Differential Analysis of Directed Networks

Min Ren
Department of Statistics
Purdue University
West Lafayette, IN
ren80@purdue.edu

Dabao Zhang
Department of Statistics
Purdue University
West Lafayette, IN
zhangdb@purdue.edu

Abstract

We developed a novel statistical method to identify structural differences between networks characterized by structural equation models. We propose to reparameterize the model to separate the differential structures from common structures, and then design an algorithm with calibration and construction stages to identify these differential structures. The calibration stage serves to obtain consistent prediction by building the ℓ_2 regularized regression of each endogenous variables against pre-screened exogenous variables, correcting for potential endogeneity issue. The construction stage consistently selects and estimates both common and differential effects by undertaking ℓ_1 regularized regression of each endogenous variable against the predicts of other endogenous variables as well as its anchoring exogenous variables. Our method allows easy parallel computation at each stage. Theoretical results are obtained to establish non-asymptotic error bounds of predictions and estimates at both stages, as well as the consistency of identified common and differential effects. Our studies on synthetic data demonstrated that our proposed method performed much better than independently constructing the networks. A real data set is analyzed to illustrate the applicability of our method.

1 INTRODUCTION

It is of great importance and interest to detect sparse structural differences or differential structures between two cognate networks. For instance, the gene regulatory networks of diseased and healthy individuals may differ slightly from each other [West et al., 2012], and identifying the subtle difference between them helps design specific drugs. Social networks evolve over times, and monitoring their abrupt changes may serve as surveillance to economic stability or disease epidemics [Pianese et al., 2013, Berkman and Syme, 1979]. However, addressing such practical problems demands differential analysis of large networks, calling for development of efficient statistical method to infer and compare complex structures from high dimensional data. In this paper, we focus on differential analysis of directed acyclic or even cyclic networks which can be described by structural equation models (SEMs).

Many efforts have been made towards construction of a single network via SEM. For example, both Xiong et al. [2004] and Liu et al. [2008] employed a genetic algorithm to search for the best SEM using different information criteria. Most recently, Ni et al. [2017, 2018] employed a hierarchical Bayes approach to construct the SEM based networks. However, these approaches were designed for small or medium scale networks. For large-scale networks that the number of endogenous variables p exceeds the sample size n, Cai et al. [2013] proposed a regularization approach to fit a sparse model. Because this method suffers from incapability of parallel computation, it may not be feasible for large networks. Logsdon and Mezey [2010] proposed another penalization approach to fit the model in a node-wise fashion which alleviates the computational burden. Most recently, Lin et al. [2015], Zhu [2018], and Chen et al. [2017] each proposed a two-stage approach to construct SEMs, with different algorithms designed at different stages. As shown by Chen et al. [2017], such a two-stage approach can have superior performance compared to other methods.

To the best of our knowledge, no algorithm has been proposed to conduct differential analysis of directed networks characterized by SEM. While a naive approach would separately construct each individual network and identify common and differential structures, this ap-

proach fails to take advantage of the commonality as well as sparse differential structures of the paired networks, leading to higher false positive rate or lower power. In this light, we introduce a novel statistical method, specially in the directed network regime, to conduct differential analysis of two networks via appropriate reparameterization of the corresponding models. There are two major features of our method. Firstly, we jointly model the commonality and difference between two networks explicitly. This helps us to gain dramatic performance improvements over the naive construction method. Secondly, benefiting from the flexible framework of SEMs, we are able to conduct differential analysis of directed networks. Most importantly, our method allow for both acyclic and cyclic networks. Compared to the other methods, directionality and allowing for cyclicity are crucial for many network studies, especially in constructing gene regulatory networks. As far as we know, our method is the first work on differential analysis of directed networks that enjoys the two promising features.

The rest of this paper is organized as follows. We first introduce the model and its identifiability condition in Section 2.1 and Section 2.2, respectively. Then, we present our proposed method of **Re**parameterization-based **D**ifferential analysis of directed **Net**works, termed as **ReDNet**, in Section 2.3. The theoretical justification of the proposed method is described in Section 2.4. Section 3 includes our studies on synthetic data showing the superior performance of our method, as well as an analysis of the Genotype-Tissue Expression (GTEx) data sets. We conclude our paper with brief discussion in Section 4.

2 METHODS

Here we first introduce the model and its identification condition, and then describe our proposed **ReDNet** method for identifying common and differential structures between two directed networks, followed with its theoretical justification.

2.1 THE MODEL

We consider two networks, each describing the dependencies among a common set of variables or nodes in a unique population. For each node $i \in \{1, 2, \ldots, p\}$ in network $k \in \{1, 2\}$, its regulation structure can be represented by the following equation,

$$\underbrace{\mathbf{Y}_i^{(k)}}_{\text{node } i} = \underbrace{\mathbf{Y}_{-i}^{(k)}\boldsymbol{\gamma}_i^{(k)}}_{\text{regulation by others}} + \underbrace{\mathbf{X}^{(k)}\boldsymbol{\phi}_i^{(k)}}_{\text{anchoring regulation}} + \underbrace{\boldsymbol{\epsilon}_i^{(k)}}_{\text{error}}, \quad (1)$$

where $\mathbf{Y}_i^{(k)}$ is the i-th column of $\mathbf{Y}^{(k)}$ and $\mathbf{Y}_{-i}^{(k)}$ is the submatrix of $\mathbf{Y}^{(k)}$ by excluding $\mathbf{Y}_i^{(k)}$, with $\mathbf{Y}^{(k)}$ a $n^{(k)} \times p$ matrix. $\mathbf{X}^{(k)}$ is a $n^{(k)} \times q$ matrix with each column standardized to have ℓ_2 norm $\sqrt{n^{(k)}}$. The vectors $\boldsymbol{\gamma}_i^{(k)}$ and $\boldsymbol{\phi}_i^{(k)}$ encode the inter-nodes and anchoring regulatory effects, respectively. The index set of non-zeros of $\boldsymbol{\phi}_i^{(k)}$ is known and denoted by $\mathcal{A}_i^{(k)}$, in other words, $\mathcal{A}_i^{(k)} = \text{supp}(\boldsymbol{\phi}_i^{(k)})$. The support set $\mathcal{A}_i^{(k)}$ indexes the direct causal effects for the i-th node, and can be prespecified based on the domain knowledge. However, the size of nonzero effect $\boldsymbol{\phi}_i^{(k)}$ is unknown and can be estimated. Further property of $\mathcal{A}_i^{(k)}$ will be discussed in Section 2.2. All elements of the error term are independently distributed following a normal distribution with mean zero and standard deviation $\sigma_i^{(k)}$. We assume that the matrix $\mathbf{X}^{(k)}$ and the error term $\boldsymbol{\epsilon}_i^{(k)}$ are independent of each other. However $\mathbf{Y}_{-i}^{(k)}$ and $\boldsymbol{\epsilon}_i^{(k)}$ may correlate with each other. $\mathbf{Y}^{(k)}$ and $\mathbf{X}^{(k)}$ include observed endogenous variables and exogenous variables, respectively.

