A prototype knockoff filter for group selection with FDR control

Jiajie Chen

School of Mathematical Sciences, Peking University, Beijing 100871, China
cjiajie@pku.edu.cn

Anthony Hou

Department of Statistics, Harvard University, Cambridge, MA 02138, USA
ahou@college.harvard.edu

and

Thomas Y. Hou[†]

Applied and Computational Mathematics, California Institute of Technology,
Pasadena, CA 91125, USA
[†]Corresponding author. Email: hou@cms.caltech.edu

In many applications, we need to study a linear regression model that consists of a response variable and a large number of potential explanatory variables, and determine which variables are truly associated with the response. In Foygel Barber & Candès (2015, *Ann. Statist.*, **43**, 2055–2085), the authors introduced a new variable selection procedure called the knockoff filter to control the false discovery rate (FDR) and proved that this method achieves exact FDR control. In this paper, we propose a prototype knockoff filter for group selection by extending the Reid–Tibshirani (2016, *Biostatistics*, **17**, 364–376) prototype method. Our prototype knockoff filter improves the computational efficiency and statistical power of the Reid–Tibshirani prototype method when it is applied for group selection. In some cases when the group features are spanned by one or a few hidden factors, we demonstrate that the Principal Component Analysis (PCA) prototype knockoff filter outperforms the Dai–Foygel Barber (2016, *33rd International Conference on Machine Learning (ICML 2016)*) group knockoff filter. We present several numerical experiments to compare our prototype knockoff filter with the Reid–Tibshirani prototype method and the group knockoff filter. We have also conducted some analysis of the knockoff filter. Our analysis reveals that some knockoff path method statistics, including the Lasso path statistic, may lead to loss of power for certain design matrices and a specially designed response even if their signal strengths are still relatively strong.

Keywords: variable selection; false discovery rate (FDR); group variable selection; knockoff filter; linear regression.

1. Introduction

In many scientific endeavors, we need to determine from a response variable together with a large number of potential explanatory variables which variables are truly associated with the response. In order for this study to be meaningful, we need to make sure that the discoveries are indeed true and replicable. Thus, it is highly desirable to obtain exact control of the false discovery rate (FDR) within a certain prescribed level. In [1], Foygel Barber and Candès introduced a new variable selection procedure called the knockoff filter to control the FDR for a linear model. This method achieves exact FDR control

in finite sample settings and does not require any knowledge of the noise level. A key observation is that by constructing knockoff variables that mimic the correlation structure found within the existing variables one can obtain accurate FDR control. The method is very general and flexible. It can be applied to a number of statistics and has more statistical power (the proportion of true signals being discovered) than existing selection rules in some cases.

1.1 *A brief review of the knockoff filter*

Throughout this paper, we consider the following linear regression model: $y = X\beta + \epsilon$, where the feature matrix X is an $n \times p$ ($n \geq 2p$) matrix with full rank, its columns are normalized to be unit vectors in the l^2 norm and ϵ is a Gaussian noise $\varepsilon \sim N(0, \sigma^2 I_n)$. We first provide a brief overview of the knockoff filter introduced in [1]. The knockoff filter begins with the construction of a knockoff matrix $\tilde{X}$ that obeys

$$\tilde{X}^T \tilde{X} = X^T X, \quad \tilde{X}^T X = X^T X - diag(s), \tag{1}$$

where $s_i \in [0, 1]$. The positive definiteness of the Gram matrix $[X\tilde{X}]^T [X\tilde{X}]$ requires

$$diag(s) \preceq 2X^T X. \tag{2}$$

The first condition in (1) ensures that $\tilde{X}$ has the same covariance structure as the original feature matrix X. The second condition in (1) guarantees that the correlations between distinct original and knockoff variables are the same as those between the originals. To ensure that the method has good statistical power to detect signals, we should choose s_j as large as possible to maximize the difference between X_j and its knockoff $\tilde{X}_j$. Once $diag(s)$ is obtained, $\tilde{X}$ can be constructed in terms of $X, diag(s)$ and an orthonormal matrix $U \in R^{n \times p}$ with $U^T X = 0$. The existence of U requires $n \geq 2p$. The next step is to calculate a statistic, W_j, for each pair X_j and $\tilde{X}_j$ using the Gram matrix $[X \tilde{X}]^T [X\tilde{X}]$ and the marginal correlation $[X \tilde{X}]^T y$. In addition, W_j satisfies a flip-coin property, which implies that swapping arbitrary pair X_j and $\tilde{X}_j$ only changes the sign of W_j, but keeps the sign of other W_i ($i \neq j$) unchanged. The construction of the knockoff features and the symmetry of the test statistic are important to achieve a crucial property that the signs of the W_j's are i.i.d. random for the null hypotheses. This property plays a crucial role in obtaining exact FDR control by using a supermartingale argument.

The final step is to run the knockoff (knockoff+) selection procedure at the target FDR level q. A large positive W_j gives evidence that variable j is a non-null. Then select the model $\hat{S} \triangleq \{j : W_j \geq T\}$ using a data-dependent threshold T defined below:

$$T_i \triangleq \min \left\{ t > 0 : \frac{i + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}, \quad i = 0, 1. \tag{3}$$

T_0 and T_1 are used in the knockoff and knockoff+ selection procedure, respectively.

There are several ways to construct a statistic W_j. Among them, the Lasso path (LP) statistic is discussed in detail in [1]. It first fits a Lasso regression of y on $[X\tilde{X}]$ for a list of regularizing parameters λ in a descending order and then calculates the first λ at which a variable enters the model, i.e. $Z_j \triangleq \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$ for feature X_j and $\tilde{Z}_j = \sup\{\lambda : \tilde{\beta}_j(\lambda) \neq 0\}$ for its knockoff $\tilde{X}_j$. The LP statistic is defined as $W_j = \max(Z_j, \tilde{Z}_j) \cdot sign(Z_j - \tilde{Z}_j)$. If X_j is a non-null, it has a non-trivial effect on y

and should enter the model earlier than its knockoff $\tilde{X}_j$, resulting in a large positive W_j. In this case, the corresponding feature is likely to be selected by the knockoff filter (3). If X_j is a null, it is likely that X_j enters the model later, resulting in a small positive or negative W_j. The corresponding feature is likely to be rejected according to (3).

The main result in [1] is that the knockoff procedure and knockoff+ procedure has exact control of modified false discovery rate (mFDR) and FDR, respectively,

$$mFDR \triangleq E\left[\frac{\#\{j \in \hat{S} : \beta_j = 0\}}{\#\{j \in \hat{S}\} + q^{-1}}\right] \le q, \quad FDR \triangleq E\left[\frac{\#\{j \in \hat{S} : \beta_j = 0\}}{\#\{j \in \hat{S}\} \vee 1}\right] \le q.$$

In a subsequent paper [2], Foygel Barber and Candès developed a framework for high-dimensional linear model with $p \ge n$. In this framework, the observations are split into two groups, where the first group is used to screen for a set of potentially relevant variables, whereas the second is used for inference over this reduced set of variables. The authors also developed strategies for leveraging information from the first part of the data at the inference step for greater power. They proved that this procedure controls the directional FDR in the reduced model controlling for all screened variables.

The knockoff filter has been further generalized to the model-free framework in [7]. Whereas the knockoffs procedure is constrained to linear models, the model-X knockoffs provide valid inference from finite samples in settings in which the conditional distribution of the response is arbitrary and completely unknown. Furthermore, this holds independent of the number of covariates. They achieved correct inference by constructing knockoff variables probabilistically instead of geometrically. The rigorous FDR control of the model-free knockoffs procedure is also established.

The research development for the knockoff filter has inspired a number of follow-up works; see e.g. [9,11,14].

1.2 *A prototype knockoff filter for group selection*

Group selection is an effective way to perform statistical inference when features within each group are highly correlated, but the correlation among different groups is relatively weak. Inspired by the prototype method developed by Reid and Tibshirani in [14], we propose a prototype knockoff filter for group selection that has exact group FDR control (defined in Theorem 2.1) for strongly correlated features. Assume that X can be clustered into k groups $X = (X_{C_1}, X_{C_2}, ..., X_{C_k})$ in such a way that the within-group correlation is relatively strong. As in [14], we split the data (X, y) by rows into two disjoint parts $(X^{(1)}, y^{(1)})$ and $(X^{(2)}, y^{(2)})$, and extract prototype X_{P_i} for each group $|C_i|$ using the first part of the data $(X^{(1)}, y^{(1)})$. We then construct the knockoff matrix $\tilde{X}_P^{(2)}$ only for the prototype features in the second part of the design matrix and run the knockoff selection. Finally, select group i if P_i is selected by the knockoff filter. We have also developed a *Principal Component Analysis (PCA) prototype filter* and proved that both of these prototype knockoff filters have exact group FDR control. We compare these two prototype filters with a variant of the Reid–Tibshirani prototype method for group selection and the Dai–Foygel Barber group knockoff filter [9], and provide numerical experiments to demonstrate the effectiveness of our methods.

