The knockoff filter for FDR control in group-sparse and multitask regression

Ran Dai   RANDAI@UCHICAGO.EDU
Department of Statistics, University of Chicago, Chicago IL 60637 USA

Rina Foygel Barber   RINA@UCHICAGO.EDU
Department of Statistics, University of Chicago, Chicago IL 60637 USA

Abstract

We propose the group knockoff filter, a method for false discovery rate control in a linear regression setting where the features are grouped, and we would like to select a set of relevant groups which have a nonzero effect on the response. By considering the set of true and false discoveries at the group level, this method gains power relative to sparse regression methods. We also apply our method to the multitask regression problem where multiple response variables share similar sparsity patterns across the set of possible features. Empirically, the group knockoff filter successfully controls false discoveries at the group level in both settings, with substantially more discoveries made by leveraging the group structure.

1. Introduction

In a high-dimensional regression setting, we are faced with many potential explanatory variables (features), often with most of these features having zero or little true effect on the response. Model selection methods can be applied to find a small submodel containing the most relevant features, for instance, via sparse model fitting methods such as the lasso (Tibshirani, 1996), or in a setting where the sparsity respects a grouping of the features, the group lasso (Yuan & Lin, 2006). In practice, however, we may not be able to determine whether the set of features (or set of groups of features) selected might contain many false positives. For the (non-grouped) sparse setting, the knockoff filter (Barber & Candès, 2015) creates "knockoff copies" of each variable to act as a control group, detecting whether the lasso (or another model selection method) is successfully controlling the false discovery rate (FDR), and tuning this

method to find a model as large as possible while bounding FDR. In this work, we will extend the knockoff filter to the group sparse setting, and will find that by considering features, and constructing knockoff copies, at the group-wise level, we are able to improve the power of this method at detecting true signals. Our method can also extend to the multitask regression setting (Obozinski et al., 2006), where multiple responses exhibit a shared sparsity pattern when regressed on a common set of features. As for the knockoff method, our work applies to the setting where $n \geq p$.

2. Background

We begin by giving background on several models and methods underlying our work.

2.1. Group sparse linear regression

We consider a linear regression model, $Y = X\beta + z$, where $y \in \mathbb{R}^n$ is a vector of responses and $X \in \mathbb{R}^{n \times p}$ is a known design matrix. In a grouped setting, the p features are partitioned into m groups of variables, $G_1, \ldots, G_m \subseteq \{1, \ldots, p\}$, with group sizes $p_1, \cdots, p_m$. The noise distribution is assumed to be $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. We assume sparsity structure in that only a small portion of β_{G_i}'s are nonzero, where $\beta_{G_i} \in \mathbb{R}^{p_i}$ is the subvector of β corresponding to the ith group of features. When not taking group into consideration, a commonly used method to find a sparse vector of coefficients β is the lasso (Tibshirani, 1996), an ℓ_1-penalized linear regression, which minimizes the following objective: function

$$\widehat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\} . \quad (1)$$

To utilize the feature grouping, so that an entire group of features is selected simultaneously, Yuan & Lin (2006) proposed following grouped lasso penalties:

$$\widehat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_{\text{group}} \right\} . \quad (2)$$

where $\|\beta\|_{\text{group}} = \sum_{i=1}^m \|\beta_{G_i}\|_2$. This penalty promotes sparsity at the group level; for large λ, few groups will

be selected (i.e. β_{G_i} will be zero for many groups), but within any selected group, the coefficients will be dense (all nonzero). The ℓ_2 norm penalty on β_{G_i} may sometimes be rescaled relative to the size of the group.

2.2. Multitask learning

In a multitask learning problem with a linear regression model, we consider the model

$$Y = XB + E \tag{3}$$

where the response $Y \in \mathbb{R}^{n \times r}$ contains r many response variables measured for n individuals, $X \in \mathbb{R}^{n \times p}$ is the design matrix, $B \in \mathbb{R}^{p \times r}$ is the coefficient matrix, and $E \in \mathbb{R}^{n \times r}$ is the error matrix, with rows $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ for some unknown covariance structure $\Sigma \in \mathbb{R}^{r \times r}$. If the number of features p is large, we may believe that only a few of the features are relevant; in that case, most rows of B will be zero—that is, B is row-sparse.

In a low-dimensional setting, we may consider the multivariate normal model, with likelihood determined by both the coefficient matrix B and the covariance matrix Σ. in a high-dimensional setting, combining this likelihood with a sparsity-promoting penalty may be computationally challenging, and so a common approach is to ignore the covariance structure of the noise and to simply use a least-squares loss together with a penalty,

$$\widehat{B} = \arg\min_B \left\{ \frac{1}{2} \|Y - XB\|_{\text{Fro}}^2 + \lambda \|B\|_{\ell_1/\ell_2} \right\}, \tag{4}$$

where $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm, and where the ℓ_1/ℓ_2 norm in the penalty is given by $\|B\|_{\ell_1/\ell_2} = \sum_i \sqrt{\sum_j B_{ij}^2}$. This penalty promotes row-wise sparsity of B.

