Analysis of Variance of Multiple Causal Networks

Zhongli Jiang
Department of Statistics, Purdue University, West Lafayette, IN
jiang548@purdue.edu

Dabao Zhang
Department of Epidemiology & Biostatistics, University of California, Irvine, CA
dabao.zhang@uci.edu

Abstract

Constructing a directed cyclic graph (DCG) is challenged by both algorithmic difficulty and computational burden. Comparing multiple DCGs is even more difficult, compounded by the need of identifying perturbational causalities across graphs. We propose to unify multiple DCGs with a single structural model and develop a limited-information-based method to simultaneously construct multiple networks and infer their disparities, which can be visualized by appropriate correspondence analysis. The algorithm provides DCGs with robust non-asymptotic theoretical properties. It is designed with two sequential stages, each of which involves parallel computation tasks that are scalable to the network complexity. Taking advantage of high-performance clusters, our method makes it possible to evaluate the statistical significance of DCGs using the bootstrap method. We demonstrated the effectiveness of our method by applying it to synthetic and real datasets.

1 Introduction

Study of causal networks described by directed cyclic graphs (DCGs) becomes increasingly popular in a variety of fields, such as biomedical and social sciences [10, 3]. Detecting perturbational structures across networks sheds light on mechanistic dynamics, with promising applications. For example, gene regulatory networks related to various types of cancer differ from that of healthy cells in a complicated way and discovery of deviated regulations helps identify cancer-related biological pathways and hence design target-specific drugs [23, 18]. Interpersonal networks in social media have become a favored way for people to get informed of breaking news and natural disasters nowadays. Understanding their evolution over time may provide promising information to businesses, educators, and governments [7]. However, the complex structure and large number of nodes impose challenges on constructing even a single network, not to mention modelling and comparing multiple networks.

Structural models are widely used to describe causal networks and conduct causal inference [8, 15]. With available instrumental variables (IVs) from genomic variations, many methods have been developed to construct large causal networks of gene regulation by integrating transcriptomic and genotypic data. For instance, Liu et al. [14] proposed to assemble driver-responder relationships identified from local structural models, a formidable task due to enormous numbers of possible driver-responder pairs. Cai et al. [4] developed a sparsity-aware maximum likelihood (SML) method by putting ℓ_1 penalty on causal effects. It may reach local maximum, leading to the identifiability issue. Chen et al. [5] proposed a two-stage penalized least square (2SPLS) approach which allows parallel computing and shows superior performance.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

To infer sparse differences between two large networks, Ren and Zhang [16] recently developed a parallel algorithm based on 2SPLS and achieved better performance than separately constructing the two networks using 2SPLS. Zhou and Cai [25] proposed a fused sparse SEM (FSSEM) method to learn the differences through maximum likelihood inference of joint networks which is computationally infeasible with large networks. Li et al. [13] employed a Bayesian scheme in their method BFDSEM, unrealistically assuming equal sample sizes from both cohorts.

To the best of our knowledge, there is no algorithm available for investigating causal effects varying across multiple cohorts while identifying stable ones. In this paper, we develop a novel algorithm, designed for parallel computation, to construct a unified structural model for multiple causal networks. While each causal network may be depicted by a DCG, we deliver an analysis of variance (ANOVA) of these networks to identify causalities that are different across networks as well as important drivers and responders. Our algorithm is scalable in several aspects. Firstly, it is scalable to data size including sample size and number of variables. It would still be efficient and powerful under high dimension scheme. Secondly, it is scalable to the computational environment with its parallel computation. The fast computation allows to attack a daunting task in studying multiple networks, i.e., calculating p-values via the bootstrap method and thus controlling the false discovery rate. Thirdly, the algorithm is scalable to model complexity as it is able to infer causal networks beyond directed acyclic graphs (DAGs), such as DCGs which are demanded to depict gene regulatory networks [14].

The rest of the paper is organized as follows. We first state the model and discuss its identifiability guaranteed by available instrumental variables in Section 2. In Section 3, we present our proposed algorithm **AN**alysis **Of** **VA**riance of directed **Net**works, termed as **NetANOVA**, and introduce measures to quantify an endogenous variable's contribution as either drivers or responders. The theoretical justification is shown in Section 4 with detailed proofs in Supplemental Material. We demonstrate the feasibility and promise of our algorithm with a large-scale simulation study shown in Section 5 and a real data analysis to compare gene regulatory networks in healthy lung tissues and lung tissues with two types of cancer in Section 6. We conclude the paper with a discussion in Section 7.

2 A Unified Structural Model of Multiple Causal Networks

2.1 Model specification

We consider K pertinent causal networks, each for a cohort. For each, say k-th, cohort we have $n^{(k)}$ observations in $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)})$, where $\mathbf{Y}^{(k)}$ is an $n^{(k)} \times p$ matrix including values of p endogenous variables, and $\mathbf{X}^{(k)}$ is an $n^{(k)} \times q$ matrix including values of q exogenous variables. Each variable is assumed to have mean zero. For the k-th causal network, we consider each, say i-th, endogenous variable is causally affected by other variables as follows,

$$\mathbf{Y}_i^{(k)} = \mathbf{Y}_{-i}^{(k)} \boldsymbol{\gamma}_i^{(k)} + \mathbf{X}_{\mathcal{I}_i}^{(k)} \boldsymbol{\phi}_{\mathcal{I}_i}^{(k)} + \boldsymbol{\epsilon}_i^{(k)}, \tag{1}$$

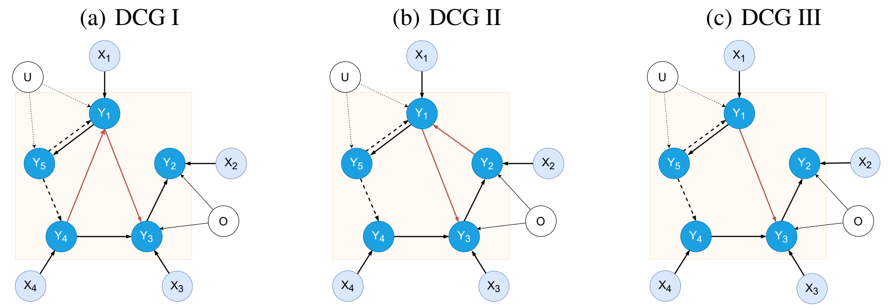
where $\mathbf{Y}_i^{(k)}$ is the i-th column of $\mathbf{Y}^{(k)}$, and $\mathbf{Y}_{-i}^{(k)}$ is the submatrix of $\mathbf{Y}^{(k)}$ excluding the i-th column. The $(p-1)$-dimensional column vector $\boldsymbol{\gamma}_i^{(k)}$ includes the direct causal effects of other endogenous variables (drivers) on the i-th one (responder). The set $\mathcal{I}_i$ includes the indices of exogenous variables that serve as IVs of the i-th endogenous variable so $\mathbf{X}_{\mathcal{I}_i}^{(k)}$ is a $n^{(k)} \times |\mathcal{I}_i|$ matrix including values of corresponding IVs. Each IV has a direct causal effect on its corresponding endogenous variable and can be identified through domain knowledge. $\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)}$ includes causal effects from corresponding IVs. $\boldsymbol{\epsilon}_i^{(k)}$ includes disturbance errors which are independently distributed with mean zero and standard deviation $\sigma_i^{(k)}$.

Pooling together equation (1) for each endogenous variable, we model the k-th causal network as,

$$\mathbf{Y}^{(k)} = \mathbf{Y}^{(k)} \boldsymbol{\Gamma}^{(k)} + \mathbf{X}^{(k)} \boldsymbol{\Phi}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \tag{2}$$

which describes a DCG as shown in Figure 1.

The $p \times p$ matrix $\boldsymbol{\Gamma}^{(k)}$ describes the causal relationships between each pair of endogenous variables and has zero diagonal elements to prohibit self-regulation. The $q \times p$ matrix $\boldsymbol{\Phi}^{(k)}$ encodes the causal effects of exogenous variables on endogenous variables. The $n^{(k)} \times p$ matrix $\boldsymbol{\epsilon}^{(k)}$ includes all disturbance errors. We assume that both $\mathbf{Y}^{(k)}$ and $\mathbf{X}^{(k)}$ have been appropriately centralized within the cohort, so no intercepts are needed in the above model.


(a) DCG I

(b) DCG II

(c) DCG III

Figure 1: DCGs of three causal networks. The nodes outside the shaded regions are exogenous variables with IVs in light blue and confounding variables (unobservable U and observed O) in white but disturbance errors are omitted. Causal relations in red differ across networks, and dashed ones cannot be revealed by available data due to unobservable U or unavailable IV for Y_5.

2.2 A unified structural model

Many algorithms have been developed to construct a structural model for single causal network [14, 4, 5]. We will show that, in the interest of comparing multiple causal networks, we can also unify them into a single structural model and hence develop an appropriate algorithm to construct all networks at the same time to deliver an ANOVA of multiple networks.

Suppose we have a baseline, say K-th, network and are interested in others' deviation from the baseline. We first reparameterize the causal effects of all endogenous variables in (1) with

$$\boldsymbol{\beta}_i^{(k)} = \boldsymbol{\gamma}_i^{(k)} - \boldsymbol{\gamma}_i^{(K)}, \text{ for } k = 1, 2, \ldots, K-1; \quad \boldsymbol{\beta}_i^{(K)} = \boldsymbol{\gamma}_i^{(K)}. \tag{3}$$

Therefore $\boldsymbol{\beta}_i^{(K)}$ includes baseline effects and, for $k \neq K$, $\boldsymbol{\beta}_i^{(k)}$ includes deviated effects of k-th network from the baseline. When it is of interest to compare the networks with each other, we can similarly reparameterize to consider the deviated effects of each network from the average effects.

For any K sets of $l \times m$ matrices, say $\mathbf{A}_i$ with $i = 1, 2, \cdots, K$, we define a matrix-valued function,

$$\mathcal{T}(\mathbf{A}_1, \cdots, \mathbf{A}_K) = (\mathbf{A}_{1:K}, (\operatorname{diag}(\mathbf{A}_1^T, \cdots, \mathbf{A}_{K-1}^T), \mathbf{0}_{m(K-1)})^T),$$

where $\mathbf{A}_{1:K}$ is constructed by stocking $\mathbf{A}_1$, $\cdots$, $\mathbf{A}_K$ row-wisely, and $\mathbf{0}_{m(K-1)}$ is a $m(K-1)$-dimensional column vector with all elements zero. Further denote

$$\begin{aligned}
\mathbf{Y}_i &= (\mathbf{Y}_i^{(1)T}, \mathbf{Y}_i^{(2)T}, \cdots, \mathbf{Y}_i^{(K)T})^T,\\
\boldsymbol{\beta}_i &= (\boldsymbol{\beta}_i^{(K)T}, \boldsymbol{\beta}_i^{(1)T}, \boldsymbol{\beta}_i^{(2)T}, \cdots, \boldsymbol{\beta}_i^{(K-1)T})^T,\\
\boldsymbol{\phi}_{\mathcal{I}_i} &= (\boldsymbol{\phi}_{\mathcal{I}_i}^{(1)T}, \boldsymbol{\phi}_{\mathcal{I}_i}^{(2)T}, \cdots, \boldsymbol{\phi}_{\mathcal{I}_i}^{(K)T}),\\
\boldsymbol{\epsilon}_i &= (\boldsymbol{\epsilon}_i^{(1)T}, \boldsymbol{\epsilon}_i^{(2)T}, \cdots, \boldsymbol{\epsilon}_i^{(K)T})^T,
\end{aligned}$$

Notice that $\boldsymbol{\gamma}_i^{(k)} = \boldsymbol{\beta}_i^{(K)} + \boldsymbol{\beta}_i^{(k)}$ for $k = 1, 2, \ldots, K-1$ from (3), we have a unified structural model for K causal networks described by (2),

$$\mathbf{Y}_i = \mathcal{T}(\mathbf{Y}_{-i}^{(1)}, \mathbf{Y}_{-i}^{(2)}, \cdots, \mathbf{Y}_{-i}^{(K)})\boldsymbol{\beta}_i + \operatorname{diag}(\mathbf{X}_{\mathcal{I}_i^{(1)}}, \mathbf{X}_{\mathcal{I}_i^{(2)}} \cdots, \mathbf{X}_{\mathcal{I}_i^{(K)}})\boldsymbol{\phi}_{\mathcal{I}_i} + \boldsymbol{\epsilon}_i. \tag{4}$$

2.3 Instrumental Variables, model identifiability, and confounding variables

As shown in Figure 1, the ability to reveal causal effects relies on the available IVs. An IV is an exogenous variable that only affects directly the driver but not the responder, except indirectly through its effect on the driver. This means that IVs can be used to address the model identifiability issue, which is complicated by the model's inherent endogeneity, and allow us to isolate the causal effects of endogenous variables. We specify the following assumption on available IVs following the rank condition [20], a necessary and sufficient condition for model identification.

Assumption 1. a. $\mathbf{X}^{(k)} \perp\!\!\!\perp \boldsymbol{\epsilon}^{(k)}$; **b**. $\exists \mathcal{C} \subset \{1, 2, \cdots, p\}$ but $\mathcal{C} \neq \emptyset$; **c**. $\forall i \in \mathcal{C}$, $\exists \mathcal{I}_i \subset \{1, 2, \cdots, q\}$ with $\mathcal{I}_i \neq \emptyset$ and $\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)} \neq 0$; **d**. $\forall i, j \in \mathcal{C}$ with $i \neq j$, $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$; **e**. $\forall j \notin \mathcal{C}$, $\gamma_{ij}^{(k)} = 0$.

Assumption 1.a specifies IVs observed in $\mathbf{X}^{(k)}$ that are uncorrelated to the disturbance errors. Assumption 1.b specifies the endogenous variables in $\mathcal{C}$ that have available IVs and Assumptions 1.c and 1.d state the restrictions on the available IVs, i.e., their availability and uniqueness, respectively. While $\gamma_{ij}^{(k)}$ is the j-th component of $\gamma_i^{(k)}$, Assumption 1.e states that we are not able to identify any causal relation with j-th endogenous variable as a driver if it has no IV available, e.g., Y_5 in Figure 1.

A striking advantage of the IV method is its robustness to confounding but exogenous variables, which may be observable or unobservable, as shown in Figure 1. As will be demonstrated by our later algorithm, available IVs help disassociate an endogenous variable from disturbance errors of other endogenous variables in model fitting. Therefore, we may explicitly incorporate observable confounding variables to improve causal inference. However, when there are unobservable confounding variables, we can still effectively construct multiple networks with available IVs.

3 Model Building and Interpretation

The single structural model in (4) makes it possible to employ a limited-information approach to construct multiple causal networks even in the case of a large number of endogenous variables. Therefore we develop a two-stage algorithm NetANOVA which allows parallel computing for fast computation and hence bootstrapping the data for significance assessment. Another daunting but necessary task is to understand the multiple networks and catch the important variables in the networks and important perturbational causalities across them. We therefore propose coefficents of determination and cause as well as correspondence analysis.

3.1 The algorithm NetANOVA

The disassociation stage: We first predict each endogenous variable in $\mathcal{C}$, solely based on all available exogenous variables, to disassociate it from the disturbance errors of other endogenous variables, following aforementioned assumption 1. For this purpose, we rearrange the terms in (2) to obtain the reduced model,

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\pi}^{(k)} + \boldsymbol{\xi}^{(k)}, \tag{5}$$

where $\boldsymbol{\pi}^{(k)} = \boldsymbol{\Phi}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$ and $\boldsymbol{\xi}^{(k)} = \epsilon^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$.