1.3 *Alternating sign effect*

In this paper, we have also performed some analysis of the knockoff filter with certain path method statistics such as the LP and the forward selection statistics. According to (3), the knockoff filter

threshold T is determined by the ratio of the number of large positive and negative W_j. Large, negative W_j's may result in a large T and fewer selected features. For certain design matrix X and a specially designed response y with strong signal strengths, our analysis shows that for the knockoff filter with certain path method statistics, e.g. the LP statistic, some knockoff variable $\tilde{X}_j$ can enter the model earlier than its original feature X_j. This could lead to large negative W_j and reduce the power. We discuss some possible mechanism under which the path method statistic may suffer from this potential challenge for certain design matrices and the response. A possible scenario is when some features are positively (negatively) correlated, but their contribution to the response y has the opposite (same) sign, e.g. $X_j^T X_k > 0, \beta_j > 0, \beta_k < 0$. We call this mechanism the *alternating sign effect*. In general, the chance that such potential challenge arises is quite rare. But it gives a healthy warning that this potential challenge could occur for certain statistics and we must use them with care.

1.4 *Extension of the sufficiency property*

The sufficiency property of a knockoff statistic W in [1] states that W depends only on the Gram matrix $[X\tilde{X}]^T[X\tilde{X}]$ and the feature–response product $[X\tilde{X}]^T y$. In this definition, only part of the information of the response variable y, i.e. $[X\tilde{X}]^T y$, is utilized. We generalize the sufficiency property such that W can depend on the remaining information of y. This generalization maintains the FDR control of the knockoff procedure. As an application, we show that the classical noise estimate obtained by using the least squares can be incorporated in the knockoff filter without violating FDR control. We remark that in a recent work [5], the sufficiency property is also relaxed by allowing W to be a function of $||y||_2$. The definition we propose is more general than the one used in [5].

The main motivation for us to study the prototype knockoff filter for group selection is to alleviate the difficulty in feature selection of highly correlated features. In a related work [8], we have developed a pseudo knockoff filter in which we relax one of the knockoff constrains. Our numerical study indicates that the pseudo knockoff filter could give high statistical power for certain statistic when the features have relatively strong correlation. Although we cannot establish rigorous FDR control for the pseudo knockoff filter as the original knockoff filter, we provide some partial analysis of the pseudo knockoff filter with the half Lasso statistic, and establish a uniform False Discovery Proportion(FDP) bound and an expectation inequality.

The rest of the paper is organized as follows. In Section 2, we introduce our prototype knockoff filter for highly correlated features. We compare it to other group selection methods and provide numerical experiments to demonstrate the performance of various methods. In Section 3, we discuss the potential challenge of the knockoff filter with certain path method statistics due to the alternating sign effect, and generalize the sufficiency property of a knockoff statistic.

2. Prototype knockoff filters

In this section, we propose two prototype group selection methods with group FDR control to overcome the difficulty associated with strong within-group correlation. It is well known that the grouping strategy provides an effective way to handle strongly correlated features. Our work is inspired by Reid–Tibshirani's prototype method [14], Dai–Foygel Barber's group knockoff filter [9] and Foygel Barber–Candès' high-dimensional knockoff filter [2]. We provide a brief summary of the first two methods below before introducing our prototype filters.

2.1 Reid–Tibshirani's prototype method

In [14], Reid and Tibshirani introduced a prototype method for prototype selection. It can be applied directly to group selection and consists of the following steps. First, cluster columns of X into K groups, $\{C_1, ..., C_K\}$. Then split the data by rows into two (roughly) equal parts $y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix}$ and $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$. Choose a prototype for each group via the maximal marginal correlation, using only the first part of the data $y^{(1)}$ and $X^{(1)}$. This generates the prototype set $\hat{P}$. Next, form a knockoff matrix $\tilde{X}^{(2)}$ from $X^{(2)}$ and perform the knockoff filter using $y^{(2)}, [X_{\hat{P}}^{(2)} \ \tilde{X}_{\hat{P}}^{(2)}]$. Finally, group C_i is selected if and only if $X_{P_i}^{(2)}$ is chosen in the filter process. The group FDR control is a direct result of Lemma 6.1 [14]. We remark that strong within-group correlation results in small difference between the prototype X_{P_i} and its knockoff pair $\tilde{X}_{P_i}$. Suppose that X_{P_i} and X_j, $j \neq P_i$ are in the same group and strongly correlated, i.e. $||X_{P_i} - X_j||_2$ is small (X_j and X_{P_i} are normalized). The knockoff constraint $(X_{P_i} - \tilde{X}_{P_i})^T X_j = 0$ implies

$$||\tilde{X}_{P_i} - X_j||_2 = ||X_{P_i} - X_j||_2, \quad ||X_{P_i} - \tilde{X}_{P_i}||_2 \leq ||\tilde{X}_{P_i} - X_j||_2 + ||X_{P_i} - X_j||_2 = 2||X_{P_i} - X_j||_2. \quad (4)$$

Thus, $||X_{P_i} - \tilde{X}_{P_i}||_2$ is forced to be small. Hence, applying this method directly to group selection may lose power for strongly correlated features. Our numerical experiments confirm this.

2.2 Dai–Foygel Barber's group knockoff filter

In [9], Dai and Foygel Barber introduced a group-wise knockoff filter, which is a generalization of the knockoff filter. Assume that the columns of X can be divided into k groups $\{X_{G_1}, X_{G_2}, ..., X_{G_k}\}$. The authors construct the group knockoff matrix according to $\tilde{X}^T \tilde{X} = X^T X$, $\tilde{X}^T X = \Sigma - S$ and $\Sigma = X^T X$, where $S \succeq 0$ is group-block-diagonal, i.e. $S_{G_i, G_j} = 0$ for any two distinct groups $i \neq j$. In the equi-correlated construction, $S = diag(S_1, S_2, ..., S_k)$, $S_i = \gamma \Sigma_{G_i, G_i} = \gamma X_{G_i}^T X_{G_i}$ and $i = 1, 2, ..., k$. The constraint $S \preceq 2\Sigma$ implies $\gamma \cdot diag(\Sigma_{G_1, G_1}, \Sigma_{G_2, G_2}, ..., \Sigma_{G_k, G_k}) = S \preceq 2\Sigma$. In order to maximize the difference between X and $\tilde{X}$, γ is chosen as large as possible: $\gamma = \min\{1, 2 \cdot \lambda_{\min}(D\Sigma D)\}$, where $D = diag(\Sigma_{G_1, G_1}^{-1/2}, \Sigma_{G_2, G_2}^{-1/2}, ..., \Sigma_{G_k, G_k}^{-1/2})$. In the later numerical experiments, we will also use the semidefinite programming (SDP) construction, which was not considered in [9], by solving $\max \sum_i^k \gamma_i$ subject to $S = diag(S_1, S_2, ..., S_k) \preceq 2\Sigma$, $S_i = \gamma_i \Sigma_{G_i, G_i}$ and $0 \leq \gamma_i \leq 1$. This construction can be viewed as an extension of the SDP knockoff construction in [1]. Due to the extra cost in solving the optimization problem, the SDP construction is more expensive than the equi-correlated construction.

The group-wise statistic introduced in [9] can be obtained after the construction of the group knockoff matrix. The construction above guarantees the group-wise exchangeability. Finally, group FDR control, i.e. $FDR_{group} \triangleq E\left[\frac{\#\{i:\beta_{G_i}=0, i\in\hat{S}\}}{(|\hat{S}|\vee 1)}\right] \leq q$, is a result of the group-wise exchangeability. Here $\hat{S} = \{j : W_j \geq T\}$ is the set of selected groups for a chosen group statistic W_j.

2.3 Prototype knockoff filters

2.3.1 *Prototype using the data* In this subsection, we propose a prototype knockoff filter that takes advantage of the prototype features, and improves the computational efficiency in the construction of knockoffs and statistical power of the Reid–Tibshirani prototype method. We assume that X can be clustered into k groups $X = (X_{C_1}, X_{C_2}, ..., X_{C_k})$ in such a way that within-group correlation is relatively

J. CHEN *ET AL.*

strong. We select the prototype features using a procedure similar to Reid–Tibshirani's prototype method.

