Electronic Journal of Statistics
Vol. 14 (2020) 4133–4164
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1767

Reconstruction of a directed acyclic graph with intervention*

Si Peng and Xiaotong Shen

School of Statistics, University of Minnesota
313 Ford Hall, 224 Church St SE
Minneapolis, MN 55455
e-mail: pengx179@umn.edu; xshen@umn.edu

Wei Pan

Division of Biostatistics, University of Minnesota
420 Delaware St. S.E.
Minneapolis, MN 55455
e-mail: panxx014@umn.edu

Abstract: Identification of causal relations among variables is central to many scientific investigations, as in regulatory network analysis of gene interactions and brain network analysis of effective connectivity of causal relations between regions of interest. Statistically, causal relations are often modeled by a directed acyclic graph (DAG), and hence that reconstruction of a DAG's structure leads to the discovery of causal relations. Yet, reconstruction of a DAG's structure from observational data is impossible because a DAG Gaussian model is usually not identifiable with unequal error variances. In this article, we reconstruct a DAG's structure with the help of interventional data. Particularly, we construct a constrained likelihood to regularize intervention in addition to adjacency matrices to identify a DAG's structure, subject to an error variance constraint to further reinforce the model identifiability. Theoretically, we show that the proposed constrained likelihood leads to identifiable models, thus correct reconstruction of a DAG's structure through parameter estimation even with unequal error variances. Computationally, we design efficient algorithms for the proposed method. In simulations, we show that the proposed method enables to produce a higher accuracy of reconstruction with the help of interventional observations.

MSC2020 subject classifications: Primary 62-09.
Keywords and phrases: Causal relations, constrained likelihood, intervention, reconstruction identifiability.

Received May 2020.

Contents

*The authors thank the editor, the associate editor and anonymous referees for helpful comments and suggestions. Research supported in part by NSF grants DMS-1712564, DMS-1721216, DMS-1952539, and NIH grants 1R01GM126002, 2R01HL105397, 1R01AG065636, R01AG069895.

1. Introduction

Directed acyclic graph (DAG) models are useful to describe pairwise causal relations between random variables, defined by a certain Markov property [5], with each node representing one variable and each directed edge representing the corresponding pairwise causal relation. DAG models have been widely used in gene and social networks [7, 19]. To identify causal relations, intervention observations are usually collected in addition to observational attributes [16]. The central topic this article addresses is the reconstruction of a DAG model based on interventional data and pertinent issues with respect to the effect of the intervention on the reconstruction of a DAG's structure.

In the literature, it is generally believed that interventions may help the reconstruction of a DAG's structure, particularly when a DAG model is not identifiable from data, that is, DAGs in a Markov equivalence class are not distinguishable based on observational data alone [16]. In biological experiments, for instance, intervention occurs in a form of randomized treatments in a clinical trial or a form of gene knockdown or knockout experiments in systems biology. In such a situation, some or all system variables are controlled, permitting direction estimation of ambiguous edges connecting to these controlled variables. Yet exactly how intervention impacts reconstruction of a DAG's structure remains unknown. Consequently, it is practically important to design a reconstruction

method for interventional data, permitting the identification of a DAG's structure. Most existing methods for intervention [6, 10, 8] assume known intervention, that is, affected variables of the intervention are known *a priori* before data collection; see [9] for references therein. However, assuming known intervention is impractical, as in system biology, where the effect of various chemicals intervening a system cannot be precisely known. To our knowledge, one exception is a Bayesian method of [4], which is designed for a low-dimensional problem without theoretical guarantee, due to the super-exponential complexity in the number of nodes.

In this article, we propose a novel variance constraint on interventions to fully identify the DAG structure in the framework of unknown intervention. Theoretically, we show in Theorem 1 that a DAG structure is fully identifiable under the constraint, which is otherwise only possible when all the error variances are not the same [18]. Moreover, we propose a constrained maximum likelihood to seek the most efficient interventions by sparsity pursuit, leading to an identifiable reconstruction of a DAG's structure. Computationally, we develop an efficient algorithm to solve nonconvex minimization subject to the quadratic variance constraint based on the alternating direction method of multipliers (ADMM) [2]. In simulations, we investigate the impact of the intervention on reconstruction and compare the proposed method with its counterpart without intervention. Overall, the proposed method performs well.

This article is organized into seven sections. Section 2 introduces the proposed method and discusses the issue of identifiability due to intervention, followed by the computational development in Section 3. Section 4 establishes the consistency theorems of the proposed method. Section 5 performs some simulations to study the intervention effect, and two real datasets are analyzed. Section 6 summarizes the results. Finally, the Appendix A contains technical details and proofs.