By combining the p linear equations in (1), we can rewrite the two sets of linear equations in a systematic fashion as two structural equation models below,

$$\begin{cases} \mathbf{Y}^{(1)} = \mathbf{Y}^{(1)}\boldsymbol{\Gamma}^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\Phi}^{(1)} + \mathcal{E}^{(1)}, \\ \mathbf{Y}^{(2)} = \mathbf{Y}^{(2)}\boldsymbol{\Gamma}^{(2)} + \mathbf{X}^{(2)}\boldsymbol{\Phi}^{(2)} + \mathcal{E}^{(2)}, \end{cases} \quad (2)$$

where each matrix $\boldsymbol{\Gamma}^{(k)}$ is $p \times p$ with zero diagonal elements and represents the inter-nodes regulatory effects in the corresponding network. Specifically, excluding the i-th element (which is zero) from the i-th column of $\boldsymbol{\Gamma}^{(k)}$ leads to $\boldsymbol{\gamma}_i^{(k)}$. The $q \times p$ matrix $\boldsymbol{\Phi}^{(k)}$ contains the anchoring regulatory effects and its i-th column is $\boldsymbol{\phi}_i^{(k)}$. Each error term $\mathcal{E}^{(k)}$ is $n^{(k)} \times p$ and has the error term $\boldsymbol{\epsilon}_i^{(k)}$ as its i-th column.

Figure 1 gives an illustrative example of networks with three nodes and one anchoring regulation per node for the structural equations in (2). For example, with anchoring regulation on node Y_1, X_1 has a direct effect on node Y_1 but indirect effects on node Y_2 and Y_3 via Y_1.

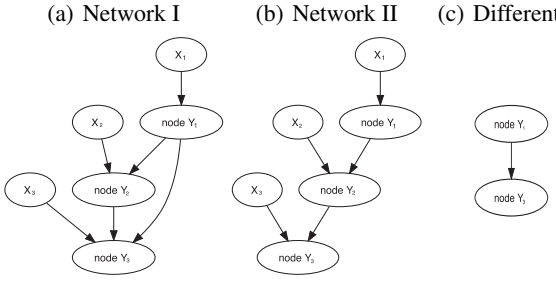

Figure 1: An Illustrative Example of Differential Network Between Two Directed Networks. The error term for each node is not shown for simplicity.

For each network k, its full model in (2) can be further transformed into the reduced form as follows,

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\pi}^{(k)} + \boldsymbol{\xi}^{(k)}, \tag{3}$$

where the $q \times p$ matrix $\boldsymbol{\pi}^{(k)} = \boldsymbol{\Phi}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$ and the transformed error term $\boldsymbol{\xi}^{(k)} = \mathcal{E}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$. The reduced model (3) reveals variables observed in $\mathbf{X}^{(k)}$ as instrumental variables which will be used later to correct for the endogeneity issue. Otherwise, directly applying any regularization based regression to equation (1) will result in non-consistent or suboptimal estimation of model parameters [Fan and Liao, 2014, Chen et al., 2017, Lin et al., 2015, Zhu, 2018].

2.2 THE MODEL IDENTIFIABILITY

Here we introduce an identifiability assumption which helps to infer an identifiable system (2) from available data. We assume that each endogenous variable is directly regulated by a unique set of exogenous variables as long as it regulates other endogenous variables. That is, any regulatory node needs at least one anchoring exogenous variable to distinguish the corresponding regulatory effects from association. Explicitly let $\mathcal{M}_{i0}^{(k)}$ denote the index set of endogenous variables which either directly or indirectly regulate the i-th endogenous variable in the k-th network. Thus, $\mathcal{A}_i^{(k)} \subseteq \mathcal{M}_{i0}^{(k)}$. The model identification condition can be stated in the below.

Assumption 1. For any $i = 1, \cdots, p$, $\mathcal{A}_i^{(k)} \neq \emptyset$ if there exists j such that $i \in \mathcal{M}_{j0}^{(k)}$. Furthermore, $\mathcal{A}_i^{(k)} \cap \mathcal{A}_j^{(k)} = \emptyset$ as long as $i \neq j$.

This assumption is slightly less restrictive than the one employed by Chen et al. [2017], and is a sufficient condition for model identifiability as it satisfies the rank condition in Schmidt [1976]. It can be further relaxed to allow nonempty $\mathcal{A}_i^{(k)} \cap \mathcal{A}_j^{(k)}$ as long as each regulatory node has its own unique anchoring exogenous variables.

The above identifiability assumption not only identifies $\boldsymbol{\gamma}_i^{(k)}$ in model (1) from $\boldsymbol{\pi}^{(k)}$ in model (3) but also helps reveal regulatory directionality of the networks. As illustrated in Figure 2, we can not recover the directionality between nodes Y_1 and Y_2 without the extra information provided by the direct causal factors X_1 and X_2 because all four sub-networks consisting of Y_1 and Y_2 (without X_1 and X_2) will be Markov equivalent. The known set $\mathcal{A}_j^{(k)}$ serves as external prior knowledge which helps recover the directionality. In our two-stage construction of the differential network, the additional anchors X_1 and X_2 serve as instrumental variables in the calibration stage, since both X_1 and X_2 are independent of the error terms. The present direct causal effects from $\mathbf{X}^{(k)}$ together with Assumption 1 differentiates our approach from the classical graphical models [Meinshausen and Bühlmann, 2006, Yuan and Lin, 2007] or the PC algorithm approaches [Spirtes et al., 2000, Kalisch and Bühlmann, 2007], since those methods either cannot recover edge directions or do not allow for cyclic structures due to lack of additional direct causal effects from $\mathbf{X}^{(k)}$.

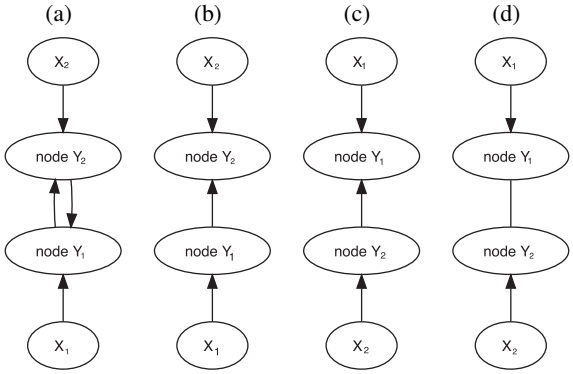

Figure 2: An Illustrative Example of Networks Which Are Not Markov Equivalent. However, without X_1 and X_2, sub-networks consisting of only node Y_1 and Y_2 will be Markov equivalent.