It is common to reformulate this ℓ_1-penalized multitask linear regression as a group lasso problem. We form a vector response $y \in \mathbb{R}^{nr}$ by stacking the columns of Y:

$$y = \text{vec}(Y) = (Y_{11}, \ldots, Y_{n1}, \ldots, Y_{1r}, \ldots, Y_{nr})^\top \in \mathbb{R}^{nr},$$

and a new larger design matrix by repeating X in blocks:

$$\mathbb{X} = \mathbf{I}_r \otimes X = \begin{pmatrix} X & 0 & \ldots & 0 \\ 0 & X & \ldots & 0 \\ & & \ldots & \\ 0 & 0 & \ldots & X \end{pmatrix} \in \mathbb{R}^{nr \times pr}.$$

(Here $\otimes$ is the Kronecker product.) Define the coefficient vector $\beta = \text{vec}(B) \in \mathbb{R}^{pr}$ and noise vector $\epsilon = \text{vec}(E) \in \mathbb{R}^{nr}$. Then the multitask model (3) can be rewritten as

$$y = \mathbb{X}\beta + \epsilon, \tag{5}$$

where ϵ follows a Gaussian model, $\epsilon \sim \mathcal{N}(0, \mathbf{\Sigma})$, for

$$\mathbf{\Sigma} = \Sigma \otimes \mathbf{I}_n = \begin{pmatrix} \Sigma_{11}\mathbf{I}_n & \ldots & \Sigma_{1r}\mathbf{I}_n \\ \ldots & \ldots & \ldots \\ \Sigma_{r1}\mathbf{I}_n & \ldots & \Sigma_{rr}\mathbf{I}_n \end{pmatrix}.$$

The group sparse structure of β is determined by groups

$$G_j = \{j, j+p, \ldots, j+(r-1)p\}$$

for $j = 1, \ldots, p$; this corresponds to the row sparsity of B in the original formulation (3). Then, the multitask learning problem has been reformulated into a group-sparse regression problem—and so, the multitask lasso (4) can equivalently be solved by the group lasso optimization problem

$$\widehat{\beta} = \arg\min_\beta \left\{ \frac{1}{2} \|y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_{\text{group}} \right\}. \tag{6}$$

2.3. The group false discovery rate

The original definition of false discovery rate (FDR) is the expected proportion of incorrectly selected features among all selected features. When the group rather than individual feature is of interest, we prefer to control the false discovery rate at the group level. Mathematically, we define the group false discovery rate (FDR$_{\text{group}}$) as

$$\text{FDR}_{\text{group}} = \mathbb{E}\left[\frac{\#\{i : \beta_i = 0, i \in \widehat{S}\}}{\#\{i : i \in \widehat{S}\} \vee 1} \right] \tag{7}$$

the expected proportion of selected groups which are actually false discoveries. Here $\widehat{S} = \{i : \widehat{\beta}_i \neq 0\}$ is the set of all selected group of features, while $a \vee b$ denotes $\max\{a, b\}$.

2.4. The knockoff filter for sparse linear regression

In the sparse (rather than group-sparse) setting, the lasso (1) provides an accurate estimate for the coefficients in a sparse linear model, but performing inference on the results, for testing the accuracy of these estimates or the set of features selected, remains a challenging problem. The knockoff filter (Barber & Candès, 2015) addresses this question, and provides a method controlling the false discovery rate (FDR) of the selected set at some desired level q (e.g. $q = 0.2$).

To run this method, there are two main steps: constructing knockoffs, and filtering the results. First a set of p knockoff features is constructed: for each feature X_j, $j = 1, \ldots, p$, it is given a knockoff copy $\widetilde{X}_j$, where the matrix of knockoffs $\widetilde{X} = [\widetilde{X}_1 \ldots \widetilde{X}_p]$ satisfies, for some vector $s \geq 0$,

$$\widetilde{X}^\top \widetilde{X} = X^\top X, \ \widetilde{X}^\top X = X^\top X - \text{diag}\{s\}. \tag{8}$$

Next, the lasso is run on an augmented data set with response y and $2p$ many features $X_1, \ldots, X_p, \widetilde{X}_1, \ldots, \widetilde{X}_p$:

$$\widehat{\beta}(\lambda) = \arg\min_{b \in \mathbb{R}^{2p}} \left\{ \|y - [X \ \widetilde{X}]b\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

This is run over a range of λ values decreasing from $+\infty$ (a fully sparse model) to 0 (a fully dense model). If X_j is a

true signal—that is, it has a nonzero effect on the response y—then this should be evident in the lasso: X_j should enter the model earlier (for larger λ) than its knockoff copy $\widetilde{X}_j$. However, if X_j is null—that is, $\beta_j = 0$ in the true model— then it is equally likely to enter before or after $\widetilde{X}_j$.

Next, to filter the results, let λ_j and $\widetilde{\lambda}_j$ be the time of entry into the lasso path for each feature and knockoff:

$$\lambda_j = \sup\{\lambda : \widehat{\beta}(\lambda)_j \neq 0\}, \widetilde{\lambda}_j = \sup\{\lambda : \widehat{\beta}(\lambda)_{j+p} \neq 0\},$$

and let $\widehat{S}(\lambda), \widetilde{S}(\lambda) \subseteq \{1, \ldots, p\}$ be the sets of original features, and knockoff features, which have entered the lasso path before time λ, and before their counterparts:

$$\widehat{S}(\lambda) = \{j : \lambda_j > \widetilde{\lambda}_j \vee \lambda\} \text{ and } \widetilde{S}(\lambda) = \{j : \widetilde{\lambda}_j > \lambda_j \vee \lambda\}.$$

Estimate the proportion of false discoveries in $\widehat{S}(\lambda)$ as

$$\text{FDP}(\lambda) \approx \widehat{\text{FDP}}(\lambda) = \frac{|\widetilde{S}(\lambda)|}{|\widehat{S}(\lambda)| \vee 1}. \tag{9}$$