As shown in the real data analysis in constructing gene regulatory networks, the number of exogenous variables, i.e., the dimension q for $\mathbf{X}^{(k)}$, can be much larger than the sample size $n^{(k)}$, rising issues of prediction consistency and computational time. To address them, we first apply the iterative sure independence screening (ISIS; [9]) to screen for exogenous variables and then perform regression with ℓ_2 penalty. As later theoretical analysis shows, it allows for $q \lesssim exp(n^{(k)\theta})$ for some $\theta > 0$.

Specifically for each endogenous variable $i \in \mathcal{C}$ and $k \in \{1, \ldots, K\}$, we apply ISIS with computational complexity $O(n^{(k)}p)$ to screen for a set of exogenous variables, indexed by set $\mathcal{M}_i^{(k)}$. Using these $d = |\mathcal{M}_i^{(k)}|$ exogenous variables, we apply ridge regression to obtain the ridge estimator,

$$\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} = (\mathbf{X}^T_{\mathcal{M}_i^{(k)}}\mathbf{X}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)}I)^{-1}\mathbf{X}^T_{\mathcal{M}_i^{(k)}}\mathbf{Y}_i^{(k)} \tag{6}$$

and predict $\mathbf{Y}_i^{(k)}$ with

$$\hat{\mathbf{Y}}_i^{(k)} = \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\,\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}, \tag{7}$$

where $\lambda_i^{(k)}$ is a tuning parameter that can be selected via generalized cross validation [11].

The inference stage: We use the predicted endogenous variables in the previous stage to identify and estimate the causal effects, hence constructing the causal networks. We first calculate the projection matrix,

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

Multiplying the projection matrix $\mathbf{P}_i = \mathrm{diag}\{\mathbf{P}_i^{(1)}, \mathbf{P}_i^{(2)}, \ldots, \mathbf{P}_i^{(K)}\}$ to both sides of model (4), we can eliminate the exogenous variables from the model and get,

$$\mathbf{P}_i\mathbf{Y}_i = \mathbf{P}_i\boldsymbol{\Upsilon}_{-i}\boldsymbol{\beta}_i + \mathbf{P}_i\boldsymbol{\epsilon}_i, \tag{8}$$

where $\boldsymbol{\Upsilon}_{-i} = \mathcal{T}(\mathbf{Y}_{-i}^{(1)}, \mathbf{Y}_{-i}^{(2)}, \cdots, \mathbf{Y}_{-i}^{(K)})$. Since $\mathbf{P}_i \boldsymbol{\Upsilon}_{-i}$ and $\boldsymbol{\epsilon}_i$ are correlated, we instead regress $\mathbf{P}_i \mathbf{Y}_i$ against $\mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i}$ with $\hat{\boldsymbol{\Upsilon}}_{-i} = \mathcal{T}(\hat{\mathbf{Y}}_{-i}^{(1)}, \hat{\mathbf{Y}}_{-i}^{(2)}, \cdots, \hat{\mathbf{Y}}_{-i}^{(K)})$ which is disassociated from $\boldsymbol{\epsilon}_i$. With possibly high-dimensional $\boldsymbol{\beta}_i$, we apply adaptive lasso [26] to obtain its estimator,

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

where ν_i is a tuning parameter, and $\hat{\boldsymbol{\omega}}_i = |\hat{\boldsymbol{\beta}}_{0i}|^{-\gamma}$ for some $\gamma > 0$ where $|\hat{\boldsymbol{\beta}}_{0i}|_1$ contains the absolute values of $\hat{\boldsymbol{\beta}}_{0i}$, which is a preliminary estimator. The original networks can be recovered by calculating $\hat{\gamma}_i^{(k)} = \hat{\boldsymbol{\beta}}_i^{(K)} + \hat{\boldsymbol{\beta}}_i^{(k)}$ for $k = 1, 2, \ldots, K-1$.

In summary of these two stages, we summarize the algorithm in Algorithm 1.

Algorithm 1 ANalysis **O**f **VA**riance of directed **Net**works (NetANOVA)

```
Input: $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)})$, $k \in \{1, 2, ..., K\}$, with each variable centralized within cohort but scaled according to the baseline cohort; Predefined index set $\mathcal{I}_i$ for each $i \in \mathcal{C}$; $d \leftarrow O(n_{\min}^{1-\theta})$.
STAGE 1
for $i \in \mathcal{C}$ do
    1. Reduce the dimension of $\mathbf{X}^{(k)}$ by ISIS to get $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$; Set $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} = \mathbf{X}^{(k)}$ if $q \le n^{(k)}$.
    2. Estimate $\hat{\mathbf{Y}}_i^{(k)}$ by regressing $\mathbf{Y}_i^{(k)}$ against $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ with ridge regression.
end for
STAGE 2
for $i = 1, 2, \ldots, p$ do
    1. Calculate projection matrices $\mathbf{P}_i$.
    2. Predict $\hat{\boldsymbol{\Upsilon}}_{-i}$ from $\hat{\mathbf{Y}}_i^{(k)}$, $i \in \mathcal{C}$.
    3. Estimate $\hat{\boldsymbol{\beta}}_i$ by regressing $\mathbf{P}_i \mathbf{Y}_i$ against $\mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i}$ with adaptive lasso.
end for
Output: $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$ which contains the baseline and differential causal effects.
```

3.2 Coefficient of determination and coefficient of cause

For each endogenous variable i, we can calculate its coefficient determination to measure the proportion of its variation due to the effects of its drivers, i.e., $\mathbf{Y}_{-i}^{(k)} \gamma_i^{(k)}$ in (1), for cohort k,

$$R_i^{2(k)} = 1 - ||\mathbf{Y}_i^{(k)} - \mathbf{Y}_{-i}^{(k)} \hat{\gamma}_i^{(k)}||_2^2 / ||\mathbf{Y}_i^{(k)}||_2^2,$$

On the other hand, we can also calculate the coefficient of cause for each endogenous variable i which summarizes proportions that it contributes to the variation of its responders, for cohort k,

$$C_i^{2(k)} = \sum_{j=1}^{p} \left(1 - ||\mathbf{Y}_j^{(k)} - \mathbf{Y}_i^{(k)} \hat{\gamma}_{ji}^{(k)}||_2^2 / ||\mathbf{Y}_j^{(k)}||_2^2 \right).$$

While we name the above coefficient of determination as causal R^2 with its value between zero and one, we will simply call the coefficient of cause C^2 which is positive but may be greater than one.

3.3 Correspondence analysis of causal effects

We can conduct a correspondence analysis of causal effects in $\boldsymbol{\Gamma}^{(k)}$ to reveal clusters of drivers, responders, or driver-responder pairs which outstand from the rest in each cohort k. We can also conduct a correspondence analysis of deviated causal effects to reveal these clusters which deviate the most from a baseline or the rest. However, each causal effect or its deviation may vary differently and we need to standardize them, based on the bootstrap results, before correspondence analysis. For example, to compare cohorts k and l for their causal effects, we should instead obtain the following summary statistic,

$$z_{ij}^{(k,l)} = \left(\hat{\gamma}_{ij}^{(k)} - \hat{\gamma}_{ij}^{(l)}\right) \Big/ \left(\sum_{b=1}^{B} (\hat{\gamma}_{ij}^{(k,b)} - \hat{\gamma}_{ij}^{(l,b)})^2 / (B-1)\right)^{1/2}.$$

where $\hat{\gamma}_{ij}^{(k)}$ and $\hat{\gamma}_{ij}^{(k,b)}$ are estimates of corresponding effects from the observed data and b-th set of bootstrap data, respectively; we also let $z_{ij}^{(k)} = 0$ when both of its numerator and denominator are sufficiently small. We can define similar statistics to investigate effects' variation from their means.

A singular value decomposition (SVD) of $Z^{(k,l)} = (z_{ij}^{(k,l)})_{p \times p}$ can obtain its left and right singular vectors, say $\{U_i^{(k,l)}\}$ and $\{V_i^{(k,l)}\}$, respectively. While the left singular vectors help identify responders that differ between the two cohorts, the right singular vectors help identify drivers that differ between the two cohorts. We can overlay the plot of $U_2^{(k)}$ vs. $U_1^{(k)}$ with the plot of $V_2^{(k)}$ vs. $V_1^{(k)}$ to identify responder-driver pairs for their causalities varying the most across networks.

4 Theoretical Analysis

In this section, we will establish non-asymptotic guarantees and show that our constructed DCGs have good theoretical properties. We characterize the properties with a prespecified sequence $\delta^{(k)} = \exp\{-o(n^{(k)})\}$ with $\delta^{(k)} \to 0$ as $n^{(k)} \to \infty$, i.e., each $\delta^{(k)}$ approaches zero slower than $\exp\{-n^{(k)}\}$. We denote $\delta_{min} = \min_{1 \le k \le K} \delta^{(k)}$ and only present main results here, with details including general notations, assumptions, and proofs in Supplementary Material.

Theorem 4.1. *Let $n = \sum_{k=1}^{K} n^{(k)}$, $\mathcal{S}_i = supp(\boldsymbol{\beta}_i)$, and g_n is a function of n and δ_{min} specified in (9) in Supplemental Material. Then under Assumption 1 and Assumptions B.1–B.4 in Supplementary Material, we have that, with probability at least $1 - \delta_{min} - \delta$ with $\delta = p \sum_{k=1}^{K} \delta^{(k)}$,*

1. *(Bounded Errors) $||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_2^2 \lesssim |S_i| \; g_n \; \{d \vee \log(pK/\delta) \vee \|\boldsymbol{\pi}\|_{2max}^2 \vee \log(d/\delta_{min})\}/n;$*
2. *(Causality Consistency) $sign(\hat{\boldsymbol{\beta}}_i) = sign(\boldsymbol{\beta}_i)$.*

We will next show that, with proper choice of $\{\delta^{(k)}\}$, we can bound the estimation errors system-wise with ultra high probability. We can pick $\{\delta^{(k)}\}$ such that $\delta_{min} \asymp e^{-n^t}$ where $t \in (0, min(\theta, 1-\theta))$. Then we have $f_n \lesssim n^{(2-\theta)/2}$, and the restriction on the number of true signals can be reduced to $|\mathcal{S}_i| \lesssim n^{\theta/2}$, which is a requirement that can be fulfilled especially in a sparse model. We have $d = n^{1-\theta}$, so $g_n \lesssim n^{(1-2\theta)/2}$. As a result, the bound can be dominated by the order of $|S_i| n^{(1-4\theta)/2}$.

Note that the error for the whole system of p nodes can be controlled by a similar bound, replacing each occurrence by $|S_i|$ with $\underset{1 \le i \le p}{max} |S_i|$, with probability at least $1 - p(\delta_{min} + \delta)$. Following the former calculation, we learn that, when $\underset{1 \le i \le p}{max} |S_i|$ grow slower than $n^{(4\theta-1)/2}$, the error will approach zero when $n \to \infty$. To achieve the bound with high probability, we only need $p\delta$ to diminish. That is, we could have the error well controlled even if the dimension grows up to e^{n^s} where $s \in (0, \frac{t}{2})$.

Given the auspicious estimation performance, we will further discuss the benefit of pooling cohorts compared to that of learning in a single-cohort analysis, where the estimators $\beta_i^{(k)}$ for each cohort are estimated separately using the algorithm without re-parameterization. For each k, we can separately conduct the analysis on each cohort and obtain the following property for single task estimation.

Corollary 4.1.1. *Under the same conditions as Theorem 4.1, we have that, with probability at least $1 - \delta_{min} - \delta/K$,*

$$||\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)}||_2^2 \lesssim |S_i| \; g_n \; \{d \vee \log(pK/\delta) \vee \|\boldsymbol{\pi}\|_{2max}^2 \vee \log(d/\delta_{min})\}/n^{(k)}.$$

The proof directly follows by treating each cohort as a baseline cohort in Theorem 4.1. Note that the denominator is making a crucial difference in inferring the bound. The algorithm indicates that when we aggregate the samples that does not differ too much, the estimator tends to converge faster, especially when $n \gg n^{(k)}$, where we have a plethora of samples in addition to the samples for the base cohort.

With the same choice of δ_{min} and δ as the above theorem, the probability for consistent variable selection can approach one, with a less stringent requirement in the size of S_i. The above theorem implies that our proposed method can identify both baseline and differential regulatory effects, among all of the K networks with a sufficiently large probability, not only in terms of the set of true signals but also the sign of signals. In the case of gene regulatory networks, for instance, our method could correctly distinguish the up and down regulations between genes.

Next, we will present a theorem for the coefficient of determination and coefficient of cause defined in Section 3.2. Denote the statistics R^2 and C^2 calculated with real parameters as, respectively,

$$R_{0i}^{2(k)} = 1 - ||\mathbf{Y}_i^{(k)} - \mathbf{Y}_{-i}^{(k)}\gamma_i^{(k)}||_2^2/||\mathbf{Y}_i^{(k)}||_2^2, \quad C_{0i}^{2(k)} = \sum_{j=1}^{p}(1 - ||\mathbf{Y}_j^{(k)} - \mathbf{Y}_i^{(k)}\gamma_{ji}^{(k)}||_2^2/||\mathbf{Y}_j^{(k)}||_2^2).$$

Then we can derive the following properties.

Theorem 4.2. *(Coefficients of Determination/Cause Consistency) Under the same conditions as in Theorem 4.1, with* $h_n = \sqrt{n^{(k)}} + \sqrt{-\log(\delta_{min})} + 1$*, we have that,*

1. *With probability at least* $1 - p(2\delta_{min} + \delta/K)$*,*

$$\sum_{i=1}^{p}|R_i^{2(k)} - R_{0i}^{2(k)}| \quad \lesssim \quad \frac{|\mathcal{S}_i|g_n\{d \vee \log(p/\delta) \vee \|\boldsymbol{\pi}\|_{2max}^2 \vee \log(d/\delta_{min})\}}{n^{(k)3/4}}(1 + ||\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)}||_2 h_n);$$

2. *With probability at least* $1 - p(\delta_{min} + \delta/K)$*,*

$$\sum_{i=1}^{p}|C_i^{2(k)} - C_{0i}^{2(k)}| \quad \lesssim \quad \frac{|\mathcal{S}_i|g_n\{d \vee \log(p/\delta) \vee \|\boldsymbol{\pi}\|_{2max}^2 \vee \log(d/\delta_{min})\}}{n^{(k)}}(1 + ||B||_1 \vee 1).$$

As considered in Theorem 4.1, we set $\delta_{min} \asymp \delta/(pK) \asymp e^{-n^t}$. Then $h_n \lesssim \sqrt{n^{(k)}}$, and $\sum_{i=1}^{p}|R_i^{2(k)} - R_{0i}^{2(k)}| \lesssim \underset{1\leq i\leq p}{max}|S_i|n^{(1-4\theta)/2} + (\underset{1\leq i\leq p}{max}|S_i|)^{1/2}n^{\frac{-2\theta-1}{2}}$ which would vanish if $\underset{1\leq i\leq p}{max}|S_i| \lesssim n^{(4\theta-1)/2}$. Similarly, $\sum_{i=1}^{p}|C_i^{2(k)} - C_{0i}^{2(k)}| \lesssim \underset{1\leq i\leq p}{max}|S_i|n^{(1-4\theta)/4}$ and will converge to 0 when $\underset{1\leq i\leq p}{max}|S_i| \lesssim n^{(4\theta-1)/4}$. From the above discussion, both R^2, C^2 for all the nodes in the whole system is ℓ_1 consistent. It is worth noting that, although our algorithm fits a model for each responder, estimate of each node's explanatory power as a driver is also well controlled over the system, but with a slightly larger bound as shown in the above theorem.

