Statistica Sinica Preprint No: SS-2023-0099	
Title	Reproducible Learning in Large-Scale Multiple Graphical Models
Manuscript ID	SS-2023-0099
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0099
Complete List of Authors	Jia Zhou, Guangming Pan, Zeming Zheng and Changchun Tan
Corresponding Authors	Jia Zhou
E-mails	tszhjia@mail.ustc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Statistica Sinica

REPRODUCIBLE LEARNING IN LARGE-SCALE MULTIPLE GRAPHICAL MODELS

Jia Zhou[1], Guangming Pan*[2], Zeming Zheng[3] and Changchun Tan[1]

[1]*Hefei University of Technology,* [2]*Nanyang Technological University and* [3]*University of Science and Technology of China*

Abstract: Reproducible learning of the underlying structure among large-scale network data is important in many contemporary applications. Despite the fast-growing literature on this subject, the practical issue of data heterogeneity has rarely been addressed. In this paper, we propose a new method called the multiple graphical knockoff filter to efficiently recover the underlying sparse connected structure of a general population from a high-dimensional heterogeneous dataset. We provide theoretical justification on the asymptotic false discovery rate control, and the theory for the power analysis is also established. To the best of our knowledge, this is the first formal theoretical result on the power for the graphical knockoffs procedure. Our new methodology and results are evidenced by numerical studies.

Key words and phrases: False discovery rate, Heterogeneity, Multiple graphical models, High-dimensionality, Power .

*Corresponding author

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

1. Introduction

The surge of big data in an unprecedented scale has brought us an enormous amount of information that makes large-scale network analysis increasingly frequent in many contemporary applications, such as biology, economics, and social science (e.g. Giudici and Alessanfro (2016), and Shin et al. (2014)). It is often of practical interest to uncover the underlying network formed by a large number of individuals that are sparsely related. As a popular choice, Gaussian graphical models provide a flexible way to specify the conditional independence structure among a large number of nodes. There is a growing literature on Gaussian graphical models, mainly focusing on the problem of support recovery and link strength estimation; see for example, Friedman et al. (2008); Fan and Lv (2016); Cheng et al. (2017); Zhou et al. (2022), and among many others.

To obtain a reliable outcome and alleviate reproducibility issues, controlling the false discovery rate (FDR) which is defined as the expected proportion of false discoveries among all the discoveries proposed by Benjamini and Hochberg (1995) has gained much attention recently. There have been several studies proposed focusing on FDR control in structure learning for Gaussian graphical models. One class of methods is based on multiple testing approaches. For example, in low-dimensional settings,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Drton and Perlman (2007) suggested testing pairwise partial correlations. After obtaining the corresponding p-values, the BH procedure (Benjamini and Yekutieli, 2001) can be applied to recover the graph structure with finite sample FDR control without any additional assumptions. In high-dimensional settings, Liu (2013) proposed a structure learning algorithm named as GFC based on a certain test statistic and its asymptotic distribution, providing asymptotic FDR control under some regularity conditions.

Another class of methods is based on the knockoff idea which was originally proposed by Barber and Candès (2015) for low-dimensional linear models called fixed-X knockoff and later extended to high-dimensional regression models with random design called model-X knockoff (Candès et al., 2018). The nice properties of the model-X knockoff procedure, such as having no restrictions on dimensions and the conditional distribution, make this procedure widely used and developed (Barber et al., 2020; Fan et al., 2020; Liu et al., 2022). Based on fixed-X knockoff and model-X knockoff framework, Li and Maathuis (2021) and Zhou et al. (2022) proposed graphical knockoff methods to control the FDR of low-dimensional graphical models and high-dimensional graphical models respectively. Additionally, a new method that achieves the FDR control by estimating two independent parameters via data splitting and then obtaining symmetric statistics has also

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

been extended to Gaussian graphical models (Dai et al., 2023).

Among those endeavors, they all focus on a single graphical model which assumed that the dataset is homogeneous. However, in some real applications such as climate research, disease diagnosis, text mining, and so on, high-dimensional heterogeneous datasets are popularly observed (Guo et al., 2011; Lee and Liu, 2015; Ma and Michailidis, 2016), wherein the dataset comprises multiple subpopulations. As a motivation example, consider a gene expression dataset of breast cancer derived from the METABRIC, which are discussed in detail in Section 5. Literature (Johnson et al., 2021) suggests that breast cancer can be categorized into four molecular subtypes, each exhibiting critical differences in incidence, survival rates, and imaging characteristics. It is more realistic to acknowledge that gene expression level distributions may vary across these subtypes, resulting in dataset heterogeneity. Ignoring the heterogeneity and employing existing FDR control methods designed for homogeneous datasets to reconstruct gene network structures may lead to a loss of power, rendering some significant edges undetectable. Hence, in this paper, we introduce a novel procedure called multiple graphical knockoff filter (MGKF) to address this challenge.

The major contributions of this paper are threefold. First of all, to the best of our knowledge, this is the first work to attempt to address the chal-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

lenging issue of heterogeneity in reproducible learning of graphical models. In the face of the heterogeneous data, our method can efficiently recover the underlying connectivity patterns of a general population of interest with guaranteed FDR control and high power. Secondly, we provide theoretical justifications on the asymptotic false discovery rate control, and the power analysis is also established. It's worth pointing out that this is the first formal theoretical result on the power for the graphical knockoffs procedure. Last but not least, benefiting from the tuning-free property of the heterogeneous group square-root Lasso algorithm (Ren et al., 2019), our procedure can deal with large-scale datasets with high computational efficiency.

The rest of the paper is organized as follows. Section 2 presents the problem setup and our new methodology. We establish the theoretical properties of the proposed method including FDR control and the power analysis in Section 3. Simulation studies and a real data analysis are provided in Sections 4 and Section 5, respectively. Section 6 discusses implications and extensions of our work. Additional technical details and all the proofs are relegated to supplementary material.

Notations: For any $a \in \{1, \ldots, p\}$, write $[-a] = \{1, \ldots, p\} \backslash \{a\}$, and we abbreviate it as $-a$ when it appears in the subscript. For any $a \in \{1, \ldots, p\}$, $\mathbf{x}_{-a}^{(t)}$ denotes the subvector of $\mathbf{x}^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ by ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

cluding $X_a^{(t)}$. Moreover, the notation $\mathbf{X}_a^{(t)}$ denotes the ath column of $\mathbf{X}^{(t)}$, and $\mathbf{X}_{-a}^{(t)}$ denotes the submatrix of $\mathbf{X}^{(t)}$ with the columns in $[-a]$. For any vector $\mathbf{v}$, $\|\mathbf{v}\|_d$ denotes the l_d norm of $\mathbf{v}$ for $d \geq 0$. For any matrix $\mathbf{M} \in \mathbb{R}^{p_1 \times p_2}$ and any subsets $A \subset \{1, \ldots, p_1\}$, $B \subset \{1, \ldots, p_2\}$, $\mathbf{M}_{A,B}$ denotes the submatrix of $\mathbf{M}$ with the rows in A and columns in B. The notation $\|\mathbf{M}\|_2$ denotes the spectral norm of $\mathbf{M}$. Moreover, for a sequence of matrices $\mathbf{M}_1, \ldots, \mathbf{M}_p$, $\text{bdiag}\{\mathbf{M}_1, \ldots, \mathbf{M}_p\}$ denotes the block diagonal matrix consisting of $\mathbf{M}_1, \ldots, \mathbf{M}_p$. Denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest eigenvalues of a given symmetric matrix, respectively. The notation $|\cdot|$ denotes the cardinality of a set, E^c is the complement of E, and $d_1 \vee d_2 = \max\{d_1, d_2\}$.

2. The FDR control of multiple graphs

2.1 Model settings

Motivated by the prevalence of heterogeneous datasets, we focus on the setting of multiple networks with Gaussian graphical models to encode the connectivity patters among p features $X_1, \ldots, X_p$ measured on k subpopulations of a general population. For each class $1 \leq t \leq k$, consider the Gaussian graphical model $G^{(t)} = (V, E^{(t)})$ for a p-variate random vector

$$\mathbf{x}^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})^\top \sim N(\mathbf{0}, \mathbf{\Sigma}^{(t)}) \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

where the superscript (t) means that these p features are measured on the tth subpopulation, $\mathbf{\Sigma}^{(t)}$ is $p \times p$ covariance matrix, and $G^{(t)}$ is an undirected graph associated with $\mathbf{x}^{(t)}$. Here, $V = \{1, \ldots, p\}$ the set of vertices and $E^{(t)} \subseteq \{(i,j) : 1 \leq i < j \leq p\}$ the set of the edges between vertices of tth graph. The lack of an edge (j,k) in tth graph is characterized by $X_j^{(t)} \perp\!\!\!\perp X_k^{(t)} | \mathbf{x}_{-\{j,k\}}^{(t)}$, where $\mathbf{x}_{-\{j,k\}}^{(t)}$ represents the set of all variables in $\mathbf{x}^{(t)}$ except for X_j and X_k. The connectivity patters of the general population can be characterized by $E = \cup_{t=1}^{k} E^{(t)}$.

