Nodewise Knockoffs: False Discovery Rate Control for Gaussian Graphical Models

Jinzhou Li
Seminar für Statistik, ETH Zürich, Switzerland
and
Marloes H. Maathuis
Seminar für Statistik, ETH Zürich, Switzerland
September 2, 2019

Abstract

Controlling the false discovery rate (FDR) is important for obtaining reliable and reproducible conclusions in scientific research. This paper considers the problem of controlling the finite sample FDR in learning the structure of a Gaussian graphical model (GGM). Most state-of-the-art structure learning methods do not ensure the FDR control, and those that do are all based on p-values and multiple testing procedures. In this paper, we tackle this problem from a different angle by using the recently proposed knockoff idea of Barber and Candès. Our approach consists of two steps: (a) constructing knockoffs and feature statistics nodewisely; (b) applying a graph-wide rule in choosing the thresholds for each node and then recovering the structure of the graph. The finite sample FDR control property of this approach is shown. In addition, we use a sample-splitting-recycling procedure that first uses half of the sample to select hyperparameters, then learns the structure of the graph using all samples in a certain way such that the FDR control property still holds. Finally, we examine our methodology using simulations and a real data set.

Keywords: knockoffs, false discovery rate (FDR), Gaussian graphical model (GGM), sample splitting, sample recycling, structure learning.
1 Introduction

Gaussian graphical models (GGMs) are popular tools used to explore the conditional independence relationships among variables in many scientific fields such as biology, economics and social science. Formally, for a multivariate Gaussian random vector $x = (x_1, \ldots, x_p)^T$ with mean $\mu$ and covariance matrix $\Sigma$, we can assign it an undirected graph $G = (V, E)$, where the node set $V = [p] = \{1, \ldots, p\}$ and the edge set $E = \{(i, j) \in [p]^2 : x_i \not\perp x_j | x_{[p]\setminus\{i,j\}}\}$ and $i \neq j$. Here, $x_{[p]\setminus\{i,j\}}$ stands for the vector $x$ after removing $x_i$ and $x_j$, and $x_i \not\perp x_j | x_{[p]\setminus\{i,j\}}$ indicates that $x_i$ and $x_j$ are dependent conditional on $x_{[p]\setminus\{i,j\}}$.

Structure learning in GGM aims to obtain an estimator $\hat{E}$ of the true edge set $E$ from $n$ i.i.d. observations $x_1, \ldots, x_n$. To obtain a reliable outcome and alleviate the reproducibility issue (see, for example, Ioannidis (2005), Begley and Ellis (2012) and Baker (2016)) in scientific research, it is crucial to develop statistical methods with confidence statements about false discoveries. In this paper, we propose a structure learning approach that guarantees the finite sample false discovery rate (FDR, see Benjamini and Hochberg (1995)) control when $n \geq p$. That is, our estimated edge set $\hat{E}$ satisfies

$$\text{FDR} = \mathbb{E}\left[\frac{|\hat{F}|}{|\hat{E}| \lor 1}\right] \leq q,$$

where $q \in [0, 1]$ is a preset nominal FDR level, $\hat{F} = \hat{E}\setminus E$ is the set of the falsely discovered edges, $|\cdot|$ calculates the number of the edges for a given edge set, and $|\hat{E}| \lor 1 = \max(|\hat{E}|, 1)$.

Let $\Omega = \Sigma^{-1}$ be the precision matrix of $x$. It is well-known (see, for example, Lauritzen (1996)) that for any pair $(i, j) \in [p]^2$ and $i \neq j$,

$$(i, j) \notin E \iff x_i \perp x_j | x_{[p]\setminus\{i,j\}} \iff \Omega_{i,j} = 0 \iff \rho_{ij\|\{i,j\}} = 0 \iff \beta_{ij} = \beta_{ji} = 0,$$

where $\rho_{ij\|\{i,j\}}$ is the partial correlation between $x_i$ and $x_j$ given $x_{[p]\setminus\{i,j\}}$, and $\beta_{ij}$ is the regression
coefficient of $x_j$ in the regression of $x_i$ on $x_{[p]\{i\}}$, i.e.,

$$x_i = \alpha_i + \sum_{j \neq i} \beta_{ji}^i x_j + z_i,$$

where

$$\begin{cases}
\alpha_i = \mu_i - \Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}\mu_{-i}, \\
\beta_{ji}^i = -\Omega_{ji}/\Omega_{ii}, \\
z_i \sim N(0, \Sigma_{ii} - \Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}\Sigma_{-i,i}) \text{ and } z_i \perp x_{-i}.
\end{cases}$$

Here, $\mu_{-i}$ is the vector $\mu$ after deleting its $i$th element $\mu_i$, $\Sigma_{i,-i}$ is the $i$th row of $\Sigma$ without $i$th element, $\Sigma_{i,i}$, $\Sigma_{-i,-i}$ is the matrix $\Sigma$ after removing its $i$th row and $i$th column, and other notations are similarly defined.

Previous research on GGM structure learning problem has mainly focused on developing applicable methods and studying their theoretical properties in high-dimensional settings, and various approaches have been proposed based on one of the equivalences in (2). For example, Meinshausen and B"uhlmann (2006) made use of relationship (3) to recover the underlying graph nodewise by using LASSO. Yuan and Lin (2007), Friedman et al. (2008) and d'Aspremont et al. (2008) proposed different algorithms to estimate the graph by minimizing the negative log-likelihood plus an $l_1$-penalty on the entries of the precision matrix, and Rothman et al. (2008) and Raskutti et al. (2009) studied the corresponding theoretical properties. Fan et al. (2009) suggested using the nonconvex Smoothly Clipped Absolute Deviation (SCAD) penalty instead of the $l_1$-penalty in the penalized likelihood approach, and Cai et al. (2011) proposed minimizing the $l_1$-norm of the precision matrix, rather than the negative log-likelihood, under certain constraints. However, none of these methods controls the FDR of the estimated graph.

A different approach to learn the graph structure relies on multiple testing. In low-dimensional cases, Drton and Perlman (2007) suggested testing $\rho_{(i\|p\setminus{i})} = 0$ for all $i, j \in [p]$ by using sample partial correlations. After obtaining the corresponding p-values, the Benjamini-Yekutieli (BY)
procedure (see Benjamini and Yekutieli (2001)) can be applied to recover the graph structure with the FDR control. The Benjamini-Hochberg (BH) procedure (see Benjamini and Hochberg (1995)) is also commonly used to control the FDR based on p-values while it requires the so-called Positive Regression Dependency on each one from a Subset (PRDS) assumption (see Benjamini and Yekutieli (2001)). In high-dimensional settings, Liu et al. (2013) proposed a test statistic and derived its asymptotic distribution. A structure learning procedure was then developed with the asymptotic FDR control under some regularity conditions.

Recently, Barber and Candès (2015) developed an interesting and innovative framework called fixed-X knockoffs. This framework was originally designed for low-dimensional Gaussian linear regression models with fixed design, and it achieves the finite sample FDR control in variable selection without resorting to hypothesis testings and p-values. The underlying idea is to construct artificial variables (i.e. fixed-X knockoffs) that mimic the correlation structure of the original predictors, then use these knockoffs as controls in the procedure that follows, ensuring the FDR control. Various research has been done based on the fixed-X knockoffs framework: Dai and Barber (2016) generalized it to group-variable selection and multitask learning; Janson et al. (2016) adapted it to control the k-familywise error rate; Barber and Candès (2016) extended it to high-dimensional linear models through a sample-splitting-recycling procedure and developed novel theory showing that the directional FDR is controlled; Katsevich et al. (2019) proposed new procedure building on it for multilayer controlled variable selection tasks and shown that this procedure controls the FDR at both individual-variable and group-variable levels; and Weinstein et al. (2017) studied its theoretical power under an i.i.d. Gaussian design with Lasso statistics.

Another closely related framework is the so-called model-X knockoffs proposed by Candès et al. (2018). Unlike fixed-X knockoffs which are constructed algebraically, model-X knockoffs are generated in a probabilistic manner, when the joint distribution of the predictors is assumed completely known. Beyond low-dimensional linear settings, the model-X knockoff procedure en-
sures the finite sample FDR control for high-dimensional and non-linear cases. One of the most important issues in this framework is to find an efficient way to generate knockoffs. To achieve such efficiency, Sesia et al. (2018), Gimenez et al. (2018) and Bates et al. (2019) proposed different knockoff generation methods when the predictors follow hidden Markov models, Bayesian networks and graphical models, respectively. Jordon et al. (2018), Romano et al. (2018) and Liu and Zheng (2018) considers the problem of generating approximate knockoffs when the distribution of predictors is unknown. In the theoretical part, Barber et al. (2018) and Fan et al. (2019) studied the robustness of the model-X knockoff procedure, and the second paper also discussed the power of this method when the distribution of the predictors is characterized by a Gaussian graphical model. Beyond variable selection, Zheng et al. (2018) proposed recovering the structure of a GGM by using the knockoff idea. However, their approach aims to control only the FDR of the estimated neighborhood of each node, rather than of the estimated graph. Therefore, to the best of our knowledge, the problem of generalizing the knockoff framework from variable selection settings to the more involved structure learning tasks and achieving (1) is still open.

In this paper, we proposed the so-called nodewise knockoff procedure by building on the knockoff framework introduced by Barber and Candès (2015). The nodewise knockoff method proceeds in two steps: (a) treating each variable as response and constructing knockoffs and feature statistics based on the linear model (3); (b) applying a graph-wide rule to choose the thresholds for each node, then estimating the neighborhoods and recovering the structure of the graph. We proved that this approach controls the finite sample FDR of the estimated graph.

As shown in Barber and Candès (2015) and Candes et al. (2018), the knockoff framework is fairly flexible as there are many different choices to construct knockoffs and feature statistics. Hence essentially, there are many FDR control procedures in our hands. Although all of these procedures control the FDR, their statistical powers are different, and it is possible that the most powerful procedure in one setting is powerless in another setting. Therefore, a natural and prac-
tical question arises: which FDR control procedure should we choose for a given problem? We would like to emphasize that such issue is not restricted to our structure learning tasks, and we believe that all (fixed-X and model-X) knockoff based approaches face the same problem. For structure learning, one extra thing we will need to decide is the rule used to recover the graph from estimated neighborhoods. In this paper, we used the sample-splitting-recycling approach as in [Barber and Candès (2016)]. The idea is to split the sample into two halves, and select a graph recovering rule and procedures used to construct knockoffs and feature statistics in the first half of sample. Then, based on the selected rule and procedures, we implement the nodewise knockoff method using all samples in a certain manner. Specifically, the knockoffs of the first half of sample is simply identical to the original sample. Compared with the sample-splitting procedure, sample-splitting-recycling gains more statistical power while the finite sample FDR control property still holds.

The remainder of this paper is organized as follows. In Section 2, we give a brief review of the fixed-X knockoffs framework in linear regression models. In Section 3, we introduce the nodewise knockoff approach and the sample-splitting-recycling procedure, and prove the finite sample FDR control guarantee. Numerical performance of the proposed procedure is evaluated by simulation studies and a real data example in Section 4.

2 Preliminaries of the fixed-X knockoffs framework

The fixed-X knockoff procedure is tailored to Gaussian linear regression models in which the number of observations $n$ is at least as large as the number of variables $p$. Formally, consider the linear model

$$y = X\beta + z,$$
where \( y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n, \) \( X = [x^{(1)}, \ldots, x^{(p)}] \in \mathbb{R}^{n \times p} \) is a fixed and known design matrix with \( x^{(i)} = (x^{(i)}_1, \ldots, x^{(i)}_n)^T \) and \( x^{(i)^T} x^{(i)} = 1, \) for \( i = 1, \ldots, p, \) \( z = (z_1, \ldots, z_n)^T \in \mathbb{R}^n \) is a noise vector with multivariate Gaussian distribution \( N_n(0, \sigma^2 I) \), and we assume that \( (X^T X)^{-1} \) exists.

We denote the index set of the non-null variables by \( S = \{ i \in [p] : \beta_i \neq 0 \} \) and the index set of the null variables by \( N_0 = [p] \backslash S \). The goal is to estimate \( S \) with the FDR control guarantee.

In variable selection, one of the reasons for a null variable to be wrongly selected is that the selection procedure might be deceived by the correlations between this null variable and other non-null variables. Therefore, for each variable, knockoff procedure considers to create an artificial variable (i.e. knockoff) which possesses the same correlation relationships with other variables as its original counterpart, then to use this knockoff as negative control. We summarize the knockoff procedure at a high level in Algorithm [1] then give more details and discussions.

In the Step 1 of Algorithm [1], we construct knockoffs. The knockoff matrix \( \tilde{X} \) is constructed to mimic the correlation structure of the original variables. Formally,

\[
[X \tilde{X}]^T [X \tilde{X}] = \begin{pmatrix} X^T X & X^T \tilde{X} \\ \tilde{X}^T X & \tilde{X}^T \tilde{X} \end{pmatrix} = \begin{pmatrix} X^T X & X^T X - \text{diag}(s) \\ X^T X - \text{diag}(s) & X^T X \end{pmatrix}, \tag{6}
\]

where \( s \in \mathbb{R}^p_\geq 0 \) and \( \text{diag}(s) \) denotes the diagonal matrix with \( s \) on the diagonals.