2. Method

Consider a causal model consisting of p random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ described by a DAG, with each node representing one variable and directed edges encoding causal relations between any two variables, where $^\top$ denotes the transpose. This model factorizes the joint distribution of $\boldsymbol{Y}$, $P(\boldsymbol{Y})$ into a product of conditional distributions of each variable given its parents: $P(\boldsymbol{Y}) = \prod_{j=1}^{p} P(Y_j|\mathbf{pa}_j)$, where $\mathbf{pa}_j$ denotes the parent set of Y_j and is defined to be empty if Y_j has no parents. This factorization is known as the local Markov property [5].

A DAG is modeled by structural equations as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{D}), \tag{1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p)^\top$ represents the latent or unexplained error, $\boldsymbol{D} = \mathrm{Diag}(\sigma_1^2, \cdots, \sigma_p^2)$ is the error covariance matrix and $\boldsymbol{A} = (A_{ij})_{p\times p}$ is an adjacency matrix that uniquely determines a DAG. Here $A_{ij} \neq 0$ encodes an

edge from node j to node i. In (1), the inverse covariance matrix of $\boldsymbol{Y}$ is $\boldsymbol{\Omega} = (\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{A})$, where $\boldsymbol{I}$ is the identity matrix.

When $\sigma_1 = \cdots = \sigma_p$, (1) is identifiable [18].Then a DAG's structure can be reconstructed by estimating $\boldsymbol{A}$. However, when $\sigma_1 = \cdots = \sigma_p$ breaks down, (1) is usually not identifiable, which means that $\boldsymbol{A}$ is not estimable.

2.1. Intervention or covariate models and variance constraint

To treat non-identifiability in an observational study, consider a model consisting of W intervention variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_W)^\top$, where the outcome of $\boldsymbol{Y}$ is observed with intervention variables $\boldsymbol{X}$ that may be non-informative.

After incorporating the intervention variables into (1), we obtain that

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \boldsymbol{D}), \tag{2}$$

where $\boldsymbol{A}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{D}$ are defined as in (1), $\boldsymbol{B} = (B_{jw})_{p\times W}$ is an unknown intervention coefficient matrix, whose jwth entry B_{jw} indicates the directional strength of the intervention of X_w on Y_j. When $B_{jw} = 0$; $j = 1, \cdots, p$, there is no intervention of X_w on Y_i, and thus X_w is non-informative. Note that (2) becomes a causal model with covariates $\boldsymbol{X}$.

In the situation of unequal error variances, with the help of the intervention, we may impose constraints to achieve model identifiability, which otherwise is impossible [18]. Assume that $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma}_X)$, which is independent of $\boldsymbol{\epsilon}$. Without loss of generality, assume that $\boldsymbol{\Sigma}_X = \boldsymbol{I}$ subsequently because we can reparametrize $\boldsymbol{X}$ as $\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{X}$. Then under (2), $\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega} = (\boldsymbol{I} - \boldsymbol{A})^\top(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{D})^{-1}(\boldsymbol{I} - \boldsymbol{A})$. In (2), we impose the variance constraint:

$$\mathrm{Var}(\boldsymbol{B}\boldsymbol{X} + \boldsymbol{\epsilon}) = \theta\boldsymbol{I}, \text{ or } \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{D} = \theta\boldsymbol{I}, \tag{3}$$

where $\theta > 0$ is a parameter to be estimated.

More details are deferred to Sections 3 and 4.

Theorem 1 (Identifiability)**.** *Assume that* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ *is independent of* $\boldsymbol{\epsilon}$ *in* (2)*, and* $A_{jk} \neq 0$ *for all* k *which is a parent of* j*;* $j = 1, 2, \cdots, p$*. Under* (3)*,* $\boldsymbol{A}$ *is identifiable from the distribution of* $(\boldsymbol{Y}, \boldsymbol{X})$*.*

As suggested by Theorem 1, $\boldsymbol{A}$ in (2) becomes identifiable when (3) is imposed on $\boldsymbol{B}$, which is otherwise impossible. Note, however, that interventions $\boldsymbol{B}$ leading to identifiable $\boldsymbol{A}$ may not be unique. In what is to follow, we impose a sparsity constraint to identify a most sparse $\boldsymbol{B}$ in terms of the number of nonzero elements of $\boldsymbol{B}$.

