Information and Inference: A Journal of the IMA (20xx) **00**, 1–29
doi:10.1093/imaiai/iay012

A pseudo knockoff filter for correlated features

Jiajie Chen
Applied and Computational Mathematics, Caltech, Pasadena, CA 91125, USA
jchen@caltech.edu

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

and

Thomas Y. Hou†
Applied and Computational Mathematics, Caltech, Pasadena, CA 91125, USA
†Corresponding author. Email: hou@cms.caltech.edu

[Received on 30 August 2017; revised on 14 May 2018; accepted on 7 June 2018]

In 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. Inspired by the work by Barber & Candès (2015, *Ann. Statist.*, **43**, 2055–2085), we propose a pseudo knockoff filter that inherits some advantages of the original knockoff filter and has more flexibility in constructing its knockoff matrix. Moreover, we perform a number of numerical experiments that seem to suggest that the pseudo knockoff filter with the half Lasso statistic has FDR control and offers more power than the original knockoff filter with the Lasso Path or the half Lasso statistic for the numerical examples that we consider in this paper. Although we cannot establish rigourous FDR control for the pseudo knockoff filter, we provide some partial analysis of the pseudo knockoff filter with the half Lasso statistic and establish a uniform false discovery proportion bound and an expectation inequality.

Keywords: knockoff filter; feature selection; FDR control; correlated features.

1. Introduction

In many applications, we need to study a statistical 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 [1], Barber and Candès introduce the knockoff filter to control the false discovery rate (FDR) in a statistical linear model. More specifically, the knockoff filter constructs knockoff variables that mimic the correlation structure of the true feature variables to obtain exact FDR control in finite sample settings. It has been demonstrated that this method has more power than existing selection rules when the proportion of null variables is high.

1.1 *A brief review of the knockoff filter*

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 have been normalized to be unit vectors in the l^2 norm and ϵ

is a Gaussian noise $N(0, \sigma^2 I_n)$. The knockoff filter begins with the construction of a knockoff matrix $\tilde{X}$ that obeys

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

where $s_i \in [0, 1], i = 1, 2, \ldots, p$. The positive definiteness of the Gram matrix $[X\tilde{X}]^T[X\tilde{X}]$ requires

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

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 original variables. The power (the expected proportion of true discoveries) of the knockoff filter depends critically on the value of s_i. A general guideline in constructing the knockoff matrix is to choose s_j as large as possible to maximize the difference between X_j and its knockoff $\tilde{X}_j$. Next, we choose a statistic, W_j, for each pair $X_j, \tilde{X}_j$ by using the Gram matrix $[X\ \tilde{X}]^T[X\tilde{X}]$ and the marginal correlation $[X\ \tilde{X}]^Ty$. In addition, W_j satisfies a flip-coin property that swapping arbitrary pair $X_j, \tilde{X}_j$ 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 ensure 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.

One of the knockoff statistics considered in [1] is the Lasso path statistic, which is defined as $W_j = \max(Z_j, \tilde{Z}_j) \cdot \text{sign}(Z_j - \tilde{Z}_j)$, where Z_j and $\tilde{Z}_j$ are the solutions of the Lasso path problem given below:

$$\left(\hat{\beta}(\lambda), \tilde{\beta}(\lambda)\right) = \underset{(b,\tilde{b})}{\text{argmin}} \left\{ \frac{1}{2}\|y - Xb - \tilde{X}\tilde{b}\|_2^2 + \lambda\left(\|b\|_1 + \|\tilde{b}\|_1\right) \right\},$$

$$Z_j = \sup\left\{\lambda : \hat{\beta}_j(\lambda) \neq 0\right\},\ \tilde{Z}_j = \sup\left\{\lambda : \tilde{\beta}_j(\lambda) \neq 0\right\}.$$

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 positive W_j. A large positive W_j implies that there is a high probability that the variable j is a non-null. This consideration suggests that we select the variable j with positive W_j larger than a data-dependent threshold T, $\hat{S} \triangleq \{j : W_j \geq T\}$, where T is defined below

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

The false discovery proportion (FDP) of the knockoff filter and its estimate at threshold t are given by

$$FDP(t) \triangleq \frac{\#\{j : W_j \geq t\ \&\ \beta_j = 0\}}{\#\{j : W_j \geq t\} \vee 1}, \quad \widehat{FDP}(t) \triangleq \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}. \tag{4}$$

The FDR is the expectation of FDP. The i.i.d. signs for the null W_j enable one to construct a supermartingale M_t with respect to an appropriate backward filtration $\mathcal{F}_t$ such that

$$\frac{FDP(t)}{\widehat{FDP(t)}} \leq \frac{\#\{j : W_j \geq t\ \&\ \beta_j = 0\}}{1 + \#\{j : W_j \leq -t\ \&\ \beta_j = 0\}} \triangleq M_t, \quad E[M_t] \leq 1. \tag{5}$$

The threshold T defined in (3) gives a stopping time. Using the definition of T and the stopping time theorem, the authors in [1] obtained $E\left[FDP(T)/q\right] \leq E\left[FDP(T)/\widehat{FDP(T)}\right] \leq E[M_T] \leq 1$. The main result in [1] is that the knockoff procedure controls the FDR

$$FDR \triangleq E[FDP(T)] \leq q.$$

In a subsequent paper [2], Barber and Candès developed a framework for high-dimensional linear model with $p \geq n$. The knockoff filter has been further generalized to the model-free framework in [5]. The model-free knockoffs provide valid inference from finite samples in settings in which the conditional distribution of the response is arbitrary and completely unknown. This research has inspired a number of follow-up works, such as [6,7,9,17,18]. There are several other feature selection methods that offer some level of FDR control (e.g. [3,4,8,12–15]). We refer to [1] for a thorough comparison between the knockoff filter and these other approaches.

1.2 *Pseudo knockoff filter*

In this paper, we propose a pseudo knockoff filter that inherits some advantages of the original knockoff filter and have greater flexibility in constructing their pseudo knockoff matrix. The first condition that we impose on the pseudo knockoff matrix is the following orthogonality condition:

$$(X + \tilde{X})^T(X - \tilde{X}) = 0. \tag{6}$$

It can be shown that this condition is equivalent to $X^TX = \tilde{X}^T\tilde{X},\ X^T\tilde{X} = \tilde{X}^TX$.

We consider three classes of pseudo knockoffs that have different additional constraint. For the first class of pseudo knockoff filters, the pseudo knockoff matrix $\tilde{X}$ is chosen to be orthogonal to X, i.e. $X^T\tilde{X} = \tilde{X}^TX = 0$. We call this pseudo knockoff the *orthogonal pseudo knockoff*. It maximizes the difference between the pseudo knockoff matrix $\tilde{X}$ and its original design matrix X. The orthogonality condition makes X_j and its knockoff orthogonal regardless of the correlation structure of X.

The second class of pseudo knockoff filters is called the *block diagonal pseudo knockoff*. We begin by constructing a block diagonal matrix $\mathbf{B}$ that satisfies the property $\mathbf{B} \succeq \Sigma^{-1}$. We can then solve for $\tilde{X}$ from the relationship $\mathbf{B} = 4[(X - \tilde{X})^T(X - \tilde{X})]^{-1}$ where $\mathbf{B} = 2\text{diag}(S_{11}^{-1}, S_{22}^{-1}, \ldots, S_{kk}^{-1})$. The condition (6) and $4[(X - \tilde{X})^T(X - \tilde{X})]^{-1} = 2\text{diag}(S_{11}^{-1}, S_{22}^{-1}, \ldots, S_{kk}^{-1})$ imply that

$$X^TX = \tilde{X}^T\tilde{X}, \quad X^TX - X^T\tilde{X} = \text{diag}(S_{11}, S_{22}, \ldots, S_{kk}).$$

We construct $\mathbf{B}$ by adapting it to the structure of X. One of the guiding principles is to make it as small as possible so that we maximize the difference between X and $\tilde{X}$.

The third class of the pseudo knockoff filter is called the general pseudo knockoff by constructing $\mathbf{B}$ whose principal submatrices are diagonal. The construction is similar to the case when $\mathbf{B}$ is a block diagonal matrix.

1.3 *A half Lasso statistic*

We propose to use a half-penalized method to construct the statistics of our pseudo knockoff filter. More specifically, the pseudo knockoff statistic is based on the solution of the following half-penalized

optimization problem

$$\min_{\hat{\beta},\tilde{\beta}} \frac{1}{2}\|y - X\hat{\beta} - \tilde{X}\tilde{\beta}\|_2^2 + P(\hat{\beta} + \tilde{\beta}), \tag{7}$$

where $P(x)$ is an even non-negative and non-decreasing function in each coordinate of x. An important consequence of the orthogonality condition (6) is that we can reformulate the half-penalized problem into two sub-problems equivalently

$$\min_{\hat{\beta}+\tilde{\beta}} \left\{ \frac{1}{2} \left\| \frac{X+\tilde{X}}{2}\left(\hat{\beta}^{ls} + \tilde{\beta}^{ls} - \hat{\beta} - \tilde{\beta}\right) \right\|_2^2 + P(\hat{\beta} + \tilde{\beta}) \right\} + \min_{\hat{\beta}-\tilde{\beta}} \left\{ \frac{1}{2} \left\| \frac{X-\tilde{X}}{2}\left(\hat{\beta}^{ls} - \tilde{\beta}^{ls} - (\hat{\beta} - \tilde{\beta})\right) \right\|_2^2 \right\}, \tag{8}$$

where $\hat{\beta}^{ls}$ and $\tilde{\beta}^{ls}$ are the least-squares coefficients by regressing y on the augmented feature matrix $[X,\ \tilde{X}]$. If we choose $P = \lambda||\cdot||_{l^1}$, we obtain a *half Lasso* method. We will mainly focus on the half Lasso statistic in this paper. Once we solve the half-penalized problem, we can construct the pseudo knockoff statistic as follows:

$$W_j \triangleq (\hat{\beta}_j + \tilde{\beta}_j) \cdot \text{sign}(\hat{\beta}_j - \tilde{\beta}_j) \text{ or } W_j = \max\left(|\hat{\beta}_j|, |\tilde{\beta}_j|\right) \cdot \text{sign}\left(|\hat{\beta}_j| - |\tilde{\beta}_j|\right).$$

We then apply a procedure similar to the knockoff filter (3) to select features.

We have carried out a number of numerical experiments for different design matrices with various correlation structures to test the performance of the three classes of pseudo knockoff filters and compare their performance with that of the knockoff filter. For the examples that we consider in this paper, our numerical experiments indicate that all three classes of pseudo knockoff filters with the half Lasso statistic have FDR control. Moreover, the orthogonal and the general pseudo knockoff filter seem to offer more power than that of the knockoff filter with the Lasso Path or the half Lasso statistic, especially when the features are highly correlated.

1.4 *Uniform FDP bounds*

There has been some recent progress in obtaining uniform FDP bounds in [**10**,**11**]. Using (3), (4) and (5), we can divide the control of FDR into three steps. First of all, we construct an estimate of FDP. We then choose a data-dependent threshold T that achieves some adaptivity. The final step is to obtain an estimate for $E[FDP(T)/\widehat{FDP(T)}]$ for this adaptive threshold, T. In [**10**], the authors showed that the above strategy of controlling FDR provides a general strategy for a variety of existing procedures that offer FDR control under some assumptions. In [**11**], the authors established a uniform bound across all possible threshold for the knockoff filter

$$E\left[\sup_{t>0} \frac{FDP(t)}{\widehat{FDP(t)}}\right] \leq E\left[\sup_{t>0} \frac{\#\{j : W_j \geq t \ \&\ \beta_j = 0\}}{1 + \#\{j : W_j \leq -t \ \&\ \beta_j = 0\}}\right] \leq 1.93. \tag{9}$$

In [**10**], the above uniform FDP bounds are established for several FDR procedures under some independence assumption similar to the *i.i.d. signs for the nulls* in the knockoff filter.

Inspired by the work of [**10**,**11**], we establish a uniform FDP bound under an assumption weaker than the independence assumption on the conditional distribution of the statistic W. Specifically, we prove the following theorem.

THEOREM 1.1 Let $\mathcal{F}$ be a σ field that satisfies the following: (a) $|W_i|$ is $\mathcal{F}$ measurable for all null i; (b) conditional on $\mathcal{F}$, W_{S_0} can be divided into m groups $C_1, C_2, \ldots, C_m$ ($C_i \subset S_0$) such that the elements of $\text{sign}(W_{C_i})$ are mutually independent with $P(\text{sign}(W_j) = 1) = P(\text{sign}(W_j) = -1) = 1/2$ for $j \in C_i$. For any $t > 0$, we have

$$E\left[\frac{\#\{j \in S_0 : W_j \geq t\}}{\#\{j \in S_0 : W_j \leq -t\} + m}\middle| \mathcal{F}\right] \leq 1. \tag{10}$$

Moreover, if W_{S_0} further satisfies $W_{S_0} \overset{d}{=} -W_{S_0}$ conditional on $\mathcal{F}$, we have

$$E\left[\sup_{t>0}\frac{\#\{j \in S_0 : W_j \geq t\}}{\#\{j \in S_0 : W_j \leq -t\} + m}\middle| \mathcal{F}\right] \leq 3.9. \tag{11}$$

Although Theorem 1.1 does not provide FDR control for the pseudo knockoff filter, it provides some partial understanding of the pseudo knockoff filter. For the block diagonal and the general pseudo knockoff filters, we verify that the pseudo knockoff statistic W_j satisfies the assumption in Theorem 1.1 for some appropriate σ field $\mathcal{F}$. For the orthogonal pseudo knockoff filter, the pseudo knockoff statistic W_j does not satisfy the assumption in Theorem 1.1. To gain some understanding of the orthogonal pseudo knockoff filter, we obtain a relatively tight upper bound for the distribution function of $\frac{\#\{j\in S_0:\ W_j\geq t\}}{\#\{j\in S_0:\ W_j\leq -t\}}$ for fixed t when Σ^{-1} is diagonally dominated or when Σ^{-1} has some special structure.

The rest of the paper is organized as follows. In Section 2, we introduce the three classes of pseudo knockoff filters and discuss some essential properties of the pseudo knockoff filters. In Section 3, we present a number of numerical experiments to demonstrate the effectiveness of the proposed methods. In Section 4.1, we prove (10) and outline the proof of (11) in Theorem 1.1. In Section 4.2, we provide some partial analysis of the orthogonal pseudo knockoff filter.

2. A pseudo knockoff filter

In this section, we describe how to construct the three classes of pseudo knockoff filters and the half Lasso statistic. We will also discuss some of the essential properties of these pseudo knockoff filters and the half Lasso statistic.

2.1 *The Basic constraint and a symmetry property*

Given a design matrix $X \in R^{n\times p}$ with $n > 2p$, the basic constraint of the pseudo knockoff matrix is given by

$$\tilde{X}^T\tilde{X} = X^TX, \quad X^T\tilde{X} = \tilde{X}^TX. \tag{12}$$

We can prove that (12) and (6) are equivalent. It is obvious that (12) implies (6). If (6) holds, we have $X^TX - \tilde{X}^T\tilde{X} = X^T\tilde{X} - \tilde{X}^TX$. Note that the right-hand side is a symmetric matrix, while the left-hand side is a skew-symmetric matrix. It follows that $X^TX - \tilde{X}^T\tilde{X}$ is symmetric and skew-symmetric. Thus we must have $X^TX - \tilde{X}^T\tilde{X} = 0$, which further implies $X^T\tilde{X} - \tilde{X}^TX = 0$. These two equations establish (12). The orthogonality condition (6) is the foundation of the pseudo knockoff filter, and leads to the conditional independence between the amplitude of the null statistic $|W_{S_0}|$ and its sign $\text{sign}(W_{S_0})$.