2.3 TWO-STAGE DIFFERENTIAL ANALYSIS OF NETWORKS

Here we intend to develop a regularized version of the two-stage least squares. We first screen for exogenous variables and conduct ℓ_2 regularized regression of each endogenous variable against screened exogenous variables to obtain its good prediction which helps address the endogeneity issue in the following stage. At the second stage, we reparametrize the model to explicitly model the common and differential regulatory effects and identify them via the adaptive lasso method.

2.3.1 The Calibration Stage

To address the endogeneity issue, we aim for good prediction of each endogenous variable following the reduced model in (3). However, in the high-dimensional setting, the dimension q of $\mathbf{X}^{(k)}$ can be much larger than the sample size $n^{(k)}$, and any direct prediction with all exogenous variables may not produce consistent prediction. Note that both Lin et al. [2015] and Zhu [2018] proposed to conduct variable selection with lasso or its variants and predict with selected exogenous variables. We here instead propose to first screen for exogenous variables with ISIS [Fan and Lv, 2008], and then apply ridge regression to predict the endogenous variables with screened exogenous variables. While variable screening is more robust

and provides higher coverage of true variables than variable selection, its combination with ridge regression puts less computational burden. Furthermore, as shown by Chen et al. [2017], ridge regression performs well in predicting the endogenous variables.

Let $\mathcal{M}_i^{(k)}$ denotes the selected index set for i-th node in k-th network from the variable screening which reduces the dimension from q to $d = |\mathcal{M}_i^{(k)}|$. The *Sure Independence Screening Property* in Fan and Lv [2008] can be directly applied in our case to guarantee that $\mathcal{M}_i^{(k)}$ covers the true set $\mathcal{M}_{i0}^{(k)}$ with a large probability.

Assumption 2. $n^{(1)}$ and $n^{(2)}$ are at the same order, i.e., $n_{\min} = \min(n^{(1)}, n^{(2)}) \asymp n^{(1)} \asymp n^{(2)}$, and $p \asymp q$.

Theorem 1. *Assuming Conditions 1-4 in the supplemental materials which restrict positive $\tilde{\tau}$ and $\tilde{\kappa}$, under Assumption 2, there exists some $\theta \in (0, 1 - 2\tilde{\kappa} - \tilde{\tau})$ such that, when $d = |\mathcal{M}_i^{(k)}| = O((n_{\min})^{1-\theta})$, we have, for some constant $C > 0$,*

$$\mathbb{P}(\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}) = 1 - \mathcal{O}\left(\exp\left\{-\frac{C(n^{(k)})^{1-2\tilde{\kappa}}}{\log(n^{(k)})}\right\}\right).$$

Hereafter we assume that $\mathcal{M}_i^{(k)}$ successfully covers the true set $\mathcal{M}_{i0}^{(k)}$ for convenience of stating the following assumptions and theorems. That is, the probability of successful screening is not incorporated into our assumptions or theorems in the below.

For node i in network k, let $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ denotes the submatrix of $\mathbf{X}^{(k)}$ with prescreened columns which are indexed by $\mathcal{M}_i^{(k)}$. With $\boldsymbol{\pi}_i^{(k)}$ denoting the i-th column of $\boldsymbol{\pi}^{(k)}$, the subvector of $\boldsymbol{\pi}_i^{(k)}$ indexed by $\mathcal{M}_i^{(k)}$ will be simply denoted by $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$ without confusion. Such simplified notations will apply to other vectors and matrices in the rest of this paper.

With d pre-screened exogenous variables, we can apply ridge regression to the model

$$\mathbf{Y}_i^{(k)} = \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} + \boldsymbol{\xi}_i^{(k)}, \tag{4}$$

to obtain the estimates $\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}$ of $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$, and predict $\mathbf{Y}_i^{(k)}$ with $\hat{\mathbf{Y}}_i^{(k)} = \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} \hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}$.

2.3.2 The Construction Stage

With known $\mathcal{A}_i^{(k)}$, we can rewrite model (1) as,

$$\mathbf{Y}_i^{(k)} = \mathbf{Y}_{-i}^{(k)} \boldsymbol{\gamma}_i^{(k)} + \mathbf{X}^{(k)}_{\mathcal{A}_i^{(k)}} \boldsymbol{\phi}^{(k)}_{\mathcal{A}_i^{(k)}} + \boldsymbol{\epsilon}_i^{(k)}. \tag{5}$$

Before we use the predicted $\mathbf{Y}^{(k)}$ to identify both common and differential regulatory effects across the two networks, we first reparametrize the model so as to define differential regulatory effects explicitly,

$$\boldsymbol{\beta}_i^- = \frac{\boldsymbol{\gamma}_i^{(1)} - \boldsymbol{\gamma}_i^{(2)}}{2},\ \boldsymbol{\beta}_i^+ = \frac{\boldsymbol{\gamma}_i^{(1)} + \boldsymbol{\gamma}_i^{(2)}}{2}. \tag{6}$$

Here $\boldsymbol{\beta}_i^-$ represents the **differential regulatory effects** between the two networks. We need compare $\boldsymbol{\beta}_i^+$ with $\boldsymbol{\beta}_i^-$ to identify the **common regulatory effects**, that is, effects of all regulations with nonzero values in $\boldsymbol{\beta}_i^+$ but zero values in $\boldsymbol{\beta}_i^-$.

Note that other differential analysis of networks may suggest a different reparametrization to identify common and differential regulatory effects. For example, in a typical case-control study, we may expect few structures in the case network mutated from the control network. While we are interested in identifying differential structures in the case network, we may be also interested in identifying baseline network structures in the control network. Therefore we may reparametrize the model with the regulatory effects in the control network, as well as the differential regulatory effects defined as the difference of regulatory effects between case and control networks. We want to point out that the method described here still applies and we can also derive similar theoretical results as follows.

