Controlling the False Discovery Rate via Knockoffs

Rina Foygel Barber[a] and Emmanuel Candès[b]

[a] Department of Statistics, University of Chicago, Chicago, IL 60637, U.S.A.
[b] Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

April 2014

Abstract

In many fields of science, we observe a response variable together with a large number of potential explanatory variables, and would like to be able to discover which variables are truly associated with the response. At the same time, we need to know that the false discovery rate (FDR)—the expected fraction of false discoveries among all discoveries—is not too high, in order to assure the scientist that most of the discoveries are indeed true and replicable. This paper introduces the *knockoff filter*, a new variable selection procedure controlling the FDR in the statistical linear model whenever there are at least as many observations as variables. This method achieves exact FDR control in finite sample settings no matter the design or covariates, the number of variables in the model, and the amplitudes of the unknown regression coefficients, and does not require any knowledge of the noise level. As the name suggests, the method operates by manufacturing knockoff variables that are cheap—their construction does not require any new data—and are designed to mimic the correlation structure found within the existing variables, in a way that allows for accurate FDR control, beyond what is possible with permutation-based methods. The method of knockoffs is very general and flexible, and can work with a broad class of test statistics. We test the method in combination with statistics from the Lasso for sparse regression, and obtain empirical results showing that the resulting method has far more power than existing selection rules when the proportion of null variables is high.

Keywords. Variable selection, false discovery rate (FDR), sequential hypothesis testing, martingale theory, permutation methods, Lasso.

1 Introduction

Understanding the finite sample inferential properties of procedures that select and fit a regression model to data is possibly one of the most important topics of current research in theoretical statistics. This paper is about this problem, and focuses on the accuracy of variable selection in the classical linear model under arbitrary designs.

1.1 The false discovery rate in variable selection

Suppose we have recorded a response variable of interest y and many potentially explanatory variables X_j on n observational units. Our observations obey the classical linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}, \tag{1.1}$$

where as usual, $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of responses, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a known design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of coefficients, and $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is Gaussian noise. Because we are interested in valid inference from finitely many samples, we shall mostly restrict our attention to the case where $n \geq p$ as otherwise the model would not even be identifiable. Now in modern settings, it is often the case that there are typically just a few relevant variables among the many that have been recorded. In genetics, for instance, we typically expect that only a few genes are associated with a phenotype y of interest. In terms of the linear model (1.1), this means that only a few components of the parameter $\boldsymbol{\beta}$ are expected to be nonzero. While there certainly is no shortage of data fitting strategies, it is not always clear whether any of these offers real guarantees on the accuracy of the selection with a finite sample size. In this paper, we propose controlling the false discovery rate (FDR) among all the selected variables, i.e. all the variables included in the model, and develop novel and very concrete procedures, which provably achieve this goal.

Informally, the FDR is the expected proportion of falsely selected variables, a false discovery being a selected variable not appearing in the true model. Formally, the FDR of a selection procedure returning a subset $\hat{S} \subset \{1, \ldots, p\}$ of variables is defined as

$$\mathsf{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right]. \tag{1.2}$$

(The definition of the denominator above, sets the fraction to zero in the case that zero features are selected, i.e. $\hat{S} = \emptyset$; here we use the notation $a \vee b = \max\{a, b\}$.) We will say that a selection rule controls the FDR at level q if its FDR is guaranteed to be at most q no matter the value of the coefficients $\boldsymbol{\beta}$. This definition asks to control the type I error averaged over the selected variables, and is both meaningful and operational. Imagine we have a procedure that has just made 100 discoveries. Then roughly speaking, if our procedure is known to control the FDR at the 10% level, this means that we can expect at most 10 of these discoveries to be false and, therefore, at least 90 to be true. In other words, if the collected data were the outcome of a scientific experiment, then we would expect that most of the variables selected by the knockoff procedure correspond to real effects that could be reproduced in followup experiments.

In the language of hypothesis testing, we are interested in the p hypotheses $H_j : \beta_j = 0$ and wish to find a multiple comparison procedure able to reject individual hypotheses while controlling the FDR. This is the reason why we will at times use terminology from this literature, and may say that H_j has been rejected to mean that feature j has been selected, or may say that the data provide evidence against H_j to mean that the jth variable likely belongs to the model.

1.2 The knockoff filter

This paper introduces a general FDR controlling procedure that is guaranteed to work under *any* fixed design $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, as long as $n \geq p$ and the response $\boldsymbol{y}$ follows a linear Gaussian model as in (1.1). An important feature of this procedure is that it does not require any knowledge of the noise level σ. Also, it does not assume any knowledge about the number of variables in the model, which can be arbitrary. We now outline the steps of this new method.

Step 1: Construct knockoffs. For each feature $\boldsymbol{X}_j$ in the model (i.e. the jth column of $\boldsymbol{X}$), we construct a "knockoff" feature $\tilde{\boldsymbol{X}}_j$. The goal of the knockoff variables is to imitate the correlation structure of the original features, in a very specific way that will allow for FDR control.

Specifically, to construct the knockoffs, we first calculate the Gram matrix $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$ of the original features,[1] after normalizing each feature such that $\Sigma_{jj} = \|\boldsymbol{X}_j\|_2^2 = 1$ for all j. We will ensure that these knockoff features obey

$$\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \boldsymbol{\Sigma}, \qquad \boldsymbol{X}^\top \tilde{\boldsymbol{X}} = \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\}, \tag{1.3}$$

where $\boldsymbol{s}$ is a p-dimensional nonnegative vector. In words, $\tilde{\boldsymbol{X}}$ exhibits the same covariance structure as the original design $\boldsymbol{X}$, but in addition, the correlations between distinct original and knockoff variables are the same as those between the originals (because $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\}$ are equal on off-diagonal entries):

$$\boldsymbol{X}_j^\top \tilde{\boldsymbol{X}}_k = \boldsymbol{X}_j^\top \boldsymbol{X}_k \text{ for all } j \neq k.$$

However, comparing a feature $\boldsymbol{X}_j$ to its knockoff $\tilde{\boldsymbol{X}}_j$, we see that

$$\boldsymbol{X}_j^\top \tilde{\boldsymbol{X}}_j = \Sigma_{jj} - s_j = 1 - s_j,$$

while $\boldsymbol{X}_j^\top \boldsymbol{X}_j = \tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{X}}_j = 1$. To ensure that our method has good statistical power to detect signals, we will see that we should choose the entries of $\boldsymbol{s}$ as large as possible so that a variable $\boldsymbol{X}_j$ is not too similar to its knockoff $\tilde{\boldsymbol{X}}_j$.

A strategy for constructing $\tilde{\boldsymbol{X}}$ is to choose $\boldsymbol{s} \in \mathbb{R}_+^p$ satisfying $\operatorname{diag}\{\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}$, and construct the $n \times p$ matrix $\tilde{\boldsymbol{X}}$ of knockoff features as

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1} \operatorname{diag}\{\boldsymbol{s}\}) + \tilde{\boldsymbol{U}}\boldsymbol{C}; \tag{1.4}$$

here, $\tilde{\boldsymbol{U}}$ is an $n \times p$ orthonormal matrix that is orthogonal[2] to the span of the features $\boldsymbol{X}$, and $\boldsymbol{C}^\top \boldsymbol{C} = 2\operatorname{diag}\{\boldsymbol{s}\} - \operatorname{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\boldsymbol{s}\}$ is a Cholesky decomposition (whose existence is guaranteed by the condition $\operatorname{diag}\{\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}$; see Section 2.1.1 for details).

[1] We assume throughout that $\boldsymbol{\Sigma}$ is invertible as the model would otherwise not be identifiable.

[2] In this version of the construction, we are implicitly assuming $n \geq 2p$. Section 2.1.2 explains how to extend this method to the regime $p \leq n < 2p$.

Step 2: Calculate statistics for each pair of original and knockoff variables. We now wish to introduce statistics W_j for each $\beta_j \in \{1, \dots, p\}$, which will help us tease apart apart those variables that are in the model from those that are not. These W_j's are constructed so that large positive values are evidence against the null hypothesis $\beta_j = 0$.

In this instance, we consider the Lasso model [22], an ℓ_1-norm penalized regression that promotes sparse estimates of the coefficients $\boldsymbol{\beta}$, given by

$$\hat{\boldsymbol{\beta}}(\lambda) = \text{arg min}_{\boldsymbol{b}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\}. \tag{1.5}$$

For sparse linear models, the Lasso is known to be asymptotically accurate for both variable selection and for coefficient or signal estimation (see for example [3, 5, 23, 24]), and so even in a non-asymptotic setting, we will typically see $\hat{\boldsymbol{\beta}}(\lambda)$ including many signal variables and few null variables at some value of the penalty parameter λ. Taking Z_j to be the point λ on the Lasso path at which feature $\boldsymbol{X}_j$ first enters the model,

$$Z_j = \sup \{\lambda : \hat{\beta}_j(\lambda) \neq 0\}, \tag{1.6}$$

we then hope that Z_j is large for most of the signals, and small for most of the null variables. However, to be able to quantify this and choose an appropriate threshold for variable selection, we need to use the knockoff variables to calibrate our threshold. With this in mind, we instead compute the statistics (1.6) on the augmented $n \times 2p$ design matrix $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$ (this is the columnwise concatenation of $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$), so that $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$ replaces $\boldsymbol{X}$ in (1.5). This yields a $2p$-dimensional vector $(Z_1, \dots, Z_p, \tilde{Z}_1, \dots, \tilde{Z}_p)$. Finally, for each $j \in \{1, \dots, p\}$, we set

$$W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j, \\ -1, & Z_j < \tilde{Z}_j \end{cases} \tag{1.7}$$

(we can set W_j to zero in case of equality $Z_j = \tilde{Z}_j$). A large positive value of W_j indicates that variable $\boldsymbol{X}_j$ enters the Lasso model early (at some large value of λ) and that it does so before its knockoff copy $\tilde{\boldsymbol{X}}_j$. Hence, this is an indication that this variable is a genuine signal and belongs in the model. We may also consider other alternatives for constructing the W_j's; for instance, instead of recording the variables' entry into the Lasso model, we can consider forward selection methods and record the order in which the variables are added to the model (see Section 2.2 for this and other alternatives).

In Section 2, we discuss a broader methodology, where the statistics W_j may be defined in any manner that satisfies the sufficiency property and the antisymmetry property, which we will define later on; the construction above is a specific instance that we find to perform well empirically.

Step 3: Calculate a data-dependent threshold for the statistics. We wish to select variables such that W_j is large and positive, i.e. such that $W_j \geq t$ for some $t > 0$. Letting q be the target FDR, define a data-dependent threshold T as:

$$T = \min \left\{ t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\} \quad (\text{or } T = +\infty \text{ if this set is empty}) \,, \tag{1.8}$$

where $\mathcal{W} = \{|W_j| : j = 1, \dots, p\} \backslash \{0\}$ is the set of unique nonzero[3] values attained by the $|W_j|$'s. We shall see that the fraction appearing above, is an estimate of the proportion of false discoveries if we were to select all features j's with $W_j \geq t$. For this reason, we will often refer to this fraction as the "knockoff estimate of FDP".

For a visual representation of this step, see Figure 1, where we plot the point $(Z_j, \tilde{Z}_j)$ for each feature j, with black dots denoting null features and red squares denoting true signals. Recall that W_j is positive if the original variable is selected before its knockoff (i.e. $Z_j > \tilde{Z}_j$), and is negative otherwise (i.e. $Z_j < \tilde{Z}_j$). Therefore a feature j whose point lies below the dashed diagonal line in Figure 1 then has a positive value of W_j, while points above the diagonal are assigned negative W_j's. For a given value of t, the numerator and denominator of the fraction appearing in (1.8) above are given by the numbers of points in the two gray shaded regions of the figure (with nulls and non-nulls both counted, since in practice we do not know which features are null).

With these steps in place, we are ready to define our procedure:

[3] If $W_j = 0$ for some feature $\boldsymbol{X}_j$, then this gives no evidence for rejecting the hypothesis $\beta_j = 0$, and so our method will never select such variables.

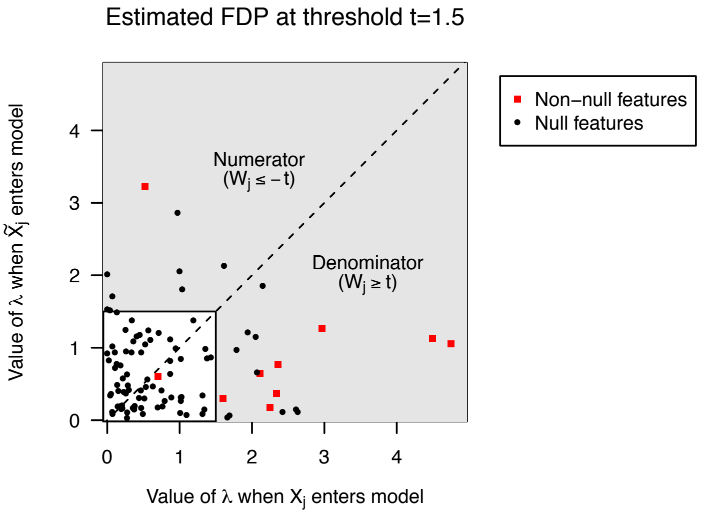


Figure 1: Representation of the knockoff procedure plotting pairs $(Z_j, \tilde{Z}_j)$. Black dots correspond to null hypotheses ($\beta_j = 0$) while red squares are nonnulls ($\beta_j \neq 0$). Setting $t = 1.5$, the number of points in the shaded region below the diagonal is equal to $\#\{j : W_j \ge t\}$, the number of selected variables at this threshold, while the number of points in the shaded region above the diagonal is equal to $\#\{j : W_j \le -t\}$. Observe that the true signals (red squares) are primarily below the diagonal, indicating $W_j > 0$, while the null features (black dots) are roughly symmetrically distributed across the diagonal.