Least-squares coefficients. Consider the least-squares coefficients $(\hat{\beta}^{ls}, \tilde{\beta}^{ls})$ of regressing y on the augmented design matrix $[X\ \tilde{X}]$. It is easy to obtain that $(\hat{\beta}^{ls} + \tilde{\beta}^{ls}, \hat{\beta}^{ls} - \tilde{\beta}^{ls})$ are the least-squares coefficients of regressing $y = X\beta + \varepsilon$ on $\left[\frac{X+\tilde{X}}{2}\ \frac{X-\tilde{X}}{2}\right]$. Using the orthogonality condition (6), we have a simple expression of the least-squares coefficients,

$$\begin{pmatrix} \hat{\beta}^{ls} + \tilde{\beta}^{ls} - \beta \\ \hat{\beta}^{ls} - \tilde{\beta}^{ls} - \beta \end{pmatrix} = \begin{pmatrix} \left[\left(\frac{X+\tilde{X}}{2}\right)^T \frac{X+\tilde{X}}{2}\right]^{-1} \left(\frac{X+\tilde{X}}{2}\right)^T \epsilon \\ \left[\left(\frac{X-\tilde{X}}{2}\right)^T \frac{X-\tilde{X}}{2}\right]^{-1} \left(\frac{X-\tilde{X}}{2}\right)^T \epsilon \end{pmatrix} \triangleq \begin{pmatrix} \epsilon^{(1)} \\ \epsilon^{(2)} \end{pmatrix}. \tag{13}$$

The above relationship will be used repeatedly throughout the paper. Denote

$$\eta \triangleq \hat{\beta}^{ls} + \tilde{\beta}^{ls} = \beta + \epsilon^{(1)}, \xi \triangleq \hat{\beta}^{ls} - \tilde{\beta}^{ls} = \beta + \epsilon^{(2)}. \tag{14}$$

From the orthogonality property (6), we know that $\left(\frac{X+\tilde{X}}{2}\right)^T \epsilon$ and $\left(\frac{X-\tilde{X}}{2}\right)^T \epsilon$ have independent multivariate normal distributions. Using (13), we know that $\epsilon^{(1)}$ and $\epsilon^{(2)}$, $\eta = \hat{\beta}^{ls} + \tilde{\beta}^{ls}$ and $\xi = \hat{\beta}^{ls} - \tilde{\beta}^{ls}$ are also independent.

The pseudo-knockoff statistics and their properties. According to (8), we can solve $\hat{\beta}+\tilde{\beta}$ and $\hat{\beta}-\tilde{\beta}$ in the half-penalized problem (7) separately. Thus the solution can be expressed as

$$\hat{\beta} + \tilde{\beta} = f(\hat{\beta}^{ls} + \tilde{\beta}^{ls}) = f(\eta), \quad \hat{\beta} - \tilde{\beta} = \hat{\beta}^{ls} - \tilde{\beta}^{ls} = \xi, \tag{15}$$

for some function $f : R^p \to R^p$. We construct the pseudo knockoff statistic as follows:

$$W_j \triangleq (\hat{\beta}_j + \tilde{\beta}_j) \cdot \operatorname{sign}(\hat{\beta}_j - \tilde{\beta}_j) \quad \text{or} \quad W_j \triangleq \max\left(|\hat{\beta}_j|, |\tilde{\beta}_j|\right) \cdot \operatorname{sign}\left(|\hat{\beta}_j| - |\tilde{\beta}_j|\right). \tag{16}$$

The pseudo knockoff statistic satisfies the following two properties.

Amplitude property. The amplitude of W is determined by $\hat{\beta} + \tilde{\beta} = f(\eta)$ and $|\hat{\beta} - \tilde{\beta}| = |\xi|$. In fact, using the definition of W and (15), we have

$$|W| = |\hat{\beta} + \tilde{\beta}| = |f(\eta)| \quad \text{or} \quad |W| = |\hat{\beta}| \vee |\tilde{\beta}| = \frac{1}{2}\left(|\hat{\beta} + \tilde{\beta} + |\hat{\beta} - \tilde{\beta}\| \vee |\hat{\beta} + \tilde{\beta} - |\hat{\beta} - \tilde{\beta}\|\right).$$

Sign property. The sign of W is determined by $\operatorname{sign}(\hat{\beta} + \tilde{\beta})$ and $\operatorname{sign}(\hat{\beta} - \tilde{\beta})$. Since $\operatorname{sign}(|\hat{\beta}| - |\tilde{\beta}|) = \operatorname{sign}(|\hat{\beta}|^2 - |\tilde{\beta}|^2)$, for both definitions of W, we have

$$\operatorname{sign}(W) = \operatorname{sign}(\hat{\beta} + \tilde{\beta}) \cdot \operatorname{sign}(\hat{\beta} - \tilde{\beta}) = \operatorname{sign}(f(\eta)) \cdot \operatorname{sign}(\xi).$$

Now we show that the pseudo-knockoff statistic satisfies a symmetry property.

PROPOSITION 2.1 Conditional on η, we have $W_{S_0} \overset{d}{=} -W_{S_0}$, where $S_0 \triangleq \{j : \beta_j = 0\}$ and the pseudo knockoff statistic W_j is defined in (16). Consequently, for any threshold $t > 0$, we have

$$\#\{j : \beta_j = 0 \quad \text{and} \quad W_j \geq t\} \overset{d}{=} \#\{j : \beta_j = 0 \quad \text{and} \quad W_j \leq -t\}. \tag{17}$$

Proof. According to (14) and (15), the solution of the half-penalized problem can be expressed as

$$\hat{\beta} + \tilde{\beta} = f(\eta) = f\big(\beta + \epsilon^{(1)}\big), \quad \hat{\beta} - \tilde{\beta} = \xi = \beta + \epsilon^{(2)}.$$

Next, we replace $(\epsilon^{(1)}, \epsilon^{(2)})$ by $(\epsilon^{(1)}, -\epsilon^{(2)})$ to generate a new pair of solutions $(\hat{\beta}^{new}, \tilde{\beta}^{new})$. From (14), changing $\epsilon^{(2)}$ to $-\epsilon^{(2)}$ does not change η. Thus, we obtain

$$\hat{\beta}^{new} + \tilde{\beta}^{new} = f(\eta) = \hat{\beta} + \tilde{\beta}, \quad \hat{\beta}^{new} - \tilde{\beta}^{new} = \beta - \epsilon^{(2)}.$$

The amplitude and sign properties of W imply $|W^{new}_{S_0}| = |W_{S_0}|$ and

$$\text{sign}\big(W^{new}_{S_0}\big) = \text{sign}((f(\eta))_{S_0} \cdot (-\varepsilon^{(2)})_{S_0}) = -\text{sign}((f(\eta))_{S_0} \cdot (\varepsilon^{(2)})_{S_0}) = -\text{sign}(W_{S_0}).$$

Hence $W^{new}_{S_0} = -W_{S_0}$.

Recall that W_{S_0} is generated by $\epsilon^{(1)}, \epsilon^{(2)}$ and that $\epsilon^{(1)}$, $\epsilon^{(2)}$ have independent multivariate normal distributions with zero mean. Conditional on η (or equivalently $\epsilon^{(1)}$), we have

$$\big(\epsilon^{(1)}, \epsilon^{(2)}\big) \overset{d}{=} \big(\epsilon^{(1)}, -\epsilon^{(2)}\big) \Longrightarrow W_{S_0} \overset{d}{=} W^{new}_{S_0} = -W_{S_0}.$$

(17) is a direct result of $W_{S_0} \overset{d}{=} -W_{S_0}$. □

A half Lasso statistic. We assume that $n > 2p$ and choose $P(x) = \lambda||x||_1$ in (7) to obtain a *half Lasso* optimization problem:

$$\min_{\hat{\beta}, \tilde{\beta}} \frac{1}{2}\|y - X\hat{\beta} - \tilde{X}\tilde{\beta}\|_2^2 + \lambda\|\hat{\beta} + \tilde{\beta}\|_1. \tag{18}$$

We then define the pseudo knockoff statistic according to (16). It satisfies the symmetry property in Proposition 2.1. We have conducted many simulations with different design matrices and signal sparsity, and found that the half Lasso statistic offers robust performance when the tuning parameter λ is of the same order as the noise level. Thus we can choose the tuning parameter λ by $\lambda = \mu||U^T y||_2/\sqrt{n - 2p}$, where $U \in R^{n\times(n-2p)}$ is an orthonormal matrix such that $[X\tilde{X}]^T U = 0$. In fact, $U^T y$ is exactly the residue of regressing y onto $[X\ \tilde{X}]$. From our numerical study, we also observe that the power of the half Lasso statistic is not very sensitive to μ for a small range of μ centred at $\mu = 1$, and the numerical results seem to suggest that $\mu = 0.75$ is among the optimal choice. Thus we choose $\lambda = 0.75||U^T y||_2/\sqrt{n - 2p}$ as the default tuning parameter. One can verify the symmetry property of the pseudo knockoff statistic using a similar argument.

2.2 *Construction of the pseudo-knockoff matrix*

In the previous subsection, we described the basic constraint (12) for the pseudo knockoff matrix. In this subsection, we impose an additional constraint on $\tilde{X}$ so that we can obtain another important property for the pseudo knockoff statistic. In particular, we are interested in three classes of pseudo knockoff matrices, namely the *orthogonal*, the *block diagonal* and the general pseudo knockoff matrices.

From (13) and (14), we know that the covariance matrix of $\epsilon^{(2)}$, or equivalently ξ, is given by

$$\mathbf{B} \triangleq 4\big[(X-\tilde{X})^T(X-\tilde{X})\big]^{-1}. \tag{19}$$

We can design $\mathbf{B}$ in such a way that we obtain some special correlation structure on ξ. To increase the power of the pseudo knockoff filter, we would like to construct $\tilde{X}$ such that the difference between $\tilde{X}_j$ and X_j is large. Since $||X_j - \tilde{X}_j||_2^2 = ((\mathbf{B}/4)^{-1})_{jj}$, we aim to design $\mathbf{B}$ as small as possible. Due to the existing constraint (12) or (6), the covariance matrix $\mathbf{B}$ cannot be chosen arbitrarily. We give a necessary and sufficient condition on $\mathbf{B}$ to find $\tilde{X}$ that satisfies (6) and (19).

Necessary condition on B. Assume that there exists some $\tilde{X}$ that satisfies (6) and (19), and $X - \tilde{X}$ has full rank. Performing singular value decomposition (SVD) on $(X-\tilde{X})/2$, we have $(X-\tilde{X})/2 = \mathbf{PM}^{-1}$ for some orthonormal matrix $\mathbf{P} \in R^{n\times p}$ and some invertible matrix $\mathbf{M} \in R^{p\times p}$. As a result, we get $\mathbf{B} = [(\mathbf{PM}^{-1})^T(\mathbf{PM}^{-1})]^{-1} = \mathbf{MM}^T$ and $\tilde{X} = X - 2\mathbf{PM}^{-1}$. Substituting the last equation into the orthogonal condition $(X+\tilde{X})^T(X-\tilde{X}) = 0$ (see (6)), we obtain

$$\begin{aligned} 4(X-\mathbf{PM}^{-1})^T\mathbf{PM}^{-1} = 0 \quad &\Longleftrightarrow \quad \mathbf{M}^{-T}\mathbf{M}^{-1} = \mathbf{M}^{-T}\mathbf{P}^TX \\ \Longleftrightarrow \quad \mathbf{M}^{-1} = P^TX \quad &\Longrightarrow \quad \mathbf{B} = (X^TPP^TX)^{-1}. \end{aligned}$$

Since $P \in R^{n\times p}$ is orthonormal, we have

$$X^TPP^TX \preceq X^T\mathbf{I}X = X^TX = \Sigma \quad \Longrightarrow \quad \mathbf{B} = (X^TPP^TX)^{-1} \succeq \Sigma^{-1}. \tag{20}$$

Sufficiency. If B satisfies (20), we have $B - \Sigma^{-1} \succeq 0$ and we can construct $\tilde{X}$ as follows:

$$\tilde{X} = X(\mathbf{I} - 2\Sigma^{-1}\mathbf{B}^{-1}) + 2\mathbf{UCB}^{-1}, \tag{21}$$

where $\mathbf{C} \in R^{p\times p}$ satisfies $\mathbf{C}^T\mathbf{C} = \mathbf{B} - \Sigma^{-1}$ and $\mathbf{U} \in R^{n\times p}$ is an orthonormal matrix with $\mathbf{U}^TX = 0$. We will show that $\tilde{X}$ constructed from (21) satisfies (6) and (19) in the end of Appendix A.

2.2.1 *An orthogonal construction* The simplest construction is to choose $\mathbf{B} = 2\Sigma^{-1}$, which is equivalent to the following:

$$X^TX = \tilde{X}^T\tilde{X}, \quad X^T\tilde{X} = \tilde{X}^TX = 0. \tag{22}$$

We call this special pseudo knockoff the *orthogonal pseudo knockoff* since $\tilde{X}$ and X are orthogonal. To construct an orthogonal pseudo knockoff matrix $\tilde{X}$, we first find the SVD of $X \in R^{n\times p}$: $X = UDV^T$, $U \in Orth^{n\times p}$, $D = \text{diag}\{\sigma_1, \ldots, \sigma_p\}$ and $V \in Orth^{p\times p}$. We then choose any orthonormal matrix $W \in R^{n\times p}$, whose column space is orthogonal to that of X (i.e. $X^TW = 0$), and construct the pseudo knockoff matrix $\tilde{X}$ as $\tilde{X} = WDV^T$. It is easy to verify that $\tilde{X}$ satisfies (22).

2.2.2 *A block diagonal construction* **A Block Diagonal Construction.** Consider a block diagonal matrix $\mathbf{B} = 2\text{diag}(S_{11}^{-1}, S_{22}^{-1}, \ldots, S_{kk}^{-1})$, where S_{ii}s are invertible matrices. The constraint on $\mathbf{B}$ is equivalent to

$$2\mathbf{B}^{-1} = \text{diag}(S_{11}, S_{22}, \ldots, S_{kk}) \preceq 2\Sigma. \tag{23}$$

Hence $(X - \tilde{X})^T(X - \tilde{X}) = 4\mathbf{B}^{-1} = 2\text{diag}(S_{11}, S_{22}, \ldots, S_{kk})$. Using this relationship together with the basic constraint (12), i.e. $X^TX = \tilde{X}^T\tilde{X}, X^T\tilde{X} = \tilde{X}^TX$, we obtain

$$X^TX = \tilde{X}^T\tilde{X}, \quad X^TX - X^T\tilde{X} = \text{diag}(S_{11}, S_{22}, \ldots, S_{kk}). \tag{24}$$

Assume that X can be clustered into $(X_{G_1}, X_{G_2}, \ldots, X_{G_k})$. Inspired by the group knockoff construction in [7], we first choose $S_{ii} \triangleq \gamma\Sigma_{G_i,G_i} = \gamma X_{G_i}^T X_{G_i}$, $i = 1, 2, \ldots, k$. The constraint (23) implies $\gamma \cdot \text{diag}(\Sigma_{G_1,G_1}, \Sigma_{G_2,G_2}, \ldots, \Sigma_{G_k,G_k}) \preceq 2\Sigma$. In order to maximize the difference between X and $\tilde{X}$, γ should be chosen as large as possible: $\gamma \leq \min\{1, 2\cdot\lambda_{\min}(D\Sigma D)\}$, where $D = \text{diag}\big(\Sigma_{G_1,G_1}^{-1/2}, \Sigma_{G_2,G_2}^{-1/2}, \ldots, \Sigma_{G_k,G_k}^{-1/2}\big)$. To ensure that the matrix $(X + \tilde{X})^T(X + \tilde{X})$ is non-singular, we choose $\gamma = \frac{1}{1.2}\min\{1, 2 \cdot \lambda_{\min}(D\Sigma D)\}$ in our numerical experiments. Once we construct $\mathbf{B}$, we can generate the pseudo knockoff matrix via the procedure described earlier. This construction is useful if the features X_j are clustered.

