for the significance of an added variable by conditioning on the fact that the added variable was not prespecified, but rather selected from a pool of potential variables. G'Sell *and others* (2016) provide an important extension to this work by deriving a rule, forwardStop, for selecting the stopping point along this sequence of sequential tests to preserve a specified false discovery rate.

False discoveries are straightforward to define in single-variable hypothesis testing: a false discovery is one that is independent of the outcome. In regression models, however, this idea is complicated by various kinds of conditional independence. Typically, in regression a feature X_j is considered a false discovery if it is independent of the outcome Y given all other features; symbolically, if $X_j \perp\!\!\!\perp Y | \{X_k\}_{k \neq j}$. This is the perspective adopted by most work in this area, including sample splitting; here, we will refer to this as the *fully conditional* perspective. The pathwise approaches use a different definition, which we will call the *pathwise conditional* perspective. Letting $\mathcal{M}_j$ denote the set of variables with non-zero coefficients in the model at the point in the path where feature j is selected, in these approaches a feature j is considered a false discovery if $X_j \perp\!\!\!\perp Y | \{X_k : k \in \mathcal{M}_j\}$.

This article introduces a weaker definition of false discovery than those considered in the existing literature. Rather than the conditional definitions given in the previous paragraph, here we consider a marginal perspective in which a selected feature j is false if it is marginally independent of the outcome, the definition used in single-feature testing: $X_j \perp\!\!\!\perp Y$. Adopting a simpler definition makes it possible to estimate the expected number of false discoveries as well as their rate, which we call the marginal false discovery rate (mFDR). Our goal here is not to argue that the mFDR is always superior to either of the conditional FDRs; clearly, there are times where a conditional FDR is an important quantity of interest. However, the mFDR is also an interesting and useful summary of feature selection accuracy. This is especially true in high dimensions where attempting to control a conditional false discovery rate is often hopeless.

2. Marginal false discovery rates

Consider the linear model with normally distributed errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \qquad \varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \text{N}(0, \sigma^2),$$

where $\mathbf{y}$ denotes the response and $\mathbf{X}$ the $n \times p$ design matrix, with n denoting the number of independent observations and p the number of features. Without loss of generality, we assume throughout that the response and covariates are centered so that the intercept term can be ignored, and that the features are standardized so that $\frac{1}{n}\sum_i x_{ij}^2 = 1$ for all j. We are interested in studying the lasso estimator $\widehat{\boldsymbol{\beta}}$, defined as the quantity that minimizes

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.1}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$. Here, the least-squares loss measures the fit of the model and the L_1 norm is used to penalize large values of $\boldsymbol{\beta}$, with λ controlling the tradeoff between the two.

An appealing aspect of using the lasso to estimate $\boldsymbol{\beta}$ is that the resulting estimates are sparse: some coefficients will be non-zero, but for many, $\widehat{\beta}_j = 0$. In this article, we say that a feature for which $\widehat{\beta}_j \neq 0$ has been "selected." Note that $\widehat{\boldsymbol{\beta}}$ changes with λ although we will suppress this in the notation; for a large enough value of λ, $\widehat{\beta}_j = 0$ for all j, but as we lower λ, more variables will be included. To address the expected number of features included in a lasso model by chance alone, we begin by considering the orthonormal case, then turn our attention to the general case.

2.1. *Orthonormal case*

For a given value of the regularization parameter λ, let $\mathbf{r} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ denote the residuals. The lasso (2.1) is a convex optimization problem, so the Karush–Kuhn–Tucker (KKT) conditions are both necessary and sufficient for any solution $\widehat{\boldsymbol{\beta}}$:

$$\frac{1}{n}\mathbf{x}_j'\mathbf{r} = \lambda \, \mathrm{sign}(\widehat{\beta}_j) \qquad \text{for all } \widehat{\beta}_j \neq 0$$

$$\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}\right| \leq \lambda \qquad \text{for all } \widehat{\beta}_j = 0.$$

Letting $\mathbf{X}_{-j}$ and $\boldsymbol{\beta}_{-j}$ denote the portions of the design matrix and coefficient vector that remain after removing the jth feature, let $\mathbf{r}_j = \mathbf{y} - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}$ denote the partial residuals with respect to feature j. The KKT conditions thus imply that

$$\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| > \lambda \qquad \text{for all } \widehat{\beta}_j \neq 0$$

$$\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| \leq \lambda \qquad \text{for all } \widehat{\beta}_j = 0 \tag{2.2}$$

and therefore the probability that variable j is selected is

$$\mathbb{P}(\widehat{\beta}_j \neq 0) = \mathbb{P}\left(\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| > \lambda\right).$$

This indicates that if we are able to characterize the distribution of $\frac{1}{n}\mathbf{x}_j'\mathbf{r}_j$ under the null, we can estimate the number of false discoveries in the model. Indeed, this is straightforward in the case of orthonormal design ($\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{I}$):

$$\begin{aligned}
\frac{1}{n}\mathbf{x}_j'\mathbf{r}_j &= \frac{1}{n}\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_{-j}\boldsymbol{\beta}_{-j} + \mathbf{x}_j\beta_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}) \\
&= \frac{1}{n}\mathbf{x}_j'\boldsymbol{\varepsilon} + \beta_j \\
&\sim N(\beta_j, \sigma^2/n).
\end{aligned} \tag{2.3}$$

Thus, if $\beta_j = 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| > \lambda\right) = 2\Phi(-\lambda\sqrt{n}/\sigma).$$

These results are related to the expected number of false discoveries in the following theorem, the proof of which follows directly from the above by summing $\mathbb{P}(\widehat{\beta}_j \neq 0)$ over the set of null variables.

