Summary and discussion of: "Controlling the False Discovery Rate via Knockoffs"

Statistics Journal Club, 36-825

Sangwon "Justin" Hyun and William "Willie" Neiswanger

1 Paper Summary

1.1 Quick intuitive summary of this paper:

- In many settings we have a number of explanatory variables (i.e. "features") and a response variable.
- And we want to know which explanatory variables are *truly* associated with the response variable.
- While doing this, we want to bound the *false discovery rate* (FDR)—the expected fraction of false discoveries.
- This paper introduces the *knockoff filter*, a variable selection method that controls the FDR in linear models when the number of observations is greater than the number of variables.
- The basic idea is that you construct extra variables, called "knockoff variables", which have certain similar correlation structure as existing variables and allow for FDR control when existing variable selection methods are applied.
- The main advantages of the knockoff filter over competing methods are that it achieves exact FDR control in finite sample settings, no matter the design matrix, covariates, number of variables, or size of regression coefficients, and that it empirically shows more power than existing methods.

1.2 Background, Terminology, and Definitions

Assume a linear model of the form $y = X\beta + z$, where:

- We have a vector of responses $y \in \mathbb{R}^n$
- We have a design matrix $X \in \mathbb{R}^{n \times p}$
- We have an (unknown) vector of coefficients $\beta \in \mathbb{R}^p$.
- We have Gaussian noise $z \sim \mathcal{N}(0, \sigma^2 I)$

Suppose we have a variable selection method that returns a set of indices $\hat{S} \subset \{1, \ldots, p\}$. In the context of variable selection, we define the false discovery rate (FDR) to be the expected fraction of falsely selected variables, which can be written:

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

where the notation $a \vee b = \max\{a, b\}$. When we say that the FDR is controlled at level q, we mean that the FDR $\leq q$ for all coefficients β.

1.3 Knockoffs Overview

The goal of the knockoff filter is to provide a variable selection method that controls the FDR at level q for any finite sample of data whenever the number of observations n is larger than the number of variables p. This method should hold for any fixed design matrix X and and coefficients β.

1.3.1 Step 1: constructing the knockoffs

We now describe a general method for constructing knockoffs.

For each feature X_j (i.e. the j^{th} column of design matrix X), construct a "knockoff copy" $\tilde{X}_j$ such that:

1. $\tilde{X}^\top \tilde{X} = \Sigma$
2. $X^\top \tilde{X} = \Sigma - \text{diag}\{s\}$

where we have the Gram matrix $\Sigma = X^\top X$ (after X has been normalized such that each feature satisfies $\|X_j\|_2^2 = 1$), and nonnegative vector $s \in \mathbb{R}^{\text{l}}_+$. Note that the above conditions imply the following two conditions:

1. $X_j^\top \tilde{X}_k = X_j^\top X_k$ (for all $j \neq k$)
2. $X_j^\top \tilde{X}_j = \Sigma_{jj} - s_j = 1 - s_j$

Note that $X_j^\top X_j = \tilde{X}_j^\top \tilde{X}_j = 1$.

One basic strategy for computing knockoffs is to carry out the following two steps:

1. Choose $s \in \mathbb{R}_+^p$ such that $\text{diag}\{s\} \preceq 2\Sigma$
2. Set $\tilde{X} = X\left(I - \Sigma^{-1}\text{diag}\{s\}\right) + \tilde{U}C$

where $\tilde{U} \in \mathbb{R}^{n \times p}$ is an orthonormal matrix that is orthogonal to the span of the features of X, and C is a Cholesky decomposition of the Schur complement of the Gram matrix of the augmented knockoff design matrix, which we know exists since $\text{diag}\{s\} \preceq 2\Sigma$.

1.3.2 Step 2: calculate statistics for each pair of original/knockoff variables

We define statistics W_j for each variable $j = 1, \ldots, p$ such that large positive values give evidence that variable β_j is a true signal.

To calculate these statistics, we need an existing variable selection method. A running example throughout this document will be the LASSO method, which returns the coefficients

$$\hat{\beta} = \text{argmin}_b \left\{ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|b\|_1 \right\} \tag{2}$$

as a function of the penalty parameter λ.

We define a statistic Z_j to be the point λ where variable j "enters the model" (i.e. where the LASSO method first returns a non-zero value for β_j), i.e.

$$Z_j = \sup \left\{ \lambda : \hat{\beta}_j(\lambda) \neq 0 \right\} \tag{3}$$

To use the knockoff method, we first run the variable selection method (LASSO, in this case) on an augmented matrix, which consists of concatenating the original design and knockoff design matrices, which can be written $\left[X, \tilde{X}\right]$. Running LASSO on this augmented matrix gives us the $2p$ statistis $(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$.