2.2.3 *A general construction* In general, we first divide the features X_j into m groups $C_1, C_2, .., C_m$ such that the correlation within each group is relatively weak. We remark that this criterion of partition is different from the grouping strategy in the block diagonal construction. The motivation of this partition is that $(\Sigma^{-1})_{C_jC_j}$ may be close to a diagonal matrix, which can be useful for the later construction of $\mathbf{B}$.

We give two examples to illustrate why this partition may give rise to $(\Sigma^{-1})_{C_jC_j}$ that is close to a diagonal matrix. For example, if each X_j is only strongly correlated with its neighbours X_{j+i} for $|i|$ small, we can choose $C_k = \{im + k : i = 0, 1, ..\}$ for $k = 1, 2, .., m$. If $\Sigma_{ij} = X_i^TX_j = \rho^{|i-j|}$ for some $\rho > 0$, Σ^{-1} is tridiagonal, and thus $(\Sigma^{-1})_{C_j,C_j}$ is a diagonal matrix. Another example is that if X can be clustered into several groups such that the within-group correlation is stronger than the between-group correlation and the maximal group size is m, then we can pick C_i as the ith element in each group for $1 \leq i \leq m$. If the between group correlation is 0, $(\Sigma^{-1})_{C_j,C_j}$ is also a diagonal matrix.

We construct a diagonal matrix $\mathbf{S}_j$ that majorizes $(\Sigma^{-1})_{C_j,C_j}$ using a semidefinite program (SDP)

$$\text{minimize } \textit{trace}(\mathbf{S}_j) \quad \text{subject to} \quad \gamma(\Sigma^{-1})_{C_j,C_j} \preceq \mathbf{S}_j, \quad 2 \leq (\mathbf{S}_j)_{ii}.$$

The above SDP is similar to the SDP in the knockoff construction [1] and can be solved very efficiently. $\gamma > 1$ is some parameter to be determined. If $(\Sigma^{-1})_{C_j,C_j}$ is close to a diagonal matrix, we can construct a S_j such that their entries are not too large. Next, we construct $\mathbf{B}$ as follows:

$$\mathbf{B}_{C_i,C_i} = \mathbf{S}_i, \quad \mathbf{B}_{C_i,C_j} = \gamma(\Sigma^{-1})_{C_i,C_j} \quad 1 \leq i \neq j \leq m. \tag{25}$$

The difference between $\mathbf{B}$ and Σ^{-1} is on the diagonal. The above $\mathbf{B}$ satisfies constraint (20)

$$\begin{aligned}\mathbf{B} - \gamma\Sigma^{-1} &= \text{diag}(B_{C_1,C_1} - \gamma(\Sigma^{-1})_{C_1,C_1}, \ldots, B_{C_m,C_m} - \gamma(\Sigma^{-1})_{C_m,C_m}) \\ &= \text{diag}(\mathbf{S}_1 - \gamma(\Sigma^{-1})_{C_1,C_1}, \ldots, \mathbf{S}_m - \gamma(\Sigma^{-1})_{C_m,C_m}) \succeq 0 \quad \Rightarrow \quad \mathbf{B} \succeq \gamma\Sigma^{-1} \succeq \Sigma^{-1}.\end{aligned}$$

We choose $\gamma = 1.2$ to ensure that $(X + \widetilde{X})^T(X + \widetilde{X})$ is non-singular.

Among three constructions of the pseudo knockoff matrix, we choose the general construction as the default construction. After we construct the pseudo knockoff matrix $\tilde{X}$, we use $y, [X\ \tilde{X}]$ to calculate the half Lasso statistic and finally apply the knockoff+ filter (3) with the target FDR level q to selection features.

Relation to the knockoff filter. If $m = 1$, $\tilde{X}$ constructed via the block diagonal or the general construction is exactly a knockoff matrix of X [1]. The constraint (1) in the original knockoff filter implies that $(X - \tilde{X})^T(X - \tilde{X})$ is a diagonal matrix, which in turn forces $[(X - \tilde{X})^T(X - \tilde{X})]^{-1} = \mathbf{B}/4$ to be a diagonal matrix. In the construction of the pseudo knockoff matrix (23) or (25), we only require that $\mathbf{B}$ be a block diagonal matrix or some submatrices of $\mathbf{B}$ be diagonal. In this case, we can consider the pseudo knockoff filter as a generalization of the knockoff filter.

By comparing our block diagonal pseudo knockoff construction with the group knockoff filter in [7], we can see that the pseudo knockoff matrix, $\tilde{X}$, in (24) is actually a group knockoff matrix of X. The group knockoff filter is originally designed for group selection with group FDR control while our block diagonal pseudo knockoff filter is designed for feature selection.

3. Numerical results for the pseudo knockoff filter

In this section, we perform a number of numerical experiments to test the robustness of the pseudo knockoff filter and study the performance of various methods.

Notations. $\beta_i \overset{i.i.d}{\sim} \{\pm A\}$ means that β_i takes value A or $-A$ independently with equal probability $1/2$. We denote the orthogonal pseudo knockoff, the pseudo knockoff with the block diagonal construction and the pseudo knockoff with general construction as *orthogonal (OPK), block diagonal (BDPK)* and *general (GPK)* pseudo knockoff.

Data. Given some covariance matrix Σ, we first draw the rows of the design matrix $X \in R^{n\times p}$ from a multivariate normal distribution $N(0, \Sigma)$ and then normalize the columns of X. The pseudo knockoff matrix is generated according to Section 2.2. To generate the signal strength $\beta \in R^p$, we choose k coefficients $\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_k}$ randomly and set $\beta_{i_j} \overset{i.i.d}{\sim} \{\pm A\}$. Finally, the response variable $y \in R^n$ is generated from $y = X\beta + \epsilon, \varepsilon \sim N(0, I_n)$. Unless we specify otherwise, we will use the following default setup, i.e. the sample size is $p = 500$, $n = 1500$, the sparsity is $k = 30$, the signal amplitude is $A = 3.5$ and the covariance matrix is $\Sigma = I_p$.

Methods. The methods that we focus on include the OPK, BDPK and GPK filters with the half Lasso statistic ($\lambda = 0.75$). We use the knockoff+ filter (3) with nominal FDR level $q = 20\%$. We assume that every five features form a group and then construct the BDPK matrix. We choose $C_k = \{im + k : i = 0, 1, ..\}$ with $m = 2, 3, 5$ to construct the GPK matrix. After obtaining the fitted value $\hat{\beta}, \tilde{\beta}$ in the half Lasso problem, we have two choices to construct the statistic, W, in (16). Denote $W_j^{(1)} = (\hat{\beta}_j + \tilde{\beta}_j) \cdot \text{sign}(\hat{\beta}_j - \tilde{\beta}_j)$ and $W_j^{(2)} = |\hat{\beta}_j| \vee |\tilde{\beta}_j| \cdot \text{sign}(|\hat{\beta}_j| - |\tilde{\beta}_j|)$. For the OPK, we use $W^{(2)}$, which seems to offer more power with OPK; for other pseudo knockoff filters, we consider both constructions of W in (16). There are nine methods in total.

3.1 *Numerical evidence of FDR control for the pseudo knockoff filter*

In this subsection, we perform extensive numerical experiments to test whether the pseudo knockoff filter has FDR control. For this purpose, we apply it to select features in the linear model $y = X\beta + \epsilon$ with different design matrices under various extreme conditions.

The default simulated data is discussed at the beginning of Section 3 and we vary one of the default settings in each experiment as follows (one setting is varied while keeping the others unchanged).

(a) *Sparsity*: k varies from $10, 20, 30, \ldots, 90, 100$.

(b) *Signal amplitude*: A varies from $2.8, 2.9, \ldots, 4.2$.

(c) *Correlation Structure*: We use the covariance matrix $\Sigma \in R^{500\times 500}$, $\Sigma_{ij} = \rho^{|i-j|}$ and vary the correlation level $\rho = 0, 0.1, \ldots, 0.9$.

(d) *The sample size*: We vary the sample size $n = 150l$, $p = 50l$ and sparsity $k = 10l$ with $l \in \{2, 3, \ldots, 12\}$.

Group Structure: We assume that the features X_j can be clustered into 100 groups with five features in each group. To generate a different group structure, we choose the covariance matrix $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ for $i \neq j$ in the same group and $\Sigma_{ij} = \gamma \cdot \rho$ for $i \neq j$ in different groups, and generate the design matrix X as in the previous discussion.

(e) *The within-group correlation*: $\gamma = 0$ is fixed and ρ varies from $0, 0.1, 0.2, \ldots, 0.9$.

(f) *The between-group correlation*: $\rho = 0.5$ is fixed and γ varies from $0, 0.1, 0.2, \ldots, 0.9$.

We pay particular attention to the *FDR* (the mean FDP), the *power* (the expected proportion of true discoveries) and the *expectation*, which is defined as the expectation of $\frac{\#\{j:\ W_j \geq T\ \&\ \beta_j = 0\}}{\#\{j:\ W_j \leq -T\ \&\ \beta_j = 0\}+1}$. Each experiment is repeated 200 times to calculate these quantities. The design matrix X and the pseudo knockoff matrices $\tilde{X}$ are fixed over these trials. We plot the results of OPK and BDPK ($m = 5$), GPK ($m = 2$) with $W^{(1)} = (\hat{\beta} + \tilde{\beta}) \cdot \mathrm{sign}(\hat{\beta} - \tilde{\beta})$ in Figs 1 and 2.

The dotted line in Figs 1 and 2 represents the prescribed FDR q or constant 1 as a reference. In all figures, we observe that the FDR is controlled by $q = 20\%$. From the results of the expectation, we observe that all of them are close to or less than 1. Other six methods described before Section 3.1 control FDR in the above examples. In Section 4, we will provide partial analysis to gain some understanding of the pseudo knockoff filter.

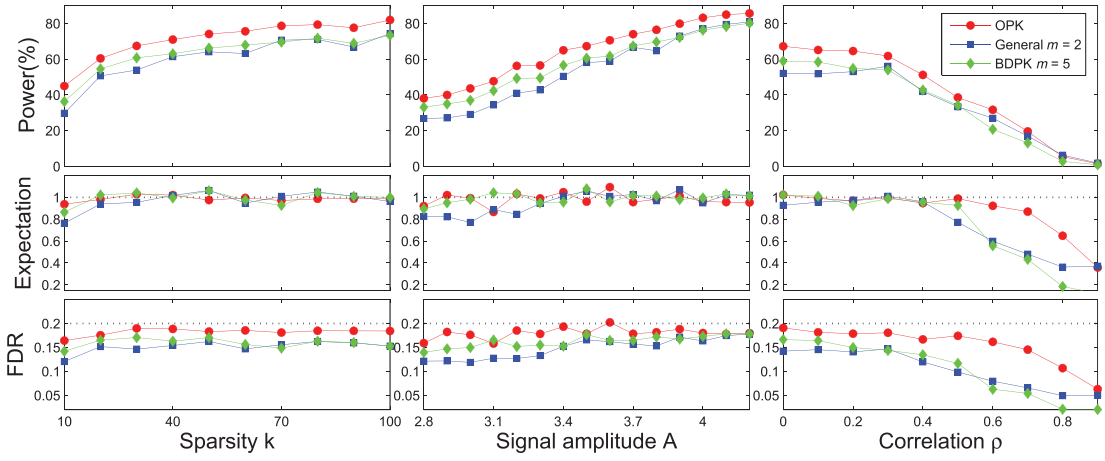

FIG. 1. Testing the orthogonal, the block diagonal and the general pseudo knockoff+ at a nominal FDR $q = 20\%$ by varying the sparsity, the signal amplitude or the feature correlation.

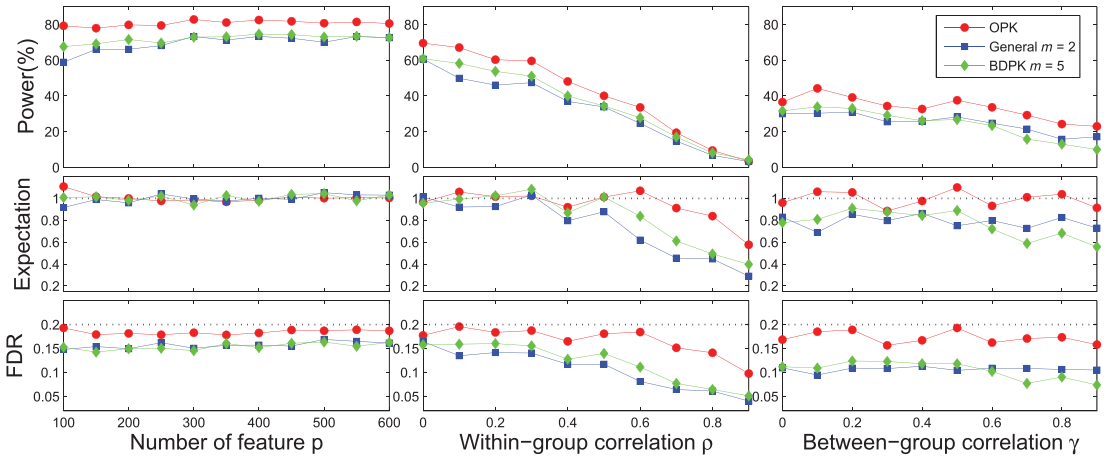

FIG. 2. Testing several pseudo knockoff+ filters at a nominal FDR $q = 20\%$ by varying the number of features p, the within-group correlation or the between-group correlation.

3.2 *The pseudo knockoff filter in some correlated scenarios*

Due to the constraints on the knockoff matrix in the original knockoff filter, strongly correlated features force the s_i to be small [**6**], which may lead to loss of some power. A main advantage of the pseudo knockoff filter is that it relaxes the constraint of $\tilde{X}$ in (12). In some correlated scenarios with some special structure, we can construct the pseudo knockoff matrix that is adapted to such structure and improve the power. To illustrate the effectiveness of the pseudo knockoff filter, we compare the knockoff filter using various statistics with various pseudo knockoff constructions using the half Lasso statistic.