2.2. Constrained maximum likelihood

This section estimates $\boldsymbol{A}$ subject to the DAG requirement while seeking a most sparse $\boldsymbol{B}$ subject to (3). Consequently, the smallest set of informative intervention variables can be identified through $\boldsymbol{B}$.

Under (2), two data matrices $(y_{ij})_{n\times p}$ and $(x_{iw})_{n\times W}$ are observed, with n representing the sample size. Then the negative log likelihood is

$$l(\boldsymbol{A},\boldsymbol{B},\boldsymbol{D}) = \sum_{j=1}^{p}\Big[-\frac{n}{2}\log\sigma_j^2 + \frac{1}{2\sigma_j^2}\sum_{i=1}^{n}\Big(y_{ij} - \sum_{k\neq j}A_{jk}y_{ik} - \sum_{w=1}^{W}B_{jw}x_{iw}\Big)^2\Big]. \quad (4)$$

To identify nonzero entries of $\boldsymbol{A}$ and $\boldsymbol{B}$, we impose sparsity constraints to regularize:

$$\sum_{1\le j\neq l\le p} I(A_{jl}\neq 0)\le K_1, \qquad \sum_{1\le j\le p,1\le l\le W} I(B_{jl}\neq 0)\le K_2, \quad (5)$$

where K_1 and K_2 are nonnegative integer-valued tuning parameters. Note that the constraint on $\boldsymbol{B}$ removes zero entries thus zero-columns of $\boldsymbol{B}$, which can be regarded as selection of intervention variables. To reinforce the DAG requirement, we impose acyclic constraints [27] to reinforce the DAG requirement to ensure no loops to occur:

$$\lambda_{ik} + I(j\neq k) - \lambda_{jk} \ge I(A_{ij}\neq 0); i,j,k = 1,\cdots,p, i\neq j, \quad (6)$$

where $\boldsymbol{\lambda} = \{\lambda_{jl}\}_{p\times p}$ is a dual variable matrix.

For computation, we replace the indicator functions in (5) and (6) by its computational surrogate $J_\tau(z) = \min(\frac{|z|}{\tau}, 1)$ [20] to circumvent the difficulty of non-discontinuity in optimization. This yields

$$\sum_{1\le j<l\le p} \mathrm{J}_\tau(A_{jl}) \le K_1, \qquad \sum_{1\le j\le p,1\le l\le W} \mathrm{J}_\tau(B_{jl}) \le K_2, \quad (7)$$

$$\lambda_{ik} + I(j\neq k) - \lambda_{jk} \ge J_\tau(A_{ij}); i,j,k = 1,\cdots,p, i\neq j, \quad (8)$$

where $\mathrm{J}_\tau(z)$ approximates the indicator function as $\tau\to 0^+$.

Minimizing (4) in $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{D})$ subject to (7), (8), and (3) yields the constrained maximum likelihood estimate (CMLE):

$$\begin{gathered}\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{D}} l(\boldsymbol{A},\boldsymbol{B},\boldsymbol{D}) = \\ \sum_{j=1}^{p}\Big[-\tfrac{n}{2}\log\sigma_j^2 + \tfrac{1}{2\sigma_j^2}\sum_{i=1}^{n}\Big(y_{ij} - \sum_{k\neq j}A_{jk}y_{ik} - \sum_{w=1}^{W}B_{jw}x_{iw}\Big)^2\Big], \\ \text{subj to} \quad \sum_{1\le j<l\le p}\mathrm{J}_\tau(A_{jl}) \le K_1, \quad \sum_{1\le j\le p,1\le l\le W}\mathrm{J}_\tau(B_{jl}) \le K_2, \\ \lambda_{ik} + I(j\neq k) - \lambda_{jk} \ge J_\tau(A_{ij}); i,j,k = 1,\cdots,p, i\neq j, \\ \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{D} = \theta\boldsymbol{I}, \end{gathered} \quad (9)$$

where (K_1, K_2, τ) are tuning parameters.

3. Computation

This section develops a computational strategy to solve (9). First, θ is estimated by $\hat\theta$ through an estimate $\hat{\boldsymbol{A}}$ of $\boldsymbol{A}$ from the method in [27], that is,

$$\hat\theta = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\Big(y_{ij} - \sum_{k\neq j,k=1}^{p}\hat A_{jk}y_{ik}\Big)^2, \quad (10)$$

where $\{\hat{A}_{jk}\}_{j,k=1,\cdots,p}$ are obtained by a constrained maximum likelihood estimate with the sparsity constraint and acyclic constraint, based on the structural equation model (1) with observational data $\boldsymbol{Y}$ alone. Then we solve (9) with θ replaced by $\hat{\theta}$ using a blockwise coordinate descent alternating between two blocks $(\boldsymbol{A}, \boldsymbol{B})$ and $\boldsymbol{D}$ until convergence. In particular, the $(\boldsymbol{A}, \boldsymbol{B})$-block is solved via a difference convex (DC) programming followed by an alternating direction method of multipliers (ADMM), while the $\boldsymbol{D}$-block is updated by gradient descent. More details are further discussed subsequently.