We can then compute the p statistics $W_1, \ldots, W_p$ as

$$W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & \text{if } Z_j > \tilde{Z}_j \\ -1 & \text{if } Z_j < \tilde{Z}_j \end{cases} = \begin{cases} Z_j & \text{if } Z_j > \tilde{Z}_j \\ -\tilde{Z}_j & \text{if } \tilde{Z}_j > Z_j \end{cases} \tag{4}$$

1.3.3 Step 3: calculate a threshold of the W_j's

The W_j's are constructed such that higher values give evidence that j is a true signal. We define a data-dependent threshold T to allow for performing a final variable selection (with the FDR guarantees).

For a specified level q, we define the data-dependent threshold to be

$$T = \min \left\{ t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\} \tag{5}$$

or $T = +\infty$ if this set is empty, where $\mathcal{W}$ is the set of unique nonzero $|W_j|$'s, i.e. $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\} \setminus \{0\}$.

1.3.4 The Knockoff and Knockoff+ Procedures

All together, the **knockoff** procedure can be written as the following: construct the knockoff matrix $\tilde{X}$, statistics W_j, and threshold T, as in steps 1-3 above, then select the model $\hat{S}$ with

$$\hat{S} = \{j : W_j \geq T\}. \tag{6}$$

The first main result (called **Theorem 1** in the paper) states that the above method will control a quantity known as the "modified FDR". More formally, for any $q \in [0, 1]$, the knockoff method above satisfies

$$\mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} + q^{-1}}\right] \leq q \tag{7}$$

where the expectation is taken over the Guassian noise z, while treating the design matrices X and $\tilde{X}$ as fixed.

This paper also provides a procedure known as **knockoff+**, which can be written as the following: construct the knockoff matrix $\tilde{X}$ and statistics W_j as in steps 1-3 above, and chose the data-dependent threshold T to be

$$T = \min\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\} \tag{8}$$

or $T = +\infty$ if this set is empty. After, select the model $\hat{S}$ with

$$\hat{S} = \{j : W_j \geq T\}. \tag{9}$$

Note that the threshold T chosen by the knockoff+ procedure is always higher (or equal to) that chosen by the knockoff procedure, and is thus a more conservative procedure.

The second main result (called **Theorem 2** in the paper) states that the above method controls the exact FDR. More formally, for any $q \in [0, 1]$, the knockoff+ method above satisfies

$$\mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right] \leq q \tag{10}$$

where the expectation is taken over the Guassian noise z, while treating the design matrices X and $\tilde{X}$ as fixed.

1.4 Theory and Extensions

- Why $\tilde{X}$? (and why the conditions?)
  The knockoff dataset $\tilde{X}$ can be thought of as the solution to the desired correlation structure:
  $$[X\ \tilde{X}]^T[X\ \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \text{diag}\{s\} \\ \Sigma - \text{diag}\{s\} & \Sigma \end{bmatrix} = G$$
  for some vector $s \in \mathbb{R}^p$.
  We can see that $\tilde{X}$ exists iff $G \succeq 0$. Let us see why. Take G's Schur complement:[1]
  $$\begin{aligned} A &= \Sigma - (\Sigma - \text{diag}\{s\})\Sigma^{-1}(\Sigma - \text{diag}\{s\}) \\ &= 2\text{diag}\{s\} - \text{diag}\{s\}\Sigma^{-1}\text{diag}\{s\} \end{aligned}$$
  and we know that $G \succeq 0$ if and only if $A \succeq 0$. We can also see that A is the Schur's complement of another matrix (G'):
  $$G' = \begin{bmatrix} \Sigma & \text{diag}\{s\} \\ \text{diag}\{s\} & 2\text{diag}\{s\} \end{bmatrix}$$

$G' \succeq 0$ if and only if $\text{diag}\{s\} \succeq 0$ and $\Sigma - \text{diag}\{s\}\frac{1}{2}(\text{diag}\{s\})^{-1}\text{diag}\{s\} \succeq 0$ (this comes from another Schur complement of G'). This is the main condition for the existence of $\tilde{X}$ in the paper. $\cdots\cdots$ **Cond1**

- Constructing $\tilde{X}$: how and why?