5 Simulation Study

We examine the performance of NetANOVA by simulating data of sample size 200, 500 and 1000 from each of three pertinent DCGs, consisting of 1000 endogenous variables. However, only 50 endogenous variables are involved with causality, with each regulated by 3 others on average. Each of the three networks has 5 unique causal effects. First two networks share five causal effects but with opposite signs to that of the third one, resulting in the first two networks having 15 different effects from the third one and the first network have 20 different effects from the second one. The size for causal effects is taken from a uniform distribution over $[-0.8, -0.3] \bigcup [0.3, 0.8]$. The IVs are generated from a multinomial distribution with 3 outcomes 0, 1, 2, with probabilities 0.25, 0.5, 0.25 respectively. The disturbance errors are sampled independently from $N(0, 0.1^2)$.

For each sample size, we simulated 100 data sets and applied NetANOVA to each by bootstrapping 100 times to calculate p values and construct Receiver Operating Characteristic (ROC) curves as shown in Figure 2. For causal effects of all three DCGs, Figure 2.a shows an almost perfect area under curve (AUC) for each sample size. When comparing causal effects of the first two DCGs vs. the baseline one, i.e., the third DCG, the AUC bottoms at 0.867 with sample size at 200, but reaches one with sample size at 1000, showing excellent performance of NetANOVA.

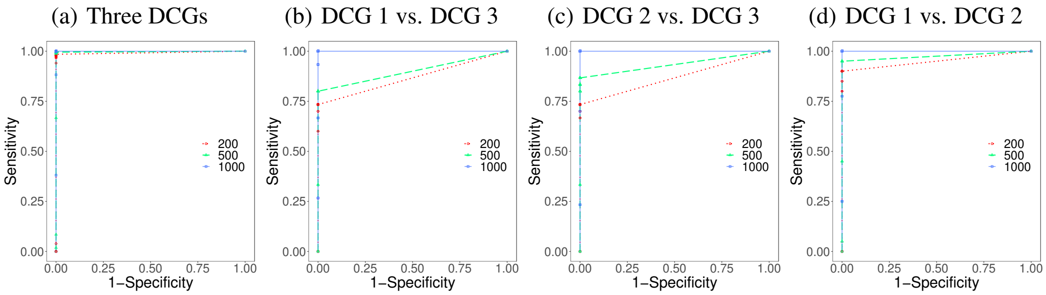


Figure 2: ROC curves of **NetANOVA** on simulated data. (a) ROC curves of causal effects constructed in all three networks; (b) ROC curves of deviated effects of the first network from the third one; (c) ROC curves of deviated effects of the second network from the third one; (d) ROC curves of deviated effects of the first network from the second one.

Taking one simulated dataset with sample size at 500, we evaluated our proposed statistics R^2 and C^2 by plotting estimated statistics against the statistics calculated with true causal effects in Figure 3. For each of the three networks, we observe linear trends with slope almost one for both R^2 and C^2 statistics, indicating their excellent performance.

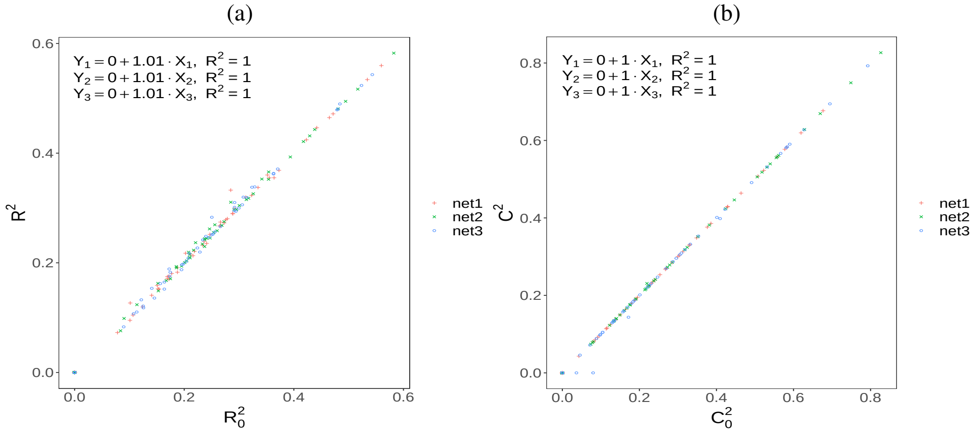
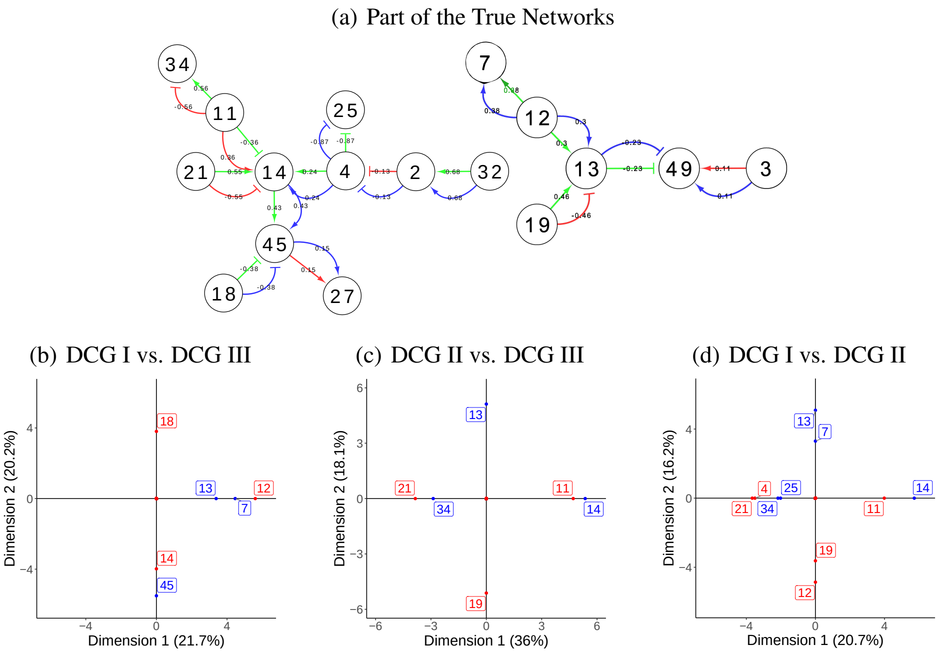


Figure 3: Statistics R^2 and C^2 on a simulated data with sample size 500. We plot each statistic (R^2 or C^2) vs. its value calculated with true causal effects (R_0^2 or C_0^2) for each network. We simply label the three networks as 1, 2, and 3. Shown in the top-left corners are the fitted linear regression models between the pairs as well as the corresponding coefficients of determination.

We also took one simulated data with sample size 500 and conducted our proposed correspondence analysis to identify drivers and responders that show important perturbational causality across the three DCGs. When comparing DCG I vs. DCG III, Figure 4.b shows a driver-responder pair (14, 45) with different causal effects, which is verified by observing 14 significantly up regulates 45 in DCG I vs. DCG III in Figure 4.a (blue connections). However, we also see the driver 18 and responder 45 stay opposite on the y-axis, because 18 significantly down regulates 45 implying in DCG I vs. DCG III in Figure 4.a. On x-axis, we see pairs (12,13) and (12,7) which correspond to another cluster of causal networks as shown in the right of Figure 4.a.

6 Real Data Analysis

We applied NetANOVA to investigate the gene regulatory networks of lung tissues of healthy individuals and patients with lung adenocarcinoma (LUAD) or lung squamous cell carcinoma (LUSC). We obtained transcriptomic and genotypic data of healthy lungs ($n = 482$) from the Genotype-Tissue Expression (GTEx) project [6] and of both LUAD ($n = 485$) and LUSC ($n = 406$) from The Cancer


(a) Part of the True Networks

(b) DCG I vs. DCG III

(c) DCG II vs. DCG III

(d) DCG I vs. DCG II

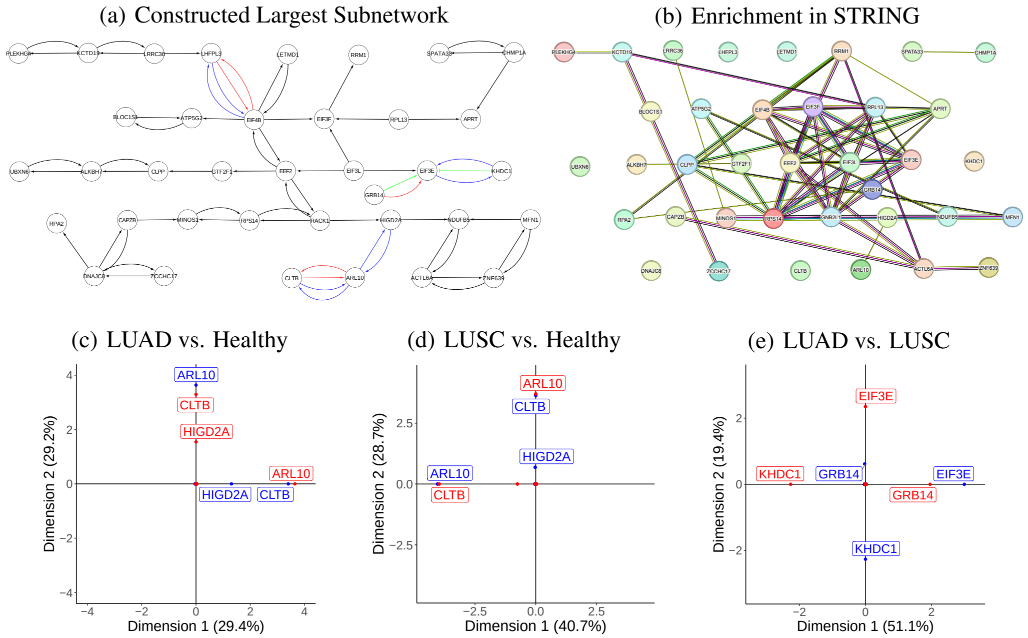
Figure 4: Correspondence analysis of **NetANOVA** on a simulated data with sample size 500. (a) Plot of some deviated causal effects among the three true networks (shown in Figure 6) with DCG I vs. DCG III, DCG II vs. DCG III, and DCG I vs. DCG II in blue, red, and green, respectively. There are a total of 1000 endogenous variables with the first 50 involved in causal relations. (b)-(d) Plots for correspondence analysis with drivers in red and responders in blue.

Genome Atlas (TCGA) project [22]. After pre-processing, there are 15,135 genes and 427,820 single nucleotide polymorphisms (SNPs) being shared by three cohorts. Cis-eQTL mapping identified 7059 genes with at least one significant SNPs inside their genetic regions (p-value< 0.05), i.e., valid IVs. We bootstrapped the data 100 times to assess the significance of all effects with results shown in Table 1.

Table 1: Summary of causal effects identified by **NetANOVA**. Shown in the columns are the results from the original data, different bootstrap cutoffs (80% - 100%), and adjusted by Benjamini-Hochberg adjustment (BH), respectively.

Type of Effects	Original	80%	90%	95%	100%	BH
Healthy Tissue	79833	16594	11481	8760	4602	3165
LUAD vs Healthy Tissue	13711	1185	848	670	458	296
LUSC vs Healthy Tissue	13104	1385	980	768	477	274
LUSC vs LUAD	18139	1615	976	665	289	38

Controlling adjusted p-value at 0.1, we have the largest subnetwork bearing differential structures shown in Figure 5.a, which is verified via STRING [21] as in Figure 5.b. STRING reports a protein-protein interaction (PPI) enrichment with p-value $< 10^{-16}$, implying significant causal effects between the genes shown in Figure 5.a. We identified many previous validated relationships, such as the connected pair RPS14 and RACK1 (GNB2L1) which were experimentally verified by [1]. Our results suggest new findings, particularly causal relationship rather than mere association, e.g., deviated regulations between ARL10 and CLTB from healthy lung tissues in both LUAD and LUSC. Our correspondence analysis of these deviated effects, shown in Figure 5.c-e, confirms these important driver-responder pairs, and the calculated statistics R^2 and C^2 in Table 2 also show such deviation between the three cohorts.



(a) Constructed Largest Subnetwork

(b) Enrichment in STRING

(c) LUAD vs. Healthy

(d) LUSC vs. Healthy

(e) LUAD vs. LUSC

Figure 5: Partial results of gene regulatory networks for human lung. (a) The largest subnetworks of gene regulation with deviated causal effects between healthy, LUAD, and LUSC lung tissues. Shown in black is the gene regulatory network for healthy lung tissue. Shown in blue, red, green are significantly deviated causal effects of LUAD vs. healthy, LUSC vs. healthy, and LUAD vs. LUSC, respectively. (b) Enrichment of the largest subnetwork in STRING. (c)-(e) Correspondence analysis of the deviated effects between the largest subnetworks with drivers in red and responders in blue.

7 Discussion

Though having promising applications, construction and perturbational study of multiple causal networks are computationally challenging in practice due to the involved large systems. We develop the algorithm **NetANOVA** which avoids optimizing full-information objective functions of the whole system and instead takes limited-information objective functions, each for revealing all drivers of a single responder. Such a responder-focused method allows to deploy parallel computation in two sequential stages and makes it possible to be computationally scalable to the number of involved responders. With available clusters of computers, we can further take the bootstrap method for the usually infeasible task, i.e., evaluating the significance of the constructed causalities in each network and perturbational causalities across networks. Our theoretical analysis and simulation study demonstrate the utility and efficiency of NetANOVA.

Table 2: Statistics R^2 and C^2 for Important Genes

	Cohort	ARL10	CLTB	EIF3E	GRB14	HIGD2A	KHDC1
R^2	LUAD	0.90	0.74	0.73	0.58	0.71	0.60
	LUSC	0.79	0.79	0.77	0.75	0.54	0.00
	Healthy	0.00	0.00	0.56	0.00	0.30	0.00
C^2	LUAD	1.41	0.74	1.18	0.00	0.86	0.60
	LUSC	0.79	0.79	0.75	0.75	0.16	0.00
	Healthy	0.00	0.00	0.00	0.00	0.34	0.00

With the model complexity of multiple causal networks, especially DCGs that NetANOVA supports, it is a daunting task to visualize them and comprehend each network and variation across multiple networks. We have proposed statistics R^2 and C^2 to quantify the contributions of responders and drivers within a causal network, respectively. While comparing these two statistics across networks may help identify responders and drivers involved in network dynamics, we develop a correspondence analysis to visualize the key players. While our simulation study and real data analysis show promising results, correspondence analysis could be further developed to realize its full potential.

Acknowledgments and Disclosure of Funding

This research is partially supported by the Office of Research of Purdue University and the National Cancer Institute under Grant R03 CA235363. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. 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 analyses described in this manuscript were obtained from the GTEx Portal on 10/03/2021 and dbGaP accession number phs000424.v8.p2 on 7/23/2019.