Step 1 Split the data by rows into two parts $y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix}$ and $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$, where $y^{(1)} \in R^{n_1}, y^{(2)} \in R^{n_2}, X^{(1)} \in R^{n_1 \times p}$ and $X^{(2)} \in R^{n_2 \times p}$, and then choose a prototype P_i for each group via the marginal correlation $P_i = \arg\max_{j \in C_i} |X_j^{(1)T} y^{(1)}|$, using the first part of the data.

Step 2 Let $Q = \{1, 2, \ldots, p\} \backslash P$. The knockoff matrix $\widetilde{X^{(2)}} = \left(\widetilde{X^{(2)}}_P, X_Q^{(2)} \right)$ obeys

$$\widetilde{X^{(2)}}_P^T \widetilde{X^{(2)}}_P = X_P^{(2) \, T} X_P^{(2)}, \quad X_P^{(2) \, T} X_P^{(2)} - X_P^{(2) \, T} \widetilde{X^{(2)}}_P = diag(s_P), s_P \in R^{|P|}, \tag{5}$$

$$\left(\widetilde{X^{(2)}}_{P_i} - X_{P_i}^{(2)} \right)^T X_{C_i^c}^{(2)} = 0, \quad \text{for } i = 1, 2, \ldots, k, \tag{6}$$

where $X_{C_i^c} = (X_{C_1}, \ldots, X_{C_{i-1}}, X_{C_{i+1}}, \ldots, X_{C_k})$. The construction of Reid–Tibshirani's prototype method [14] requires $\left(\widetilde{X^{(2)}}_{P_i} - X_{P_i}^{(2)} \right)^T X_{P_i^c}^{(2)} = 0$, which implies $\left(\widetilde{X^{(2)}}_{P_i} - X_{P_i}^{(2)} \right)^T X_{C_i \backslash P_i}^{(2)} = 0$. If the within-group correlation is strong, $\widetilde{X^{(2)}}_{P_i}$ is forced to be close to $X_{P_i}^{(2)}$ (see (4)). For group selection, the within-group constraints are not necessary, and we do not impose the constrains between $\widetilde{X^{(2)}}_{P_i}$ and $X_{C_i \backslash P_i}^{(2)}$ in (6). Thus, we can construct $\widetilde{X^{(2)}}_{P_i}$ that can maximize its difference with the original feature $X_{P_i}^{(2)}$:

$$\overline{X^{(2)}}_{P_i} = X_{P_i}^{(2)} - X_{C_i^c}^{(2)} \left(X_{C_i^c}^{(2) \, T} X_{C_i^c}^{(2)} \right)^{-1} X_{C_i^c}^{(2) \, T} X_{P_i}^{(2)}, \quad W_i = \overline{X^{(2)}}_{P_i} / ||\overline{X^{(2)}}_{P_i}||_2^2, \quad W \in R^{n_2 \times k}. \tag{7}$$

We consider two constructions of s_P: the equi-correlated construction $s_{P_i} = 2\lambda_{\min}((W^T W)^{-1}) \wedge ||X_{P_i}^{(2)}||_2^2$ for all i and the SDP construction

$$\text{maximize} \sum_{i=1}^{k} s_{P_i} \quad \text{subject to} \quad diag(s_P) \preceq 2(W^T W)^{-1}, \quad 0 \leq s_{P_i} \leq ||X_{P_i}^{(2)}||_2^2. \tag{8}$$

We then construct $\widetilde{X^{(2)}}_P = X_P^{(2)} - W diag(s_P) + UC$; where $U \in R^{n_2 \times k}$ is an orthonormal matrix with $U^T X^{(2)} = 0$ and $C^T C = 2 diag(s_P) - diag(s_P) W^T W diag(s_P)$, $C \in R^{k \times k}$ is the Cholesky decomposition. The existence of U and C requires $n_2 \geq p + k$ and $diag(s_P) \preceq 2(W^T W)^{-1}$.

Step 3 Recycle the first part data and select features with recycling. We concatenate the original design matrix on the first part with the knockoff matrix on the second part $\tilde{X} = \begin{pmatrix} X^{(1)} \\ \widetilde{X^{(2)}} \end{pmatrix}$ and $\tilde{X}_P = \begin{pmatrix} X_P^{(1)} \\ \widetilde{X^{(2)}}_P \end{pmatrix}$. One can verify that (5) still holds true for $(X, \tilde{X})$ with the same s_P

$$\tilde{X}_P^T \tilde{X}_P = X_P^T X_P, \quad X_P^T X_P - X_P^T \tilde{X}_P = diag(s_P), s_P \in R^{|P|}. \tag{9}$$

Finally, run the knockoff filter on y and $(X_P, \tilde{X}_P)$ as (3) to obtain $\hat{S}$, and select group i if $P_i \in \hat{S}$.

The recycling procedure in knockoff filter was developed in [2], in which the authors showed that it could raise the power substantially. We also observed this improvement in our numerical simulation. In

order to select the prototype in step 1 efficiently and retain a large difference between X_{P_i} and $\tilde{X}_{P_i}$, we choose $n_1 = 0.2n \vee 5 \max_i |C_i|$ and $n_2 = n - n_1$. The requirement $n_2 \geq p + k$ implies $n \geq 1.25(p + k)$.

The main result of the prototype knockoff filter is that it controls the group FDR.

THEOREM 2.1   For any $q \in [0, 1]$, the prototype knockoff filter using the knockoff and knockoff+ filter controls the group mFDR and group FDR, respectively,

$$
mFDR_{group} \triangleq E\left[\frac{\#\{i : \ \beta_{C_i} = 0, \ i \in \hat{S}\}}{\#\{i : \ i \in \hat{S}\} + q^{-1}}\right] \leq q, \quad FDR_{group} \triangleq E\left[\frac{\#\{i : \ \beta_{C_i} = 0, \ i \in \hat{S}\}}{\#\{i : \ i \in \hat{S}\} \vee 1}\right] \leq q.
$$

The result is a consequence of the lemma below and the super-martingale argument [1].

LEMMA 2.2   (i.i.d. signs for the null clusters). Let $\eta \in \{\pm 1\}^k$ be a sign sequence independent of W_P, with $\eta_j = +1$ for all non-null clusters j and $\eta_j \overset{i.i.d.}{\sim} \{\pm 1\}$ for null clusters j. Conditional on $y^{(1)}$, $(W_{P_1}, ..., W_{P_k}) \overset{d}{=} (W_{P_1}\eta_1, ..., W_{P_k}\eta_k)$.

Proof.   Recall that the features in the final knockoff screening process (step 3) is $(X_P, \tilde{X}_P)$. The statistic W_P can be written as $W_P = W([X_P, \tilde{X}_P]^T[X_P, \tilde{X}_P], [X_P, \tilde{X}_P]^T y)$. Since the true model is $y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$, we have $y^{(i)} = X^{(i)}\beta + \varepsilon^{(i)}, i = 1, 2$, where $\varepsilon^{(1)}$ is the first n_1 components of ε and $\varepsilon^{(2)}$ consists of the remaining components. In particular, the prototype set P and $y^{(1)}$ are independent of $\varepsilon^{(2)}$ and conditional on $y^{(1)}$; the randomness of W_P comes from $\varepsilon^{(2)}$ only. Following the analysis for the original knockoff filter [1], we just need to verify the exchangeability for the features and the response. The exchangeability for the features comes from (9). Conditional on $y^{(1)}$, the exchangeability for the response is guaranteed by the invariance of $Var([X_P \ \tilde{X}_P]^T_{swap(S')} y \mid y^{(1)})$ for any S', which is a result of (5), (9) and $E[(X_{P_i} - \tilde{X}_{P_i})^T y \mid y^{(1)}] = 0$ for null clusters i:

$$
E\big[(X_{P_i} - \tilde{X}_{P_i})^T y \mid y^{(1)}\big] = E\big[(X_{P_i}^{(2)} - \tilde{X}_{P_i}^{(2)})^T y^{(2)} \mid y^{(1)}\big] = E\big[\big(X_{P_i}^{(2)} - \tilde{X}_{P_i}^{(2)}\big)^T (X^{(2)}\beta + \varepsilon^{(2)}) \mid y^{(1)}\big]
$$
$$
= E\big[\big(X_{P_i}^{(2)} - \tilde{X}_{P_i}^{(2)}\big)^T X^{(2)}\beta \mid y^{(1)}\big] = (X_{P_i}^{(2)} - \tilde{X}_{P_i}^{(2)})^T X^{(2)}\beta = (X_{P_i}^{(2)} - \tilde{X^{(2)}}_{P_i})^T (X_{C_i^c}^{(2)} \beta_{C_i^c}) = 0.
$$