THEOREM 2.1 Suppose $\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Then, letting $\mathcal{S} = \{j : \widehat{\beta}_j \neq 0\}$ denote the set of selected features and $\mathcal{N} = \{j : \beta_j = 0\}$ the set of null features, for any value of λ, we have

$$\mathbb{E}\,|\mathcal{S} \cap \mathcal{N}| = 2\,|\mathcal{N}|\,\Phi(-\lambda\sqrt{n}/\sigma).$$

To use this as an estimate, the unknown quantities $|\mathcal{N}|$ and σ^2 must be estimated. One straightforward possibility is to replace $|\mathcal{N}|$ by p and estimate the variance σ^2 by

$$\hat{\sigma}^2 = \frac{\mathbf{r}'\mathbf{r}}{n - |\mathcal{S}|}.$$

Dividing the residual sum of squares by the degrees of freedom of the lasso (Zou *and others*, 2007) is a simple approach for estimating the residual variance, but other possibilities exist (e.g., Fan *and others*, 2012). This implies the following estimate for the expected number of false discoveries:

$$\widehat{\text{FD}} = 2p\Phi(-\sqrt{n}\lambda/\hat{\sigma}) \tag{2.4}$$

and, as an estimate of the false discovery rate:

$$\widehat{\text{FDR}} = \frac{\widehat{\text{FD}}}{|\mathcal{S}|}; \tag{2.5}$$

note that this estimate will be somewhat conservative, as we are replacing $|\mathcal{N}|$ by its upper bound, p, rather than attempting to estimate it.

This straightforward estimate of the false discovery rate is made possible by the orthonormality condition assumed in (2.3). When the features that comprise $\mathbf{X}$ are correlated, however, the distribution of $\frac{1}{n}\mathbf{x}_j'\mathbf{r}_j$ is considerably more complex, as is the proper definition of what is meant by a "false discovery." Nevertheless, as the rest of this article will show, with a suitable definition of false discovery, estimator (2.5) can still be used to estimate false discovery rates for variable selection in non-orthonormal settings.

2.2. *Definition in the general case*

Consider the causal diagram presented below. In this situation, variable A could never be considered a false discovery: it has a direct causal relationship with the outcome Y. Likewise, if variable N were selected, this would obviously count as a false discovery – N has no relationship, direct or indirect, to the outcome.
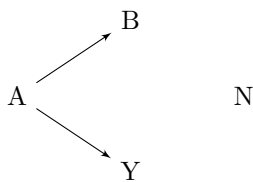

Variable B, however, occupies a gray area as far as false discoveries are concerned. From a marginal perspective, B would not be considered a false discovery, since it is not independent of Y. However, B

and Y are conditionally independent given A, so from a fully conditional regression perspective, B is a false discovery. Lastly, from a pathwise conditional perspective, B may or may not be a false discovery, depending on whether A is already included in the model or not at the point at which B enters.

The central argument of this article is that estimating the number of false selections arising from variables like B is inherently complicated and requires complex approaches; however, simple approaches like the one derived in Section 2.1 may still be used to estimate the number of false selections arising from variables like N.

Here, I define a *noise feature* to be a variable like N, that has no causal path (direct or indirect) between it and the outcome, and the mFDR as the proportion of selected features that are noise features. Again, this is consistent with how false discoveries are defined in univariate testing, but differs from the existing literature in regression modeling.

It is worth addressing a technical point here: it is possible for two variables to be conditionally dependent despite being marginally independent. Such a relationship would appear in a directed acyclic graph as the following:

$$X_1 \searrow \qquad Y \swarrow$$
$$X_2$$
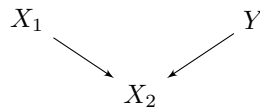
As the figure indicates, such a relationship would require a feature (or features) to be caused by the outcome Y. This is contrary to the usual framework in which regression models are applied, where Y is (or is assumed to be) a consequence of the features used for prediction (e.g., a genetic variant can lead to heart disease, but heart disease will not change an individual's genetic sequence). Thus, although we are actually making a stronger assumption than mere marginal independence here, in the typical regression setting the two are equivalent. At any rate, although the term may be slightly imprecise from a technical perspective, we feel that the name "marginal false discovery" is easily understood and best distinguishes this quantity from false discoveries under the various conditional perspectives.

The proposed definition has several advantages. First, when two variables (like A and B in the first diagram) are correlated, it is often difficult to distinguish between which of them is driving changes in Y and which is merely correlated with Y. As we will see, this causes methods using a conditional perspective to be conservative, especially in high dimensions.

Second, in many scientific applications, discovering variables like B is not particularly problematic. For example, two genetic variants in close proximity to each other on a chromosome will be highly correlated. Although it is obviously desirable to identify which of the two is the causal variant, locating a nearby variant is also scientifically valuable, as it narrows the search to a small region of the genome for future follow-up studies.

The final advantage is clarity of interpretation. Whether or not a feature would qualify as a marginal false discovery depends only on the relationship between it and the outcome, not on whether another feature has been included in the model. In contrast, interpreting the results from pathway-based tests such as those in Lockhart *and others* (2014) and Tibshirani *and others* (2016) can be challenging, as the definition of the null hypothesis is constantly changing with λ.

3. Independent noise features

The orthonormal conditions of Section 2.1 clearly do not hold in most settings for which penalized regression is typically used. Thankfully, they can be relaxed in two important ways that make the results more widely applicable. First, the predictors do not have to be strictly orthogonal in order for the estimator to work; they can simply be uncorrelated. Second, this condition of being uncorrelated applies only to the

noise features—i.e., the variables like N in the diagram from Section 2.1; variables like A and B can have any correlation structure. These statements are justified theoretically in Section 3.1 and via simulation in 3.2.

3.1. *Theory*

To make these statements concrete, let $\mathcal{A}, \mathcal{N}$ partition $\{1, 2, \ldots, p\}$ such that $\beta_j = 0$ for all $j \in \mathcal{N}$ and the following condition holds:

$$\lim_{n \to \infty} \frac{1}{n} \mathbf{X}'\mathbf{X} = \left[\begin{array}{cc} \Sigma_{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\mathcal{N}} \end{array} \right].$$

The opening remarks of this section can now be stated precisely in the following theorem, the proof of which appears in the Appendix.