Following the reparametrization in (6), we can rewrite model (5) as follows,

$$\begin{pmatrix}\mathbf{Y}_i^{(1)}\\ \mathbf{Y}_i^{(2)}\end{pmatrix} = \begin{pmatrix}\mathbf{Y}_{-i}^{(1)} & \mathbf{Y}_{-i}^{(1)}\\ \mathbf{Y}_{-i}^{(2)} & -\mathbf{Y}_{-i}^{(2)}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_i^+\\ \boldsymbol{\beta}_i^-\end{pmatrix} + \begin{pmatrix}\mathbf{X}^{(1)}_{\mathcal{A}_i^{(1)}} & 0\\ 0 & \mathbf{X}^{(2)}_{\mathcal{A}_i^{(2)}}\end{pmatrix}\begin{pmatrix}\boldsymbol{\phi}^{(1)}_{\mathcal{A}_i^{(1)}}\\ \boldsymbol{\phi}^{(2)}_{\mathcal{A}_i^{(2)}}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}_i^{(1)}\\ \boldsymbol{\epsilon}_i^{(2)}\end{pmatrix}. \tag{7}$$

Denote

$$\mathbf{Y}_i = \begin{pmatrix}\mathbf{Y}_i^{(1)}\\ \mathbf{Y}_i^{(2)}\end{pmatrix},\quad \mathbf{Z}_{-i} = \begin{pmatrix}\mathbf{Y}_{-i}^{(1)} & \mathbf{Y}_{-i}^{(1)}\\ \mathbf{Y}_{-i}^{(2)} & -\mathbf{Y}_{-i}^{(2)}\end{pmatrix},$$
$$\boldsymbol{\beta}_i = \begin{pmatrix}\boldsymbol{\beta}_i^+\\ \boldsymbol{\beta}_i^-\end{pmatrix},\quad \boldsymbol{\epsilon}_i = \begin{pmatrix}\boldsymbol{\epsilon}_i^{(1)}\\ \boldsymbol{\epsilon}_i^{(2)}\end{pmatrix}.$$

Further define the projection matrix for each network,

$$\mathbf{H}_i^{(k)} = I_{n^{(k)}} - \mathbf{X}^{(k)}_{\mathcal{A}_i^{(k)}}\left(\mathbf{X}^{(k)T}_{\mathcal{A}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{A}_i^{(k)}}\right)^{-1}\mathbf{X}^{(k)T}_{\mathcal{A}_i^{(k)}}.$$

Applying the projection matrix $\mathbf{H}_i = \mathrm{diag}\{\mathbf{H}_i^{(1)}, \mathbf{H}_i^{(2)}\}$ to both sides of model (7), we can remove the exogenous variables from the model and obtain,

$$\mathbf{H}_i\mathbf{Y}_i = \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i. \tag{8}$$

Algorithm 1 Reparameterization-Based Differential Analysis of Network (ReDNet)

Input: For $k \in \{1,2\}$, $\mathbf{Y}^{(k)}$, $\mathbf{X}^{(k)}$, index set $\mathcal{A}_i^{(k)}$ for each $i \in \{1,2,\ldots,p\}$. Set $d = O(n_{\min}^{1-\theta})$.

for $i \to 1$ **to** p **do**

Stage 1.a. Screen for a submatrix $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ of $\mathbf{X}^{(k)}$ for $\mathbf{Y}_i^{(k)}$ versus $\mathbf{X}^{(k)}$ and set $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} = \mathbf{X}^{(k)}$ if $q \le n^{(k)}$.

Stage 1.b. Apply ridge regression to regress $\mathbf{Y}_i^{(k)}$ against $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ to obtain prediction $\hat{\mathbf{Y}}_i^{(k)}$.

end for

for $i \to 1$ **to** p **do**

Stage 2. Apply adaptive lasso to regress $\mathbf{H}_i\mathbf{Y}_i$ against $\mathbf{H}_i\hat{\mathbf{Z}}_{-i}$ to obtain coefficients estimate $\hat{\boldsymbol{\beta}}_i$.

end for

Output: The common and differential regulatory effects in $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$.

To address the endogeneity issue, we predict $\mathbf{Z}_{-i}$ by replacing its component $\mathbf{Y}_{-i}^{(k)}$ with the predicted value $\hat{\mathbf{Y}}_{-i}^{(k)}$ from the previous stage, and then regressing $\mathbf{H}_i\mathbf{Y}_i$ against $\mathbf{H}_i\hat{\mathbf{Z}}_{-i}$ with the adaptive lasso to consistently estimate $\boldsymbol{\beta}_i$. That is, an optimal $\boldsymbol{\beta}_i$ can be obtained as,

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i} \left\{ \frac{1}{n} ||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2^2 + \lambda_i \boldsymbol{\omega}_i^T |\boldsymbol{\beta}_i|_1 \right\},$$

where $|\boldsymbol{\beta}_i|_1$ is a vector taking elementwise absolute values of $\boldsymbol{\beta}_i$, $\boldsymbol{\omega}_i$ is the adaptive weights whose components are inversely proportional to the components of an initial estimator of $\boldsymbol{\beta}_i$, and λ_i is the adaptive tuning parameter.

The two-stages algorithm is summarized in Algorithm 1. With the estimator $\hat{\boldsymbol{\beta}}_i$ from the second stage, we can accordingly obtain estimators $\hat{\boldsymbol{\gamma}}_i^{(1)} = \hat{\boldsymbol{\beta}}_i^{+} + \hat{\boldsymbol{\beta}}_i^{-}$ and $\hat{\boldsymbol{\gamma}}_i^{(2)} = \hat{\boldsymbol{\beta}}_i^{+} - \hat{\boldsymbol{\beta}}_i^{-}$.

2.4 THEORETICAL ANALYSIS

As shown in Theorem 1, a screening method like ISIS [Fan and Lv, 2008] can identify $\mathcal{M}_i^{(k)}$ with size $d = O(n_{\min}^{1-\theta})$ which covers the true set $\mathcal{M}_{i0}^{(k)}$ with a sufficiently large probability. For the sake of simplicity and without loss of generality, in the following we assume $\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}$.

We first investigate the consistency of predictions from the first stage. The consistency properties will be characterized by prespecified sequences $f^{(k)} = o(n^{(k)})$ but $f^{(k)} \to \infty$ as $n^{(k)} \to \infty$. We also denote $f_{\max} = f^{(1)} \vee f^{(2)}$, i.e., $\max\{f^{(1)}, f^{(2)}\}$.

The following assumption is required for the consistency properties.

Assumption 3. For each network k, the singular values of $\mathbf{I} - \boldsymbol{\Gamma}^{(k)}$ are positively bounded from below, and there exist some positive constants $c_1^{(k)}$ and $c_2^{(k)}$ such that, for each node i, $\max_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \le c_1^{(k)}$ and $\min_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \ge c_2^{(k)}$. Furthermore, the ridge parameter $\lambda_i^{(k)} = o(n_{\min})$.

For the ease of exposition, we will omit the subscript $\mathcal{M}_i^{(k)}$ from $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ henceforth, and accordingly use $\boldsymbol{\pi}_i^{(k)}$ and $\hat{\boldsymbol{\pi}}_i^{(k)}$ which include the zero components of excluded predictors.