In high-dimensional settings where the number of covariates p is comparable to or exceeds the number of observation, the connectivity patters among the set of covariates are usually assumed sparse. It's reasonable to assume that these k graphs share a similar sparsity structure as they belong to a general population, where most of the pairs are not connected in all graphs. Meanwhile, the rest are connected in all or some graphs with the connectivity strengths between nodes and the variability of nodes change across subpopulations due to the specificity, which cause the heterogeneity of the observed dataset $\mathbf{X} = \left((\mathbf{X}^{(1)})^\top, \ldots, (\mathbf{X}^{(k)})^\top \right)^\top$ with $\mathbf{X}^{(t)} = (\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{n^{(t)}}^{(t)})^\top$, where $\{\mathbf{x}_i^{(t)}\}_{i=1}^{n^{(t)}}$ are independent and identically distributed (i.i.d.) copies from model (2.1). In addition, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are assumed to be independent.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Denote by $\widehat{E}$ the estimated edge set by some selection procedure based on the heterogeneous dataset $\mathbf{X}$. The FDR and power of the selection procedure with respect to the edge set E is defined as

$$\mathrm{FDR}(\widehat{E}) = \mathbb{E}\left[\frac{|\widehat{E} \cap E^c|}{|\widehat{E} \vee 1|}\right], \quad \text{and} \quad \mathrm{Power}(\widehat{E}) = \mathbb{E}\left[\frac{|\widehat{E} \cap E|}{|E|}\right].$$

Our goal is to develop a procedure to recover the edge set E with guaranteed FDR control, meanwhile enjoying high power. Existing high-dimensional graphical FDR control methods, such as GCF (Liu, 2013) and HGKF (Zhou et al., 2022), are unsuitable for our models since they are designed for homogeneous datasets. Additionally, learning the structure of each subpopulation individually and then merging them to estimate E is inefficient, as it fails to leverage the common structure across different groups. Therefore, we propose a new method called the multiple graphical knockoff filter to achieve this goal.

2.2 Review of model-X knockoffs framework

Our suggested procedure falls in the general framework of model-X knockoffs (Candès et al., 2018), which we briefly review in this section. The key ingredient of the model-X knockoffs framework is the construction of the so-called model-X knockoff variables that are defined as follows.

Definition 1. (Model-X knockoff variables (Candès et al., 2018)) For a

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

set of random variables $\mathbf{x} = (X_1, \ldots, X_p)$, a new set of random variables $\widetilde{\mathbf{x}} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$ is called a set of model-X knockoff variables if it satisfies the following properties: (1) For any subset $S \subset \{1, \ldots, p\}$, we have $(\mathbf{x}, \widetilde{\mathbf{x}})_{\text{swap}(S)} \overset{d}{=} (\mathbf{x}, \widetilde{\mathbf{x}})$, where $\overset{d}{=}$ denotes equal in distribution and the vector $(\mathbf{x}, \widetilde{\mathbf{x}})_{\text{swap}(S)}$ is obtained by swapping X_j and $\widetilde{X}_j$ for each $j \in S$. (2) Conditional on $\mathbf{x}$, $\widetilde{\mathbf{x}}$ is independent of response Y, if there is a response Y.

With the constructed knockoff variables $\widetilde{\mathbf{x}}$, the next step is to construct knockoff statistic $W_j = f_j(Z_j, \widetilde{Z}_j)$ for each $1 \leq j \leq p$, where Z_j and $\widetilde{Z}_j$ represent feature importance measure for jth covariate X_j and its knockoff counterpart $\widetilde{X}_j$, respectively, and $f_j(\cdot, \cdot)$ is an antisymmetric function satisfying $f_j(z_j, \tilde{z}_j) = -f_j(\tilde{z}_j, z_j)$. For example, in linear regression model, one can choose Z_j and $\widetilde{Z}_j$ as the Lasso regression coefficients of X_j and $\widetilde{X}_j$, respectively, and a widely used knockoff statistic called Lasso Coefficient Difference (LCD) is defined as $W_j = |Z_j| - |\widetilde{Z}_j|$. Observe that all model-X knockoff variables $\widetilde{X}_j's$ are just noise features by the second property in Definition 1. Thus, intuitively, a large positive value of knockoff statistic W_j indicates that jth covariate X_j is important, while a small magnitude of W_j usually corresponds to noise features. The final step of the knockoffs inference framework is to sort $|W_j's|$ from high to low and select features whose $W_j's$ are at or above some threshold T.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

2.3 Multiple graphical knockoff filter

Recall that for each subpopulation $1 \leq t \leq k$, the p features follows the multiple Gaussian distribution $N(\mathbf{0}, \mathbf{\Sigma}^{(t)})$ (2.1). It's well-known (See, e.g. Lauritzen (1996)) that there is no edge between nodes a and j in tth graph if and only if $\beta_{aj}^{(t)} = \beta_{ja}^{(t)} = 0$, where $\beta_{aj}^{(t)}$ is the regression coefficient of $X_j^{(t)}$ in the regression of $X_a^{(t)}$ on $\mathbf{x}_{-a}^{(t)}$, that is

$$X_a^{(t)} = \mathbf{x}_{-a}^{(t)} \boldsymbol{\beta}_a^{(t)} + \eta_a^{(t)} \tag{2.2}$$

where $\boldsymbol{\beta}_a^{(t)} = (\beta_{aj}^{(t)}, j \in [-a])^\top \in \mathbb{R}^{p-1}$, and $\eta_a^{(t)}$ is the random noise independent of $\mathbf{x}_{-a}^{(t)}$. To estimate the set of edges E, it's equivalent to finding out these pairs $\{(i,j), 1 \leq i < j \leq p\}$ that satisfy $\boldsymbol{\beta}_{i(j)}^\top = (\beta_{ij}^{(1)}, \ldots, \beta_{ij}^{(t)}) \neq \mathbf{0}$.

Equation (2.2) builds the relationship between a Gaussian graphical model and linear models. It's therefore natural to consider make use of the knockoff framework for FDR controlled graph estimation in Gaussian graphical models. Following model-X knockoff framework, our method also has three steps.

Step 1: Construct knockoffs. For any subgroup $1 \leq t \leq k$, given a node a, we treat the other $p-1$ nodes as predictors. The ideal knockoff variables $\widetilde{\mathbf{x}}_{-a}^{(t)} = (\widetilde{X}_j^{(t)}, j \in [-a])^\top \in \mathbb{R}^{p-1}$ can be constructed by sampling

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

from the conditional distribution $\widetilde{\mathbf{x}}_{-a}^{(t)}|\mathbf{x}_{-a}^{(t)} \sim N(\boldsymbol{\mu}_a^{(t)}, \boldsymbol{\Upsilon}_a^{(t)})$ with

$$\begin{aligned}\boldsymbol{\mu}_a^{(t)} &= \left(\mathbf{I}_{p-1} - \text{diag}\{\mathbf{s}_a^{(t)}\}\boldsymbol{\Omega}_{-a}^{(t)}\right)\mathbf{x}_{-a} \quad \text{and} \\ \boldsymbol{\Upsilon}_a^{(t)} &= 2\text{diag}\{\mathbf{s}_a^{(t)}\} - \text{diag}\{\mathbf{s}_a^{(t)}\}\boldsymbol{\Omega}_{-a}^{(t)}\text{diag}\{\mathbf{s}_a^{(t)}\}, \end{aligned} \tag{2.3}$$

where $\mathbf{I}_{p-1}$ is the $(p-1)\times(p-1)$ identity matrix, $\boldsymbol{\Omega}_{-a}^{(t)} = (\boldsymbol{\Sigma}_{-a,-a}^{(t)})^{-1}$ is the precision matrix of $\mathbf{x}_{-a}$, $\text{diag}\{\mathbf{s}_a^{(t)}\}$ is a $(p-1)\times(p-1)$ diagonal matrix with the vector $\mathbf{s}_a^{(t)} = \{s_{aj}^{(t)}\}_{j\in[-a]}$ being the non-negative diagonal entries such that $\boldsymbol{\Sigma}_{-a,-a}^{(t)} - 2^{-1}\text{diag}\{\mathbf{s}_a^{(t)}\}$ is positive semidefinite. We will adopt the semidefinite programme construction (SDP) (Candès et al., 2018) to obtain an appropriate $\mathbf{s}_a^{(t)}$. Similar to that in Fan et al. (2020), we will treat it as a nuisance parameter throughout our theoretical analysis.