Theorem 3.1 Suppose $\Sigma_{\mathcal{N}} = \mathbf{I}$. Then for any $j \in \mathcal{N}$ and for λ_n such that the sequence $\sqrt{n}\lambda_n$ is bounded,

$$\frac{1}{\sqrt{n}} \mathbf{x}'_j \mathbf{r}_j \xrightarrow{\text{d}} N(0, \sigma^2).$$

Theorem 3.1 shows that if the noise features are uncorrelated, then in the limit $\frac{1}{n} \mathbf{x}'_j \mathbf{r}_j$ behaves just as it did in Section 2.1 (equation 2.3). Thus, estimators (2.4) and (2.5) still provide approximate estimators for the mFDR in this setting, at an accuracy that improves with the sample size. The technical condition requiring $\sqrt{n}\lambda_n$ to be bounded is only necessary so that the estimate $\widehat{\boldsymbol{\beta}}$ will be $\sqrt{n}$-consistent. Without it (i.e., for large values of λ), $\frac{1}{n} \mathbf{x}'_j \mathbf{r}_j$ will converge to a random variable with a variance larger than σ^2 due to underfitting. Note that Theorem 3.1 treats p as fixed; extending this result to allow $p \to \infty$ would be of interest as future work, but lies outside the scope of this article.

3.2. *Simulation*

To illustrate the consequences of Theorem 3.1, let us carry out the following simulation study, with both a "low-dimensional" ($n > p$) and "high-dimensional" ($n < p$) component. As in Section 2.1, three types of features will be included:

- Causative: Six variables with $\beta_j = 1$

- Correlated: Each causative feature is correlated ($\rho = 0.5$) with m other features; $m = 2$ for the low-dimensional case and $m = 9$ for the high-dimensional case

- Noise: Independent noise features are added to bring the total number of variables up to 60 in the low-dimensional case and 600 in the high-dimensional case

The causative, correlated, and noise features correspond to variables A, B, and N, respectively, in the diagram from Section 2.1. In each setting, the sample size was $n = 100$, while the total number of causative/correlated/noise features was 6/12/42 for the low-dimensional setting and 6/54/540 for the high-dimensional setting. The results of the simulation are shown in Figure 1.

As Theorem 3.1 implies, estimators (2.4) and (2.5) are quite accurate, on average, when the noise features are independent. The estimated number of noise features selected and the mFDR are both somewhat conservative, as we would expect from using p as an upper bound for the number of noise features
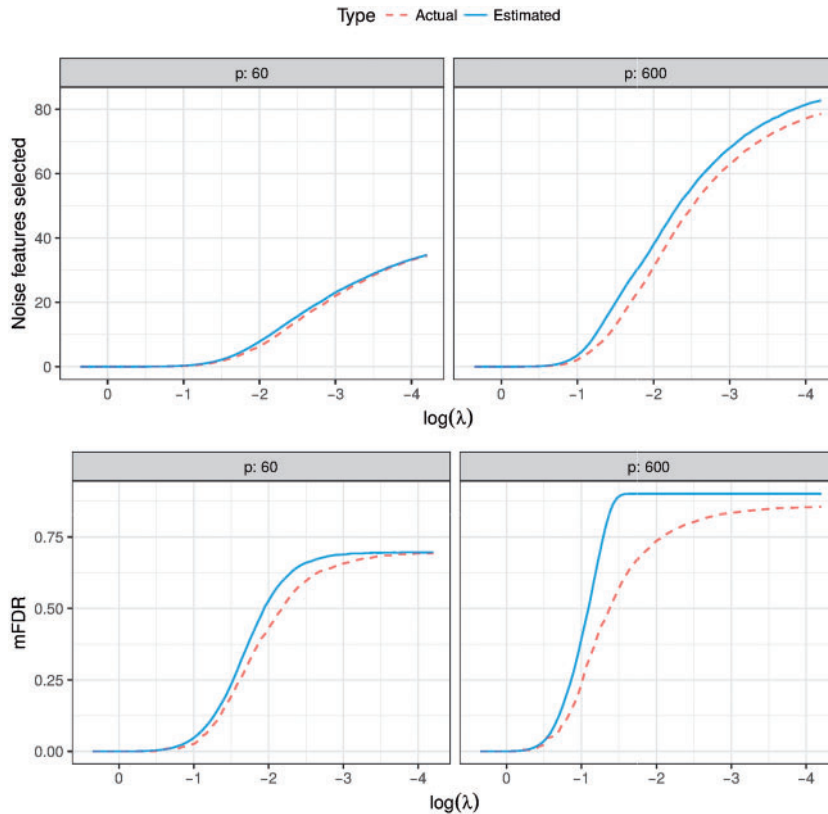
Fig. 1. Accuracy of estimators (2.4) and (2.5) in the case of independent noise features.

(e.g., in the high-dimensional case, $p = 600$ but $|\mathcal{N}| = 540$). However, the effect is slight in this setting. For example, in the high-dimensional case at $\lambda = 0.55$, the actual mFDR was 5%, while the estimated rate was 6.5%.

Being able to estimate the mFDR means we can use it to select the regularization parameter λ. For example, we could choose λ to be the smallest value of λ such that $\widehat{\text{mFDR}}(\lambda) < 0.1$. Figure 2 compares this approach with several other methods for selecting λ in terms of the number of each type of feature the method selects on average. For Lasso (mFDR), univariate testing (i.e., marginal regression), sample splitting (using the hdi package, Dezeure *and others*, 2015), and the exact conditional path-based test (using larInf and forwardStop from the selectiveInference package, Tibshirani *and others*, 2016; G'Sell *and others*, 2016), the nominal false discovery rates were set to 10%. For cross-validation (CV), the value of λ minimizing the CV error was selected.