References

[1] Francisco Acosta-Reyes, Ritam Neupane, Joachim Frank, and Israel S Fernández. The Israeli acute paralysis virus IRES captures host ribosomes by mimicking a ribosomal state with hybrid tRNAs. *The EMBO Journal*, 38(21):e102226, 2019.

[2] Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[3] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network Analysis in the Social Sciences. *Science*, 323(5916):892–895, 2009.

[4] 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.

[5] Chen Chen, Min Ren, Min Zhang, and Dabao Zhang. A Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations. *The Journal of Machine Learning Research*, 19(1):40–73, 2018.

[6] GTEx Consortium et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[7] Mário Cordeiro, Rui P Sarmento, Pavel Brazdil, and João Gama. Evolving Networks and Social Network Analysis Methods and Techniques. *Social Media and Journalism: Trends, Connections, Implications*, 101(2), 2018.

[8] Henriette Engelhardt, Hans-Peter Kohler, and Alexia Prskawetz. *Causal analysis in population studies*. Springer, 2009.

[9] 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.

[10] Ilias Georgakopoulos-Soares, Chengyu Deng, Vikram Agarwal, Candace SY Chan, Jingjing Zhao, Fumitaka Inoue, and Nadav Ahituv. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature communications*, 14(1):2333, 2023.

[11] 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.

[12] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[13] Yan Li, Dayou Liu, Tengfei Li, and Yungang Zhu. Bayesian differential analysis of gene regulatory networks exploiting genetic perturbations. *BMC Bioinformatics*, 21(1):1–13, 2020.

[14] 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.

[15] Judea Pearl. *Causality*. Cambridge university press, 2009.

[16] Min Ren and Dabao Zhang. Differential Analysis of Directed Networks. *arXiv preprint arXiv:1807.10173*, 2018.

[17] Phillippe Rigollet and Jan-Christian Hütter. High Dimensional Statistics. *Lecture notes for course 18S997*, 813:814, 2015.

[18] Da Ruan, Alastair Young, and Giovanni Montana. Differential analysis of biological networks. *BMC Bioinformatics*, 16(1):1–13, 2015.

[19] Mark Rudelson, Roman Vershynin, et al. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[20] Peter Schmidt. *Econometrics*. New York, Marcel Dekker, 1976.

[21] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[22] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45 (10):1113, 2013.

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

[24] Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[25] Xin Zhou and Xiaodong Cai. Inference of differential gene regulatory networks based on gene expression and genetic perturbation data. *Bioinformatics*, 36(1):197–204, 2020.

[26] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

A List of Notations

For a matrix A:

$\mathbf{A}_{ij}$	Entry of the i-th row and the j-th column of the matrix
$\|\|\mathbf{A}\|\|_1$	Maximum column sums of the absolute value of the matrix.
$\|\|\mathbf{A}\|\|_{-\infty}$	Minimum of the row sums of absolute values of the matrix.
$\|\|\mathbf{A}\|\|_{\infty}$	Maximum of the row sums of absolute values of the matrix.
$\lambda_{\text{max}}(\mathbf{A})$	Maximum eigenvalue of a matrix.
$\lambda_{\text{min}}(\mathbf{A})$	Minimum eigenvalue of a matrix.

For a vector ν :

$\boldsymbol{\nu}_i$	i-th element of $\boldsymbol{\nu}$.
$\|\|\boldsymbol{\nu}\|\|_q$	$\left(\sum_{i=1}^n \|\boldsymbol{\nu}_i\|^q\right)^{1/q}$ for $q = 1, 2$.

For numbers x and y:

$x \wedge y$	Minimum of numbers x and y.
$x \vee y$	Maximum of numbers x and y.
$x \lesssim y$	$x \leq cy$ for some positive constant c.
$x \gtrsim y$	$x \geq cy$ for some positive constant c.
$x \asymp y$	$cx \leq y \leq dx$ for some positive constant c, d.

B Detailed Theoretical Analysis

In this section, we will present detailed assumptions besides Assumption 1 and systematically analyze theoretical properties of our proposed NetANOVA algorithm, which lay the groundwork for proving the theorems stated in the main text.

Our theoretical analysis is conducted, assuming a prespecified sequence $\delta^{(k)} = \exp\{-o(n^{(k)})\}$ with $\delta^{(k)} \to 0$ as $n^{(k)} \to \infty$, i.e., each $\delta^{(k)}$ approaches zero slower than $\exp\{-n^{(k)}\}$. We also denote

$$\delta_{min} = \min_{1 \leq k \leq K} \delta^{(k)}.$$

B.1 General Results

To prove the consistency of the disassociation stage, we first point out that $\mathcal{M}_i^{(k)}$ obtained from ISIS [24] successfully recovers the true nonzero set $\mathcal{M}_{i*}^{(k)}$ in reduced form (5). We first state the assumption which restricts the sample size for different cohorts in the same order and ensures the sparsity of the true underlying relationships. We state the assumption by extending the conditions in Fan and Lv [9] to pave the way for Theorem B.1, the sure screening property.

Denote

$$\Sigma^{(k)} = \text{Cov}(\mathbf{X}^{(k)}), \quad W^{(k)} = (\Sigma^{(k)})^{-1/2} \mathbf{X}^{(k)T},$$

and, for any index subset $\mathcal{M} \subset \{1, 2, \cdots, q\}$,

$$\Sigma_{\mathcal{M}}^{(k)} = cov(\mathbf{X}_{\mathcal{M}}^{(k)}), \quad W_{\mathcal{M}}^{(k)} = (\Sigma_{\mathcal{M}}^{(k)})^{-1/2} \mathbf{X}_{\mathcal{M}}^{(k)T}$$

Further denote the j-th row of $\boldsymbol{Y}_i^{(k)}$, $\boldsymbol{X}_i^{(k)}$, and $\boldsymbol{\pi}_i^{(k)}$ as $Y_{ji}^{(k)}$, $X_{ji}^{(k)}$, and $\pi_{ji}^{(k)}$, respectively.

Assumption B.1.

(a) $\lambda_{max}(\Sigma^{(k)}) \lesssim (n^{(k)})^{\tau^{(k)}}$ *for some positive* $\tau^{(k)}$.

(b) $W^{(k)}$ *follows a spherically symmetric distribution with concentration property: There exist some constants* $\tilde{c}_1^{(k)} > 1$, $\tilde{c}_2^{(k)} > 1$, *and* $\tilde{c}_3^{(k)} > 0$ *such that, for any index subset* $\mathcal{M} \subset \{1, 2, \cdots, q\}$ *with* $|\mathcal{M}| \geq \tilde{c}_1^{(k)} n^{(k)}$, *we have, with probability at least* $1 - \exp(-\tilde{c}_3^{(k)} n^{(k)})$,

$$1/\tilde{c}_2^{(k)} \leq \lambda_{min}(W_M^{(k)T} W_M^{(k)}/|\mathcal{M}|) \leq \lambda_{max}(W_M^{(k)T} W_M^{(k)}/|\mathcal{M}|) \leq \tilde{c}_2^{(k)}.$$

(c) $var(Y_{ji}^{(k)}) \lesssim 1$ *and there exists* $\kappa^{(k)} \geq 0$ *such that*

$$\min_{m\in\mathcal{M}_{i*}^{(k)}} \left|\pi_{mi}^{(k)}\right| \gtrsim (n^{(k)})^{-\kappa^{(k)}} \quad \text{and} \quad \min_{m\in\mathcal{M}_{i*}^{(k)}} \left|cov\left(Y_{ji}^{(k)}, X_{jm}^{(k)}\pi_{mi}^{(k)}\right)\right| \gtrsim 1.$$

For each node i in network $k \in \{1, 2, ..., K\}$, we have the following theorem.

Theorem B.1. *Denote* $U^{(k)} = 1 - 2\kappa^{(k)} - \tau^{(k)}$, $U_{min} = \underset{1\leq k\leq K}{min} U^{(k)}$, $\kappa_{max} = \underset{1\leq k\leq K}{max} \kappa^{(k)}$, *and* $q \lesssim \exp(n_{min}^{(k)})^{\tilde{c}}$ *for some* $\tilde{c} \in (0, 1 - 2\kappa^{(k)})$ *where* $n_{\min} = \underset{1\leq k\leq K}{min} n^{(k)}$*, then under Assumption B.1, there exists some* $\theta \in (0, U_{min})$ *and some positive constant c, such that, with probability at least* $1 - c\exp\left\{-(n^{(k)})^{1-2\kappa_{max}}/\log(n^{(k)})\right\}$,

$$\mathcal{M}_{i*}^{(k)} \subseteq \mathcal{M}_i^{(k)}.$$

Theorem B.1 implies that ISIS can recover the true nonzero set $\mathcal{M}_{i*}^{(k)}$ for each network with overwhelming probability, when n tends to infinity, and thus paves the way for our subsequent analysis. We will establish the rest of the theory based on $\mathcal{M}_{i*}^{(k)} \subseteq \mathcal{M}_i^{(k)}$ for convenience.

Next, we will investigate the consistency of predictions using ridge regression in (7). To facilitate simplicity, we will exclude the subscript $\mathcal{M}_i^{(k)}$ from $\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}$ and replace it with $\mathbf{X}_i^{(k)}$ from now on. Similarly, we will refer to the true and estimated causal parameters as $\boldsymbol{\pi}_i^{(k)}$ and $\hat{\boldsymbol{\pi}}_i^{(k)}$, respectively.

Assumption B.2. $\lambda_{max}(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)}) \asymp \lambda_{min}(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)}) \asymp n^{(k)}$ *and the singular values of* $\mathbf{I} - \mathbf{\Gamma}^{(k)}$ *have a positive lower bound for each* $k \in 1, \ldots, K$.

Let

$$\mathbf{\Upsilon} = \mathcal{T}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \cdots, \mathbf{Y}^{(K)}), \quad \mathbf{\Pi} = \mathcal{T}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \cdots, \boldsymbol{\pi}^{(K)}),$$

and denote

$$\mathbf{X} = \text{diag}\{\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}, \ldots, \mathbf{X}_i^{(K)}\}.$$

We use $\hat{\mathbf{\Upsilon}}$ and $\hat{\mathbf{\Pi}}$ to denote the prediction of $\mathbf{\Upsilon}$ and $\mathbf{\Pi}$, respectively. In addition, we will use the subscript j to denote the j-th column of the corresponding matrices. We further denote

$$||\boldsymbol{\pi}||_{2max}^2 = \max_{1\leq i\leq p}\{||\boldsymbol{\pi}_i^{(1)}||_2^2 \vee ||\boldsymbol{\pi}_i^{(2)}||_2^2 \vee \cdots \vee ||\boldsymbol{\pi}_i^{(K)}||_2^2\},$$

which represents the maximum of the square of the magnitude of signals over all networks and nodes. We have the following theorem for all $j \in \{1, 2, \ldots, Kp\}$.

Theorem B.2. *Let the ridge parameter in (6) be, for each node i,* $\lambda_i^{(k)} = \sqrt{(n^{(k)})}$*. Under Assumptions B.1 and B.2, we have, with probability at least* $1 - \sum_{k=1}^K \delta^{(k)}$,

1. $||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2 \lesssim d \vee \log(1/\delta_{min}) \vee ||\boldsymbol{\pi}||_{2max}^2/n_{min}$;
2. $||\mathbf{X}(\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j)||_2^2 \lesssim d \vee \log(1/\delta_{min}) \vee ||\boldsymbol{\pi}||_{2max}^2$.

This implies that, with proper choice of the ridge parameter and sequence $\{\delta^{(k)}\}$, we can have well-bounded estimation and prediction loss, essentially ℓ_2 consistent. For example, we can pick $\delta^{(k)} \asymp e^{-n_{\min}^t}$ where $t \in min(0.1 - \theta)$, so the ℓ_2 estimation loss for each term and the MSE (Mean squared error) would tends to 0 with large sample size, as long as $||\boldsymbol{\pi}||_{2max}^2$ is bounded by a positive constant.

We could also draw conclusions in terms of the error over the whole system, i.e. the Frobenius norm. So that with probability at least $1 - p\sum_{k=1}^K \delta^{(k)}$, the systematic estimation error and MSE could reach the exact same bound for a single node. To control the loss with ultra-high probability, we only need control $p\delta^{(k)}$. Notice that each $p\delta^{(k)} \asymp pe^{-n_{\min}^t} \to 0$ whenever $p = o(e^{n_{min}^t})$. That is, the dimension can grow with a restricted exponential term, i.e, $p = e^{n_{min}^t}$.

After exploring the theory for the first stage, we will discuss the promising properties in the second stage. For each node i, we use $\mathcal{S}_i$ to denote the indices of true non-zero components of $\boldsymbol{\beta}_i$, i.e., $\mathcal{S}_i = \text{supp}(\boldsymbol{\beta}_i)$. Further denote

$$\mathbf{\Pi}_{-\boldsymbol{i}} = \mathcal{T}(\boldsymbol{\pi}_{-i}^{(1)}, \boldsymbol{\pi}_{-i}^{(2)}, \cdots, \boldsymbol{\pi}_{-i}^{(K)}).$$

Following Bickel et al. [2], we impose the restricted eigenvalue condition to constrain the projected design matrix.

Assumption B.3. $||\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_{-i}\beta_i||_2^2 \geq 2\lambda_0 n||\beta_i||_2^2$ *whenever* $||\beta_{S_i^c}||_1 \leq 3||\beta_{S_i}||_1$ *for some positive constant* λ_0*. In addition,* $\|\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}\|_\infty \leq \|\hat{\boldsymbol{\omega}}_{\mathcal{S}_i^c}\|_{-\infty}$.

The latter assumption is intuitive. Recall $\hat{\boldsymbol{\omega}}_i = 1/|\hat{\boldsymbol{\beta}}_{0i}|^\gamma$, so we need $\|\hat{\boldsymbol{\beta}}_{0\mathcal{S}_i}\|_{-\infty} \geq \|\hat{\boldsymbol{\beta}}_{0\mathcal{S}_i^c}\|_\infty$, where the estimation for true signal should always dominate the noise. This assumption is milder than the selection consistency of estimators which can be achieved by a proper estimator such as Lasso under mild conditions.

Denote

$$\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p],$$

and define

$$\begin{cases} f_n = \sqrt{n}||\mathbf{\Pi}||_1\sqrt{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2} + d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2, \\ g_n = (C_\Pi\sqrt{\frac{d}{n_{min}}(d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2)} + ||\mathbf{\Pi}||_1)(||\mathbf{B}||_1 \vee 1). \end{cases} \tag{9}$$

Next we will show that under some mild conditions, for each node i, the estimation and prediction loss could be bounded with large probability. Denote $\sigma_{max} = \max\limits_{1\leq i\leq p} \max\limits_{1\leq k\leq K} \sigma_i^{(k)}$ and $\tilde{\sigma}_{max} = \max\limits_{1\leq i\leq p} \max\limits_{1\leq k\leq K} \tilde{\sigma}_i^{(k)}$.