3.1. Optimization subject to the variance constraint

To deal with the variance constraint in (9), we first consider a general constrained minimization subject to the variance constraint as follows:

$$\min_{\boldsymbol{B}} f(\boldsymbol{B}), \quad \text{subj to } \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\Lambda}, \tag{11}$$

where $f(\boldsymbol{B})$ is a cost function and $\boldsymbol{\Lambda}$ is a diagonal matrix.

For (11), we work with its equivalent form by introducing a dual matrix $\boldsymbol{C}$ to decouple $\boldsymbol{B}$ and the constraint: $\min_{(\boldsymbol{B},\boldsymbol{C})} f(\boldsymbol{B})$, subject to $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Lambda}$ and $\boldsymbol{C} - \boldsymbol{B} = \boldsymbol{0}$. Then we apply the alternating direction method of multipliers (ADMM) [2] to obtain its augmented Lagrangian: $L_\rho(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{y}) = f(\boldsymbol{B}) + \boldsymbol{y}^\top \text{vec}(\boldsymbol{C} - \boldsymbol{B}) + \frac{\rho}{2}||\boldsymbol{C} - \boldsymbol{B}||_F^2$, where $\boldsymbol{y} \in \mathbb{R}^{pW}$ is the Lagrangian multiplier for constraint $\boldsymbol{C} - \boldsymbol{B} = \boldsymbol{0}$, and $\rho > 0$ is the augmented Lagrangian parameter. Then it is further simplified by introducing a dual variable matrix $\boldsymbol{V} = \{V_{jl}\}_{p \times W}$ to incorporate $\boldsymbol{y}^\top \text{vec}(\boldsymbol{C} - \boldsymbol{B})$ into the quadratic form

$$\min_{(\boldsymbol{B},\boldsymbol{C},\boldsymbol{V})} L_\rho(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{V}) = f(\boldsymbol{B}) + \frac{\rho}{2}||\boldsymbol{C} - \boldsymbol{B} + \boldsymbol{V}||_F^2, \quad \text{subj to } \boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Lambda}. \tag{12}$$

Now we apply ADMM to iterate through three steps to solve (12) until convergence. In the kth iteration,

$$\boldsymbol{C}^{(k+1)} = \text{argmin}_{\boldsymbol{C}} \frac{\rho}{2}||\boldsymbol{C} - \boldsymbol{B}^{(k)} + \boldsymbol{V}^{(k)}||_F^2, \quad \text{subj to } \boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Lambda}, \tag{13}$$

$$\boldsymbol{B}^{(k+1)} = \text{argmin}_{\boldsymbol{B}} f(\boldsymbol{B}) + \frac{\rho}{2}||\boldsymbol{C}^{(k+1)} - \boldsymbol{B} + \boldsymbol{V}^{(k)}||_F^2, \tag{14}$$

$$\boldsymbol{V}^{(k+1)} = \boldsymbol{V}^{(k)} + \boldsymbol{C}^{(k+1)} - \boldsymbol{B}^{(k+1)}. \tag{15}$$

In (13)–(15), the variance constraint $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Lambda}$ enters only in (13). Next we provide a closed-form solution of (13) in Lemma 1.

Lemma 1. *The solution of* (13) *can be written as* $\boldsymbol{C}^{(k+1)} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{P}\boldsymbol{O}^\top$, *where a singular value decomposition of* $(\boldsymbol{B}^{(k)} - \boldsymbol{V}^{(k)})^\top \boldsymbol{\Lambda}^{1/2}$ *gives* $\boldsymbol{O}\boldsymbol{E}\boldsymbol{P}^\top$, $\boldsymbol{O} \in \mathbb{R}^{W \times p}$, $\boldsymbol{E}, \boldsymbol{P} \in \mathbb{R}^{p \times p}$ *are the left, diagonal, and right matrices in the decomposition.*

3.2. Algorithm for solving (9)

After plugging $\hat{\theta}$ into (9), we begin with the update of the $(\boldsymbol{A}, \boldsymbol{B})$-block by fixing $\boldsymbol{D}$ at an initial value $\boldsymbol{D}_0$ and optimize (9) with regard to $(\boldsymbol{A}, \boldsymbol{B})$. A good

estimate of $\boldsymbol{D}_0$ can be obtained by solving (9) without the variance constraint, details are given in the Appendix.