It is worth noting that both Lasso (mFDR) and univariate testing limit the fraction of selections due to noise features (to 5% and 7%, respectively, in the $p = 60$ simulation) to the nominal rate, but claim nothing about the fraction arising from correlated features. This is obvious from the figure for univariate testing; for Lasso (mFDR), the fraction of selected features coming from either the Correlated or Noise groups (i.e., all features with $\beta_j = 0$) was 17% in the low-dimensional setting and 23% in the high-dimensional setting.

Nevertheless, compared to univariate testing, the Lasso (mFDR) approach has two major advantages, despite using the same definition of a false discovery. First, although unable to precisely quantify the
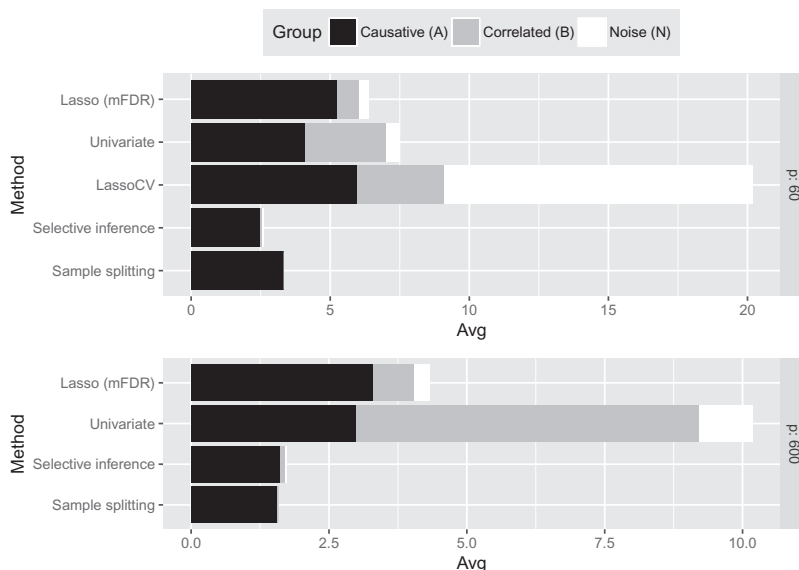
Fig. 2. Average number of each type of feature selected by various methods for the simulation setup in Section 3.2. Cross-validation is omitted from the high-dimensional ($p = 600$) plot because of the very large number of noise features it selects, which dominates the plot when included.

number of indirect associations that have been selected, by regressing on features directly related to the outcome, the lasso greatly reduces the number of these associations that are identified as important. For example, in the $p = 600$ case, 82% of the features identified by lasso were causally related to the outcome, compared with just 32% for univariate testing. Second, by successfully explaining variation in the outcome and thereby reducing noise, the lasso has greater power. For example, in the $p = 60$ case, the lasso is able to identify 5.3 causative features on average, compared with 4.1 for univariate testing.

Among the penalized regression approaches, CV is a considerable outlier, with no protection against the selection of large numbers of noise features. In the high-dimensional case, CV selected over 30 noise features on average, and is omitted from the plot so as not to obscure the performance of the other methods. Of course, the goals of CV are very different: to obtain the model with the best predictive accuracy, regardless of the number of noise features it may contain.

The sample splitting and selective inference approaches, on the other hand, greatly limit the number of both noise features and correlated features selected by the lasso model. The more restrictive definition of a false discovery used by these approaches, however, causes them to behave quite conservatively, and have considerably less power to detect the causative features than any of the other methods considered.

4. CORRELATED NOISE FEATURES

The simulation results of Section 3.2 are something of a best-case scenario for the proposed method, since the variables in $\mathcal{N}$ were independent and Theorem 3.1 establishes that the estimator is consistent in this case. As we will see, when noise features are correlated, estimator 2.5 becomes conservative. The case of correlated noise features is significantly less mathematically tractable, making it harder to construct a rigorous proof of this claim, but here we offer some insight into why this is true and investigate the performance of the proposed method via simulation.
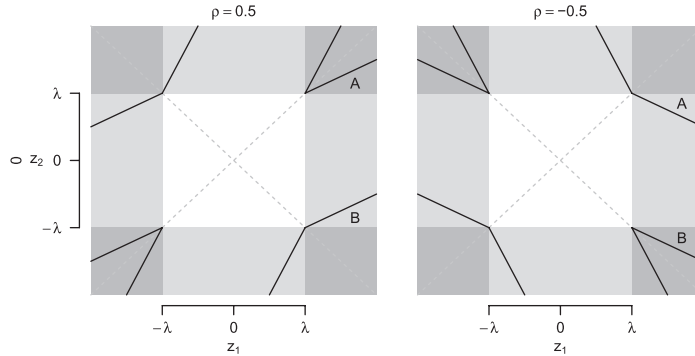
Fig. 3. Illustration of how correlation affects selection in the bivariate case. Features $\mathbf{x}_1$ and $\mathbf{x}_2$ are positively correlated ($\rho = 0.5$) on the left and negatively correlated ($\rho = -0.5$) on the right.

To illustrate why the proposed mFDR estimator becomes conservative when noise features are correlated, let us consider the $p = 2$ case. Letting $z_j = \frac{1}{n}\mathbf{x}'_j\mathbf{y}$, we begin by noting that $\mathbf{z}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and

$$\text{Var}(\mathbf{z}) = \frac{\sigma^2}{n}\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where ρ denotes the correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$, and

$$2\,|\mathcal{N}|\,\Phi(-\lambda\sqrt{n}/\sigma) = \mathbb{P}(|z_1| > \lambda) + \mathbb{P}(|z_2| > \lambda). \tag{4.6}$$

A visual representation of this quantity is given in Figure 3.

The shading represents whether 0 (white), 1 (light gray), or 2 (dark gray) features will be selected under this orthogonal approximation; the quantity in (4.6) can be found by integrating these constant regions with respect to the joint density of $\mathbf{z}$.