Theorem B.3. *Suppose that the adaptive lasso at the second stage takes the tuning parameter*

$$\nu_i = \frac{4}{\sqrt{n}||\hat{\boldsymbol{\omega}}_{S_i}||_\infty} g_n max\{3\sqrt{2}\sqrt{\log(\frac{4Kd}{\delta_{min}})}(\sigma_{max} \vee \tilde{\sigma}_{max}), \sqrt{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2}\},$$

and $|\mathcal{S}_i| \leq \frac{\lambda_0}{C}\frac{n}{f_n}$ *for some constant* C*. Then under Assumptions B.1–B.3, we have that, with probability at least* $1 - \delta_{min} - p\sum_{k=1}^K \delta^{(k)}$*,*

1. $||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_2^2 \lesssim \frac{|S_i|}{n} g_n\{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2 \vee \log\frac{d}{\delta_{min}}\}$*;*
2. $\frac{1}{n}||\mathbf{P}_i\hat{\mathbf{\Upsilon}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \lesssim \frac{|S_i|}{n} g_n\{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2 \vee \log\frac{d}{\delta_{min}}\}$.

For each k, we can separately conduct the analysis on each cohort and obtain the following property for single task estimation. That is, letting

$$f(n^{(k)}) = \sqrt{n^{(k)}}||\boldsymbol{\pi}^{(k)}||_1\sqrt{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)}} + d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)},$$

$$g(n^{(k)}) = \left(C_\pi\sqrt{\frac{d}{n^{(k)}}(d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)})} + ||\boldsymbol{\pi}^{(k)}||_1\right)(||\mathbf{B}||_1 \vee 1),$$

we have the following result for each node i.

Corollary B.3.1. *Suppose that the adaptive lasso at the inference stage takes the tuning parameter*

$$\nu_i = \frac{4}{\sqrt{n^{(k)}}||\hat{\boldsymbol{\omega}}_{S_i}||_\infty} g(n^{(k)}) max\{3\sqrt{2}\sqrt{\log(\frac{4d}{\delta_{min}})}(\sigma_{max} \vee \tilde{\sigma}_{max}), \sqrt{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)}}\},$$

and $|\mathcal{S}_i| \leq \frac{\lambda_0}{C}\frac{n^{(k)}}{f(n^{(k)})}$ *for some constant* C*. Then under Assumptions B.1–B.3, we have that, with probability at least* $1 - \delta_{min} - p\delta^{(k)}$*,*

1. *$||\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)}||_2^2 \lesssim \frac{|S_i|}{n^{(k)}} g(n^{(k)})\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee \log \frac{d}{\delta_{min}}\}$;*
2. *$\frac{1}{n^{(k)}}||\mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)})||_2^2 \lesssim \frac{|S_i|}{n^{(k)}} g(n^{(k)})\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee \log \frac{d}{\delta_{min}}\}$.*

The proof directly follows by treating each cohort as a baseline cohort in theorem (4.1). Note that the denominator is making a crucial difference in inferring the bound. The algorithm indicates that when we aggregate the samples that does not differ too much, the estimator tends to converge faster, especially when $n \gg n^{(k)}$, where we have a plethora of samples in addition to the samples for the base cohort.

Moving forward, we will examine the selection consistency of the estimator. Denote the covariance matrix of $\mathbf{P}_i \mathbf{X} \boldsymbol{\Pi}_{-i}$ as

$$\Sigma_i = \frac{1}{n} \boldsymbol{\Pi}_{-i}^T \mathbf{X}^T \mathbf{P}_i \mathbf{X} \boldsymbol{\Pi}_{-i},$$

and correspondingly denote

$$\hat{\Sigma}_i = \frac{1}{n} \hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \mathbf{X} \hat{\boldsymbol{\Pi}}_{-i},$$

for $\mathbf{P}_i \mathbf{X} \hat{\boldsymbol{\Pi}}_{-i}$. We will use subscripts to denote the corresponding rows and columns in above matrices. For instance, Σ_{S_i, S_i} represents the rows and columns of Σ_i, both indexed by S_i.

In order to investigate the selection consistency of causal effects, we impose the irrepresentable condition introduced by Zhao and Yu [24].

Assumption B.4. *For each $i \in \{1, 2, \cdots, p\}$, $\Sigma_{\mathcal{S}_i, \mathcal{S}_i}$is invertible and $||\Sigma_{\mathcal{S}_i^c, \mathcal{S}_i} \Sigma_{\mathcal{S}_i, \mathcal{S}_i}^{-1}||_\infty < 1 - \eta$ for some constant $\eta \in (0, 1)$.*

Theorem B.4. *(Causality Selection Consistency) Suppose that $|\mathcal{S}_i| \leq \frac{\eta}{(\eta+2)\tau_i} \frac{n}{f_n}$ with $\tau_i = ||\Sigma_{\mathcal{S}_i, \mathcal{S}_i}^{-1}||_\infty$ and $\min\limits_{j \in \mathcal{S}_i} |\boldsymbol{\beta}_{ij}| > \frac{\nu_i \tau_i ||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty}{2-\eta}$ for each node i. Then under Assumptions B.1–B.4, we have probability at least $1 - \delta_{min} - p \sum_{k=1}^K \delta^{(k)}$, such that $sign(\hat{\boldsymbol{\beta}}_i) = sign(\boldsymbol{\beta}_i)$.*

Note that estimation consistency does not imply sign consistency and vice versa. We further establish sign consistency to guarantee our causal variables are selected with the correct sign. From the discussion in Theorem 4.1, it directly follows that $\nu_i \lesssim n^{(1-3\theta)/2}$, which would infer $\frac{\nu_i \tau ||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty}{2-\eta} \lesssim n^{\frac{1-3\theta}{2}}$. Thus we are putting really mild assumption on the smallest true signal size, requiring at most a gap of $n^{\frac{2-3\theta}{2}}$ between the signal and the error decay rate of $n^{-\frac{1}{2}}$, As for the number of true significant variables, we can grant the growth at most to $n^{\frac{\theta}{2}}$.

B.2 Proofs of the Theorems

B.2.1 Proof of Theorem B.1

It can be inferred from [9] that there exists some $\theta^{(k)} \in (0, U^{(k)})$ such that, when $|\mathcal{M}_i^{(k)}| \lesssim (n^{(k)})^{1-\theta^{(k)}}$ and for some positive constant $\tilde{c}_4^{(k)}$, we have that, with probability at least $1 - \tilde{c}_4^{(k)} \exp\left\{-(n^{(k)})^{1-2\kappa^{(k)}} / \log(n^{(k)})\right\}$,

$$\mathcal{M}_{i*}^{(k)} \subseteq \mathcal{M}_i^{(k)}.$$

Denote

$$\theta = \min_{1 \leq k \leq K} \theta^{(k)},$$

then, for $|\mathcal{M}_i^{(k)}| \lesssim n_{\text{min}}^{1-\theta}$, with probability at least $1 - \tilde{c}_4^{(k)} \exp\{-(n^{(k)})^{1-2\kappa_{max}} / \log(n^{(k)})\}$, we have that

$$\mathcal{M}_{i*}^{(k)} \subseteq \mathcal{M}_i^{(k)}$$

B.2.2 Proof of Theorem B.2

In the proof of theorem 2, we will first give the bound of estimation loss and predictions loss for ridge regression on each network, after which we utilize the result to prove the bound for constructed networks.

Lemma B.1. *For each $k \in \{1, 2, \dots, K\}$, set $\lambda_i^{(k)} = \sqrt{n^{(k)}}$. Under Assumptions B.1 and B.2, we have, with probability at least $1 - \delta^{(k)}$,*

1. $||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 \lesssim \frac{d \vee \log(1/\delta^{(k)}) \vee ||\boldsymbol{\pi}_i^{(k)}||_2^2}{n^{(k)}}$;
2. $||\mathbf{X}_i^{(k)}(\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})||_2^2 \lesssim d \vee \log(1/\delta^{(k)}) \vee ||\boldsymbol{\pi}_i^{(k)}||_2^2$.

Proof of Lemma B.1. We first link the ridge regression estimator with the ordinary least squres (OLS) estimator denoted as

$$\hat{\boldsymbol{\pi}}_i^{*(k)} = \big(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)})^{-1}\mathbf{X}_i^{(k)T}\mathbf{Y}_i^{(k)}.$$

We write ridge estimator $\hat{\boldsymbol{\pi}}_i^{(k)}$ as

$$\hat{\boldsymbol{\pi}}_i^{(k)} = \big(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)} + \lambda_i^{(k)} I_d\big)^{-1}\mathbf{X}_i^{(k)T}\mathbf{Y}_i^{(k)} = L\hat{\boldsymbol{\pi}}_i^{*(k)},$$

where

$$\begin{aligned} L = L^T &= (\lambda_i^{(k)}(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)})^{-1} + I_d\big)^{-1} = \frac{1}{\lambda_i^{(k)}}((\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)})^{-1} + \frac{1}{\lambda_i^{(k)}} I_d\big)^{-1} \\ &= I_d - \lambda_i^{(k)}(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)} + \lambda_i^{(k)} I_d\big)^{-1}, \end{aligned}$$

using Woodbury's identity.

The ℓ_2 loss for estimation can be then decomposed as

$$\begin{aligned} &||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 \\ = \;& (L\hat{\boldsymbol{\pi}}_i^{*(k)} - \boldsymbol{\pi}_i^{(k)})^T(L\hat{\boldsymbol{\pi}}_i^{*(k)} - \boldsymbol{\pi}_i^{(k)}) \\ = \;& \underbrace{(\hat{\boldsymbol{\pi}}_i^{*(k)} - \boldsymbol{\pi}_i^{(k)})^T L^T L(\hat{\boldsymbol{\pi}}_i^{*(k)} - \boldsymbol{\pi}_i^{(k)})}_{T_{21}} + \underbrace{2(\hat{\boldsymbol{\pi}}_i^{*(k)} - \boldsymbol{\pi}_i^{(k)})^T L^T(L - I_d)\boldsymbol{\pi}_i^{(k)}}_{T_{22}} \\ & + \underbrace{\boldsymbol{\pi}_i^{(k)T}(L - I_d)^T(L - I_d)\boldsymbol{\pi}_i^{(k)}}_{T_{23}}. \end{aligned}$$

We will derive the bound for T_{21}, T_{22} and T_{23} with respectively, after which the estimation loss could be bound easily using the property of eigenvalues of $\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)}$.

We write

$$\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)} = Q_i^{(k)}V_i^{(k)}Q_i^{(k)T},$$

using eigendecomposition, where $Q_i^{(k)}$ is unitary, and $V_i^{(k)}$ is a diagonal matrix with diagonal entries to be eigenvalues v_{ij}, where $v_{ij} \asymp n^{(k)}$ according to Assumption B.2, for each $j \in \{1, 2, \dots, d\}$. Then

$$L = \frac{1}{\lambda_i^{(k)}} Q_i^{(k)}\big(V_i^{(k)-1} + \frac{1}{\lambda_i^{(k)}} I_d\big)^{-1}Q_i^{(k)T} = I_d - \lambda_i^{(k)}Q_i^{(k)}\big(V_i^{(k)} + \lambda_i^{(k)} I_d\big)^{-1}Q_i^{(k)T}.$$

Denote

$$\mathrm{var}(\epsilon_{ji}^{(k)}) = \sigma_i^{(k)2}, \quad \mathrm{var}(\xi_{ji}^{(k)}) = \tilde{\sigma}_i^{(k)2}.$$

Note that

$$\hat{\boldsymbol{\pi}}_i^{*(k)} - \boldsymbol{\pi}_i^{(k)} \sim \mathcal{N}(0, \tilde{\sigma}_i^{(k)2}(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)})^{-1}),$$

which follows the sub-Gaussian distribution. Following Assumption B.2, the singular value of $\mathbf{I} - \boldsymbol{\Gamma}^{(k)}$ is bounded from below, then we have $\tilde{\sigma}_{max} \lesssim \sigma_{max}$. We then employ the Hanson-Wright inequality [19] to bound its tail. For any t, there is some positive constant t_1, such that

$$\begin{aligned}
\mathbb{P}(T_{21} \geq \mathbb{E}(T_{21}) + t) &\leq t_1 \exp\left(-\frac{t^2}{\tilde{\sigma}_i^{(k)4}/n^{(k)2}||L^TL||_F^2} \wedge \frac{t}{\tilde{\sigma}_i^{(k)2}/n^{(k)}||L^TL||_{op}}\right). && (10)\\
\mathbb{E}(T_{21}) &= \tilde{\sigma}_i^{(k)2}\mathrm{tr}\left(LL^T\mathbf{X}_i^{(k)}\tilde{\mathbf{X}}_i^{(k)T}\right)\\
&= \tilde{\sigma}_i^{(k)2}\mathrm{tr}\left(\frac{1}{\lambda_i^{(k)2}}Q_i^{(k)}\big(V_i^{(k)-1} + \frac{1}{\lambda_i^{(k)}}I_d\big)^{-2}V_i^{(k)-1}Q^{(k)T}\right)\\
&= \tilde{\sigma}_i^{(k)2}\sum_{j=1}^{d}\frac{v_{ij}}{(v_{ij}+\lambda_i^{(k)})^2} \lesssim \frac{d}{n^{(k)}}. && (11)\\
||LL^T||_F^2 &= \mathrm{tr}\left(LL^TLL^T\right)\\
&= \mathrm{tr}\left(\frac{1}{\lambda_i^{(k)4}}Q_i^{(k)}\big(V_i^{(k)-1} + \frac{1}{\lambda_i^{(k)}}I_d\big)^{-4}Q^{(k)T}\right)\\
&= \sum_{j=1}^{d}\frac{v_{ij}^4}{(v_{ij}+\lambda_i^{(k)})^4} \lesssim d. && (12)\\
||LL^T||_{op} &= \sqrt{\lambda_{max}\left(LL^TLL^T\right)} \lesssim 1. && (13)
\end{aligned}$$

Let

$$t = \sqrt{\tilde{\sigma}_i^{(k)4}/n^{(k)2}||LL^T||_F^2 log(\frac{K}{\delta^{(k)}})/exp(t_1)} \vee \left(\tilde{\sigma}_i^{(k)2}/n^{(k)}||LL^T||_{op}log(\frac{K}{\delta^{(k)}})/exp(t_1)\right),$$

with (10), (11), (12), and (13), we have that, with probability at least $1 - \delta^{(k)}/K$,

$$T_{21} \lesssim \frac{d \vee \sqrt{dlog(\frac{1}{\delta^{(k)}})} \vee log(\frac{1}{\delta^{(k)}})}{n^{(k)}}. \tag{14}$$

Next we will bound T_{22} using Gaussian tail inequality. Denote

$$\boldsymbol{\pi}_i^{q(k)} = Q^{(k)T}\boldsymbol{\pi}_i^{(k)},$$

then

$$||\boldsymbol{\pi}_i^{q(k)}||_2^2 = ||\boldsymbol{\pi}_i^{(k)}||_2^2.$$

For any positive t,

$$\mathbb{P}\left(T_{22} \geq t\right) \leq \exp\left(-\frac{1}{2}\frac{t^2}{\mathrm{var}(T_{22})}\right).$$

where

$$\begin{aligned}
\mathrm{var}(T_{22}) &= 4\tilde{\sigma}_i^{(k)2}\boldsymbol{\pi}_i^{(k)T}(L-I_d)^TL(\mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)})^{-1}L^T(L-I_d)\boldsymbol{\pi}_i^{(k)}\\
&= 4\tilde{\sigma}_i^{(k)2}\boldsymbol{\pi}_i^{(k)T}Q_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}(V_i^{(k)-1}+\frac{1}{\lambda_i^{(k)}}I_d)^{-1}V_i^{(k)-1} \times\\
&\quad (V_i^{(k)-1}+\frac{1}{\lambda_i^{(k)}}I_d)^{-1}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}Q_i^{(k)T}\boldsymbol{\pi}_i^{(k)}\\
&= 4\tilde{\sigma}_i^{(k)2}\sum_{j=1}^{d}\frac{\pi_i^{q(k)2}v_{ij}\lambda_i^{(k)2}}{(v_{ij}+\lambda_i^{(k)})^4} \lesssim \frac{||\boldsymbol{\pi}_i^{(k)}||_2^2}{n^{(k)2}}.
\end{aligned}$$