To solve (9) with a fixed $\boldsymbol{D}$, we follow [27] to convert (9) to its equivalent dual form. The procedure contains two steps. First, we apply a DC programming method and decompose the nonconvex constraint function of the nonconvex constraints (7) and (8) into a difference of two convex functions, based on which we construct a sequence of convex approximation of nonconvex constrained sets iteratively, the details are given in the Appendix. Then at the mth iteration, we solve a relaxed subproblem (16). The iteration process continues until a termination criterion is met.

The mth subproblem amounts to

$$\begin{aligned}
&\min_{(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\lambda},\boldsymbol{\xi})} l(\boldsymbol{A},\boldsymbol{B}) \\
&+\tfrac{\mu_1}{\tau}\textstyle\sum_{1\le j\ne l\le p}|A_{jl}|w_{jl}^{(m-1)}+\tfrac{\mu_2}{\tau}\sum_{1\le j\le p,1\le l\le W}|B_{jl}|v_{jl}^{(m-1)}, \\
&\text{subj to}\lambda_{js}+\tau I(l\ne s)-\lambda_{ls}=|A_{jl}|_1 w_{jl}^{(m-1)}+\tau(1-w_{jl}^{(m-1)})+\xi_{jls}; \\
&j,l,s=1,\dots,p, j\ne l, \xi_{jls}\ge 0, \\
&\boldsymbol{B}\boldsymbol{B}^\top+\boldsymbol{D}=\hat{\theta}\boldsymbol{I},
\end{aligned} \tag{16}$$

where $\boldsymbol{\xi}=\{\xi_{jls}\}_{p\times p\times p}$, $\xi_{jls}\ge 0$ is a slack variable tensor, and $w_{jl}^{(m-1)}=I(|\hat{A}_{jl}^{(m-1)}|\le\tau)$ and $v_{jl}^{(m-1)}=I(|\hat{B}_{jl}^{(m-1)}|\le\tau)$ are obtained from the $(m-1)$th iteration.

For (16), we apply ADMM method by decoupling $(\boldsymbol{A},\boldsymbol{B})$ in the likelihood from the rest part of the cost function and the acyclic constraint. As in Section 3.1, we introduce dual variable tensor $\boldsymbol{y}=\{y_{jls}\}_{p\times p\times p}$, dual variable matrix $\boldsymbol{U}=\{U_{jl}\}_{p\times(p+W)}$ and $\boldsymbol{V}=\{V_{jl}\}_{p\times W}$. Then we minimize the augmented Lagrangian under the variance constraint:

$$\begin{aligned}
&\min_{(\boldsymbol{A},\boldsymbol{C},\boldsymbol{B},\boldsymbol{F},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{U},\boldsymbol{V})} L_\rho(\boldsymbol{A},\boldsymbol{C},\boldsymbol{B},\boldsymbol{F},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{U},\boldsymbol{V})=l(\boldsymbol{A},\boldsymbol{B}) \\
&+\tfrac{\mu_1}{\tau}\textstyle\sum_{1\le j\ne l\le p}|F_{jl}|w_{jl}^{(m-1)}+\tfrac{\mu_2}{\tau}\sum_{1\le j\le p,1\le l\le W}|B_{jl}|v_{jl}^{(m-1)} \\
&+\textstyle\sum_{1\le s\le p}\sum_{1\le j\ne l\le p}\frac{\rho}{2}\Big(|F_{jl}|w_{jl}^{(m-1)}+\tau(1-w_{jl}^{(m-1)})+\xi_{jls}-\lambda_{jl} \\
&\qquad -\tau I(l\ne s)+\lambda_{ls}+y_{jls}\Big)^2 \\
&+\tfrac{\rho}{2}\textstyle\sum_{1\le j,l\le p}(A_{jl}-F_{jl}+U_{jl})^2 \\
&+\tfrac{\rho}{2}\textstyle\sum_{1\le j\le p,1\le l\le W}(B_{jl}-F_{j,l+p}+U_{j,l+p})^2 \\
&+\tfrac{\rho}{2}\textstyle\sum_{1\le j\le p,1\le l\le W}(C_{jl}-B_{jl}+V_{jl})^2, \\
&\text{subj to}\ \ \boldsymbol{C}\boldsymbol{C}^\top+\boldsymbol{D}=\hat{\theta}\boldsymbol{I}.
\end{aligned} \tag{17}$$

Again, we solve (17) over blocks $(\boldsymbol{A},\boldsymbol{C},\boldsymbol{B},\boldsymbol{F},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{U},\boldsymbol{V})$ iteratively until convergence, where analytic updating formulas are given in the Appendix. After ADMM iterations converge, we continue the DC iterations until converge, then the current $(\boldsymbol{A},\boldsymbol{B})$-block is updated.