In the case where $\mathbf{x}_1$ and $\mathbf{x}_2$ are correlated, the white region (where no features are selected) remains the same, but the conditions under which both features are selected (the light and dark gray regions) change. By working through the KKT conditions for the bivariate case under various conditions, we can determine these selection boundaries. For example, the boundary for $\widehat{\beta}_2 > 0$ given that $\widehat{\beta}_1 > 0$ (i.e., given that $z_1 > \lambda$) is

$$\frac{1}{n}\left|\mathbf{x}'_2\mathbf{r}_2\right| > \lambda \implies z_2 - \rho\widehat{\beta}_1 > \lambda$$
$$\implies z_2 > \rho z_1 + \lambda(1 - \rho).$$

These boundaries are drawn in black on Figure 3; the preceding equation is the line just above the letter "A". As the figure illustrates, when $\rho > 0$ the region over which $\beta_1, \beta_2 \neq 0$ narrows in the upper right and lower left quadrants, and widens in the other two quadrants. Considering the difference $\mathbb{E}\,|\mathcal{S} \cap \mathcal{N}| - \{\mathbb{P}(|z_1| > \lambda) + \mathbb{P}(|z_2| > \lambda)\}$, we see that in many regions, the terms cancel out. The differences lie in four pairs of triangular regions; one such pair is labeled "A" and "B". The region A is where two variables are selected under orthogonality, but only one in the correlated case, while region B is where one variable is selected under orthogonality, but two are selected in the correlated case. However, because z_1 and z_2 are positively correlated in this scenario, the density at each point in A is higher than its corresponding point in B, and therefore estimator (2.4) is conservative in the sense that it overestimates the expected number of false
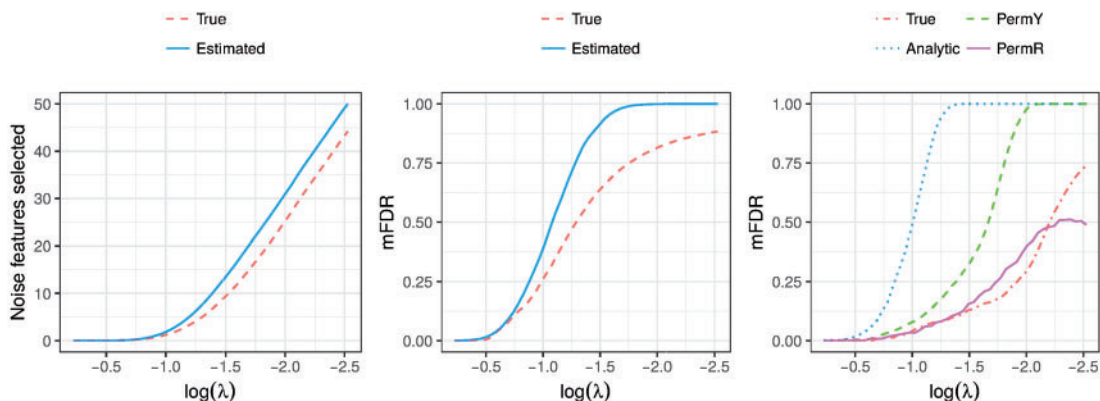
Fig. 4. Accuracy of the analytic (2.5), permute-the-outcome ("PermY"), and permute-the-residuals ("PermR") estimators in the case of correlated noise features. Left, Center: Autoregressive correlation. Right: Exchangeable correlation.

discoveries. The opposite scenario happens for $\rho < 0$, where region B now has the higher probability density, again leading to a larger value under orthogonality, with equality between the two quantities occurring only at $\rho = 0$.

In terms of Theorem 3.1, this argument implies that when $\Sigma_{\mathcal{N}} \neq \mathbf{I}$, the quantity $\frac{1}{\sqrt{n}}\mathbf{x}_j'\mathbf{r}_j$ converges to a distribution with thinner tails than $N(0, \sigma^2)$. Intuitively, this makes sense: if two noise features are correlated, a regression-based method such as the lasso will tend to select a single feature rather than both. Thus, the uncorrelated case is not just mathematically convenient, it also represents a worst-case scenario with respect to the number of noise features that we can expect to be selected by chance alone.

To investigate the robustness of the proposed mFDR estimator in the presence of moderate correlation, let us carry out the following simulation. The generating model contains six independent causative features and 494 correlated noise features, with a 1:1 signal-to-noise ratio ($n = 100$, $p = 500$, $R^2 = 0.5$). The noise features are given an autoregressive correlation structure with $\text{Cor}(\mathbf{x}_j, \mathbf{x}_k) = 0.8^{|j-k|}$. The results of the simulation are shown in Figure 4.

Compared with Figure 1, the mFDR estimates are somewhat more conservative in this case, although still quite accurate—certainly accurate enough to be useful in practice. For example, at $\lambda = 0.43$, the true inclusion rate for noise variables was 14%, while the estimated rate according to (2.4) was 20%.

This simulation illustrates that although its derivation is based on independent noise features, the proposed mFDR estimator is reasonably robust to the presence of correlation. Furthermore, it provides a conservative estimate of the mFDR, suggesting that the approach provides control over the mFDR, at least on average.

5. PERMUTATION APPROACH

As correlation between noise features increases, the proposed estimator becomes more conservative. As an extreme example, suppose that the correlation structure from Section 4 was exchangeable instead of autoregressive: $\text{Cor}(\mathbf{x}_j, \mathbf{x}_k) = 0.8$ for all j, k. The results of this modification to the simulation (all other aspects remaining the same) are shown in the right panel of Figure 4.

As one might expect, the estimates are far more conservative in this case. For example, at $\lambda = 0.43$, the estimated mFDR is 23% even though the true mFDR is only 1%. This is a consequence of the independence approximation, where (2.5) operates under the simplification that the selection of one noise feature does

not affect the probability of other noise features being selected. This is reasonably accurate in many cases, but not in this situation, where all noise features are highly correlated with each other. As a result, the lasso tends to select only a single noise feature from this highly correlated set, while the independence-based estimate (2.4) indicates that it has likely selected, say, seven or eight.