To understand why, note that since X_j and $\widetilde{X}_j$ are equally likely to enter in either order if X_j is null (no real effect), then j is equally likely to fall into either $\widehat{S}(\lambda)$ or $\widetilde{S}(\lambda)$. Therefore, the numerator $|\widetilde{S}(\lambda)|$ should be an (over)estimate of the number of nulls in $\widehat{S}(\lambda)$—thus, the ratio estimates the FDP. Alternately, we can choose a more conservative definition

$$\text{FDP}(\lambda) \approx \widehat{\text{FDP}}_+(\lambda) = \frac{1 + |\widetilde{S}(\lambda)|}{|\widehat{S}(\lambda)| \vee 1}. \tag{10}$$

Finally, the knockoff filter selects $\widehat{\lambda} = \min\{\lambda : \widehat{\text{FDP}}(\lambda) \leq q\}$, where q is the desired bound on FDR level, and then outputs the set $\widehat{S}(\widehat{\lambda})$ as the set of "discoveries". The knockoff+ variant does the same with $\widehat{\text{FDP}}_+(\lambda)$. Theorems 1 and 2 of (Barber & Candès, 2015) prove that the knockoff procedure bounds a modified form of the FDR, $\text{mFDR} = \mathbb{E}\left[\frac{(\# \text{ of false discoveries})}{(\# \text{ of discoveries}) + q^{-1}}\right]$, while the knockoff+ procedure bounds the FDR.

3. The knockoff filter for group sparsity

In this section, we extend the knockoff method to the group sparse setting. This involves two key modifications: the construction of the knockoffs at a group-wise level rather than for individual features, and the "filter" step where the knockoffs are used to select a set of discoveries. Throughout the remainder of the paper, "knockoff" refers to the original knockoff method, while "group knockoff" (or, later on, "multitask knockoff") refers to our new method.

3.1. Group knockoff construction

The original knockoff construction requires that $\widetilde{X}^\top X = X^\top X - \text{diag}\{s\}$, that is, all off-diagonal entries are equal. When the features are highly correlated, this construction is only possible for vectors s with extremely small entries; that is, $\widetilde{X}_j$ and X_j are themselves highly correlated, and the knockoff filter then loses power as it is hard to distinguish between a real signal X_j and its knockoff copy $\widetilde{X}_j$.

In a group-sparse setting, we will see that we can relax this requirement on $\widetilde{X}^\top X$, thereby improving our power. In particular, the best gain will be in situations where within-group correlations are high but between-group correlations are low; this may arise in many applications, for example, when genes related to the same biological pathways are grouped together, we expect to see the largest correlations occuring within groups rather than between genes in different groups.

To construct the group knockoffs, we require the following condition on the matrix $\widetilde{X} \in \mathbb{R}^{n \times p}$:

$$\widetilde{X}^\top \widetilde{X} = \Sigma := X^\top X, \text{ and } \widetilde{X}^\top X = \Sigma - S,$$
$$\text{where } S \succeq 0 \text{ is group-block-diagonal}, \tag{11}$$

meaning that $S_{G_i, G_j} = 0$ for any two distinct groups $i \neq j$. Abusing notation, write $S = \text{diag}\{S_1, \ldots, S_m\}$ where $S_i \succeq 0$ is the $p_i \times p_i$ matrix for the ith group block, meaning that $S_{G_i, G_i} = S_i$ for each i while $S_{G_i, G_j} = 0$ for each $i \neq j$. Extending the construction of (Barber & Candès, 2015),[1] we construct these knockoffs by first selecting $S = \text{diag}\{S_1, \ldots, S_m\}$ that satisfies the condition $S \preceq 2\Sigma$, then setting $\widetilde{X} = X(\mathbf{I}_p - \Sigma^{-1}S) + \widetilde{U}C$, where $\widetilde{U}$ is a $n \times p$ orthonormal matrix orthogonal to the span of X, while $C^\top C = 2S - S\Sigma^{-1}S$ is a Cholesky decomposition. Now, we still need to choose the matrix $S \succeq 0$, which has group-block-diagonal structure, so that the condition $S \preceq 2\Sigma$ is satisfied (this condition ensures the existence of the Cholesky decomposition defining C). To do this, we choose the following construction: we set $S = \text{diag}\{S_1, \ldots, S_m\}$ where we choose $S_i = \gamma \cdot \Sigma_{G_i, G_i}$; the scalar $\gamma \in [0, 1]$ is chosen to be as large as possible so that $S \preceq 2\Sigma$ still holds, which amounts to choosing

$$\gamma = \min\left\{1, 2 \cdot \lambda_{\min}\left(D\Sigma D\right)\right\}$$

where $D = \text{diag}\{\Sigma_{G_1, G_1}^{-1/2}, \ldots, \Sigma_{G_m, G_m}^{-1/2}\}$. This construction can be viewed as an extension of the "equivariant" knockoff construction of Barber & Candès (2015); their SDP construction, which gains a slight power increase in the non-grouped setting, may also be extended to the grouped setting but we do not explore this here.

---

[1]This construction is for the setting $n \geq 2p$; see (Barber & Candès, 2015) for a simple trick to extend to $n \geq p$.

Looking back at the group knockoff matrix condition (11), we see that any knockoff matrix $\widetilde{X}$ satisfying (8) would necessarily also satisfy this group-level condition. However, the group-level condition is weaker; it allows more flexibility in constructing $\widetilde{X}$, and therefore, will enable more separation between a feature X_j and its knockoff $\widetilde{X}_j$, which in turn can increase power to detect the true signals.