By solving (6), we can obtain the formula (4) in which \( \tilde{U} \in \mathbb{R}^{n \times p} \) has column space orthogonal to that of \( X \) and can be obtained when \( n \geq 2p \), \( C \in \mathbb{R}^{p \times p} \) satisfies \( C^T C = 2\text{diag}(s) - \text{diag}(s)(X^T X)^{-1}\text{diag}(s) \) and can be obtained by Cholesky-Decomposition. As for \( s \), Barber and Candès [2015] suggested two approaches, “Equi” and “SDP”, to compute it. “SDP” calculates \( s \)

\footnotetext[1]{In the paper of Barber and Candès [2015], knockoff refers to the procedure controlling mFDR and knockoff+ refers to the procedure controlling FDR. In this paper, we will use knockoff rather than knockoff+ to indicate procedures controlling FDR.}

\footnotetext[2]{In the case of \( p \leq n < 2p \), problems arise because such \( \tilde{U} \) does not exist. One remedy for this is to artificially enlarge the design matrix, see Barber and Candès [2015].}
Algorithm 1: Knockoff procedure (Barber and Candès (2015))

**Input**: \((X, y, q, O', \mathcal{P})\), where \(X \in \mathbb{R}^{n \times p}\) is the design matrix, \(y \in \mathbb{R}^p\) is the response vector, \(q \in [0, 1]\) is the nominal FDR level, \(O' \in \{\text{Equi (8)}, \text{SDP (7)}\}\) is the optimization strategy used to compute \(s\), and \(\mathcal{P}\) is the procedure used to construct feature statistics.

**Output**: Estimated set of the non-null variables \(\hat{S}\).

1. **Step 1.** Construct the knockoff matrix \(\tilde{X} = [\tilde{x}^{(1)}, \cdots, \tilde{x}^{(p)}] \in \mathbb{R}^{n \times p}\) by
   \[
   \tilde{X}(X, O') = X \left( I - (X^T X)^{-1} \text{diag}(s) \right) + UC. \tag{4}
   \]

2. **Step 2.** Create feature statistics \(W(X, \tilde{X}, y, \mathcal{P}) = \mathcal{P}(X, \tilde{X}, y) = (W_1, \cdots, W_p)\).

3. **Step 3.** Calculate the threshold \(\hat{T}\) by
   \[
   \hat{T}(W, q) = \min \left\{ t \in \{|W_i| : i = 1, \cdots, p\} \setminus \{0\} : \frac{1 + \# \{ i \in [p] : W_i \leq -t \}}{\# \{ i \in [p] : W_i \geq t \}} \geq 1 \right\} > 0 \tag{5}
   \]
   and set \(\hat{T} = +\infty\) if the above set is empty. Finally, estimate the set of the non-null variables by \(\hat{S} = \{ i \in [p] : W_i \geq \hat{T} \}\).

by solving the following convex optimization problem:
\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{p} |1 - s_i| \\
\text{subject to} & \quad s \geq 0, \\
& \quad 2X^T X - \text{diag}(s) \geq 0.
\end{align*}
\tag{7}
\]

The above constraints are necessary conditions for the existence of \(\tilde{X}\) in equation (6), and this optimization problem is a semidefinite program (SDP) which can be solved efficiently. “Equi” adds an extra restriction that all \(s_i\) are the same to reduce computational expense, and leads to an analytic solution:
\[
\begin{align*}
\hat{s}_i &= \min(2\lambda_{\min}(X^T X), 1), \quad \forall i = 1, \cdots, p, \\
\end{align*}
\tag{8}
\]
where $\lambda_{\min}(X^T X)$ denotes the minimal eigenvalue of $X^T X$. We point out that there is no dominance relationship between “SDP” and “Equi” with respect to the statistical power. From our experience, “Equi” is computationally more efficient but is generally less powerful than “SDP”.

In the Step 2 of Algorithm 1 we construct for each predictor $X_i$ a feature statistic $W_i$ which possesses the so-called sign-flip property defined as follows.

**Definition 2.1. (Sign-flip property on O, Barber and Candès (2015))**

Let $W = (W_1, \cdots, W_p)$ be a random vector and $O \subseteq [p]$ be an index set. Let $\epsilon = (\epsilon_1, \cdots, \epsilon_p)$ be a sign sequence independent of $W$ with $\epsilon_i = +1$ for $i \notin O$ and $\epsilon_i$ i.i.d. from a Rademacher distribution for $i \in O$. Then we say that $W$ possesses sign-flip property on $O$ if

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

As shown in Algorithm 1, the construction of $W$ depends on some procedure $\mathcal{P}$ and is very flexible. The only two conditions are:

1. **Sufficiency property**: $W$ depends only on the Gram matrix $[X \bar{X}]^T [X \bar{X}]$ and the feature-response inner product $[X \bar{X}]^T y$. That is, $W = \mathcal{P}([X \bar{X}]^T [X \bar{X}], [X \bar{X}]^T y)$.

2. **Antisymmetry property**: Swapping $x_i$ and $\bar{x}_i$ causes a sign switch of $W_i$. That is, for all $O \subseteq \{1, \cdots, p\}$,

$$W_i([X \bar{X}]_{\text{swap}(O)}, y) = \begin{cases} W_i([X \bar{X}], y), & i \notin O, \\ -W_i([X \bar{X}], y), & i \in O, \end{cases}$$

where $[X \bar{X}]_{\text{swap}(O)}$ means that we swap the columns $x^{(i)}$ and $\bar{x}^{(i)}$ if $i \in O$.

Now we describe the concrete steps to construct feature statistics $W$. It is helpful to think that we first compute statistics $Z_i$ and $\bar{Z}_i$ measuring the importance of $X_i$ and $\bar{X}_i$ to the response $y$. 

9
respectively, and then construct \( W_i \) based on \( Z_i \) and \( \tilde{Z}_i \). One example to construct feature statistics is based on the elastic net regularization (Zou and Hastie (2005))

\[
\hat{\beta}(\lambda, \alpha) = \arg\min_{\beta \in \mathbb{R}^p} \left( \frac{1}{2}\|y - [X \ X] \beta\|^2_2 + \lambda \left[ (1 - \alpha)\|\beta\|^2_2 / 2 + \alpha\|\beta\|_1 \right] \right), \quad \alpha \in [0, 1], \quad \lambda \geq 0. \tag{9}
\]

For a preset \( \lambda \geq 0 \), we can use

\[
Z_i = \sup\{\lambda \geq 0 : \hat{\beta}_i(\lambda, \alpha) \neq 0\} \quad \text{and} \quad \tilde{Z}_i = \sup\{\lambda \geq 0 : \hat{\beta}_{i+p}(\lambda, \alpha) \neq 0\}
\]

to measure the importance of \( X_i \) and \( \tilde{X}_i \) for the response \( y \). With an extra preset tuning parameter \( \lambda \geq 0 \), we can also use the regression coefficients and let

\[
Z_i = \hat{\beta}_i(\lambda, \alpha) \quad \text{and} \quad \tilde{Z}_i = \hat{\beta}_{i+p}(\lambda, \alpha).
\]

Then, \( W_i \) can be obtained by

\[
W_i = (Z_i \lor \tilde{Z}_i) \cdot \text{sign}(Z_i - \tilde{Z}_i) \quad \text{or} \quad W_i = Z_i - \tilde{Z}_i.
\]

It can be verified that the above constructed \( W \) satisfies the required two conditions which are sufficient for \( W \) to possess the sign-flip property as the following lemma shows.

**Lemma 2.1.** *(Sign-flip property on \( N_0 \) with fixed design matrix, Barber and Candès (2015))*

Let \( X \in \mathbb{R}^{n \times p} \) be a fixed design matrix in a liner model, \( N_0 \) be the index set of null variables and \( W \) be a feature statistic vector satisfying sufficiency and antisymmetry properties. Then \( W \) possesses the sign-flip property on \( N_0 \).

\[^3\text{In order for the feature statistics to satisfy the antisymmetry and sufficiency properties, cross-validation cannot be used to choose tuning parameters } \alpha \text{ and } \lambda.\]

\[^4\text{In Candes et al. (2018), the feature statistics can possess the sign-flip property without satisfying the sufficiency property due to a different and more demanding way to construct knockoffs. In this paper, the term feature statistics always denote those satisfying both sufficiency and antisymmetry properties.}\]
From the proof of Lemma 1 in [Barber and Candès (2015)], it is easy to show that sign-flip property also holds for random design matrix which is the case we are facing with GGM, thus we write it here as a lemma for clearness.

**Lemma 2.2. (Sign-flip property on $N_0$ with random design matrix)**

Let $X \in \mathbb{R}^{n \times p}$ be a random design matrix in a liner model, $N_0$ be the index set of null variables and $W$ be a feature statistic vector satisfying sufficiency and antisymmetry properties. Then $W$ possesses the sign-flip property on $N_0$.

In the Step 3 of Algorithm [1] we obtain the data-dependent threshold $\hat{T}$ based on $W$ and $q$. By our construction, $W_i$ possesses an interpretation that a positive value of $W_i$ means that the variable selection method would choose $x_i$ before $\tilde{x}_i$, and a negative value of $W_i$ indicates that $\tilde{x}_i$ would be chosen prior to $x_i$. As we already know that $\tilde{x}_i$ is a null variable, we will never choose the variable with a negative $W$. Hence we select variables by $\tilde{S} = \{i \in [p] : W_i \geq \hat{T}\}$. Based on the sign-flip property of $W$ on $N_0$, [Barber and Candès (2015)] proved the finite sample FDR control guarantee

$$FDR = \mathbb{E} \left[ \frac{\# \{i \in N_0 : W_i \geq \hat{T}\}}{\# \{i \in [p] : W_i \geq \hat{T}\} \lor 1} \right] \leq q.$$ 

**3 Nodewise knockoff method and sample-splitting-recycling**

In this section, we present our main procedure for learning the structure of a GGM and show that it controls the finite sample FDR. A sample-splitting-recycling approach originally proposed by [Barber and Candès (2016)] is also introduced to choose data-driven graph recovering rule and feature statistics construction procedure. First, we introduce some notation and discuss two naive ideas.

Let $X = [x_1, \cdots, x_n]^T = [x^{(1)}, \cdots, x^{(p)}] \in \mathbb{R}^{n \times p}$ be the data matrix with $n$ i.i.d. observations $x_1, \cdots, x_n$ from a $p$-dimensional Gaussian distribution $N_p(0, \Omega^{-1})$ (we assume that the mean
\( \mu = 0 \) without loss of generality, and let \( G = (V,E) \) be the corresponding undirected graph. Denote the matrix obtained after deleting the \( i \)th column of \( X \) by \( X^{(-i)} \in \mathbb{R}^{n \times (p-1)} \), then by equation (3) we have

\[
x^{(i)} = \sum_{j \neq i} \beta_j^{(i)} x^{(j)} + z^{(i)} = X^{(-i)} \beta^{(i)} + z^{(i)}, \quad i = 1, \cdots, p, \tag{10}
\]

where \( \beta_j^{(i)} = -\Omega_{ji}/\Omega_{ii}, \beta^{(i)} \in \mathbb{R}^{p-1} \) and \( z^{(i)} \in \mathbb{R}^n \).

To estimate the graph structure, one naive idea is to directly apply the fixed-X knockoff procedure (Algorithm 1) for each node \( i \in \{1, \cdots, p\} \) at some nominal FDR level \( q \in [0, 1] \). One then obtains the estimated neighborhood \( \hat{\text{ne}}_i \) of node \( i \), where the true neighborhood is defined as \( \text{ne}_i = \{ j \in [p] \setminus \{i\} : (i, j) \in E \} = \{ j \in [p] \setminus \{i\} : \beta_{ji} \neq 0 \} \). Then, by either implementing the AND rule or the OR rule displayed in (11), the estimated graph can be obtained:

\[
\text{AND rule: } \hat{E}_{\text{AND}} = \{ (i, j) \in [p]^2 : i \in \hat{\text{ne}}_j \text{ and } j \in \hat{\text{ne}}_i \},
\]

\[
\text{OR rule: } \hat{E}_{\text{OR}} = \{ (i, j) \in [p]^2 : i \in \hat{\text{ne}}_j \text{ or } j \in \hat{\text{ne}}_i \}. \tag{11}
\]

However, it is not hard to see from simulations that \( \hat{\text{ne}}_i \)'s, \( i = 1, \cdots, p, \) with the FDR controlled at level \( q \), does not necessary result in an estimated edge set \( \hat{E} \) with the FDR controlled at level \( q \), independent of whether the AND rule or the OR rule is applied.

Another approach is to use knockoff procedure with nominal FDR level \( q/p \) to obtain \( \hat{\text{ne}}_i \), and then apply the OR rule to get the estimated edge set. This approach indeed controls the FDR for \( \hat{E}_{\text{OR}} \) at level \( q \). To see this, let \( \hat{F}_i = \hat{\text{ne}}_i \setminus \text{ne}_i \) be the falsely discovered neighbors for node \( i \) and \( \hat{F}_{\text{OR}} = \hat{E}_{\text{OR}} \setminus E \) be the set of the falsely discovered edges. Then because \( |\hat{F}_{\text{OR}}| \leq \sum_{i=1}^p |\hat{F}_i| \) and \( |\hat{E}_{\text{OR}}| \geq |\hat{\text{ne}}_i| \), we have

\[
\text{FDR} = \mathbb{E} \left[ \frac{|\hat{F}_{\text{OR}}|}{|\hat{E}_{\text{OR}}| \lor 1} \right] \leq \sum_{i=1}^p \mathbb{E} \left[ \frac{|\hat{F}_i|}{|\hat{\text{ne}}_i| \lor 1} \right] \leq \sum_{i=1}^p \frac{q}{p} = q,
\]

where the last inequality holds due to our choice of the nominal FDR control level of the knockoff procedure. However, we will now show that this approach is powerless and never makes a
discovery. For \( i \in \{1, \cdots, p\} \), let \( W_{1}^{(i)}, \cdots, W_{p}^{(i)} \) be the feature statistics corresponding to linear model \((10)\). Note that for any \( t > 0 \),

\[
\frac{1 + \# \{ j \in [p] \setminus \{i\} : W_{j}^{(i)} \leq -t \} \lor 1}{\# \{ j \in [p] \setminus \{i\} : W_{j}^{(i)} \geq t \} \lor 1} \geq \frac{1}{p} \geq \frac{q}{p}, \quad \forall q \in [0, 1],
\]

thus the threshold \( \hat{T}_{i} \) calculated by the Step 3 of Algorithm 1 will always be \( \hat{T}_{i} = +\infty \). Hence \( \hat{\mathbf{e}}_{i} = \emptyset \) for all \( i = 1, \cdots, p \) and \( \hat{E}_{OR} = \emptyset \).