Since $\widetilde{\mathbf{x}}_{-a}^{(t)}$ is constructed without looking at $X_a^{(t)}$ and

$$\begin{pmatrix}\mathbf{x}_{-a}^{(t)} \\ \widetilde{\mathbf{x}}_{-a}^{(t)}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{-a,-a}^{(t)} & \boldsymbol{\Sigma}_{-a,-a}^{(t)} - \text{diag}\{\mathbf{s}_a^{(t)}\} \\ \boldsymbol{\Sigma}_{-a,-a}^{(t)} - \text{diag}\{\mathbf{s}_a^{(t)}\} & \boldsymbol{\Sigma}_{-a,-a}^{(t)}\end{pmatrix}\right),$$

the ideal knockoff variables obviously satisfy Definition 1. For each $1 \le i \le n^{(t)}$, denote by $\mathbf{x}_{i,-a}^{(t)}$ the ith row of $\mathbf{X}_{-a}^{(t)}$. The ith row of the ideal knockoff matrix $\widetilde{\mathbf{X}}_{-a}^{(t)}$ can be constructed as above (2.3) using $\mathbf{x}_{i,-a}$ and $\boldsymbol{\Omega}_{-a}^{(t)}$ for $1 \le a \le p, 1 \le t \le k$.

However, the true matrices $\boldsymbol{\Omega}_{-a}^{(t)}$ for all $1 \le a \le p, 1 \le t \le k$ used to construct the ideal knockoff variables (2.3) are generally unknown. We

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

will replace $\mathbf{\Omega}_{-a}^{(t)}$ with its some consistent estimate to generate approximate knockoff variables in practice and show that FDR control can still be guaranteed (See Section 3 for details).

Step 2: Calculate statistics. Based on the knockoffs, we will extend the LCD statistic mentioned in Section 2.2 to our model to construct the knockoff statistic $W_{a,j}$ measuring the importance of node j to node a. Specifically, let $[(\boldsymbol{\zeta}_a^{(t)})^\top, (\widetilde{\boldsymbol{\zeta}}_a^{(t)})^\top]$ be an estimated coefficient vector of $\mathbf{X}_a^{(t)}$ regression on $[\mathbf{X}_{-a}^{(t)}, \widetilde{\mathbf{X}}_{-a}^{(t)}]$, where $(\boldsymbol{\zeta}_a^{(t)})^\top = (\zeta_{aj}, j \in [-a]) \in \mathbb{R}^{p-1}$ and $(\widetilde{\boldsymbol{\zeta}}_a^{(t)})^\top = (\widetilde{\zeta}_{aj}, j \in [-a]) \in \mathbb{R}^{p-1}$. Since the importance of node j to node a is characterized by k graphs, the extended LCD statistic has the form of

$$W_{a,j} = f_j(Z_j, \widetilde{Z}_j) = |Z_j| - |\widetilde{Z}_j| = \|\boldsymbol{\zeta}_{a(j)}\|_l - \|\widetilde{\boldsymbol{\zeta}}_{a(j)}\|_l, \tag{2.4}$$

where $\boldsymbol{\zeta}_{a(j)} = (\zeta_{aj}^{(1)}, \ldots, \zeta_{aj}^{(k)})$, $\widetilde{\boldsymbol{\zeta}}_{a(j)} = (\widetilde{\zeta}_{aj}^{(1)}, \ldots, \widetilde{\zeta}_{aj}^{(k)})$, and $0 < l < \infty$.

The crucial point now is how to obtain an efficient estimate of the regression coefficient vectors. It would be beneficial to borrow the strength across all k classes of data to achieve more accurate estimation since these k subpopulations share some common structure. In our procedure, we novelly stack all k multiple linear regressions which combine original variables and knockoff variables by introducing a large-scale block diagonal design matrix, taking into account both heterogeneity and common information among different graphs. The similar connected patterns between different graphs

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

make the coefficient vector of the large-scale linear regression have a group sparsity structure, then the heterogeneous group square-root Lasso (HGSL) (Ren et al., 2019) can be applied to obtain an efficient estimator of the coefficient vector, so as to obtain competitive statistics.

Specifically, Let

$$\mathcal{Y}_a = \left((\mathbf{X}_a^{(1)})^\top, \ldots, (\mathbf{X}_a^{(k)})^\top\right), \ \mathcal{X}_{-a} = \text{bdiag}\left\{(\mathbf{X}_{-a}^{(1)}, \widetilde{\mathbf{X}}_{-a}^{(1)}), \ldots, (\mathbf{X}_{-a}^{(k)}, \widetilde{\mathbf{X}}_{-a}^{(k)})\right\}$$

with k blocks. We run the heterogeneous group square-root Lasso (HGSL) proposed by Ren et al. (2019) with the response $\mathcal{Y}_a$ and the combined design matrix $\mathcal{X}_{-a}$ to jointly estimate of regression coefficients, that is,

$$\widehat{\boldsymbol{\beta}}_a^{\text{aug}} = \arg\min_{\mathbf{b}^{\text{aug}} \in \mathbb{R}^{2(p-1)k}} \left\{ \sum_{t=1}^{k} Q_t(\mathbf{b}^{(t)}, \widetilde{\mathbf{b}}^{(t)}) + \lambda \left(\sum_{j \neq a} \|\mathbf{D}_{a(j)}^{1/2} \mathbf{b}_{(j)}\|_2 + \sum_{j \neq a} \|\widetilde{\mathbf{D}}_{a(j)}^{1/2} \widetilde{\mathbf{b}}_{(j)}\|_2 \right) \right\}, \tag{2.5}$$

where

$$Q_t(\mathbf{b}^{(t)}, \widetilde{\mathbf{b}}^{(t)}) = \frac{\left\| \mathbf{X}_a^{(t)} - \left(\mathbf{X}_{-a}^{(t)}, \widetilde{\mathbf{X}}_{-a}^{(t)}\right) \begin{pmatrix} \mathbf{b}^{(t)} \\ \widetilde{\mathbf{b}}^{(t)} \end{pmatrix} \right\|_2}{\sqrt{n}},$$

with $n = \min_{1 \leq t \leq k}\{n^{(t)}\}$ and $\lambda > 0$ is the regularization parameter. Additionally, $\mathbf{b}^{(t)} = (b_j^{(t)}, j \in [-a])^\top$, $\widetilde{\mathbf{b}}^{(t)} = (\widetilde{b}_j^{(t)}, j \in [-a])^\top$, $\mathbf{b}_{(j)} = (b_j^{(1)}, \ldots, b_j^{(k)})^\top$, $\widetilde{\mathbf{b}}_{(j)} = (\widetilde{b}_j^{(1)}, \ldots, \widetilde{b}_j^{(k)})^\top$, and $\mathbf{b}^{\text{aug}} = \left((\mathbf{b}^{(1)})^\top, (\widetilde{\mathbf{b}}^{(1)})^\top \ldots, (\mathbf{b}^{(k)})^\top, (\widetilde{\mathbf{b}}^{(k)})^\top\right)$. Moreover, the notations $\mathbf{D}_{a(j)}$ and $\widetilde{\mathbf{D}}_{a(j)}$ denote $k \times k$ diagonal matrices with their t-th diagonal entry equal to $(\mathbf{X}_j^{(t)})^\top \mathbf{X}_j^{(t)}/n^{(t)}$ and $(\widetilde{\mathbf{X}}_j^{(t)})^\top \widetilde{\mathbf{X}}_j^{(t)}/n^{(t)}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

which are introduced to scale the design matrix. As we can see that in equation (2.5) $Q_t(\mathbf{b}^{(t)}, \widetilde{\mathbf{b}}^{(t)})$ can be written as

$$Q_t(\mathbf{b}^{(t)}, \widetilde{\mathbf{b}}^{(t)}) = \frac{\left\| \mathbf{X}_a^{(t)} - \left(\mathbf{X}_{-a}^{(t)} (\mathbf{D}_a^{(t)})^{-1/2}, \widetilde{\mathbf{X}}_{-a}^{(t)} (\widetilde{\mathbf{D}}_a^{(t)})^{-1/2} \right) \begin{pmatrix} (\mathbf{D}_a^{(t)})^{1/2} \mathbf{b}^{(t)} \\ (\widetilde{\mathbf{D}}_a^{(t)})^{1/2} \widetilde{\mathbf{b}}^{(t)} \end{pmatrix} \right\|_2}{\sqrt{n}},$$

where $\mathbf{D}_a^{(t)} = \mathrm{diag}\left((\mathbf{X}_{-a}^{(t)})^\top \mathbf{X}_{-a}^{(t)} / n^{(t)} \right)$ and $\widetilde{\mathbf{D}}_a^{(t)} = \mathrm{diag}\left((\widetilde{\mathbf{X}}_{-a}^{(t)})^\top \widetilde{\mathbf{X}}_{-a}^{(t)} / n^{(t)} \right)$ are diagonal scaling matrices.