Let

$$t = \sqrt{2\mathrm{var}(T_{22})\log(K/\delta^{(k)})},$$

we have that, with probability at least $1 - \delta^{(k)}/K$,

$$T_{22} \lesssim \frac{||\boldsymbol{\pi}_i^{(k)}||_2}{n^{(k)}}\sqrt{\log(1/\delta^{(k)})}. \tag{15}$$

Finally, we can bound T_{23} as

$$\begin{aligned} T_{23} &= \boldsymbol{\pi}_i^{(k)T} Q_i^{(k)} (V_i^{(k)} + \lambda_i^{(k)} I_d)^{-2} Q_i^{(k)T} \boldsymbol{\pi}_i^{(k)} \\ &= \lambda_i^{(k)2} \sum_{j=1}^{d} \frac{\pi_i^{q(k)2}}{(v_{ij} + \lambda_i^{(k)})^2} \lesssim \frac{||\boldsymbol{\pi}_i^{(k)}||_2^2}{n^{(k)}}. \end{aligned} \tag{16}$$

Combining the bound (14), (15) and (16), we have that, with probability at least $1 - \delta^{(k)}$,

$$||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 \lesssim \frac{d \vee log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i^{(k)}||_2^2}{n^{(k)}}.$$

Note that

$$\frac{(\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})^T \mathbf{X}_i^{(k)T} \mathbf{X}_i^{(k)} (\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})}{(\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})^T (\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})} \leq \lambda_{max}(\mathbf{X}_i^{(k)T} \mathbf{X}_i^{(k)}).$$

So we directly have

$$||\mathbf{X}_i^{(k)} (\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})||_2^2 \leq \lambda_{max}(\mathbf{X}_i^{(k)T} \mathbf{X}_i^{(k)}) ||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 \lesssim d \vee log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i^{(k)}||_2^2.$$

which concludes the proof of Lemma B.1. □

Next, we will prove the theorem by repeatedly applying Lemma B.1. We write the error loss for each $j \in \{1, 2, \ldots, Kp\}$ as

$$||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2 = \begin{cases} \sum_{k=1}^K ||\hat{\boldsymbol{\pi}}_{j|p}^{(k)} - \boldsymbol{\pi}_{j|p}^{(k)}||_2^2, & j \leq p, \\ ||\hat{\boldsymbol{\pi}}_{j|p}^{(k-1)} - \boldsymbol{\pi}_{j|p}^{(k-1)}||_2^2, & j > p, \end{cases} \tag{17}$$

where $j|d$ is the remainder of j divided by p, denoting the corresponding column index. Applying the bounds in Lemma B.1 to all the networks, we have that, with probability at least $1 - \sum_{k=1}^K \delta^{(k)}$,

$$||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2 \lesssim \sum_{k=1}^{K} \frac{d \vee \log(1/\delta^{(k)}) \vee ||\boldsymbol{\pi}_i^{(k)}||_2^2}{n^{(k)}} \lesssim \frac{d \vee \log(1/\delta_{min}) \vee ||\boldsymbol{\pi}||_{2max}^2}{n_{min}}.$$

Using the same approach we can bound the prediction loss. We have that, with probability at least $1 - \sum_{k=1}^K \delta^{(k)}$,

$$\begin{aligned} ||\mathbf{X}(\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j)||_2^2 &= \sum_{k=1}^{K} ||X_i^{(k)} (\hat{\boldsymbol{\pi}}_{j|p}^{(k)} - \boldsymbol{\pi}_{j|p}^{(k)})||_2^2 \\ &\lesssim \sum_{k=1}^{K} d \vee \log(1/\delta^{(k)}) \vee ||\boldsymbol{\pi}_i^{(k)}||_2^2 \lesssim d \vee \log(1/\delta_{min}) \vee ||\boldsymbol{\pi}||_{2max}^2, \end{aligned}$$

which concludes the proof of Theorem B.2.

B.2.3 Proof of Theorem B.3

Lemma B.2. *Suppose that for each node i and a properly chosen positive constant C,*

$$|\mathcal{S}_i| \leq \frac{\lambda_0}{C} \frac{n}{f_n}. \tag{18}$$

for function f_n defined in (9). Under Assumptions B.1–B.3, we have that, with probability at least $1 - p \sum_{k=1}^K \delta^{(k)}$,

$$||\mathbf{P}_i \mathbf{X} \hat{\mathbf{\Pi}}_{-i} \beta_i||_2^2 \geq \lambda_0 n ||\beta_i||_2^2,$$

whenever $||\beta_{S_i^c}||_1 \leq 3 ||\beta_{S_i}||_1$.

Proof of Lemma B.2. We will bound the maximum value in the matrix

$$|(\mathbf{P}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i})^T(\mathbf{P}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i}) - (\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_{-i})^T(\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_{-i})|,$$

whose value in l-th row and r-th column can be expressed as

$$\begin{aligned}
&|\hat{\mathbf{\Pi}}_l^T\mathbf{X}\mathbf{P}_i\mathbf{X}\hat{\mathbf{\Pi}}_r - \mathbf{\Pi}_l^T\mathbf{X}\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_r| \\
\leq\ & |\underbrace{(\hat{\mathbf{\Pi}}_l - \mathbf{\Pi}_l)^T\mathbf{X}^T\mathbf{P}_i\mathbf{X}(\hat{\mathbf{\Pi}}_r - \mathbf{\Pi}_r)}_{T_{31}}| + |\underbrace{(\hat{\mathbf{\Pi}}_l - \mathbf{\Pi}_l)^T\mathbf{X}^T\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_r}_{T_{32}}| \\
+\ & |\underbrace{\mathbf{\Pi}_l^T\mathbf{X}^T\mathbf{P}_i\mathbf{X}(\hat{\mathbf{\Pi}}_r - \mathbf{\Pi}_r)}_{T_{33}}|.
\end{aligned}$$

We will further bound T_{31}, T_{32} and T_{33}. Note that, since $\mathbf{P}_i$ is a projection matrix,

$$\lambda_{max}(\mathbf{P}_i) = 1.$$

Then by applying Theorem B.2, we have that, with probability at least $1 - \sum_{k=1}^K \delta^{(k)}$,

$$\begin{aligned}
T_{31} &\leq ||\mathbf{P}_i\mathbf{X}(\hat{\mathbf{\Pi}}_l - \mathbf{\Pi}_l)||_2 \times ||\mathbf{P}_i\mathbf{X}(\hat{\mathbf{\Pi}}_r - \mathbf{\Pi}_r)||_2 \\
&\leq \lambda_{max}^2(\mathbf{P}_i)||\mathbf{X}(\hat{\mathbf{\Pi}}_l - \mathbf{\Pi}_l)||_2 \times ||\mathbf{X}(\hat{\mathbf{\Pi}}_r - \mathbf{\Pi}_r)||_2 \\
&\lesssim d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2.
\end{aligned}$$

With equation (17), we have that,

$$||\mathbf{X}\,\mathbf{\Pi}_r||_2^2 \leq \sum_{k=1}^K ||X_i^{(k)}\boldsymbol{\pi}_{r|p}^{(k)}||_2^2 \lesssim \sum_{i=1}^K n^{(k)}||\boldsymbol{\pi}_{r|p}^{(k)}||_2^2 \lesssim n\sum_{i=1}^K ||\boldsymbol{\pi}_{r|p}^{(k)}||_2^2 \lesssim n||\mathbf{\Pi}||_1^2.$$

Thus T_{32} could be bounded as

$$T_{32} \leq ||\mathbf{X}\mathbf{\Pi}_r||_2||\mathbf{P}_i\mathbf{X}(\hat{\mathbf{\Pi}}_l - \mathbf{\Pi}_l)||_2 \lesssim \sqrt{n}||\mathbf{\Pi}||_1\sqrt{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2}. \tag{19}$$

In a similar manner, we have

$$T_{33} \lesssim \sqrt{n}||\mathbf{\Pi}||_1\sqrt{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2}. \tag{20}$$

Then

$$T_{31} + T_{32} + T_{33} \lesssim \sqrt{n}||\mathbf{\Pi}||_1\sqrt{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2} + d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2 = f_n.$$

Let, for certain positive constant C,

$$T_{31} + T_{32} + T_{33} \leq Cf_n,$$

then we have,

$$\begin{aligned}
&\beta_i^T((\mathbf{P}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i})^T(\mathbf{P}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i}) - (\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_{-i})^T(\mathbf{P}_i\mathbf{X}\mathbf{\Pi}_{-i}))\beta_i \\
\leq\ & ||\beta_i||_1^2(|T_{31}| + |T_{32}| + |T_{33}|) \leq C|\mathcal{S}_i|||\beta_i||_2^2 f_n \leq \lambda_0 n||\beta_i||_2^2.
\end{aligned} \tag{21}$$

Along with the restriction of Assumption B.3, we have

$$||\mathbf{P}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i}\beta_i||_2^2 \geq \lambda_0 n||\beta_i||_2^2,$$

whenever $||\beta_{S_i^c}||_1 \leq 3||\beta_{S_i}||_1$ by applying triangular inequality. The proof of Lemma B.2 is now complete. □

Lemma B.3. *For each $i \in \{1, 2, \ldots, p\}$, ν_i is the tuning parameter in the inference stage. Under Assumptions B.1–B.3, we have, with probability at least $1 - \delta_{min} - p\sum_{k=1}^K \delta^{(k)}$,*

$$||\hat{\mathbf{\Upsilon}}_{-i}^T\mathbf{P}_i[\boldsymbol{\epsilon}_i + (\mathbf{\Upsilon}_{-i} - \hat{\mathbf{\Upsilon}}_{-i})\boldsymbol{\beta}_{\boldsymbol{i}}]||_\infty \leq \frac{1}{4}n\nu_i||\hat{\boldsymbol{\omega}}_{S_i}||_\infty. \tag{22}$$

Proof of Lemma B.3. We have $\boldsymbol{\Upsilon}_{-i} = \mathbf{X}\boldsymbol{\Pi}_{-i} + \boldsymbol{\xi}_{-i}$ and $\hat{\boldsymbol{\Upsilon}}_{-i} = \mathbf{X}\hat{\boldsymbol{\Pi}}_{-i}$ by definition, where

$$\boldsymbol{\xi_{-i}} = \mathcal{T}(\boldsymbol{\xi}_{-i}^{(1)}, \boldsymbol{\xi}_{-i}^{(2)}, \cdots, \boldsymbol{\xi}_{-i}^{(K)}).$$

Note that

$$\begin{aligned} ||\hat{\boldsymbol{\Upsilon}}_{-i}^T \mathbf{P}_i[\boldsymbol{\epsilon}_i + (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i})\boldsymbol{\beta}_i]||_\infty &\leq \underbrace{||\hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \boldsymbol{\epsilon}_i||_\infty}_{T_{34}} + \underbrace{||\hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \mathbf{X}(\boldsymbol{\Pi}_{-i}^T - \hat{\boldsymbol{\Pi}}_{-i}^T)\boldsymbol{\beta}_i||_\infty}_{T_{35}} \\ &+ \underbrace{||\hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \boldsymbol{\xi}_{-i} \boldsymbol{\epsilon}_i||_\infty}_{T_{36}}. \end{aligned}$$

Firstly, by applying Theorem B.2, we have that, with probability at least $1 - p\sum_{k=1}^K \delta^{(k)}$,

$$\begin{aligned} \|(\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j)^T\|_{-\infty}^2 &= \max_{1\leq j\leq Kp} ||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_1^2 \leq \max_{1\leq j\leq Kp} \Big(Kd||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_2^2\Big) \\ &\lesssim \frac{d}{n_{min}}(d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2). \end{aligned}$$

Denote $t_\Pi = \|(\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j)^T\|_{-\infty} \leq C_\Pi \sqrt{\frac{d}{n_{min}}(d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2)}$,

We further have $X_{\cdot j}^T X_{\cdot j} \asymp n$ after standardization, where $X_{\cdot j}$ is the j-th column of $\mathbf{X}$. Therefore, $X_{\cdot j}^T \mathbf{P}_i \boldsymbol{\epsilon}_i$ follows Gaussian distribution with variance bounded as

$$\text{var}\left(X_{\cdot j}^T \mathbf{P}_i \boldsymbol{\epsilon}_i\right) \lesssim X_{\cdot j}^T \mathbf{P}_i X_{\cdot j} \lesssim n.$$

We have

$$\begin{aligned} &\mathbb{P}\left(||\hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \boldsymbol{\epsilon}_i||_\infty \geq t\right) \\ &\leq \mathbb{P}\left((||(\hat{\boldsymbol{\Pi}}_{-i} - \boldsymbol{\Pi}_{-i})^T||_\infty + ||\boldsymbol{\Pi}_{-i}^T||_\infty)||\mathbf{X}^T \mathbf{P}_i \boldsymbol{\epsilon}_i||_\infty \geq t\right) \\ &\leq \mathbb{P}\left(||\mathbf{X}^T \mathbf{P}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{t}{t_\Pi + ||\boldsymbol{\Pi}||_1}\right) \\ &\leq 2Kd \exp\left\{-\left(\frac{t}{t_\Pi + ||\boldsymbol{\Pi}||_1}\right)^2 / (2n\sigma_{max}^2)\right\}. \end{aligned} \tag{23}$$

Using Theorem B.2 and Assumption B.2, the second term could be bounded as,

$$\begin{aligned} &||\hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \mathbf{X}(\boldsymbol{\Pi}_{-i}^T - \hat{\boldsymbol{\Pi}}_{-i}^T)\boldsymbol{\beta}_i||_\infty \\ &\leq (||(\hat{\boldsymbol{\Pi}}_{-i} - \boldsymbol{\Pi}_{-i})^T||_\infty + ||\boldsymbol{\Pi}_{-i}^T||_\infty)||\mathbf{X}^T \mathbf{P}_i \mathbf{X}(\boldsymbol{\Pi}_{-i}^T - \hat{\boldsymbol{\Pi}}_{-i}^T)\boldsymbol{\beta}_i||_\infty \\ &\leq (t_\Pi + ||\boldsymbol{\Pi}||_1)||\mathbf{B}||_1 \underset{j_1,j_2}{max} |\mathbf{X}_{j_1}^T \mathbf{P}_i \mathbf{X}(\boldsymbol{\Pi}_{j_2}^T - \hat{\boldsymbol{\Pi}}_{j_2}^T)| \\ &\leq \sqrt{n}(t_\Pi + ||\boldsymbol{\Pi}||_1)||\mathbf{B}||_1 \underset{j_2}{max} ||\mathbf{P}_i \mathbf{X}(\boldsymbol{\Pi}_{j_2}^T - \hat{\boldsymbol{\Pi}}_{j_2}^T)||_2 \\ &\leq \sqrt{n}(t_\Pi + ||\boldsymbol{\Pi}||_1)||\mathbf{B}||_1 \sqrt{d \vee \log(\frac{1}{\delta_{min}}) \vee ||\boldsymbol{\pi}||_{2max}^2}, \end{aligned} \tag{24}$$