Definition 1 (Knockoff). *Construct $\tilde{\boldsymbol{X}}$ as in* (1.4)*, and calculate statistics W_j satisfying the sufficiency and antisymmetry properties (defined in Section 2;* (1.7) *above gives an example of a statistic satisfying these properties). Then select the model*

$$\hat{S} = \{j : W_j \ge T\},$$

where T is the data-dependent threshold (1.8)*.*

A main result of this paper is that this procedure controls a quantity nearly equal to the FDR:

Theorem 1. *For any $q \in [0,1]$, the knockoff method satisfies*

$$\mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} + q^{-1}}\right] \le q,$$

where the expectation is taken over the Gaussian noise $\boldsymbol{z}$ in the model (1.1)*, while treating $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ as fixed.*

The "modified FDR" bounded by this theorem is very close to the FDR in settings where a large number of features are selected (as adding q^{-1} in the denominator then has little effect), but it sometimes may be preferable to control the FDR exactly. For this, we propose a slightly more conservative procedure:

Definition 2 (Knockoff+). *Select a model as in Definition 1 but with a data-dependent threshold T defined as*

$$T = \min\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le q\right\} \quad \textit{(or } T = +\infty \textit{ if this set is empty)}\,, \tag{1.9}$$

Note that the threshold T chosen by knockoff+ is always higher (or equal to) than that chosen in (1.8) by the knockoff filter, meaning that knockoff+ is (slightly) more conservative.

Our second main result shows that knockoff+ controls the FDR.

Theorem 2. *For any $q \in [0,1]$, the knockoff+ method satisfies*

$$\mathsf{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \textit{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right] \le q,$$

where the expectation is taken over the Gaussian noise $\boldsymbol{z}$ in the model (1.1)*, while treating $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ as fixed.*

We have explained why a large positive value of W_j bears some evidence against the null hypothesis $\beta_j = 0$, and now give a brief intuition for how our specific choice of threshold allows control of FDR (or of the modified FDR). The way in which $\boldsymbol{W}$ is constructed implies that the signs of the W_j's are i.i.d. random for the "null hypotheses"; that is, for those j's such that $\beta_j = 0$. Therefore, for any threshold t,

$$\#\{j : \beta_j = 0 \text{ and } W_j \geq t\} \overset{d}{=} \#\{j : \beta_j = 0 \text{ and } W_j \leq -t\}, \tag{1.10}$$

where $\overset{d}{=}$ means equality in distribution. In Figure 1, for instance, $\#\{j : \beta_j = 0 \text{ and } W_j \geq t\}$ is the number of null points (black dots) in the shaded region below the diagonal, while $\#\{j : \beta_j = 0 \text{ and } W_j \leq -t\}$ is the number of null points in the shaded region above the diagonal. Note that the null points are distributed approximately symmetrically across the diagonal, as described by (1.10).

Hence, we can estimate the false discovery proportion (FDP) at the threshold t as

$$\frac{\#\{j : \beta_j = 0 \text{ and } W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j : \beta_j = 0 \text{ and } W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} =: \widehat{\mathsf{FDP}}(t), \tag{1.11}$$

where $\widehat{\mathsf{FDR}}(t)$ is the knockoff estimate of FDP. The knockoff procedure can be interpreted as finding a threshold via $T = \min\{t \in \mathcal{W} : \widehat{\mathsf{FDP}}(t) \leq q\}$, with the convention that $T = +\infty$ if no such t exists; this is the most liberal threshold with the property that the estimated FDP is under control. In fact, the inequality in (1.11) will usually be tight because most strong signals will be selected before their knockoff copies (in Figure 1, we see that most of the red squares lie below the diagonal, i.e. $W_j \geq 0$); this means that our estimate of FDP will probably be fairly tight unless the signal strength is weak. (We will see later that the additional "+1" appearing in the knockoff+ method, yielding a slightly more conservative procedure, is necessary both theoretically and empirically to control FDR in scenarios where extremely few discoveries are made.)

1.3 Outline of the paper

The rest of this paper is organized as follows:

- Section 2 introduces the more general form of our variable selection procedure, and gives some theoretical properties of the procedure that will allow for FDR control.
- Section 3 discusses some related methods and strategies for FDR control for variable selection. We compare our proposal to permutation-based methods, to the Benjamini-Hochberg (BHq) procedure and some of its variants, and to other methods. In particular, Section 3.3 presents simulations demonstrate that the method is effective in practice, and performs well compared to the BHq and related procedures.
- In Section 4, we present an application of the knockoff method to real data where the task is to find mutations in the HIV-1 protease or reverse transcriptase that are associated with drug resistance.
- Section 5 moves to a more general problem of sequential hypothesis testing—we show that our approach is an example of a procedure for controlling FDR in a sequential hypothesis testing problem.
- Some proofs are deferred to Section 6.
- In Section 7 we close the paper with a discussion outlining possible extensions of this work.

2 Knockoffs and FDR Control

2.1 The knockoff features

We begin with the construction of the knockoff features $\tilde{\boldsymbol{X}}_j$ and set $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$ as before. We first present the method in the natural setting where $n \geq 2p$ before presenting ways of extending the construction to the range $p \leq n < 2p$. To ease readability, vectors and matrices are boldfaced throughout the paper whereas scalars are not.

2.1.1 The natural setting $n \geq 2p$

As introduced earlier, the matrix $\tilde{\boldsymbol{X}}$ must obey

$$\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma}\end{bmatrix} := \boldsymbol{G}, \tag{2.1}$$

where $\boldsymbol{s} \in \mathbb{R}^p$ is some vector. A necessary and sufficient condition for $\tilde{\boldsymbol{X}}$ to exist is that $\boldsymbol{G}$ is positive semidefinite. Indeed, recall that $\boldsymbol{G} \succeq \boldsymbol{0}$ if and only if the Schur complement

$$\boldsymbol{A} = \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\}) = 2\operatorname{diag}\{\boldsymbol{s}\} - \operatorname{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\boldsymbol{s}\} \tag{2.2}$$

is positive semidefinite. In turn, standard Schur complement calculations show that $\boldsymbol{A} \succeq \boldsymbol{0}$ if and only if

$$\begin{bmatrix}\boldsymbol{\Sigma} & \operatorname{diag}\{\boldsymbol{s}\} \\ \operatorname{diag}\{\boldsymbol{s}\} & 2\operatorname{diag}\{\boldsymbol{s}\}\end{bmatrix} \succeq \boldsymbol{0} \quad \iff \quad \begin{array}{l}\operatorname{diag}\{\boldsymbol{s}\} \succeq \boldsymbol{0} \\ 2\boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} \succeq \boldsymbol{0}\end{array}$$

as claimed earlier. Now let $\tilde{\boldsymbol{U}} \in \mathbb{R}^{n \times p}$ be an orthonormal matrix whose column space is orthogonal to that of $\boldsymbol{X}$ so that $\tilde{\boldsymbol{U}}^\top \boldsymbol{X} = \boldsymbol{0}$: such a matrix exists because $n \geq 2p$. Since $\boldsymbol{A} \succeq \boldsymbol{0}$, we can factorize it as $\boldsymbol{A} = \boldsymbol{C}^\top \boldsymbol{C}$, where $\boldsymbol{C}$ is $p \times p$. A simple calculation then shows that setting

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\boldsymbol{s}\}) + \tilde{\boldsymbol{U}}\boldsymbol{C} \tag{2.3}$$

gives the correlation structure specified in (2.1).

Now that we understand the condition on $\boldsymbol{s}$ necessary for knockoff features with the desired correlation structure to exist, it remains to discuss which one we should construct, that is, to specify a choice of $\boldsymbol{s}$. Returning to the example of statistic from Section 1.2, we will have a useful methodology only if those variables that truly belong to the model tend to be selected before their knockoffs as we would otherwise have no power. Imagine that variable $\boldsymbol{X}_j$ is in the true model. Then we wish to have $\boldsymbol{X}_j$ enter before $\tilde{\boldsymbol{X}}_j$. To make this happen, we need the correlation between $\tilde{\boldsymbol{X}}_j$ and the true signal to be small, so that $\tilde{\boldsymbol{X}}_j$ does not enter the Lasso model early. In other words, we would like $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ to be as orthogonal to each other as possible. In a setting where the features are normalized, i.e. $\Sigma_{jj} = 1$ for all j, we would like to have $\tilde{\boldsymbol{X}}_j^\top \boldsymbol{X}_j = 1 - s_j$ as close to zero as possible. Below, we consider two particular types of knockoffs:

- *Equi-correlated knockoffs*: Here, $s_j = 2\lambda_{\text{min}}(\boldsymbol{\Sigma}) \wedge 1$ for all j, so that all the correlations take on the identical value
$$\langle \boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j \rangle = 1 - 2\lambda_{\text{min}}(\boldsymbol{\Sigma}) \wedge 1. \tag{2.4}$$
Among all knockoffs with this equi-variant property, this choice minimizes the value of $|\langle \boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j \rangle|$.

- *SDP knockoffs*: Another possibility is to select knockoffs so that the average correlation between an original variable and its knockoff is minimum. This is done by solving the convex problem
$$\begin{array}{ll}\text{minimize} & \sum_j |1 - s_j| \\ \text{subject to} & s_j \geq 0 \\ & \operatorname{diag}\{\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}.\end{array}$$
This optimization problem is a highly structured semidefinite program (SDP), which can be solved very efficiently [4]. At the optimum, $s_j \leq 1$ for all j (the reason is that if $s_j > 1$, then setting $s_j = 1$ maintains feasibility and reduces the value of the objective), so that this problem is equivalent to the simpler SDP
$$\begin{array}{ll}\text{minimize} & \sum_j (1 - s_j) \\ \text{subject to} & 0 \leq s_j \leq 1 \\ & \operatorname{diag}\{\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}.\end{array} \tag{2.5}$$

2.1.2 Extensions to $p \leq n < 2p$

When $n < 2p$, we can no longer find a subspace of dimension p which is orthogonal to $\boldsymbol{X}$, and so we cannot construct $\tilde{U}$ as above. We can still use the knockoff filter, however, as long as the noise level σ is known or can be estimated—for instance, under the Gaussian noise model (1.1), we can use the fact that the residual sum of squares from the full model is distributed as $\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}}\|_2^2 \sim \sigma^2 \cdot \chi^2_{n-p}$, where $\hat{\boldsymbol{\beta}}^{\mathsf{LS}}$ is the vector of coefficients in a least-squares regression. Now letting $\hat{\sigma}$ be our estimate of σ, draw a $(n-p)$-dimensional vector $\boldsymbol{y}'$ with i.i.d. $\mathcal{N}(0, \hat{\sigma}^2)$ entries. If $n-p$ is large, then $\hat{\sigma}$ will be an extremely accurate estimate of σ, and we can proceed as though σ and $\hat{\sigma}$ were equal. We then augment the response vector $\boldsymbol{y}$ with the new $(n-p)$-length vector $\boldsymbol{y}'$, and augment the design matrix $\boldsymbol{X}$ with $n-p$ rows of zeros. Then approximately,

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{\beta}, \sigma^2 \mathbf{I} \right) .$$

We now have a linear model with p variables and $2p$ observations, and so we can apply the knockoff filter to this row-augmented data using the method described for the $n \geq 2p$ setting.

We present one other alternative method for the $p < n < 2p$ regime, which may be used if one prefers not to estimate the noise level σ. Specifically, we can use the "gap" $n - p$ between the number of observations and the number of variables, to test only a subset of $n-p$ of the variables. To do this, select $2p - n$ variables for which $\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j$ so that these are simply duplicated, and construct knockoffs for the remaining $n-p$; that is, we choose the vector $\boldsymbol{s}$ in (2.1) such that only $n-p$ entries are nonzero. Then since the matrix $\boldsymbol{A}$ in (2.2) has rank $n-p$, we can write $\boldsymbol{A} = \boldsymbol{C}^\top \boldsymbol{C}$ for some $\boldsymbol{C} \in \mathbb{R}^{(n-p)\times p}$. Finally, we find $\tilde{\boldsymbol{U}} \in \mathbb{R}^{n \times (n-p)}$ orthogonal to $\boldsymbol{X}$, and then define $\tilde{\boldsymbol{X}}$ as in (2.3) and proceed as before. Clearly, while this approach will control the FDR, it will have no power in detecting those variables in the duplicated set. To address this, then, we can cycle through the duplicates; for clarity, we present the idea for $n = 3p/2$. Partition the p variables as $J_1 \cup J_2$, with each J_i of cardinality $p/2$. In the first round, duplicate variables in J_1 and construct knockoffs for those in J_2. In the second, reverse the roles of J_1 and J_2. Now if in each round, we run the knockoff method controlling the FDR at level $q/2$, the overall FDR is guaranteed to be controlled at level q.

In Section 4, we analyze real HIV data with most cases of the form $n < 2p$ and, thereby, show that the basic knockoff method can be adapted to situations in which $n \geq p$ even if $n \not\geq 2p$, so that it applies all the way to the limit of model identifiability.

2.2 Symmetric statistics

We next consider a statistic $W\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}, \boldsymbol{y}\right) \in \mathbb{R}^p$ with large positive values of W_j giving evidence that $\beta_j \neq 0$, and introduce two simple properties.