This phenomenon is not unique to penalized regression; substantial correlation among features causes problems with conventional false discovery rates as well (Efron, 2007). One widely used approach for controlling error rates while preserving correlation structures is to use a permutation approach (Westfall and Young, 1993), and a similar strategy may be applied in order to calculate mFDRs for penalized regression as well. The primary advantage of this approach is that it reduces the conservatism of the analytic approach developed in Section 2, while the primary disadvantage is a greatly increased computational burden. In this section, I describe two permutation-based methods and apply them to the simulation shown in Figure 4.

5.1. *Permuting the outcome*

The simplest approach is simply to randomly permute the outcome $\mathbf{y}$, creating new outcomes $\tilde{\mathbf{y}}^{(b)}$ for $b = 1, 2, \ldots, B$. Then, for each permutation b, solve for the lasso path $\tilde{\boldsymbol{\beta}}^{(b)}(\lambda; \mathbf{X}, \tilde{\mathbf{y}}^{(b)})$, estimate the average number of noise features included in the model for a given value of λ

$$\widehat{\text{FD}}(\lambda) = \frac{\sum_b \#\{\tilde{\beta}_j^{(b)}(\lambda) \neq 0\}}{B},$$

and the mFDR by $\widehat{\text{FD}}(\lambda)/S(\lambda)$, where $S(\lambda)$ denotes the number of variables selected by the lasso using the original (i.e., not permuted) data.

This method is applied to the case of extreme correlation in Figure 4. The method is considerably less conservative than the analytic approach, although still conservative for reasons that will be discussed in Section 5.2. For example, at $\lambda = 0.43$, the average estimated mFDR is 4%, where the true value was 1% and the analytic approach yielded 23%.

By permuting Y, this approach constructs realizations of the data in which all features belong to $\mathcal{N}$, the set of noise features. Like (2.5), it limits the number of noise features selected but cannot be used to control the number of false discoveries in the fully conditional sense. In the hypothesis testing literature, this is referred to as "weak control" over the error rate.

5.2. *Permuting the residuals*

As seen in Figure 4, permuting the outcomes is still conservative in its estimation of mFDR. Partitioning the variance of Y into signal and noise, ideally we would permute only the noise, but by permuting the outcome we permute the signal as well. This has the effect of overestimating the noise present in the model. One alternative is, rather than permuting the outcome, to permute the residuals, $\mathbf{r}(\lambda) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)$, of the original lasso fit. The method is otherwise identical to that described in Section 5.1, with the notable exception that the residuals, unlike the outcomes, depend on λ and thus, B separate lasso solutions $\tilde{\boldsymbol{\beta}}^{(b)}(\lambda; \mathbf{X}, \tilde{\mathbf{r}}^{(b)})$ must be calculated at each value of λ (in Section 5.1, the same solutions could be used for all values of λ), substantially increasing the computational burden.

Nevertheless, this increased computational cost does offer a benefit, as seen in Figure 4. Unlike the analytic and permute-the-outcome approaches, permuting the residuals provides an essentially unbiased estimate of the true false discovery rate except at very small λ. For example, at $\lambda = 0.29$, the average estimated mFDR is 8%, as was the true mFDR, whereas the permute-the-outcome and analytic methods estimated mFDRs of 16% and 90%, respectively.

The purpose of this manuscript is primarily to investigate the analytic approach to calculating false discovery rates outlined in Section 2, with relatively less emphasis on the permutation approaches. In the opinion of the author, the analytic approach is almost always useful, since it can be calculated instantly. Whether one wishes to take the time to carry out the permutation approach, however, depends on context: how correlated the features are, how problematic a somewhat conservative estimate of mFDR is, and how far along in the analytic process one is (initial exploration or final publication results). For a more detailed comparison of the analytic and permutation approaches in the context of genetic association studies, see Yi *and others* (2015).

6. CASE STUDIES

6.1. *Breast cancer gene expression study*

As a case study in applying the proposed method to real data, we will analyze data on gene expression in breast cancer patients from The Cancer Genome Atlas (TCGA) project, available at http://cancergenome.nih.gov. In this dataset, expression measurements of 17 814 genes, including BRCA1, from 536 patients are recorded on the log scale. BRCA1 is a well studied tumor suppressor gene with a strong relationship to breast cancer risk. Because BRCA1 is likely to interact with many other genes, including other tumor suppressors and regulators of the cell division cycle, it is of interest to find genes with expression levels related to that of BRCA1.

For this analysis, 491 genes with missing data were excluded, resulting in a design matrix with $p = 17\,322$ predictors. The resulting mFDR estimates for the lasso solution path are presented in Figure 5. Here, the mFDR estimates indicate that many genes are predictive of BRCA1 expression: we can safely select 55 variables before the false discovery rate exceeds 10%. This makes sense scientifically, as a large number of genes are known to affect BRCA1 expression through a variety of mechanisms, and the sample size here is sufficient that we should be able to identify many of them. For the sake of comparison, univariate testing of each feature separately identifies 7903 genes that have a significant correlation with BRCA1 at a false discovery rate of 10%.

In contrast, the selective inference and sample-splitting procedures each select just a single feature. Although the mFDR estimates are slightly conservative as discussed in Section 4, this matters far less in practice than whether one adopts a marginal or conditional perspective with respect to false discoveries. This example clearly illustrates how conservative the conditional perspective is in practice, especially with high-dimensional data.

The mFDR approach is also vastly more convenient than sampling splitting or selecting inference from the standpoint of computational burden. Calculating the mFDR is essentially instantaneous after fitting the lasso path; the entire analysis requires just 1.3 seconds. Meanwhile, sample splitting required 9 minutes, and applying the pathwise conditional test with the forwardStop rule using the selectiveInference package required 1.5 hours.