Denote $\mathbf{X} = \mathrm{diag}\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\}$, and

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Y}^{(1)} & \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} & -\mathbf{Y}^{(2)} \end{pmatrix}, \quad \boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{\pi}^{(1)} & \boldsymbol{\pi}^{(1)} \\ \boldsymbol{\pi}^{(2)} & -\boldsymbol{\pi}^{(2)} \end{pmatrix}.$$

We use $\boldsymbol{\Pi}_j$ to denote the j-th column of the matrix $\boldsymbol{\Pi}$ and $\boldsymbol{\pi}_j^{(k)}$ to denote the j-th column of the matrix $\boldsymbol{\pi}^{(k)}$. We also use $\hat{\mathbf{Z}}$ and $\hat{\boldsymbol{\Pi}}$ to denote the prediction of $\mathbf{Z}$ and estimate of $\boldsymbol{\Pi}$, respectively. Note that, with the ridge parameter $\lambda_i^{(k)}$ for the ridge regression taken on node i in network k, we have $r_i^{(k)} = (\lambda_i^{(k)})^2||\boldsymbol{\pi}_i^{(k)}||_2^2/n^{(k)}$ and hence define $r_{\max} = \max_{1\le i\le p}[r_i^{(1)} \vee r_i^{(2)}]$. Then the estimation and prediction losses at the first stage can be summarized in the following theorem.

Theorem 2. *Under Assumptions 1-3, for each $j \in \{1,2,\ldots,2p\}$, there will exist some constant C_1 and C_2 such that, with probability at least $1 - e^{-f^{(1)}} - e^{-f^{(2)}}$,*
1. $||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_2^2 \le C_1 (d \vee r_{\max} \vee f_{\max}) / n_{\min}$;
2. $||\mathbf{X}(\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j)||_2^2 \le C_2 (d \vee r_{\max} \vee f_{\max})$.

The proof is detailed in the supplemental materials.

Note that these two sets of losses can be controlled by the same upper bounds across the two networks with probability at least $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$. Therefore, $f^{(k)}$ can be selected such that $f^{(k)} - log(p) \to \infty$, which will provide a probability approaching one to have the network-wide losses approaching zero.

Furthermore, the dimension p can be divergent up to an exponential order, say $p = e^{n_{\min}^c}$ for some $c \in (0,1)$. We can set $f^{(1)} = f^{(2)} = n_{\min}^{(1+c)/2}$ and, apparently, $f^{(k)} = o(n_{\min})$ but $f^{(k)} - log(p) = n_{\min}^{(1+c)/2} - n_{\min}^c \to \infty$.

Since the ridge parameter $\lambda_i^{(k)} = o(n_{\min})$, $r_i^{(k)} = ||\boldsymbol{\pi}_i^{(k)}||_2^2 \times o(n_{\min})$. Therefore, when all $||\boldsymbol{\pi}_i^{(k)}||_2$ are uniformly bounded, we have $r_{\max} = o(n_{\min})$. Otherwise, the ridge parameter $\lambda_i^{(k)}$ should be adjusted accordingly

to control both estimation and prediction losses.

Before we characterize the consistency of estimated regulatory effects on the second stage, we first introduce the following concept of restricted eigenvalue which is used to present an assumption.

Definition 2.1. *The restricted eigenvalue of a matrix* $\mathbf{A}$ *on an index set* $\mathcal{S}$ *is defined as*

$$\phi_{re}(\mathbf{A}, \mathcal{S}) = \min_{||\delta_{\mathcal{S}^c}||_1 \leq 3||\delta_{\mathcal{S}}||_1} \frac{||\mathbf{A}\delta||_2}{\sqrt{n}||\delta_{\mathcal{S}}||_2}. \quad (9)$$

For the i-th node, we use $\mathcal{S}_i$ to denote the non-zero indices of $\boldsymbol{\beta}_i$, i.e., $\mathcal{S}_i = \text{supp}(\boldsymbol{\beta}_i)$. Further denote

$$\boldsymbol{\Pi}_{-i} = \begin{pmatrix} \boldsymbol{\pi}_{-i}^{(1)} & \boldsymbol{\pi}_{-i}^{(1)} \\ \boldsymbol{\pi}_{-i}^{(2)} & -\boldsymbol{\pi}_{-i}^{(2)} \end{pmatrix}.$$

As in Bickel et al. [2009], we impose the following restricted eigenvalue condition on the design matrix in (8).

Assumption 4. There exists a constant $\boldsymbol{\phi}_0 > 0$ such that $\phi_{\text{re}}(\mathbf{H}_i\mathbf{X}\boldsymbol{\Pi}_{-i}, \mathcal{S}_i) \geq \boldsymbol{\phi}_0$.

Let $n = n^{(1)} + n^{(2)}$, $c_{\max} = c_1^{(1)} \vee c_1^{(2)}$, and $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p]$. The matrix norms $||\cdot||_1$ and $\|\cdot\|_\infty$ are the maximum of column and row sums of absolute values of the matrix, respectively. For a vector, we define $\|\cdot\|_\infty$ and $\|\cdot\|_{-\infty}$ to be the maximum and minimum absolute values of its components. Then, we can derive the following loss bounds for the estimation and prediction at the second stage on the basis of Theorem 2.

Theorem 3. *Suppose that, for node* i*, the adaptive lasso at the second stage takes the tuning parameter* $\lambda_i \asymp \|\boldsymbol{\omega}_i\|_{-\infty}^{-1}||\mathbf{B}||_1||\boldsymbol{\Pi}||_1\sqrt{(d \vee r_{\max} \vee f_{\max})\log(p)/n_{min}}$*, and* $\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\boldsymbol{\Pi}||_1 \leq \sqrt{c_{\max}^2||\boldsymbol{\Pi}||_1^2 + \phi_0^2/(64C_2|\mathcal{S}_i|)}$*. Let* $h_n = (||\mathbf{B}||_1^2 \wedge 1) \times \big((n||\boldsymbol{\Pi}||_1^2/d) \wedge (d \vee r_{\max} \vee f_{\max})\big)\log(p)$*. Under Assumptions 1-4, there exist positive constants* C_3 *and* C_4 *such that, with probability at least* $1 - 3e^{-C_3h_n + \log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$*,*

1. Estimation Loss:

$$||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 \leq 8C_4|\mathcal{S}_i| \times \frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty^2||\mathbf{B}||_1||\boldsymbol{\Pi}||_1}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}\sqrt{\frac{(d \vee r_{max} \vee f_{\max})\log(p)}{n_{\min}}};$$

2. Prediction Loss:

$$\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \leq C_4^2|\mathcal{S}_i| \times \frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty^2||\mathbf{B}||_1^2||\boldsymbol{\Pi}||_1^2}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}\frac{(d \vee r_{max} \vee f_{\max})\log(p)}{n_{\min}}.$$

The main idea of the proof is to take advantage of the commonly used restricted eigenvalue condition and irrepresentable condition for lasso-type estimator. However, the design matrix in our case includes predicted values instead of the original one, which complicates the proof. We claim that the restricted eigenvalue and irrepresentable condition still hold for the predicted design matrix as long as the estimation and prediction losses are well controlled at the calibration stage. The proof is detailed in the supplemental materials.