When updating the $\boldsymbol{D}$-block, we conduct a gradient descent update for each of $(\sigma_1^2, \cdots, \sigma_p^2)$ based on the current values of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$. The gradients for the jth dimension is $l'_j = -\frac{n}{2\sigma_j^2} - \frac{1}{2\sigma_j^4}\sum_{i=1}^n (y_{ij} - \sum_{k\neq j}\hat{A}_{jk}y_{ik} - \sum_{w=1}^W \hat{B}_{jw}x_{iw})^2$. Then σ_j^2 is updated by

$$\sigma_j^2 = \sigma_j^2 - \alpha l'_j, \tag{18}$$

where $\alpha > 0$ is the step size.

The computational strategy is summarized in Algorithm 1.

Algorithm 1: Constrained maximum likelihood

Step 1. Obtain an estimate $\hat{\theta}$ of θ by (10), then plug $\hat{\theta}$ into (9).
Step 2. Fix $\boldsymbol{D}$ at an initial value $\boldsymbol{D}_0$. Set pre-specified error tolerance ϵ_0 and the maximum number of iterations M_0 for blockwise coordinate descent.
Step 3. (($\boldsymbol{A}, \boldsymbol{B}$)-block) Initialize $\boldsymbol{A}$ and $\boldsymbol{B}$. Set pre-specified error tolerance ϵ_1 and the maximum number of DC iterations M_1.
  Step 3.1. At the mth DC iteration, compute $\hat{\boldsymbol{A}}^{(m)}$ and $\hat{\boldsymbol{B}}^{(m)}$ by cycling through the ADMM updating steps until convergence.
  Step 3.2. When $|l(\hat{\boldsymbol{A}}^{(m+1)}, \hat{\boldsymbol{B}}^{(m+1)}, \hat{\boldsymbol{D}}) - l(\hat{\boldsymbol{A}}^{(m)}, \hat{\boldsymbol{B}}^{(m)}, \hat{\boldsymbol{D}})| \leq \epsilon_1$ or $m = M_1$, stop the DC loop and output $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = (\hat{\boldsymbol{A}}^{(m)}, \hat{\boldsymbol{B}}^{(m)})$.
Step 4. ($\boldsymbol{D}$-block) Update $\boldsymbol{D}$ according to (18).
Step 5. Iterate through steps 3 through 4 until $\sum_{j=1}^p |l'_j| \leq \epsilon_0$ or the number of iterations equals to M_0, stop and output $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$.

The computation complexity of Algorithm 1 is of order $O(M_0M_1M_2p^2(p+W)+n(p+W)^2+p(p+W)^3)$, where M_0, M_1 and M_2 are the maximum number of iterations for $(\boldsymbol{A}, \boldsymbol{B})$-step, DC and ADMM, respectively. For each ADMM iteration, the computation complexity is $O(p^2(p+W))$. The preparation phase has $O(n(p+W)^2 + p(p+W)^3)$ complexity. In practice, the DC loop usually converges in a few iterations, which has finite termination property, c.f., Lemma 2 in [21].

4. Theory

This section develops a theory quantifying the reconstruction error of the CMLE defined in (9). In particular, we first show the estimated $\hat{\theta}$ recovers the optimal parameter estimation of the oracle estimator $\hat{\theta}_O$, which is defined as the maximum likelihood estimate in [27] provided that the true non-zero pattern of the DAG is given. Then we establish reconstruction consistency of the true DAG's structure defined by adjacency matrix $\boldsymbol{A}$ in the presence of intervention variables $\boldsymbol{X}$ in (2).