Statistics. We use the half Lasso statistic with $\lambda = 0.75||U^T y||_2/\sqrt{n-2p}$ $(n > 2p)$ for the pseudo knockoff filter. We also consider the corresponding statistics in the knockoff filter for comparison. Specifically, we consider the knockoff filter with the half Lasso or Lasso using the same tuning parameter $(\lambda = 0.75||U^T y||_2/\sqrt{n-2p},\ n > 2p)$ and the sign max statistic $W^{(2)}$. In addition, we have tested the knockoff filter with other statistics, including the Lasso path and the Orthogonal Matching Pursuit (OMP) statistics. The knockoff matrix is generated by the SDP construction introduced in [**1**]. In the following examples, we use a slightly larger signal amplitude $A = 5$. For these methods, we use the knockoff+ filter (3) with nominal FDR level $q = 20\%$. Throughout all the examples in this section, we repeat the experiment 200 times to obtain the FDR and the averaged power.

Group structure. We consider a design matrix $X \in R^{1500\times 500}$ with a group structure and two sparsity cases: $k = 30$ and $k = 100$. In particular, we consider experiment (e) in Section 3.1. The within-group correlation factor ρ varies from $0.5, 0.55, , \ldots, 0.95$ and the between-group correlation factor is $\gamma = 0$. In all other settings, we use the default values. By taking advantage of the *a priori* knowledge of the correlation structure of X, we construct the BDPK and GPK with $m = 5$. We also implement the OPK with $W^{(2)}$ statistic for comparison.

In Fig. 3, the pseudo knockoff filters control FDR and outperform the knockoff filter with the OMP or the Lasso path statistic. The BDPK with $W^{(1)}$ statistic (not plotted) also outperforms the knockoff filter with two statistics, but offers less power than that of the OPK or the GPK.

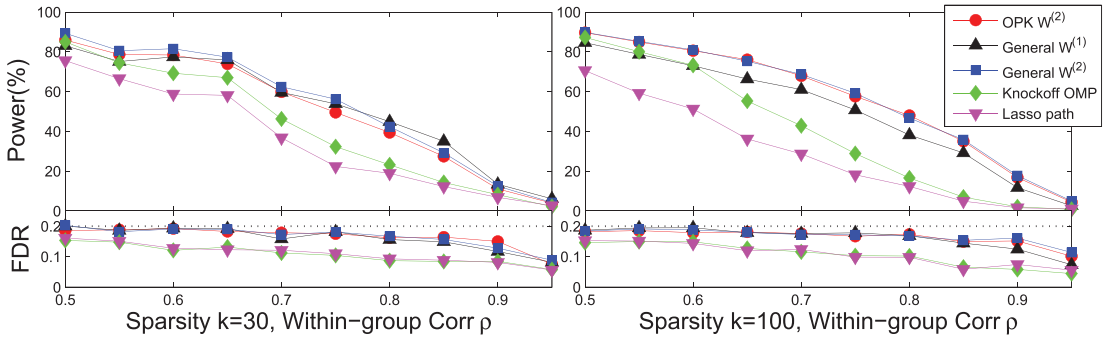

FIG. 3. Comparing the orthogonal, the general pseudo knockoff+ filter and the knockoff+ filter with several statistics at nominal FDR $q = 20\%$ by varying the within-group correlation. Here, the general $W^{(i)}$ means the method using the general pseudo knockoff construction and $W^{(i)}$ statistic.

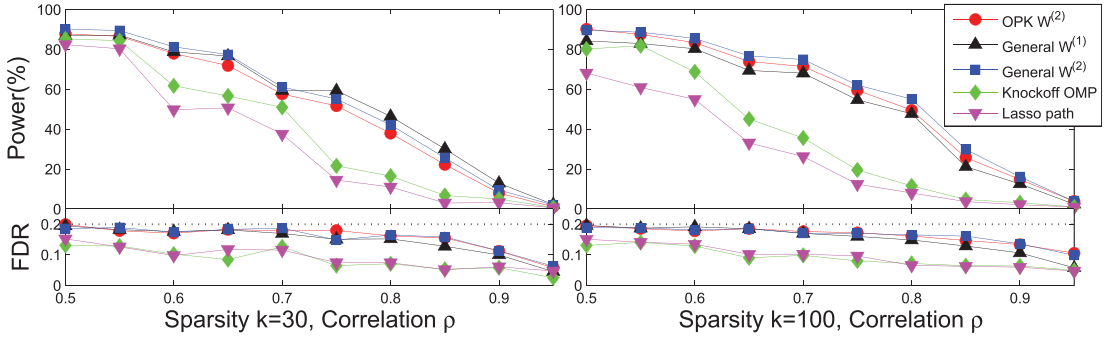

FIG. 4. Comparing the orthogonal, the general pseudo knockoff+ filter and the knockoff+ filter at nominal FDR $q = 20\%$ by varying the pairwise correlation.

Decaying structure. We consider a design matrix $X \in R^{1500\times 500}$ with some decaying structure and two sparsity cases: $k = 30$ and $k = 100$. Specifically, the design matrix X is generated from $N(0, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$, where ρ varies from $0.5, 0.55, \ldots, 0.95$. Other settings use the default values. We know *a priori* that the off-diagonal elements of Σ^{-1} decay rapidly. Thus, we apply the GPK with parameter $m = 5$. We also implement the OPK with $W^{(2)}$ statistic for comparison.

In Fig. 4, we again observe that in both figures the pseudo knockoff filters control FDR and outperform the knockoff filter with the OMP or the Lasso path statistic. We also implement the GPK with parameter $m = 2$ and two statistics $W^{(1)}$ and $W^{(2)}$. Its performance is still better than that of the knockoff filter with the OMP or the Lasso path statistic.

In these two examples with group or decaying structure, the knockoff filter with the Lasso sign max statistic $W^{(2)}$ or with the half Lasso statistic ($W^{(1)}$ version) offers more power than that of the OMP or the Lasso path statistic. Their powers are comparable to that of the OPK or the GPK. The tuning parameter $\lambda = 0.75||U^T y||_2/\sqrt{n-2p}$, which was designed for the pseudo knockoff filter with half Lasso statistic, works equally well for the knockoff filter with the Lasso or the half Lasso statistic in these two examples.

Exploring the special structure in the precision matrix. Next, we investigate how we can design an effective pseudo knockoff filter by taking advantage of the special structure in the precision matrix Σ^{-1}. We consider three examples: (a) (Σ^{-1}) is a block diagonal matrix with equal block size 5 and $(\Sigma^{-1})_{ii} = 1$, $(\Sigma^{-1})_{ij} = \rho$ for $i \neq j$ in the same block and 0 otherwise; (b) $(\Sigma^{-1})_{ij} = \rho^{|i-j|}$; and (c) $(\Sigma^{-1})_{ii} = 1$ and $(\Sigma^{-1})_{ij} = \rho$ for $i \neq j$. We then generate X from the multivariate normal distribution $N(0, \Sigma)$ as in the previous numerical examples. We vary ρ from $0.5, 0.55, \ldots, 0.95$ in example (a), (b) and from $0, 0.1, 0.2, .., 0.9$ in example (c). We consider the sparsity level $k = 30$ and focus on the pseudo knockoff filter with the half Lasso statistic and the knockoff filter with the Lasso and the half Lasso statistics. The special structure of the precision matrix suggests that choosing $m = 5$ for the GPK would be a reasonable choice for these examples. We also implement the OPK for comparison.

We observe that when we construct the knockoff matrix $\tilde{X}$ using the original knockoff filter, the difference between some feature X_i and its knockoff $\tilde{X}_i$ generated by the SDP construction is very small for some cases in example (b), (c) when ρ is large. We compute the mean s_i (see (1)) in example (c) for 10 different values of ρ that we use in this example. Their mean values for $\rho = 0, 0.1, \ldots, 0.9$ are $0.426, 0.031, 0.013, 0.007, 0.005, 0.003, 0.0019, 0.0013, 0.0007, 0.0003$, respectively. In our computation, we have used the glmnet package in MATLAB [16] to solve the Lasso optimization problem, $(\hat{\beta}, \tilde{\beta}) = \arg\min_{(\hat{b},\tilde{b})} \frac{1}{2}||y - X\hat{b} - \tilde{X}\tilde{b}||_2^2 + \lambda||(\hat{b}, \tilde{b})||_1$. The original results that we have obtained are a bit surprising in the sense that the Lasso statistic constructed this way fails to control FDR in this extreme example. To gain some understanding what goes wrong, we found that the numerical solution of this Lasso optimization problem 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 optimization problem except that we have swapped the order of the input variables $(X, \tilde{X})$. This numerical error may be attributed to the extremely small difference between X_i and $\tilde{X}_i$ for some i and the degeneracy of the augmented design matrix $[X\ \tilde{X}]$. This numerical error may lead to the violation of the flip-coin property of the knockoff statistic W constructed from the numerical solution $(\hat{\beta}, \tilde{\beta})$, which may explain why we could lose FDR control in this extreme case. To overcome this difficulty, we turn off the knockoff $\tilde{X}_i$ for X_i if s_i is small when we construct the knockoff Lasso sign max statistic. More specifically, we define an index set, $P \triangleq \{i : s_i \geq 0.001\}$. We first solve $(\hat{\beta}, \tilde{\beta}_P) = \arg\min_{\hat{b},\tilde{b}} \frac{1}{2}\|y - X\hat{b} - \tilde{X}_P\tilde{b}_P\|_2^2 + \lambda\|(\hat{b}, \tilde{b}_P)\|_1$. We then construct $W_P^{(2)} = (|\hat{\beta}_P| \vee |\tilde{\beta}_P|) \cdot \text{sign}(|\hat{\beta}_P| - |\tilde{\beta}_P|)$ and set $W_{P^c}^{(2)} = 0$. The numerical results that we present in Fig. 5 for the Lasso $W^{(2)}$ statistic are obtained using this slightly modified procedure in constructing the knockoff Lasso statistic.

In three subfigures in Fig. 5, the OPK and the GPK with the half Lasso statistic control FDR and outperform the knockoff filter with the half Lasso statistic $W^{(1)}$ (the half Lasso with $W^{(2)}$ offers less power than the half Lasso with $W^{(1)}$) and the Lasso sign max statistic. The Lasso with $W^{(1)}$ statistic offers performance similar to that of $W^{(2)}$. We have implemented the knockoff filter with the OMP and the Lasso path statistics in example (c) and found that these statistics perform poorly, which may be attributed to the smallness of s_i in this example. In general, from $0 \prec \text{diag}(s) \preceq 2X^TX \Rightarrow (X^TX)^{-1} \preceq 2(\text{diag}(s))^{-1}$, one can show that the slow decay of the off-diagonal elements of $(X^TX)^{-1}$ forces s_i to be extremely small, which could lead to a significant loss of power of the knockoff filter. The OPK with the half Lasso statistic maintains a high power in example (c), which may be attributed to the orthogonal property between X and its pseudo knockoff $\tilde{X}$. We have also tested the OPK with the least-squares statistic in example (c). Because of the slow decay of the off-diagonal elements of $(X^TX)^{-1}$, $\text{sign}(W_j^{ls})$ $1 \leq j \leq p$ are correlated for large ρ and we found that the least-squares statistic fails to control the FDR in these cases.

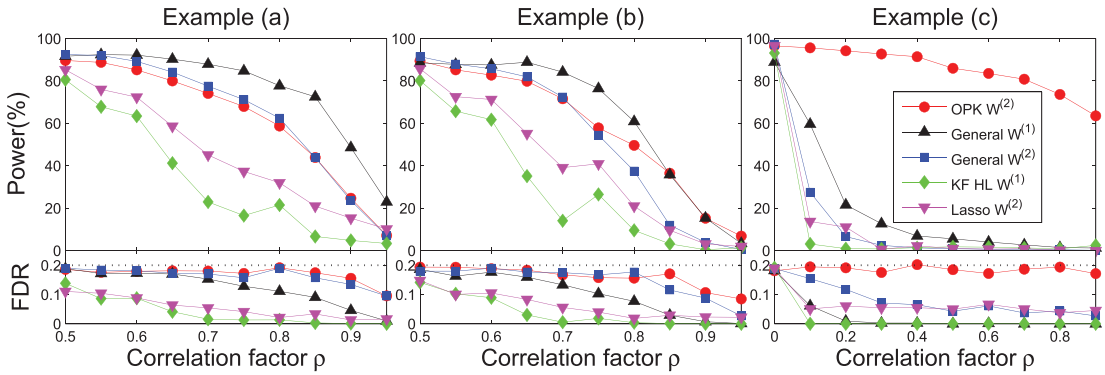

FIG. 5. Comparing the pseudo knockoff+ filter with the knockoff+ filter with several statistics at nominal FDR $q = 20\%$ by varying ρ in various precision matrices. The left, middle and right subfigures correspond to example (a), (b) and (c), respectively. *KF HL* is short for the knockoff filter with the half Lasso and *Lasso* is short for the knockoff filter with the Lasso statistic.

In these examples, we find that in the sparse case, the GPK with $W^{(1)}$ offers more power than the GPK with $W^{(2)}$, while in the non-sparse case, $W^{(2)}$ offers more power than $W^{(1)}$. In Section 4.1, we show that the GPK with $W^{(1)}$ statistic satisfies the assumptions in Theorem 1.1. Although we cannot verify these assumptions for $W^{(2)}$ statistic due to the fact that $|W_{S_0}|$ and $\text{sign}(W_{S_0})$ are not independent, we expect that Theorem 1.1 is approximately true for $W^{(2)}$ due to the sign property $\text{sign}(W^{(1)}) = \text{sign}(W^{(2)})$ and the similarity between $W^{(1)}$ and $W^{(2)}$.

4. Some analysis of the pseudo knockoff filter

In this section, we will provide some partial analysis for the pseudo knockoff filter, which may provide some understanding regarding the performance of the pseudo knockoff filter.

4.1 *A uniform FDP bound*

In the knockoff filter, the following expectation inequality,

$$E\left[\frac{\#\{j : W_j \geq T, \beta_j = 0\}}{\#\{j : W_j \leq -T, \beta_j = 0\} + 1}\right] \leq 1, \tag{26}$$

plays an important role in obtaining the exact FDR control of the knockoff filter.

The numerical experiments in Section 3 show that the pseudo knockoff with the half Lasso statistic offers FDR control and the expectation (26) is approximately valid. Since we relax one of the constraints in the knockoff filter, we cannot apply the supermartingale argument to obtain (26) for the pseudo knockoff filter. To gain some understanding why (26) may be valid for the pseudo knockoff with the half Lasso statistic, we would like to estimate the expectation (26) for fixed t and the suprema over all t in Theorem 1.1. For a technical reason, we still cannot prove (26) right now. Instead we prove a weaker version of (26) by replacing 1 in the denominator by m.