- The statistic $\boldsymbol{W}$ is said to obey the *sufficiency property* if $\boldsymbol{W}$ depends only on the Gram matrix and on feature-response inner products; that is, we can write

  $$\boldsymbol{W} = f\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \boldsymbol{y}\right)$$

  for some $f : S^+_{2p} \times \mathbb{R}^{2p} \to \mathbb{R}^p$, where S^+_{2p} is the cone of $2p \times 2p$ positive semidefinite matrices. We allow ourselves to call this the sufficiency property since under Gaussian noise $\boldsymbol{X}^\top \boldsymbol{y}$ is a sufficient statistic for $\boldsymbol{\beta}$.

- The statistic $\boldsymbol{W}$ is said to obey the *antisymmetry property* if swapping $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ has the effect of switching the sign of W_j—that is, for any $S \subseteq \{1, \ldots, p\}$,

  $$W_j\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}, \boldsymbol{y}\right) = W_j\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}, \boldsymbol{y}\right) \cdot \begin{cases} +1, & j \notin S, \\ -1, & j \in S. \end{cases}$$

  Here, we write $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}$ to mean that the columns $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ have been swapped in the matrix $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$, for each $j \in S$. Formally, if $\boldsymbol{V} \in \mathbb{R}^{n \times 2p}$ with columns $\boldsymbol{V}_j$, then for each $j = 1, \ldots, p$,

  $$(\boldsymbol{V}_{\mathsf{swap}(S)})_j = \begin{cases} \boldsymbol{V}_j, & j \notin S, \\ \boldsymbol{V}_{j+p}, & j \in S, \end{cases} \quad (\boldsymbol{V}_{\mathsf{swap}(S)})_{j+p} = \begin{cases} \boldsymbol{V}_{j+p}, & j \notin S, \\ \boldsymbol{V}_j, & j \in S. \end{cases}$$

The statistic $\boldsymbol{W}$ we examined in Section 1.2, given in equation (1.7), obeys these two properties. The reason why the sufficiency property holds is that the Lasso (1.5) is equivalent to

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} - \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} + \lambda \|\boldsymbol{b}\|_1,$$

and thus depends upon the problem data $(\boldsymbol{X}, \boldsymbol{y})$ through $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ only.[4] Note that the antisymmetry property in (1.7) is explicit. This is only one example of a statistic of this type—other examples include:

1. $W_j = \boldsymbol{X}_j^\top \boldsymbol{y} - \tilde{\boldsymbol{X}}_j^\top \boldsymbol{y}$. Under the Gaussian model for $\boldsymbol{y}$, one can show that $\boldsymbol{W} \sim \mathcal{N}(\operatorname{diag}\{\boldsymbol{s}\}\boldsymbol{\beta}, 2\sigma^2 \operatorname{diag}\{\boldsymbol{s}\})$, which means that the p statistics are independent. Rescaling things so that $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{\beta}, 2\sigma^2 \operatorname{diag}\{\boldsymbol{s}^{-1}\})$, this distribution can be obtained by simply taking $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{y} + \mathcal{N}(\boldsymbol{0}, \sigma^2(2\operatorname{diag}\{\boldsymbol{s}^{-1}\} - \boldsymbol{\Sigma}^{-1}))$, where the terms in the sum are independent. For a large signal $|\beta_j|$, however, W_j may be positive or negative depending on the sign of β_j.

2. $W_j = |\boldsymbol{X}_j^\top \boldsymbol{y}| - |\tilde{\boldsymbol{X}}_j^\top \boldsymbol{y}|$, which resolves the issue of sign above.

3. $W_j = |\hat{\beta}_j^{\mathsf{LS}}| - |\hat{\beta}_{j+p}^{\mathsf{LS}}|$ or $W_j = |\hat{\beta}_j^{\mathsf{LS}}|^2 - |\hat{\beta}_{j+p}^{\mathsf{LS}}|^2$, where $\hat{\boldsymbol{\beta}}^{\mathsf{LS}}$ is the least-squares solution obtained by regressing $\boldsymbol{y}$ on the augmented design, $\hat{\boldsymbol{\beta}}^{\mathsf{LS}} = \left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}\right)^{-1} \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \boldsymbol{y}$.

4. Define Z_j as in Section 1.2, $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$ for $j = 1, \ldots, 2p$ where $\hat{\boldsymbol{\beta}}(\lambda)$ is the solution to the augmented Lasso model regressing $\boldsymbol{y}$ on $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$. We may then take $W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p})$, but may also consider other options such as $W_j = Z_j - Z_{j+p}$, or alternately we can take $W_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$ for some fixed value of λ. (For consistency with the rest of this section, the notation here is slightly different than in Section 1.2, with Z_{j+p} instead of $\tilde{Z}_j$ giving the λ value when $\tilde{\boldsymbol{X}}_j$ entered the Lasso model.)

5. The example above of course extends to all penalized likelihood estimation procedures of the form

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda P(\boldsymbol{b}),$$

   where $P(\cdot)$ is a penalty function—the Lasso being only one such example. We can again define $\boldsymbol{W}$ by finding the λ values at which each feature enters the model, or by fixing λ and comparing coefficients in $\hat{\beta}$.

6. We can also consider a forward selection procedure [8]: initializing the residual as $\boldsymbol{r}_0 = \boldsymbol{y}$, we iteratively choose variables via

$$j_t = \arg\max_j |\langle \boldsymbol{X}_j, \boldsymbol{r}_{t-1}\rangle|$$

   and then update the residual $\boldsymbol{r}_t$ by either regressing the previous residual $\boldsymbol{r}_{t-1}$ on $\boldsymbol{X}_{j_t}$ and taking the remainder, or alternately using orthogonal matching pursuit [16] where after selecting j_t we define $\boldsymbol{r}_t$ to be the residual of the least square regression of $\boldsymbol{y}$ onto $\{\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_t}\}$. As before, however, we apply this procedure to the augmented design matrix $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$. Next let $Z_1, \ldots, Z_{2p}$ give the reverse order in which the $2p$ variables (the originals and the knockoffs) entered the model, i.e. $Z_j = 2p$ if $\boldsymbol{X}_j$ entered first; $Z_j = 2p - 1$ if $\boldsymbol{X}_j$ entered second; etc. The statistics $W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p})$ then reflect the time at which the original variable $\boldsymbol{X}_j$ and the knockoff variable $\tilde{\boldsymbol{X}}_j$ entered the model.

Clearly, the possibilities are endless.

2.3 Exchangeability results

Despite the fact that the statistics W_j, $j \in \{1, \ldots, p\}$, are dependent and have marginal distributions that are complicated functions of the unknown parameter vector $\boldsymbol{\beta}$, our selection procedure provably controls the false discovery rate, as stated in Theorems 1 and 2 from Section 1.2. In this section, we establish a property of the statistics W_j that we will use to prove our main results on FDR control.

In fact, the construction of the knockoff features, and the symmetry of the test statistic, are in place to achieve a crucial property: namely, that the signs of the W_j's are i.i.d. random for the "null hypotheses".

[4] If we would like to include an intercept term in our model, i.e. $\boldsymbol{y} = \beta_0 \boldsymbol{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}$, then the Lasso also depends on $\boldsymbol{X}^\top \boldsymbol{1}$ and $\boldsymbol{y}^\top \boldsymbol{1}$. In this case, we can apply our method as long as the knockoffs additionally satisfy $\tilde{\boldsymbol{X}}^\top \boldsymbol{1} = \boldsymbol{X}^\top \boldsymbol{1}$.

Lemma 1 (i.i.d. signs for the nulls). *Let $\boldsymbol{\epsilon} \in \{\pm 1\}^p$ be a sign sequence independent of $\boldsymbol{W}$, with $\epsilon_j = +1$ for all non-null j and $\epsilon_j \overset{\text{i.i.d.}}{\sim} \{\pm 1\}$ for null j. Then*

$$(W_1, \dots, W_p) \overset{d}{=} (W_1 \cdot \epsilon_1, \dots, W_p \cdot \epsilon_p).$$

This property fully justifies our earlier statement (1.10) that $\#\{j : \beta_j = 0, W_j \leq -t\}$ is distributed as $\#\{j : \beta_j = 0, W_j \geq t\}$. Indeed, conditional on $|\boldsymbol{W}| = (|W_1|, \dots, |W_p|)$, both these random variables follow the same binomial distribution, which implies that their marginal distributions are identical. In turn, this gives that $\widehat{\mathsf{FDP}}(t)$ from Section 1.2 is an estimate of the true false discovery proportion $\mathsf{FDP}(t)$.

The i.i.d. sign property for the nulls is a consequence of two exchangeability properties for $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$.

Lemma 2 (Pairwise exchangeability for the features). *For any subset $S \subset \{1, \dots, p\}$,*

$$\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)} = \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}.$$

That is, the Gram matrix of $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$ is unchanged when we swap $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ for each $j \in S$.

Proof. This follows trivially from the definition of $\boldsymbol{G} = \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$ in (2.1). □

Lemma 3 (Pairwise exchangeability for the response). *For any subset S of nulls,*

$$\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}^\top \boldsymbol{y} \overset{d}{=} \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \boldsymbol{y}.$$

That is, the distribution of the product $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \boldsymbol{y}$ is unchanged when we swap $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ for each $j \in S$, as long as none of the swapped features appear in the true model.

Proof. Since $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$, for any S', we have

$$\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S')}^\top \boldsymbol{y} \sim N\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S')}^\top \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S')}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S')}\right).$$

Next we check that the mean and variance calculated here are the same for $S' = S$ and for $S' = \emptyset$. Lemma 2 proves that the variances are equal. For the means, since $\boldsymbol{X}_j^\top \boldsymbol{X}_i = \tilde{\boldsymbol{X}}_j^\top \boldsymbol{X}_i$ for all $i \neq j$, and support$(\boldsymbol{\beta}) \cap S = \emptyset$, we see that $\boldsymbol{X}_j^\top \boldsymbol{X}\boldsymbol{\beta} = \tilde{\boldsymbol{X}}_j^\top \boldsymbol{X}\boldsymbol{\beta}$ for all $j \in S$, which is sufficient. □

Proof of Lemma 1. For any set $S \subset \{1, \dots, p\}$, let $\boldsymbol{W}_{\mathsf{swap}(S)}$ be the statistic we would get if we had replaced $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$ with $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}$ when calculating $\boldsymbol{W}$. The anti-symmetry property gives

$$\boldsymbol{W}_{\mathsf{swap}(S)} = (W_1 \cdot \epsilon_1, \dots, W_p \cdot \epsilon_p), \quad \epsilon_j = \begin{cases} +1, & j \notin S, \\ -1, & j \in S. \end{cases}$$

Now let $\boldsymbol{\epsilon}$ be as in the statement of the lemma and let $S = \{j : \epsilon_j = -1\}$. Since S contains only nulls, Lemmas 2 and 3 give

$$\boldsymbol{W}_{\mathsf{swap}(S)} = f\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}^\top (\boldsymbol{X}\ \tilde{\boldsymbol{X}})_{\mathsf{swap}(S)}, \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}_{\mathsf{swap}(S)}^\top \boldsymbol{y}\right) \overset{d}{=} f\left(\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}^\top \boldsymbol{y}\right) = \boldsymbol{W}.$$

This proves the claim. □

2.4 Proof sketch for main results

With the exchangeability property of the W_j's in place, we now turn to the proof of our main results, Theorems 1 and 2, which establish FDR control for the knockoff filter (specifically, approximate FDR control for the knockoff method, and exact FDR control for knockoff+). In this section, we sketch the main ideas behind the proof, restricting our attention to the knockoff+ method for simplicity. The full details will be presented later, in Sections 5 and 6, where we will see that our methods can be framed as special cases of a sequential hypothesis testing procedure. Such sequential procedures are not specifically about the regression problem we consider here, and this is the reason why we prefer postponing their description as not to distract from the problem at hand.

To understand how the knockoff+ method controls FDR, we consider Step 3 of the method, where after calculating the statistics W_j, we must now choose a data-dependent threshold T given by

$$T = \min\left\{ t \in \mathcal{W} : \widehat{\mathsf{FDP}}(t) := \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\}} \le q \right\} .$$

Assume without loss of generality that $|W_1| \ge \cdots \ge |W_p|$. Then, to find the smallest possible value of T such that the knockoff estimate of FDP is $\le q$, we can simply test the smallest value $t = |W_p|$, then move to $t = |W_{p-1}|$, and so on, stopping as soon as we find a value of t that satisfies $\widehat{\mathsf{FDP}}(t) \le q$. From this description, it appears that T is a stopping time with respect to a filtration we shall introduce in Section 6, and in fact, the main step of our proof is to show that T is a stopping time for a supermartingale that is given by

$$\frac{\#\{j : \beta_j = 0 \text{ and } W_j \le -t\}}{1 + \#\{j : \beta_j = 0 \text{ and } W_j \ge t\}} .$$

By the Optional Stopping Time theorem, therefore, the expected value of this supermartingale at the random time $t = T$ is bounded by its expected value at time $t = 0$: letting p_0 be the number of null features and writing $Y = \#\{j : \beta_j = 0 \text{ and } W_j \le 0\}$, we have

$$\mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } W_j \le -T\}}{1 + \#\{j : \beta_j = 0 \text{ and } W_j \ge T\}}\right] \le \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } W_j \le 0\}}{1 + \#\{j : \beta_j = 0 \text{ and } W_j \ge 0\}}\right] = \mathbb{E}\left[\frac{Y}{1 + p_0 - Y}\right] \le 1 ,$$

where the last step comes from a property of the binomial distribution proved in Section 6; note that since $\mathsf{sign}(W_j) \overset{\text{i.i.d.}}{\sim} \{\pm 1\}$ for the null features j, then $Y = \#\{j : \beta_j = 0 \text{ and } W_j \le 0\}$ is distributed as a $\mathsf{Binomial}(p_0, 1/2)$ random variable (for the purposes of this proof sketch, we assume here that $W_j \neq 0$ for all j for simplicity). With this bound in place, we are ready to prove FDR control for knockoff+. We have

$$\begin{aligned}
\mathsf{FDR} &= \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } W_j \ge T\}}{\#\{j : W_j \ge T\} \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } W_j \ge T\}}{1 + \#\{j : \beta_j = 0 \text{ and } W_j \le -T\}} \cdot \frac{1 + \#\{j : W_j \le -T\}}{\#\{j : W_j \ge T\} \vee 1}\right] \\
&\le \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } W_j \ge T\}}{1 + \#\{j : \beta_j = 0 \text{ and } W_j \le -T\}} \cdot q\right] \\
&\le q ,
\end{aligned}$$

where the first inequality applies the definition of T to bound the latter fraction in the expectation, while the second inequality applies our martingale bound from above. The proof for the knockoff method is similar, and we refer the reader to Sections 5 and 6 for details.