Based on $\widehat{\boldsymbol{\beta}}_a^{\mathrm{aug}} = \left((\widehat{\boldsymbol{\beta}}_a^{(1)})^\top, (\widetilde{\boldsymbol{\beta}}_a^{(1)})^\top, \ldots, (\widehat{\boldsymbol{\beta}}_a^{(k)})^\top, (\widetilde{\boldsymbol{\beta}}_a^{(k)})^\top \right)$, where $\widehat{\boldsymbol{\beta}}_a^{(t)} = (\widehat{\beta}_{aj}^{(t)}, j \in [-a])^\top \in \mathbb{R}^{p-1}$ and $\widetilde{\boldsymbol{\beta}}_a^{(t)} = (\widetilde{\beta}_{aj}^{(t)}, j \in [-a])^\top \in \mathbb{R}^{p-1}$, we specifically use l_2 norm as in (2.4) to simplify the power analysis that

$$W_{a,j} = \|\widehat{\boldsymbol{\beta}}_{a(j)}\|_2 - \|\widetilde{\boldsymbol{\beta}}_{a(j)}\|_2, \text{ for } 1 \le a \ne j \le p \tag{2.6}$$

where $\widehat{\boldsymbol{\beta}}_{a(j)} = (\widehat{\beta}_{aj}^{(1)}, \ldots, \widehat{\beta}_{aj}^{(k)})$ and $\widetilde{\boldsymbol{\beta}}_{a(j)} = (\widetilde{\beta}_{aj}^{(1)}, \ldots, \widetilde{\beta}_{aj}^{(k)})$.

Step 3: Find the global threshold vector. For convenience, hereafter we called the statistic $W_{i,j}$ as the null statistic if there is no edge between the node pair (i, j). We take a similar form as that in Li and Maathuis (2021) and Zhou et al. (2022) to obtain a global threshold vector. Specifically, for each node $a \in \{1, \ldots, p\}$, we denote by NE_a the true neighborhood of node a, which is defined as $\mathrm{NE}_a = \{j \in [-a] : \boldsymbol{\beta}_{a(j)} = (\beta_{aj}^{(1)}, \ldots, \beta_{aj}^{(k)}) \ne \mathbf{0}\}$. Given a positive threshold vector $\mathbf{T} = (T_1, \ldots, T_p)$, for each node a, we can estimate NE_a by $\widehat{\mathrm{NE}}_a(\mathbf{T}) = \{j \in [-a] : W_{a,j} \ge T_a\}$. The estimated edge set

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

$\widehat{E}(\mathbf{T})$ is given by

$$\widehat{E}(\mathbf{T}) = \Big\{(a,j) : j \in \widehat{\mathrm{NE}}_a(\mathbf{T}) \text{ or } a \in \widehat{\mathrm{NE}}_j(\mathbf{T}), 1 \le a < j \le p\Big\}. \tag{2.7}$$

Now given a pre-specified FDR level q, we will obtain an appropriate threshold vector $\widehat{\mathbf{T}}$ as follows.

$$\begin{aligned}\widehat{\mathbf{T}} =&(\widehat{T}_1, \ldots, \widehat{T}_p) = \underset{\mathbf{T}=\{T_1,\ldots,T_p\}}{\arg\max} \ |\widehat{E}(\mathbf{T})| \\ &\text{subject to } \ \frac{\gamma + |\{j : j \in [-a], W_{a,j} \le -T_a\}|}{|\widehat{E}(\mathbf{T})| \vee 1} \le \tilde{q}_a := \frac{q}{c_\gamma p} \\ &\text{and } T_a \in \{|W_{a,j}|, j \in [-a]\} \cup \{+\infty\} \backslash \{0\} \ \text{ for all } \ a \in \{1, \ldots, p\}.\end{aligned} \tag{2.8}$$

Similar to that in Li and Maathuis (2021) and Zhou et al. (2022), we provide two alternative pairs $(1, 1.93)$ and $(0.01, 102)$ for the choice of (γ, c_γ). We set $\widehat{\mathbf{T}} = (+\infty, \ldots, +\infty)$ if there is no feasible solution. Based on the global threshold vector $\widehat{\mathbf{T}}$, we can obtain the estimate edge set $\widehat{E}(\widehat{\mathbf{T}})$ via equation (2.7). For convenience, we will abbreviate $\widehat{E}(\widehat{\mathbf{T}})$ as $\widehat{E}$ in the following context, which denotes the final estimated edge set of our procedure with true precision matrices.

Note that as we mentioned in step 1, the true precision matrices used to generate ideal knockoff matrices $\widetilde{\mathbf{X}}_{-a}^{(t)}$ are generally unknown, and we will use some estimated precision matrices to replace the true ones in (2.3) to generate approximate knockoff matrices, then proceed steps 2-3. In short,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

we will summarize that for any given sequence of symmetric positive denite matrices in $(p-1)$ dimensions $\{\mathbf{\Gamma}_{-a}^{(t)}\}_{1\leq a\leq p, 1\leq t\leq k} := \theta$, our procedure with parameter θ can be proceed as follow.

Procedure 1 (MGKF): **1.** For each node a, to generate knockoff matrix $\widetilde{\mathbf{X}}_{-a}^{(t)}(\theta)$ following the Step 1 by replacing $\mathbf{\Omega}_{-a}^{(t)}$ with $\mathbf{\Gamma}_{-a}^{(t)}$.

2. For each node a, to obtain the estimated regression coefficient vector $\widehat{\boldsymbol{\beta}}_a^{\text{aug}}(\theta)$ via equation (2.5) by replacing $\widetilde{\mathbf{X}}_{-a}^{(t)}$ with $\widetilde{\mathbf{X}}_{-a}^{(t)}(\theta)$. Then calculating the knockoff statistic $W_{a,j}(\theta)$ via (2.6) based on $\widehat{\boldsymbol{\beta}}_a^{\text{aug}}(\theta)$.

3. Based on $\{W_{a,j}(\theta), 1 \leq a \neq j \leq p\}$, the threshold vector $\widehat{\mathbf{T}}(\theta)$ is found by step 3 via (2.8), and the final edge set estimation $\widehat{E}(\theta)$ is obtained via (2.7) with the $\widehat{\mathbf{T}}(\theta)$.

3. Theoretical properties

In this section, we investigate the theoretical properties of our proposed procedure, including asymptotic FDR control and power analysis. Throughout our theoretical analysis, we consider the regularization parameter fixed at $\lambda = C_\lambda[\frac{k+\log(p)}{n}]^{1/2}$ with C_λ is some positive constant. Therefore, we will drop the dependence of various quantities on λ whenever there is no confusion. We begin with introducing some technical conditions that will be used in our theoretical analysis.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Condition 1. For $1 \leq t \leq k$, the eigenvalues of $\mathbf{\Omega}^{(t)}$ are uniformly bounded within the interval $[1/M_1, M_1]$ for some constant $M_1 \geq 1$.

Condition 2. It holds that $n^{(1)} \asymp \cdots \asymp n^{(k)}$ with $\max_{1\leq t\leq k}\{n^{(t)}\}/n \leq M_2$, where $\asymp$ means the same order, $n = \min_{1\leq t\leq k}\{n^{(t)}\}$, and M_2 is some positive constant.

Condition 3. For some positive constants M_3, δ and $b_n \to 0$ as $n \to \infty$, with probability at least $1 - p^{-\delta}$, $\|\widehat{\mathbf{\Omega}}^{(t)}_{-a} - \mathbf{\Omega}^{(t)}_{-a}\|_2 \leq M_3 b_n$ holds uniformly over $1 \leq a \leq p$, $1 \leq t \leq k$.

Condition 1 is typical, which is also used in Fan et al. (2020) and Zhou et al. (2022). Similar to Ren et al. (2019), we assume that in Condition 2 that our sample is balanced with sample sizes of each of the k classes comparable to each other. Condition 3 is quite flexible, which is also introduced in Zhou et al. (2022), Fan et al. (2020), and Fan et al. (2015). This condition holds for many existing approaches, such as CLIME (Cai et al., 2011), ISEE (Fan and Lv, 2016), and Glasso (Friedman et al., 2008), as long as the estimators are sparse and enjoy some typical entry-wise estimation accuracy under mild regularity conditions.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

3.1 FDR control guarantee

To develop the theory for FDR control, we begin with an important lemma that motivates the basic theoretical framework. For ease of presentation, let $\mathbf{U}_{-a}^{(t)}(\theta) = [\mathbf{X}_{-a}^{(t)}, \widetilde{\mathbf{X}}_{-a}^{(t)}(\theta)]^\top [\mathbf{X}_{-a}^{(t)}, \widetilde{\mathbf{X}}_{-a}^{(t)}(\theta)]/n$ and $\mathbf{V}_{-a}^{(t)}(\theta) = [\mathbf{X}_{-a}^{(t)}, \widetilde{\mathbf{X}}_{-a}^{(t)}(\theta)]^\top \mathbf{X}_{a}^{(t)}/n$. Denote by

$$\mathbf{U}_{-a}(\theta) = \left(\mathbf{U}_{-a}^{(1)}(\theta), \ldots, \mathbf{U}_{-a}^{(k)}(\theta)\right), \text{ and } \mathbf{V}_{-a}(\theta) = \left(\mathbf{V}_{-a}^{(1)}(\theta), \ldots, \mathbf{V}_{-a}^{(k)}(\theta)\right).$$

Let $\mathbf{H}_a(\theta) = [\mathbf{U}_{-a}(\theta), \mathbf{V}_{-a}(\theta)]$. Considering the node a from 1 to p, we focus on the large matrix $\mathbf{H}(\theta) = \text{bdiag}\{\mathbf{H}_1(\theta), \ldots, \mathbf{H}_p(\theta)\}$.