According to the assumption of W_{S_0} in Theorem 1.1, in the extreme but highly unlikely case, W_{S_0} can be m copies of $(\eta_1, \eta_2, .., \eta_L)$ where η_j are independent and symmetric random variables. Then (11) reduces to (9) with an upper bound that is about twice as large as the upper bound in (9) and (10)

reduces to $E\left[\frac{\#\{j:\ \eta_j\geq t\}}{\#\{j:\ \eta_j\leq -t\}+1}\middle|\mathcal{F}\right]\leq 1$. Since $\mathrm{sign}(\eta_j)$ are i.i.d. Rademacher random variables, the latter expectation is $1-2^{-n}$, where $n=\#\{j:\ |\eta_j|\geq t\}$. Both results in Theorem 1.1 are relatively tight. For the half Lasso statistic, this extreme scenario is very unlikely to occur since the l^1 regularization imposes sparsity and forces $\hat{\beta}_j+\tilde{\beta}_j$ to be zero for many features X_j in a correlated group. As a result, W_j is zero for many features X_j in a correlated group, and thus it is very unlikely that such an extreme scenario can be realized for the half Lasso statistic. In Section 3.2, we consider some highly correlated examples, including the cases with 0.95 within-group correlation and with 0.95 correlation between X_i and X_{i+1} for each i. These highly correlated examples in principle could generate strongly correlated W_{S_0}, pseudo knockoff filter with the half Lasso statistic still offers FDR control.

Proof of (10). Let $N_t\triangleq\{j\in S_0:|W_j|\geq t\}$. By assumption of $\mathcal{F}$, N_t is determined and we can divide N_t into m groups $C_1,C_2,\ldots,C_m$ ($C_i\subset S_0$) such that the elements of $\mathrm{sign}(W_{C_i})$ are mutually independent. Obviously, $|N_t|=\sum_{i=1}^m|C_i|$. Using the following Cauchy–Schwarz inequality

$$\sum_{i=1}^m\frac{a_i^2}{b_i}\sum_{i=1}^m b_i\geq\left(\sum_{i=1}^m a_i\right)^2\iff\frac{1}{\sum_{i=1}^m a_i}\sum_{i=1}^m\frac{a_i^2}{b_i}\geq\frac{\sum_{i=1}^m a_i}{\sum_{i=1}^m b_i},\quad a_i,b_i>0,$$

with $a_i=|C_i|+1, b_i=\#\{j\in C_i:W_j\leq -t\}+1$, we obtain

$$\begin{aligned}
&E\left[\frac{\#\{j\in S_0:W_j\geq t\}}{\#\{j\in S_0:W_j\leq -t\}+m}\middle|\mathcal{F}\right]+1=E\left[\frac{|N_t|+m}{\sum_{i=1}^m\left(\#\{j\in C_i:W_j\leq -t\}+1\right)}\middle|\mathcal{F}\right]\\
&\leq E\left[\frac{1}{|N_t|+m}\sum_{i=1}^m\frac{(|C_i|+1)^2}{\#\{j\in C_i:W_j\leq -t\}+1}\middle|\mathcal{F}\right]=\sum_{i=1}^m\frac{|C_i|+1}{|N_t|+m}E\left[\frac{|C_i|+1}{\#\{j\in C_i:W_j\leq -t\}+1}\middle|\mathcal{F}\right]\\
&=\sum_{i=1}^m\frac{|C_i|+1}{|N_t|+m}\left\{1+E\left[\frac{\#\{j\in C_i:W_j\geq t\}}{\#\{j\in C_i:W_j\leq -t\}+1}\middle|\mathcal{F}\right]\right\}.
\end{aligned}\tag{27}$$

In the above derivation, we have used $\#\{j\in C_i:W_j\leq -t\}+1+\#\{j\in C_i:W_j\geq t\}=|C_i|+1$ to obtain the first and the last equalities, and used the fact that $|N_t|$ and $|C_i|$ are measurable with respect to $\mathcal{F}$ to yield the second equality. From the assumption (b), $\mathbf{1}_{W_j>0}$ with $j\in C_i$ are mutually independent and each obeys a binomial distribution. We yield

$$E\left[\frac{\#\{j\in C_i:W_j\geq t\}}{\#\{j\in C_i:W_j\leq -t\}+1}\middle|\mathcal{F}\right]=E\left[\frac{\#\{j\in C_i:W_j>0\}}{\#\{j\in C_i:W_j<0\}+1}\middle|\mathcal{F}\right]=1-2^{-|C_i|}\leq 1.$$

Therefore, the last line in (27) is bounded by

$$\frac{1}{|N_t|+m}\sum\nolimits_{i=1}^m 2\big(|C_i|+1\big)=\frac{2}{|N_t|+m}\cdot\big(|N_t|+m\big)=2.$$

Subtracting 1 on both sides of (27) concludes the proof of (10). □

The proof of (11) is more technical and we need the following concentration inequality.

LEMMA 4.1 Assume that the σ field $\mathcal{F}$ satisfies the conditions in Theorem 1.1 and $|W_{S_0}|$ are in decreasing order : $|W_{i_1}| \geq |W_{i_2}| \geq .. \geq |W_{i_l}| > 0$, where W_{i_k}, $1 \leq k \leq l$ are all non-zero elements in W_{S_0}. Denote $V_j^{\pm} = \#\{i_k : (\pm)W_{i_k} \geq |W_{i_j}|\} = \#\{i_k : (\pm)W_{i_k} > 0 \ \& \ k \leq j\}$. For any $t > 1$ and $i < j \leq ti$, we have

$$P\left(\frac{V_j^+}{V_i^- + m} > t\right) \leq \inf_{\theta>0} \exp\left(-\theta\left(\frac{t \cdot i - j}{2} + tm\right)\right) \cdot \left(\frac{\exp(m\theta/2) + \exp(-m\theta/2)}{2}\right)^{(j-i)/m} \cdot \left(\frac{\exp((1+t)m\theta/2) + \exp(-(1+t)m\theta/2)}{2}\right)^{i/m}. \tag{28}$$

Roughly speaking, the above probability decays exponentially fast with respect to i and t. To prove (28), we first apply the Hölder inequality to decouple correlated terms and then establish a bound of the moment generating function (MGF) of $V_j^+ + tV_i^+$ similar to the Heoffding MGF bound. Finally, we apply the Laplace transform method. We will use (28) and a slicing method to control the suprema in (11). We defer the proof of (11) and Lemma 4.1 to Appendix A.

Next, we show that the pseudo knockoff statistic satisfies the assumptions in Theorem 1.1.

Independence of ξ. Let m be the largest block size of $\mathbf{B}$ in the block diagonal construction or the parameter in the general construction. Recall that the covariance matrix of $\xi = \hat{\beta}^{ls} - \tilde{\beta}^{ls}$ is $\mathbf{B}$. Since $\mathbf{B}_{C_j,C_j}$ is a diagonal matrix in the general construction, thus $\xi_i, i \in C_j$ are mutually independent.

For the block diagonal construction, we can choose C_j to be the collection of the jth element in each block if there exists such an element. Then $\xi_i, i \in C_j$ are also mutually independent.

The general construction. For $\tilde{X}$ generated by the general construction, we choose $W_i = (\hat{\beta}_i + \tilde{\beta}_i) \cdot \text{sign}(\hat{\beta}_i - \tilde{\beta}_i) = f(\eta)_i \text{sign}(\xi_i)$ (16). Let $\mathcal{F}$ be the σ field generated by η. According to the amplitude property of W, $|W|$ is $\mathcal{F}$ measurable. Since η and $\xi = \beta + \varepsilon^{(2)}$ are independent and $\xi_i, i \in C_j \cap S_0$ are mutually independent, we conclude that

$$\text{sign}(W_i) = \text{sign}\big(f(\eta)_i\big) \cdot \text{sign}\big(\beta_i + \varepsilon_i^{(2)}\big) = \text{sign}\big(f(\eta)_i\big) \cdot \text{sign}\big(\varepsilon_i^{(2)}\big), \quad i \in C_j \cap S_0$$

are symmetric and mutually independent conditional on $\mathcal{F}$. This verifies condition (b) in Theorem 1.1. The additional condition $W_{S_0} \overset{d}{=} -W_{S_0}$ follows from Proposition 2.1.

In the numerical experiments that we presented in Section 3, we have also used $W = |\hat{\beta}| \vee |\tilde{\beta}| \cdot \text{sign}(|\hat{\beta}| - |\tilde{\beta}|)$. Although we cannot prove that this statistic satisfies the assumption in Theorem 1.1, our numerical experiments seem to suggest that the FDR control is not sensitive to the choice of statistic in (16).

Block diagonal construction. If $\tilde{X}$ is generated by the block diagonal construction, we show that both statistics in (16) satisfy the assumptions in Theorem 1.1. For $W_i = (\hat{\beta}_i + \tilde{\beta}_i) \cdot \text{sign}(\hat{\beta}_i - \tilde{\beta}_i)$, we can use the same argument as above. For $W_i = |\hat{\beta}_i| \vee |\tilde{\beta}_i| \cdot \text{sign}(|\hat{\beta}_i| - |\tilde{\beta}_i|)$, $\mathcal{F}$ is the σ field generated by η and $|\xi|_{S_0}$. The amplitude property implies $|W_i|$ is $\mathcal{F}$ measurable for null i. The symmetry property of W_{S_0} follows from Proposition 2.1. It remains to verify that conditional on $\mathcal{F}$, $\text{sign}(W_i)$ are mutually independent for $i \in C_j \cap S_0$.

Note that $\mathrm{Var}(\xi) = \mathbf{B} = \mathrm{diag}(S_{11}, S_{22}, \ldots, S_{kk})$, $\xi_{S_0} = \epsilon^{(2)}_{S_0}$ and the elements of C_i come from different blocks. We can change the sign of $\epsilon^{(2)}_{S_0}$ in any block $S_{i_1 i_1}, S_{i_2 i_2}, \ldots, S_{i_j i_j}$ without changing $|\xi_{S_0}|$ and the joint distribution of $\epsilon^{(2)}_{S_0}$. Consequently, conditional on $\mathcal{F}$, $\mathrm{sign}(\xi_i)$ are mutually independent for $i \in C_j \cap S_0$. Using the independence of $\mathrm{sign}(\xi_{C_j \cap S_0})$, the sign property and the symmetry property of W_{S_0}, we verify the condition (b) in Theorem 1.1.

4.2 *Partial analysis of the orthogonal pseudo knockoff*

From the previous numerical results, we observe that the orthogonal pseudo knockoff is among the most powerful pseudo knockoffs and still maintains robust FDR control. One of the main reasons is that $\tilde{X}_j$ in OPK is orthogonal to X_j, and thus the difference between them is maximized. In this subsection, we provide some partial analysis of the orthogonal pseudo knockoff with $W^{(1)}$ statistic and expect that similar results also hold for OPK with $W^{(2)}$ statistic. First we discuss several properties of the orthogonal pseudo knockoff.

Symmetry Property. Since $X^T \tilde{X} = 0$ is symmetric, the symmetry property stated in Proposition 2.1 holds for the orthogonal pseudo knockoff.

Recall $W_j = (\hat{\beta}_j + \tilde{\beta}_j)\mathrm{sign}(\hat{\beta}_j - \tilde{\beta}_j) = f(\eta)\mathrm{sign}(\xi)$. We introduce the following notations:

$$\Sigma = X^T X, \quad D = \mathrm{diag}(\Sigma^{-1}) = \mathrm{diag}(d_1, d_2, \ldots, d_p), \quad \widetilde{\Sigma^{-1}} = D^{-1/2}\Sigma^{-1}D^{-1/2}. \tag{29}$$

By definition, we have $(\widetilde{\Sigma^{-1}})_{ii} = 1$. Let $\mathcal{F}$ be the σ field generated by η. Conditional on $\mathcal{F}$, $|W_{S_0}|$ is determined. We assume that $|W_{S_0}|$ is arranged in a decreasing order and use the same notation $V_i^{\pm}$ as in Lemma 4.1. Similar to (10) or (28), we estimate the ratio V_i^+/V_i^-.

THEOREM 4.2 For any $\delta \in (0,1)$ and $j \geq 1$, conditional on $\mathcal{F} = \sigma(\eta)$, the OPK satisfies

$$P\left(\frac{V_j^+}{V_j^-} \geq \frac{1+\delta}{1-\delta} \,\middle|\, \mathcal{F}\right) \leq \frac{(1+3\pi)\lambda_{\max}\left(\widetilde{\Sigma^{-1}}_{N_j N_j}\right)}{\pi\delta^2 j}, \tag{30}$$

where $N_j \triangleq \{i_k : |W_{i_k}| \geq |W_{i_j}|\}$ and $\lambda_{\max}(\widetilde{\Sigma^{-1}}_{N_j N_j})$ is the largest eigenvalue of the submatrix $\widetilde{\Sigma^{-1}}_{N_j N_j}$.

REMARK 4.3 Note that the diagonal elements of $\widetilde{\Sigma^{-1}}_{N_j N_j}$ are all 1 and $|N_j| = j$. Note that $j = \mathrm{Tr}(\widetilde{\Sigma^{-1}}_{N_j N_j}) = \sum_{i=1}^{j} \lambda_i(\widetilde{\Sigma^{-1}}_{N_j N_j})$ and $\lambda_i(\widetilde{\Sigma^{-1}}_{N_j N_j}) > 0$. Thus, we have $\lambda_{\max}(\widetilde{\Sigma^{-1}}_{N_j N_j}) < j$.

From the sign property of W, we know $\mathrm{sign}(W_{S_0}) = \mathrm{sign}((f(\eta))_{S_0}) \cdot \mathrm{sign}(\xi_{S_0})$. Denote $Y_i = \mathbf{1}_{W_i > 0}$. We first analyse the covariance of each pair (Y_i, Y_j), $i, j \in S_0$.

LEMMA 4.4 Conditional on η, for any null variable i, j, we have

$$\mathrm{Cov}(Y_i, Y_j|\eta) \leq \frac{1}{2\pi}(\widetilde{\Sigma^{-1}})_{ij}\left(\mathbf{1}_{(f(\eta))_i > 0} - \mathbf{1}_{(f(\eta))_i < 0}\right)\left(\mathbf{1}_{(f(\eta))_j > 0} - \mathbf{1}_{(f(\eta))_j < 0}\right) + \frac{3}{2}(\widetilde{\Sigma^{-1}})_{ij}^2. \tag{31}$$

We will defer the proof to Appendix B.