The first equality holds because X_{P_i} agrees with $\tilde{X}_{P_i}$ in the first n_1 components. The third equality holds due to the fact that the first part of the data, $y^{(1)}$, is independent of $\varepsilon^{(2)}$, which generates the second part of the data, $y^{(2)}$, and has zero mean. The final equality follows from (6).   □

2.3.2   *A PCA prototype filter*   In this subsection, we propose a PCA prototype filter for group selection for some special cases. The PCA prototype filter works well under the following assumptions: (i) the within-group correlation is relatively strong; (ii) the features within each group are positively correlated; and (iii) we know *a priori* that the signals in each group have the same sign. Assume that X can be clustered into k groups $X = (X_{C_1}, X_{C_2}, ...X_{C_k})$ as in the previous subsection. The PCA prototype filter follows steps similar to those described in the previous subsection. First of all, we calculate the first principal component V_i for each group $X_{C_i}, 1 \leq i \leq k$. Secondly, we construct the knockoff matrix following a similar procedure

$$
\bar{V}_i = V_i - X_{C_i^c}^{(2)}(X_{C_i^c}^{(2) \ T} X_{C_i^c}^{(2)})^{-1} X_{C_i^c}^{(2) \ T} V_i, \quad W_i = \bar{V}_i/||\bar{V}_i||_2^2, \quad \tilde{V} = V - Wdiag(s_P) + UC, \tag{10}
$$

where $P = \{1, 2, \ldots, k\}$ and s_P is obtained from the equi-correlated construction or the SDP (8) with a slightly different constraint $0 \leq s_i \leq 1$. U is an orthonormal matrix with $U^T X = 0$ and C is obtained by the same formula. Finally, we run the knockoff filter on y and $[V, \tilde{V}]$ to obtain $\hat{S}$ and select group i if $i \in \hat{S}$. Theorem 2.1 holds true for the PCA prototype filter and the proof is similar.

REMARK 2.3   In this paper, we focus on selecting one prototype for each group. The prototype knockoff filter can also be generalized to include a few prototypes for each group.

2.3.3   *Computing the projection*   For $X \in R^{n \times p}, n > p$, we design a recursive procedure to calculate $\bar{X}_{C_i} = X_{C_i} - X_{C_i^c}(X_{C_i^c}^T X_{C_i^c})^{-1} X_{C_i^c}^T X_{C_i}$ for all i with $O(np^2)$ flops. As a result, we can obtain W in (7) with $O(np^2)$ flops. Similar result holds true for (10). For simplicity, we assume $|C_i| = l$ for all i.
Algorithm: Projection (X, k, l)

1. If $k = 1$, return X.

2. Else: divide X into two parts $a = \lfloor k/2 \rfloor$, $X_1 = X(:, 1 : al)$ and $X_2 = X(:, (al + 1) : kl)$. Then compute the projection recursively: $W_1 = X_1 - X_2(X_2^T X_2)^{-1} X_2^T X_1$ and $W_2 = X_2 - X_1(X_1^T X_1)^{-1} X_1^T X_2$. $\bar{X}_1 = $ **Projection**(W_1, a, l), $\bar{X}_2 = $ **Projection**$(W_2, k - a, l)$, return $(\bar{X}_1, \bar{X}_2)$.

For fixed n and l, let a_k be the total flops. From the recursion, we have $a_1 \leq C$ and $a_{2^m} \leq 2a_{2^{m-1}} + C2n(2^{m-1}l)^2$ for some universal constant C. Simple calculation yields $a_{2^m} \leq Cn(2^m l)^2$. The monotonicity of a_k implies $a_k \leq 4Cn(kl)^2 = O(np^2)$. Similar algorithm and analysis can be applied to X with different size of groups.

2.4   *Numerical comparison study of different knockoff group selection methods*

In this subsection, we perform several numerical experiments to compare Reid–Tibshirani's prototype method, the prototype knockoff filter, the PCA prototype filter and Dai–Foygel Barber's group knockoff filter. Throughout the section, the group size is 5, the noise level is 1, the nominal FDR is 20% and we use the adaptive threshold T_1 defined in (3) and the knockoff+ selection procedure.

Simulated signals with no cancelation. We use the numerical example in [9] to compare several methods. The design matrix $X \in R^{3000 \times 1000}$ is clustered into 200 groups with 5 features in each group. The rows of X follow the $N(0, \Sigma)$ distribution with columns normalized, where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$ in the same group and $\Sigma_{ij} = \gamma \cdot \rho$ for $i \neq j$ in a different group. We choose 20 groups ($l = 20$) with one signal in each group. Specifically, we first choose l groups $i_1, i_2, ..., i_l$ randomly and then generate the signals β_j at indices $j = C_{i_1,1}, C_{i_2,1}, \ldots, C_{i_l,1}$ (the first feature in the selected groups) $\overset{\text{i.i.d.}}{\sim} \{\pm M\}$ and $\beta_j = 0$ for other indices. The signal amplitude M is 3.5.
For Reid–Tibshirani's prototype method and our prototype knockoff filter, we choose $n_1 = 0.2$ and $n = 600$, and split the data y and X into two parts as described in both methods. Apply the first part of the data to obtain the prototype for each group and then construct the SDP knockoffs on $\tilde{X}^{(2)}$. For these two methods and the PCA prototype filter, we use the orthogonal matching pursuit (OMP) statistic [1,13] and use the following short-hand notations *Reid–Tibshirani knockoff+*, *Prototype knockoff+* and *PCA knockoff+* in the following figures. For the group knockoff filter, the first method is to construct the equi-correlated group knockoffs and then apply the group LP (GLP) statistic, which is the method discussed in [9]. We also consider two other group selection methods based on the SDP group knockoffs and the OMP statistic. After constructing the SDP group knockoffs $\tilde{X}^g$, we extract the first principal

component of each group $X_{C_i}, \tilde{X}_{C_i}^g, i = 1, 2, \ldots, k$, which form $V, \tilde{V} \in R^{n \times k}$. We then run the knockoff filter on y and $[V, \tilde{V}]$ with the OMP statistic. This method is an analog of the PCA prototype filter with a different construction on $\tilde{V}$. Meanwhile, it is equivalent to the group knockoff with a special group knockoff statistic and thus the group FDR control follows from [9]. The other method uses a group version of the OMP statistic defined below. After selecting group $j_t, t \geq 0$, we define r_t to be the residual of the least square regression of y onto $\{X_{C_{j_1}}, X_{C_{j_2}}, \ldots, X_{C_{j_t}}\}$ ($r_0 = y$) and choose group j_{t+1} via $j_{t+1} = \arg\max_j ||r_t^T X_{C_j}||_2$. Let Z_j and $\tilde{Z}_j$ be the reversed order when the variable X_{C_j} or $\tilde{X}_{C_j}$ enters the model, e.g. Z_j ($\tilde{Z}_j$) $= 2k$ if X_{C_j} ($\tilde{X}_{C_j}$) enters the first, where k is the number of groups. We then define the group OMP statistic as $W_j = \max(Z_j, \tilde{Z}_j) \cdot sign(Z_j - \tilde{Z}_j), j = 1, 2, \ldots, k$. We use the following short-hand notations *Group knockoff+ GLP*, *PCA Group knockoff+* and *Group knockoff+ OMP* to stand for these three group knockoff methods in the following figures.

To study the effect of within-group correlation, we fix the between-group correlation factor $\gamma = 0$ and vary $\rho = 0, 0.1, \ldots, 0.9$. To study the effect of between-group correlation, we choose the within-group correlation factors $\rho = 0.5$ and $\rho = 0.9$ while varying $\gamma = 0, 0.1, \ldots, 0.9$. Each experiment is repeated 100 times. The GLP is calculated via the *gglasso* package [15] in R with number of λ equal to 1000.