3.2. Filter step

After constructing the group knockoff matrix, we then select a set of discoveries (at the group level) as follows. First, we apply the group lasso (2) to the augmented data set,

$$\widehat{\beta} = \arg\min_{b \in \mathbb{R}^{2p}} \left\{ \|y - [X \ \widetilde{X}]b\|_2^2 + \lambda\|b\|_{\text{group}} \right\}.$$

Here, with the augmented design matrix $[X \ \widetilde{X}]$, we now have $2m$ many groups: one group G_i for each group in the original design matrix, and one group $\widetilde{G}_i = \{j + p : j \in G_i\}$ corresponding to the same group within the knockoff matrix; the penalty norm is then defined as $\|b\|_{\text{group}} = \sum_{i=1}^m \|b_{G_i}\|_2 + \sum_{i=1}^m \|b_{\widetilde{G}_i}\|_2$.

The filter process then proceeds exactly as for the original knockoff method, with groups of features in place of individual features. First we record the time when each group or knockoff group enters the lasso path,

$$\lambda_i = \sup\{\lambda : \widehat{\beta}(\lambda)_{G_i} \neq 0\}, \widetilde{\lambda}_i = \sup\{\lambda : \widehat{\beta}(\lambda)_{\widetilde{G}_i} \neq 0\},$$

then define the selected groups and knockoff groups as

$$\widehat{S}(\lambda) = \{i : \lambda_i > \widetilde{\lambda}_i \vee \lambda\} \text{ and } \widetilde{S}(\lambda) = \{i : \widetilde{\lambda}_i > \lambda_i \vee \lambda\}$$

(note that these sets are subsets of $\{1, \ldots, m\}$, the list of groups, rather than counting individual features). Finally, estimate the proportion of false discoveries in $\widehat{S}(\lambda)$ exactly as in (9), and define $\widehat{\lambda} = \min\{\lambda : \widehat{\text{FDP}}(\lambda) \leq q\}$ as before; the final set of discovered groups is given by $\widehat{S}(\widehat{\lambda})$. (For group knockoff+, we use the more conservative estimate of the group FDP, as for the knockoff.)

3.3. Theoretical Results

Here we turn to a more general framework for the group knockoff, working with the setup introduced in Barber & Candès (2015). Let $W \in \mathbb{R}^m$ be a vector of statistics, one for each group, with large positive values for W_i indicating strong evidence that group i may have a nonzero effect (i.e. $\beta_{G_i} \neq 0$). W is defined as a function of the augmented design matrix $[X \ \widetilde{X}]$ and the response y, which we write as $W = w([X \ \widetilde{X}], y)$. In the group lasso setting described above, the statistic is given by $W_i = (\lambda_i \vee \widetilde{\lambda}_i) \cdot \text{sign}(\lambda_i - \widetilde{\lambda}_i)$. In general, we require two properties for this statistic: sufficiency and group-antisymmetry. The first is exactly as for

(non-group) knockoffs; the second is a modification moving to the group sparse setting.

Definition 1. The statistic W is said to obey the sufficiency property if it only depends on the Gram matrix and feature-response inner products, that is, for any $X, \widetilde{X}, y$,

$$w([X \ \widetilde{X}], y) = f([X \ \widetilde{X}]^\top[X \ \widetilde{X}], [X \ \widetilde{X}]^\top y) \quad (12)$$

for some function f.

Before defining the group-antisymmetry property, we introduce some notation. For any group $i = 1, \ldots, m$, let $[X \ \widetilde{X}]_{\text{swap}(i)}$ be the matrix with

$$\left([X \ \widetilde{X}]_{\text{swap}(i)}\right)_j = \begin{cases} X_j, & \text{if } 1 \leq j \leq p \text{ and } j \notin G_i, \\ \widetilde{X}_j, & \text{if } 1 \leq j \leq p \text{ and } j \in G_i, \end{cases}$$

and

$$\left([X \ \widetilde{X}]_{\text{swap}(i)}\right)_{j+p} = \begin{cases} \widetilde{X}_j, & \text{if } 1 \leq j \leq p \text{ and } j \notin G_i, \\ X_j, & \text{if } 1 \leq j \leq p \text{ and } j \in G_i, \end{cases}$$

for each $j = 1, \ldots, p$. In other words, the columns corresponding to G_i in the original component X, are swapped with the same columns of $\widetilde{X}$.

Definition 2. The statistic W is said to obey the group-antisymmetry property if swapping two groups X_i and $\widetilde{X}_i$ has the effect of switching the sign of W_i with no other change to W, that is,

$$w([X \ \widetilde{X}]_{\text{swap}(i)}, y) = \mathbf{I}_i^\pm \cdot w([X \ \widetilde{X}], y),$$

where $\mathbf{I}_i^\pm$ is the diagonal matrix with a -1 in entry (i, i) and $+1$ in all other diagonal entries.

Next, to run the group knockoff or group knockoff+ method, we proceed exactly as in (Barber & Candès, 2015); we change notation here for better agreement with the group lasso setting. Define

$$\widehat{S}(t) = \{i : W_i \geq t\} \text{ and } \widetilde{S}(t) = \{i : W_i \leq -t\}.$$

Then estimate the FDP as in (9) for the knockoff method, or as in (10) for knockoff+ (with parameter t in place of the lasso penalty path parameter λ); then find $\widehat{t}$, the minimum $t \geq 0$ with $\widehat{\text{FDP}}(t)$ (or $\widehat{\text{FDP}}_+(t)$) no larger than q, and output the set $\widehat{S} = \widehat{S}(\widehat{t})$ of discovered groups.