The available anchoring regulators as required by Assumption 1 implies that both $||\mathbf{B}||_1 > 0$ and $||\boldsymbol{\Pi}||_1 > 0$, so $h_n/\log(p) \to \infty$. That is, these loss bounds hold with a sufficient large probability with properly chosen $f^{(k)}$.

The two sets of losses in Theorem 3 can also be controlled across the whole system by the same upper bounds defined by replacing $|\mathcal{S}_i|$ with $s_{\max} = \max_i |\mathcal{S}_i|$, with probability at least $1 - 3e^{-C_3h_n + \log(4q) + 2\log(p)} - e^{-f^{(1)}+2\log(p)} - e^{-f^{(2)}+2\log(p)}$. When both p and q are divergent up to an exponential order, say $p \asymp q \asymp e^{n_{\min}^c}$ for some $c \in (0,1)$, we can set $f^{(1)} = f^{(2)} = n_{\min}^{(1+c)/2}$ to guarantee the bounds at a sufficient large probability. However, the bounds are determined by $(d \vee r_{\max} \vee f_{\max})\log(p)$ which is $o(n_{\min})$ only when $c < \min(1/3, \theta)$. Therefore, if $s_{\max}$ also diverges up to $n_{\min}^{\tilde{c}}$ with $\tilde{c} < \min(1/4, \theta/2, 1-\theta)$, the losses can be well controlled for $c < \min((1-4\tilde{c})/3, \theta - 2\tilde{c})$.

Note that, with properly chosen $f^{(1)}$ and $f^{(2)}$, these losses are well controlled at $o(n_{\min})$, revealing the fact that we need to have sufficient observations for each network for consistent differential analysis of the two networks.

Let $W_i = diag\{\boldsymbol{\omega}_i\}$. Denote $\mathcal{I}_i = \frac{1}{n}\boldsymbol{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}\boldsymbol{\Pi}_{-i}$ and $\hat{\mathcal{I}}_i = \frac{1}{n}\hat{\boldsymbol{\Pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}\hat{\boldsymbol{\Pi}}_{-i}$. Let $\mathcal{I}_{i,11}$ be a submatrix of $\mathcal{I}_i$ with rows and columns both indexed by $\mathcal{S}_i$, and $\mathcal{I}_{i,21}$ be a submatrix of $\mathcal{I}_i$ with rows and columns indexed by $\mathcal{S}_i^c$ and $\mathcal{S}_i$, respectively. $\hat{\mathcal{I}}_{i,11}$ and $\hat{\mathcal{I}}_{i,21}$ are similarly defined from $\hat{\mathcal{I}}_i$. We further define the minimal signal strength $b_i = \max_{j \in \mathcal{S}_i}|\boldsymbol{\beta}_{ij}|$ and $\psi_i = ||\mathcal{I}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty$.

The following assumption, reminiscent of the *adaptive irrepresentable condition* in Huang et al. [2008], helps investigate the selection consistency of regulatory effects.

Assumption 5. (Weighted Irrepresentable Condition) There exists a constant $\tau \in (0,1)$ such that $||W_{\mathcal{S}_i^c}^{-1}\mathcal{I}_{i,21}^{-1}\mathcal{I}_{i,11}W_{\mathcal{S}_i}||_\infty < 1 - \tau$.

Theorem 4. *(Variable Selection Consistency) Denote* $\hat{\mathcal{S}} = \{j : \hat{\boldsymbol{\beta}}_{ij} \neq 0, j \neq i\}$*. Suppose that, for node* i*,* $\hat{\mathcal{I}}_{i,11}$ *is invertible,* $b_i > \lambda_i\psi_i/(2-\tau)$*, and* $\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\boldsymbol{\Pi}||_1 \leq$

$\sqrt{c_{\max}^2||\mathbf{\Pi}||_1^2 + \min(\phi_0^2/64, \tau(4-\tau)^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}/\psi_i)/(C_2|\mathcal{S}_i|)}$.
Under Assumptions 1-5, there exists some constant $C_5 > 0$ *such that* $\hat{\mathcal{S}}_i = \mathcal{S}_i$ *with probability at least* $1 - 3e^{-C_5 h_n + \log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$.

This theorem implies that our proposed method can identify both common and differential regulatory effects between the two networks with a sufficiently large probability. On the other hand, the assumed weighted irrepresentable condition means that the true signal should not correlate too much with irrelevant predictors so as to conduct a successful differential analysis. The corresponding proof is detailed in the supplemental materials.

3 EXPERIMENTS

3.1 SYNTHETIC DATA EVALUATION

Here we report on experiments with synthetic data to show the superior performance of our method. We compare the method **ReDNet** to a naive differential analysis which employs the 2SPLS method proposed by [Chen et al., 2017] to construct each network separately. Note that the 2SPLS method is modified here by applying ISIS to screen exogenous variables before conducting ridge regression to predict endogenous variables, making the naive differential analysis comparable to **ReDNet**.

Synthetic data are generated from both acyclic and cyclic networks involving 1000 endogenous variables, with the sample size from 200 to 300. Each network includes a subnetwork of 50 endogenous variables, whose shared and differential structures will be investigated against its pair. On average, each endogenous variable has one regulatory effect in a sparse subnetwork, and three regulatory effects in a dense network. While each pair of subnetworks in comparison share many identical regulatory effects, they also share five regulatory effects but with opposite signs, and each network has five unique regulatory effects (so the total number of differential regulatory effects is 15). The nonzero regulatory effects were independently sampled from a uniform distribution over the range $[-0.8, -0.3] \cup [0.3, 8]$. While assuming each node is directly regulated by one exogenous variable, each exogenous variable was sampled from discrete values 0,1 and 2 with probabilities 0.25, 05 and 0.25, respectively. All of the noise terms were independently sampled from the normal distribution $N(0, 0.1^2)$. We also conducted differential analysis between two networks with both $\mathbf{X}^{(1)} \neq \mathbf{X}^{(2)}$ and $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$ as in practice the paired networks may or may not share identically valued exogenous variables.

We evaluate the the performance in terms of the false discovery rate (FDR), power and Matthews correlation coefficient (MCC) [Matthews, 1975]. Let TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. MCC is defined as,

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}.$$

Here we refer nonzero effects as positives and zero effects as negatives. The MCC varies from 0 to 1 with larger values implying better variable selection.

In each differential analysis, the ridge regression employed the generalized cross validation [Golub et al., 1979] to select the ridge parameter, and the adaptive lasso used 10-fold cross-validation to choose its tuning parameter. Following the recommendation by Fan and Lv [2008], $(n^{(k)})^{0.9}$ variables are screened by ISIS.