Proof of Theorem 4.2. According to the symmetry property (Proposition 2.1) of OPK, for $i \in N_j (\subset S_0)$, we have

$$E(Y_i|\eta) = E(\mathbf{1}_{W_i>0}|\eta) = 1/2, \quad E\big(V_j^+|\eta\big) = E\big(V_j^-|\eta\big) = j/2, \quad V_j^+ + V_j^- = j. \tag{32}$$

Denote $w_i \triangleq \mathbf{1}_{f_i(\eta)>0} - \mathbf{1}_{f_i(\eta)<0}$. Using (31) and $\widetilde{\Sigma^{-1}}_{N_jN_j} \preceq \lambda_{\max}(\widetilde{\Sigma^{-1}}_{N_jN_j})\mathbf{I}$, we obtain

$$\begin{aligned}\mathrm{Var}(V_j^+|\eta) &\le \sum_{s,t\in N_j} \mathrm{Cov}(Y_s, Y_t|\eta) \le \sum_{s,t\in N_j} \frac{1}{2\pi}(\widetilde{\Sigma^{-1}})_{st} w_s w_t + \frac{3}{2}(\widetilde{\Sigma^{-1}})_{st}^2 \\ &= \frac{w_{N_j}^T (\widetilde{\Sigma^{-1}})_{N_jN_j} w_{N_j}}{2\pi} + \frac{3}{2}\mathrm{Tr}\big((\widetilde{\Sigma^{-1}})_{N_jN_j}^2\big) \le \frac{\lambda_{\max}\big(\widetilde{\Sigma^{-1}}_{N_jN_j}\big)j}{2\pi} + \frac{3\lambda_{\max}\big(\widetilde{\Sigma^{-1}}_{N_jN_j}\big)j}{2}. \end{aligned} \tag{33}$$

Conditional on η, we apply (32), (33) and the Chebyshev inequality to yield

$$\begin{aligned} P\big(V_j^- \le (1-\delta)j/2 \,|\, \eta\big) &= P\big(V_j^- \ge (1+\delta)j/2 \,|\, \eta\big) = \frac{1}{2}P\big(|V_j^- - j/2| \ge \delta j/2 \,|\, \eta\big) \\ &\le \frac{2\mathrm{Var}\big(V_j^- \,|\, \eta\big)}{(\delta j)^2} \le \frac{(1+3\pi)\lambda_{\max}\big(\widetilde{\Sigma^{-1}}_{N_jN_j}\big)}{\pi\delta^2 j}. \end{aligned} \tag{34}$$

The first identity holds since the symmetry property (17) implies that $V_j^- \overset{d}{=} V_j^+ = j - V_j^-$. The estimate (30) follows from integrating the last inequality in (34). □

For some design matrices that have certain special structure in $\widetilde{\Sigma^{-1}}$, e.g. the design matrices to be considered in the next subsection, we can show using (33) that $\mathrm{Var}(V_j^+|\eta) = O(j)$. Conditional on η, if $\mathrm{sign}(W_i)$ $i \in N_j$ are independent, which is true if we use the knockoff statistic, we have $\mathrm{Var}(V_j^+|\eta) = j/4$. In this case, $\mathrm{Var}(V_j^+|\eta)$ is the same order as that in the knockoff for all j.

4.3 *Some special design matrices*

For some special design matrices, we can improve the estimate of $\mathrm{Var}(V_j^+|\eta)$ in (33) and get better control of V_j^+/V_j^-. In our simulations, we observe that the OPK offers robust FDR control. We would like to offer a partial explanation of this phenomenon.

A diagonally dominated case. Let $X \in R^{n\times p}$ and $\Sigma = X^TX$. We consider several classes of design matrices described below.

(a) For any $i \ne j$, $\langle X_i, X_j\rangle \triangleq X_i^TX_j = \rho$, $\rho \in [0, 1)$.

(b) Assume that X can be clustered into k groups, $X = (X_{C_1}, X_{C_2}, \ldots, X_{C_k})$. The within-group correlation of group i is ρ_i for some $\rho_i \in [0, 1)$ and the between-group correlation is zero.

(c) The sizes of different groups are equal. The within-group correlation is ρ and the between-group correlation is $\gamma \cdot \rho$.

Case (a) corresponds to setting (a), (b) and (d) with $\rho = 0$ in Section 3.1; case (b) and (c) correspond to setting (e) and (f) in Section 3.1. Denote $\mathbf{E} \triangleq \Sigma^{-1}$ for convenience. From (29),

$(\widetilde{\Sigma^{-1}})_{ij} = \mathbf{E}_{ij}/(\mathbf{E}_{ii}^{1/2}\mathbf{E}_{jj}^{1/2})$. For the design matrices described above, we can show that Σ^{-1} is diagonally dominated, i.e. $\sum_{j\neq i}|(\Sigma^{-1})_{ij}| < \Sigma_{ii}^{-1}$. The proof is a bit technical and tedious. We will omit the proof here. From Lemma 4.4, we have

$$\mathrm{Cov}(Y_i, Y_j|\eta) \le \frac{1}{2\pi}(\widetilde{\Sigma^{-1}})_{ij}w_i w_j + \frac{3}{2}(\widetilde{\Sigma^{-1}})_{ij}^2 \le c_0\big|(\widetilde{\Sigma^{-1}})_{ij}\big|, \quad c_0 = \frac{1}{2\pi} + \frac{3}{2} < 2. \tag{35}$$

Since Σ^{-1} is diagonally dominated, we can improve the estimate of $\mathrm{Var}(V_j^+|\eta)$ in (33)

$$\begin{aligned}\mathrm{Var}(V_j^+|\eta) &\le \sum_{s,t\in N_j} c_0|(\widetilde{\Sigma^{-1}})_{st}| = c_0\sum_{s,t\in N_j}\frac{|\mathbf{E}_{st}|}{\mathbf{E}_{ss}^{1/2}\mathbf{E}_{tt}^{1/2}} \le c_0\left(\sum_{s,t\in N_j}\frac{|\mathbf{E}_{st}|}{\mathbf{E}_{ss}}\right)^{1/2}\left(\sum_{s,t\in N_j}\frac{|\mathbf{E}_{st}|}{\mathbf{E}_{tt}}\right)^{1/2}\\ &= c_0\left(\sum_{s\in N_j}\frac{1}{\mathbf{E}_{ss}}\sum_{t\in N_j}|\mathbf{E}_{st}|\right)^{1/2}\left(\sum_{t\in N_j}\frac{1}{\mathbf{E}_{tt}}\sum_{s\in N_j}|\mathbf{E}_{st}|\right)^{1/2} \le c_0\left(\sum_{s\in N_j}\frac{2\mathbf{E}_{ss}}{\mathbf{E}_{ss}}\right)^{1/2}\left(\sum_{t\in N_j}\frac{2\mathbf{E}_{tt}}{\mathbf{E}_{tt}}\right)^{1/2}\\ &= 2c_0 j.\end{aligned}$$

Here, we have used $\mathbf{E}_{st} = \mathbf{E}_{ts}$ and the diagonal dominated assumption to yield $\sum_{t\in N_j}|\mathbf{E}_{st}| \le \sum_{t=1}^{p}|\mathbf{E}_{st}| \le 2\mathbf{E}_{ss}$. With this refined estimate of $\mathrm{Var}(V_j^+|\eta)$, the upper bound in Theorem 4.2 can be reduced to $\frac{2+6\pi}{\pi\delta^2 j}$.

Exponentially Decaying Class. Assume that $|(\Sigma^{-1})_{ij}| \le C\rho^{|i-j|}$ for $\rho \in [0,1)$ and some constant C. The design matrix in setting (c) in Section 3.1 has a similar structure. One can prove that $(\widetilde{\Sigma^{-1}})_{ii} \ge 1$ using the fact that $\Sigma_{ii} = 1$ and Σ is positive definite. By our assumption, we have $|(\widetilde{\Sigma^{-1}})_{ij}| \le |(\Sigma^{-1})_{ij}| \le C\rho^{|i-j|}$. Hence, we have $\lambda_{\max}(\widetilde{\Sigma^{-1}}) \le ||\widetilde{\Sigma^{-1}}||_{l_1} \le 2C/(1-\rho)$. Denote $c_0 = (1+3\pi)/(2\pi)$. Using (33) and Theorem 4.2, we yield

$$\mathrm{Var}(V_j^+|\eta) \le c_0\lambda_{\max}(\widetilde{\Sigma^{-1}})j \le \frac{2c_0 Cj}{1-\rho}, \quad P\left(\frac{V_j^+}{V_j^-} \ge \frac{1+\delta}{1-\delta}\middle|\mathcal{F}\right) \le \frac{4c_0 C}{\delta^2(1-\rho)j}.$$

Therefore, for all the design matrices that we considered in Section 3.1 (up to randomness), we have $\mathrm{Var}(V_j^+|\eta) = O(j)$ for all j. This may offer some partial explanation why we observe robust FDR control of OPK in these examples.

5. Concluding remarks

In this paper, we proposed a pseudo knockoff filter for feature selection with correlated features. Both the BDPK and the GPK filters preserve some essential features of the original knockoff filter, but offer more flexibility in constructing the knockoff matrix. We also proposed the OPK filter. Our numerical experiments seem to suggest that the pseudo knockoff filters have FDR control in the numerical examples that we considered in this paper. Moreover, the OPK and GPK filters seem to offer more power than the knockoff filter with the Lasso Path and the half Lasso statistics in these examples, especially

when the features are highly correlated. For the BDPK and the GPK filters, we provided an estimate for the expectation of the ratio $\frac{\#\{j\in S_0:\ W_j\geq t\}}{\#\{j\in S_0:\ W_j\leq -t\}+m}$ for any fixed threshold in (10) and its suprema over all possible thresholds in (11) under weaker assumptions on the conditional distribution of the statistic. For the orthogonal pseudo knockoff filter, we provided some estimate of the distribution function (30). This estimate provides a relatively tight upper bound when Σ^{-1} is diagonally dominated or when Σ^{-1} has some special structure. Although our analysis does not lead to FDR control, it may offer some partial understanding of the pseudo knockoff filter.

We would like to emphasize that our understanding of the pseudo knockoff filter is still quite limited. In some extreme cases, we found that the orthogonal pseudo knockoff filter with the least square statistic fails to control FDR. Although we have better understanding of the OPK with the half Lasso statistic and obtained better theoretical results for the GPK, these results do not provide a satisfactory explanation for the robust performance of the pseudo knockoff filter with the half Lasso statistic that we observed numerically. In our future study, we would like to further investigate whether one can find some appropriate conditions on the design matrices under which we can obtain exact FDR control for the pseudo knockoff filter with the half Lasso statistic. This question seems to be extremely difficult. Some new method of analysis needs to be developed to give an affirmative answer to this question.

Acknowledgements

The research of J.C. was performed during his visit to ACM at Caltech. We would like to thank Professor Emmanuel Candes for his many valuable comments and suggestions to our work. We would also like to thank Professor Lucas Janson for his interest and comments on the earlier version of this manuscript and Dr. Pengfei Liu for the discussions on the pseudo knockoff.

Funding

National Science Foundation (DMS 1318377 and DMS 1613861).

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. *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. Candès, E. J., Fan, Y., Janson, L. & Lv, J. (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. B*, **80**, 551–577. doi:10.1111/rssb.12265.
6. Chen, J., Hou, A. & Hou, T. Y. (2017) Some analysis of the knockoff filter and its variants. arXiv:1706.03400.
7. 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.
8. G'Sell, M. G., Wager, S., Chouldechova, A. & Tibshirani, R. (2016) Sequential selection procedures and false discovery rate control. *J. R. Stat. Soc. B*, **78**, 423–444. doi:10.1111/rssb.12122.

9. Janson, L. & Su, W. (2016) Familywise error rate control via knockoffs. *Electron. J. Statist.*, **10**, 960–975. doi:10.1214/16-EJS1129.

10. Katsevich, E. & Ramdas, A. (2018) Towards 'simultaneous selective inference': post-hoc bounds on the false discovery proportion. arXiv:1803.06790.

11. Katsevich, E. & Sabatti, C. (2017) Multilayer knockoff filter: controlled variable selection at multiple resolutions. arXiv:1706.09375.

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

13. Meinshausen, N. & Bühlmann, P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **72** 417–473.

14. Miller, A. J. (1984) Selection of subsets of regression variables. *J. Roy. Statist. Soc. Ser. A*, **147**, 389–425.

15. Miller, A. (2002) *Subset Selection in Regression*, vol. 95 of Monographs on Statistics and Applied Probability, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

16. Qian, J., Hastie, T., Friedman, J., Tibshirani, R. & Simon, N. (2013) *Glmnet for Matlab*. http://www.stanford.edu/∼hastie/glmnet_matlab/.

17. Reid, S. & Tibshirani, R. (2016) Sparse regression and marginal testing using cluster prototypes. *Biostat.*, **17**, 364–376.

18. Su, W., Qian, J. & Liu, L. (2015) Communication-efficient false discovery rate control via knockoff aggregation. arXiv:1506.05446v2.

Appendices

A. Proof of Theorem 1.1

The derivations in this Appendix are conditional on $\mathcal{F}$ and we drop the notation of conditional expectation for simplicity.

Proof of Lemma 4.1. We first estimate the MGF of $V_j^+ + tV_i^+$ and then apply the Laplace transform method to establish (28). Denote $\xi_k = \mathbf{1}_{W_{i_k}>0} - 1/2$, $\lambda_k = 1 + t$ for $k \le i$ and $\lambda_k = 1$ for $i < k \le j$. Since $|W_{i_l}|$ is decreasing, we obtain

$$V_j^+ + tV_i^+ - \frac{j+ti}{2} = \#\{k \le j : W_{i_k} > 0\} + t\#\{k \le i : W_{i_k} > 0\} - \frac{j+ti}{2} = \sum_{k\le j} \lambda_k \xi_k, \quad V_i^- = i - V_i^+.$$

According to the assumption of W_{S_0} in Theorem 1.1, conditional on $\mathcal{F}$, we can divide W_{S_0} into m groups $C_1, C_2, \ldots, C_m$ such that $\mathrm{sign}(W_k), k \in C_l$ are independent. We can use the Hölder inequality to decouple correlated terms and estimate the MGF of $V_j^+ + tV_i^+ - \frac{j+ti}{2}$ for any $\theta > 0$ as follows:

$$G(\theta) = E\exp\left(\theta\left(V_j^+ + tV_i^+ - (j+ti)/2\right)\right) = E\exp\left(\sum_{l=1}^{m} \sum_{k\in C_l,\ k\le j} \lambda_k \xi_k \theta\right)$$

$$\le \prod_{l=1}^{m} \left\{ E\exp\left(\sum_{k\in C_l,\ k\le j} m\lambda_k \xi_k \theta\right)\right\}^{1/m} = \prod_{l=1}^{m} \prod_{k\in C_l,\ k\le j} \left(E\exp(m\lambda_k \xi_k \theta)\right)^{1/m},$$

where we have used the fact that $\xi_k = \mathbf{1}_{W_{i_k}>0} - 1/2,\ k \in C_l$ are independent to yield the last equality. The symmetry assumption of $\mathrm{sign}(W_j)$ in Theorem 1.1 implies $\xi_k \sim \{\pm 1/2\}$. Using the definition of λ_k,

we obtain

$$G(\theta) \le \prod_{l=1}^{m} \prod_{k\in C_l,\, k\le j} \left(E\exp(m\lambda_k\xi_k\theta)\right)^{1/m} = \prod_{k=1}^{j} \left(E\exp(m\lambda_k\xi_k\theta)\right)^{1/m}$$
$$= \left(\frac{\exp\left((1+t)m\theta/2\right) + \exp\left(-(1+t)m\theta/2\right)}{2}\right)^{i/m} \left(\frac{\exp(m\theta/2)+\exp(-m\theta/2)}{2}\right)^{(j-i)/m}. \tag{A.1}$$

To simplify the notations, we define $B(x,y,t,\xi,s)$ as follows:

$$\exp\left(-\xi\left(\frac{tx-y}{2}+s\right)\right)\left(\frac{\exp\left((1+t)\xi/2\right)+\exp\left(-(1+t)\xi/2\right)}{2}\right)^{x}\left(\frac{\exp(\xi/2)+\exp(-\xi/2)}{2}\right)^{y-x}. \tag{A.2}$$

For fixed $t,\xi > 0$, it is easy to verify that $B(x,y,t,\xi,s)$ is a monotonically decreasing function of x,s and an increasing function of y. Choosing $\xi = m\theta, x = i/m$ and $y = j/m$, we can simplify (A.1) as follows:

$$E\exp\left(\theta\left(V_j^+ + tV_i^+ - (j+ti)/2\right)\right) = G(\theta) \le \exp\left(m\theta\left(\frac{ti-j}{2m}+s\right)\right) B\left(\frac{i}{m},\frac{j}{m},t,m\theta,s\right) \tag{A.3}$$

for any s. Using (A.3) and applying the Markov inequality to $\exp(\theta(V_j^+ + tV_i^+ - (j+ti)/2))$ for any $\theta > 0$, we yield

$$P\left(V_j^+ > tV_i^- + sm\right) = P\left(V_j^+ + tV_i^+ - \frac{ti+j}{2} > \frac{ti-j}{2} + sm\right)$$
$$\le \inf_{\theta>0} \exp\left(-\theta\left(\frac{ti-j}{2}+sm\right)\right) E\left\{\exp\left(\theta\left(V_j^+ + tV_i^+ - \frac{j+ti}{2}\right)\right)\right\}$$
$$\le \inf_{\theta>0} B\left(\frac{i}{m},\frac{j}{m},t,m\theta,s\right) = \inf_{\xi>0} B\left(\frac{i}{m},\frac{j}{m},t,\xi,s\right). \tag{A.4}$$

Choosing $s = t$ in (A.4) establishes (28). □

One can obtain the following Hoeffding-type concentration inequality using a similar argument:

$$P\left(\frac{V_i^+}{V_i^- + m} > t\right) = P\left(V_i^+ - \frac{i}{2} > \frac{(t-1)i+2tm}{2(1+t)}\right) \le \exp\left(-\left(\frac{(t-1)i+2tm}{2(1+t)}\right)^2 \frac{2}{mi}\right), \tag{A.5}$$

where $t > 1$. The key insight is that for large i, the above probability decays exponentially fast. Thus it is possible to estimate the suprema of $V_i^+/(V_i^- + m)$ using some covering argument.