This procedure offers the following theoretical guarantee:

Theorem 1. *If the vector of statistics W satisfies the sufficiency and group-antisymmetry assumption, then the group knockoff procedure controls a modified group FDR,*

$$\text{mFDR}_{\text{group}} = \mathbb{E}\left[\frac{\#\{i : \beta_i = 0, i \in \widehat{S}\}}{\#\{i : i \in \widehat{S}\} + q^{-1}}\right] \leq q,$$

while the group knockoff+ procedure controls the group FDR, $\text{FDR}_{\text{group}} \leq q$.

The proof of this result follows the original knockoff proof of Barber & Candès (2015), and we do not reproduce it here; the result is an immediate consequence of their main lemma, moved into the grouped setting:

Lemma 1. *Let $\epsilon \in \{\pm 1\}^m$ be a sign sequence independent of W, with $\epsilon_i = 1$ for all non-null groups i and $\epsilon_i \sim \{\pm 1\}$ independently with equal probability for all null groups i. Then we have*

$$(W_1, \cdots, W_m) =_d (W_1 \epsilon_1, \cdots, W_m \epsilon_m), \qquad (13)$$

where $=_d$ denotes equality in distribution.

This lemma can be proved via the sufficiency and group-antisymmetry properties, exactly as for the individual-feature-level result of Barber & Candès (2015).

4. Knockoffs for multitask learning

For the multitask learning problem, the reformulation as a group lasso problem (6) suggests that we can apply the group-wise knockoffs to this problem as well. However, there is one immediate difficulty: the model for the noise ϵ in (6) has changed—the entries of ϵ are not independent, but instead follow a multivariate Gaussian model with co-variance Σ. In fact, we will see shortly that we can work even in this more general setting. Reshaping the data to form a group lasso problem as in (5), we will work with the vectorized response $y \in \mathbb{R}^{nr}$ and the repeated-block design matrix $\mathbb{X} \in \mathbb{R}^{nr \times pr}$. We will also construct a repeated-block knockoff matrix,

$$\widetilde{\mathbb{X}} = \mathbf{I}_r \otimes \widetilde{X} = \begin{pmatrix} \widetilde{X} & 0 & \ldots & 0 \\ 0 & \widetilde{X} & \ldots & 0 \\ & & \ldots & \\ 0 & 0 & \ldots & \widetilde{X} \end{pmatrix},$$

where $\widetilde{X} \in \mathbb{R}^{n \times p}$ is any matrix satisfying the original knockoff construction conditions (8) with respect to the original design matrix X. Applying the group knockoff methodology with this data $(\mathbb{X}, y)$ and knockoff matrix $\widetilde{\mathbb{X}}$, we obtain the following result:

Theorem 2. *For the multitask learning setting with an arbitrary covariance structure $\Sigma \in \mathbb{R}^{r \times r}$, the knockoff or knockoff+ methods control the modified group FDR or the group FDR, respectively, at the level q.*

Proof. In order to apply the result for the group-sparse setting to this multitask scenario, we need to address two questions: first, whether $\widetilde{\mathbb{X}}$ satisfies the group knockoff matrix conditions (11), and second, how to handle the issue of the non-i.i.d. structure of the noise ϵ.

We first check the conditions (11) for $\widetilde{\mathbb{X}}$. let $\widetilde{X} \in \mathbb{R}^{n \times p}$ be a knockoff matrix for X, satisfying (8), and let $\Sigma = X^\top X$.

Then we see that

$$\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}} = \mathbf{I}_r \otimes (\widetilde{X}^\top \widetilde{X}) = \mathbf{I}_r \otimes \Sigma = \mathbb{X}^\top \mathbb{X}, \text{ and}$$
$$\widetilde{\mathbb{X}}^\top \mathbb{X} = \mathbf{I}_r \otimes (\widetilde{X}^\top X) = \mathbf{I}_r \otimes (\Sigma - \text{diag}\{s\})$$
$$= \mathbb{X}^\top \mathbb{X} - \mathbf{I}_r \otimes \text{diag}\{s\}$$

where s is defined as in (8). Since the difference $\mathbf{I}_r \otimes \text{diag}\{s\}$ is a diagonal matrix, we see that $\widetilde{\mathbb{X}}$ satisfies the group knockoff condition (11); in fact, it satisfies the stronger (ungrouped) knockoff condition (8).

Next we turn to the issue of the non-identity covariance structure $\mathbb{\Sigma}$ for the noise term $\epsilon \in \mathbb{R}^{nr}$. First, write

$$\mathbb{\Sigma}^{-1/2} = \Sigma^{-1/2} \otimes \mathbf{I}_n$$

to denote an inverse square root for $\mathbb{\Sigma}$. Note also that

$$\mathbb{\Sigma}^{-1/2} \cdot \mathbb{X} = (\Sigma^{-1/2} \otimes \mathbf{I}_n) \cdot (\mathbf{I}_r \otimes X) = \Sigma^{-1/2} \otimes X$$
$$= (\mathbf{I}_r \otimes X) \cdot (\Sigma^{-1/2} \otimes \mathbf{I}_p) = \mathbb{X} \cdot \mathbb{\Sigma}_*^{-1/2}, \quad (14)$$

for $\mathbb{\Sigma}_* = \Sigma \otimes \mathbf{I}_p$. Taking our vectorized multitask regression model (5), multiplying both sides by $\mathbb{\Sigma}^{-1/2}$ on the left, and applying (14), we obtain a "whitened" reformulation of our model,

$$y^{\text{wh}} = \mathbb{X} \cdot (\mathbb{\Sigma}_*^{-1/2} \beta) + \epsilon^{\text{wh}} \text{ for } \begin{cases} y^{\text{wh}} = \mathbb{\Sigma}^{-1/2} y, \\ \epsilon^{\text{wh}} = \mathbb{\Sigma}^{-1/2} \epsilon, \end{cases} \quad (15)$$