For each type of networks, 100 synthetic data sets were generated, and the differential analysis results are summarized in Figure 3. Overall, both **ReDNet** and the naive approach maintain high power in identifying both differential and common regulatory effects. However, the naive approach tends to report high FDR, especially over 80% false discoveries of differential regulatory effects. Such a tendency to report false positives by the naive approach results in lower MCC, with dramatic decrease in identifying differential regulatory effects.

While both methods performed stably across networks with $\mathbf{X}^{(1)} \neq \mathbf{X}^{(2)}$ and $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$, **ReDNet** performed better in identifying both common and differential regulatory effects from dense networks than sparse networks in terms of FDR and MCC. However, the naive approach tends to report even higher FDR and so much lower MCC when identifying differential regulatory effects from dense networks, although reporting lower FDR and higher MCC when identifying common regulatory effects from dense networks.

We also calculated the standard errors (SE) of the reported FDR, power, and MCC over 100 synthetic data sets (the results are not shown). They are all small with most at the scale of thousandth and others at the scale of hundredth. Therefore, **ReDNet** performed robustly in differential analysis of networks, and the 2SPLS approach by Chen et al. [2017] performed also robustly in constructing single networks.

3.2 THE GENOTYPE-TISSUE EXPRESSION DATA

We performed differential analysis of gene regulatory networks on two sets of genetic genomics data from the Genotype-Tissue Expression (GTEx) project [Carithers et al., 2015], with one collected from human whole blood

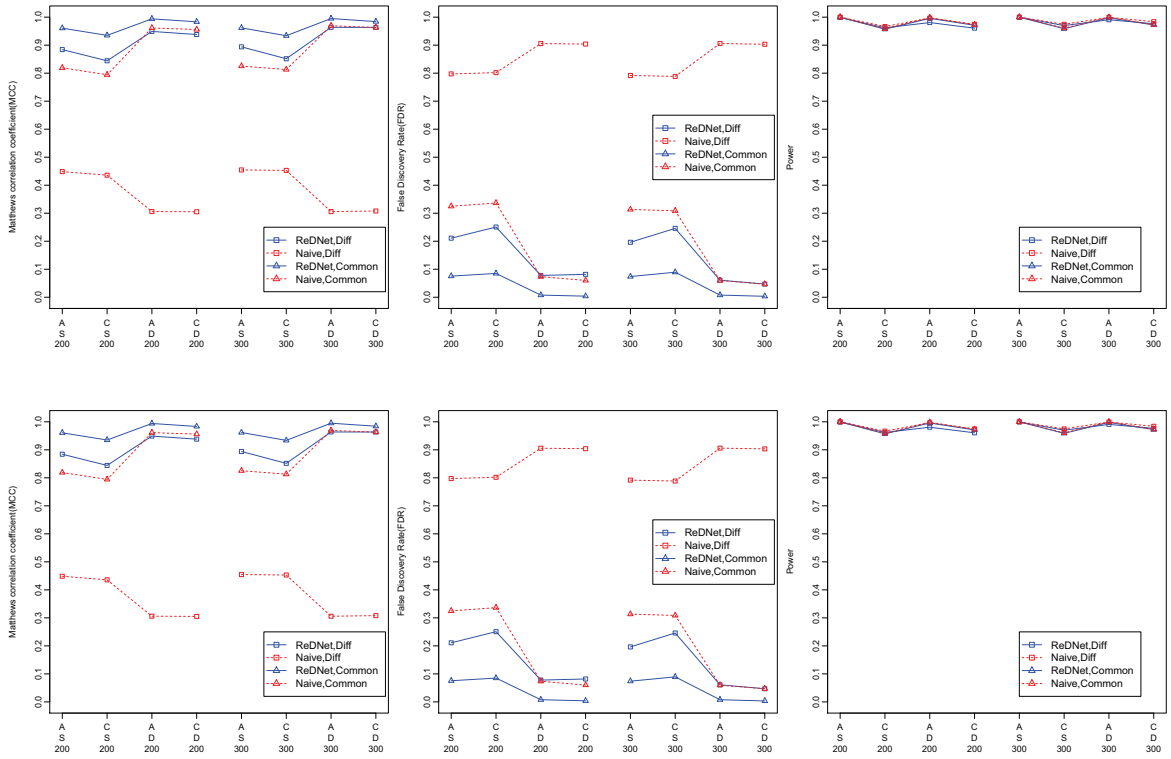

Figure 3: Performance of ReDNet Versus the Naive Approach (i.e. two networks are constructed independently). The results average over 100 synthetic data sets for different types of networks, with letters *A, C, S, D* in the x-axis denoting Acyclic, Cyclic, Sparse and Dense networks, respectively. "Diff" and "Common" summarize the performance on differential and common regulatory effects, respectively. The sample size $n^{(2)} = n^{(2)}$ is either 200 or 300.

(WB) and another one from human muscle skeletal (MS). The WB and MS data included genome-wide genetic and genotypic values from 350 and 367 healthy subjects, respectively. Both data sets were preprocessed following Carithers et al. [2015] and Stegle et al. [2010], resulting in a total of 15,899 genes and 1,083,917 single nucleotide polymorphisms (SNPs) being shared by WB and MS.

Expression quantitative trait loci (eQTL) mapping [Gilad et al., 2008] was conducted and identified 9875 genes with at least one marginally significant cis-eQTL (with p-value$<$ 0.05). For each gene, we further filtered its set of cis-eQTL by controlling the pairwise correlation under 0.9 and keeping up to three cis-eQTL which have the strongest association with the corresponding gene expression. These cis-eQTL serve as anchoring exogenous variables for the genes, and expression levels of different genes are endogenous variables. At completion of preprocessing data, we have 9,875 endogenous variables and 23,920 exogenous variables.

We applied **ReDNet** to infer the differential gene regulation on a set of eighty target genes, which had largest changes on gene-gene correlation between the two tissues. We identified a total of 711 common and 572 differential regulations on the eighty target genes. To evaluate the significance of identified regulations, we bootstrapped 100 data sets, and conducted differential analysis on each bootstrap data set. As summarized in Table 1, 50, 43 and 34 differential regulatory effects were identified in over 70%, 80% and 90% of the bootstrap data sets, respectively.

Table 1: Summary of Regulations Identified in Over 70%, 80%, 90% of the Bootstrap Data Sets by **ReDNet** From GTEx Data. Shown under "Original" are for those identified from the original data.