The above discussion shows that the relationship between nodewise and graph-wide FDR control is rather unclear. Although there is one procedure that controls the graph-wide FDR, it requires applying an excessively conservative nominal FDR level to each node, making it powerless. Thus, it is not immediately clear about how to get a non-powerless structure learning procedure that controls the graph-wide FDR.

We circumvent this problem by not directly setting nodewise nominal FDR levels, instead we set a graph-wide one over the entire feature statistic matrix \( \mathbf{W} \) obtained by treating every variable as response. Our approach involves two main steps: a nodewise construction of knockoffs and feature statistics, and a global filtering step to estimate the graph. We assume \( n \geq 2p \) throughout this paper, but this method still applicable when \( p < n < 2p \) as the original fixed-X knockoff method, see Barber and Candès (2015) for details.

### 3.1 Nodewise knockoff method

#### 3.1.1 Construct knockoffs and feature statistic matrix nodewisely

The first step of our approach is to construct knockoffs and feature statistics nodewisely by treating each \( x_{i} \) as response and the remaining variables as predictors. Specifically, based on the linear relationship \((10)\) and formular \((4)\), we construct the knockoffs by

\[
\bar{\mathbf{X}}^{(-i)} = \mathbf{X}^{(-i)}(I - (\mathbf{X}^{(-i)\top} \mathbf{X}^{(-i)})^{-1} \text{diag}(s_{i})) + \bar{\mathbf{U}}_{i} \mathbf{C}_{i}, \quad i = 1, \cdots, p,
\]

13
with $s_i$ calculated by using the pre-decided optimization strategy “Equi” (8) or “SDP” (7).

Then, for each $i = 1, \cdots, p$, we can construct the feature statistics by treating $x^{(i)}$ as the response vector and $[X^{(-i)} X^{(-i)}]$ as the design matrix. For example, based on the elastic net regularization (9), we can use

\begin{equation}
\begin{aligned}
Z^{(i)}_j &= \sup_{\lambda \geq 0} \{ \lambda : \hat{\beta}^{(i)}_j (\lambda, \alpha) \neq 0 \} \\
\tilde{Z}^{(i)}_j &= \sup_{\lambda \geq 0} \{ \lambda : \hat{\beta}^{(i)}_{j+p} (\lambda, \alpha) \neq 0 \}
\end{aligned}
\end{equation}

with a preset $\alpha \in [0, 1]$ or

\begin{equation}
\begin{aligned}
Z^{(i)}_j &= \hat{\beta}^{(i)}_j (\lambda, \alpha) \\
\tilde{Z}^{(i)}_j &= \hat{\beta}^{(i)}_{j+p} (\lambda, \alpha).
\end{aligned}
\end{equation}

with preset $\alpha \in [0, 1]$ and $\lambda \geq 0$, then construct the feature statistics by

\begin{align*}
W^{(i)}_j = Z^{(i)}_j - \tilde{Z}^{(i)}_j, & \quad \text{for } j \in [p] \setminus \{i\} \\
W^{(i)}_i = 0.
\end{align*}

Ultimately, we acquire a feature statistic matrix $W$:

\[
W = \begin{pmatrix} W^{(1)} & W^{(2)} & \cdots & W^{(p-1)} & W^{(p)} \end{pmatrix}^T = \\
\begin{pmatrix}
0 & W^{(1)}_2 & \cdots & W^{(1)}_{p-1} & W^{(1)}_p \\
W^{(2)}_1 & 0 & \cdots & W^{(2)}_{p-1} & W^{(2)}_p \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
W^{(p-1)}_1 & W^{(p-1)}_2 & \cdots & 0 & W^{(p-1)}_p \\
W^{(p)}_1 & W^{(p)}_2 & \cdots & W^{(p)}_{p-1} & 0
\end{pmatrix}
\]

Informally, all information the original knockoff method requires for variable selection with FDR control is in fact gathered in the feature statistics. Similarly, the feature statistic matrix $W$ contains all information we need to achieve FDR control in learning the structure of a GGM. As was previously the case, a large and positive $W^{(i)}_j$ indicates that it is likely that there exists an edge between node $i$ and $j$. 

14
The i\(^{th}\) row \(W^{(i)}\) of the feature statistic matrix \(W\) corresponds to the linear model (10) with response vector \(x^{(i)}\) and random design matrix \(X^{(-i)}\). By Lemma 2.2, \(W^{(i)}\) possesses the sign-flip property on \([p] \setminus ne_i\). This sign-flip property for each row of \(W\) is the key to the FDR control guarantee of our proposed procedure. We would like to point out that one could use different procedures to construct different rows of \(W\), and the rowwise sign-flip property and the FDR control guarantee would still hold. However, in this paper we will restrict ourselves to use the same procedure for the construction of all rows of \(W\).

### 3.1.2 Structure learning with the FDR control

Using the feature statistic matrix \(W\) as a starting point, we now design a strategy that is capable of recovering the underlying graph with the FDR control. The general strategy to estimate the edge set \(E\) is as follows:

- **Step 1.** Choose a threshold \(\hat{T}_i > 0\) for each node \(i\).

- **Step 2.** Estimate the neighborhood of node \(i\) by \(\hat{ne}_i = \{j \in [p] \setminus \{i\} : W^{(i)}_{j} \geq \hat{T}_i\}\).

- **Step 3.** Combine the estimated neighborhoods \(\hat{ne}_i, i = 1, \ldots, p\), by using either the AND or the OR rule to get \(\hat{E}\).

Setting proper thresholds in Step 1 is the key to control the finite sample FDR while not losing all power. The previously mentioned two naive approaches choose the threshold \(\hat{T}_i\) by a local way, which involves only a single row \(W^{(i)}\). Our approach, however, considers the whole feature statistic matrix, and obtain the thresholds through a global manner.

We now introduce some notation. Given the feature statistic matrix \(W\) and the threshold
vector $\widehat{T} = (\widehat{T}_1, \cdots, \widehat{T}_p)$, we define

$$\widehat{F}_i = \{ j \in \widehat{\mathcal{N}}_i : j \notin \mathcal{N}_i \}, \quad \widehat{M}_i = \{ j \in [p] \setminus \{i\} : W_{ij}^{(i)} \leq -\widehat{T}_i \}, \quad \widehat{N}_i = \{ j \in \widehat{M}_i : j \notin \mathcal{N}_i \},$$

$$\widehat{F}_{\text{AND}} = \{(i, j) \in \widehat{E}_{\text{AND}} : (i, j) \notin E\}, \quad \widehat{F}_{\text{OR}} = \{(i, j) \in \widehat{E}_{\text{OR}} : (i, j) \notin E\}. \quad (15)$$

Here, $\widehat{F}_i \subseteq \widehat{\mathcal{N}}_i$ is the set of the falsely discovered neighborhoods of $x_i$. $\widehat{F}_{\text{OR}}$ and $\widehat{F}_{\text{AND}}$ are the sets of the falsely discovered edges when using the AND and the OR rule, respectively. Note that both $\widehat{F}_i$ and $\widehat{N}_i$ are unobservable, while $\widehat{\mathcal{N}}_i$ and $\widehat{M}_i$ are observable. All of the above terms depend on either some threshold $\widehat{T}_i$ and one row $W(i)$ or the threshold vector $\widehat{T}$ and the whole matrix $W$, but we will suppress these dependences when it is not needed for simplicity.

It is easy to see the following relationships

$$|\widehat{F}_{\text{OR}}| = \sum_{i=1}^p |\widehat{F}_i| - |\widehat{F}_{\text{AND}}|, \quad |\widehat{F}_{\text{AND}}| = \frac{1}{2} \left( \sum_{i=1}^p |\widehat{F}_i| - |\widehat{F}_{\text{OR}} \setminus \widehat{F}_{\text{AND}}| \right). \quad (16)$$

With equation (16), we now introduce two methods to choose threshold vector $\widehat{T}$ based on the AND or the OR rule.

When using the AND rule in recovering the graph, we have

$$\text{FDR} = \mathbb{E} \left[ \frac{|\widehat{F}_{\text{AND}}|}{|\widehat{E}_{\text{AND}}| \lor 1} \right] \leq \frac{1}{2} \mathbb{E} \left[ \frac{1}{|\widehat{E}_{\text{AND}}| \lor 1} \sum_{i=1}^p \frac{|\widehat{F}_i|}{1 + |\widehat{N}_i|} \cdot \frac{1 + |\widehat{M}_i|}{|\widehat{E}_{\text{AND}}| \lor 1} \right].$$

Note that the term

$$\frac{1 + |\widehat{M}_i(\widehat{T}_i)|}{|\widehat{E}_{\text{AND}}(\widehat{T})| \lor 1}$$

is observable, thus it can be used to calculate the threshold vector $\widehat{T}$ if the upper bound of

$$\mathbb{E} \left[ \frac{|\widehat{F}_i(\widehat{T}_i, W^{(i)})|}{1 + |\widehat{N}_i(\widehat{T}_i, W^{(i)})|} \right]$$

(17)
is known. Although it seems that the term \( T_i \) depends only on \( W(i) \) which possesses the nice sigh-flip property, we cannot apply the martingale arguments used in Barber and Candès (2015) to obtain an upper bound for this term, due to the fact that \( \hat{T}_i \) depends on the entirety of \( W \). Katsevich et al. (2019) encountered the same issue in their multilayer controlled variable selection problem, and they proposed to bound this term by taking a supremum inside the expectation with respect to the threshold, which lead to an upper bound 1.93 as the following lemma shows.

**Lemma 3.1.** (Reformulated version of Lemma 4 in Katsevich et al. (2019))

Let \( O \subseteq \{1, \ldots, p\} \), \( W = (W_1, \ldots, W_p) \) be a random vector possesses the sign-flip property on \( O \), \( \hat{F}(t) = \{i \in O : W_i \geq t\} \) and \( \hat{N}(t) = \{i \in O : W_i \leq -t\} \). Then

\[
\mathbb{E}\left[ \sup_{t>0} \frac{\hat{|F(t)|}}{1 + |\hat{N}(t)|} \right] \leq 1.93.
\]

Therefore, as long as the data-dependent threshold vector \( \hat{T} = (\hat{T}_1, \ldots, \hat{T}_p) \) satisfies

\[
\frac{1 + |\hat{M}_i(\hat{T}_i)|}{|\hat{E}_{\text{AND}}(\hat{T})| \lor 1} \leq q_i, \quad i = 1, \ldots, p,
\]

for some preset \( q_1, \ldots, q_p \in [0, 1] \) with \( \sum_{i=1}^p q_i = 2q/1.93 \), the FDR is controlled at level \( q \).

To achieve the optimal statistical power, the thresholds \( \hat{T}_1, \ldots, \hat{T}_p \) should be computed through the following optimization problem:

\[
\text{maximize} \quad |\hat{E}_{\text{AND}}| \\
\text{subject to} \quad \frac{1 + |\hat{M}_i(\hat{T}_i)|}{|\hat{E}_{\text{AND}}(\hat{T})| \lor 1} \leq q_i = \frac{2q}{1.93p}, \quad i = 1, \ldots, p. \quad (18)
\]

We set \( \hat{T} = (+\infty, \ldots, +\infty) \) if there is no feasible \( \hat{T} \) satisfying the above constraints. Here we make the choice to set \( q_1 = \cdots, q_p = 2q/(1.93p) \) for two reasons: The first is that this is the
simplest and the most natural choice when we have no prior information about which choice of \( q = (q_1, \ldots, q_p) \) may lead to the largest \( |E_{\text{AND}}| \). The second reason is that setting all \( q_i, i = 1, \ldots, p \), to the same value, as you will see later, allows for a simple algorithm to solve this combinatorial optimization problem quickly and guarantees to find the global optimal. We emphasize again that \( q \) should be preset and not depend on the data in order to guarantee the FDR control.

Similarly, when using the OR rule to recover the graph, we have

\[
\text{FDR} = E \left[ \frac{|\hat{F}_{\text{OR}}|}{|\hat{E}_{\text{OR}}| + 1} \right] \leq E \left[ \sum_{i=1}^{p} \frac{|\hat{F}_i|}{|\hat{E}_{\text{OR}}| + 1} \right] = \sum_{i=1}^{p} E \left[ \frac{|\hat{F}_i|}{1 + |\hat{N}_i|} \cdot \frac{1 + |\hat{M}_i|}{|\hat{E}_{\text{OR}}| + 1} \right]
\]

The corresponding optimization problem then becomes

\[
\begin{align*}
\text{maximize} & \quad |\hat{E}_{\text{OR}}| \\
\text{subject to} & \quad \frac{1 + |\hat{M}_i(\hat{T}_i)|}{|\hat{E}_{\text{OR}}(\hat{T})| + 1} \leq q_i = \frac{q}{1.93p}, \quad i = 1, \ldots, p. \\
& \quad \hat{T}_i \in \{|W_j^i|, j \in [p]\} \cup \{+\infty\} \setminus \{0\}, \quad i = 1, \ldots, p,
\end{align*}
\]

and we set \( \hat{T} = (+\infty, \ldots, +\infty) \) if no feasible point is found.