Proof of Theorem 1.1 We estimate the distribution function of the suprema. Since $V_i^+ \le i$, we get

$$P\left(\sup_{s>0} \frac{\#\{j\in S_0 : W_j \ge s\}}{\#\{j\in S_0 : W_j \le -s\} + m} > t\right) = P\left(\sup_{i\ge 1} \frac{V_i^+}{V_i^- + m} > t\right) = P\left(\sup_{i>\lfloor tm\rfloor} \frac{V_i^+}{V_i^- + m} > t\right). \tag{A.6}$$

Once we obtain the above estimate, we can integrate t from 0 to ∞ to yield (11). Based on (A.4) or (A.5), for any fixed $t > 1$, $P(V_i^+/(V_i^- + m) > t)$ should be exponentially small for large i. For small i,

the concentration inequality is not sharp and we can use the symmetry of the joint distribution of W_{S_0}, i.e. $W_{S_0} \overset{d}{=} -W_{S_0}$, to obtain a better estimate. Since V_i^+, V_i^- are monotone, the supremum over an interval $\sup_{i_1<i\le i_2} V_i^+/(V_i^- + m)$ can be bounded by $V_{i_2}^+/(m+V_{i_1+1}^-)$. We will split $i > \lfloor tm \rfloor$ into several intervals with well-chosen end points $(i_k, i_{k+1}]$ and then apply (A.4). □

Estimate for small $t < t^* = 4$. Denote $a = 2\lfloor tm \rfloor + 1$. We split the distribution into two parts:

$$P\left(\sup_{i>\lfloor tm\rfloor} \frac{V_i^+}{V_i^- + m} > t\right) \le P\left(\sup_{i>\lfloor tm\rfloor} \frac{V_i^+}{V_i^- + m} > t, V_a^- \ge \frac{a+1}{2}\right) + P\left(V_a^- \le \frac{a-1}{2}\right) \triangleq I + II. \quad \text{(A.7)}$$

The probability that the supremum over some small i is larger than t is high. Fortunately, we can use II to take care of the contribution of small i. Since $W_{S_0} \overset{d}{=} -W_{S_0}$, we have $V_a^- \overset{d}{=} V_a^+ = a - V_a^-$, and thus $II = 1/2$. In I, $V_a^- \ge (a+1)/2$ implies that the denominator is not small. Thus $V_i^+/(V_i^- + m) > t$ cannot be true for small i. In fact, the monotonicity of $V_i^\pm$ and $V_i^+ + V_i^- = i$ implies

$$V_a^- \ge \frac{a+1}{2} = \lfloor tm \rfloor + 1 \Rightarrow \sup_{\lfloor tm\rfloor < i \le a} \frac{V_i^+}{V_i^- + m} \le \frac{V_a^+}{m} \le t, \quad \sup_{a<i\le \lfloor t^2 m\rfloor + a} \frac{V_i^+}{V_i^- + m} \le \frac{\lfloor t^2 m\rfloor + \lfloor tm\rfloor}{\lfloor tm\rfloor + 1 + m} \le t.$$

Therefore, the term I mainly takes care of the contribution of large i and can be reduced to

$$I = P\left(\sup_{i>\lfloor t^2 m\rfloor + a} \frac{V_i^+}{V_i^- + m} > t, V_a^- \ge \frac{a+1}{2}\right) \le P\left(\sup_{i>\lfloor t^2 m\rfloor + a} \frac{(V_i^+ - V_a^+) + \frac{a+1}{2}}{(V_i^- - V_a^-) + \frac{a+1}{2} + m} > t\right).$$

We freeze $V_a^\pm$ and introduce a new random process $U_j^\pm = V_{j+a}^\pm - V_a^\pm$. According to the definition of $V_j^\pm$ in Lemma 4.1, we have $U_j^\pm = \#\{i_k : (\pm)W_{i_k} > 0 \ \&\ a < k \le j + a\}$ and it is the same as $V_j^\pm$ after throwing away $W_{i_1}, W_{i_2}, .., W_{i_a}$. Thus, the random process $\{U_j^\pm\}_{j\ge1}$ and $\{V_j^\pm\}_{j\ge1}$ have the same properties and the concentration inequality (A.4) holds true for $U_j^\pm$. For any increasing sequence $\{s_i\}_{i\ge1}$ with $s_1 = t^2$, we obtain

$$I \le P\left(\sup_{i>\lfloor t^2 m\rfloor} \frac{U_i^+ + \frac{a+1}{2}}{U_i^- + \frac{a+1}{2} + m} > t\right) \le \sum_{k\ge1} P\left(\sup_{\lfloor s_k m\rfloor < i \le \lfloor s_{k+1} m\rfloor} \frac{U_i^+ + \frac{a+1}{2}}{U_i^- + \frac{a+1}{2} + m} > t\right)$$

$$\le \sum_{k\ge1} P\left(\frac{U_{\lfloor s_{k+1} m\rfloor}^+ + \frac{a+1}{2}}{U_{\lfloor s_k m\rfloor+1}^- + \frac{a+1}{2} + m} > t\right) = \sum_{k\ge1} P\left(U_{\lfloor s_{k+1} m\rfloor}^+ > tU_{\lfloor s_k m\rfloor+1}^- + (t-1)\frac{a+1}{2} + tm\right).$$

Denote $r_k = (t-1)\frac{a+1}{2} + tm$. Applying (A.4) with $s = \frac{r_k}{m}$ gives

$$P\left(U_{\lfloor s_{k+1} m\rfloor}^+ > tU_{\lfloor s_k m\rfloor+1}^- + (t-1)\frac{a+1}{2} + tm\right) \le \inf_{\xi>0} B\left(\frac{\lfloor s_k m\rfloor + 1}{m}, \frac{\lfloor s_{k+1} m\rfloor}{m}, t, \xi, \frac{r_k}{m}\right).$$

Recall that $a = 2\lfloor tm \rfloor + 1$. We obtain

$$\frac{r_k}{m} = (t-1)\frac{a+1}{2m} + t > (t-1)t + t = t^2.$$

Using the monotonicity of B in x, y, s variables in (A.2), we obtain

$$I \le \sum_{k\ge 1} \inf_{\xi>0} B\left(\frac{\lfloor s_k m\rfloor + 1}{m}, \frac{\lfloor s_{k+1} m\rfloor}{m}, t, \xi, \frac{r_k}{m}\right) \le \sum_{k\ge 1} \inf_{\xi>0} B\left(s_k, s_{k+1}, t, \xi, t^2\right).$$

The upper bound of I, II is independent of m. Thus we can estimate (A.6) uniformly for m.

Estimate for large $t \ge t^* = 4$**.** For large t, $i > \lfloor tm\rfloor$ is large and (A.4) can be sharp. Choosing any increasing sequence $\{s_k\}_{k\ge 1}$ with $s_1 = t$ and then applying (A.4) with $s = t$, we obtain

$$\begin{aligned} P\left(\sup_{i>\lfloor tm\rfloor} \frac{V_i^+}{V_i^- + m} > t\right) &\le \sum_{k\ge 1} P\left(\sup_{\lfloor s_k m\rfloor < i \le \lfloor s_{k+1} m\rfloor} \frac{V_i^+}{V_i^- + m} > t\right) \le \sum_{k\ge 1} P\left(\frac{V^+_{\lfloor s_{k+1} m\rfloor}}{V^-_{\lfloor s_k m\rfloor + 1} + m} > t\right) \\ &\le \sum_{k\ge 1} \inf_{\xi>0} B\left(\frac{\lfloor s_k m\rfloor + 1}{m}, \frac{\lfloor s_{k+1} m\rfloor}{m}, t, \xi, t\right) \le \sum_{k\ge 1} \inf_{\xi>0} B\left(s_k, s_{k+1}, t, \xi, t\right), \end{aligned} \tag{A.8}$$

where we have used the monotonicity of B in x, y and s variables (A.2) to obtain the last inequality.

Choosing s_k**.** For a fixed t, we use a greedy strategy to optimize the selection of s_k so that we have a sharp upper bound. Assume that $s_k, k \ge 1$ is obtained. The candidate values of s_{k+1} are $C = \{s_k + ih : i = 1, , 2.., 19\}$, $h = (t-1)s_k/20$. For each $s \in C$, we construct an arithmetic sequence $a_i \triangleq (s - s_k)\cdot i + s_k, i = 0, 1, \ldots, 30$. Then we choose s_{k+1} as follows:

$$s_{k+1} = \arg\min_{s\in C}\left(\sum\nolimits_{i=1}^{30} \min_{\xi\in D_i} B\left(a_{i-1}, a_i, t, \xi, \eta_t\right)\right), \tag{A.9}$$

where $\eta_t = t^2$ for small $t < t^*$ and $\eta_t = t$ for $t \ge t^*$. Using $\mathrm{e}^x + \mathrm{e}^{-x} \le 2\exp(x^2/2)$, we know

$$B(x, y, t, \xi, s) \le \exp\big(-\xi\big((tx - y)/2 + s\big)\big)\exp\big(\xi^2(y - x)/8\big)\exp\big((1 + t)^2\xi^2 x/8\big).$$

The minimizer of the right-hand side is $\xi^*(x, y, t, s) = \frac{2(tx-y)+4s}{y-x+(t+1)^2 x}$. We choose

$$D_i = \left\{\xi^*/3 + 0.01\, j : \xi^*/3 + 0.01\, j \in \left[\xi^*/3, 3\xi^*\right]\right\}, \quad \xi^* = \xi^*(a_{i-1}, a_i, t, \eta_t),$$

in (A.9) and approximate $\inf_{\xi>0} B\left(a_{i-1}, a_i, t, \xi, \eta_t\right)$ by $\min_{\xi\in D_i} B\left(a_{i-1}, a_i, t, \xi, \eta_t\right)$. We stop constructing s_k if $s_k > 150$. We denote by $k_t \in Z$ the subindex of the last term and then $s_{k_t} > 150$.

The remaining part. The remaining part can be arbitrary small if we construct s_k over a large range and calculate large t in the last step numerically. For $2.4 \le t \le 15$, we use the above procedure to estimate $P(\sup_{\lfloor tm\rfloor < i \le 150m} V_i^+/(V_i^- + m) > t)$. To estimate the remaining part $P(\sup_{150m<i} V_i^+/(V_i^- + m) > t)$, we choose $\tilde{s}_i \triangleq s_{k_t - 150 + i + 1} = i$ for $i \ge 150$. From (A.2), we know

$$\begin{aligned} B\left(\tilde{s}_i, \tilde{s}_{i+1}, t, \xi, t\right) &= \exp\left(-\xi\left(\frac{ti - i - 1}{2} + t\right)\right)\frac{\mathrm{e}^{\xi/2} + \mathrm{e}^{-\xi/2}}{2}\cdot\left(\frac{\mathrm{e}^{(1+t)\xi/2} + \mathrm{e}^{-(1+t)\xi/2}}{2}\right)^i \\ &= \mathrm{e}^{-\xi t}\cdot\frac{\mathrm{e}^{\xi} + 1}{2}\left(\frac{\mathrm{e}^{\xi} + \mathrm{e}^{-t\xi}}{2}\right)^i = c_\xi\cdot\mathrm{e}^{-t\xi}\cdot r(t, \xi)^i, \end{aligned} \tag{A.10}$$

where $r(t,\xi) \triangleq \frac{e^{\xi}+e^{-t\xi}}{2}$. We can choose $\xi \in (0,1]$ such that $r(t,\xi) < 1-\varepsilon$ uniformly for $t \geq 2.4$ and some $\varepsilon > 0$. It follows that the tail $B(\tilde{s}_i, \tilde{s}_{i+1}, t, \xi, t)$ decays exponentially fast with respect to i,t. To estimate $P(\sup_{150m<i} V_i^+/(V_i^- + m) > t)$, we choose $\xi = 0.2$ for $t \in [2.4, 15]$ and obtain $r(\xi,t) < 0.93$.

For $t > 15$, we choose $s_i = t + i - 1$ for $i \geq 1$, $\xi = 0.5$ and yield $r(t,\xi) < 0.83$. Note that (A.10) still holds true after replacing $(\tilde{s}_i, \tilde{s}_{i+1}, i)$ by (s_i, s_{i+1}, s_i). Thus we can estimate the distribution function $P\left(\sup_{i>\lfloor tm \rfloor} V_i^+/(V_i^- + m) > t\right)$ in (A.8) directly, which decays exponentially fast with respect to t.