Let $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{D}$ represent model parameters under (2). Let $E = \{(i,j) : A_{ij} \neq 0\}$ be the set of non-zero edges in the graph $\mathcal{G}$, and $|E|$ denote the size of the set. Let $\boldsymbol{\Omega} = (\boldsymbol{I}-\boldsymbol{A})^\top(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{D})^{-1}(\boldsymbol{I}-\boldsymbol{A}) = (\boldsymbol{I}-\boldsymbol{A})^\top(\boldsymbol{I}-\boldsymbol{A})/\theta$ be the precision matrix of $\boldsymbol{Y}$. In what follows, we will assume the variance constraint (3), and by Theorem 1, $\boldsymbol{A}$ is identifiable from the distribution of $(\boldsymbol{Y}, \boldsymbol{X})$, and

thus the graph structure of $\mathcal{G}$ is identifiable. Let $G^0, \boldsymbol{A}^0, \boldsymbol{B}^0, \boldsymbol{R}^0, E^0, \boldsymbol{\Omega}^0, \theta^0$ denote the truth. Let $\hat{\boldsymbol{A}}_O$ and $\hat{\theta}_O$ denote the oracle estimator, or the maximum likelihood estimate provided that the true set E^0 of non-zero edges is given. Let $\hat{\boldsymbol{\Omega}}_O = (\boldsymbol{I} - \hat{\boldsymbol{A}}_O)^\top (\boldsymbol{I} - \hat{\boldsymbol{A}}_O)/\hat{\theta}_O$ be the oracle estimator for $\boldsymbol{\Omega}$.

For the observational model, let $\boldsymbol{A}^{obs}$ be the model parameter, then the precision matrix is $\boldsymbol{\Omega}^{obs} = (\boldsymbol{I} - \boldsymbol{A}^{obs})^\top (\boldsymbol{I} - \boldsymbol{A}^{obs})/\theta$, whose oracle estimator is $\hat{\boldsymbol{\Omega}}_O^{obs} = (\boldsymbol{I} - \hat{\boldsymbol{A}}_O^{obs})^\top (\boldsymbol{I} - \hat{\boldsymbol{A}}_O^{obs})/\hat{\theta}_O$ where $\hat{\boldsymbol{A}}_O^{obs}$ is the oracle estimator of $\boldsymbol{A}^{obs}$.

The degree of reconstructability is defined as

$$C_{\min}(\boldsymbol{\Omega}^0) = \inf_{\{\boldsymbol{A} \neq \boldsymbol{A}^0, |E| \leq |E^0|, \boldsymbol{A} \text{ induces a DAG}\}} \frac{-\log(1 - h^2(\boldsymbol{\Omega}, \boldsymbol{\Omega}^0))}{\max(|E^0 \setminus E|, 1)},$$

where $h^2(\boldsymbol{\Omega}, \boldsymbol{\Omega}^0)$ is the Hellinger distance between $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^0$ under (2) and the variance constraint (3), and $E_1 \setminus E_2$ denote the set difference between E_1 and E_2. The degree of reconstructability measures the difficulty of reconstructing the graph, and we require it to be larger than a certain level in order for our proposed method to be consistent in reconstructing the graph structure. For a detailed discussion about the degree of reconstructability of a graph, c.f., [27].

Assumption A.1 (Boundedness)**.** For some positive constants M_1 and M_2, $\inf_{\boldsymbol{\Omega}} c_{\min}(\boldsymbol{\Omega}) \geq M_1$, $\sup_{1 \leq k \leq p} |\Omega_{kk}| \leq M_2$, where $c_{\min}(\boldsymbol{\Omega})$ is the smallest eigenvalue of $\boldsymbol{\Omega}$ and Ω_{kk} is the kth diagonal element of $\boldsymbol{\Omega}$.

Assumption A.2 (Boundedness)**.** For some positive constants M_3 and M_4, $\inf_{\boldsymbol{\Gamma}} c_{\min}(\boldsymbol{\Gamma}) \geq M_3$ and $\sup_{1 \leq k \leq p+W} |\Gamma_{kk}| \leq M_4$, where $\boldsymbol{\Pi}$ is the covariance matrix of the joint distribution of $(\boldsymbol{Y}, \boldsymbol{X})$ and $\boldsymbol{\Gamma} = \boldsymbol{\Pi}^{-1}$.

Assumption B (Degree of reconstructability)**.** $C_{\min}(\boldsymbol{\Omega}^0) \geq 4 d_0^{-1} n^{-1} \times \max(\log p, |E^0|)$, for some positive constant $d_0 > 0$, say $d_0 = \frac{2}{27}\frac{1}{963}$.