Remark 3.1. Intuitively, one would expect the method based on the AND rule to be more conservative than based on the OR rule, since it uses the AND rule to recover the graph. However, when comparing (19) and (20), one can see that (19) uses \( 2q/(1.93p) \) as the graph-wide constraint, whereas (20) uses a smaller value \( q/(1.93p) \). Thus in some settings, methods based on the AND rule might be more statistically powerful, if, for example, there is no feasible point in the optimization problem corresponding to the OR rule.
Remark 3.2. As illustrated in next subsection, (19) and (20) can be efficiently solved, so there is no computational issue with respect to the optimization problems. However, through simulation studies we note that such methodology would suffer from the situation that no feasible point exists for (19) and (20) when the “quality” of the feature statistic matrix $W$ is not good enough, so no discovery can be made. In addition, it seems that too sparse graph can not be well handled by this approach. For example, consider (19) (similar arguments hold for (20) ) and the case that there exists $\hat{T}$ for $W$ such that $|\hat{E}_{AND}(\hat{T})| = |E| = \gamma p(p - 1)/2$ and $|\hat{M}_i(\hat{T})| = 0$ for all $i = 1, \cdots, p$, where $\gamma \in [0, 1]$ is the sparsity level of the true graph. Then to fulfill the constraints in (19), we require $\gamma \geq 1.93/((p - 1)q)$. This indicates that in order for a feasible point existing, the true underlying graph can not be too sparse and the minimal sparsity level depends on the dimension $p$ and the nominal FDR level $q$. We point out that such issue widely exists for knockoff based approaches. For example, for the original knockoff method in linear model with dimension $p$ and sparsity level $\gamma$, if we consider the a similar case as for GGM that there exists $\hat{T}$ for $W$ such that $|\hat{S}(\hat{T})| = |S| = \gamma p$ and $\#\{i \in [p] : W_i \leq -\hat{T}\} = 0$, then in order for a feasible point existing in (5) of Algorithm 1 we have $\gamma \geq 1/(pq)$. One can see that the lower bound for the sparsity level $\gamma$ is quite similar to the one corresponds to the nodewise knockoff method, albeit more restrictive by a constant for the latter as the price for ensuring the FDR control of the more complex structure learning problem.

We summarize the nodewise knockoff procedure in Algorithm 2.

3.1.3 Algorithm used to solve the optimization problems (19) and (20)

At first glance, (19) and (20) are both combinatorial optimization problems, and it is seemingly infeasible to solve them through a brute-force search when $p$ is large. However, due to the structure of the constraints, the search space can be significantly restricted and there is in fact a simple algorithm that is guaranteed to efficiently find the global optimal. We restrict ourselves to
Algorithm 2: Nodewise knockoff

**Input:** \((X, q, \mathcal{R}, \mathcal{O}^x, \mathcal{P})\), where \(X\) is a data matrix, \(q \in [0, 1]\) is the nominal FDR level, \(\mathcal{R} \in \{\text{AND, OR}\}\) is the rule used to recover the edge set from the estimated neighborhoods, \(\mathcal{O}^x \in \{\text{Equi (8), SDP (7)}\}\) is the optimization strategy used for the construction of knockoffs, and \(\mathcal{P}\) is the procedure used to construct feature statistics.

**Output:** Estimated edge set \(\widehat{E}\).

1. **Step 1.** Generate feature statistic matrix \(W = (W^{(1)}, \cdots, W^{(p)})^T\) using \(X\) and \(\mathcal{P}\):
   
   for \(i = 1, \cdots, p\) do Treat \(x^{(i)}\) as a response vector and \(X^{(-i)}\) as a design matrix, and apply the first two steps of Algorithm 1 with input \((X^{(-i)}, x^{(i)}, q, \mathcal{O}^x, \mathcal{P})\) to get feature statistics \(W^{(i)}\) with an extra element \(W^{(i)}_i = 0\).

2. **Step 2.** Solving (19) or (20) to get thresholds \(\widehat{T^i} = (\widehat{T^1}_1, \cdots, \widehat{T^p}_p)\) based on \(W\), \(q\) and \(\mathcal{R}\), then obtain \(\widehat{e}_i = \{j \in [p]\setminus\{i\} : W^{(i)}_j \geq \widehat{T^i}_j\}\) for \(i = 1, \cdots, p\).

3. **Step 3.** Recover the estimated edge set \(\widehat{E}\) based on \(\widehat{e}_1, \cdots, \widehat{e}_p\) and \(\mathcal{R}\).

(19) and give two observations that allow for a significant restriction of the search space. Similar arguments hold for (20).

Firstly, since the maximal number of edges in any undirected graph is \(|E_{\text{max}}| = p(p - 1)/2\), \(\widehat{T} = (\widehat{T^1}_1, \cdots, \widehat{T^p}_p)\) can only be feasible for this optimization problem if

\[
|\widehat{M}_i(\widehat{T}_j)| \leq |E_{\text{max}}| \cdot 2q/(1.93p) - 1 = q(p - 1)/1.93 - 1 =: m_{\text{max}}, \quad \forall i = 1, \cdots, p.
\]

Therefore, if there is an element \(\widehat{T}_j\) of \(\widehat{T}\) such that \(|\widehat{M}_i(\widehat{T}_j)| > m_{\text{max}}\), we can conclude that any \(\widehat{T}^*\) with at least one element \(\widehat{T}^*_j \leq \widehat{T}_j\) cannot be feasible as \(|\widehat{M}_j(\widehat{T}^*_j)| \geq |\widehat{M}_j(\widehat{T}_j)| > m_{\text{max}}\), and stop searching in this direction.

Secondly, the left part of the constraint

\[
\frac{1 + |\widehat{M}_i(\widehat{T}_i)|}{|\widehat{E}_{\text{AND}}(\widehat{T})| + 1} \leq \frac{2q}{1.93p}
\]

20
is decreasing with respect to $|\hat{E}_{AND}(\hat{T})|$, and for different $i = 1, \ldots, p$, the only difference of the constraints lies in $|\hat{M}_i(\hat{T}_i)|$. Hence, if for a given vector $\hat{T} = (\hat{T}_1, \ldots, \hat{T}_p)$ with maximal value $|\hat{M}_i(\hat{T}_i)|$ the above constraint is satisfied, we can immediately conclude that the vector $\hat{T}^* = (\hat{T}_1^*, \ldots, \hat{T}_p^*)$ with $\hat{T}_j^* = \min\left\{ T \in \{ |W_j^{(i)}|, i \in [p]\} \cup \{ +\infty \} \setminus \{0\} : |\hat{M}_j(T)| \leq |\hat{M}_i(\hat{T}_i)| \right\}, \forall j = 1, \ldots, p$, is also a feasible point. And since $\hat{T}_j^* \leq \hat{T}_j, \forall j = 1, \ldots, p$, we must have $|\hat{E}_{AND}(\hat{T})| \leq |\hat{E}_{AND}(\hat{T}^*)|$. Therefore, in order to find the global optimal of the original optimization problem (19), it suffices to check the feasibility of $\hat{T} = (\hat{T}_1, \ldots, \hat{T}_p)$ with $\hat{T}_i = \min\left\{ T \in \{ |W_j^{(i)}|, j \in [p]\} \cup \{ +\infty \} \setminus \{0\} : |\hat{M}_i(T)| \leq m \right\}$, for $m = m_{\text{max}}, m_{\text{max}} - 1, \ldots, 0$, and at most $m_{\text{max}}$ points have to be considered.

We summarize the algorithm for solving (19) and (20) in Algorithm 3.

**Algorithm 3 : Algorithm to solve the optimization problem (19) and (20)**

**Input:** Feature statistic matrix $W$ and $m_{\text{max}}$.

**Output:** The global optimal $\hat{T} = (\hat{T}_1, \ldots, \hat{T}_p)$ or $\hat{T} = (+\infty, \ldots, +\infty)$.

1. for $m = m_{\text{max}}, m_{\text{max}} - 1, \ldots, 0$ do Let $\hat{T} = (\hat{T}_1, \ldots, \hat{T}_p)$ with $\hat{T}_i = \min\left\{ T \in \{ |W_j^{(i)}|, j \in [p]\} \cup \{ +\infty \} \setminus \{0\} : |\hat{M}_i(T)| \leq m \right\}, \forall i = 1, \ldots, p$.
2. if $\hat{T}$ satisfy the constraints of the optimization problem, then stop and return $\hat{T}$.
3. if there is no feasible point, then return $\hat{T} = (+\infty, \ldots, +\infty)$.

### 3.1.4 Control the mFDR

Similar to the original fixed-X knockoff method in [Barber and Candes (2015)](https://arxiv.org/abs/1501.01418), we can also control a modified FDR instead of the real FDR of the estimated graph with slightly altered versions of our methods. Specifically, by replacing

$$\frac{1 + |\hat{M}_i(\hat{T}_i)|}{|\hat{E}_{AND}(\hat{T})|} \cup 1 \leq \frac{2q}{1.93p} \quad \text{and} \quad \frac{1 + |\hat{M}_i(\hat{T}_i)|}{|\hat{E}_{OR}(\hat{T})|} \cup 1 \leq \frac{q}{1.93p}$$

21
with
\[
\frac{|\hat{M}_i(T_i)|}{|\hat{E}_{\text{AND}}(T)|} \lor 1 \leq \frac{2q}{1.93p} \quad \text{and} \quad \frac{|\hat{M}_i(T_i)|}{|\hat{E}_{\text{OR}}(T)|} \lor 1 \leq \frac{q}{1.93p}
\] (21)
in (19) and (20), respectively, we can obtain the methods that control the following two modified false discovery rates $\text{mFDR}_{\text{AND}}$ and $\text{mFDR}_{\text{OR}}$, respectively:

\[
\text{mFDR}_{\text{AND}} = \mathbb{E}\left[\frac{|\hat{F}_{\text{AND}}|}{|\hat{E}_{\text{AND}}| + 1.93p/(2q)}\right], \quad \text{mFDR}_{\text{OR}} = \mathbb{E}\left[\frac{|\hat{F}_{\text{OR}}|}{|\hat{E}_{\text{OR}}| + 1.93p/q}\right].
\]

By controlling the mFDR instead of the FDR, more discoveries would be made. However, these two modified FDRs can be much bigger than the real FDR, and are only close to the real FDR if $|\hat{E}_{\text{AND}}| \gg 1.93p/(2q)$ and $|\hat{E}_{\text{OR}}| \gg 1.93p/q$, respectively.

### 3.2 Nodewise knockoffs with sample-splitting-recycling

#### 3.2.1 Motivation

The nodewise knockoff procedure (Algorithm 2) requires deciding hyperparameters $\mathcal{R}$, $\mathcal{O}$, and $\mathcal{P}$ before applying it. However, it is generally unknown about how to set these hyperparameters to achieve a good statistical power for a practical problem. More importantly, as you will see later, different choices of $(\mathcal{R}, \mathcal{O}, \mathcal{P})$ can result in different, even opposite, statistical powers in different settings.

In this paper, we consider using $\mathcal{R} \in \{\text{AND, OR}\}$, $\mathcal{O} \in \{\text{Equi (8), SDP (7)}\}$ and $\mathcal{P} \in \{(12) + (14), (13) + (14)\}$. For the preset $\alpha$ in (12) and (13), we use $\alpha \in \{0.001, 0.1, \cdots, 0.9, 1\}$. For the $\lambda$ in (13), it is not practical to directly set some values as for $\alpha$ because the range of $\lambda$ within which there are nonzero regression coefficients is generally unknown in advance. Therefore, we will choose the value of $\lambda$ after fitting an elastic net and obtaining the corresponding range of $\lambda$ in which there are nonzero regression coefficients. Specifically, let $I$ be the $\lambda$ vector returned by using R package “glmnet” with some preset $\alpha$, we then set hyperparameter $\lambda$ to be the
{0.05, 0.1, …, 0.95, 1} quantile of \( I \). One can verify that the feature statistics constructed in this way satisfy the antisymmetry and sufficiency properties. We consider all combinations of the above choices, obtaining 924 different choices of \((\mathcal{R}, O^s, \mathcal{P}(\alpha, \lambda))\) in total.

These 924 different \((\mathcal{R}, O^s, \mathcal{P})\) can be viewed as 924 different FDR control procedures. To evaluate their performances, we apply them in 4 different settings by simulations. We find that all of the methods control the FDR successfully, but their powers are quite different as illustrated in Figure [1a] and [1b]. Each point in one plot represents the powers of one procedure in two different settings. One can see that there are no procedure that is uniformly the best in two settings. And the most powerful procedure in setting 1 (or setting 3) is even powerless in setting 2 (or setting 4), and vice versa. We also red-colored the approaches with \( \mathcal{P} \) used in the simulation part of Barber and Candès (2015) (that is, using \( \mathcal{R} \in \{ \text{AND, OR} \} \), \( O^s \in \{ \text{Equi, SDP} \} \) and \( \mathcal{P} = (12) + (14) \) with \( \alpha = 1 \)), and one can see that they are powerless in setting 2 and setting 4. The main message conveyed by the above plots is that different choices of \((\mathcal{R}, O^s, \mathcal{P})\) can result in different statistical powers in different settings, and it is clearly not a good idea to stick on one particular choice for practical usage.

Hence, one natural, practical and important question arises: given a bunch of FDR control procedures, which one should we choose for a given problem? Most of the classical research about the FDR control problems focus on developing different procedures in different scenarios given p-values. However, the situation is reversed under the knockoff framework. Analogy to multiple testing context, now it is like that the procedure to obtain the final result based on p-values is fixed, but there are many different approaches to obtain p-values, and the goal is to find the best p-values among many choices while controlling the FDR. We guess that this is one of the reasons that we didn’t find many related research on this problem.
(a) Powers of 924 different FDR control procedures in setting 1 and setting 2. The red points denote the powers of procedures with $P$ as in Barber and Candès (2015). The blue point denotes the power of the nodewise knockoff method with sample-splitting-recycling.