Using the same way above, we have

$$\text{var}(X_{\cdot j}^T \mathbf{P}_i \boldsymbol{\xi}_i) \lesssim n.$$

Then,

$$\begin{aligned} &\mathbb{P}\left(||\hat{\boldsymbol{\Pi}}_{-i}^T \mathbf{X}^T \mathbf{P}_i \boldsymbol{\xi}_{-i} \boldsymbol{\epsilon}_i \boldsymbol{\beta}_i||_\infty \geq t\right) \\ &\leq \mathbb{P}\left((||(\hat{\boldsymbol{\Pi}}_{-i} - \boldsymbol{\Pi}_{-i})^T||_\infty + ||\boldsymbol{\Pi}_{-i}^T||_\infty)||\mathbf{B}||_1||\mathbf{X}^T \mathbf{P}_i \boldsymbol{\xi}_{-i} \boldsymbol{\epsilon}_i||_\infty \geq t\right) \\ &\leq \mathbb{P}\left(||\mathbf{X}^T \mathbf{P}_i \boldsymbol{\xi}_{-i} \boldsymbol{\epsilon}_i||_\infty \geq \frac{t}{(t_\Pi + ||\boldsymbol{\Pi}||_1)||\mathbf{B}||_1}\right) \\ &\leq 2Kd\ \exp\left\{-\left(\frac{t}{(t_\Pi + ||\boldsymbol{\Pi}||_1)||\mathbf{B}||_1}\right)^2 / (2n\tilde{\sigma}_{\max}^2)\right\}. \end{aligned} \tag{25}$$

With ν_i defined in Theorem B.3 and

$$t = \frac{1}{12} n\nu_i \|\hat{\boldsymbol{\omega}}_{S_i}\|_\infty$$

in the above inequalities, we then have

$$\begin{aligned}
&\mathbb{P}(T_{34} \geq \frac{1}{12} n\nu_i \|\hat{\boldsymbol{\omega}}_{S_i}\|_\infty) \leq \delta_{min}/2, \\
&\mathbb{P}(T_{36} \geq \frac{1}{12} n\nu_i \|\hat{\boldsymbol{\omega}}_{S_i}\|_\infty) \leq \delta_{min}/2, \\
&T_{35} \leq \frac{1}{12} n\nu_i \|\hat{\boldsymbol{\omega}}_{S_i}\|_\infty.
\end{aligned}$$

Conditioning on t_Π and using the bound obtained at the beginning of the discussion. We have that

$$\mathbb{P}(T_{34} + T_{35} + T_{36} \leq \frac{1}{4} n\nu_i \|\hat{\boldsymbol{\omega}}_{S_i}\|_\infty) \geq (1 - \delta_{min} - p \sum_{k=1}^{K} \delta^{(k)}).$$

This concludes the proof of Lemma B.3. □

Next, we follow the techniques in [17] to bound the prediction loss. Following the definition of the adaptive lasso, we have

$$\frac{1}{n} ||\mathbf{P}_i \mathbf{Y}_i - \mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i} \hat{\boldsymbol{\beta}}_i||_2^2 + \nu_i \hat{\boldsymbol{\omega}}_i^T |\hat{\boldsymbol{\beta}}_i|_1 \leq \frac{1}{n} ||\mathbf{P}_i \mathbf{Y}_i - \mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i} \boldsymbol{\beta}_i||_2^2 + \nu_i \hat{\boldsymbol{\omega}}_i^T |\boldsymbol{\beta}_i|_1. \tag{26}$$

We can rewrite the inequality as

$$\begin{aligned}
&\frac{1}{n} ||\mathbf{P}_i (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i}) \boldsymbol{\beta}_i + \mathbf{P}_i \boldsymbol{\epsilon}_i + \mathbf{P} \hat{\boldsymbol{\Upsilon}}_{-i} (\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i)||_2^2 + \nu_i \hat{\boldsymbol{\omega}}_i^T |\hat{\boldsymbol{\beta}}_i|_1 \\
&\qquad \leq \frac{1}{n} ||\mathbf{P}_i (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i}) \boldsymbol{\beta}_i + \mathbf{P}_i \boldsymbol{\epsilon}_i||_2^2 + \nu_i \hat{\boldsymbol{\omega}}_i^T |\boldsymbol{\beta}_i|_1, \\
&\frac{1}{n} ||\mathbf{P}_i (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i}) \boldsymbol{\beta}_i||_2^2 \\
&\qquad \leq \frac{2}{n} \{\hat{\boldsymbol{\Upsilon}}_{-i}^T \mathbf{P}_i [\boldsymbol{\epsilon}_i + (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i}) \boldsymbol{\beta_i}]\}^T (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) + \nu_i \hat{\boldsymbol{\omega}}_i^T |\boldsymbol{\beta}_i|_1 - \nu_i \hat{\boldsymbol{\omega}}_i^T |\hat{\boldsymbol{\beta}}_i|_1.
\end{aligned}$$

Adding $\frac{1}{2} \nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1$, and then multiplying n to both sides, together with Lemma B.3 and Assumption B.3, we have that

$$\begin{aligned}
&||\mathbf{P} \hat{\boldsymbol{\Upsilon}}_{-i} (\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i)||_2^2 + \frac{n}{2} \nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 \\
\leq\ & 2||\{\hat{\boldsymbol{\Upsilon}}_{-i}^T \mathbf{P}_i [\boldsymbol{\epsilon}_i + (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i}) \boldsymbol{\beta_i}]\}^T ||_\infty ||(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_1 + \frac{n}{2} \nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 \\
& + n\nu_i \hat{\boldsymbol{\omega}}_i^T |\boldsymbol{\beta}_i|_1 - n\nu_i \hat{\boldsymbol{\omega}}_i^T |\hat{\boldsymbol{\beta}}_i|_1 \\
\leq\ & n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 + n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\boldsymbol{\beta}_{S_i}||_1 - n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty (||\hat{\boldsymbol{\beta}}_{S_i}||_1 + ||\hat{\boldsymbol{\beta}}_{S_i^C}||_1) \\
=\ & n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_{S_i} - \boldsymbol{\beta}_{S_i}||_1 + n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\boldsymbol{\beta}_{S_i}||_1 - n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_{S_i}||_1 \\
\leq\ & 2n\nu_i ||\hat{\boldsymbol{\omega}}_{S_i}||_\infty ||\hat{\boldsymbol{\beta}}_{S_i} - \boldsymbol{\beta}_{S_i}||_1.
\end{aligned} \tag{27}$$

Comparing the two sides, we have

$$||\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i||_1 \leq 4||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1, \tag{28}$$

$$||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c} - \boldsymbol{\beta}_{\mathcal{S}_i^c}||_1 \leq 3||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1. \tag{29}$$

which indicates that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ satisfies the condition in Lemma B.2. So we have

$$\begin{aligned}
&||\mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 \\
\leq\ & \frac{3}{2} \nu_i n \|\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}\|_\infty ||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_1 \leq \frac{3}{2} \nu_i n \|\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}\|_\infty \sqrt{|\mathcal{S}_i|} ||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_2 \\
\leq\ & \frac{3}{2} \nu_i n \|\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}\|_\infty \sqrt{|\mathcal{S}_i|} \frac{||\mathbf{P}_i \hat{\boldsymbol{\Upsilon}}_{-i} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2}{\sqrt{n\lambda_0}}.
\end{aligned} \tag{30}$$

Thus we can bound the error term as,

$$\frac{1}{n}||\mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2 \leq \frac{9}{4}\frac{|\mathcal{S}_i|}{\lambda_0}||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty^2\nu_i^2.$$

Using the value of ν_i, we get that

$$\begin{aligned}&\frac{1}{n}||\mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2\\ \lesssim\ &\frac{|S_i|}{n}(\sqrt{d\frac{d\vee\log(\frac{1}{\delta_{min}})\vee||\boldsymbol{\pi}||_{2max}^2}{n_{min}}}\\ &+||\boldsymbol{\Pi}||_1)(||\mathbf{B}||_1\vee 1)\times d\vee\log(\frac{1}{\delta_{min}})\vee||\boldsymbol{\pi}||_{2max}^2\vee log\frac{d}{\delta_{min}}.\end{aligned}$$

Applying Lemma B.2 again, we derive that

$$\begin{aligned}&||\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i||_2^2\\ \leq\ &\frac{1}{\lambda_0 n}||\mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2\\ \lesssim\ &\frac{|S_i|}{n}(\sqrt{d\frac{d\vee\log(\frac{1}{\delta_{min}})\vee||\boldsymbol{\pi}||_{2max}^2}{n_{min}}}\\ &+||\boldsymbol{\Pi}||_1)\times(||\mathbf{B}||_1\vee 1)d\vee\log(\frac{1}{\delta_{min}})\vee||\boldsymbol{\pi}||_{2max}^2\vee log\frac{d}{\delta_{min}}.\end{aligned}$$

The above prediction and estimation bounds condition on the bound of t_Π and restricted eigenvalue condition for prediction matrices, which hold with probability at least $1-\delta_{min}-p\sum_{k=1}^K\delta^k$. The proof of Theorem B.3 is then completed.

B.2.4 Proof of Theorem B.4

Lemma B.4. *Suppose that, for each node i and function f_n defined in (9),*

$$|\mathcal{S}_i|\leq\frac{\eta}{(\eta+2)\tau_i}\frac{n}{f_n}. \tag{31}$$

Under Assumptions B.1–B.4, we have that, with the probability at least $1-p\sum_{k=1}^K$*,*

$$||\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}^{-1}||_\infty\leq 1-\eta^2/2.$$

Proof of Lemma B.4. Following the proof of Lemma B.2, we have showed that, with probability at least $1-p\sum_{k=1}^K\delta^{(k)}$,

$$\max_{l,r}\frac{1}{n}|\hat{\Sigma}_{l,r}-\Sigma_{l,r}|\leq f_n/n.$$

where the subscript l,r denotes the elements in the l-th row and r-th column in the corresponding matrix.

Consider the part indexed by set $\mathcal{S}_i$ with the assumption in the lemma, we have that

$$||\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}-\Sigma_{\mathcal{S}_i,\mathcal{S}_i}||_\infty\leq|\mathcal{S}_i|\frac{f_n}{n}\leq\frac{\eta}{(\eta+2)\tau_i}.$$

In a similar manner, we have,

$$||\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}-\Sigma_{\mathcal{S}_i^c,\mathcal{S}_i}||_\infty\leq|\mathcal{S}_i|\frac{f_n}{n}\leq\frac{\eta}{(\eta+2)\tau_i}. \tag{32}$$

We bound the error of matrix inversion as described in Horn and Johnson [12] and obtain that,

$$\begin{aligned}||\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}^{-1}||_\infty &\leq ||\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}^{-1}-\Sigma_{\mathcal{S}_i,\mathcal{S}_i}^{-1}||_\infty+||\Sigma_{\mathcal{S}_i,\mathcal{S}_i}^{-1}||_\infty\\ &\leq \frac{\tau_i^2||\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}-\Sigma_{\mathcal{S}_i,\mathcal{S}_i}||_\infty}{1-\tau_i||\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}-\Sigma_{\mathcal{S}_i,\mathcal{S}_i}||_\infty}+\tau_i\\ &\leq \frac{\tau_i}{1-\tau_i|\mathcal{S}_i|f_n/n}\leq\frac{\eta+2}{2}\tau_i. \end{aligned}\tag{33}$$

Note the following decomposition,

$$\begin{aligned}\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i^c,\mathcal{S}_i}\Sigma^{-1}_{\mathcal{S}_i,\mathcal{S}_i} &= (\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i^c,\mathcal{S}_i})\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i} \\ &+ \Sigma_{\mathcal{S}_i^c,\mathcal{S}_i}\Sigma^{-1}_{\mathcal{S}_i,\mathcal{S}_i}\left(\Sigma_{\mathcal{S}_i,\mathcal{S}_i} - \hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}\right)\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}.\end{aligned}$$

Then, collecting (32), (32), (33) and Assumption B.4, we have that

$$\begin{aligned}||\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i^c,\mathcal{S}_i}\Sigma^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty &\leq ||\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i^c,\mathcal{S}_i}||_\infty||\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty \\ &+ ||\Sigma_{\mathcal{S}_i^c,\mathcal{S}_i}\Sigma^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty||\hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}||_\infty||\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty \\ &\leq \eta - \frac{1}{2}\eta^2.\end{aligned}$$

Together with Assumption B.4, we derive that

$$||\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty \leq 1 - \frac{1}{2}\eta^2.$$

The proof of Lemma B.4 is then completed. □

Denote $W_i = diag(\hat{\omega}_i)$. Applying the KKT condition, we get that

$$-\frac{2}{n}\hat{\boldsymbol{\Upsilon}}^T_{-i}\mathbf{P}_i(\mathbf{P}_i\mathbf{Y}_i - \mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}\hat{\boldsymbol{\beta}}_i) + \nu_i W_i \alpha_i = 0, \tag{34}$$

where $\alpha_i \in \mathbb{R}^{Kp-K}$, satisfying $\alpha_{ij}\ I(\hat{\boldsymbol{\beta}}_{ij} \neq 0) = sign(\hat{\boldsymbol{\beta}}_{ij})$.
Using the equation $\mathbf{P}_i\mathbf{Y}_i = \mathbf{P}_i\boldsymbol{\Upsilon}_{-i}\boldsymbol{\beta}_i + \mathbf{P}_i\boldsymbol{\epsilon}_i$, we have that

$$\begin{aligned}&\hat{\boldsymbol{\Upsilon}}^T_{-i}\mathbf{P}_i(\mathbf{P}_i\mathbf{Y}_i - \mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}\hat{\boldsymbol{\beta}}_i) \\ =\ & \hat{\boldsymbol{\Upsilon}}^T_{-i}\mathbf{P}_i[\mathbf{P}_i\boldsymbol{\epsilon}_i + \mathbf{P}_i(\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i})\hat{\boldsymbol{\beta}}_i) + \mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}(\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i)] \\ =\ & \underbrace{\hat{\boldsymbol{\Upsilon}}^T_{-i}\mathbf{P}_i(\boldsymbol{\epsilon}_i + (\boldsymbol{\Upsilon}_{-i} - \hat{\boldsymbol{\Upsilon}}_{-i})\hat{\boldsymbol{\beta}}_i)}_{T_{41}} - \underbrace{\hat{\boldsymbol{\Upsilon}}^T_{-i}\mathbf{P}_i\hat{\boldsymbol{\Upsilon}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)}_{T_{42}}.\end{aligned} \tag{35}$$

With the definition of $\hat{\Sigma}_i$, (34) implies that,

$$\frac{1}{n}T_{41} - \hat{\Sigma}_i(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) = \frac{1}{2}\nu_i W_i \alpha_i, \tag{36}$$

We consider the rows indexed by $\mathcal{S}_i$ and $\mathcal{S}_i^c$ in both sides of equation (36) which can be decomposed as

$$\begin{cases}\frac{1}{n}T_{41,\mathcal{S}_i} - \hat{\Sigma}_{\mathcal{S}_i,\mathcal{S}_i}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}) = \frac{1}{2}\nu_i W_{\mathcal{S}_i}\alpha_{\mathcal{S}_i}, \\ \frac{1}{n}T_{41,\mathcal{S}_i^c} - \hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}) = \frac{1}{2}\nu_i W_{\mathcal{S}_i^c}\alpha_{\mathcal{S}_i^c},\end{cases} \tag{37}$$

where $T_{41,\mathcal{S}_i}$ and $T_{41,\mathcal{S}_i^c}$ denote the $\mathcal{S}_i$ and $\mathcal{S}_i^c$ rows for T_{41}, respectively.
From the first equation of (37), we can equate the error of estimation indexed by $\mathcal{S}_i$ as,