where $\epsilon^{\text{wh}} \sim \mathcal{N}(0, \mathbf{I}_{nm})$ is the "whitened" noise. Now we are back in a standard linear regression setting, and can apply the knockoff method—note that we are working with a new setup: while the design matrix $\mathbb{X}$ is the same as in (5), we now work with response vector y^{wh} and coefficient vector $\mathbb{\Sigma}_*^{-1/2} \beta$. The group sparsity of the coefficient vector has not changed, due to the block structure of $\mathbb{\Sigma}^{-1/2}$; we have

$$(\mathbb{\Sigma}_*^{-1/2} \beta)_{G_j} = \mathbb{\Sigma}_{G_j, G_j}^{-1/2} \beta_{G_j}$$

for each $j = 1, \ldots, p$, and so the "null groups" for the original coefficient vector β (i.e. groups j with $\beta_{G_j} = 0$) are preserved in this reformulated model.

We need to check only that the group lasso output, namely $\widehat{\beta}$, depends on the data only through the sufficient statistics $\mathbb{X}^\top \mathbb{X}$ and $\mathbb{X}^\top y^{\text{wh}}$; here we use the "whitened" response y^{wh} rather than the original response vector y since the knockoff theory applies to linear regression with i.i.d. Gaussian noise, as in the model (15) for y^{wh}. When we apply the group lasso, as in the optimization problem (6), it is clear that the minimizer $\widehat{\beta}$ depends on the data $\mathbb{X}, y$ only through $\mathbb{X}^\top \mathbb{X}$ and $\mathbb{X}^\top y$. Furthermore, we can write

$$\mathbb{X}^\top y = \mathbb{X}^\top \mathbb{\Sigma}^{1/2} y^{\text{wh}} = \mathbb{\Sigma}_*^{1/2} \cdot (\mathbb{X}^\top y^{\text{wh}}),$$
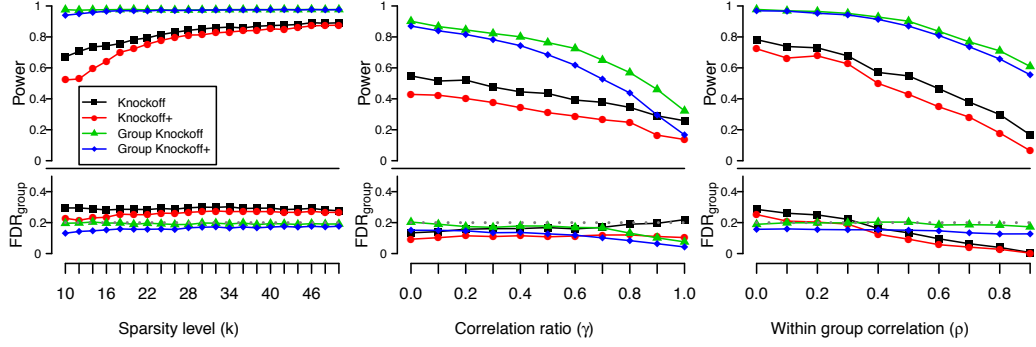
Figure 1. Results for the group-sparse regression simulation, comparing group knockoff and knockoff+ against the original knockoff and knockoff+ methods.

where we can show $\Sigma^{1/2} \cdot \mathbb{X} = \mathbb{X} \cdot \Sigma_*^{1/2}$ exactly as in (14) before. Therefore, $\widehat{\beta}$, depends on the data only through the sufficient statistics $\mathbb{X}^\top \mathbb{X}$ and $\mathbb{X}^\top y^{\mathrm{wh}}$, as desired.

Our statistics for the knockoff filter therefore will satisfy the sufficiency property. The group-antisymmetry property is obvious from the definition of the method. Therefore, applying our main result Theorem 1 for the group-sparse setting to the whitened model (15), we see that the (modified or unmodified) group FDR control result holds for this setting. ☐

5. Simulated data experiments

We test our methods in the group sparse and multitask settings. All experiments were carried out in Matlab (MAT-LAB, 2015) and R (R Core Team, 2015), including the `grpreg` package in R (Breheny & Huang, 2015).

5.1. Group sparse setting

To evaluate the performance of our method in the group sparse setting, we compare it empirically with the (non-group) knockoff using simulated data from a group sparse linear regression, and examine the effects of sparsity level and feature correlations within and between groups.

5.1.1. DATA

To generate the simulation data, we use the sample size $n = 3000$ with number of features $p = 1000$. In our basic setting, the number of groups is $m = 200$ with corresponding number of features per group set as $p_i = 5$ for each group i. To generate features, as a default we use an uncorrelated setting, drawing the entries of X as i.i.d. standard normals, then normalize the columns of X. Our default sparsity level is $k = 20$ (that is, 20 groups with nonzero signal); β_j, for each j inside a signal group, i.i.d. chosen randomly from $\{\pm 3.5\}$.

To study the effects of sparsity level and feature correlation, we then vary these default settings as follows (in each experiment, one setting is varied while the others remain at their default level):

- Sparsity level: we vary the number of groups with nonzero effects, $k \in \{10, 12, 14, \ldots, 50\}$.