For each design matrix X and its knockoff $\tilde{X}$, we consider the following measurements: average of $||X_{P_i} - \tilde{X}_{P_i}||_2^2/2 = s_{P_i}$ over all prototype features for the Reid–Tibshirani's prototype knockoff, the prototype knockoff, the PCA prototype knockoff, $||X_i - \tilde{X}_i^g||_2^2/2 = 2\lambda_{\min}(D\Sigma D)$ uniformly for the equi-correlated group knockoff and average of $||X_i - \tilde{X}_i^g||_2^2/2$ (i.e. average γ_i) over all features for the SDP group knockoff. In three experiments, the mean values of this average ($\bar{s}_{P_i}^{RT}, \bar{s}_{P_i}^{Prototype}, \bar{s}_{P_i}^{PCA}, \lambda_{\min}^{group}, \bar{\gamma}_i^g$) (10 design matrices in each experiment) are $(0.14, 0.50, 0.79, 0.36, 0.38)$, $(0.13, 0.39, 0.49, 0.29, 0.31)$ and $(0.03, 0.30, 0.44, 0.26, 0.28)$. We have performed numerical experiments for a general class of design matrices and found that the PCA prototype knockoff has the largest average difference, and the average difference $||X_{P_i} - \tilde{X}_{P_i}||_2^2$ of the prototype knockoff is slightly larger than that of the group knockoff. These two prototype methods overcome the problem of strong within-group correlation and improve the power of the Reid–Tibshirani's prototype method significantly.

The performance of the prototype knockoff filter is comparable to that of the group knockoff with the GLP and the OMP statistics. When the within-group correlation is strong and the between-group correlation factor γ is relatively small (the left and the right subfigures), the first principal component captures most of the information, and the PCA prototype filter outperforms the group knockoff filter with the GLP or the OMP statistic due to the larger difference between the prototype feature and its knockoff. In the middle subfigure where the within-group correlation is 0.5, the PCA prototype filter offers slightly less power than that of the group knockoff filter with the OMP statistic due to the relatively weak within-group correlation. Comparing the performance of the PCA knockoff+ and the PCA group knockoff+ in Fig. 1, we find that the PCA knockoff+ consistently offers more power than that of the PCA group knockoff+, which justifies that the knockoff construction that we propose indeed increases the power. The advantage of larger s_{P_i} in the PCA prototype knockoff can be exploited if we know *a priori* that the signals in each correlated group have the same sign. We have also constructed the equi-correlated group knockoff and then applied the group OMP statistic. It offers power similar to that of the Group knockoff+ OMP; see Fig. 1. For the later numerical experiments, we only focus on the power of different methods since these methods are guaranteed to control FDR.
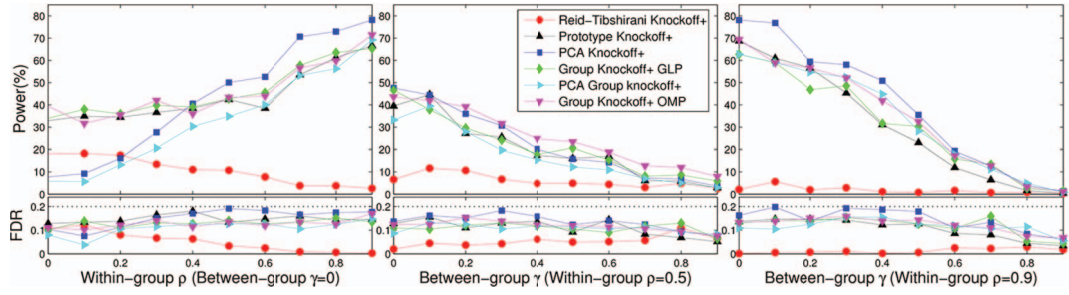
FIG. 1. Testing Reid–Tibshirani's prototype method, the prototype knockoff filter, the PCA prototype filter and the group knockoff filter with varying within-group correlation or between-group correlation. Here, we use the knockoff+ selection procedure.
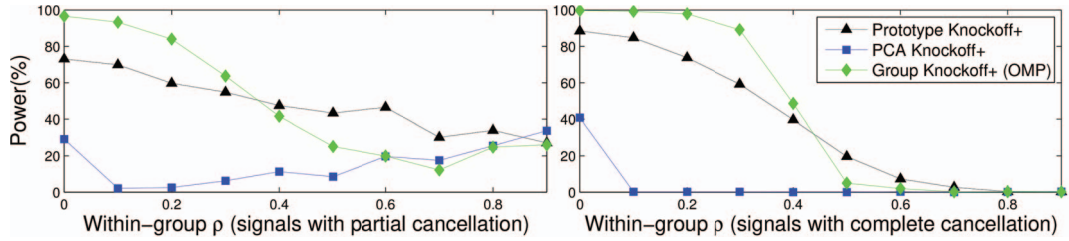

FIG. 2. Comparing the powers of the prototype knockoff filter, the PCA prototype knockoff filter and the group knockoff filter using signals with cancelation.

Simulated signals with cancelation. In the second example, we first generate $X \sim N(0, \Sigma)$ with within-group correlation $\rho = 0, 0.1, \ldots, 0.9$ and between-group correlation 0 as in the previous example. We then generate signals with cancelations in the first 20 correlated groups. We consider two settings of signal amplitude: (a) partial cancelation: $\beta_{C_i} = (a_i, -0.5a_i, 0, 0.., 0)$; (b) complete cancelation: $\beta_{C_i} = (a_i, -a_i, 0, 0.., 0)$, where $a_j \overset{\text{i.i.d.}}{\sim} \{\pm 5\}, i = 1, 2, \ldots, 20$. Other settings remain the same as in the previous example. We construct the SDP knockoff for the prototype knockoff filter and the SDP group knockoff, and focus on the prototype knockoff filter, the PCA prototype knockoff filter and the group knockoff filter with the OMP statistic.

In Fig. 2, when the within-group correlation is weak, the group knockoff filter captures more signals, and offers more power than that of the prototype knockoff filter and the PCA prototype knockoff filter. For large ρ, two signals almost merge into one signal and the prototype knockoff filter slightly outperforms the group knockoff. The PCA prototype filter loses considerable power in this case since the projection of the signals within the group onto the first principal component direction suffers from the cancelation. Without any prior knowledge about the signal, the first principal component may not be a good prototype, and we recommend to construct a data-dependent prototype via marginal correlation for the prototype knockoff filter.

In the following numerical experiments, without specification, we use the SDP construction for prototype knockoff filters and the group knockoff. We focus on the prototype knockoff filter, the PCA prototype knockoff filter, the group knockoff filter with the associated OMP statistic and the group knockoff filter with the GLP statistic.
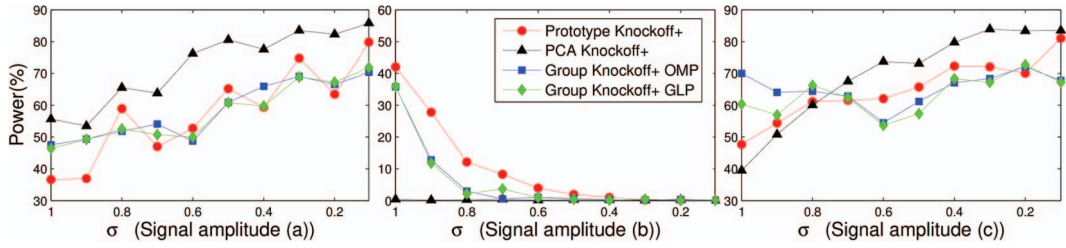
FIG. 3. Comparing the powers of the prototype knockoff filter, the PCA prototype knockoff filter and the group knockoff filter by varying within-group correlation and signal amplitude. Here, the group features are generated by one hidden factor.

Group features spanned by one hidden factor. For $j = 1, 2, \ldots, 100$, we generate group features $X_{C_j}(\xi) = (1, 1, \ldots, 1) \cdot \cos j\xi \in R^5$ with group size 5. In total, we have $p = 500$ features. We then generate $n = 3p = 1500$ i.i.d. realizations of ξ, $\xi \sim Unif[0, 2\pi]$, assemble $X_{C_j}(\xi)$ by rows and normalize the columns to obtain $X_{C_j} \in R^{n \times 5}$ and $X_0 = (X_{C_1}, X_{C_2}, \ldots, X_{C_{100}}) \in R^{n \times p}$. To avoid linear dependence, we perturb X_0 by some white noise: $X_0 + \sigma \cdot \tilde{G}$, where $\tilde{G} \in R^{n \times p}$ is obtained by normalizing the columns of $G \in R^{n \times p}$, $G_{ij} \overset{\text{i.i.d.}}{\sim} N(0, 1)$. We then normalize the columns of $X_0 + \sigma \cdot \tilde{G}$ to obtain the design matrix X and the modified group features matrix X_{C_j}. We select the last 15 groups (group with high frequency features) and generate the signal β_{C_j} in the selected group via one of the settings of signal amplitude (a) $\beta_{C_j} = (a_j, 0, 0, 0, 0)^T$, $a_j \overset{\text{i.i.d.}}{\sim} \{\pm 3.5\}$; (b) $\beta_{C_j} = (a_j, -a_j, 0, 0, 0)^T$, $a_j \overset{\text{i.i.d.}}{\sim} \{\pm 6\}$; (c) $\beta_{C_j} = (a_j, a_j, -a_j, 0, 0)^T$, $a_j \overset{\text{i.i.d.}}{\sim} \{\pm 3.5\}$. We then generate y as follows: $y = \sum_{j=86}^{100} X_{C_j} \beta_{C_j} + \varepsilon$, $\varepsilon \sim N(0, I_n)$. Note that the noise level is 1 and in setting (b) and (c), there are cancelations in the signals.