Lemma 1. *The estimated edge set $\widehat{E}(\theta)$ defined in Procedure 1 depends only on $\mathbf{H}(\theta)$.*

This lemma suggests that the statistical $\mathbf{H}(\theta)$ is crucial to the final selection result $\widehat{E}(\theta)$. Based on this lemma we would like to sketch the main ideas for deriving the theoretical guarantee on asymptotic FDR control of our procedure. According to the lemma and the definition of FDR, the FDR can be written as $\mathbb{E}[\text{FDP}\{\mathbf{H}(\theta)\}]$. Note that if $\mathbf{H}(\theta)$ is replaced by $\mathbf{H}(\theta_0)$, which is formed by ideal knockoff matrices, the FDR of our procedure with true precision matrices, i.e. $\mathbb{E}[\text{FDP}\{\mathbf{H}(\theta_0)\}]$, is perfectly controlled to be no larger than q (See Lemma S1.3 in the Supplementary Material for details). Intuitively, for a sequence of estimated precision ma-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

trices $\{\widehat{\mathbf{\Omega}}_{-a}^{(t)}\}_{1\leq a\leq p,1\leq t\leq k} := \hat{\theta}$, the $\mathbb{E}[\mathrm{FDP}\{\mathbf{H}(\hat{\theta})\}]$ is close to $\mathbb{E}[\mathrm{FDP}\{\mathbf{H}(\theta_0)\}]$ if the $\mathbf{H}(\hat{\theta})$ is asymptotically equivalent to $\mathbf{H}(\theta_0)$ with large probability.

Nevertheless, note that $\mathrm{FDP}(\cdot)$ is a discontinuous function, which makes it challenging to establish the convergence of its expectation. Similar to that in Fan et al. (2020) we also need an algorithmic stability assumption to remedy the issue caused by the discontinuity of $\mathrm{FDP}(\cdot)$. Drawing on the analytical techniques in Fan et al. (2020), we will focus on a subspace of $\mathbf{H}(\theta)$ to facilitate introducing the algorithmic stability assumption. For any subset $\mathcal{A}_a \subset [-a]$, let $\mathbf{H}_{\mathcal{A}_a}(\theta) = [\mathbf{U}_{\mathcal{A}_a}(\theta), \mathbf{V}_{\mathcal{A}_a}(\theta)]$, where

$$\mathbf{U}_{\mathcal{A}_a}(\theta) = \left(\mathbf{U}_{\mathcal{A}_a}^{(1)}(\theta), \ldots, \mathbf{U}_{\mathcal{A}_a}^{(k)}(\theta)\right), \text{ and } \mathbf{V}_{\mathcal{A}_a}(\theta) = \left(\mathbf{V}_{\mathcal{A}_a}^{(1)}(\theta), \ldots, \mathbf{V}_{\mathcal{A}_a}^{(k)}(\theta)\right)$$

with

$$\mathbf{U}_{\mathcal{A}_a}^{(t)} = \frac{1}{n}\left[\mathbf{X}_{\mathcal{A}_a}^{(t)}, \widetilde{\mathbf{X}}_{\mathcal{A}_a}^{(t)}(\theta)\right]^\top \left[\mathbf{X}_{\mathcal{A}_a}^{(t)}, \widetilde{\mathbf{X}}_{\mathcal{A}_a}^{(t)}(\theta)\right] \text{ and } \mathbf{V}_{\mathcal{A}_a}^{(t)} = \frac{1}{n}\left[\mathbf{X}_{\mathcal{A}_a}^{(t)}, \widetilde{\mathbf{X}}_{\mathcal{A}_a}^{(t)}(\theta)\right]^\top \mathbf{X}_a^{(t)}.$$

Let $\mathbf{H}_{\mathcal{A}}(\theta) = \mathrm{bdiag}\left\{\mathbf{H}_{\mathcal{A}_1}(\theta), \ldots, \mathbf{H}_{\mathcal{A}_p}(\theta)\right\}$. For the sequence of $\{\mathcal{A}_a\}_{a=1}^p$, we simply denote it as $\mathcal{A}$, that is $\{\mathcal{A}_a\}_{a=1}^p := \mathcal{A}$. Define a mapping $E_{\mathcal{A}}(\mathbf{H}_{\mathcal{A}}(\theta))$ which represents the outcome of first restricting ourselves to the smaller set of neighbors indexed by $\{\mathcal{A}_a\}_{a=1}^p$, and then applying our procedure to $\mathbf{H}_{\mathcal{A}}(\theta)$ to further select neighbors of each node $1 \leq a \leq p$ from set $\mathcal{A}_a$. The following Lemma 2 provides the foundation to simplify our theoretical analysis into a lower-dimensional space.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Lemma 2. *Under Conditions 1-3, for any sequence of* $\{\mathcal{A}_a\}_{a=1}^p$ *that satisfies* $\mathcal{A}_a \supset \mathcal{A}_a^*(\theta)$ *of each node* $1 \le a \le p$*, we have* $E_{\mathcal{F}}(\mathbf{H}(\theta)) = E_{\mathcal{A}}(\mathbf{H}_{\mathcal{A}}(\theta))$*, where* $\mathcal{F} = \{[-a]\}_{a=1}^p$*, and* $\mathcal{A}_a^*(\theta)$ *denotes the support of knockoff statistics* $\mathbf{W}_a(\theta)$.

A similar lemma was introduced in Fan et al. (2020), demonstrating that when $\mathbf{W}_a(\theta)$ is sparse, the theoretical analysis of our procedure can be simplified to a lower-dimensional space. Condition 4 guarantees the sparsity of $\mathbf{W}_a(\theta)$ for all $1 \le a \le p$. As in Fan et al. (2020), the dimensionality reduction to a smaller model characterized by $\{\mathcal{A}_a\}_{a=1}^p$ serves to facilitate theoretical analysis, and our procedure does not require any prior knowledge of such a sequence $\{\mathcal{A}_a\}_{a=1}^p$.

Condition 4. For any $1 \le a \le p$, the HGSL estimator $\widehat{\boldsymbol{\beta}}_a^{\text{aug}}(\theta)$ satisfies $|\{l : 1 \le l \ne a \le p, \widehat{\boldsymbol{\beta}}_{a(l)}(\theta) \ne \mathbf{0} \text{ or } \widetilde{\boldsymbol{\beta}}_{a(l)}(\theta) \ne \mathbf{0}\}| < d/2$ for some positive integer $d \le n \wedge p$ which may diverge with n.

This condition puts a constraint on the group sparsity of the HGSL solution, which means that the sparsity level of $\mathbf{W}_a(\theta)$ is no larger than $d/2$. It is mild and can always be achieved since users have the freedom to choose the size of the HGSL model. Similar constrain and justifications are provided in Fan et al. (2020) for Lasso solution.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

For any given $\{\mathcal{A}_a\}_{a=1}^p$, it's convenient to define $\mathbf{H}^0_{\mathcal{A}_a} = \mathbf{H}_{\mathcal{A}_a}(\theta_0) \in \mathbb{R}^{d_{a1}\times d_{a2}}$, where $d_{a1} = 2|\mathcal{A}_a|$, and $d_{a2} = 2k|\mathcal{A}_a| + k$. Denote by

$$\mathbb{I}_{\mathcal{A}} = \left\{ \mathcal{H} = \text{bdiag}\{\mathbf{R}_1, \ldots, \mathbf{R}_p\} \text{ with } \mathbf{R}_a \in \mathbb{R}^{d_{a1}\times d_{a2}} : \max_{1\le a\le p} \|\mathbf{R}_a - \mathbf{H}^0_{\mathcal{A}_a}\|_2 \le a_{np} \right\},$$

where $a_{np} \to 0$ as $n \to \infty$.

Condition 5. (Algorithmic stability). For any sequence of $\{\mathcal{A}_a\}_{a=1}^p$ with $\mathcal{A}_a \subset [-a]$ for all $1 \le a \le p$ that satisfy $\max_{a=1}^p |\mathcal{A}_a| \le d \le n \wedge p$, there exists a positive sequence $\rho_{np} \to 0$ as $n \wedge p \to \infty$ such that

$$\sup_{\max_{a=1}^p |\mathcal{A}_a|\le d} \sup_{\mathcal{H}_1,\mathcal{H}_2\in\mathbb{I}_{\mathcal{A}}} \frac{|E_{\mathcal{A}}(\mathcal{H}_1)\triangle E_{\mathcal{A}}(\mathcal{H}_2)|}{|E_{\mathcal{A}}(\mathcal{H}_1)| \wedge |E_{\mathcal{A}}(\mathcal{H}_2)|} = O(\rho_{np}),$$

where $\triangle$ stands for the symmetric difference between two sets.