3 Comparison with other Variable Selection Techniques

There are of course many other variable selection techniques, based on ideas from Benjamini and Hochberg or perhaps based on permuted designs rather than knockoffs, which may be designed with the goal of keeping FDR under control. In this section, we review some of these procedures and compare some of them empirically.

3.1 Comparing to a permutation method

To better understand the ideas behind our method, we next ask whether we could have constructed the matrix of knockoff features $\tilde{\boldsymbol{X}}$ with a simple permutation. Specifically, would the above results hold if, instead of constructing $\tilde{\boldsymbol{X}}$ as above, we use a matrix $\boldsymbol{X}^{\pi}$, with entries given by

$$\boldsymbol{X}^{\pi}_{i,j} = \boldsymbol{X}_{\pi(i),j}$$

for some randomly chosen permutation π of the sample indices $\{1,\ldots,n\}$? In particular, the matrix $\boldsymbol{X}^\pi$ will always satisfy

$$\boldsymbol{X}^{\pi\top}\boldsymbol{X}^\pi = \boldsymbol{X}^\top\boldsymbol{X},$$

and so the permuted covariates display the same correlation structure as the original covariates, while breaking association with the response $\boldsymbol{y}$ due to the permutation.

Permutation methods are widely used in applied research. While they may be quite effective under a global null, they may fail to yield correct answers in cases other than the global null; see also [6, 7] for other sources of problems associated with permutation methods. Consequently, inference in practical settings, where some signals do exist, can be quite distorted. In the linear regression problem considered here, a permutation-based construction can dramatically underestimate the FDP in cases where $\boldsymbol{X}$ displays nonvanishing correlations. To understand why, suppose that the features $\boldsymbol{X}_j$ are centered. Then the Gram of the augmented design matrix $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{X}^\pi\end{bmatrix}$ is equal to

$$\begin{bmatrix}\boldsymbol{X} & \boldsymbol{X}^\pi\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{X} & \boldsymbol{X}^\pi\end{bmatrix} \approx \begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}\end{bmatrix},$$

where the approximately-zero off-diagonal blocks arise from the fact that $\mathbb{E}_\pi\left[\boldsymbol{X}_i^\top\boldsymbol{X}_j^\pi\right] = 0$ when the features are centered. In particular, the exchangeability results (Lemmas 2 and 3) will not hold for the augmented matrix $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{X}^\pi\end{bmatrix}$, and this can lead to extremely poor control of FDR.

To see this empirically, consider a setting high correlation between features. We let $n = 300$ and $p = 100$, and generate each row of $\boldsymbol{X}$ i.i.d. from a $\mathcal{N}(\boldsymbol{0},\boldsymbol{\Theta})$ distribution, where $\Theta_{ii} = 1$ for all i and $\Theta_{ij} = 0.3$ for all $i \neq j$. We then center and normalize the columns of $\boldsymbol{X}$, and define

$$\boldsymbol{y} = 3.5\cdot(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_{30}) + \boldsymbol{z} \text{ where } \boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0},\mathbf{I}_n). \tag{3.1}$$

Next we fit the Lasso path (1.5) for the response $\boldsymbol{y}$ and the augmented design matrix $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{X}^\pi\end{bmatrix}$. Let Z_j and Z_j^π denote the largest λ values when feature $\boldsymbol{X}_j$ and permuted feature $\boldsymbol{X}_j^\pi$, respectively, enter the Lasso path (exactly as for the knockoff method, but with $\boldsymbol{X}^\pi$ replacing $\tilde{\boldsymbol{X}}$). Figure 2 shows the values of Z_j and Z_j^π for a single trial of this simulation. We see that while many of the original null features enter the model at moderate values of λ, the permuted features do not enter the Lasso path until λ is extremely small; the difference arises from the fact that only the original null features are correlated with the signals $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{30}$. In other words, the $\boldsymbol{X}_j^\pi$'s are not good knockoffs of the $\boldsymbol{X}_j$'s for the nulls $j = 31,\ldots,100$—they behave very differently in the Lasso regression. Next we test the effect of these issues on FDR control. Using either the knockoff features $\tilde{\boldsymbol{X}}_j$ or the permuted features $\boldsymbol{X}_j^\pi$, we proceed to construct the statistics $\boldsymbol{W}$ (1.7) and the threshold T (1.8) as with the knockoff method (i.e. after computing Lasso models using either $\begin{bmatrix}\boldsymbol{X} & \tilde{\boldsymbol{X}}\end{bmatrix}$ or $\begin{bmatrix}\boldsymbol{X} & \boldsymbol{X}^\pi\end{bmatrix}$). We obtain the following FDR, when the target FDR is set at $q = 20\%$:

	FDR over 1000 trials (nominal level $q = 20\%$)
Knockoff method	12.29%
Permutation method	45.61%

We note that there are many ways in which the permuted features $\boldsymbol{X}^\pi$ may be used to try to estimate or control FDR, but in general such methods will suffer from similar issues arising from the lack of correlation between the permuted and the original features.

3.2 The Benjamini-Hochberg procedure and variants

The Benjamini-Hochberg (BHq) procedure [1] is a hypothesis testing method known to control FDR under independence. Given z-scores $Z_1,\ldots,Z_p$ corresponding to p hypotheses being tested so that $Z_j \sim \mathcal{N}(0,1)$ if the jth hypothesis is null, the procedure[5] rejects a hypothesis whenever $|Z_j| \geq T$, where T is a data-dependent threshold given by

$$T = \min\left\{t : \frac{p\cdot\mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}}{\#\{j : |Z_j| \geq t\}} \leq q\right\} \quad \text{(or } T = +\infty \text{ if this set is empty)}, \tag{3.2}$$

[5] We present BHq in the notation from [19], and use z-scores rather than p-values, to facilitate comparison with our methods.

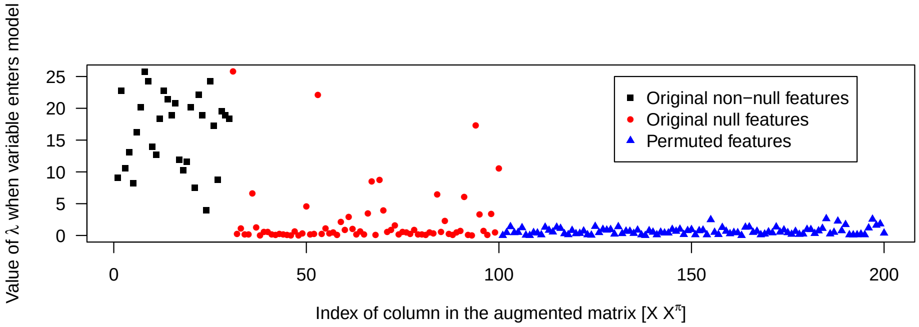


Figure 2: Results of the Lasso path, with simulated data specified in (3.1). Many of the null features $\boldsymbol{X}_j$ for $j = 31, \dots, 100$ enter the Lasso model earlier (i.e. at higher values of λ) than most of the permuted features, leading to loss of FDR control.

for a desired FDR level q. Note that for any t, the number of null hypotheses with $|Z_j| \geq t$ can be estimated by $\pi_0 p \cdot \mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}$, where $\pi_0 p$ is the total number of null hypotheses. For $\pi_0 < 1$, then, the fraction in the definition of (3.2) is an overestimate of the FDP by a factor of $(\pi_0)^{-1}$, see [9] and references therein.

Turning to the problem of variable selection in regression, the BHq procedure may be applied by calculating the least-squares estimate,

$$\hat{\boldsymbol{\beta}}^{\mathsf{LS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

For Gaussian noise as in (1.1), these fitted coefficients follow a $\mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Sigma}^{-1})$ distribution, where we recall that $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$. Therefore, setting

$$Z_j = \frac{\hat{\beta}^{\mathsf{LS}}_j}{\sigma \sqrt{(\boldsymbol{\Sigma}^{-1})_{jj}}}$$

yields z-scores, i.e. marginally $Z_j \sim \mathcal{N}(0,1)$ whenever $\beta_j = 0$. Variables are then selected using the data-dependent threshold given in (3.2).

Under orthogonal designs in which $\boldsymbol{X}^\top \boldsymbol{X}$ is a diagonal matrix, the Z_j's are independent; in this setting, Benjamini and Hochberg [1] prove that the BHq procedure controls FDR at the level $\pi_0 \cdot q$ (see Section 3.4 for a comparison of knockoff methods with BHq in the orthogonal design setting). Without the assumption of independence, however, there is no such guarantee. In fact, it is not hard to construct designs with only two variables such that the BHq procedure does not control the FDR at level q. FDR control has been established for test statistics obeying the positive regression dependence on a subset property (PRDS) introduced in [2]. The problem is that the PRDS property does not hold in our setting, for two reasons. First, for one sided tests where the alternative is $\beta_j > 0$, say, one would reject for large values of Z_j. Now for the PRDS property to hold we would need to have $(\boldsymbol{\Sigma}^{-1})_{ij} \geq 0$ for all nulls i and all j, which rarely is in effect. Second, since the signs of the coefficients β_j are in general unknown, we are performing two-sided tests where we reject for large values of $|Z_j|$ rather than Z_j; these absolute-value statistics are not known to be PRDS either even under positive values of $(\boldsymbol{\Sigma}^{-1})_{ij}$.

Against this background, Benjamini and Yekutieli in [2] show that the BHq procedure yields FDR bounded by $\pi_0 q \cdot S(p)$ regardless of the dependence among the z-scores, where

$$S(p) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{p} \approx \log p + 0.577. \tag{3.3}$$

Therefore, if we define T as in (3.2) but with $q/S(p)$ in place of q, then we are again guaranteed a bound on FDR.

Finally, as another option, we can "whiten the noise" in $\hat{\boldsymbol{\beta}}$ before applying BHq. Specifically, let $\boldsymbol{Z}' \sim \mathcal{N}(0, \sigma^2 \cdot (\lambda_0^{-1} \mathbf{I} - \boldsymbol{\Sigma}^{-1}))$ be drawn independently from the data, where $\lambda_0 = \lambda_{\mathsf{min}}(\boldsymbol{\Sigma})$. Then

$$\hat{\boldsymbol{\beta}} + Z' \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \lambda_0^{-1} \mathbf{I}), \tag{3.4}$$

and we can then apply BHq to the z-scores given by

$$Z_j = \frac{\hat{\boldsymbol{\beta}}_j + Z'_j}{\sigma / \sqrt{\lambda_0}}.$$

Method	FDR (%) (nominal level $q = 20\%$)	Power (%)	Theoretical guarantee of FDR control?
Knockoff+ (equivariant construction)	**14.40**	**60.99**	**Yes**
Knockoff (equivariant construction)	17.82	66.73	No
Knockoff+ (SDP construction)	**15.05**	**61.54**	**Yes**
Knockoff (SDP construction)	18.72	67.50	No
Benjamini-Hochberg (BHq) [1]	18.70	48.88	No
BHq + log-factor correction [2]	**2.20**	**19.09**	**Yes**
BHq with whitened noise	**18.79**	**2.33**	**Yes**

Table 1: FDR and power in the setting of Section 3.3.1 with $n = 3000$ observations, $p = 1000$ variables and $k = 30$ variables in the model with regression coefficients of magnitude 3.5. Bold face font highlights those methods that are known theoretically to control FDR at the nominal level $q = 20\%$.

Since these z-scores are now independent, applying BHq yields an FDR of at most $\pi_0 q$.

For all of the variants of BHq considered here, FDR control is estimated or guaranteed to be at a level of $\pi_0 q$, which is lower than the nominal level q, i.e. the method is more conservative than desired. However, here we are primarily interested in a sparse setting where $\pi_0 \approx 1$, and so this will not have a strong effect.

3.3 Empirical comparisons with the Benjamini-Hochberg method and variants

We now test our method in a range of settings, comparing to BHq and its variants, and examining the effects of sparsity level, signal magnitude, and feature correlation.

3.3.1 Comparing methods

We begin with a comparison of seven methods: the equi-variant and the SDP constructions for the knockoff and knockoff+ filters; the BHq procedure; the BHq procedure with the log-factor correction [2] to guarantee FDR control with dependent z-scores (i.e. this applies BHq with $q/S(p)$ replacing q, as in (3.3)); and the BHq procedure with whitened noise, as in (3.4). To summarize our earlier discussions, the equi-variant and SDP constructions for knockoff+, the BHq method with the log-factor correction, and the BHq method with whitened noise are all guaranteed to control FDR at the nominal level q; the other methods do not offer this exact guarantee.