(b) Powers of 924 different FDR control procedures in setting 3 and setting 4. The red points denote the powers of procedures with $P$ as in Barber and Candès (2015). The blue point denotes the power of the nodewise knockoff method with sample-splitting-recycling.

Figure 1: Two plots illustrating the powers of 924 different FDR control procedures and our data-splitting-recycling procedure in 4 different settings.

### 3.2.2 Sample-splitting-recycling for nodewise knockoff method

Which FDR control procedure should we choose for a given problem? The most naive idea is to choose the procedure giving maximal discoveries for this problem, but it is easy to see that such approach would lose the FDR control. Another idea is to implement sample-splitting: use the first part of sample to choose the procedure and implement the chosen procedure in the remaining sample to get the final result. In this way, the FDR is controlled but statistical power would be damaged because the first part of sample is unused to obtain the final result.
In general, reuse of sample would lead to severe statistical issue which makes inference not valid anymore. In particular, the FDR might not be controlled. However, it is possible to reuse sample in some degree while controlling the FDR for the knockoff framework, as shown in Barber and Candès (2016). Their paper considers using knockoffs to select variables in a high-dimensional linear model, and the sample-splitting is used to reduce the dimensionality to obtain a low-dimensional linear model based on the first part of sample. Then, instead of selecting variables based only on the second half of sample, they use all of sample in a particular way that ensures the FDR control. In our case, we will use the same sample-splitting-recycling approach proposed by Barber and Candès (2016) to select the FDR control procedure.

The key for reusing sample lies in the construction of knockoffs corresponding to the used sample. Formally, let \( X = [x^{(1)}, \ldots, x^{(p)}] \in \mathbb{R}^{n \times p} \) be the original sample matrix, \( X^1 = [x^{1(1)}, \ldots, x^{1(p)}] = [x_1^1, \ldots, x_m^1]^T \in \mathbb{R}^{n_1 \times p} \) and \( X^2 = [x^{2(1)}, \ldots, x^{2(p)}] = [x_1^2, \ldots, x_n^2]^T \in \mathbb{R}^{n_2 \times p} \) be two subsample matrices obtained by randomly splitting \( X \) with \( n_1 + n_2 = n \). In this paper, we will use \( n_1 = \lfloor n/2 \rfloor \) and \( n_2 = n - n_1 \), where \( \lfloor n/2 \rfloor \) denotes the biggest integer that is smaller than \( n/2 \). Let \( X^{re} = [x^{re(1)}, \ldots, x^{re(p)}] = \begin{pmatrix} X^1 \\ X^2 \end{pmatrix} \in \mathbb{R}^{n \times p} \) be the collection of these two subsamples, and denote the matrices obtained after deleting the \( i \)th column of \( X^1, X^2 \) and \( X^{re} \) by \( X^{1(-i)} \in \mathbb{R}^{n_1 \times (p-1)}, X^{2(-i)} \in \mathbb{R}^{n_2 \times (p-1)} \) and \( X^{re(-i)} \in \mathbb{R}^{n \times (p-1)} \), respectively.

In the first step, we use \( X^1 \) to select one procedure among \( m \) candidate procedures. Concretely, we run Algorithm 2 with all \( m \) choices of \((R, \mathcal{O}, \mathcal{P})\) to obtain \( \hat{E}_i, i = 1, \ldots, m \). Then we choose the procedure \((R^*, \mathcal{O}^*, \mathcal{P}^*)\) that gives the maximal number of edges \( |\hat{E}^*| \) (if there is a tie, we randomly choose one). For sample-splitting approach, one would then implement the chosen \((R^*, \mathcal{O}^*, \mathcal{P}^*)\) to the remaining data set \( X^2 \) to get the final estimated graph. Sample-splitting-recycling approach, however, also makes use of the used data \( X^1 \) but in a particular way: instead of directly implementing the original nodewise knockoff procedure (Algorithm 2) to \( X^{re} \), we gen-
erate the knockoffs in a different way: for each $i = 1, \cdots, p$, we construct the knockoff matrix by
\[
\tilde{X}_{re(-i)} = \begin{pmatrix}
X^{1(-i)} \\
\tilde{X}^{2(-i)}
\end{pmatrix} \in \mathbb{R}^{n \times (p-1)},
\] (22)
where $\tilde{X}^{2(-i)}$ is the knockoff matrix constructed using (4) based on $X^{*2(-i)}$. Note that here we didn’t construct the knockoff matrix $\tilde{X}_{re(-i)}$ directly based on $X^{re(-i)}$ using (4), and this is the only one modification for the sample-splitting-recycling approach, the remaining procedure is the same as in Algorithm 2 except we use $(X^{re(-i)}, \tilde{X}^{re(-i)})$, $i = 1, \cdots, p$, to construct the feature statistic matrix $W^{re} = (W^{re(1)}, \cdots, W^{re(p)})^T$ with $W^{re(i)} = f([X^{re(-i)} \tilde{X}^{re(-i)}]^T [X^{re(-i)} \tilde{X}^{re(-i)}], [X^{re(-i)} \tilde{X}^{re(-i)}]^T X^{re(i)}, \mathcal{P}(X^*))$.

We summarize the nodewise knockoff procedure with sample-splitting-recycling in Algorithm 4:

**Algorithm 4 : Nodewise knockoff with sample-splitting-recycling**

**Input:** $(X, q, \mathcal{P})$, where $X$ is a data matrix, $q \in [0,1]$ is the nominal FDR level and $\mathcal{P} = \{\mathcal{P}_1, \cdots, \mathcal{P}_m\}$ is the set of candidate procedures to construct feature statistics.

**Output:** Estimated edge set $\tilde{E}$.

1. **Step 1.** Randomly split data matrix $X$ into two parts $X^{*1}$ and $X^{*2}$.

2. **Step 2.** For $i, j = 1, 2$ and $k = 1, \cdots, m$, implement Algorithm 2 with input $(X^{*1}, q, \mathcal{R}_i, \mathcal{O}_k^*, \mathcal{P}_k)$ to obtain $\tilde{E}(X^{*1}, q, \mathcal{R}_i, \mathcal{O}_k^*, \mathcal{P}_k)$. Here $\mathcal{R}_1$ = AND, $\mathcal{R}_2$ = OR, $\mathcal{O}_1^*$ = Equi and $\mathcal{O}_2^*$ = SDP. Let $\tilde{E}(X^{*1}, q, \mathcal{R}^*, \mathcal{O}^*, \mathcal{P}^*)$ be the estimated edge set with the maximal number of edges.

3. **Step 3.** Generate feature statistic matrix $W^{re} = (W^{re(1)}, \cdots, W^{re(p)})^T$ using $X^{re}, \mathcal{O}^*$ and $\mathcal{P}^*$:

4. **for** $i = 1, \cdots, p$ **do** Construct knockoff matrix $\tilde{X}^{re(-i)}$ as in (22) based on $X^{re(-i)}$ and $\mathcal{O}^*$, then apply the Step 2 of Algorithm 1 with $(X^{re(-i)}, \tilde{X}^{re(-i)}, X^{re(i)}, \mathcal{P}^*)$ as input to obtain feature statistics $W^{re(i)}$.

5. **Step 4.** Apply the Step 2 and the Step 3 of Algorithm 2 with $W^{re}$ and $\mathcal{R}^*$ to obtain $\tilde{E}$.
As a quick illustration of this sample-splitting-recycling procedure, we implement it in the previous setting 1 to setting 4 and blue-colored its powers. Figure 1a shows a successful outcome of this approach, in which the blue point lies in the top right corner with almost the best powers in both settings. But of course, it can be expected that there are situations as in Figure 1b in which half of the sample is not enough to choose a good procedure that is powerful in setting 1, so reuse of the sample does not help a lot. Sample-splitting-recycling procedure provides a possible answer to our original question about which FDR control procedure we should choose, but better solution is desired and we leave it as future research.

Finally, we would like to point out that these 924 candidate procedures are our own choices in this paper, one can flexibly use more (or less) candidate procedures (and constructed them in ways different from what we have done). Without any prior information, we recommend to use as many candidate procedures as possible within an affordable computational expense to achieve a good statistical power.

3.2.3 Finite sample FDR control guarantee

At last, we give the theoretical FDR control guarantee of our proposed methods.

Theorem 3.1. (Finite sample error control in structure learning of GGM)

Let \( x_1, \cdots, x_n \) be \( n \) i.i.d. observations from a \( p \)-dimensional Gaussian distribution \( N_p(0, \Sigma) \). Let \( G = (V, E) \) be the corresponding undirected graph and assume \( n \geq 2p \). Then, for any \( q \in [0, 1] \),

- if \( \hat{E} \) is the estimated edge set obtained using Algorithm 2 or 4 and \( \hat{F} = \hat{E} \backslash E \) is the set of falsely discovered edges, then we have

\[
FDR = \mathbb{E} \left[ \frac{\left| \hat{F} \right|}{\left| \hat{E} \right| \lor 1} \right] \leq q;
\]
if $\hat{E}$ is the estimated edge set obtained using Algorithm 2 or 4 based on the first (or second) constraints in (21), and $\hat{F}_{\text{AND}} = \hat{E}_{\text{AND}} \setminus E$ (or $\hat{F}_{\text{OR}} = \hat{E}_{\text{OR}} \setminus E$) is the set of falsely discovered edges, then we have

$$mFDR_{\text{AND}} = \mathbb{E}\left[\frac{|\hat{F}_{\text{AND}}|}{|E_{\text{AND}}| + 1.93 p/(2q)}\right] \leq q \quad \text{or} \quad mFDR_{\text{OR}} = \mathbb{E}\left[\frac{|\hat{F}_{\text{OR}}|}{|E_{\text{OR}}| + 1.93 p/q}\right] \leq q.$$  

### 4 Simulations and a real data example

In this section we examine the performance of our proposed nodewise knockoff method with sample-splitting-recycling in a range of settings and compare it with other methods. All simulations were carried out in R.

#### 4.1 Methods and simulation setups

Most of the existing structure learning algorithms do not ensure the finite sample FDR control, the most powerful method we aware that guarantees this is the Benjamini & Yekutieli (BY) procedure (see Benjamini and Yekutieli (2001)) based on the p-values obtained from testing $\rho_{ij \mid |p| \setminus \{i,j\}} = 0$ for all $i, j \in [p], i \neq j$. Besides, since the Benjamini & Hochberg (BH) procedure (see Benjamini and Hochberg (1995)) is commonly used in practice to control the FDR, we will also compare our method with it even though its FDR control property relies on the PRDS assumption (see Benjamini and Yekutieli (2001)) which may not hold for GGM structure learning problems. We use the R-package “ppcor” developed by Kim (2015) to calculate the p-values in our simulations. In addition, we also consider the GFC-L and the GFC-SL (GGM estimation with FDR control using Lasso or scaled Lasso) proposed by Liu et al. (2013). The FDR control guarantee of these two methods are established in the asymptotic sense under some regularity conditions. In other words, there is no finite sample FDR control guarantee for them. We use the R-package
“SILGGM” developed by [Zhang et al. (2018)] to implement GFC-L and GFC-SL in our simulations. For our proposed nodewise knockoff method with sample-splitting-recycling, we run Algorithm 4 with 924 candidate procedures described in Section 3.2. Besides, we also report the most powerful procedure among these 924 candidate procedures, so that we can see the best performance of the nodewise knockoff method if we know the oracle of the best hyperparameters for a given data set. In all, we consider six methods and denote them by GFC-L, GFC-SL, BH, BY, NK-Re and NK-Oracle, respectively.

According to Remark 3.2, knockoff based methods would be powerless for very sparse graphs. Therefore, we will not consider such setups in our simulations and we point out that the other four multiple testing based procedures may perform well in such settings. We consider the following three types of graphs in our simulations:

- **Band graph**: $\Omega^0_{i,i} = 1$ for $i = 1, \cdots, p$ and $\Omega^0_{i,j} = \text{sign}(a) \cdot |a|^{i-j}/c \cdot 1_{|i-j| \leq \rho}$ for all $i \neq j$. We set $c = 1.5$ to make sure that the entries do not decay too fast.

- **Erdős-Rényi Random graph**: $\Omega^0_{i,i} = 1$ for $i = 1, \cdots, p$ and $\Omega^0_{i,j} = \omega_{i,j} \cdot \phi_{i,j}$ where $\omega_{i,j} \overset{i.i.d.}{\sim} \text{Unif}(C)$ with $C$ some given set and $\phi_{i,j} \overset{i.i.d.}{\sim} \text{Bernoulli}(\delta)$ for all $i, j \in [p], i \neq j$.

- **Graph adapted from real data**: Based on the log and nonparanormal transformed real data in Section 4.3, we keep the first $p$ variables with the maximal sample variances. Then we compute the sample precision matrix $\hat{\Omega} = (\hat{\Sigma})^{-1}$ where $\hat{\Sigma}$ is the sample covariance matrix. Because such precision matrix is dense and the corresponding graph is a complete graph, we apply a thresholding rule with threshold $T$ and let $\Omega^0_{i,j} = \hat{\Omega}_{i,j} \cdot 1_{|\hat{\Omega}_{i,j}| \geq T}$.