Interestingly, although sample splitting and pathwise conditional testing each identified a single feature, they did not select the same feature. Sample splitting selected the gene NBR2, which is unquestionably associated with BRCA1 expression—the two genes are adjacent to one another on chromosome 17 and share a promoter (Di *and others*, 2010). However, NBR2 is the third gene to enter the lasso model. The forwardStop rule based on pathwise conditional testing stops after the first variable is added to the model, and thus fails to identify NBR2.

Finally, it is worth comparing these results to the selection of λ by CV. For the TCGA data, $\lambda = 0.0436$ minimizes the CV error. The estimated mFDR at this value, however, is above 90%, indicating that although this value of λ may be useful for prediction, we cannot be confident that the variables selected by the model are truly related to the outcome. Indeed, it has long been recognized from a theoretical perspective that
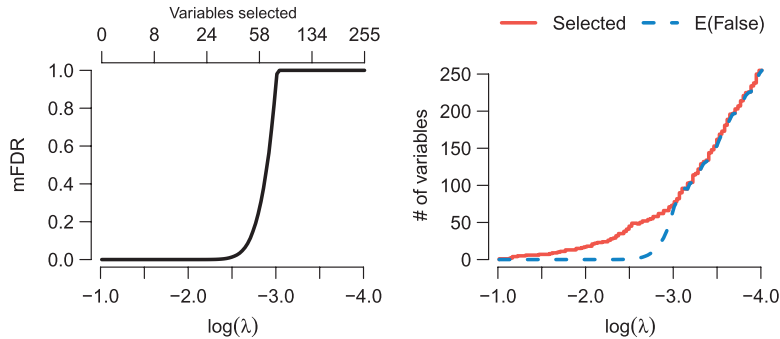
Fig. 5. False discovery rate estimates for a lasso model applied to the breast cancer TCGA data.

while the lasso has attractive variable selection and prediction properties, it cannot achieve both those aims simultaneously (Fan and Li, 2001). The mFDR estimates illustrate this concretely: $\lambda = 0.0436$ produces accurate predictions, but a larger value, $\lambda = 0.0681$, is required in order to have confidence that no more than 10% of the variables selected are false discoveries.

6.2. *Minimax concave penalty*

The mFDR estimator follows directly from the KKT conditions for a given penalty, and it is straightforward to extend to other penalties. In fact, the KKT conditions for many penalties lead to the same expression (2.2) and therefore the same mFDR estimator (although strictly speaking, for non-convex optimization problems these are no longer KKT conditions; they are still necessary but no longer sufficient).

One such penalty is the minimax concave penalty or MCP (Zhang, 2010). Briefly, the MCP produces sparse estimates like the lasso, but modifies the penalty such that the selected variables are estimated with less shrinkage towards zero (see the original article for details). The main consequence of this is that MCP solutions tend to be more sparse, with larger regression coefficients, than those of the lasso, thereby requiring fewer features to achieve the same predictive ability. To put it differently, for the lasso to estimate large coefficients accurately, it must lower λ, thereby increasing the number of noise features in the model and increasing the mFDR. MCP, on the other hand, can estimate large coefficients accurately while still screening out the noise features, at least asymptotically (this is known as the "oracle property").

To see how this works in practice, we can fit an MCP model to the TCGA data and compare it to the lasso. For the value of λ minimizing CV error, the lasso had a somewhat higher predictive accuracy (cross-validated $R^2 = 0.59$ for the lasso and $R^2 = 0.54$ for MCP), although the MCP model used far fewer features (38 compared with 96 for the lasso). Consequently, for these values of λ the estimated mFDR for MCP was much lower than that of the lasso: 44% compared with the lasso's mFDR of above 90%. Alternatively, if we restrict each model to a 10% mFDR, the lasso can achieve a predictive accuracy of $R^2 = 0.55$ compared with $R^2 = 0.52$ for MCP, illustrating the moderate sacrifice in predictive accuracy we can expect if we require the model to have a low FDR.

This is representative of the relationship between lasso and MCP: the lasso can typically achieve slightly better prediction accuracy, but in doing so selects a large number of noise features. This pattern has been observed in simulation studies (e.g., Breheny and Huang, 2011), but mFDR estimates offer a way to observe and assess this tradeoff in the analysis of real data.

This example also illustrates the ease with which the proposed estimator can be extended to other penalties. In addition to MCP, the SCAD (Fan and Li, 2001) and elastic net (Zou and Hastie, 2005) also

have similar KKT conditions leading to (2.2) and thus require only trivial modifications to the mFDR estimator.

6.3. *Genetic association study of cardiac fibrosis*

As an additional example, let us also look at a genetic association study of cardiac fibrosis. The data come from the Myocardial Applied Genomics Network (MAGNet), which collected tissue and gene expression data on 313 human hearts along with genetic data on $p = 660\,496$ SNPs. Here, we use the ratio of cardiomyocytes to fibroblasts in the heart tissue, measured on the log scale, as the outcome. An abundance of fibroblasts is indicative of cardiac fibrosis, which leads to heart failure. In this analysis, the goal is to discover single nucleotide polymorphisms (SNPs) associated with increased fibrosis. Neither sample splitting nor covariance testing was attempted here due to the size of the dataset.

In contrast to the TCGA data, no features can be selected with any degree of confidence in the MAGNet data. For all values of λ in which features are selected, the mFDR estimate is 100%. This is consistent with a traditional genome-wide association analysis, which also fails to identify any SNPs that are significant following a Bonferroni correction for multiple testing.

This negative example illustrates the usefulness of mFDR estimates in terms of guarding against false positives. The efficiency with which methods like the lasso extend to high dimensions make analyses like this (with $n = 313$ and $p = 660\,496$) feasible from a computational standpoint. However, there are important statistical considerations here that need to be accounted for; namely, with p so large, the probability of selecting noise features is so great that one cannot place any trust in the variables that the lasso selects. These considerations are clearly illustrated by the estimation of mFDR, but are lost if one simply uses the lasso to identify the top $|\mathcal{S}|$ SNPs (as in, e.g., Wu *and others*, 2009).