	Original	70%	80%	90%
Common	711	116	108	93
Differential	572	50	43	34

The top four subnetworks bearing differential regulations on some of the eighty target genes were shown in Figure 4. We also constructed the differential networks using the naive approach (the results are not shown), and reported more regulations which cover the reported ones

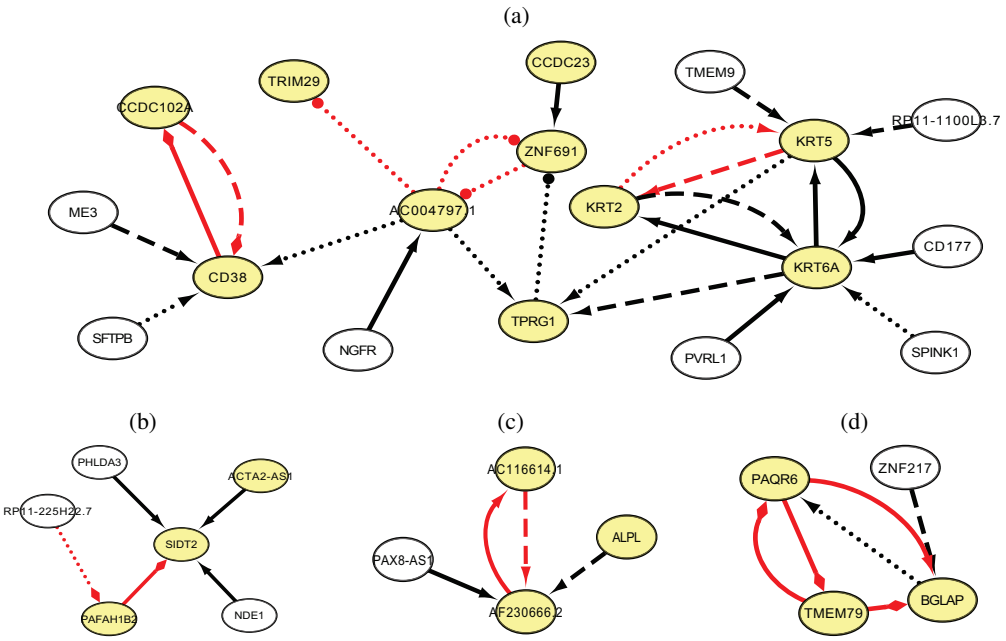

Figure 4: The Top Four Differential Subnetworks of Gene Regulation Identified by **ReDNet** From GTEx Data. The dotted, dashed, and solid lines imply regulations constructed in over 70%, 80%, and 90% of the bootstrap data sets, respectively. Highlighted in yellow are the target genes whose regulatory genes are focused in this study. The differential regulations are in red while common regulations are in black. The arrow head implies up regulation in both networks or no regulation in at most one network; the circle head implies down regulation in the whole blood but up regulation in muscle skeletal; and the diamond head implies up regulation in whole blood but down regulation muscle skeletal.

by **ReDNet**. This concurs with our observation in the synthetic data evaluation that the naive approach tends to report higher false positives, especially for differential regulatory effects.

4 CONCLUSION

We have developed a novel two-stage differential analysis method named **ReDNet**. The first stage, i.e., the calibration stage, aims for good prediction of the endogenous variables, and the second stage, i.e., the construction stage, identifies both common and differential network structures in a node-wise fashion. The key idea of **ReDNet** method is to appropriately reparametrize the independent models into a joint model so as to estimate differential and common effects directly. This approach can dramatically reduce the false discovery rate. In the experiments with synthetic data, we demonstrated the effectiveness of our method, which outperformed the naive approach with a large margin. Note that **ReDNet** allows independently conducting all ℓ_2 regularized regressions at the same time at the first stage, and all ℓ_1 regularized regressions at the same time at the second stage. Therefore, **ReDNet** not only permits parallel computation but also allows for fast subnetwork construction to avoid potential huge computational demands from differential analysis of large networks.

There are some interesting directions for future research. Firstly, it is worthwhile to explore other re-parametrization approaches such as baseline reparametrizaiton in a case-control study. Secondly, while we only consider differential analysis of two networks, it is possible to generalize our method to compare multiple networks, demanding more complex reparametrization. Finally, applying the proposed method for fully differential analysis of 53 tissues in the GTEx project still provides challenging computational and methodological issues.

Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analysis described in this paper were obtained from dbGaP accession number phs000424.v7.p2 on 08/18/2017.

References

Lisa F Berkman and S Leonard Syme. Social networks, host resistance, and mortality: a nine-year follow-up study of alameda county residents. *American Journal of Epidemiology*, 109(2):186–204, 1979.

Peter J Bickel, Yaacov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Xiaodong Cai, Juan Andrés Bazerque, and Georgios B Giannakis. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology*, 9 (5):e1003068, 2013.

Latarsha J Carithers, Kristin Ardlie, Mary Barcus, Philip A Branton, Angela Britton, Stephen A Buia, Carolyn C Compton, David S DeLuca, Joanne Peter-Demchok, Ellen T Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and Biobanking*, 13(5):311–319, 2015.

Chen Chen, Min Zhang, and Dabao Zhang. A two-stage penalized least squares method for constructing large systems of structural equations. *arXiv preprint arXiv:1511.00370v2*, 2017.

Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872, 2014.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in Genetics*, 24(8):408–415, 2008.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8 (Mar):613–636, 2007.

Wei Lin, Rui Feng, and Hongzhe Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110 (509):270–288, 2015.

Bing Liu, Alberto de La Fuente, and Ina Hoeschele. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178 (3):1763–1776, 2008.

Benjamin A Logsdon and Jason Mezey. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology*, 6(12):e1001014, 2010.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

Yang Ni, Yuan Ji, Peter Müller, et al. Reciprocal graphical models for integrative gene regulatory network analysis. *Bayesian Analysis*, 2017. doi: doi:10.1214/17-BA1087.

Yang Ni, Peter Mller, Yitan Zhu, and Yuan Ji. Heterogeneous reciprocal graphical models. *Biometrics*, 2018. doi: 10.1111/biom.12791.

Fabio Pianese, Xueli An, Fahim Kawsar, and Hiroki Ishizuka. Discovering and predicting user routines by differential analysis of social network traces. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–9, June 2013.

Peter Schmidt. *Econometrics*. New York,Marcel Dekker, 1976.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Computational Biology*, 6(5):e1000770, 2010.

James West, Ginestra Bianconi, Simone Severini, and Andrew E Teschendorff. Differential network entropy reveals cancer system hallmarks. *Scientific Reports*, 2, 2012.

Momiao Xiong, Jun Li, and Xiangzhong Fang. Identification of genetic networks. *Genetics*, 166(2):1037–1052, 2004.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1): 19–35, 2007.

Ying Zhu. Sparse linear models and ℓ_1-regularized 2SLS with high-dimensional endogenous regressors and instruments. *Journal of Econometrics*, 202(2):196 – 213, 2018.