After obtaining the upper bound of the distribution function for $t_i = 2.4 + 0.005i \in [2.4, 15]$ and any $t > 15$, we use the monotonicity of the distribution function and integrate (A.6) to conclude

$$E\left[\sup_{i\geq 1} \frac{V_i^+}{V_i^- + m}\right] = \int_0^\infty P\left(\sup_{i>\lfloor tm \rfloor} \frac{V_i^+}{V_i^- + m} > t\right) \mathrm{d}t \leq 2.4 + \int_{2.4}^\infty P\left(\sup_{i>\lfloor tm \rfloor} \frac{V_i^+}{V_i^- + m} > t\right) \mathrm{d}t \leq 3.9.$$

Verification of the construction (21) of $\tilde{X}$. Direct calculations show that

$$\begin{aligned}
(X-\tilde{X})^T(X-\tilde{X}) &= \left[(2X\Sigma^{-1} - 2\mathbf{UC})\mathbf{B}^{-1}\right]^T\left[(2X\Sigma^{-1} - 2\mathbf{UC})\mathbf{B}^{-1}\right] \\
&= 4\mathbf{B}^{-1}(\Sigma^{-1}X^TX\Sigma^{-1} + \mathbf{C}^T\mathbf{U}^T\mathbf{UC})\mathbf{B}^{-1} = 4\mathbf{B}^{-1}(\Sigma^{-1} + \mathbf{C}^T\mathbf{C})\mathbf{B}^{-1} = 4\mathbf{B}^{-1}. \\
(X+\tilde{X})^T(X-\tilde{X}) &= \left[X(2\mathbf{I} - 2\Sigma^{-1}\mathbf{B}^{-1}) + 2\mathbf{UCB}^{-1}\right]^T\left[2X\Sigma^{-1}\mathbf{B}^{-1} - 2\mathbf{UCB}^{-1}\right] \\
&= 4(\mathbf{I} - \Sigma^{-1}\mathbf{B}^{-1})^TX^TX\Sigma^{-1}\mathbf{B}^{-1} - 4\mathbf{B}^{-1}\mathbf{C}^T\mathbf{U}^T\mathbf{UCB}^{-1} \\
&= 4(\mathbf{I} - \mathbf{B}^{-1}\Sigma^{-1})\mathbf{B}^{-1} - 4\mathbf{B}^{-1}C^TC\mathbf{B}^{-1} \\
&= 4(\mathbf{I} - \mathbf{B}^{-1}\Sigma^{-1})\mathbf{B}^{-1} - 4\mathbf{B}^{-1}(\mathbf{B} - \Sigma^{-1})\mathbf{B}^{-1} = 0.
\end{aligned}$$

Here we use $\mathbf{U}^TX = X^T\mathbf{U} = 0$. The first identity implies (19) and the second is exactly (6).

B. Proof of Lemma 4.4

Conditional on η, we can determine $N_\eta = \{j \in S_0 : W_j \neq 0\}$. Recall that ξ and η are independent and $\xi_{S_0} \overset{d}{=} -\xi_{S_0}$. We have $E(\mathbf{1}_{\xi_i>0}|\eta) = E(\mathbf{1}_{\xi_i>0}) = 1/2, i \in S_o$. For any $i,j \in N_\eta$, we get

$$E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j<0}) - \frac{1}{4} + E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4} = E(\mathbf{1}_{\xi_i>0}) - \frac{1}{2} = 0.$$

Similarly, we have $E(\mathbf{1}_{\xi_i<0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4} = -(E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4}), \forall\, i,j \in S_0$. Meanwhile, the symmetry of ξ_{S_0} implies $E(\mathbf{1}_{\xi_i<0}\mathbf{1}_{\xi_j<0}) = E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0})$. Therefore, we obtain

$$
\begin{aligned}
\mathrm{Cov}(Y_i, Y_j|\eta) &= E(Y_iY_j|\eta) - E(Y_i|\eta)E(Y_j|\eta) = E(Y_iY_j|\eta) - \frac{1}{4} \\
&= E\big[\big(\mathbf{1}_{f_i(\eta)>0}\mathbf{1}_{\xi_i>0} + \mathbf{1}_{f_i(\eta)<0}\mathbf{1}_{\xi_i<0}\big)\cdot\big(\mathbf{1}_{f_j(\eta)>0}\mathbf{1}_{\xi_j>0} + \mathbf{1}_{f_j(\eta)<0}\mathbf{1}_{\xi_j<0}\big)|\,\eta\big] - \frac{1}{4} \\
&= \mathbf{1}_{f_i(\eta)>0}\mathbf{1}_{f_j(\eta)>0}\left[E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4}\right] + \mathbf{1}_{f_i(\eta)>0}\mathbf{1}_{f_j(\eta)<0}\left[E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j<0}) - \frac{1}{4}\right] \\
&\quad + \mathbf{1}_{f_i(\eta)<0}\mathbf{1}_{f_j(\eta)>0}\left[E(\mathbf{1}_{\xi_i<0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4}\right] + \mathbf{1}_{f_i(\eta)<0}\mathbf{1}_{f_j(\eta)<0}\left[E(\mathbf{1}_{\xi_i<0}\mathbf{1}_{\xi_j<0}) - \frac{1}{4}\right] \\
&= \left(E\big(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}\big) - \frac{1}{4}\right)\big(\mathbf{1}_{f_i(\eta)>0,f_j(\eta)>0} - \mathbf{1}_{f_i(\eta)>0,f_j(\eta)<0} - \mathbf{1}_{f_i(\eta)<0,f_j(\eta)>0} + \mathbf{1}_{f_i(\eta)<0,f_j(\eta)<0}\big) \\
&= \left(E\big(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}\big) - \frac{1}{4}\right)\big(\mathbf{1}_{f_i(\eta)>0} - \mathbf{1}_{f_i(\eta)<0}\big)\big(\mathbf{1}_{f_j(\eta)>0} - \mathbf{1}_{f_j(\eta)<0}\big) \\
&= \left(E\big(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}\big) - \frac{1}{4}\right) w_iw_j, \qquad \text{(B.1)}
\end{aligned}
$$

where $w_i \triangleq \mathbf{1}_{f_i(\eta)>0} - \mathbf{1}_{f_i(\eta)<0}$. By definition, $w_i = 1$ or -1. From $\mathrm{Cov}(\xi) = \mathbf{B} = 2\Sigma^{-1}$, we know that $\begin{pmatrix}\xi_i\\ \xi_j\end{pmatrix} \sim N\left(0, \begin{pmatrix}\mathbf{B}_{ii} & \mathbf{B}_{ij}\\ \mathbf{B}_{ji} & \mathbf{B}_{jj}\end{pmatrix}\right)$. Since normalizing ξ_i, ξ_j does not change their sign, we assume that $\begin{pmatrix}\xi_i\\ \xi_j\end{pmatrix} \sim N\left(0, \begin{pmatrix}1 & \mu_{ij}\\ \mu_{ij} & 1\end{pmatrix}\right)$, where $\mu_{ij} = \mathbf{B}_{ij}/\big(\mathbf{B}_{ii}^{1/2}\mathbf{B}_{jj}^{1/2}\big) = (\widetilde{\Sigma^{-1}})_{ij}$ (see (29)). Define $\mu = \mu_{ij} = (\widetilde{\Sigma^{-1}})_{ij}$ and let $P(\xi_i, \xi_j)$ and $P_s(\cdot)$ be the probability distribution function of (ξ_i, ξ_j) and the standard normal distribution, respectively. Using

$$
0 \le \mathrm{e}^x - 1 - x \le \frac{x^2}{2}\big(\mathrm{e}^x 1_{x>0} + 1\big), \quad x \triangleq -\frac{\mu^2\xi_i^2 + \mu^2\xi_j^2 - 2\mu\xi_i\xi_j}{2(1-\mu^2)},
$$

we expand $P(\xi_i, \xi_j) - P_s(\xi_i)P_s(\xi_j)$ up to μ^2

$$
\begin{aligned}
\big[P(\xi_i, \xi_j) - P_s(\xi_i)P_s(\xi_j)\big]w_iw_j &= \frac{P_s(\xi_i)P_s(\xi_j)}{\sqrt{1-\mu^2}}\left(1 + x - \sqrt{1-\mu^2} + \mathrm{e}^x - 1 - x\right)w_iw_j \\
&\le \frac{P_s(\xi_i)P_s(\xi_j)}{\sqrt{1-\mu^2}}\left(\left(1 + x - \sqrt{1-\mu^2}\right)w_iw_j + \mathrm{e}^x - 1 - x\right) \\
&\le \frac{P_s(\xi_i)P_s(\xi_j)}{\sqrt{1-\mu^2}}\left(\left(1 + x - \sqrt{1-\mu^2}\right)w_iw_j + \frac{x^2(\mathrm{e}^x\mathbf{1}_{x>0} + 1)}{2}\right) \\
&= \frac{P_s(\xi_i)P_s(\xi_j)}{\sqrt{1-\mu^2}}\left(\left(1 + x - \sqrt{1-\mu^2}\right)w_iw_j + \frac{x^2}{2}\right) + P(\xi_i, \xi_j)\frac{x^2\mathbf{1}_{x>0}}{2}.
\end{aligned}
$$

Integrating both sides with respect to ξ_i, ξ_j in the region $\xi_i, \xi_j > 0$ gives

$$\left(E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4}\right) w_i w_j \le \int_{\xi_i,\xi_j>0} \frac{P_s(\xi_i)P_s(\xi_j)}{\sqrt{1-\mu^2}} \left(\left(1 + x - \sqrt{1-\mu^2}\right) w_i w_j + \frac{x^2}{2}\right) \mathrm{d}\xi_i \, \mathrm{d}\xi_j$$
$$+ \int_{\xi_i>0,\xi_j>0} P(\xi_i, \xi_j) \frac{x^2 \mathbf{1}_{x>0}}{2} \, \mathrm{d}\xi_i \, \mathrm{d}\xi_j \triangleq \mathbf{I} + \mathbf{II} + \mathbf{III}. \tag{B.2}$$

Since $P_s(\cdot)$ is a standard Gaussian distribution and $x = -\frac{\mu^2\xi_i^2 + \mu^2\xi_j^2 - 2\mu\xi_i\xi_j}{2(1-\mu^2)}$, we can calculate all the moments in **I** and **II** explicitly. For **I**, we have

$$\begin{aligned}
\mathbf{I} &= \left(\frac{1}{4}\left(\frac{1}{\sqrt{1-\mu^2}} - 1\right) + \frac{\mu}{2\pi(1-\mu^2)^{3/2}} - \frac{1}{4}\frac{\mu^2}{(1-\mu^2)^{3/2}}\right) w_i w_j \\
&\le \frac{\mu w_i w_j}{2\pi} + \left|\frac{\mu}{2\pi}\left((1-\mu^2)^{-3/2} - 1\right)\right| + \left|\frac{1}{4}\left(\frac{1}{\sqrt{1-\mu^2}} - 1\right) - \frac{1}{4}\frac{\mu^2}{(1-\mu^2)^{3/2}}\right| \\
&= \frac{\mu w_i w_j}{2\pi} + \frac{\mu^2}{2\pi}\left|\frac{(1-\mu^2)^{-3/2} - 1}{\mu}\right| + \frac{\mu^2}{4\sqrt{1-\mu^2}}\left|\frac{1}{1-\mu^2} - \frac{1}{1+\sqrt{1-\mu^2}}\right| \triangleq \frac{\mu w_i w_j}{2\pi} + c_1(\mu)\mu^2,
\end{aligned} \tag{B.3}$$

where $c_1(\mu) \ge 0$ collects the coefficients of μ^2 and is bounded near $\mu = 0$. We use $E(\xi \mathbf{1}_{\xi>0}) = 1/\sqrt{2\pi}, E(\xi^2 \mathbf{1}_{\xi>0}) = \frac{1}{2}$ for the standard Gaussian ξ to obtain the first equality, and $|w_i| = |w_j| = 1$ to obtain the inequality. For the second term, we get

$$\mathbf{II} = \frac{\mu^2}{8(1-\mu^2)^{5/2}} \int_{\xi_i,\xi_j>0} P_s(\xi_i)P_s(\xi_j)\left(2\xi_i\xi_j - \mu\left(\xi_i^2 + \xi_j^2\right)\right)^2 \mathrm{d}\xi_i \, \mathrm{d}\xi_j = \frac{\mu^2\left(1 - \frac{8}{\pi}\mu + 2\mu^2\right)}{8(1-\mu^2)^{5/2}} \triangleq c_2(\mu)\mu^2, \tag{B.4}$$

where $c_2(\mu) = \frac{(1 - \frac{8}{\pi}\mu + 2\mu^2)}{8(1-\mu^2)^{5/2}} \ge 0$ is bounded near $\mu = 0$. Since $\xi_i, \xi_j > 0$ and $x = -\frac{\mu^2\xi_i^2 + \mu^2\xi_j^2 - 2\mu\xi_i\xi_j}{2(1-\mu^2)}$, $\mu \le 0$ implies $x \le 0$, or equivalently $\mathbf{1}_{x>0} \le \mathbf{1}_{\mu>0}$. Note that

$$x = -\frac{\mu^2\xi_i^2 + \mu^2\xi_j^2 - 2\mu\xi_i\xi_j}{2(1-\mu^2)} \le -\frac{2\mu^2\xi_i\xi_j - 2|\mu|\xi_i\xi_j}{2(1-\mu^2)} = \frac{|\mu|\xi_i\xi_j}{\left(1 + |\mu|\right)}, \ \forall\, \xi_i, \xi_j > 0.$$

For $\xi_i, \xi_j > 0$, we have $x^2 \mathbf{1}_{x>0} \le \left(\frac{|\mu|\xi_i\xi_j}{1+|\mu|}\right)^2 \mathbf{1}_{\mu>0}$. Therefore, we obtain

$$\begin{aligned}
\mathbf{III} &= \frac{1}{2}E\left(x^2 \mathbf{1}_{x>0}\mathbf{1}_{\xi_i,\xi_j>0}\right) \le \frac{1}{2}\frac{|\mu^2\mathbf{1}_{\mu>0}|}{(1+|\mu|)^2}E\left(\xi_i^2\xi_j^2\mathbf{1}_{\xi_i,\xi_j>0}\right) \\
&\le \frac{1}{2}\frac{|\mu^2\mathbf{1}_{\mu>0}|}{(1+|\mu|)^2}\left(E\left(\xi_i^4\mathbf{1}_{\xi_i>0}\right)E\left(\xi_j^4\mathbf{1}_{\xi_j>0}\right)\right)^{1/2} = \frac{1}{2}\frac{|\mu^2\mathbf{1}_{\mu>0}|}{\left(1+|\mu|\right)^2}\frac{3}{2} \triangleq \mathbf{1}_{\mu>0}c_3(\mu)\mu^2,
\end{aligned} \tag{B.5}$$

where $c_3(\mu) = \frac{3}{4(1+|\mu|)^2}$ is bounded near $\mu = 0$. Combining (B.2), (B.3), (B.4) and (B.5), we yield

$$\begin{aligned}\mathrm{Cov}(Y_i, Y_j|\eta) &= \left[E(\mathbf{1}_{\xi_i>0}\mathbf{1}_{\xi_j>0}) - \frac{1}{4}\right] w_i w_j \le \frac{\mu}{2\pi} w_i w_j + (c_1(\mu) + c_2(\mu) + c_3(\mu)\mathbf{1}_{\mu>0})\mu^2 \\ &\triangleq \frac{\mu}{2\pi} w_i w_j + c(\mu)\mu^2.\end{aligned}$$

Here, $c(\mu) = c_1(\mu) + c_2(\mu) + c_3(\mu)\mathbf{1}_{\mu>0}$. Since $c_i(\mu)$ is a non-negative and an explicit function of μ, it is not difficult to show that $c(\mu) < \frac{3}{2}$ for $|\mu| < \frac{1}{2}$. For $|\mu| > 1/2$, we use the estimate $\mathrm{Cov}(Y_i, Y_j|\eta) \le 1/4 \le -\frac{\mu}{2\pi} + \frac{3}{2}\mu^2$. Finally, we conclude

$$\mathrm{Cov}(Y_i, Y_j|\eta) \le \left(\frac{\mu}{2\pi} w_i w_j + c(\mu)\mu^2\right) \wedge \frac{1}{4} \le \frac{\mu}{2\pi} w_i w_j + \frac{3}{2}\mu^2,$$

where $\mu = (\widetilde{\Sigma^{-1}})_{ij}$. This proves Lemma 4.4.