7. DISCUSSION

False discovery rates are not the only consideration, or even the most important consideration, in choosing a model. Depending on the application, the predictive accuracy of a model, as well its ability to identify all of the important features (the false non-discovery rate; Genovese and Wasserman, 2002), are also important. Certainly, depending on the scientific goals of the analysis, one might prefer a model with a very high false discovery rate to a model with a low false discovery rate, if the former is more useful for prediction. However, unless one is only interested in the model only as a black box for prediction, we would typically also want to know how much we can trust that the features chosen by the model are actually related to the outcome.

This is a particularly important issue for lasso-penalized regression models. The most common method for choosing λ, CV, often selects a model with good predictive accuracy, but includes a large number of noise variables unrelated to the outcome. This does not mean the model is not useful, but it does mean that one should be very cautious about interpreting features selected in this way as being "significant" in the analysis. Quantifying the false discovery rate allows one to assess this phenomenon, and either identify a smaller set of features that can be confidently selected—even if a larger set of features are used for prediction—or perhaps to choose a model that strikes a balance between prediction error and false discovery rate.

The false discovery rate estimator presented in this article is a straightforward, useful way of quantifying the reliability of feature selection for penalized regression models such as the lasso. Compared with other proposals such as sample splitting and the pathwise covariance and selective inference tests, mFDR uses a more relaxed definition of a false discovery, in which only variables marginally independent of the outcome are considered to be false discoveries. The relaxed definition pursued here has several advantages in practice, particularly in high dimensions. These advantages include greater power to detect features,

no added computation cost, and a simple rate with a straightforward interpretation. Although the estimate can be conservative for highly correlated features, in most realistic settings this conservatism is mild.

The proposed estimator is easy to implement and is available (along with the permutation methods of Section 5) in the R package ncvreg (Breheny and Huang, 2011), which was used to fit all of the models presented in this article (except where the hdi and selectiveInference packages were used). The following bit of code demonstrates its use:

```
fit <- ncvreg(X, y, penalty="lasso")
obj <- mfdr(fit)
plot(obj)
```

This will fit a lasso model, estimate the mFDR at each value of λ and assign it to an object, then plot the results, as in Figure 5.

The simplicity of the method makes it available at no added computational cost and very easy to generalize to new methods: the mfdr function in ncvreg also works for elastic net, SCAD, MCP, and Mnet (Huang *and others*, 2016) penalties. As with any statistic, there are limitations and one needs to be careful with interpretation, but mFDR is a convenient summary measure that offers a useful estimate for the reliability of feature selection in penalized regression models.

APPENDIX

All of the code and data to reproduce the analyses in this manuscript, with the exception of the human genetic data considered in Section 6.3, are available at https://github.com/pbreheny/lassoFDRpaper.

Proof of Theorem 3.1. Starting with the observation that $\mathbf{r}_j = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}$, for any $j \in \mathcal{N}$ we have

$$\frac{1}{\sqrt{n}}\mathbf{x}_j'\mathbf{r}_j = \frac{1}{\sqrt{n}}\mathbf{x}_j'\boldsymbol{\varepsilon} + (\tfrac{1}{n}\mathbf{x}_j'\mathbf{X}_{-j})\{\sqrt{n}(\boldsymbol{\beta}_{-j} - \widehat{\boldsymbol{\beta}}_{-j})\}$$

The first term, $\frac{1}{\sqrt{n}}\mathbf{x}_j'\boldsymbol{\varepsilon}$, follows a $N(0, \sigma^2)$ distribution by the independence of $\mathbf{x}_j$ and $\boldsymbol{\varepsilon}$. The second term, $\frac{1}{n}\mathbf{x}_j'\mathbf{X}_{-j}$, converges to zero by the assumption that $\Sigma_{\mathcal{N}} = \mathbf{I}$ ($\mathbf{x}_j$ is also independent of variables in $\mathcal{A}$ since $j \in \mathcal{N}$). Finally, the third term is bounded in probability provided that $\sqrt{n}\lambda_n$ is bounded (Fan and Li, 2001). Therefore, by Slutsky's Theorem, the entire quantity converges in distribution to $N(0, \sigma^2)$. □

REFERENCES

BREHENY, P. AND HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5**, 232–253.

DEZEURE, R., BÜHLMANN, P., MEIER, L. AND MEINSHAUSEN, N. (2015). High-dimensional inference: confidence intervals, *p*-values and R-software hdi. *Statistical Science* **30**, 533–558.

DI, L.-J., FERNANDEZ, A. G., DE SIERVI, A., LONGO, D. L. AND GARDNER, K. (2010). Transcriptional regulation of brca1 expression by a metabolic switch. *Nature Structural and Molecular Biology* **17**, 1406–1413.

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102**, 93–103.

FAN, J., GUO, S. AND HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society Series B* **74**, 37–65.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

GENOVESE, C. AND WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B* **64**, 499–517.

G'SELL, M. G., WAGER, S., CHOULDECHOVA, A. AND TIBSHIRANI, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 423–444.

HUANG, J., BREHENY, P., LEE, S., MA, S. AND ZHANG, C.-H. (2016). The Mnet method for variable selection. *Statistica Sinica* **26**, 903–923.

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. AND TIBSHIRANI, R. (2014). A significance test for the lasso. *The Annals of Statistics* **42**, 413–468.

MEINSHAUSEN, N., MEIER, L. AND BÜHLMANN, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association* **104**, 1671–1681.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. AND TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111**, 600–620.

WASSERMAN, L. AND ROEDER, K. (2009). High dimensional variable selection. *Annals of Statistics* **37**, 2178.

WESTFALL, P. H. AND YOUNG, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*. New York: Wiley.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. AND LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

YI, H., BREHENY, P., IMAM, N., LIU, Y. AND HOESCHELE, I. (2015). Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits. *Genetics* **199**, 205–222.

ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

ZOU, H. AND HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**, 301–320.

ZOU, H., HASTIE, T. AND TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics* **35**, 2173–2192.