In this simulation, we use the problem size $n = 3000$, $p = 1000$ and a number $k = 30$ of variables in the model. We first draw $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0, 1)$ entries, then normalize its columns. Next, to define $\boldsymbol{\beta}$, we choose $k = 30$ coefficients at random and choose β_j randomly from $\{\pm A\}$ for each of the k selected coefficients, where $A = 3.5$ is the signal amplitude. Finally, we draw $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta}, \mathbf{I})$. The signal amplitude $A = 3.5$ is selected because 3.5 is approximately the expected value of $\max_{1\leq j\leq p} |\boldsymbol{X}_j^\top \boldsymbol{z}|$ where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (each $\boldsymbol{X}_j^\top \boldsymbol{z}$ is approximately a standard normal variable if $\beta_j = 0$). Setting the signal amplitude to be near this maximal noise level ensures a setting where it is possible, but not trivial, to distinguish signal from noise.

Table 1 displays the resulting FDR and power obtained by each method, averaged over 600 trials. Empirically, all of the methods result in a FDR that is near or below the nominal level $q = 20\%$. Comparing the power, both constructions of knockoff and knockoff+ (power $> 60\%$) significantly outperform BHq (power $\approx 49\%$).

Comparing the equi-variant and SDP constructions for the knockoff and knockoff+ methods, the SDP construction achieves slightly higher power for both knockoff and knockoff+. Finally, the two variants of BHq considered do offer theoretical control of FDR, but empirically achieve very poor power in this simulation. From this point on, then, we restrict our attention to three methods: the knockoff method using the SDP construction given in (2.5), the knockoff+ method with the same SDP construction, and BHq.

3.3.2 Effect of sparsity level

Next, we consider the effect of varying the sparsity level k. We test values $k = 10, 20, 30, \ldots, 200$, while fixing the problem size $n = 3000$, $p = 1000$ and the signal amplitude $A = 3.5$, with data generated as in Section 3.3.1. The mean FDR and mean power over 200 trials are displayed in Figure 3. All three methods successfully control FDR at the nominal level $q = 20\%$; for extremely low values of k, knockoff+ is slightly more conservative (lower FDR) than

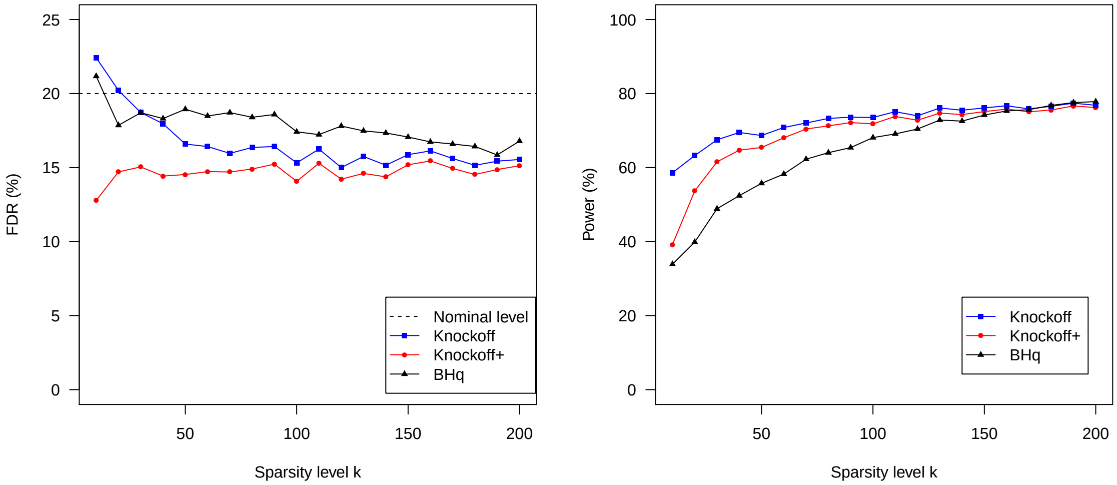


Figure 3: Testing the knockoff, knockoff+, and BHq methods at nominal level $q = 20\%$ with varying sparsity level k. Here $n = 3000$, $p = 1000$, and $A = 3.5$, and the figures show mean FDR and mean power averaged over 600 trials.

the other methods, as is expected from the fact that the total number of selected variables will be quite small when k is low. Comparing the power of the methods, we see that both knockoff and knockoff+ offer a strong advantage over BHq at low and moderate values of k—these methods successfully leverage the sparse structure of the model in this high-dimensional setting. For instance, at the lowest sparsity level $k = 10$, the power of the knockoff method is 58.60% as compared to 33.90% for BHq, while the FDR levels are nearly equal for the two methods. For higher values of k, when the problem is no longer extremely sparse, the power of BHq catches up with the knockoff and knockoff+ methods.

Looking at knockoff+, we see that this filter has higher power than BHq while having a lower type I error. In the language of multiple testing, this says that it detects more true effects while keeping the fraction of false discoveries at a lower level, which makes findings somehow more reproducible. This observation holds throughout all the experiments presented here.

3.3.3 Effect of signal amplitude

We vary the signal amplitude A to compare the three methods in regimes where signals may be easier or harder to discover. For this experiment, we test signal amplitudes $A = 2.8, 2.9, 3.0, \ldots, 4.2$ while fixing $n = 3000$, $p = 1000$, and $k = 30$, with data generated as in Section 3.3.1. The mean FDR and mean power over 200 trials are displayed in Figure 4. All three methods successfully control FDR at the nominal level $q = 20\%$, with noticeably more conservative FDR for knockoff+. Examining the power plot, the power of all three methods increases as the signal amplitude increases, with significantly higher power for the knockoff and knockoff+ methods in comparison to BHq across the entire range of signal amplitudes tested. For example, at low signal amplitude $A = 2.8$, BHq and knockoff have nearly identical FDR, but BHq achieves a power of approximately 20.65% while the power of the knockoff method is 38.28%.

3.3.4 Effect of feature correlation

Finally, we test the effect of strong feature correlations on these three methods. For this experiment, we again set $n = 3000$, $p = 1000$, $k = 30$, and $A = 3.5$, our default settings from before. However, the design matrix $\boldsymbol{X}$ is generated differently, using a tapered correlation structure. For each correlation parameter value $\rho = 0.0, 0.1, \ldots, 0.9$, we generate the rows of $\boldsymbol{X}$ from a $\mathcal{N}(0, \boldsymbol{\Theta}_\rho)$ distribution, where $(\boldsymbol{\Theta}_\rho)_{jk} = \rho^{|j-k|}$ (in the case that $\rho = 0$, we simply set $\boldsymbol{\Theta} = \mathbf{I}$ as before). We then normalize the columns of $\boldsymbol{X}$, and generate $\boldsymbol{\beta}$ and $\boldsymbol{y}$ in the same manner as

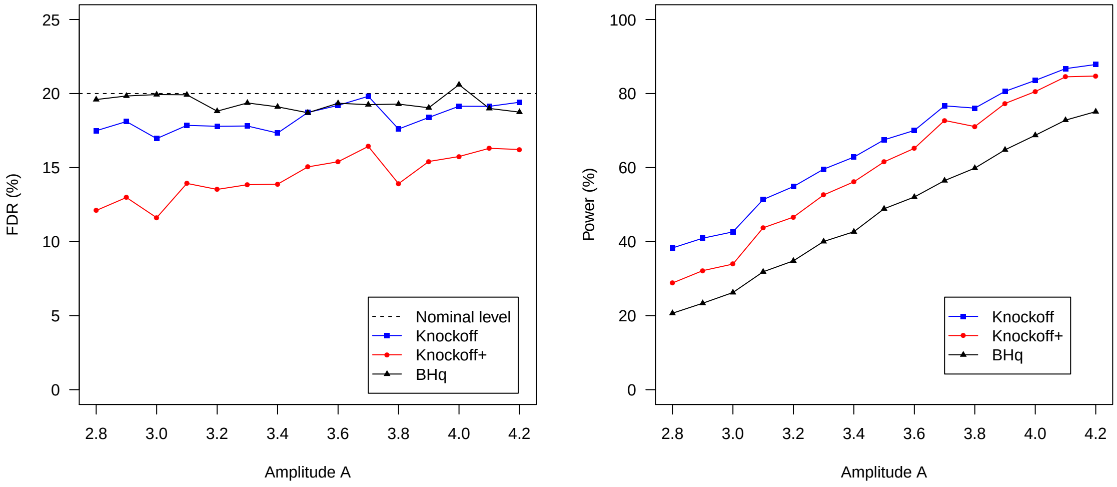


Figure 4: Testing the knockoff, knockoff+, and BHq methods at nominal level $q = 20\%$ with varying signal amplitudes A. Here $n = 3000$, $p = 1000$, and $k = 30$, and the figures show mean FDR and mean power averaged over 200 trials.

in Section 3.3.1. The mean FDR and mean power over 200 trials are displayed in Figure 5. The knockoff+ and BHq methods successfully control FDR at the nominal level $q = 20\%$ across all values of ρ. The knockoff method controls FDR for $\rho \leq 0.8$, but shows a higher FDR level of 26.67% when $\rho = 0.9$. This is consistent with our theoretical result Theorem 1, which guarantees that the knockoff method controls a modified form of the FDR that is very similar when a high number of variables are selected, but may be quite different when a small number of variables is selected (at $\rho = 0.9$, the extreme correlations between variables lead to a small number of selections from all of the three methods). Turning to the power plot, overall there is a sharp decay in power for all three methods as ρ increases, reflecting the difficulty of telling apart neighboring features that are strongly correlated. The knockoff and knockoff+ methods both achieve significantly higher power than BHq across the range of ρ values.

3.4 Relationship with the Benjamini-Hochberg procedure under orthogonal designs

The Benjamini-Hochberg (BHq) procedure is known to control FDR in the setting where the statistics for the null hypotheses are mutually independent. In the regression setting where our statistics arise from the least-squares coefficients $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$ (as in Section 3.2), the coefficients of $\hat{\boldsymbol{\beta}}$ are mutually independent if and only if $\boldsymbol{X}^\top \boldsymbol{X}$ is a diagonal matrix—the orthogonal design setting. In this section, we consider the orthogonal designs and compare the knockoff filter and BHq side-by-side to understand the similarities and differences in how these methods work. We will find that

1. The two methods both control FDR (as guaranteed by the theory), and achieve nearly identical power over a range of signal amplitudes.

2. Theoretically and empirically, regardless of signal amplitude, the FDR of BHq is given by $\pi_0 q$, where $q = 20\%$ is the nominal FDR level and π_0 is the proportion of null hypotheses, $\pi_0 = \frac{p-k}{p}$, as is shown in [1].

3. In contrast, the FDR of the knockoff method varies over the range of signal amplitudes. When the signal amplitude is high enough for the power to be substantial, the FDR of the knockoff method approaches q, rather than $\pi_0 q$—that is, the knockoff method is implicitly correcting for the proportion of nulls, and achieving the target FDR level. When the signal amplitude is so low that power is near zero, the knockoff method has an extremely low FDR (far lower than the nominal level q), which is desirable in a regime where we have little chance of finding the true signals.

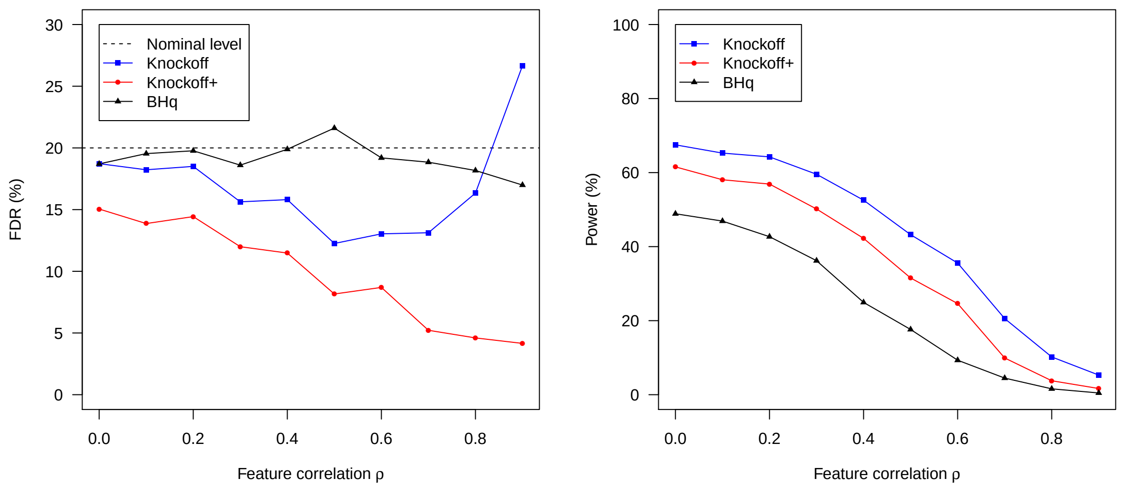
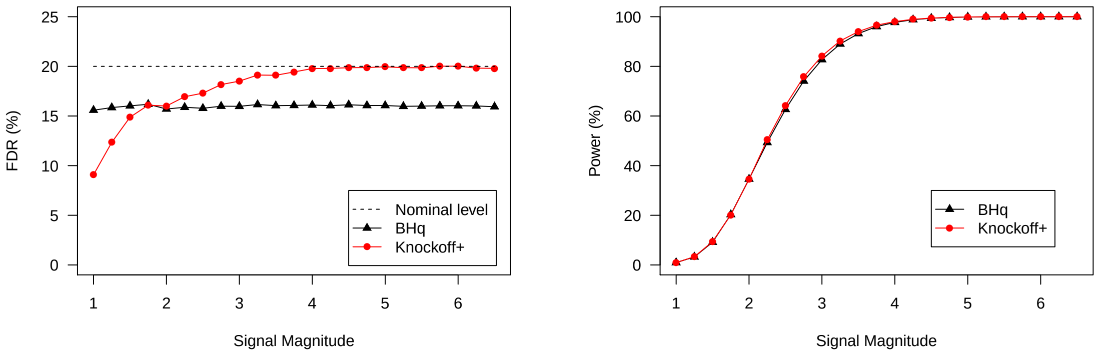


Figure 5: Testing the knockoff, knockoff+, and BHq methods at nominal level $q = 20\%$ with varying feature correlation levels. The correlation parameter ρ controls the tapered correlation structure of the design matrix, where columns $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are generated from a distribution with correlation $\rho^{|j-k|}$. Here $n = 3000$, $p = 1000$, $k = 30$, $A = 3.5$, and the figures show mean FDR and mean power averaged over 200 trials.