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i} = \hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}(\frac{1}{n}T_{41,\mathcal{S}_i} - \frac{1}{2}\nu_i W_{\mathcal{S}_i}\alpha_{\mathcal{S}_i}). \tag{38}$$

Scaling ν_i by $\frac{1}{2}\frac{4-\eta^2}{\eta^2}$ and using the same method in the proof of Lemma B.3, we have at least probability at least $1 - \delta_{min} - p\sum_{k=1}^K \delta^{(k)}$, such that

$$||T_{41}||_\infty \leq \frac{1}{2}\frac{\eta^2}{4-\eta^2}n\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty.$$

Therefore, we can bound the infinity norm of the above error as,

$$\begin{aligned}
||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_\infty &\leq ||\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty(\frac{1}{n}||T_{41}||_\infty + \frac{1}{2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty) \\
&\leq \frac{\eta+2}{2}\tau_i\frac{2}{4-\eta^2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty \\
&= \frac{\nu_i\tau_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty}{2-\eta} \leq \min_{j\in\mathcal{S}_i}|\boldsymbol{\beta}_{ij}|.
\end{aligned}$$

It indicates the largest absolute error is no larger than the minimal absolute signal, which leads to

$$sign(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}) = sign(\boldsymbol{\beta}_{\mathcal{S}_i}).$$

Next we will validate the second equation of (37) using the decomposition of error in (38), we have that

$$\begin{aligned}
&||\frac{1}{n}T_{41,\mathcal{S}_i^c} - \hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}(\frac{1}{n}T_{41,\mathcal{S}_i} - \frac{1}{2}\nu_i W_{\mathcal{S}_i}\alpha_{\mathcal{S}_i})||_\infty \\
\leq\ & \frac{1}{2}\frac{\eta^2}{4-\eta^2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty + ||\hat{\Sigma}_{\mathcal{S}_i^c,\mathcal{S}_i}\hat{\Sigma}^{-1}_{\mathcal{S}_i,\mathcal{S}_i}||_\infty(\frac{1}{2}\frac{\eta^2}{4-\eta^2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty + \frac{1}{2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty) \\
\leq\ & (\frac{1}{2}\frac{\eta^2}{4-\eta^2} + \frac{2-\eta^2}{2}\frac{2}{4-\eta^2})\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty = \frac{1}{2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i}||_\infty \\
\leq\ & \frac{1}{2}\nu_i||\hat{\boldsymbol{\omega}}_{\mathcal{S}_i^c}||_\infty.
\end{aligned}$$

From the construction above, we have proved $sign(\hat{\boldsymbol{\beta}}_i) = sign(\boldsymbol{\beta}_i)$ and thus complete the proof of Theorem B.4.

C Proofs of the Theorems in the Main Text

C.1 Proof of Theorem 4.1

Theorem 4.1 directly follows Theorem B.3 and B.4.

C.2 Proof of Theorem 4.2

Denote $\boldsymbol{\zeta}_i^{(k)} = \mathbf{X}_{\mathcal{I}_i}^{(k)}\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)} + \boldsymbol{\epsilon}_i^{(k)}$, then $\mathbf{Y}_i^{(k)} = \mathbf{Y}_{-i}^{(k)} + \boldsymbol{\zeta}_i^{(k)}$ according to (1). We have

$$\begin{aligned}
|R_i^{2(k)} - R_{0i}^{2(k)}| &= \frac{|||\mathbf{Y}_{-i}^{(k)}\hat{\boldsymbol{\gamma}}_i^{(k)} - \mathbf{Y}_i^{(k)}||_2^2 - ||\mathbf{Y}_{-i}^{(k)}\boldsymbol{\gamma}_i^{(k)} - \mathbf{Y}_i^{(k)}||_2^2|}{||\mathbf{Y}_i^{(k)}||_2^2} \\
&= \frac{[\mathbf{Y}_{-i}^{(k)}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)})]^T[\mathbf{Y}_{-i}^{(k)}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)}) - 2\boldsymbol{\zeta}_i^{(k)}]}{||\mathbf{Y}_i^{(k)}||_2^2} \\
&= \frac{|||\mathbf{Y}_{-i}^{(k)}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)})||_2^2 - 2[\mathbf{Y}_{-i}^{(k)}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)})]^T\boldsymbol{\zeta}_i^{(k)}|}{||\mathbf{Y}_i^{(k)}||_2^2}. \qquad (39)
\end{aligned}$$

By Cauchy-Schwarz inequality, we have that

$$[\mathbf{Y}_{-i}^{(k)}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)})]^T\boldsymbol{\zeta}_i^{(k)} \leq ||\mathbf{Y}_{-i}^{(k)}(\hat{\boldsymbol{\gamma}}_i^{(k)} - \boldsymbol{\gamma}_i^{(k)})||_2||\boldsymbol{\zeta}_i^{(k)}||_2, \qquad (40)$$

where

$$||\boldsymbol{\zeta}_i^{(k)}||_2^2 \lesssim ||\mathbf{X}_{\mathcal{I}_i}^{(k)}\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)}||_2||\boldsymbol{\epsilon}_i^{(k)}||_2 \leq \sqrt{n}||\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)}||_2||\boldsymbol{\epsilon}_i^{(k)}||_2. \qquad (41)$$

Note that $||\boldsymbol{\epsilon}_i^{(k)}||_2^2$ follows the χ^2 distribution, so we have that, with probability at least $1 - \delta_{min}$

$$||\boldsymbol{\epsilon}_i^{(k)}||_2 \leq \sqrt{n^{(k)} + 2\sqrt{n^{(k)}log(\frac{1}{\delta_{min}})} + 2log(\frac{1}{\delta_{min}})}. \qquad (42)$$

Furthermore, we have $||\mathbf{Y}_i||_2^{2(k)} \asymp n^{(k)}$ due to normaliztion. Then collecting Theorem (4.1), equations (39), (40), (41), (42), we have that, with probability at least $1-2\delta_{min}-p\delta^{(k)}$,

$$\begin{aligned} |R_i^{2(k)}-R_{0i}^{2(k)}| &\lesssim \frac{|S_i|g_n\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee log\frac{d}{\delta_{min}}\}}{n^{(k)2}} \\ &+ \frac{\sqrt{|S_i|g_n\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee log\frac{d}{\delta_{min}}\}\sqrt{n^{(k)}}||\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)}||_2 h_n}}{n^{(k)}} \end{aligned} \tag{43}$$

where $h_n = \sqrt{n^{(k)}+2\sqrt{n^{(k)}log(\frac{1}{\delta_{min}})}+2log(\frac{1}{\delta_{min}})}$.

We can derive the same bound for the whole system, with probability at least $1-p(2\delta_{min}+p\delta^{(k)})$, we have that

$$\sum_{i=1}^{p}|R_i^{2(k)}-R_{0i}^{2(k)}| \tag{44}$$

$$\begin{aligned} &\lesssim \frac{|S_i|g_n\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee log\frac{d}{\delta_{min}}\}}{n^{(k)}} \\ &\quad + \frac{\sqrt{|S_i|g_n\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee log\frac{d}{\delta_{min}}\}\sqrt{n^{(k)}}||\boldsymbol{\phi}_{\mathcal{I}_i}^{(k)}||_2 h_n}}{n^{(k)}}. \end{aligned} \tag{45}$$

In the following proofs, we will bound the error of the C^2 statistics.

$$\begin{aligned} &\sum_{j=1}^{p}|C_j^{2(k)}-C_{j0}^{2(k)}| \\ &= \sum_{j=1}^{p}\sum_{i=1}^{p}\frac{|||\mathbf{Y}_j^{(k)}\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\mathbf{Y}_i^{(k)}||_2^2-||\mathbf{Y}_j^{(k)}\boldsymbol{\gamma}_{ij}^{(k)}-\mathbf{Y}_i^{(k)}||_2^2|}{||\mathbf{Y}_i^{(k)}||_2^2} \\ &= \sum_{j=1}^{p}\sum_{i=1}^{p}\frac{|[\mathbf{Y}_j^{(k)}(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})]^T[\mathbf{Y}_j^{(k)}(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})-2\mathbf{Y}_i^{(k)}+2\mathbf{Y}_j^{(k)}\boldsymbol{\gamma}_{ij}^{(k)})]|}{||\mathbf{Y}_i^{(k)}||_2^2}. \end{aligned} \tag{46}$$

Note that

$$\begin{aligned} &\sum_{j=1}^{p}\sum_{i=1}^{p}|[\mathbf{Y}_j^{(k)}(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})]^T[\mathbf{Y}_j^{(k)}(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})-2\mathbf{Y}_i^{(k)}+2\mathbf{Y}_j^{(k)}\boldsymbol{\gamma}_{ij}^{(k)})]| \\ &= \sum_{j=1}^{p}\sum_{i=1}^{p}|||\mathbf{Y}_j^{(k)}(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})||_2^2-2(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})\mathbf{Y}_j^{(k)T}\mathbf{Y}_i^{(k)} \\ &\quad +2(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})\boldsymbol{\gamma}_{ij}^{(k)}\mathbf{Y}_j^{(k)T}\mathbf{Y}_j^{(k)}|, \end{aligned} \tag{47}$$

and

$$\mathbf{Y}_j^{(k)T}\mathbf{Y}_i^{(k)} \le ||\mathbf{Y}_j^{(k)}||_2||\mathbf{Y}_i^{(k)}||_2 \asymp n^{(k)}.$$

Then with equation (28), we have

$$\begin{aligned} &\sum_{j=1}^{p}|C_j^{2(k)}-C_{j0}^{2(k)}| \\ &\lesssim \sum_{j=1}^{p}\sum_{i=1}^{p}\frac{||\mathbf{Y}_j^{(k)}(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})||_2^2+n^{(k)}|\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)}|(||B||_1 \vee 1)}{n^{(k)}} \\ &\lesssim \sum_{i=1}^{p}\frac{\sum_{j=1}^{p}||\mathbf{Y}_j^{(k)}||_2^2||(\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)})||_2^2+n^{(k)}||\hat{\boldsymbol{\gamma}}_i^{(k)}-\boldsymbol{\gamma}_i^{(k)}||_1(||B||_1 \vee 1)}{n^{(k)}} \\ &\lesssim \sum_{i=1}^{p}\frac{n^{(k)}||\hat{\boldsymbol{\gamma}}_{ij}^{(k)}-\boldsymbol{\gamma}_{ij}^{(k)}||_2^2+n^{(k)}\sqrt{|\mathcal{S}_i|}||\hat{\boldsymbol{\gamma}}_i^{(k)}-\boldsymbol{\gamma}_i^{(k)}||_2(||B||_1 \vee 1)}{n^{(k)}}. \end{aligned} \tag{48}$$

Applying theorem 4.1 in the whole system, we have that, with probability at least $1 - p(\delta_{min} + p\delta^{(k)})$

$$\begin{aligned}
&\sum_{j=1}^{p} |C_j^{2(k)} - C_{j0}^{2(k)}| \\
&\quad \lesssim \quad \frac{|\mathcal{S}_i| g_n\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee log\frac{d}{\delta_{min}}\}}{n^{(k)}} \\
&\qquad + \frac{\sqrt{n^{(k)}|\mathcal{S}_i|^2 g_n\{d \vee \log(\frac{1}{\delta^{(k)}}) \vee ||\boldsymbol{\pi}_i||_2^{2(k)} \vee \log\frac{d}{\delta_{min}}\}}(||B||_1 \vee 1)}{n^{(k)}}. \qquad (49)
\end{aligned}$$

Equation (44, 49) lead to the discussion in theorem 4.2 and the equations simplifies to theorem 4.2.

D Details in the Simulation Study

In Figure 6, we present one plot of the three networks used for the simulation study that corresponds to Figure 4, showing the causal relations in the baseline network DCG III in black, the deviated causal effects DCG I vs. DCG III, DCG II vs. DCG III, and DCG I vs. DCG II in blue, red, and green, respectively.

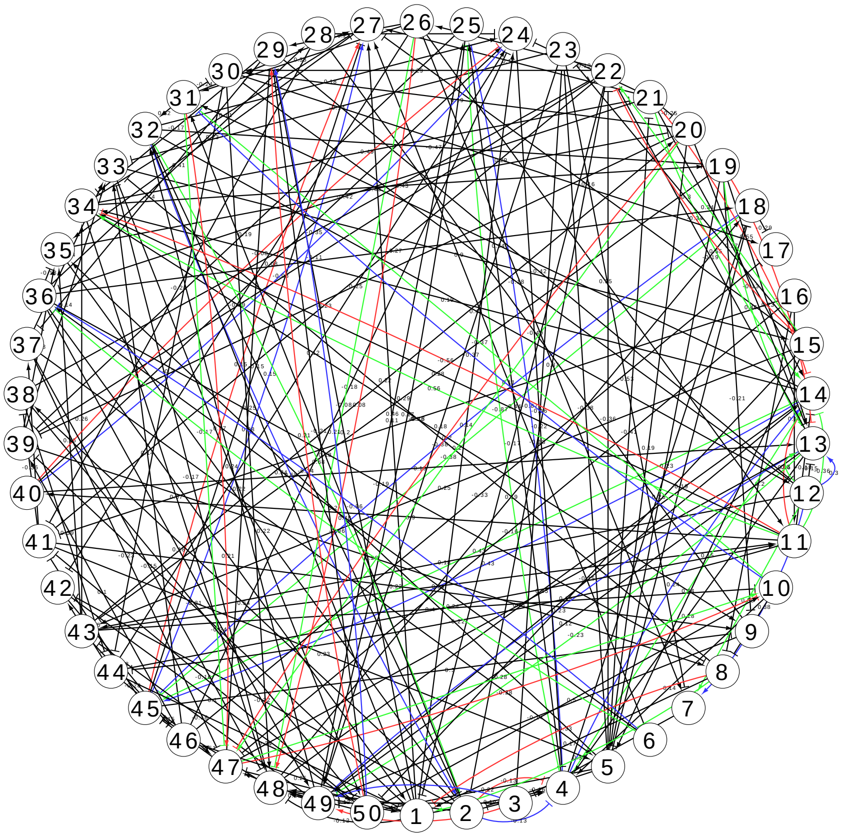


Figure 6: Plot of all causal effects among three networks used for the simulation study with baseline DCG III, DCG I vs. DCG III, DCG II vs. DCG III, and DCG I vs. DCG II in black, blue, red, and green, respectively.

E Limitations

Although we have developed a limited-information likelihood method to avoid optimizing too many model parameters as the full-information likelihood method does, the proposed method may still be challenged by large K and massive total sample size n when there are too many cohorts to compare. When K is too large, each task in the algorithm (identifying and estimating causal effects for a single responder) has to estimate $K(p-1)$ parameters with an $n \times (K(p-1))$ design matrix, possibly demanding a large amount of memory.

We have developed our algorithm for the case to compare all other networks to a single baseline network and provide theoretical analysis. In practice, we may be interested in the deviated effects of each network from the average effects. Our algorithm can be adopted for such a case. However, it is challenging to develop an appropriate theoretical analysis for this case, and deserves further study.