The final used precision matrix is $\Omega$. If the minimal eigenvalue $\lambda_{min}$ of $\Omega^0$ is negative, we let $\Omega = \Omega^0 + (|\lambda_{min}| + 0.005)I$ to make sure that $\Omega$ be positive definite. Otherwise we let $\Omega = \Omega^0$. We will give concrete values of the parameters for each graph in next section.
4.2 Simulation results

4.2.1 Power gain by sample-recycling

At first, we present the unsurprise simulation results showing that the power (the proportion of true estimated edges over the total number of the estimated edges) gain by adding sample-recycling compared to simply using sample-splitting. We consider the Band graph with dimension $p = 200$, graph parameters $a = -0.6$, $\rho = 8$, nominal FDR level $q = 0.25$ and sample size $n \in \{2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000\}$. The empirical FDR and power are computed based on 100 replications and the results are shown in Figure 2. As we can see, both approaches control the FDR successfully and the sample-splitting-recycling approach always outperforms the sample-splitting approach as long as the sample size is large enough for the methods to have nontrivial power.
4.2.2 Simulation results for Band graphs

In this part, we focus on Band graphs. We vary sample size $n$, dimension $p$, nominal FDR level $q$ and Band graph parameters $\rho$ and $a$ for a comprehensive comparison. Specifically, we consider the following settings:

- Choose the number of samples $n$ from \{1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000\} while fix $p = 200$, $q = 0.25$, $\rho = 8$ and $a = -0.6$.

- Choose dimension $p$ from \{25, 50, 100, 150, 200, 250, 300\} while fix $n = 5000$, $q = 0.25$, $\rho = 8$ and $a = -0.6$.

- Choose nominal FDR level $q$ from \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\} while fix $p = 200$, $n = 5000$, $\rho = 8$ and $a = -0.6$.

- Choose parameter $\rho$ from \{3, 4, 6, 8, 10, 12\} while fix $p = 200$, $n = 5000$, $q = 0.25$ and $a = -0.6$.

- Choose parameter $a$ from \{-0.1, -0.2, -0.3, -0.4, -0.5, -0.6, -0.7, -0.8, -0.9\} and \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\} while fix $p = 200$, $n = 5000$, $q = 0.25$ and $\rho = 8$.

The empirical FDR and power are computed based on 100 replications and are displayed in Figure 3. We first look at the FDR. As expected, NK-Re, NK-Oracle and BY controls the FDR over all settings. We note that these three methods have similar empirical FDRs that are much more smaller than the nominal FDR in most settings, and we think that such tightness is a price needed to be paid for the theoretical FDR control guarantee over all possible cases. We also note that when the magnitude of the graph parameter $a$ is large (see Figure 3e and 3f), the knockoff based methods achieve better empirical FDRs which are more close to the nominal FDR than BY. For other three methods without the finite sample FDR control guarantee, we find
that GFC-SL loses the FDR control in some settings (see Figure 3b and 3f), but GFC-L and BH successfully control the empirical FDR in these settings, and their empirical FDRs are very close to the nominal FDR.

Next, we turn to the power of the six methods. First we look at the effect of graph parameter $a$ (see Figure 3e and 3f). We note that NK-Re is especially good for the settings with big negative $a$ in which it greatly outperforms all other four p-value based methods, and its power is comparable with other methods for graphs with big positive $a$. However, when the magnitude of $a$ becomes small (so signal to noise ratio decreases), NK-Re can be powerless because the obtained feature statistic matrix $W$ is not good enough and there is no feasible point in the corresponding optimization problems. However, other four methods do not suffer from this and still have non-trivial power. Now we look at Figure 3a, 3b, 3c and 3d, we point out that for all these settings we fix $a = -0.6$, for which the knockoff method perform well, to explore the effect of other parameters. We find that in most of the settings NK-Re outperforms all other four p-value based methods, but again due to the nature of this method, it can be powerless when the sample size $n$ is not big enough, or the nominal FDR $q$ is too small, or the graph parameter $\rho$ is too small (so graph is too sparse), while other four methods are applicable to use in such situations. Finally, as expected, NK-Oracle outperforms NK-Re across all setups and can be powerful when NK-Re is powerless. The powers of these two methods can be comparable in settings where half of the sample does not bring more new information to improve the power and so can be used to select hyperparameters instead.

4.2.3 Simulation results for Erdős-Rényi random graphs

In this part, we focus on Erdős-Rényi random graphs. We consider three types of set $C$ from which the entries of $\Omega$ are sampled: $C = (-1, -0.1) \cup (0.1, 1)$, $C = (-1, 0.1)$ and $C = (0.1, 1)$. We set sample size $n = 8000$, dimension $p = 200$, nominal FDR level $q = 0.25$ and choose
(a) Empirical FDRs and powers of the six methods on Band graphs with varying sample size $n$.

(b) Empirical FDRs and powers of the six methods on Band graphs with varying dimension $p$.

(c) Empirical FDRs and powers of the six methods on Band graphs with varying nominal FDR level $q$.

(d) Empirical FDRs and powers of the six methods on Band graphs with varying graph parameter $\rho$.

(e) Empirical FDRs and powers of the six methods on Band graphs with varying negative graph parameter $a$.

(f) Empirical FDRs and powers of the six methods on Band graphs with varying positive graph parameter $a$.

Figure 3: Empirical FDRs and powers of the six methods on Band graphs.
the probability of adding an edge $\delta$ from $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The empirical FDR and power are computed based on 100 replications.

Figure 4 shows the simulation results. We note that both GFC-L and GFC-SL loses the FDR control (see Figure 4b) while other methods controls the FDR in all settings. With respect to power, we find that BH and BY perform better than the knockoff based methods for the graph with both positive and negative or only positive entries (see Figure 4a and 4c). For graph with negative entries, NK-Oracle can outperform BY and is comparable to BH while NK-Re and BY are comparable and are both less powerful than NK-Oracle and BH. we note that NK-Re becomes powerless for $\delta = 0.05$ in Figure 4a and 4c and for $\delta = 0.3$ in Figure 4b which is caused by the fact that there is no feasible point in its corresponding optimization problem.

![Figure 4](image)

(a) Empirical FDRs and powers of the six methods on Erdős-Rényi random graphs with $C = (-1, -0.1) \cup (0.1, 1)$ and varying probability $\delta$. (b) Empirical FDRs and powers of the six methods on Erdős-Rényi random graphs with $C = (-1, -0.1)$ and varying probability $\delta$. (c) Empirical FDRs and powers of the six methods on Erdős-Rényi random graphs with $C = (0.1, 1)$ and varying probability $\delta$.

Figure 4: Empirical FDRs and powers of the six methods on Erdős-Rényi random graphs.

34
4.2.4 Simulation results for graphs adapted from a real data set

In this part, we use a graph adapted from the real data set in subsection 4.3. We set the threshold $T = 0.05$ and choose the number of the reserved variables $p \in \{20, 30, 40, 50, 60, 70, 80, 90, 100\}$. The empirical FDR and power are computed based on 100 replications and are shown in Figure 5. We find that only GFC-SL loses the FDR control. GFC-L and BH are the most powerful methods and also with larger empirical FDRs compared to the remaining three methods. NK-Re outperforms BY in most of the settings, but becomes powerless when $p$ becomes large, while NK-Oracle outperforms both methods.

![Figure 5: Empirical FDRs and powers of the six methods on the graph adapted from a real data set with varying the number of the reserved variables $p$.](image)

Based on the simulation results from the above three types of graphs, we see that NK-Re outperforms BY, which is the only other method with the finite sample FDR control guarantee in our comparison, and even BH, GFC-L and GFC-SL in some settings, but of course there are also settings that BY is better. We also see that NK-Oracle is more powerful than NK-Re, which reflects the potential of knockoff based method if the hyperparameter issue can be handled in a better way. On the other hand, we note that there is an inherent restriction of the nodewise knockoff
method which is hinged to the corresponding optimization problem. Such restriction can make
this method be powerless in settings where the quality of the obtained feature statistic matrix is
not good enough. Therefore, we would think that the nodewise knockoff method is preferred
in situations where a large data set is available. In such settings, its flexibility in constructing
knockoffs and feature statistics can be utilized, and hence possibly much more powerful than BY
even BH whose p-values are computed in only one way by testing partial correlations.

4.3 Real data application

In this section, we apply the nodewise knockoff method with data-splitting-recycling to a real
data set and compare it to other four methods: GFC-L, GFC-SL, BH and BY. The real data set
is the droplet-based single-cell data set obtained by isolating pan T cells from peripheral blood
mononuclear cells of a healthy human donor. This data set was originally from [Zheng et al. (2017)]
and was analyzed by [Zhang et al. (2018)] using GGM to recover the gene network. Based on the
raw data, [Zhang et al. (2018)] first filtered out unexpressed genes and then took the first 2000 genes
with the maximal sample variances to obtain the data set “sc_pan_T.RData” (this real data set can
be downloaded from the S2 Appendix at https://journals.plos.org/ploscompbiol/article?id=10.1371/
journal.pcbi.1006369), and we will start with this data set \((n = 3555, p = 2000)\) instead of the
raw data set in our analysis.

This gene data set consists of count data, therefore, as in [Zhang et al. (2018)], we preprocessed
it by first performing a log2(counts+1) transformation then implementing a nonparanormal trans-
formation using R function “huge.npn()” of the package “huge” developed by [Zhao et al. (2012)]
to make the joint Gaussian assumption be reasonable. Besides, rather than considering all 2000
genes, we further reduce the number of genes to 50 by keeping the first 50 genes with the max-
imal sample variances. One reason to do this is that one can see from the QQ-plots of the first
200 and 50 genes with the maximal sample variances (see Figure 7 in Supplementary material)
that with 200 variables, even the marginal Gaussian assumption implied by the joint Gaussian assumption is severely violated compared to with fewer variables such as 50. By keeping only 50 variables, the Gaussian assumption seems more appropriate and the obtained network about their conditional independence relationships would be more reliable.

Like most of the sample-splitting approach, NK-Re possesses one undesired property that its outcome is random as the splitting is implemented randomly. Therefore, we randomly split 100 times to investigate how this randomness influence the outcome. Figure 6 and Table 1 displays the results obtained by 100 random splitting for the NK-Re method and the results of the GFC-SL, GFC-L, BH and BY methods. Because the ground truth of the underlying graph is unknown, we choose to compare the number of the estimated edges. We find that the GFC-SL method returns a complete graph with all edges are estimated as existence, and we think that it is very likely that it loses the FDR control due to the result we have seen in Figure 5. The GFC-L and BH methods return very similar results, and the number of discoveries of these two methods are larger than both BY and NK-Re, but note that it is possible that there are many false discoveries included in the results as shown in previous simulations. For the only two approaches with the finite sample FDR control guarantee, NK-Re report more discoveries than the BY. As expected, different sample-splittings lead to different results, but there are still around 300 edges stably estimated as existence.

| Methods     | GFC-SL | GFC-L | BH  | BY  | NK-Re          |
|-------------|--------|-------|-----|-----|----------------|
| Number of estimated edges | 1225   | 563   | 555 | 273 | 524.99 (on average) |

Table 1: The number of the estimated edges of each method.
Figure 6: Edges estimated by the nodewise knockoff method with random sample-splitting-recyclings and by the GFC-SL, GFC-L, BH, BY procedures. We color the edge if it is estimated as existence by the method. Purple, green, orange, blue and black correspond to the GFC-SL, GFC-L, BH, BY and NK-Re methods, respectively.

5 Discussion

In this paper, we consider the problem of controlling the finite sample FDR in GGM structure learning. The only existing method achieving this goal relies on p-values and multiple testing procedures. We, however, approach this problem in a different way by building on a recently proposed knockoff idea by Barber and Candès (2015). A method called nodewise knockoff with theoretical FDR control guarantee was developed. Compared to the p-value based approach for which there are in general not many ways to compute p-values, knockoffs framework is very flexible and thus is potentially more promising. Such flexibility trigger an important and practical problem: how to set the hyperparameters to achieve the best statistical power while maintain the
FDR control guarantee? We use the sample-splitting-recycling approach proposed by Barber and Candès (2016) to address the hyperparameters chosen issue in this paper. Through comprehensive simulations, we see advantages and disadvantages of the knockoff based approach.

Possible directions for future research include developing more efficient procedures to set the hyperparameters, proposing other ways to construct knockoffs and feature statistics (here the knockoffs are computed using SDP or Equi, and the feature statistics are constructed mainly based on elastic net), and extending this approach to high-dimensional GGM settings.

**Supplement material**

The supplement material contains the proofs of Lemma 2.2 and Theorem 3.1 in the main text, and two QQ-plots for real data.

**Acknowledgments**

We are grateful to Leonard Henckel for helpful discussions and commenting on drafts of the paper.

**References**

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News* 533(7604), 452.

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 43, 2055–2085.
Barber, R. F. and E. J. Candès (2016). A knockoff filter for high-dimensional selective inference. To appear in Annals of Statistics. arXiv:1602.03574.

Barber, R. F., E. J. Candès, and R. J. Samworth (2018). Robust inference with knockoffs. To appear in Annals of Statistics. arXiv:1801.03896.

Bates, S., E. Candès, L. Janson, and W. Wang (2019). Metropolized knockoff sampling. arXiv preprint arXiv:1903.00434.

Begley, C. G. and L. M. Ellis (2012). Drug development: Raise standards for preclinical cancer research. Nature 483(7391), 531.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29, 1165–1188.

Cai, T., W. Liu, and X. Luo (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association 106(494), 594–607.

Candes, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society. Series B (Methodological) 80, 551–577.

Dai, R. and R. Barber (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In International Conference on Machine Learning, pp. 1851–1859.
d’Aspremont, A., O. Banerjee, and L. El Ghaoui (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications* 30(1), 56–66.

Drton, M. and M. D. Perlman (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science* 22, 430–449.

Fan, J., Y. Feng, and Y. Wu (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics* 3(2), 521.