Figure 6 demonstrates this comparison empirically over a range of signal amplitude levels—note the contrasting FDR behavior between the two methods, even though the power is essentially identical in each setting.



(a) FDR

(b) Power

Figure 6: FDR and power of the BHq and knockoff+ methods, plotted against the size A of the regression coefficients (signal magnitude), averaged over 1000 trials. The nominal FDR level q is set to 20%. The 2000×1000 design $\boldsymbol{X}$ is orthogonal, and the number of true signals is 200 so that the fraction of nulls is $\pi_0 = 0.8$.

Next, we sketch a theoretical explanation for the different behaviors of the two methods. In order to get independent statistics, we work with a $2p \times p$ orthogonal design $\boldsymbol{X}$ and set the noise level $\sigma = 1$ without loss of generality. In this setting, $\boldsymbol{X}^\top \boldsymbol{y} := \boldsymbol{\beta} + \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\beta}, \mathbf{I})$ is the maximum likelihood estimate for the regression coefficients. Recall that the BHq procedure selects variables $\boldsymbol{X}_j$ with $|\boldsymbol{X}_j^\top \boldsymbol{y}| \geq T$ and

$$T = \min \left\{ t : t = +\infty \text{ or } \frac{p \cdot \mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}}{\#\{j : |\boldsymbol{X}_j^\top \boldsymbol{y}| = |\beta_j + z_j| \geq t\}} \leq q \right\}, \tag{3.5}$$

where the fraction above is of course the estimate of $\mathsf{FDP}(t)$. In fact, the number of null features whose statistic exceeds t will be roughly $\pi_0 p \cdot \mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}$ (since $\pi_0 p$ is the total number of null features), and so the fraction appearing in (3.5) above overestimates $\mathsf{FDP}(t)$ by a factor of $(\pi_0)^{-1}$, leading to FDR control at the level $\pi_0 q$ rather than at the nominal level q.

Next we study the behavior of the knockoff+ procedure in the same setting. By construction, both the equi-correlated and the SDP constructions must obey $\tilde{\boldsymbol{X}}^\top \boldsymbol{X} = \boldsymbol{0}$ and $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \mathbf{I}$. It follows that $\tilde{\boldsymbol{X}}^\top \boldsymbol{y} := \boldsymbol{z}'$ is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and is independent from $\boldsymbol{X}^\top \boldsymbol{y}$. Hence, our method specialized to (1.7) yields test statistics of the form

$$W_j = |\beta_j + z_j| \vee |z'_j| \cdot \operatorname{sign}(|\beta_j + z_j| - |z'_j|),$$

and our estimated knockoff+ FDP is equal to

$$\widehat{\mathsf{FDP}}(t) = \frac{1 + \#\{j : |z'_j| \geq t \text{ and } |z'_j| > |\beta_j + z_j|\}}{\#\{j : |\beta_j + z_j| \geq t \text{ and } |\beta_j + z_j| > |z'_j|\}}. \tag{3.6}$$

Now we consider the behavior of this estimated FDP under varying signal magnitude levels. Since there are $\pi_0 p$ null features, and since $|\beta_j| = A$ for the non-nulls, the expected value of the numerator in (3.6) is given by

$$\begin{aligned} &1 + \pi_0 p \cdot \mathbb{P}\left\{|z'| \geq t \text{ and } |z'| > |z|\right\} + (1-\pi_0)p \cdot \mathbb{P}\left\{|z'| \geq t \text{ and } |z'| > |A+z|\right\} \\ &\quad = 1 + \underbrace{\pi_0 p \cdot \mathbb{P}\left\{|z| \geq t\right\}\left(1 - \frac{1}{2}\mathbb{P}\left\{|z| \geq t\right\}\right)}_{\text{(Term 1)}} + \underbrace{(1-\pi_0)p \cdot \mathbb{P}\left\{|z'| \geq t \text{ and } |z'| > |A+z|\right\}}_{\text{(Term 2)}}. \end{aligned} \tag{3.7}$$

Consider a large value t. When signal amplitude is high, e.g. $A = 4$, then (Term 1) is the dominant term above, and is roughly equal to $\pi_0 p \cdot \mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}$. In this regime, we see that our procedure resembles BHq (3.5) but with a numerator adjusted to $\pi_0 p \cdot \mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}$ so that it controls the FDR nearly at the nominal level q, instead of $\pi_0 q$.

In contrast, if we consider a weak signal signal magnitude such as $A = 1$, then (Term 2) is no longer vanishing in (3.7) above—that is, the numerator in our FDP estimate may be inadvertently counting non-null features. (In the notation of Section 1.2, we may have non-null j where $W_j < 0$, due to the weakness of the signal.) In this setting, the resulting FDR of the knockoff+ method is conservative, i.e. lower than the nominal level q. However, there is no power loss relative to BHq (see Figure 6); in this low-signal regime, the distribution of the non-null statistics is very close to the null distribution and one simply cannot get any power while controlling a type I error.

3.5 Other methods

Finally, we briefly mention several other approaches that are related to the goal of this work. First, we discuss two methods presented Miller [14, 15] to control false positives in forward selection procedures for linear regression. The first method proposes creating "dummy" variables whose entries are drawn i.i.d. at random. The forward selection procedure is then applied to the augmented list of variables, and we stop when the procedure selects a dummy variable for the first time. This approach is similar in flavor to our proposed methods, but the construction of the dummy variables does not account for correlation among the existing features, and therefore may lose FDR control in a correlated setting. Whereas we do not mind a small fraction of false discoveries to gain power, Miller's approach targets a more stringent type I error control, namely, the familywise error rate (FWER), which aims to have zero false positives.

Miller [15] also proposes a second method, which makes use of a key observation that we also use extensively in our work: after selecting m variables $\boldsymbol{X}_{i_1}$ through $\boldsymbol{X}_{i_m}$, if all of the true features have already been selected, then for any rotation $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ that acts as the identity on the span of $\{\boldsymbol{X}_{i_1}, \dots, \boldsymbol{X}_{i_m}\}$, we have $\boldsymbol{y} \overset{d}{=} \boldsymbol{R}\boldsymbol{y}$ since the part of $\boldsymbol{y}$ that is orthogonal to the signal is simply Gaussian noise and, therefore, rotationally invariant. To test whether an additional $(m+1)$st variable should be included, [15] thus proposes comparing to the null distribution obtained by applying randomly chosen rotations $\boldsymbol{R}$. It is important to note, however, that this null distribution $\boldsymbol{R}$ is not valid if there are some true features that have not yet been selected. That is, if any true features are missing from $\{i_1, \dots, i_m\}$, then we might select a null feature $\boldsymbol{X}_{i_{m+1}}$ because the calculation of the null distribution is not correct. In practice, true features and null features are nearly always interspersed in the forward selection steps, and so this type of method will not achieve exact control of the FDR for this reason.

Next, we turn to recent work by G'Sell et al [10], which also gives a FDR-controlling procedure for the Lasso, without constructing additional variables, by making use of the sequence of values of λ at which new variables enter the model, whose distribution was studied in [12, 21]. Specifically, [12, 21] study the distribution of the sequence of λ values where new null variables enter the model, after all of the true signals have already been included in the model. Consequently, the sequential testing procedure of [10] controls FDR under an important assumption: the true features must all enter the model before any of the null features. Therefore, this method faces the same difficulty as the second method of Miller [15] discussed above; when signals and null features are interspersed along the Lasso path (as is generally the case in practice even when the nonzero regression coefficients are quite large), this assumption is no longer satisfied, leading to some increase of the FDR.

Finally, the stability selection approach [11, 13] controls false variable selection in the Lasso by refitting the Lasso model repeatedly for subsamples of the data, and then keeps only those variables that appear consistently in the Lasso model across most of the subsamples. These methods control false discoveries effectively in practice, and give theoretical guarantees of asymptotically consistent model selection. For a finite-sample setting, however, there is no known concrete theoretical guarantee for controlling false discoveries (with the exception of Theorem 1 in [13], which treats the specific setting where the null variables are exchangeable, i.e. in a linear model, this would be the exactly equi-variant setting where $\Sigma_{ij} = \rho$ for all $i \neq j$, for some fixed ρ). Furthermore, these methods require computing the path of Lasso models for many subsampled regressions containing p candidate variables each; in contrast, our method requires only a single computation of the Lasso path, although for a model with $2p$ variables.

4 Experiment on Real Data: HIV Drug Resistance

We apply the knockoff filter to the task of detecting mutations in the Human Immunodeficiency Virus Type 1 (HIV-1) that are associated with drug resistance. The data set, described and analyzed in [18],[6] consists of drug resistance measurements and genotype information from samples of HIV-1, with separate data sets for resistance to protease inhibitors (PIs), to nucleoside reverse transcriptase (RT) inhibitors (NRTIs), and to nonnucleoside RT inhibitors (NNRTIs). The data set sizes are:

Drug type	# drugs	Sample size	# protease or RT positions genotyped	# mutations appearing ≥ 2 times in sample
PI	6	848	99	242
NRTI	6	639	240	382
NNRTI	3	747	240	407

In each drug class, some samples are missing resistance measurements for some of the drugs, so for each drug's analysis, the sample size and the number of mutations present are slightly smaller than given in the table; we report the final n and p for each drug in Figures 7, 8, and 9 on a case-by-case basis.

We analyse each drug separately. The response y_i is given by the log-fold-increase of lab-tested drug resistance in the ith sample, while the design matrix $\boldsymbol{X}$ has entries $X_{ij} \in \{0,1\}$, indicating presence or absence of mutation #j in the ith sample (for each drug, we keep only those mutations appearing ≥ 2 times in the sample for that drug). Different mutations at the same position are treated as distinct features, and we assume an additive linear model with no interactions. To ensure that the Gaussian linear model assumption is not unreasonable, we check that the residuals of $\boldsymbol{y}$ after a full linear regression on $\boldsymbol{X}$ are approximately normally distributed. We then apply knockoff and BHq, each with $q = 20\%$, to the resulting data set. Most of the drugs have a sample size n with $p < n < 2p$ (specifically, all of the NRTI and NNRTI drugs, and one of the PI drugs; see Figures 7, 8, and 9 for the values of n and p). To apply the knockoff method for these data sets where $p < n < 2p$, we use the method described in Section 2.1.2 which extends the knockoff method beyond the original construction for the $n \geq 2p$ regime.[7]

To evaluate the results, we compare the selected mutations with existing treatment-selected mutation (TSM) panels [17]; since this is a real data experiment, the ground truth is unknown, but these panels provide a good approximation that we can use to assess the methods. For each drug class (PIs, NRTIs, NNRTIs), Rhee et al [17] create panels of mutations that are present at significantly higher frequency (after correcting for multiple comparisons) in virus samples from individuals who have been treated with that class of drug, as compared to individuals that have never received

[6] Data available online at `http://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/`

[7] Although the Gram matrix $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$ is not invertible in some cases due to rare mutations, the knockoff construction and results still hold with $\boldsymbol{\Sigma}^+$ (the Moore-Penrose pseudoinverse of $\boldsymbol{\Sigma}$) in place of $\boldsymbol{\Sigma}^{-1}$.