  1. Let $\tilde{U} \in \mathbb{R}^{n \times p}$ be an orthonormal basis to the column space spanned by X. The existence of $\tilde{U}$ is guaranteed because we assume X is full rank ($n \geq 2p$). This implies, by construction, $\tilde{U}^T X = 0$
  2. Take the Schur complement of G from above:

     $$A = 2\text{diag}\{s\} - \text{diag}\{s\}\Sigma^{-1}\text{diag}\{s\}$$

     If the conditions **Cond1** hold, then $A \succeq 0$, which is sufficient and necessary to the existence of a C such that $A = C^T C$, which can be obtained by a Cholesky decomposition.
  3. Now, construct $\tilde{X}$ as:

     $$\tilde{X} = (I - \Sigma^{-1}\text{diag}\{s\}) + \tilde{U}C$$

  4. Then, we can see that we have constructed a knockoff matrix with the desired correlation structure.

     $$[X\ \tilde{X}]^T[X\ \tilde{X}] = \begin{bmatrix} X^T X & X^T\tilde{X} \\ \tilde{X}^T X & \tilde{X}^T\tilde{X} \end{bmatrix} = \begin{bmatrix} \Sigma & \Sigma - \text{diag}\{s\} \\ \Sigma - \text{diag}\{s\} & \Sigma \end{bmatrix}$$

- Intuitive property of the knockoff variables, in our selection scheme.

  Remember, our method is useful (power is high) only if the true variables enter before the knockoffs do. Otherwise, because we cut off and only select the first consecutive k variables, we will be selecting few of the true variables, hence low power.

  Specifically, if $x \in \mathbb{S}$ for the support index set $\mathbb{S}$, then it is a desirable trait for X_j to enter before $\tilde{X}_j$! Thus we want a near-zero correlation (near-orthogonality) between the original variable X_j and its knockoff $\tilde{X}_j$:

  $$X_j \perp\!\!\!\perp \tilde{X}_j, j \in \{1, \cdots, p\}$$

  The intuition is that, *if* there is a true relationship between the original variable X_j and y, then by constructing $\tilde{X}_j$ to be near-orthogonal with the response y is *breaking* the tie between y and $\tilde{X}_j$. If X_j is a null variable, then there is no reason why either X_j or $\tilde{X}_j$ should be any more related to y than the other. These are good properties for our selection procedure! Hence, we want:

  $$X_j^T \tilde{X}_j = 1 - s_j \simeq 0$$

[1] The Schur's complement of block matrix $G = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ with respect to A is $D - BA^{-1}C$, and we have the property that $G \succeq 0$ if and only if $A \succeq 0$ and $D - BA^{-1}C \succeq 0$. Similarly, we can obtain the same condition for the Schur complement with respect to D.

- How should we choose s?

  How to set values of $s = (s_1, \cdots, s_p)$ is a user's choice. There are two ways to do this:

  1. Set all correlations $1 - s_j$ to be equally small! We can achieve this by setting:

     $$s_j = 2\lambda \min(\Sigma) \wedge 1$$

     so that for all j, the value of $\langle X_j, \tilde{X}_j \rangle = 1 - 2\lambda \min(\Sigma) \wedge 1$
  2. Set the average correlation across all variables is minimum out of all equal-correlation s! We can achieve this by solving the program:

     $$\begin{aligned} \underset{s}{\text{minimize}} \quad & \sum_j |1 - s_j| \\ \text{subject to} \quad & s_j \geq 0 \\ & \text{diag}\{s\} \succeq 2\Sigma \end{aligned}$$

     which can be solved efficiently (see paper for more details).

- Extensions to $p \leq n \leq 2p$

  There are several ways to form the knockoff variables when $p \leq n \leq 2p$. The first idea is to augment the data with some generated y' and block matrix 0 so that this is the approximate model:

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

  Then, we can use the knockoff filter just as in the $n \geq 2p$ setting; by forming the knockoffs and finding the data dependent threshold for selection! The main issue with this method is that we don't know know σ^2, so we need to use an estimate $\hat{\sigma}^2$ which comes from $\|y - X\hat{\beta}^{LS}\|_2^2 \sim \sigma^2 \chi^2_{n-p}$ so that $\hat{\sigma^2} \sim \frac{\|y - X\hat{\beta}^{LS}\|^2}{E(\chi^2_{n-p})} = \frac{\|y - X\hat{\beta}^{LS}\|^2}{n-p}$ and the variance of the estimate is $V(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p}$ so that your estimate is stable when $n - p$ is large. Then you can draw $y' \in \mathbb{R}^{n-p}$ from $\mathcal{N}(0, \hat{\sigma}^2 I)$. Plainly put, we want to generate dummy y's that have same noise structure.