By definition of X_{C_j}, the within-group correlation mainly depends on σ. Smaller σ results in larger within-group correlation. We vary $\sigma = 1, 0.9, \ldots, 0.1$ and repeat 100 times for each experiment to compare the performance of several methods. The results are plotted in Fig. 3.

In the case of signal setting (a) and of signal setting (c) with small σ ($\sigma \leq 0.7$), the PCA prototype filter offers significantly more power than that of the group knockoff filter with both statistics. The performance of the prototype knockoff filter is comparable to that of the group knockoff filter and it offers more power in the case (b). The equi-correlated group knockoff with both statistics (not plotted) offers power similar to that of the SDP group knockoff.

Group features spanned by two hidden factors. We first generate $k = 100$ low frequencies $\omega_{L,j} \overset{\text{i.i.d.}}{\sim}$ $Unif[0, 10]$ and k high frequencies $\omega_{H,j} \overset{\text{i.i.d.}}{\sim} Unif[100, 200]$. The feature in group j is a convex combination of $\cos \omega_{L,j}\xi$ and $\cos \omega_{H,j}\xi$:

$$X_{C_j}(\xi, \tau_j) = (\tau_{j,1} \cos \omega_{L,j}\xi + (1 - \tau_{j,1}) \cos \omega_{H,j}\xi \; ..., \; \tau_{j,5} \cos \omega_{L,j}\xi + (1 - \tau_{j,5}) \cos \omega_{H,j}\xi) \in R^5,$$

where $\xi \sim Unif[0, 100]$ and $\tau_{j,i} \overset{\text{i.i.d.}}{\sim} Unif[0, \tau]$ for a parameter $\tau \in [0, 1]$. By definition, the within-group correlation mainly depends on τ. If $\tau = 0$, the features in group j are spanned by one factor $\cos \omega_{H,j}\xi$. Applying a procedure similar to the one used in the previous example, we can generate
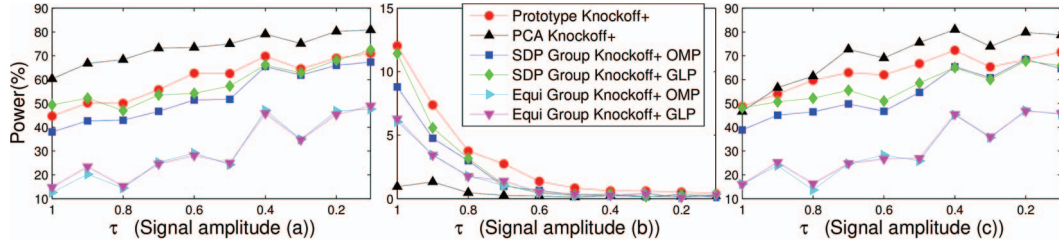
Fig. 4. Comparing the powers of the prototype knockoff filter, the PCA prototype knockoff filter and the group knockoff filter by varying within-group correlation and signal amplitude. Here, the group features are generated by two hidden factors.

$X_0 \in R^{1500 \times 500}$. To avoid linear dependence, we perturb X_0 by some white noise and normalize $X_0 + 0.3 \cdot \tilde{G}$ to obtain the design matrix X. Other settings, including 3 signal amplitudes and the sparsity (15 selected groups), remain the same as in the previous example. Due to the randomness in generating τ_j and the within-group correlation of X, we generate 5 $X_\tau \in R^{n \times p}$ for each $\tau = 1, 0.9, \ldots, 0.1$ and consider the average results. For each X_τ, we construct simulated data and repeat the experiment 100 times to obtain an average power and FDR. We plot the power averaged over five X_τ in Fig. 4.

In the case of (a) and (c), the PCA prototype filter outperforms the group knockoff filter and the prototype filter offers more power than that of the group knockoff filter almost in all cases. We can see that the equi-correlated group knockoff with the GLP or the OMP statistics (Equi Group knockoff+ GLP and OMP in Fig. 4) only offers about 50% power of the SDP group knockoff. In this example, the average of $||X_i - \tilde{X}_i^g||_2^2/2$ over all features of the equi-correlated group knockoff is 0.1402 and is much smaller than that of the SDP group knockoff, which is 0.3581. This explains the loss of considerable power.

From the last two examples, we observe that when the group features are spanned by one or a few hidden factors and the signals within each group are not canceled completely, the PCA prototype filter could offer more power than that of the group knockoff filter. If the within-group correlation is relatively strong and the signals are canceled completely as in the case (b) with $\sigma \leq 0.8$ in Figs 3 and 4, the information is lost in the measurement *y* and it is challenging to perform group selection in this case.

Computational efficiency. The computational cost of the prototype knockoff filter or the group knockoff mainly consists of the knockoff construction and the feature selection process. We can apply the equi-correlated or the SDP construction. In the equi-correlated construction, computing the smallest eigenvalue $\lambda_{\min}((W^T W)^{-1})$ or $\lambda_{\min}(D\Sigma D)$ is relatively cheap and both methods construct knockoffs in $O(np^2)$ flops (the computation of $X^T X$ is one of the bottlenecks). In the SDP construction of the prototype knockoff, (8) is a *k*-dimensional SDP (*k* is the number of groups). If $k \ll p$, the SDP construction of the prototype knockoff can be solved efficiently by exploiting its special structure [6]. We remark that the construction of the SDP knockoffs in [14] solves a *p*-dimensional problem of the same type. By default, we use the SDP construction in the prototype knockoff filter. The performance of the equi-correlated group knockoff depends on the amplitude of $\lambda_{\min}(D\Sigma D)$, which is exactly the average $||X_i - \tilde{X}_i||_2^2/4$ and could be much smaller than the corresponding average of the SDP group knockoff. In this case, the equi-correlated group knockoff could lose significant power as we have demonstrated in the last example, and one may have to construct the SDP group knockoff, which is more expensive.

The prototype knockoff filter constructs the knockoff statistic using $y, (X_P, \tilde{X}_P) \in R^{n \times 2k}$. The computational cost of many useful statistics, e.g. the LP, the OMP, the ridge regression and the Lasso statistics, is $O(nk^2)$, is relatively small compared with the cost in the knockoff construction.

3. Some observations of the knockoff filter

3.1 *Alternating sign effect*

In this section, we will perform some analysis for the knockoff filter and illustrate a potential challenge that we may encounter for some path method statistics, including the LP statistic and the forward selection statistic [1,10] for certain implementation procedure to update the residual. After performing l steps in one of the path methods (or at λ for the LP), we denote by E the set of features that have entered the model. Assume that $X_j, \tilde{X}_j \notin E$ at the lth step, but at the next step either X_j or $\tilde{X}_j$ will enter the model. After l steps, the residual is $r_l = y - X_E \hat{\beta}_E = X\beta - X_E \hat{\beta}_E + \varepsilon$. Since $X_j, \tilde{X}_j \notin E$, we have $X_j^T X_i = \tilde{X}_j X_i$, $\forall X_i \in E$. The same equality holds for $\tilde{X}_i$. For X_j and $\tilde{X}_j$, their marginal correlation with r_l determines which one of these two features will enter into the model first at the $(l+1)$st step:

$$X_j^T r_l = X_j^T (X\beta - X_E \hat{\beta}_E) + X_j^T \epsilon, \quad \tilde{X}_j^T r_l = \tilde{X}_j^T (X\beta - X_E \hat{\beta}_E) + \tilde{X}_j^T \epsilon,$$

$$(X_j - \tilde{X}_j)^T r_l = (X_j - \tilde{X}_j)^T y = s_j \beta_j + (X_j - \tilde{X}_j)^T \epsilon.$$

Assume that the noise level is relatively small. If

$$sign(\beta_j) \neq sign(X_j^T (X\beta - X_E \hat{\beta}_E)) \text{and} |X_j^T (X\beta - X_E \hat{\beta}_E)| > |s_j \beta_j|, \tag{11}$$

then $\tilde{X}_j$ will enter into the model at the $(l+1)$th step since

$$|\tilde{X}_j^T r_l| - |X_j^T r_l| \approx |X_j^T (X\beta - X_E \hat{\beta}_E) - s_j \beta_j| - |X_j^T (X\beta - X_E \hat{\beta}_E)|$$

$$= sign\left(X_j^T (X\beta - X_E \hat{\beta}_E)\right) \left[\left(X_j^T (X\beta - X_E \hat{\beta}_E) - s_j \beta_j\right) - \left(X_j^T (X\beta - X_E \hat{\beta}_E)\right) \right] = |s_j \beta_j| > 0.$$

This may reduce the power of the knockoff filter.