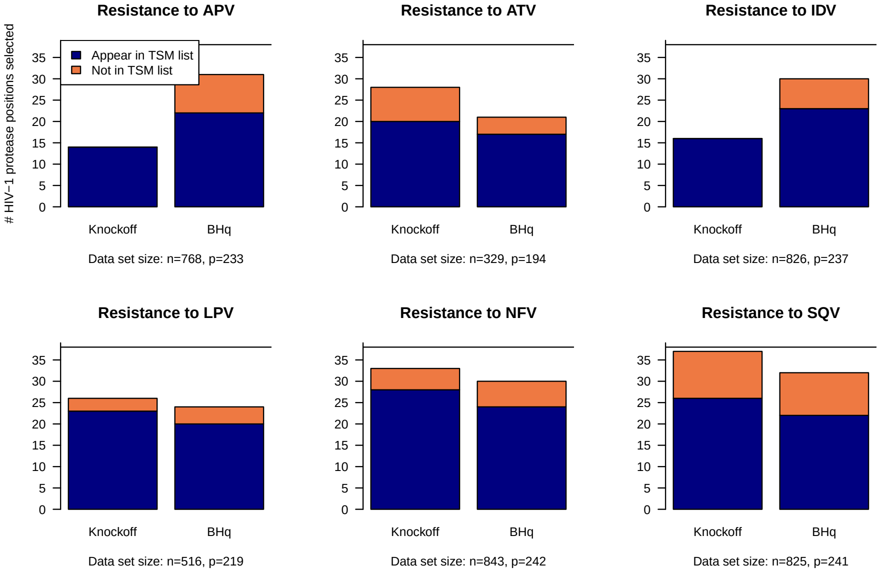


Figure 7: Results of applying the knockoff filter and BHq with $q = 20\%$ to model PI-type drug resistance of HIV-1 based on genetic mutations using data from [18]. For each PI-type treatment and for each of the three methods, the bar plots show the number of positions on the HIV-1 protease where mutations were selected. (We do not include the PI-type drug RTV as it is typically prescribed as a booster of other drugs in the PI class.) To validate the selections of the methods, dark blue indicates protease positions that appear in the treatment-selected mutation (TSM) panel for the PI class of treatments, given in Table 1 of [17], while orange indicates positions selected by the method that do not appear in the TSM list. The horizontal line indicates the total number of HIV-1 protease positions appearing in the TSM list. Note that the TSM list consists of mutations that are associated with the PI class of drugs in general, and is not specialized to the individual drugs in the class.

that class of drug. Therefore the data we use for model selection (based on lab-tested drug resistance) and the mutation panels used to validate our results (based on association with patient treatment history)[8] come from different types of studies, and we aim to see replicability; that is, we will evaluate our model selection results based on how many of the mutations identified by our analysis appear also in the TSM lists. It is known that multiple mutations at the same protease or RT position can often be associated with related drug-resistance outcomes. . Since the TSM lists are an approximation of the ground truth, we will compare only the positions of the mutations selected, with the positions of mutations on the TSM lists.

Results for the PI, NRTI, and NNRTI type drugs are displayed in Figures 7, 8, and 9, respectively. We see that both methods perform similarly for most of the drugs in the three classes, with good agreement in most cases between the positions of the selected mutations based on the lab-tested drug resistance data, and the TSM lists which are based on patient history data. Overall, the knockoff filter shows slightly better agreement with the TSM lists as compared to BHq, but there is variability in the outcomes across the different drugs. In summary we see that the FDR-controlling knockoff methods indeed select variables that mostly correspond to real (replicable) effects, as verified by the independently-created TSM lists.

[8] The reported treatment-selected mutations in [17] also include several mutations that did not pass the significance threshold after correcting for multiple comparisons, but were located at protease or RT positions known to be associated with drug resistance based on in vitro testing. We removed these positions from the lists before comparing to our results, so that we are validating our results with a list created only based on association with patient treatment history.

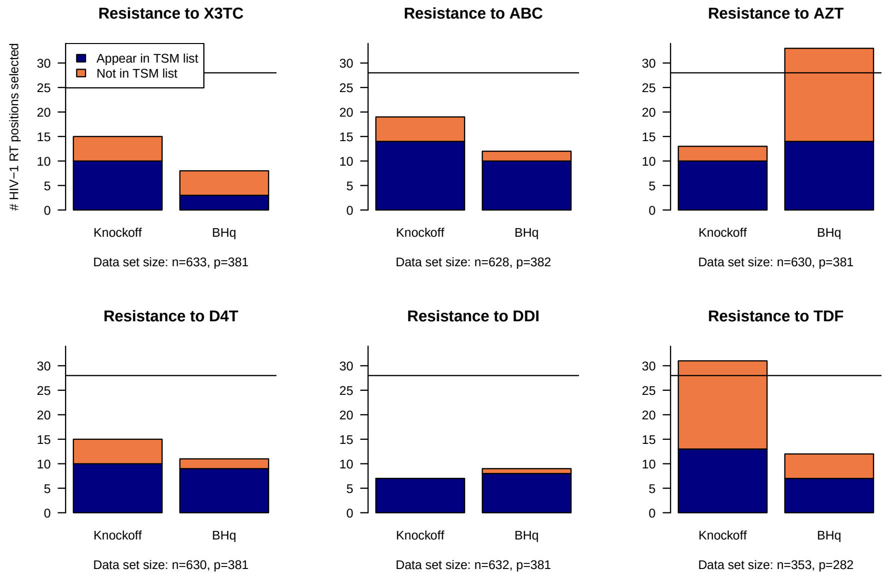


Figure 8: Same as Figure 7, but for the NRTI-type drugs, validating results agains the treatment-selected mutation (TSM) panel for NRTIs given in Table 2 of [17].

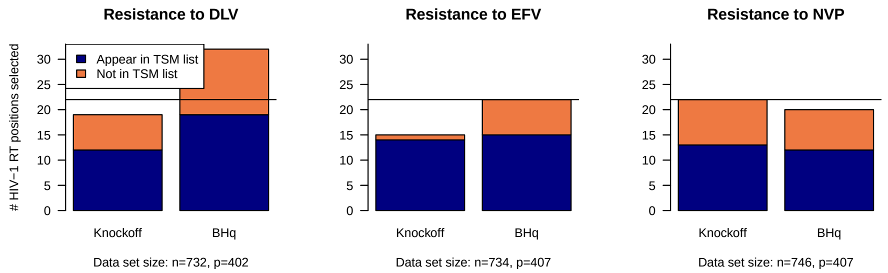


Figure 9: Same as Figure 7, but for the NNRTI-type drugs, validating results agains the treatment-selected mutation (TSM) panel for NNRTIs given in Table 3 of [17].

5 Sequential Hypothesis Testing

5.1 Two sequential testing procedures

In this section, we describe several related sequential hypothesis testing procedures, along with theoretical results for FDR control. We then relate these procedures to the knockoff and knockoff+ methods, in order to prove our main results Theorems 1 and 2.

Imagine that $p_1, \dots, p_m$ are p-values giving information about hypotheses $H_1, \dots, H_m$. These p-values obey $p_j \overset{d}{\geq} \mathsf{Unif}[0,1]$ for all null j; that is, for all null j and all $u \in [0,1]$, $\mathbb{P}\left\{p_j \leq u\right\} \leq u$. We introduce two sequential strategies, which control the FDR at any fixed level q under a usual independence property.

- **First sequential testing procedure (FSTP).** Fix any threshold $c \in (0,1)$ and any subset[9] K, and define

$$\hat{k}_0 = \max\left\{k \in K : \frac{\#\{j \leq k : p_j > c\}}{k \vee 1} \leq (1-c)\cdot q\right\},$$

[9] In many applications we would typically choose $K = [m]$, but we allow for $K \subsetneq [m]$ to help with the proof of the regression method.

and

$$\hat{k}_1 = \max\left\{k \in K : \frac{1 + \#\{j \leq k : p_j > c\}}{1+k} \leq (1-c)\cdot q\right\},$$

with the convention that we set $\hat{k}_{0/1} = 0$ if the set is empty: here $\hat{k}_{0/1}$ should be read as "$\hat{k}_0$ or $\hat{k}_1$", since we can choose which of the two definitions above to use. We then reject H_j for all $j \leq \hat{k}_{0/1}$, and thus get two distinct procedures named FSTP0 and FSTP1 hereafter.

To understand why such a sequential procedure makes sense, consider FSTP0 and assume that the null p-values are i.i.d. $\mathsf{Unif}[0,1]$. Then

$$\frac{\#\{\text{null } j \leq k\}}{k \vee 1} \approx \frac{1}{1-c}\cdot\frac{\#\{\text{null } j \leq k : p_j > c\}}{k \vee 1} \leq \frac{1}{1-c}\cdot\frac{\#\{j \leq k : p_j > c\}}{k \vee 1}$$

so that again, the procedure maximizes the number of rejections under the constraint that an estimate of FDR is controlled at level q. FSTP1 corrects FSTP0 to guarantee FDR control.

- **Second sequential testing procedure (SSTP).** Alternatively, define

$$\hat{k}_{0/1} = \max\left\{k \in K : \frac{0/1 + \#\{j \leq k : p_j > c\}}{\#\{j \leq k : p_j \leq c\} \vee 1} \leq \frac{1-c}{c}\cdot q\right\},$$

with the convention that we set $\hat{k}_{0/1} = 0$ if this set is empty. (We get two distinct procedures named SSTP0 and SSTP1 by letting the term in the numerator be 0 or 1.) Then reject H_j for all $j \leq \hat{k}_{0/1}$ such that $p_j \leq c$. Admittedly, these are not sequential testing procedures stricto senso and we are thus abusing terminology.

Again, to understand this procedure intuitively when the null p-values are i.i.d. $\mathsf{Unif}[0,1]$, we see that

$$\frac{\#\{\text{null } j \leq k : p_j \leq c\}}{\#\{j \leq k : p_j \leq c\} \vee 1} \approx \frac{c}{1-c}\cdot\frac{\#\{\text{null } j \leq : p_j > c\}}{\#\{j \leq k : p_j \leq c\} \vee 1} \leq \frac{c}{1-c}\cdot\frac{\#\{j \leq k : p_j > c\}}{\#\{j \leq k : p_j \leq c\} \vee 1}$$

so that again, the procedure maximizes the number of rejections under the constraint that an estimate of FDR is controlled at level q.

Theorem 3. *Suppose that the null p-values are i.i.d. with $p_j \geq \mathsf{Unif}[0,1]$, and are independent from the non-nulls. For each procedure considered, let V be the number of false discoveries and R the total number of discoveries.*

- *Both FSTP1 and SSTP1 control the FDR, i.e.*

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right] \leq q.$$

- *SSTP0 controls a modified FDR, i.e.*

$$\mathbb{E}\left[\frac{V}{R + \frac{c}{1-c}q^{-1}}\right] \leq q.$$

When $c = 1/2$, it therefore controls $\mathbb{E}\left[V/(R + q^{-1})\right]$.

- *FSTP0 also controls a modified FDR, i.e.*

$$\mathbb{E}\left[\frac{V}{R + \frac{1}{1-c}q^{-1}}\right] \leq q.$$

As is clear from the assumption, the order of the p-values cannot be dependent on the p-values themselves—for instance, we cannot reorder the p-values from smallest to largest, apply this procedure and expect FDR control.

5.2 Connection with knockoffs

Interestingly, the knockoff method can be cast as a special case of the second sequential hypothesis testing procedure, and the FDR controlling properties are then just a consequence of Theorem 3. We explain this connection, thereby proving Theorems 1 and 2.

Let $m = \#\{j : W_j \neq 0\}$; since our method never selects variable j when $W_j = 0$, we can ignore such variables. Assume without loss of generality that $|W_1| \geq |W_2| \geq \cdots \geq |W_m| > 0$, and set

$$p_j = \begin{cases} 1/2, & W_j > 0, \\ 1, & W_j < 0, \end{cases}$$

which can be thought of as 1-bit p-values. It then follows from Lemma 1 that the null p-values are i.i.d. with $\mathbb{P}\{p_j = 1/2\} = 1/2 = \mathbb{P}\{p_j = 1\}$ and are independent from the others, thereby obeying the assumptions of Theorem 3. Setting K to be the indices of the strict inequalities,

$$K = \{k \in [m] : |W_k| > |W_{k+1}|\} \cup \{m\},$$

one sees that the correlation-preserving proxy method is now equivalent to the second sequential testing procedure on these p-values. To see why this true, set $c = 1/2$ and observe that for any $k \in K$,

$$\frac{0/1 + \#\{j \leq k : p_j > 1/2\}}{\#\{j \leq k : p_j \leq 1/2\} \vee 1} = \frac{0/1 + \#\{j \leq k : W_j < 0\}}{\#\{j \leq k : W_j > 0\} \vee 1} = \frac{0/1 + \#\{j : W_j \leq -|W_k|\}}{\#\{j : W_j \geq |W_k|\} \vee 1}.$$

The first equality follows from the definition of p_j. The second equality holds because the absolute values of $\boldsymbol{W}$ are arranged in nonincreasing order; by definition of K, the inequality $|W_j| \geq |W_k|$ is only possible if it holds that $j \leq k$. Therefore $W_j \geq |W_k|$ (respectively, $W_j \leq -|W_k|$) is true if and only if $j \leq k$ and $W_j > 0$ (respectively, $j \leq k$ and $W_j < 0$). Hence, finding the largest k such that the ratio in the left-hand side is below q is the same as finding the smallest $|W_k|$ such that the right-hand side is below q. This is equivalent to finding the minimum $t \in \mathcal{W}$ such that

$$\frac{0/1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q,$$

which are the knockoff and knockoff+ thresholds. Finally, rejecting the p-values obeying $p_j \leq 1/2$ is the same as rejecting the positive W_j's. FDR control follows by applying Theorem 3.

6 Proofs

This section gives a proof of Theorem 3 for both procedures. We work with $K = \{1, \ldots, m\}$ as the proof for an arbitrary $K \subsetneq \{1, \ldots, m\}$ is identical.

6.1 Martingales

In [20], the authors offered a new and elegant proof of the FDR controlling property of the BHq procedure based on a martingale argument. While our argument is different, it also uses martingale theory.