Fan, Y., E. Demirkaya, G. Li, and J. Lv (2019). Rank: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 1–43.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441.

Gimenez, J. R., A. Ghorbani, and J. Zou (2018). Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine* 2(8), e124.

Janson, L., W. Su, et al. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics* 10(1), 960–975.

Jordon, J., J. Yoon, and M. van der Schaar (2018). Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks.

Katsevich, E., C. Sabatti, et al. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics* 13(1), 1–33.
Kim, S. (2015). ppcor: An r package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods* 22(6), 665.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

Liu, W. et al. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics* 41, 2948–2978.

Liu, Y. and C. Zheng (2018). Auto-encoding knockoff generator for fdr controlled variable selection. *arXiv preprint arXiv:1809.10765*.

Meinshausen, N. and P. Buhlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34, 1436–1462.

Raskutti, G., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2009). Model selection in gaussian graphical models: High-dimensional consistency of $\ell_1$-regularized mle. In *Advances in Neural Information Processing Systems*, pp. 1329–1336.

Romano, Y., M. Sesia, and E. J. Candès (2018). Deep knockoffs. *arXiv preprint arXiv:1811.06687*.

Rothman, A. J., P. J. Bickel, E. Levina, J. Zhu, et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2, 494–515.

Sesia, M., C. Sabatti, and E. Candès (2018). Gene hunting with hidden markov model knockoffs. *Biometrika* 106(1), 1–18.

Weinstein, A., R. Barber, and E. Candès (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.
Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 94, 19–35.

Zhang, R., Z. Ren, and W. Chen (2018). Silggm: An extensive r package for efficient statistical inference in large-scale gene networks. *PLoS computational biology* 14(8), e1006369.

Zhao, T., H. Liu, K. Roeder, J. Lafferty, and L. Wasserman (2012). The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research* 13(Apr), 1059–1062.

Zheng, G. X., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications* 8, 14049.

Zheng, Z., J. Zhou, X. Guo, and D. Li (2018). Recovering the graphical structures via knockoffs. *Procedia Computer Science* 129, 201–207.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67(2), 301–320.
SUPPLEMENTARY MATERIAL

A Proofs

Lemma 2.2. (Sign-flip property on \( N_0 \) with random design matrix)

Let \( X \in \mathbb{R}^{n \times p} \) be a random design matrix in a linear model, \( N_0 \) be the index set of null variables and \( W \) be a feature statistic vector satisfying sufficiency and antisymmetry properties. Then \( W \) possesses the sign-flip property on \( N_0 \).

This lemma trivially follows from the proof of Lemma 1 in Barber and Candès (2015) by first conditioning on the random design matrix then marginalizing it, here we write it in detail only for clearness. We first introduce two useful lemmas from Barber and Candès (2015).

Lemma A.1. (Pairwise exchangeability for the features, Barber and Candès (2015))

Let \( X = [x^{(1)}, \cdots, x^{(p)}] \in \mathbb{R}^{n \times p} \) be a fixed design matrix in a linear model and \( \bar{X} \in \mathbb{R}^{n \times p} \) be its knockoff counterpart. Then for any subset \( O \subset \{1, \cdots, p\} \), the Gram matrix of \([X \bar{X}]\) is unchanged when we swap \( x^{(i)} \) and \( \bar{x}^{(i)} \) for each \( i \in O \). That is,

\[
[X \bar{X}]^{\text{swap}}(O)[X \bar{X}]^{\text{swap}}(O) = [X \bar{X}]^T[X \bar{X}].
\]

Here \([X \bar{X}]^{\text{swap}}(O)\) denotes the matrix obtained by first swapping then transposing matrix \([X \bar{X}]\).

Lemma A.2. (Pairwise exchangeability for the response, Barber and Candès (2015))

Let \( y \) be a response vector, \( X = [x^{(1)}, \cdots, x^{(p)}] \in \mathbb{R}^{n \times p} \) be a fixed design matrix and \( N_0 \) be the index set of the null variables in a linear model, and let \( \bar{X} \in \mathbb{R}^{n \times p} \) be the knockoff counterpart of \( X \). Then, for any subset \( O \subseteq N_0 \), the distribution of \([X \bar{X}]^T y\) is unchanged when we swap \( x^{(i)} \) and \( \bar{x}^{(i)} \) for each \( i \in O \). That is,

\[
[X \bar{X}]^{\text{swap}}(O)y \overset{d}{=} [X \bar{X}]^Ty.
\]
Now we show Lemma 2.2. This proof follows the same outline as the proof of Lemma 1 in Barber and Candès (2015).

Proof of Lemma 2.2. Let $\epsilon = (\epsilon_1, \cdots, \epsilon_p)$ be a sign sequence independent of $W$ with $\epsilon_i = +1$ for $i \notin N_0$ and $\epsilon_i$ i.i.d. from a Rademacher distribution for $i \in N_0$. Let $S = \{i \in [p] : \epsilon_i = -1\}$ be a set depending on $\epsilon$, note that we have $S \subseteq N_0$. Then, by the antisymmetry property of $W$ and the definition of $S$, we have

$$W_{\text{swap}(S)} = (W_1 \cdot \epsilon_1, \cdots, W_p \cdot \epsilon_p).$$

Thus, to prove Lemma 2.2, it suffices to show that

$$W_{\text{swap}(S)} \overset{d}{=} W.$$

Conditional on $X$, we have

$$([X \, X]^T_{\text{swap}(S)}[X \, X]_{\text{swap}(S)}, [X \, X]^T_{\text{swap}(S)}y) \mid X \overset{d}{=} ([X \, X]^T[X \, X], [X \, X]^Ty) \mid X.$$

by Lemma A.1 and Lemma A.2. Thus for the joint distribution, we have

$$([X \, X]^T_{\text{swap}(S)}[X \, X]_{\text{swap}(S)}, [X \, X]^T_{\text{swap}(S)}y, X) \overset{d}{=} ([X \, X]^T[X \, X], [X \, X]^Ty, X).$$

Then by marginalizing $X$ out, we have

$$([X \, X]^T_{\text{swap}(S)}[X \, X]_{\text{swap}(S)}, [X \, X]^T_{\text{swap}(S)}y) \overset{d}{=} ([X \, X]^T[X \, X], [X \, X]^Ty)$$

Finally, by the sufficiency property of $W$, we have

$$W_{\text{swap}(S)} = f([X \, X]^T_{\text{swap}(S)}[X \, X]_{\text{swap}(S)}, [X \, X]^T_{\text{swap}(S)}y) \overset{d}{=} f([X \, X]^T[X \, X], [X \, X]^Ty) = W.$$

$\square$
Theorem 3.1. (Finite sample error control in structure learning of GGM)

Let $x_1, \ldots, x_n$ be $n$ i.i.d. observations from a $p$-dimensional Gaussian distribution $N_p(0, \Sigma)$. Let $G = (V, E)$ be the corresponding undirected graph and assume $n \geq 2p$. Then, for any $q \in [0, 1]$,

- if $\hat{E}$ is the estimated edge set obtained using Algorithm 2 or 4 and $\hat{F} = \hat{E} \setminus E$ is the set of falsely discovered edges, then we have
  $$\text{FDR} = \mathbb{E} \left[ \frac{|\hat{F}|}{|\hat{E} \vee 1|} \right] \leq q;$$

- if $\hat{E}$ is the estimated edge set obtained using Algorithm 2 or 4 based on the first (or second) constraints in (21), and $\hat{F}_{\text{AND}} = \hat{E}_{\text{AND}} \setminus E$ (or $\hat{F}_{\text{OR}} = \hat{E}_{\text{OR}} \setminus E$) is the set of falsely discovered edges, then we have
  $$\text{mFDR}_{\text{AND}} = \mathbb{E} \left[ \frac{|\hat{F}_{\text{AND}}|}{|\hat{E}_{\text{AND}}| + 1.93p/(2q)} \right] \leq q \quad \text{or} \quad \text{mFDR}_{\text{OR}} = \mathbb{E} \left[ \frac{|\hat{F}_{\text{OR}}|}{|\hat{E}_{\text{OR}}| + 1.93p/q} \right] \leq q.$$

To prove Theorem 3.1, all we need is the upper bound given by Lemma 3.1 (this lemma is from Katsevich et al. (2019)) which requires the sign-flip property. Hence we first show that each row of the feature statistic matrix $W$ possesses the sign-flip property on its neighborhood.

Lemma A.3. (Sign-flip property of each row of $W$)

Let $X \in \mathbb{R}^{n \times p}$ be the original sample matrix, $X^1 \in \mathbb{R}^{n_1 \times p}$ and $X^2 \in \mathbb{R}^{n_2 \times p}$ be two subsamples obtained by randomly splitting $X$ with $n_1 + n_2 = n$, and $X^{\text{re}} = \begin{pmatrix} X^1 \\ X^2 \end{pmatrix} \in \mathbb{R}^{n \times p}$ be the collection of these two subsamples.

- Let $W = (W^{(1)}, \cdots, W^{(p)})^T$ be the feature statistic matrix calculated using Algorithm 2. For any $i = 1, \cdots, p$, let $\text{ne}_i$ be the neighborhood of node $i$, then $W^{(i)} = (W_{1}^{(i)}, \cdots, W_{p}^{(i)})$ possesses the sign-flip property on $[p] \setminus \text{ne}_i \cup \{i\}$. 46
Let $W_{re} = (W_{re}^{(1)}, \ldots, W_{re}^{(p)})^T$ be the feature statistic matrix calculated using Algorithm 4. For any $i = 1, \ldots, p$, let $ne_i$ be the neighborhood of node $i$. Then conditional on $X^*$, $W_{re(i)} = (W_{re(i)}^{(1)}, \ldots, W_{re(i)}^{(p)})$ possesses the sign-flip property on $[p]\{ne_i \cup \{i\}]$. 

**Proof of Lemma A.3.** The first statement follows directly from Lemma 2.2.

Now we prove the second statement. For any $i = 1, \ldots, p$, we first show that conditional on $X^*$ and $X_{re(-i)}$, $W_{re(i)}$ possesses the sign-flip property on $[p]\{ne_i \cup \{i\}]$ using the same idea as the proof of Lemma 1 in Barber and Candès (2015). Let $S_i \subseteq [p]\{ne_i \cup \{i\}]$ be any subset, we would like to show that

$$[X_{re(-i)} \bar{X}_{re(-i)}]_{swap(S_i)}^{T} [X_{re(-i)} \bar{X}_{re(-i)}]_{swap(S_i)} = [X_{re(-i)} \bar{X}_{re(-i)}]^{T} [X_{re(-i)} \bar{X}_{re(-i)}]. \quad (23)$$

and

$$[X_{re(-i)} \bar{X}_{re(-i)}]_{swap(S_i)}^{T} x_{re(i)} = d [X_{re(-i)} \bar{X}_{re(-i)}]^{T} x_{re(i)}. \quad (24)$$

For (23), note that

$$[X_{re(-i)} \bar{X}_{re(-i)}]_{swap(S_i)} = \begin{pmatrix} [X_{1}^{*1(-i)} X_{2}^{*1(-i)}]_{swap(S_i)} \\ [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]_{swap(S_i)} \end{pmatrix},$$

so

$$[X_{re(-i)} \bar{X}_{re(-i)}]_{swap(S_i)}^{T} [X_{re(-i)} \bar{X}_{re(-i)}]_{swap(S_i)}$$

$$= \begin{pmatrix} [X_{1}^{*1(-i)} X_{2}^{*1(-i)}]_{swap(S_i)} \\ [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]_{swap(S_i)} \end{pmatrix}^{T} \begin{pmatrix} [X_{1}^{*1(-i)} X_{2}^{*1(-i)}]_{swap(S_i)} \\ [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]_{swap(S_i)} \end{pmatrix}$$

$$= [X_{1}^{*1(-i)} X_{1}^{*1(-i)}]_{swap(S_i)}^{T} [X_{1}^{*1(-i)} X_{1}^{*1(-i)}]_{swap(S_i)} + [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]_{swap(S_i)}^{T} [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]_{swap(S_i)}$$

$$= [X_{1}^{*1(-i)} X_{1}^{*1(-i)}]^{T} [X_{1}^{*1(-i)} X_{1}^{*1(-i)}] + [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]^{T} [X_{2}^{*2(-i)} X_{2}^{*2(-i)}]$$

$$= \begin{pmatrix} [X_{1}^{*1(-i)} X_{1}^{*1(-i)}] \\ [X_{2}^{*2(-i)} X_{2}^{*2(-i)}] \end{pmatrix}^{T} \begin{pmatrix} [X_{1}^{*1(-i)} X_{1}^{*1(-i)}] \\ [X_{2}^{*2(-i)} X_{2}^{*2(-i)}] \end{pmatrix}$$

$$= [X_{re(-i)} \bar{X}_{re(-i)}]^{T} [X_{re(-i)} \bar{X}_{re(-i)}],$$

47
where the third equality follows from the fact that \([XX]_{\text{swap}(S)} = [XX]\) for any matrix \(X\) and the property of knockoffs \(\widetilde{X}^{*2(-i)}\) in (6).