Intuitively the above condition assumes that the knockoff procedure is stable with respect to a small perturbation to the input $\mathcal{H}$ in any lower-dimensional subspace $\mathbb{I}_{\mathcal{A}}$. A similar condition is proposed in Fan et al. (2020) to establish the asymptotic FDR control for the high-dimensional linear regression model. Although there are p regressions in our model, we control the overall perturbation by limiting the parameters to the uniformly convergent space, so as to ensure the stability of the algorithm.

Theorem 1. *(Robust FDR control) Assume that Conditions 1-5 hold and the smallest eigenvalue of* $2diag\{\mathbf{s}_a^{(t)}\} - diag\{\mathbf{s}_a^{(t)}\}\mathbf{\Omega}^{(t)}_{-a}diag\{\mathbf{s}_a^{(t)}\}$ *is uniformly*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

bounded from below by some positive constant for all $1 \leq a \leq p$, $1 \leq t \leq k$. *If* $kb_n = o(1)$ *(*b_n *appears in Condition 3) and* $\log(p) = o(n)$, *for any pre-specified FDR level* $q \in (0, 1)$ *it holds that*

$$\mathrm{FDR}(\widehat{E}(\widehat{\theta})) = \mathbb{E}\left[\frac{|\widehat{E}(\widehat{\theta}) \cap E^c|}{|\widehat{E}(\widehat{\theta})| \vee 1}\right] \leq q + O(\rho_{np} + p^{-c_\delta})$$

where $\widehat{E}(\widehat{\theta})$ *is the estimated edge set obtained by Procedure 1 with estimated precision matrices satisfying Condition 3,* ρ_{np} *defined in Condition 5, and* c_δ *is some positive constant associated with the constant* δ *(Condition 3).*

Theorem 1 establishes the robustness of the FDR control with respect to the estimated precision matrices, which allows the number of graphs k to diverge with n as long as $kb_n \to 0$. To conduct the analysis of the robust FDR control, we generalize the analytical technique introduced by Fan et al. (2020) from a single linear regression model to our graphical models. It's non-trivial since we grapple with p different but correlated linear regression models simultaneously, and the design matrices are composed of multiple data from different distributions due to the heterogeneity. It's more complex to characterize the impact on FDR induced by estimated precision matrices and analyze the overall estimated errors.

Unlike the robust FDR control theory for graphical models discussed in Zhou et al. (2022), our theory benefits from the advantage described in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Fan et al. (2020), which does not require the independence between the estimated precision matrices and the data matrices used in the knockoff procedure. This advantage primarily stems from extending the algorithm stability technique proposed in Fan et al. (2020) to our analysis. Leveraging this condition, we can derive an upper bound for the FDP in the restricted space $\mathbb{I}_A$ that depends only on n and p, allowing us to obtain an upper bound for the FDR without needing the independence property.

3.2 Power analysis

We have established the theorems of FDR control for our procedure. Now, we will look at the other side of the cointhe power. We first impose some basic regularity conditions.

Condition 6. It holds that $\min_{(i,j)\in E} \|\boldsymbol{\beta}_{i(j)}\|_2 \geq \nu_n\{(\log(p)+k)/n\}^{1/2}$, $1 \leq i < j \leq p$, for some slowly diverging sequence $\nu_n \to \infty$ as $n \to \infty$.

Condition 7. There exists some constant $M_4 \in \left(\frac{(\gamma+1)c_\gamma p}{q|E|}, 1\right)$ such that $|\mathcal{S}_a| \geq M_4 l_a$ with $\mathcal{S}_a = \left\{j \in [-a] : \|\boldsymbol{\beta}_{a(j)}\|_2 \gg \left[\frac{l_a\{k+\log(p)\}}{n}\right]^{1/2}\right\}$ for $1 \leq a \leq p$, where $l_a = |\mathrm{NE}_a|$.

Condition 8. Let $l_m = \max\{l_i, 1 \leq i \leq p\}$. It holds that $l_m \ll M_5 n/\log(p)$ where M_5 is a positive constant, and there exists some constant $\alpha \in \left(\frac{(\gamma+1)c_\gamma}{q l_m M_4}, 1\right)$ such that $|E| \geq \alpha p l_m$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Conditions 6 and 7 impose some signal constrains. Similar conditions are also needed in Fan et al. (2020) and Fan et al. (2020) to achieve asymptotic power one. Condition 6 puts a lower bound on the minimal signal strength, which is mild. Consider a special case where the signal strength of $\|\boldsymbol{\beta}_{i(j)}\|_2$ is evenly distributed in k components. Then we only require each component to be greater than $1/\sqrt{n}$ when $k \asymp \log(p)$. Such weak signal strength requirements explain the high power of our method. Condition 7 requires some strong signals in our model. Note that the set of $\mathcal{S}_a$ is only a large enough proper subset of the NE_a which shows that our model still allows for many weak ones, and the magnitude of the strong signal is modest as $\frac{l_a(k+\log(p))}{n} = o(1)$ when $k = O\{\log(p)\}$.

The first part of Condition 8 assumed that $l_m \ll M_4 n/\log(p)$, which is a typical assumption in high-dimensional sparse graphical models (Ren et al., 2015; Fan and Lv, 2016). In addition, Condition (8) puts a lower bound on the number of the edges $|E|$, which requires the cardinality of the true edge set $|E|$ can not be too small. Nevertheless, the flexibility of this condition is evident, as it puts the constrain on the whole graph E rather than the edges set of each node. It permits the model to incorporate isolated nodes with no connections to any other nodes.

Theorem 2. *Assume that Conditions 1-3 and 6-8 hold, and the small-*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

est eigenvalue of $2diag\{\mathbf{s}_a^{(t)}\} - diag\{\mathbf{s}_a^{(t)}\}\mathbf{\Omega}_{-a}^{(t)}diag\{\mathbf{s}_a^{(t)}\}$ *is uniformly bounded from below by some positive constant for all* $1 \leq a \leq p$, $1 \leq t \leq k$. *if* $[\log(p) + \log(k)] = o(n)$ *and* $\{\widehat{\mathbf{\Omega}}_{-a}^{(t)}\}_{1\leq a\leq p, 1\leq t\leq k}$ *are independent of* $\mathbf{X}$, *the procedure 1 with the estimated precision matrices* $\{\widehat{\mathbf{\Omega}}_{-a}^{(t)}\}_{1\leq a\leq p, 1\leq t\leq k}$ *has the power satisfying*

$$\text{Power}(\widehat{E}(\widehat{\theta})) = \mathbb{E}\left[\frac{|\widehat{E}(\widehat{\theta}) \cap E|}{|E|}\right] \geq 1 - C\nu_n^{-1} - p^{-\tilde{c}_\delta} + o(\nu_n^{-1}) \to 1,$$

where C *is some positive constant, and* $\tilde{c}_\delta$ *is some positive constants related to the constant* δ *defined in Condition 3.*

Theorem 2 demonstrates that the asymptotic power guarantee of our procedure. Since parameter ν_n characterizes the signal strength, it is seen that the stronger the signal, the faster the convergence of power to one. To the best of our knowledge, this is the first formal theoretical result on the power of the graphical knockoffs procedure. Similar to the power analysis within the knockoff framework presented in Fan et al. (2020), Barber et al. (2020), and Zhou et al. (2022), our theorem also assumes that the estimated precision matrices are independent of the data matrix used in the knockoff procedure. However, it is important to note that this independence is more of a technical assumption rather than a practical necessity.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

4. Simulation studies

In this section, we conduct some simulation studies to investigate the finite-sample performance of our procedure (MGKF) in terms of FDR and power. For comparison, we apply the HGKF method proposed by Zhou et al. (2022) and the GFC method proposed by Liu (2013) to the whole dataset $\mathbf{X}_{N*p}$ with $N = \sum_{t=1}^{k} n^{(t)}$ stacked by $\{\mathbf{X}^{(t)}\}_{t=1}^{k}$ regardless heterogeneity among the different subgroups.