  Another cool way is to test a subset of size $n - p$, and cycle through them for high power. Specifically, let us take an index set I of size $|I| = 2p - n$. Then, construct knockoffs for the variables $X_i, i \in I$ in the index set as *exact* copies. and choose $s_i = 0, i \in I$. For $i \notin I$, construct the knockoff's as usual. Then, by applying the knockoff filter, this only select variables in $\{1, 2, \cdots, p\} \backslash I$! This has FDR control (by the theoretical control of FDR on any knockoff procedure) but still has low power for detecting variables in I. So, a solution is to cycle through disjoint index sets $I_1, \cdots, I_k \subseteq \{1, \cdots, p\}, \cup I_i = \{1, \cdots, p\}$ and at each time select (via knockoff filter) at

FDR target of q/k, which overall allows FDR control of q and good power. Essentially, *cycle through duplicates* with an overall q level of FDR control. [2]

- Properties of the statistic W desirable for our selection procedure.

  There are some properties that are desirable for the statistic W. Namely, we want to form W as a function of the augmented design matrix and response data, $W([X\tilde{X}], y) \in \mathbb{R}^p$, so that large positive W_j indicate $\beta_j \neq 0$ and so that two properties hold: sufficiency and anti-symmetry.

  Sufficiency means that W depends only on the gram matrix $[X\tilde{X}]^T[X\tilde{X}]$ and the feature-response inner products$[X\tilde{x}]^T y$:

$$W = f([X\ \tilde{X}]^T[X\ \tilde{X}], [X\ \tilde{X}]^T y)$$
$$f : S^+_{2p} \times \mathbb{R}^{2p} \to \mathbb{R}^p$$

  [3]

  The anti-symmetry is simple: swapping X_j, $\tilde{X}_j$ has the effect of switching the sign of the statistic W. This can be satisfied easily by constructing the statistic to reflect this exactly. For example, our default choice of the statistic

$$W = 00000$$

  has this property by construction.

  If these two properties hold, then the theoretical FDR control guarantees hold when using this statistic W for variable selection.

1.5 Simulations: Comparison with other methods

We ran simulations on toy data to compare the performance of several methods. Specifically, we see the *observed* FDR (actual proportion of false discoveries) and corresponding power (actual proportion of signals we detect) for each methods in each settings. Specifically, we generated $y \in \mathbb{R}^{300}$ and $X \in \mathbb{R}^{300\times 100}$ according to the following model:

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_k)$$

where $\beta = (\beta_1\ \beta_2\ \cdots \beta_{30}\ 0\ \cdots\ 0)^T$ and $\beta_i, i \in \{1, 2, \cdots, 30\}$ were generated as

$$\beta_i = \begin{cases} +b & w.p. \frac{1}{2} \\ -b & w.p. \frac{1}{2} \end{cases}$$

The role of b is to make the variables easy or hard to detect, depending on the size of the variables. We vary b in $\{1, 2, 5, 10, 20, 50\}$, and we refer to the order of them as

[2] Some details when forming knockoffs: we choose s to have nonzero entries in $\{1, \cdots, p\}\backslash I$ so that $A =$ Schur Complement of G has rank $n - p$, then find C such that $A = C^T C$, $C \in \mathbb{R}^{(n-p)\times p}$ and find $\tilde{U} \in \mathbb{R}^{n\times(n-p)}$ which is orthogonal to X (via SVD) and finally form $\tilde{X} = X(I - \Sigma^{-1}\text{diag}\{s\}_+\tilde{U}C$, then make data-dependent threshold and select variables!

[3] S^+_{2p} denotes a cone of $2p \times 2p$ positive semidefinite matrices.

the signal strength. To illustrate, at $b = 20$, the signal variables are very apparent (signal strength is large), so most automatic selection methods (say forward selection) should be able to detect the majority of the first 30 variables i.e. have high power, even without using the knockoff threshold. The knockoff filter is tied to our choice of variable selection. We run the knockoff filter with two variable selection methods (lasso and forward selection) the Benjamini-Hochberg procedure on linear regression coefficient estimate t-test p-values, and also the knockoff+ (which gives exact FDR control) with the Lasso. So, in sum, we compare: FS-Knockoff, lasso-Knockoff, lasso-Knockoff+, and linear-regression-BHq. Note that the Benjamini Hochberg procedure is expected to perform poorly (low power) because the p-values under the null are NOT independent.

A summary of the simulation results is in figure 1. We can see that the power for all four methods increases as we increase the size of the signal k, which agrees with our intuition. The BHq procedure seems to be controlling the signal strength for signal strengths Several pending questions remain: all four methods seem to exceed the target 0.2 FDR. knockoff+ applied on the lasso peaks at a moderately weak signal strength ($|b| = 2$). We expected the BHq on a forward selection (greedy) to perform not too well, but it is has comparable power to the and maintains exact FDR control. Also, the knockoff+ is theoretically guaranteed exact control, but there is an unexpected increase ($\sim$.6) in the observed FDR in athe moderately weak signal strength ($|b| = 2$) setting.