To understand under what condition the assumption (11) could be satisfied, we simply replace $X\beta - X_E \hat{\beta}_E$ by y. Then this assumption can be reformulated as $sign(X_j^T y) \neq sign(\beta_j)$ and $|X_j^T y| > |s_j \beta_j|$. Suppose that the correlation between the features is relatively strong. The direction X_j can capture more signals from y and thus it is likely that the marginal correlation $|X_j^T y|$ is larger than $|\beta_j| > |s_j \beta_j|$. This could occur if some features are positively (negatively) correlated, but their contribution to the response y has the opposite (same) sign, e.g. $X_j^T X_k > 0, \beta_j > 0$ and $\beta_k < 0$. We call this mechanism that could lead to (11) the *alternating sign effect*.

A numerical example. We generate the design matrix $X \in R^{900 \times 300} \sim N(0, \Sigma)$, where $\Sigma_{ij} = 0.9^{|i-j|}$, and construct its SDP knockoffs. Since some columns of X are strongly correlated, the knockoff factor $s \in R^{300}$ (defined in (1)) obtained in the SDP construction can be very small. Hence, some X_i is very close to its knockoff $\tilde{X}_i$, which may lead to the degeneracy of the augmented design matrix $[X \ \tilde{X}]$. Denote $S \triangleq \{i : s_i \leq 0.01\}$ and $L \triangleq \{i : s_i > 0.01\}$. We pick $k = 30$ features $i_1, i_2, \ldots, i_k$ randomly
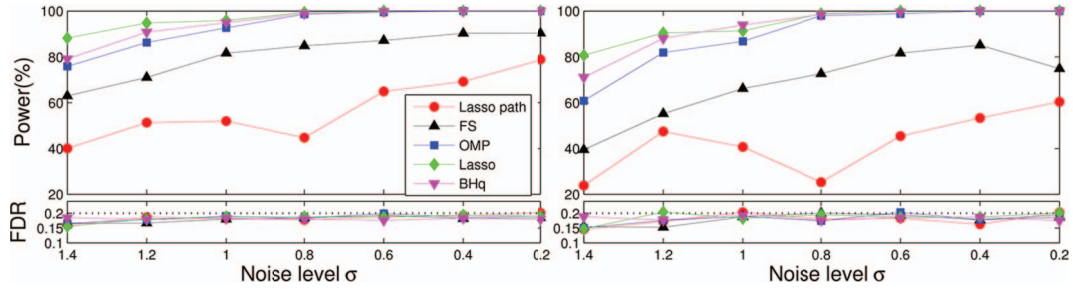
FIG. 5. Testing different knockoff statistics using example with positive correlated features and signals with different sign. In the left figure, for each σ, the power and the FDR are averaged over five cases. In the right figure, the minimal power over five cases and its associated FDR are presented.

in the set L and then generate the signal amplitude $\beta_{i_j} = \frac{0.75}{s_{i_j}}$. We then pick half of the signals β_{i_j} randomly and change their signs. By construction, 50% of the signals are positive. We construct $y = \sum_{j=1}^{k} X_{i_j}\beta_{i_j} + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$ and run the knockoff+ filter with the LP, the forward selection and the Lasso statistics with tuning parameter $\lambda = \sigma$ on y, $[X_L, \tilde{X}_L, X_S]$ to obtain the statistic W_L. We set $W_S = 0$. For the forward selection statistic, initializing $r_0 = y$, we iteratively choose X_{i_l} ($l \geq 1$) via $i_l = \arg\max_j |\langle r_{l-1}, X_j \rangle|$. According to [1], we can apply two different procedures to update the residual at step l. In the first procedure, the residual r_l is simply updated by eliminating the effect of the selected variable X_{i_l} from the previous residual r_{l-1}, while in the second procedure r_l is updated by eliminating the effect of all selected variables from y, i.e.

$$r_l^{(1)} = r_{l-1}^{(1)} - \langle r_{l-1}^{(1)}, X_{i_l}\rangle X_{i_l}, \quad r_l^{(2)} = y - P_l y, \tag{12}$$

where P_l is the orthogonal projector onto the space spanned by the l selected variables. For the forward selection statistic with these two procedures, we use short-hand notations *FS* and *OMP* since the latter uses OMP. The FS and OMP statistics can be computed by the knockoff package [1] directly. We also apply the Benjamini–Yekutieli (BHq) procedure [4] that first calculates the least-squares estimate $\hat{\beta}^{LS} = (X^T X)^{-1} X^T y$ and set $Z_j = \hat{\beta}_j^{LS}/(\sigma\sqrt{(G^{-1})_{jj}})$ to yield the z-scores, where $G = X^T X$ is the Gram matrix. Note that marginally $Z_j \sim N(0, 1)$. Variables are then selected by the threshold $T \triangleq \min\{t : 300 \cdot P(|N(0, 1)| \geq t)/\#\{j : |Z_j| \geq t\} \leq 20\%\}$.

Due to the randomness in generating the signal β, we generate five signals for each noise level $\sigma = 1.4, 1.2, \ldots, 0.2$ and consider the average and the extreme results. For each signal and σ, we repeat the experiment 100 times to obtain an average power and FDR. For each σ, we calculate the power, the FDR averaged over five cases with different signals, the minimal power and its associated FDR over five cases. The results are plotted in Fig. 5.

REMARK 3.1   In the selection procedure, we effectively turn off the knockoff $\tilde{X}_i$ for $i \in S$, i.e. s_i is small. If we run the knockoff process on y and the whole augmented design matrix M, the degeneracy of M can lead to significant numerical instability and the loss of FDR control. Note that $W_S = 0$ implies that the features in set S will not be selected by the knockoff filter. Since the signals are from L by the design of the response y, setting $W_S = 0$ will not lead to the loss of power.

In our computation, we use the knockoff package in Matlab to calculate the LP and the FS statistics. This package uses the *glmnet* package in Matlab [**12**] to solve the Lasso problem, and we also use the *glmnet* package to obtain the Lasso statistic. In an earlier example (not included in the current paper) where the features are strongly correlated, we obtained a somewhat unexpected result, i.e. the LP statistic fails to control FDR. To gain some understanding what went wrong, we found that the numerical solution of the Lasso problem $(\hat{\beta}, \tilde{\beta}) = \arg\min_{(\hat{b}, \tilde{b})} \frac{1}{2}||y - X\hat{b} - \tilde{X}\tilde{b}||_2^2 + \lambda||(\hat{b}, \tilde{b})||_1$ is significantly different from the numerical solution of $(\hat{\beta}, \tilde{\beta}) = \arg\min_{(\hat{b}, \tilde{b})} \frac{1}{2}||y - \tilde{X}\tilde{b} - X\hat{b}||_2^2 + \lambda||(\hat{b}, \tilde{b})||_1$, which is the same Lasso problem, except that we have swapped the order of the input variables $(X, \tilde{X})$. Therefore, when the features of the design matrix $[X\tilde{X}]$ are strongly correlated, the result of the Lasso-related methods obtained by the *glmnet* package could suffer from a large numerical error, which may lead to the loss of FDR control. The problem may be addressed by significantly increasing the precision of the solver, e.g. using the option $thresh = 10^{-12}$. To avoid this problem, we simply exclude $\tilde{X}_S$ when we run the knockoff selection procedure.

As the noise level σ decreases, the signal becomes relatively strong and we expect that the power increases. For $\sigma \leq 1$, the power of the BHq method is over 90% and its FDR is under control, which is a good indicator that the signal is strong enough. For $\sigma = 0.2, 0.4$, the minimal power of the knockoff with the LP $(53\%, 60\%)$ and the FS $(85\%, 75\%)$ statistics are significantly less than that of the BHq method $(100\%, 100\%)$, which suggests that these two statistics suffer from the alternating sign effect and lose significant power in this example. The power of the knockoff with the OMP or the Lasso statistic is comparable to that of the BHq method. A possible explanation of the robust performance of the OMP and the Lasso statistics in this example is that the OMP statistic can detect strong signals and then eliminate the effect of all selected variables before finding the next strong signal, and the Lasso statistic jointly estimates the effect of all regressors. Compared with the OMP, the forward selection with the first procedure in (12) (FS) fails to eliminate the effect of all selected variables, which may lead to the loss of considerable power. In strongly correlated cases, the OMP or the Lasso statistic is less likely to suffer from this effect. We have also implemented a similar test where the features are only weakly correlated: $X \sim N(0, \Sigma)$, $\Sigma_{ij} = 0.5^{|i-j|}$. The power of the knockoff with the LP and the FS statistics is comparable or more than that of the BHq method. Thus, it is unlikely that these statistics would lose power for weakly correlated features.