Throughout all numerical studies, we use ISEE (Fan and Lv, 2016) to obtain the estimated precision matrices $\widehat{\theta}$ then run our procedure, and the heterogeneous group scale-root Lasso used in our procedure is implemented using R packages `HGSL` with the suggested tuning parameter $c\sqrt{\frac{k+2\log p+2\sqrt{k\log p}}{n}}$, $c > 1$. Here, we use AIC to determine a suitable c from the sequence $[2, 4, 6, 8, 10]$ in our simulations. All codes including the implementation of MGKF and HGKF are available on GitHub: `https://github.com/zhoujia66/MGKF`. For the GFC method, we use the R-package `SILGGM` developed by Zhang et al. (2018) with scaled Lasso estimator and default values of the tuning parameters.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

4.1 Simulation study 1

We first consider a simple case where the k different graphs share the same support structure, and the heterogeneity lies in the link strength and noise level. With reference to the settings in Li and Maathuis (2021) and Ren et al. (2019), let $\mathbf{\Gamma}^{(t)0} = (\mathbf{\Gamma}^{(t)0}_{i,j})_{1\leq i,j\leq p}$ be a block diagonal matrix with m blocks. Each block represents a fully connected graph of size 20. The diagonal entries of the blocks are 1 and off-diagonal entries are generated independently from a uniform distribution over $[-0.8,-0.3]\cup[0.3,0.8]$. Further, to make the graph structure more general, we randomly permute the rows of columns of $\mathbf{\Gamma}^{(t)0}$ to obtain the matrix $\widetilde{\mathbf{\Gamma}}^{(t)0}$ (Note that for $1\leq t\leq k$, the permutation of $\mathbf{\Gamma}^{(t)0}$ is the same). The precision matrix of each subgroup is eventually given by $\mathbf{\Gamma}^{(t)} = \widetilde{\mathbf{\Gamma}}^{(t)0} + \left[|\lambda_{\min}(\widetilde{\mathbf{\Gamma}}^{(t)0})| + 0.3\right]\mathbf{I}_p$, which ensures the positive definite of the precision matrix.

For each subgroup, let $n^{(t)} = n$ for $1\leq t\leq k$, and the rows of the $n\times p$ data matrix $\mathbf{X}^{(t)}$ are i.i.d copies of $N(\mathbf{0},(\mathbf{\Gamma^{(t)}})^{-\mathbf{1}})$. In all simulations, we set the target FDR level at $q = 0.2$. To fully investigate the performance of our approach, we consider different combinations of (n,p,k).

- Case 1: Let $k = 3$ and $p = 200$ be fixed, while the sample size n for each subgroup varies between 400 and 700.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

- Case 2: Let $p = 100$ and $n = 200$ be fixed, while the number of subgroups k varies between 3 and 9.

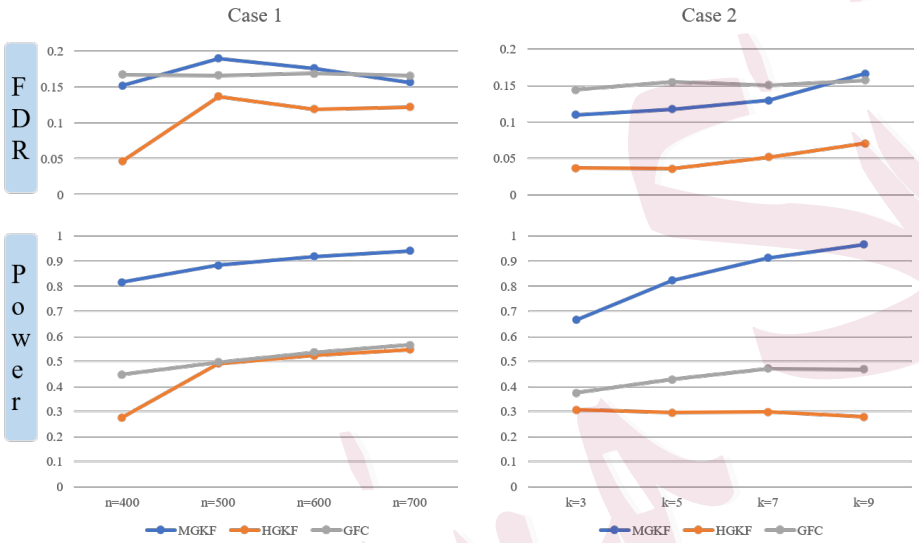

Figure 1: The empirical FDR and power of different procedures over 100 replications with $q = 0.2$.

Figure 1 shows that our procedure can control the FDR under the pre-specified level meanwhile enjoying higher power compared to HGKF and GFC methods. While the empirical FDR results of HGKF and GFC also fall below the pre-specified level, it suffers from significant power loss, a circumstance that is comprehensible. Note that the signal strength of an edge (i, j) on different subpopulations is not the same or even has different signs. Using a single regression coefficient to fit the strength can cause seri-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

ous estimation errors, resulting in relatively large thresholds in the process of controlling FDR for the HGKF and GFC. In addition, the interaction of positive and negative signals on different subgroups will make the estimated coefficient with a very small absolute value, thus losing the edge. Therefore, we can see that it makes sense to carefully deal with the heterogeneity for the heterogeneous data.

4.2 Simulation study 2

We continue to investigate the performance of our procedure under a more flexible sparsity pattern where the connection structure of k graphs is not exactly the same. Here, we employ a different data-generating scheme for entries inside the diagonal blocks. Specifically, for each entry (i, j) with $i \neq j$ inside a diagonal block $\mathbf{\Gamma}_{i,j}^{(t)0} = \gamma_{i,j}\phi_{i,j}$, where $\gamma_{i,j}$ is generated independently from the uniform distribution over $[-0.8, -0.3] \cup [0.3, 0.8]$ and $\phi_{i,j}$ is generated independently form Bernoulli (4/5). The other settings are the same as in simulation study 1. Obviously, in this way, the link structures of different graphs are not the same.

Similar to simulation study 1, here we also consider different combinations of (n, p, k) as follows:

- Case 1: Let $k = 3$ and $p = 200$ be fixed, while the sample size n for

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

each subgroup varies between 200 and 500.

- Case 2: Let $p = 100$ and $n = 200$ be fixed, while the number of subgroups k varies between 3 and 9.

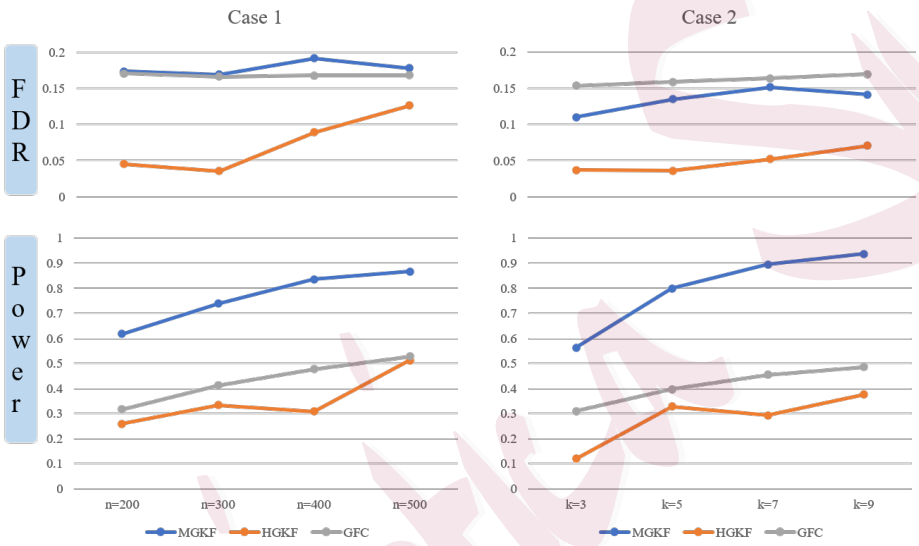

Figure 2: The empirical FDR and power of two procedures over 100 replications with $q = 0.2$.

Figure 2 shows a similar phenomenon to that in Figure 1, where our procedure can control the FDR under the pre-specified level and has an overwhelming advantage in terms of power compared to HGKF and GFC.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

5. Real data analysis

In addition to simulation examples presented in Section 4, we also demonstrate the practical utility of our MGKF procedure on a gene expression dataset of breast invasive ductal carcinoma. This dataset consists of 22605 gene expression levels of 1575 patients which is publicly available on the METABRIC repository (`https://www.cbioportal.org/study/summary?id=brca_metabric`) . It would be interesting to investigate the connectivity pattern among mutated genes of this breast cancer, which provides a stepping stone to understanding how genes affect cellular phenotypes.

Note that this dataset actually contains four subgroups corresponding to four molecular subtypes of breast cancer, namely, luminal A, luminal B, HER2-enriched, and basal-like. Previous studies (Johnson et al., 2021) have shown that these four molecular subtypes have critical differences in incidence, response to treatment, disease progression, survival, and imaging features. It would be more realistic to assume that the distribution of gene expression levels can vary from one subtype to another, which results in the heterogeneity of the whole dataset. Meanwhile, they may also share some common structure as they all belong to the breast cancer.

In this application, we select the top 100 signature genes with the highest frequency of mutation in breast invasive ductal carcinoma. Subsequent-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

ly, we eliminate samples with missing values, resulting in subsets comprising 485, 430, 180, and 266 patient samples, delineated according to four distinct subtypes based on three established indicators: ER, HER2, and Ki-67. Then, we apply our method to this refined dataset with $K = 4$. For comparison, we also apply methods of HGKF and GFC to the whole dataset regardless the possible heterogeneity. We set the target FDR values at 0.1 for all methods.