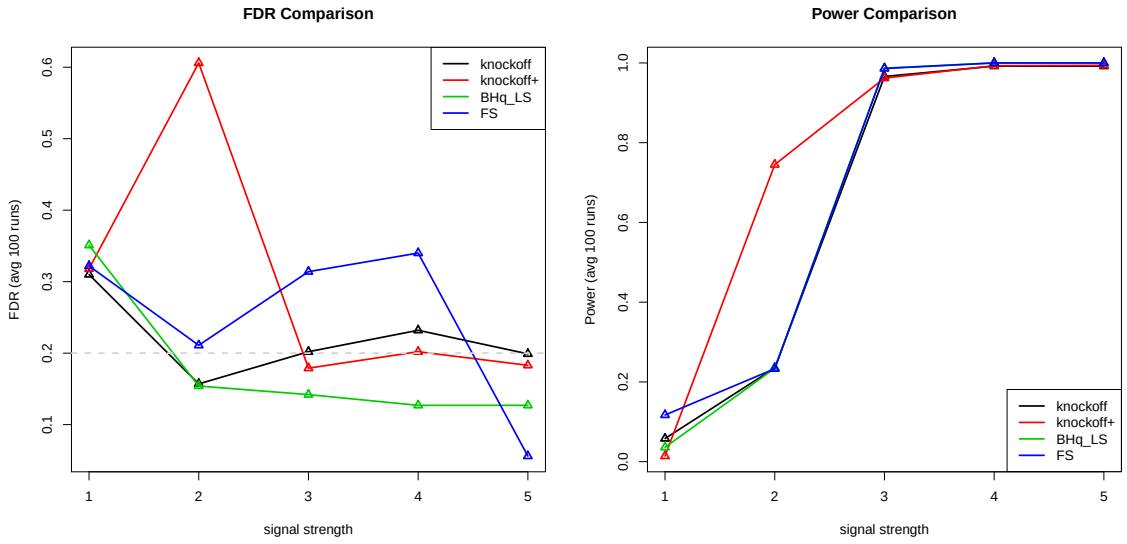

Figure 1: FDR and Power comparison to data generated from model $y \sim \mathcal{N}(X\beta, \sigma^2 I_k)$, with $y \in \mathbb{R}^{300}$ and $X \in \mathbb{R}^{300\times 100}$, where $\beta = (\beta_1\ \beta_2\ \cdots \beta_{30}\ 0\ \cdots\ 0)^T$ and $\beta_i, i \in \{1, 2, \cdots, 30\}$ were generated as $\beta_i \in \{+b, -b\}$ randomly. Each point plotted is an average over 50 iterations.

2 Discussion

Here, we list several interesting results from class discussion. The question was raised whether this data-dependent threshold was needed at all; i.e. if instead of choosing variables whose W_i are above a threshold, you could pick W's such that their absolute value was above a threshold (and perhaps have the same result). There was also a discussion of how this compares to the Benjamini Hochberg procedure on the p values in variable selection. Also, how, intuitively, do the knockoffs contribute to the problem of FDR control? An insightful answer was that the relationship (if it exists) between the response and a given variable is intentionally broken when constructing a knockoff variable, so that for signal variables, it *will* be rare (and by chance) that a knockoff will enter before its original brother, while for non-signal variables, there is a 50/50 chance they will enter. A good variable selection method should be able to successfully select original signal variables over their knockoffs, and select signal variables before the non-signals. While our method has good theoretical FDR control guarantees, if the selection method is poor in the first place the power (power to select signal variables) will be low (in the extreme case, a naive selection method that just selects variable $i = 2$ every time used with the knockoff filter will select no variables most of the time). It was also mentioned that for lasso, at each selection time, you can imagine there is a 'competition' between the remaining variables to enter the model by proving they have a high correlation with y (LARS selects the variable j that has largest correlation $y^T X_j$ out of the remaining j's), and that if the tie is broken between the knockoff and y, then the knockoff is *sure to lose* to its original brother most of the time.

References

[1] R. Foygel Barber, E. Candes, *Controlling the False Discovery Rate vis Knockoffs. ArXiv e-prints*, April 201

[2] Y. Benjamini, Y. Hochberg, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 57, No. 1 (1995), pp. 289-300

[3] Tibshirani, R. *Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1(1996), pp. 267-288