3.2 *Extension of the knockoff sufficiency property*

Let $U \in R^{n \times (n-2p)}$ be an orthonormal matrix such that $[X \tilde{X}]^T U = 0$ and $[X \tilde{X} U]$ admits a basis of R^n. (1) implies $(X + \tilde{X})^T (X - \tilde{X}) = X^T X - \tilde{X}^T \tilde{X} = 0$. Hence, R^n can be decomposed as follows:

$$R^n = span(X + \tilde{X}) \oplus span(X - \tilde{X}) \oplus span(U).$$

Our key observation is that swapping each pair of the original X_j and its knockoff $\tilde{X}_j$ does not modify these spaces: $span(X + \tilde{X})$, $span(X - \tilde{X})$ and $span(U)$. Therefore, the probability distributions of the projections of the response y onto these spaces are independent and invariant after swapping arbitrary pair X_j and $\tilde{X}_j$. Inspired by this observation, we can generalize the sufficiency property of knockoff statistic [**1**] which states that the statistic W depends only on the Gram matrix $[X\tilde{X}]^T[X\tilde{X}]$ and the feature–response product $[X\tilde{X}]^T y$.

DEFINITION 3.2 (Generalized sufficiency property). The statistic W is said to obey the generalized sufficiency property if W depends only on the Gram matrix $[X\tilde{X}]^T[X\tilde{X}]$ and the feature–response

$[X \tilde{X} U]^T y$; that is, we can write $W = f([X\tilde{X}]^T[X\tilde{X}], [X \tilde{X} U]^T y)$ for some $f : S_{2p}^+ \times R^n \to R^p$ and an orthonormal matrix $U \in R^{n \times (n-2p)}$ that satisfies $U^T[X\tilde{X}] = 0$.

The definition of the antisymmetry property remains the same: swapping X_j and $\tilde{X}_j$ has the same effect as changing the sign of W, i.e.

$$W_j([X\tilde{X}]_{swap(\hat{S})}, U, y) = W_j([X\tilde{X}], U, y) \cdot \begin{cases} +1 & j \notin \hat{S}, \\ -1 & j \in \hat{S}, \end{cases}$$

for any $\hat{S} \subset \{1, 2, \ldots, p\}$. For any knockoff matrix $\tilde{X}$ and the associated statistic W that satisfies the above definition, we call W the generalized knockoff statistic. Following the proof of Lemmas 1, 2 and 3 in [1], one can verify the pairwise exchangeability for the features and the response.

LEMMA 3.3   For any generalized knockoff statistic W and a subset $\hat{S}$ of nulls, we have

$$W_{swap(\hat{S})} = f([X\tilde{X}]_{swap(\hat{S})}^T [X\tilde{X}]_{swap(\hat{S})}, [[X\tilde{X}]_{swap(\hat{S})} U]^T y) \stackrel{d}{=} f([X\tilde{X}]^T[X\tilde{X}], [X \tilde{X} U]^T y) = W.$$

Moreover, we can show that the 'i.i.d. signs for the nulls' property still holds true for the generalized knockoff statistic.

LEMMA 3.4   (i.i.d. signs for the nulls) Let $\eta \in \{\pm 1\}^p$ be a sign sequence independent of W, with $\eta_j = +1$ for all non-null j and $\eta_j \stackrel{i.i.d.}{\sim} \{\pm 1\}$ for null j. Then $(W_1, ..., W_p) \stackrel{d}{=} (W_1 \eta_1, ..., W_p \eta_p)$.

Based on these lemmas, we can apply the same super-martingale as in [1] to establish rigorous FDR control.

Estimate of the noise level. A natural estimate of the noise level is $\hat{\sigma} = ||y - X\hat{\beta}^{ls} - \tilde{X}\tilde{\beta}^{ls}||_2/\sqrt{n - 2p}$ provided $n > 2p$, where $\hat{\beta}^{ls}$ and $\tilde{\beta}^{ls}$ are the least-squares coefficients. Let $U \in R^{n \times (n-2p)}$ be an orthonormal matrix such that $U^T[X\tilde{X}] = 0$. It is straightforward to show that

$$\hat{\sigma} = ||y - X\hat{\beta}^{ls} - \tilde{X}\tilde{\beta}^{ls}||_2/\sqrt{n - 2p} = ||U^T y||_2/\sqrt{n - 2p}.$$

Note that $\hat{\sigma}$ depends on $U^T y$ only. As an application of the generalized knockoff statistic, we can incorporate this noise estimate into the knockoff statistic without violating the FDR control.

4. Concluding remarks

In this paper, we proposed a prototype knockoff filter for group selection by extending Reid–Tibshirani's prototype method. Our prototype knockoff filter improves the computational efficiency and statistical power of the Reid–Tibshirani prototype method when it is applied for group selection. We demonstrated that when the group features are spanned by one or a few hidden factors, the PCA prototype filter can offer more power than that of the group knockoff filter. On the other hand, the PCA will not work well if the signals within each group have opposite signs and the signals in each group are canceled almost completely. Due to the improved statistical power and computational efficiency, the prototype knockoff filter with the SDP construction of the knockoff is especially attractive when the equi-correlation construction of the group knockoff gives small λ_{min} for certain design matrices. In this case, one may need to use the SDP construction for the group knockoff, which could be expensive.

We have also performed some analysis of the knockoff filter. Our analysis reveals that certain path method statistics for the knockoff filter may suffer from loss of power for certain design matrices and a specially designed response even if the signal strengths are relatively strong. We provided some partial understanding of this special phenomena and identified the alternating sign effect that could generate this phenomena. Our numerical results have confirmed that several statistics could lose significant power for certain design matrices and a specially constructed response due to the alternating sign effect.

REFERENCES

[1] BARBER, R. F. & CANDÈS, E. J. (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.

[2] BARBER, R. F. & CANDÈS, E. J. (2016) A knockoff filter for high-dimensional selective inference. *Ann. Statist.*, to appear. arXiv:1602.03574v1.

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

[4] BENJAMINI, Y. & YEKUTIELI, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

[5] BIEN, J., GAYNANOVA, I., LEDERER, J. & MÜLLER, C. L. (2018) Non-convex global minimization and false discovery rate control for the TREX. *J. Comput. Graph. Statist.*, **27**, 23–33.

[6] BOYD, S. & VANDENBERGHE, L. (2004) *Convex Optimization*. Cambridge: Cambridge University Press, MR2061575.

[7] CANDÈS, E. J., FAN, Y., JANSON, L. & LV, J. (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. Roy. Statist. Soc. Ser. B*, **80**, 551–577.

[8] CHEN, J., HOU, A. & HOU, T. Y. (2018) A pseudo knockoff filter for correlated features. *Information and Inference: A Journal of the IMA*, iay012, https://doi.org/10.1093/imaiai/iay012.

[9] DAI, R. & BARBER, R. F. (2016) The knockoff filter for FDR control in group-sparse and multitask regression. *Proceedings of The 33rd International Conference on Machine Learning, PMLR 48:1851-1859, 2016*. New York, NY, USA.

[10] DRAPER, N. R. & SMITH, H. (1981) *Applied Regression Analysis*, 2nd edn. New York: Wiley, MR0610978.

[11] JANSON, L. & SU, W. (2016) Familywise error rate control via knockoffs. *Electron. J. Stat.*, **10**, 960–975.

[12] QIAN, J., HASTIE, T., FRIEDMAN, J., TIBSHIRANI, R. & SIMON, N. (2013) *Glmnet for Matlab*. http://www.stanford.edu/hastie/glmnet_matlab/.

[13] PATI, Y. C., REZAIIFAR, R. & KRISHNAPRASAD, P. S. (1993) Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. Los Alamitos: IEEE, 40–44.

[14] REID, S. & TIBSHIRANI, R. (2016) Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, **17**, 364–376.

[15] YANG, Y. & ZOU, H. (2015) A fast unified algorithm for solving group-lasso penalize learning problems. *Statist. Comput.*, **25**, 1129–1141.