- Between-group correlation: we fix within-group correlation $\rho = 0.5$, and set the between-group correlation to be $\gamma\rho$, with $\gamma \in \{0, 0.1, 0.2, \ldots, 0.9\}$. We then draw the rows of $X \in \mathbb{R}^{n \times p}$ independently from a multivariate normal distribution with mean 0 and covariance matrix Σ, with diagonal entries $\Sigma_{jj} = 1$, within-group correlations $\Sigma_{jk} = \rho$ for $j \neq k$ in the same group, and between-group correlations $\Sigma_{jk} = \gamma\rho$ for j, k in different groups. Afterwards, we normalize the columns of X.

- Within-group correlation: as above, but we fix $\gamma = 0$ (so that between-group correlation is always zero) and vary within-group correlation, with $\rho \in \{0, 0.1, \ldots, 0.9\}$.

For each setting, we use target FDR level $q = 0.2$ and repeat each experiment 100 times.

5.1.2. RESULTS

Our results are displayed in Figure 1, which displays power (the proportion of true signals which were discovered) and FDR at the group level, averaged over all trials. We see that all four methods successfully control FDR at the desired level. Across all settings, the group knockoff is more powerful than the knockoff, showing the benefit of leveraging the group structure. The group knockoff+ and knockoff+ are each slightly more conservative than their respective methods without the "+" correction. From the experiments with zero between-group correlation and increasing within-group correlation ρ, we see that knockoff has rapidly de-
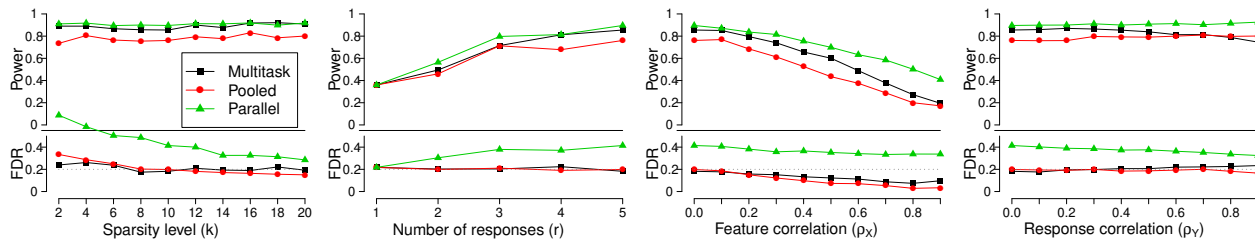
Figure 2. Results for the multitask regression simulation, comparing multitask knockoff with the pooled and parallel knockoff methods.

creasing power as ρ increases, while group knockoff does not show much power loss. This highlights the benefit of the group-wise construction of the knockoff matrix; for the original knockoff, high within-group correlation forces the knockoff features $\widetilde{X}_j$ to be nearly equal to the X_j's, but this is not the case for the group knockoff construction and the greater separation allows high power to be maintained.

5.2. Multitask regression setting

To evaluate the performance of our method in the multitask regression setting, we next perform a simulation to compare the multitask knockoff with the knockoff. (For clarity in the figures, we do not present results for the knockoff+ versions of these methods; the outcome is predictable, with knockoff+ giving slightly better FDR control but lower power.) For the multitask knockoff, we implement the method exactly as described in Section 4. The jth feature is considered a discovery if the corresponding group is selected. For the knockoff, we use reformulation (5) of the multitask model, and apply the knockoff method to the reshaped data set $(\mathbb{X}, y)$; we call this the "pooled" knockoff. We also run the knockoff separately on each of the r responses (that is, we run the knockoff with data (X, Y_j) where Y_j is the jth column of Y, separately for $j = 1, \ldots, r$). We then combine the results: the jth feature is considered a discovery if it is selected in any of the r individual regressions; this version is the "parallel" knockoff.

5.2.1. DATA

To generate the data, our default settings for the multitask model given in (3) are as follows: we set the sample size $n = 150$, the number of features $p = 50$, with $m = 5$ responses. The true matrix of coefficients B has its $k = 10$ rows nonzero, which are chosen as $2\sqrt{m}$ times a random unit vector. The design matrix X is generated by drawing i.i.d. standard normal entries and then normalizing the columns, and the entries of the error matrix E are also i.i.d. standard normal. We set the target FDR level at $q = 0.2$ and repeat all experiments 100 times. These default settings will then be varied in our experiments to examine the roles of the various parameters (only one parameter is var-

ied at a time, with all other settings at their defaults):

- Sparsity level: the number of nonzero rows of B is varied, with $k \in \{2, 4, 6, \ldots, 20\}$.

- Number of responses: the number of responses r is varied, with $r \in \{1, 2, 3, 4, 5\}$.

- Feature correlation: the rows of X are i.i.d. draws from a $N(0, \Sigma_X)$ distribution, with a tapered covariance matrix which has entries $(\Sigma_X)_{jk} = (\rho_X)^{|j-k|}$, with $\rho_X \in \{0, 0.1, 0.2, \ldots, 0.9\}$. (The columns of X are then normalized.)

- Response correlation: the rows of the noise E are i.i.d. draws from a $N(0, \Sigma_Y)$ distribution, with a equivariant correlation structure which has entries $(\Sigma_Y)_{jj} = 1$ for all j, and $(\Sigma_Y)_{jk} = \rho_Y$ for all $j \neq k$, with $\rho_Y \in \{0, 0.1, 0.2, \ldots, 0.9\}$.