Lemma 4 (Martingale process). *For $k = m, m-1, \ldots, 1, 0$, put $V^+(k) = \#\{\text{null } j : 1 \leq j \leq k, p_j \leq c\}$ and $V^-(k) = \#\{\text{null } j : 1 \leq j \leq k, p_j > c\}$ with the convention that $V^\pm(0) = 0$. Let $\mathcal{F}_k$ be the filtration defined by knowing all the non-null p-values, as well as $V^\pm(k')$ for all $k' \geq k$. Then the process*

$$M(k) = \frac{V^+(k)}{1 + V^-(k)}$$

is a super-martingale running backward in time with respect to $\mathcal{F}_k$. For any fixed q, $\hat{k}$ defined as in either sequential testing procedure is a stopping time, and as a consequence

$$\mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j \leq c\}}{1 + \#\{\text{null } j \leq \hat{k} : p_j > c\}}\right] \leq \frac{c}{1-c}. \tag{6.1}$$

Proof. Note that the filtration $\mathcal{F}_k$ informs us about whether k is null or not, since the non-null process is known exactly. On the one hand, if k is non-null, then $M(k-1) = M(k)$. On the other, if k is null, then

$$M(k-1) = \frac{V^+(k) - I}{1 + V^-(k) - (1-I)} = \frac{V^+(k) - I}{(V^-(k) + I) \vee 1}, \qquad \text{where } I = \mathbb{1}_{p_k \le c}.$$

The event $\mathcal{F}_k$ gives no further knowledge about I, and it follows from the exchangeability property of the nulls—they are i.i.d. and thus exchangeable—that $\mathbb{P}\{I = 1\} = V^+(k)/(V^+(k) + V^-(k))$. Thus in the case where k is null,

$$\mathbb{E}\left[M(k-1)|\mathcal{F}_k\right] = \frac{1}{V^+(k) + V^-(k)}\left[V_+(k)\frac{V^+(k)-1}{V^-(k)+1} + V^-(k)\frac{V^+(k)}{V^-(k) \vee 1}\right] = \begin{cases} \frac{V^+(k)}{1+V^-(k)}, & V^-(k) > 0, \\ V^+(k) - 1, & V^-(k) = 0. \end{cases}$$

In summary,

$$\mathbb{E}\left[M(k-1)|\mathcal{F}_k\right] = \begin{cases} M(k), & k \text{ non null}, \\ M(k), & k \text{ null and } V^-(k) > 0, \\ M(k) - 1, & k \text{ null and } V^-(k) = 0, \end{cases}$$

which shows that $\mathbb{E}\left[M(k-1)|\mathcal{F}_k\right] \le M(k)$. This establishes the super-martingale property.

Now $\hat{k}$ is a stopping time with respect to the backward filtration $\{\mathcal{F}_k\}$ since $\{\hat{k} \ge k\} \in \mathcal{F}_k$. The last assertion (6.1) follows from the optimal stopping time theorem for super-martingales which states that

$$\mathbb{E}M(\hat{k}) \le \mathbb{E}M(m) = \mathbb{E}\left[\frac{\#\{\text{null } j : p_j \le c\}}{1 + \#\{\text{null } j : p_j > c\}}\right].$$

Set $X = \#\{\text{null } j : p_j \le c\}$. The independence of the nulls together with the stochastic dominance $p_j \overset{d}{\ge} \mathsf{Unif}[0,1]$ valid for all nulls imply that $X \overset{d}{\le} Y$, where $Y \sim \mathsf{Binomial}(N, c)$, and N is the total number of nulls. Further, since the function $x \mapsto x/(1 + N - x)$ is nondecreasing, we have

$$\mathbb{E}\left[\frac{X}{1+N-X}\right] \le \mathbb{E}\left[\frac{Y}{1+N-Y}\right] \le \frac{c}{1-c},$$

where the last step is proved as follows:

$$\begin{aligned} \mathbb{E}\left[\frac{Y}{1+N-Y}\right] &= \mathbb{E}\left[\frac{Y}{1+N-Y} \cdot \mathbb{1}_{Y>0}\right] \\ &= \sum_{i=1}^{N} \mathbb{P}\{Y = i\} \cdot \frac{i}{1+N-i} \\ &= \sum_{i=1}^{N} c^i (1-c)^{N-i} \cdot \frac{N!}{i!(N-i)!} \cdot \frac{i}{1+N-i} \\ &= \frac{c}{1-c} \sum_{i=1}^{N} c^{i-1}(1-c)^{N-i+1} \cdot \frac{N!}{(i-1)!(N-i+1)!} \\ &= \frac{c}{1-c} \sum_{i=1}^{N} \mathbb{P}\{Y = i-1\} \\ &\le \frac{c}{1-c}. \end{aligned}$$

□

6.2 Proof of Theorem 3 for SSTP

Recall that $V = \#\{\text{null } j \le \hat{k} : p_j \le c\}$ and $R = \#\{j \le \hat{k} : p_j \le c\}$. For SSTP1, write $\hat{k} = \hat{k}_1$, and proceed as in Section 2.4:

$$\begin{aligned}
\mathbb{E}\left[\frac{V}{R \vee 1}\right] &= \mathbb{E}\left[\frac{V}{R \vee 1} \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&= \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}} \cdot \left(\frac{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}}{\#\{j \le \hat{k} : p_j \le c\} \vee 1} \cdot \mathbb{1}_{\hat{k}>0}\right)\right] \\
&\le \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}}\right] \cdot \frac{1-c}{c} \cdot q \\
&\le q,
\end{aligned}$$

where the first inequality applies the definition of $\hat{k}$ to bound the quantity in parentheses, and the second inequality applies (6.1) from Lemma 4. Similarly, consider SSTP0 and set $\hat{k} = \hat{k}_0$. Then

$$\begin{aligned}
\mathbb{E}\left[\frac{V}{\frac{c}{1-c}q^{-1} + R}\right] &= \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}} \cdot \frac{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}}{\frac{c}{1-c}q^{-1} + R}\right] \\
&\le \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}} \cdot \frac{1 + \frac{1-c}{c} \cdot qR}{\frac{c}{1-c} + qR}\right] \cdot q \\
&= \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}}\right] \cdot \frac{1-c}{c} \cdot q \\
&\le q,
\end{aligned}$$

where the first inequality applies the definition of $\hat{k}$, which gives $\#\{\text{null } j \le \hat{k} : p_j > c\} \le \frac{1-c}{c} \cdot qR$, and the last inequality applies Lemma 4. This proves Theorems 1 and 2.

6.3 Proof of Theorem 3 for FSTP

Consider FSTP1 first and set $\hat{k} = \hat{k}_1$. Then with

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right] = \mathbb{E}\left[\frac{V}{R \vee 1} \cdot \mathbb{1}_{\hat{k}>0}\right] = \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k}\}}{\hat{k} \vee 1} \cdot \mathbb{1}_{\hat{k}>0}\right], \quad \text{FDP}_+(\hat{k}) = \frac{1 + \#\{\text{null } j \le \hat{k}\}}{1 + \hat{k}},$$

we have

$$\begin{aligned}
\mathbb{E}\left[\frac{V}{R \vee 1}\right] &\le \mathbb{E}\left[\frac{1 + \#\{\text{null } j \le \hat{k}\}}{1 + \hat{k}} \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&= \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \hat{k}} \cdot \mathbb{1}_{\hat{k}>0}\right] + \mathbb{E}\left[\frac{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}}{1 + \hat{k}} \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&= \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}} \cdot \text{FDP}_+(\hat{k}) \cdot \mathbb{1}_{\hat{k}>0}\right] + \mathbb{E}\left[\text{FDP}_+(\hat{k}) \cdot \mathbb{1}_{\hat{k}>0}\right] \\
&\le \mathbb{E}\left[\frac{\#\{\text{null } j \le \hat{k} : p_j \le c\}}{1 + \#\{\text{null } j \le \hat{k} : p_j > c\}}\right] \cdot (1-c) \cdot q + (1-c) \cdot q \\
&\le \frac{c}{1-c} \cdot (1-c) \cdot q + (1-c) \cdot q = q,
\end{aligned}$$

where again the next-to-last inequality applies the definition of $\hat{k}$ and the last inequality applies Lemma 4. Moving to FSTP0 and setting $\hat{k} = \hat{k}_0$, write

$$\begin{aligned}
\mathbb{E}\left[\frac{V}{\frac{1}{1-c}q^{-1} + R}\right] &= \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k}\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&= \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j \leq c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] + \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j > c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&= \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j \leq c\}}{1 + \#\{\text{null } j \leq \hat{k} : p_j > c\}} \cdot \frac{1 + \#\{\text{null } j \leq \hat{k} : p_j > c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] + \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j > c\}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&\leq \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j \leq c\}}{1 + \#\{\text{null } j \leq \hat{k} : p_j > c\}} \cdot \frac{1 + (1-c) \cdot q\hat{k}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] + \mathbb{E}\left[\frac{(1-c) \cdot q\hat{k}}{\frac{1}{1-c}q^{-1} + \hat{k}}\right] \\
&\leq \mathbb{E}\left[\frac{\#\{\text{null } j \leq \hat{k} : p_j \leq c\}}{1 + \#\{\text{null } j \leq \hat{k} : p_j > c\}}\right] \cdot (1-c) \cdot q + (1-c) \cdot q \\
&\leq \frac{c}{1-c} \cdot (1-c) \cdot q + (1-c) \cdot q = q,
\end{aligned}$$

where once more the first inequality applies the definition of $\hat{k}$ and the last inequality applies Lemma 4.

7 Discussion

In this paper, we have proposed two variable selection procedures, knockoff and knockoff+, that control FDR in the linear regression setting and offer high power to discover true signals. We give theoretical results showing that these methods maintain FDR control under arbitrary feature correlations, even when variable selection methods such as the Lasso may select null variables far earlier than some of the weaker signals. The empirical performance of knockoff and knockoff+ demonstrates effective FDR control and excellent power in comparison to other methods such as the Benjamini-Hochberg procedure (BHq) or permutation-based methods.

A key ingredient in the knockoff and knockoff+ methods is the "one-bit p-values" obtained by comparing feature $\boldsymbol{X}_j$ with its knockoff feature $\tilde{\boldsymbol{X}}_j$, and recording which of the two was first to enter the Lasso path. This extreme discretization may be part of the reason that the methods are conservative under low signal amplitude, and could potentially be addressed by creating multiple knockoffs $\tilde{\boldsymbol{X}}_j^{(1)}, \ldots, \tilde{\boldsymbol{X}}_j^{(m)}$ for each feature $\boldsymbol{X}_j$. The resulting p-value would then rank $\boldsymbol{X}_j$ with respect to all of its knockoffs—if $\boldsymbol{X}_j$ enters the Lasso path after i of its knockoffs, then it has a rank of $i+1$ out of $m+1$, and therefore a p-value of $\frac{i+1}{m+1}$. We will investigate the potential benefits of a multiple-knockoff approach in future work.

Finally, the analysis and methods presented here rely on the assumption that $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$ is invertible, which is necessary so that $\boldsymbol{X}_j$ does not lie in the span of the remaining $(p-1)$ features and its effect on the response is, therefore, identifiable. In many modern applications, however, we are interested in a regime where $p > n$ and $\boldsymbol{\Sigma}$ is defacto non-invertible. In this type of setting, we can always write $\boldsymbol{X}_j$ as some linear combination of the other features, $\boldsymbol{X}_j = \boldsymbol{X}_{(-j)}\boldsymbol{\alpha}^{(j)}$, where $\boldsymbol{X}_{(-j)}$ is the design matrix with jth column removed and $\boldsymbol{\alpha}^{(j)} \in \mathbb{R}^{p-1}$. Hence, we cannot distinguish between a mean function that depends on $\boldsymbol{X}_j$ versus a mean function that depends on the linear combination $\boldsymbol{X}_{(-j)}\boldsymbol{\alpha}^{(j)}$. However, there are many types of common additional assumptions that may allow us to overcome this identifiability problem—for instance, sparse dependence structure among the features themselves and between the mean response and the features. We are exploring this type of approach in ongoing research.

Acknowledgements

R. F. B. was partially supported by NSF award DMS-1203762. E. C. is partially supported by AFOSR under grant FA9550-09-1-0643, by NSF via grant CCF-0963835 and by the Math + X Award from the Simons Foundation. E. C. would like to thank Małgorzata Bogdan, Lucas Janson, Chiara Sabatti and David Siegmund for helpful discussions and useful comments about an early version of the manuscript. E. C. also thanks Frédéric Benqué for a productive brainstorming session about names, Carlos Fernandez-Granda for his help in running simulations on the Solomon cluster, and Jonathan Taylor for suggesting the analysis of the HIV dataset. We are grateful to Soo-Yon Rhee and Rob Shafer for discussing the HIV dataset with us.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

[2] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001.

[3] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[6] E. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 04 2013.

[7] E. Chung and J. P. Romano. Multivariate and multiple permutation tests. Technical report, Stanford University, 2013.

[8] N. R. Draper and H. Smith. *Applied regression analysis*. John Wiley and Sons, 2nd edition, 1981.

[9] B. Efron. *Large-scale inference*, volume 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge, 2010. Empirical Bayes methods for estimation, testing, and prediction.

[10] M. G. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. False discovery rate control for sequential selection procedures, with application to the Lasso. 2013. `arXiv:1309.5352`.

[11] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst.*, 23:1432–1440, 2010.

[12] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the Lasso. 2012. To appear in *Annals of Statistics*.

[13] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[14] A. Miller. *Subset selection in regression*, volume 95 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2002.

[15] A. J. Miller. Selection of subsets of regression variables. *J. Roy. Statist. Soc. Ser. A*, 147(3):389–425, 1984.

[16] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE, 1993.

[17] S-Y. Rhee, W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, J. Taylor, D. P. Nguyen, S. Slome, D. Klein, M. Horberg, et al. HIV-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. *Journal of Infectious Diseases*, 192(3):456–465, 2005.

[18] S-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D.L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

[19] J. D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498, 2002.

[20] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205, 2004.

[21] J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Post-selection adaptive inference for least angle regression and the Lasso. 2014. `arXiv:1401.3889`.

[22] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[23] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

[24] P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.