The results show that our method identified 1056 edges, GFC identified 503 edges and HGKF identified none. This aligns with the observed power performances of the three methods as demonstrated in the preceding simulations. The gene networks reconstructed by our method and GFC are displayed in supplementary material. Notably, among the 503 edges identified by the GFC method, a remarkable 440 edges were also identified by our method. Furthermore, our approach reveals over 600 additional edges compared to GFC. Several of these edges bear substantial biological significance, as evidenced by previous studies, demonstrating the high power of our method. In the following, we will take TP53 gene as a representative to carry out specific analysis.

Focusing on TP53 gene which encodes the important protein p53, the GFC method identify 6 genes which are connected to TP53 while our meth-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

ods identify 19. For convenience, we tabulate the connected genes identified by two methods (Table 1).

Table 1: Edges identify connected with TP53.

GFC	PDE4DIP, NOTCH1, NCOR1, MAP2K4, SETD1A, PIK3R1
MGKF	PDE4DIP, NOTCH1, NCOR1, MAP2K4, SETD1A, PIK3R1, AKT1, FANCD2, TAF1, PBRM1, KDM3A, JAK1, SETDB1, FAM20C, MYH9, PTPRD, LAMB3, SF3B1, MYO3A,

In view of Table 1, we can see that all genes identified by GFC are also detected by our methods. Moreover, our approach uncovers an additional 13 edges, several of which have been corroborated by pertinent studies. For instance, Ogawara et al. (2002) shows that phosphorylation of MDM2 by Akt results in the translocation of MDM2 to the nucleus, where it promotes the ubiquitination of p53. Akter et al. (2021) has pointed out that the loss of p53 will induce Fanconi anemia group D2 protein (FANCD2) with ATRX deficiency. Additionally, Wu et al. (2014) has report that TAF1 phosphorylates p53 at Thr55, leading to dissociation of p53 from the p21 promoter and inactivation of transcription late in the DNA damage response.

The associations of TP53 with PBRM1, KDM3A, and JAK1 also appear plausible, as supported by Cai et al. (2020), Li et al. (2015), and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

Goyal et al. (2020), respectively. Although some studies may draw insights from diverse cancer types like stomach and kidney cancers, their findings can offer valuable insights into the TP53-gene connections identified within breast cancer. While some edges identified by our method have received limited attention in the current literature, it would be interesting to further investigate if such edges are biologically meaningful.

6. Discussion

In this paper, we present a novel procedure for learning the connected structure of a population from heterogeneous datasets. To our knowledge, this is the first study to extend the knockoff framework to multiple graphical models, tackling the complex issue of heterogeneity in reproducible learning. It is worth pointing out that our method will facilitate the further investigation of the heterogeneity among different subgroups. To be specific, one can use our method as a preliminary screening tool, and then recovering the connected structure for each subpopulation can be more efficient based on a reduced-dimensional space. Moreover, our work has focused on multiple Gaussian graphical models. It would be interesting to extend our idea to other multiple graphical models. The possible extensions addressing these issues are beyond the scope of the current article and will be interesting

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

topics for future research.

Supplementary Materials

Supplementary materials available online include four auxiliary lemmas, the proofs for all lemmas and Theorems 1-2, and two figures of real data analysis mentioned in Section 5.

Acknowledgments

We thank the editor, associate editor, and referees for their insightful comments. Zheng's research is supported by the National Key Research and Development Program of China (Grant No. 2022YFA1008000). Pan's research is supported by the Ministry of Education, Singapore (Grant No. MOE-T2EP20123-0007). Zhou's research is supported by the Natural Science Foundation of Hefei University of Technology (Grant No. JZ2023HGQA0085).

References

Akter, J., Y. Katai, P. Sultana, H. Takenobu, M. Haruta, R. P. Sugino, K. Mukae, S. Satoh, T. Wada, M. Ohria, K. Ando, and T. Kamijo (2021). Loss of p53 suppresses replication stress-induced dna damage in atrx-deficient neuroblastoma. *Oncogenesis 10*(73), 1–12.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

REFERENCES

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist. 43*(5), 2055–2085.

Barber, R. F., E. J. Candès, and R. J. Samworth (2020). Robust inference with knockoffs. *Ann. Statist. 48*(3), 1409–1431.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser.* B *57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29*(4), 1165–1188.

Cai, T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc. 115*(532), 1861–1872.

Cai, W., L. Su, and H. Yang (2020). Pbrm1 suppresses tumor growth as a novel p53 acetylation reader. *Mol. Cell. Oncol. 7*(3), e1729680.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: model-x knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser.* B *80*(3), 551–577.

Cheng, J., T. Li, E. Levina, and J. Zhu (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist. 26*(2), 367–378.

Dai, C., B. Lin, X. Xing, and J. S. Liu (2023). False discovery rate control via data splitting. *J. Amer. Statist. Assoc. 118*(544), 2503–2520.

Drton, M. and M. D. Perlman (2007). Multiple testing and error control in gaussian graphical

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

model selection. *Statist. Sci. 22*(3), 430–449.

Fan, Y., E. Demirkaya, G. Li, and J. Lv (2020). RANK: large-scale inference with graphical nonlinear knockoffs. *J. Amer. Statist. Assoc. 115*(529), 362–379.

Fan, Y., Y. Kong, D. Li, and Z. Zheng (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist. 43*(3), 1243–1272.

Fan, Y. and J. Lv (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *Ann. Statist. 44*(5), 2098–2126.

Fan, Y., J. Lv, M. Sharifvaghefib, and Y. Uematsua (2020). Ipad: Stable interpretable forecasting with knockoffs inference. *J. Amer. Statist. Assoc. 115*(532), 1822–1834.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Giudici, P. and S. Alessanfro (2016). Graphical network models for international financial flows. *J. Bus. Econ. Stat. 34*(1), 128–138.

Goyal, H., I. Chachoua, P. Christian, W. Vainchenker, and S. N. Constantinescu (2020). A p53-jak-stat connection involved inmyeloproliferative neoplasm pathogenesis and progression to secondary acute myeloid leukemia. *Blood Rev. 42*, 100712.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

Johnson, K. S., E. F. Conant, and M. S. Soo (2021). Molecular subtypes of breast cancer: A

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

review for breast radiologists. *Journal of Breast Imaging 3*(1), 12–24.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, New York.

Lee, W. and Y. Liu (2015). Joint estimation of multiple precision matrices with common structures. *J. Mach. Learn. Res. 16*(1), 1035–1062.

Li, J. and M. H. Maathuis (2021). GGM knockoff filter: False discovery rate control for Gaussian graphical models. *J. R. Stat. Soc. Ser.* B *83*(3), 534–558.

Li, W., S. Lin, W. Wang, X. Li, and D. Xu (2015). Kdm3a interacted with p53k372me1 and regulated p53 binding to puma in gastric cancer. *Biochem. Bophys. Res. Commun. 467*(3), 556–561.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist. 41*(6), 2948–2978.

Liu, W., Y. Ke, J. Liu, and R. Li (2022). Model-free feature screening and FDR control with knockoff features. *J. Amer. Statist. Assoc. 117*(537), 428–443.

Ma, J. and G. Michailidis (2016). Joint structural estimation of multiple graphical models. *J. Mach. Learn. Res. 17*(166), 1–48.

Ogawara, Y., S. Kishishita, T. Obata, Y. Isazawa, T. Suzuki, K. Tanaka, N. Masuyama, and Y. Gotoh (2002). Akt enhances mdm2-mediated ubiquitination and degradation of p53. *J. Biol. Chem. 277*(24), 21843–21850.

Ren, Z., Y. Kang, Y. Fan, and J. Lv (2019). Tuning-free heterogeneous inference in massive

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

network. *J. Amer. Statist. Assoc. 114*(528), 1908–1925.

Ren, Z., T. Sun, C.-H. Zhang, and H. H. Zhou (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist. 43*(3), 991–1026.

Shin, S.-Y., E. B. Fauman, A.-K. Petersen, J. Krumsiek, and et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet. 46*, 543–550.

Wu, Y., J. C. Lin, L. G. Piluso, J. M. Dhahbi, S. Bobadilla, S. R. Spindler, and X. Liu (2014). Phosphorylation of p53 by taf1 inactivates p53-dependent transcription in the dna damage response. *Mol. cell 53*(1), 63–74.

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

Zhou, J., Y. Li, Z. Zheng, and D. Li (2022). Reproducible learning in large-scale graphical models. *J. Multivariate Anal. 189*, 104934.

School of Economics, Hefei University of Technology, Hefei, Anhui, 230009, China

E-mail: tszhjia@mail.ustc.edu.cn

School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 639798, Singapore

E-mail: GMPAN@ntu.edu.sg

School of Management, University of Science and Technology of China, Hefei, Anhui, 230009, China

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0099

E-mail: Zhengzm@ustc.edu.cn

School of Economics, Hefei University of Technology, Hefei, Anhui, 230009, China

E-mail: cctan@hfut.edu.cn