5.2.2. RESULTS

Our results are displayed in Figure 2. For each method, we display the resulting FDR and power for selecting features with true effects in the model. The parallel knockoff is not able to control the FDR. This may be due to the fact that this method combines discoveries across multiple responses; if the true positives selected for each response tend to overlap, while the false positives tend to be different (as they are more random), then the false discovery proportion in the combined results may be high even though it should be low for each individual responses' selections. Therefore, while it is more powerful than the other methods, it does not lead to reliable FDR control. Turning to the other methods, both multitask knockoff and pooled knockoff generally control FDR at or near $q = 0.2$ except in the most challenging (lowest power) settings, where as expected from the theory, the FDR exceeds its target level. Across all settings, multitask knockoff is more powerful than pooled knockoff. Overall we see the advantage in the multitask formulation, with which we are able to identify a larger number of discoveries while maintaining FDR control.

6. Real data experiment

We next apply the knockoff for multitask regression to a real data problem. We study a data set that seeks to identify drug resistant mutations in HIV-1 (Rhee et al., 2006). This data set was analyzed by (Barber & Candès, 2015) using the knockoff method. Each observation, sampled from a single individual, identifies mutations along various positions in the protease or reverse transcriptase (two key proteins) of the virus, and measures resistance against a range of different drugs from three classes: protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs), and nonnucleoside reverse transcriptase inhibitors (NNRTIs). In (Barber & Candès, 2015) the data for each drug was analyzed separately; the response y was the resistance level to the drug while the features X_j were markers for the presence or absence of the jth mutation. Here, we apply the multitask knockoff to this problem: for each class of drugs, since the drugs within the class have related biological mechanisms, we expect the sparsity pattern (i.e. which mutations confer resistance to that drug) to be similar across each class. We therefore have a matrix of responses, $Y \in \mathbb{R}^{n \times r}$, where n is the number of individuals and r is the number of drugs for that class. We compare our results to those obtained with the knockoff method where drugs are analyzed one at a time (the "parallel" knockoff from the multitask simulation).

6.1. Data

Data is analyzed separately for each of the three drug types. To combine the data across different drugs, we first remove any drug with a high proportion of missing drug resistance measurements; this results in two PI drugs and one NRTI drug being removed (each with over 35% missing data). The remaining drugs all have $< 10\%$ missing data; many drugs have only $1 - 2\%$ missing data. Next we remove data from any individual that is missing drug resistance information from any of the (remaining) drugs. Finally, we keep only those mutations which appear ≥ 3 times in the sample. The resulting data set sizes are:

Class	# drugs (r)	# observations (n)	# mutations (p)
PI	5	701	198
NRTI	5	614	283
NNRTI	3	721	308

6.2. Methods

For each of the three drug types, we form the $n \times r$ response matrix Y by taking the log-transformed drug resistance measurement for the n individuals and the r drugs, and the $n \times p$ feature matrix X recording which of the p mutations are present in each of the n individuals. We then apply the multitask knockoff as described in Section 4, with
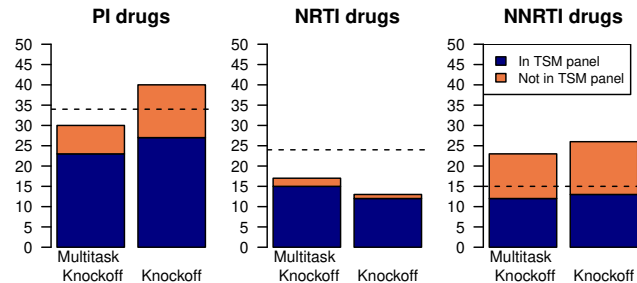

Figure 3. Results on the HIV-1 drug resistance data set. For each drug class, we plot the number of protease positions (for PI) or reverse transcriptase (RT) positions (for NRTI or NNRTI) which were selected by the multitask knockoff or knockoff method. The color indicates whether or not the selected position appears in the treatment selected mutation (TSM) panel, and the horizontal line shows the total number of positions on the TSM panel.

target FDR level $q = 0.2$. We use the equivariant construction for the knockoff matrix for both methods. We perform the analysis 100 times and we consider a mutant is selected if it has been selected by more than 60 times.

6.3. Results

We report our results by comparing the discovered mutations, within each drug class, against the treatment-selected mutation (TSM) panel (Rhee et al., 2005), which gives mutations associated with treatment by a drug from that class. As in (Barber & Candès, 2015) we report the counts by position rather than by mutation, i.e. combining all mutations discovered at a single position, since multiple mutations at the same position are likely to have related effects. To compare with the knockoff method, for each drug class we consider mutation j to be a discovery for that drug class, if it was selected for any of the drugs in that class. The results are displayed in Figure 3. In this experiment, we see that the multitask knockoff seems to show slightly better agreement with the TSM panel. As in the multitask simulation, this may be due to the fact that the knockoff combines discoveries across several drugs; a low false discovery proportion for each drug individually can still lead to a high false discovery proportion once the results are combined.

7. Discussion

We have presented a knockoff filter for the group sparse regression and multitask regression problems, where sharing information within each group or across the set of response variables allows for a more powerful feature selection method. Extending the knockoff framework to other structured estimation problems, such as non-linear regression or to low-dimensional latent structure other than sparsity, would be interesting directions for future work.

References

Barber, Rina Foygel and Candès, Emmanuel J. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 10 2015. doi: 10.1214/15-AOS1337. URL http://dx.doi.org/10.1214/15-AOS1337.

Breheny, Patrick and Huang, Jian. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, 25(2):173–187, 2015.

MATLAB. *Version 8.6.0 (R2015b)*. The MathWorks Inc., Natick, Massachusetts, 2015.

Obozinski, Guillaume, Taskar, Ben, and Jordan, Michael. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Rhee, S-Y., Fessel, W. J., Zolopa, A. R., Hurley, L., Liu, T., Taylor, J., Nguyen, D. P., Slome, S., Klein, D., Horberg, M., 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.

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

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Yuan, Ming and Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.