For (24), we have

\[
[X^{re(-i)} \widetilde{X}^{re(-i)}]^{T}_{\text{swap}(S_i)} x^{re(i)}
\]

\[
= \left( [X^{*1(-i)} X^{*1(-i)}]_{\text{swap}(S_i)} \right)^{T} x^{*1(i)}
\]

\[
= [X^{*1(-i)} X^{*1(-i)}]^{T}_{\text{swap}(S_i)} x^{*1(i)} + [X^{*2(-i)} \widetilde{X}^{*2(-i)}]^{T}_{\text{swap}(S_i)} x^{*2(i)}
\]

\[
\overset{d}{=} [X^{*1(-i)} X^{*1(-i)}]^{T} x^{*1(i)} + [X^{*2(-i)} \widetilde{X}^{*2(-i)}]^{T} x^{*2(i)}
\]

\[
= [X^{re(-i)} \widetilde{X}^{re(-i)}]^{T} x^{re(i)}.
\]

The third equation holds because \([X^{*1(-i)} X^{*1(-i)}]^{T}_{\text{swap}(S_i)} x^{*1(i)} = [X^{*1(-i)} X^{*1(-i)}]^{T} x^{*1(i)}\) and \([X^{*2(-i)} \widetilde{X}^{*2(-i)}]^{T}_{\text{swap}(S_i)} x^{*2(i)} \overset{d}{=} [X^{*2(-i)} \widetilde{X}^{*2(-i)}]^{T} x^{*2(i)}\) by Lemma A.2.

Let \(\epsilon^{(i)} = (\epsilon_1, \cdots, \epsilon_p)\) be a sign sequence independent of \(W^{re(i)}\) with \(\epsilon_j = +1\) for \(j \notin [p]\{ne_i \cup \{i\}\}\) and \(\epsilon_j\) i.i.d. from a Rademacher distribution for \(j \in [p]\{ne_i \cup \{i\}\}\). Let \(S^{(i)} = \{j \in [p] : \epsilon_j = -1\}\) be a set depending on \(\epsilon^{(i)}\), and we have \(S^{(i)} \subseteq [p]\{ne_i \cup \{i\}\}\). Then, by the antisymmetry property of \(W^{re(i)}\) and the definition of \(S^{(i)}\), we have

\[
W^{re(i)}_{\text{swap}(S^{(i)})} = (W^{(i)}_1 \cdot \epsilon_1, \cdots, W^{(i)}_p \cdot \epsilon_p)
\]

Thus it suffices to show that

\[
W^{re(i)}_{\text{swap}(S^{(i)})} \overset{d}{=} W^{re(i)}
\]

By (23) and (24), we have

\[
\left( [X^{re(-i)} \widetilde{X}^{re(-i)}]^{T}_{\text{swap}(S^{(i)})} [X^{re(-i)} \widetilde{X}^{re(-i)}], [X^{re(-i)} \widetilde{X}^{re(-i)}]^{T} x^{re(i)} \right) \overset{d}{=}
\]

\[
\left( [X^{re(-i)} \widetilde{X}^{re(-i)}]^{T}_{\text{swap}(S^{(i)})} [X^{re(-i)} \widetilde{X}^{re(-i)}]_{\text{swap}(S^{(i)})}, [X^{re(-i)} \widetilde{X}^{re(-i)}]^{T}_{\text{swap}(S^{(i)})} x^{re(i)} \right)
\]
Thus, conditional on $X^* (\text{so } \mathcal{P}(X^*) \text{ is fixed})$ and $X^{re(-i)}$, we have

$$W_{\text{swap}(S^{(i)})}^{re(i)} = f([X^{re(-i)} \tilde{X}^{re(-i)}]_\text{swap}(S^{(i)})^T [X^{re(-i)} \tilde{X}^{re(-i)}]_\text{swap}(S^{(i)})^{T}, [X^{re(-i)} \tilde{X}^{re(-i)}]_\text{swap}(S^{(i)})^{T} x^{re(i)}, \mathcal{P}(X^*))$$

$$= f([X^{re(-i)} \tilde{X}^{re(-i)}]_\text{swap}(S^{(i)})^{T} [X^{re(-i)} \tilde{X}^{re(-i)}]_\text{swap}(S^{(i)})^{T} x^{re(i)}, \mathcal{P}(X^*))$$

Therefore, $W^{re(i)}$ possesses the sign-flip property on $[p]\{ne_i \cup \{i\}\}$ conditional on $X^*$ and $X^{re(-i)}$. Then, similar to the proof of Lemma 2.2 we can show that $W_{\text{swap}(S^{(i)})}^{re(i)} \overset{d}{=} W^{re(i)}$ when $X^{re(-i)}$ is treated as random, and finally conclude that $W^{re(i)}$ possesses the sign-flip property on $[p]\{ne_i \cup \{i\}\}$ conditional on $X^*$. □

Now we proof Theorem 3.1.

**Proof of Theorem 3.1.** We first prove the results related to Algorithm 2 (that is, results without using sample-splitting-recycling). By the first statement of Lemma A.3 and Lemma 3.1, we have

$$\mathbb{E} \left[ \sup_{t \geq 0} \frac{|\tilde{F}_i(t)|}{|\tilde{N}_i(t)|} \right] \leq 1.93,$$

(25)

where $\tilde{F}_i(t) = \{ j \in [p]\{ne_i \cup \{i\}\} : W_j^{(i)} \geq t \}$ and $\tilde{N}_i(t) = \{ j \in [p]\{ne_i \cup \{i\}\} : W_j^{(i)} \leq -t \}$.

For clearness of notation, we point out that $\tilde{F}_i(t)$ and $\tilde{N}_i(t)$ defined here are the same as in (15).

Specifically, we have $\tilde{F}_i(t) = \{ j \in \tilde{M}_i : j \notin ne_i \}$ with $\tilde{M}_i = \{ j \in [p] : W_j^{(i)} \geq t \}$ and $\tilde{N}_i = \{ j \in \tilde{M}_i : j \notin ne_i \}$ with $\tilde{M}_i = \{ j \in [p] : W_j^{(i)} \leq -t \}$ because $W_j^{(i)} = 0$ and $t > 0$. 

49
For Algorithm 2 with \( R = \text{AND} \), we have

\[
\text{FDR} = \mathbb{E} \left[ \frac{\widehat{F}_{\text{AND}}}{|E_{\text{AND}}| \vee 1} \right] = \mathbb{E} \left[ \frac{\widehat{F}_{\text{AND}}}{|E_{\text{AND}}| \vee 1} \mathbbm{1}_{|E_{\text{AND}}| \geq 1} \right]
\]

where the indication function \( \mathbbm{1}_{|E_{\text{AND}}| \geq 1} \) serves as a restriction to cases with a feasible threshold, and we make this restriction in this and the following three proofs to allow for the invocation of (19), (20), and two inequalities of (21), respectively.

Similarly, for Algorithm 2 with \( R = \text{OR} \), we have

\[
\text{FDR} = \mathbb{E} \left[ \frac{\widehat{F}_{\text{OR}}}{|E_{\text{OR}}| \vee 1} \right] \leq \mathbb{E} \left[ \frac{p}{|E_{\text{OR}}| \vee 1} \sum_{i=1}^{p} \frac{\widehat{F}_i}{|E_{\text{OR}}| \vee 1} \right] \leq \frac{q}{1.93p} \sum_{i=1}^{p} \mathbb{E} \left[ \frac{\widehat{F}_i(T_i)}{1 + |N(T_i)|} \right] \leq q.
\]
For Algorithm 2 with $\mathcal{R} = \text{AND}$ and the first constraint in (21), we have

$$mFDR_{\text{AND}} = \mathbb{E}\left[ \frac{\|\hat{F}_{\text{AND}}\|}{\|\hat{E}_{\text{AND}}\| + 1.93p/2q} \right] = \mathbb{E}\left[ \frac{\|\hat{F}_{\text{AND}}\|}{\|\hat{E}_{\text{AND}}\| + 1.93p/2q} \mathbbm{1}_{\{\|\hat{E}_{\text{AND}}\| \geq 1\}} \right] \quad \text{see (16)}$$

$$\leq \frac{1}{2} \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}\|}{\|\hat{E}_{\text{AND}}\| + 1.93p/2q} \mathbb{1}_{\{\|\hat{E}_{\text{AND}}\| \geq 1\}} \right] \quad \text{see (21)}$$

$$\leq \frac{1}{2} \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}\| \cdot 1 + \|\tilde{M}_{i}\|}{1 + \|\tilde{N}_{i}\| \cdot \|\hat{E}_{\text{AND}}\| + 1.93p/2q} \mathbb{1}_{\{\|\hat{E}_{\text{AND}}\| \geq 1\}} \right]$$

$$= \frac{q}{1.93p} \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}(\tilde{T}_{i})\|}{1 + \|\tilde{N}_{i}(\tilde{T}_{i})\|} \right] \quad \leq \frac{q}{1.93p} \sum_{i=1}^{p} \mathbb{E}\left[ \sup_{t \geq 0} \frac{\|\hat{F}_{i}(t)\|}{1 + \|\tilde{N}_{i}(t)\|} \right] \quad \text{see (25)} \leq q.$$ 

At last, for Algorithm 2 with $\mathcal{R} = \text{OR}$ and the second constraint in (21), we have

$$mFDR_{\text{OR}} = \mathbb{E}\left[ \frac{\|\hat{F}_{\text{OR}}\|}{\|\hat{E}_{\text{OR}}\| + 1.93p/q} \right] \quad \text{see (16)}$$

$$\leq \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}\| \cdot 1 + \|\tilde{M}_{i}\|}{1 + \|\tilde{N}_{i}\| \cdot \|\hat{E}_{\text{OR}}\| + 1.93p/q} \mathbb{1}_{\{\|\hat{E}_{\text{OR}}\| \geq 1\}} \right] \quad \text{see (21)}$$

$$\leq \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}\| \cdot 1 + q/(1.93p) \cdot \|\hat{E}_{\text{OR}}\|}{1 + \|\tilde{N}_{i}\| \cdot \|\hat{E}_{\text{OR}}\| + 1.93p/q} \mathbb{1}_{\{\|\hat{E}_{\text{OR}}\| \geq 1\}} \right]$$

$$= \frac{q}{1.93p} \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}(\tilde{T}_{i})\|}{1 + \|\tilde{N}_{i}(\tilde{T}_{i})\|} \cdot \frac{1.93p/q + \|\hat{E}_{\text{OR}}\|}{\|\hat{E}_{\text{OR}}\| + 1.93p/q} \mathbb{1}_{\{\|\hat{E}_{\text{OR}}\| \geq 1\}} \right]$$

$$= \frac{q}{1.93p} \sum_{i=1}^{p} \mathbb{E}\left[ \frac{\|\hat{F}_{i}(\tilde{T}_{i})\|}{1 + \|\tilde{N}_{i}(\tilde{T}_{i})\|} \right] \leq \frac{q}{1.93p} \sum_{i=1}^{p} \mathbb{E}\left[ \sup_{t \geq 0} \frac{\|\hat{F}_{i}(t)\|}{1 + \|\tilde{N}_{i}(t)\|} \right] \quad \text{see (25)} \leq q.$$ 

Now we show the FDR control results related to Algorithm 4 with sample-splitting-recycling.
By the second statement of Lemma A.3 and Lemma 3.1, we have
\[
\mathbb{E}\left[ \sup_{t>0} \frac{\left| \hat{F}^{re}_i(t) \right|}{1 + \left| \hat{N}^{re}_i(t) \right|} \right] \leq 1.93, \tag{26}
\]
where \( \hat{F}^{re}_i(t) = \{ j \in [p] \ \mid \{ne_i \cup \{i\} : W_j^{re(i)} \geq t \} \) and \( \hat{N}^{re}_i(t) = \{ j \in [p] \ \mid \{ne_i \cup \{i\} : W_j^{re(i)} \leq -t \} \).

Therefore, conditional on \( X^{*1} \) and using the same argument as the first two proofs with the only modification that invoke (26) instead of (25) in the last inequality, we have
\[
\mathbb{E}\left[ \frac{\left| \hat{F}^{re}_{AND} \right|}{\left| \hat{E}^{re}_{AND} \right| \lor 1} \mid X^{*1} \right] \leq q \quad \text{and} \quad \mathbb{E}\left[ \frac{\left| \hat{F}^{re}_{OR} \right|}{\left| \hat{E}^{re}_{OR} \right| \lor 1} \mid X^{*1} \right] \leq q, \tag{27}
\]
where \( \hat{E}^{re}_{AND} \) is the estimated edge set using AND rule and \( W^{re} \), \( \hat{F}^{re}_{AND} \) is the set of falsely discovered edges in \( \hat{E}^{re}_{AND} \), and similarly for \( \hat{E}^{re}_{OR} \) and \( \hat{F}^{re}_{OR} \).

Therefore, for Algorithm 4, we have
\[
\text{FDR} = \mathbb{E}\left[ \frac{\left| \hat{F}^{re} \right|}{\left| \hat{E}^{re} \right| \lor 1} \right] = \mathbb{E}\left[ \mathbb{E}\left[ \frac{\left| \hat{F}^{re}_i \right|}{\left| \hat{E}^{re}_i \right| \lor 1} \mid X^{*1} \right] \right]
= \mathbb{E}\left[ \mathbb{E}\left[ \frac{\left| \hat{F}^{re}_{AND} \right|}{\left| \hat{E}^{re}_{AND} \right| \lor 1} 1\{\text{Choose AND rule based on } X^{*1}\} \mid X^{*1} \right] + \mathbb{E}\left[ \frac{\left| \hat{F}^{re}_{OR} \right|}{\left| \hat{E}^{re}_{OR} \right| \lor 1} 1\{\text{Choose OR rule based on } X^{*1}\} \mid X^{*1} \right] \right]
\leq q \cdot \mathbb{E}\left[ 1\{\text{Choose AND rule based on } X^{*1}\} + 1\{\text{Choose OR rule based on } X^{*1}\} \right]
= q.
\]

The control of mFDR can be proved in a similar way.

\[ \square \]

\section*{B QQ plots for real data}
(a) QQ-plots of the first 200 variables with maximal variance, here different colors correspond to different variables.

(b) QQ-plots of the first 50 variables with maximal variance, here different colors correspond to different variables.

Figure 7: Two QQ-plots for real data.