Assumption C. For some positive constants d_1, d_2 and d_3,

$$h^2(\boldsymbol{\Omega}, \boldsymbol{\Omega}^0) \geq d_1 h^2(\boldsymbol{\Omega}_\tau, \boldsymbol{\Omega}^0) - d_3 p \tau^{d_2},$$

where $\boldsymbol{\Omega}_\tau = (\boldsymbol{I} - \boldsymbol{A}_\tau)^\top (\boldsymbol{I} - \boldsymbol{A}_\tau)/\theta$ and $\boldsymbol{A}_\tau$ is a truncated version of $\boldsymbol{A}$ with its ijth element $A_{ij} I(|A_{ij}| \geq \tau)$.

Assumptions A.1 and A.2 concern the smallest eigenvalues and the maximum diagonal element of Ω. Under Assumption A.1, the likelihood function becomes bounded. Assumption B is a key condition for the consistency of reconstructed graph structure, we require the degree of reconstructability to be no less than a lower bound, which is related to the size of p or $|E^0|$. Assumption C requires the Hellinger distance to be smooth so that we approximate L_0 function with its computational surrogate, the TLP function [21] to the desired level by tuning τ.

First, we show that θ is uniquely defined regardless of the value of $\boldsymbol{B}$.

Lemma 2. *Under (2), θ uniquely satisfies (3).*

Then we show the optimal parameter estimation of θ achieved by utilizing the observational data in the Theorem 2.

Theorem 2 (Optimal parameter estimation)**.** *Under Assumptions A.1 and C, if* $K_1 = |E^0|$ *and* $\tau \leq C_{\min}(\mathbf{\Omega}^0)M_1/4p$*, then there exists a constant* $c_2 > 0$*, say* $c_2 = \frac{4}{27}\frac{1}{1926}$*, such that for any* $(n, |E^0|, p)$*,*

$$P(\hat{\theta} \neq \hat{\theta}_O) \leq P(\hat{\mathbf{\Omega}}^{obs} \neq \hat{\mathbf{\Omega}}_O^{obs}) \leq \exp(-c_2 n C_{\min}(\mathbf{\Omega}^0) + 2\log(p(p-1)+1) + 3).$$

Under Assumption B, $P(\hat{\theta} \neq \hat{\theta}_O) \to 0$ *as* $n, p, |E^0| \to \infty$*.*

The next theorem gives a reconstruction error bound, under which we obtain reconstruction consistency of the CMLE as well as its optimal parameter estimation.

Theorem 3 (Error bound and oracle properties)**.** *Under Assumptions A.1, A.2 and C, if* $K_1 = |E^0|$*,* $\tau \leq C_{\min}(\mathbf{\Omega}^0)M_1/4p$*, then there exists a constant* $c_2 > 0$*, say* $c_2 = \frac{4}{27}\frac{1}{1926}$*, such that for any* $(n, |E^0|, p)$*,*

$$P(\hat{G} \neq G^0) \leq P(\hat{\mathbf{\Omega}} \neq \hat{\mathbf{\Omega}}_O) \leq \exp(-c_2 n C_{\min}(\mathbf{\Omega}^0) + 2\log(p(p-1)+1) + 3).$$

Under Assumption B, $P(\hat{G} \neq G^0) \to 0$*,* $\frac{Eh^2(\hat{\mathbf{\Omega}}, \mathbf{\Omega}^0)}{Eh^2(\hat{\mathbf{\Omega}}_O, \mathbf{\Omega}^0)} \to 1$ *as* $n, p, |E^0| \to \infty$*.*

5. Numerical study

5.1. Simulations

This section examines the performance of the proposed method and demonstrates how intervention, as well as the variance constraint, improves the reconstructability of a DAG's structure. Seven methods are compared, including the proposed method with intervention, that without the variance constraint (3), the observational method without intervention [27], the constraint-based PC algorithm [22], the score-and-search method GES [3], and two hybrid methods, Max-Min Hill-Climbing (MMHC) [24] and ARGES [14]. For PC algorithm and GES, we use R package `pcalg`, while for MMHC we use R package `bnlearn`. For ARGES, we use the ARGES-CIG version [14] by first conducting a neighborhood selection using R package `huge` and then apply the greedy search using `pcalg`. For the other three methods, we implement in R with the main algorithm written in C, which is also available in the R package `intdag` [17] https://cran.r-project.org/web/packages/intdag/index.html.

Several performance metrics are used to measure the accuracy of reconstruction of a graph's skeleton as well as directionality. With respect to the skeleton of a graph, we use the false discovery rate (FDR) and false negative rate (FNR), defined as FDR = FP/(TP + FP) and FNR = FN/(TP + FN), where TP, FP, TN, and FN denote the true positives, false positives, true negatives and false negatives, respectively. These two metrics together measure the abilities to control false discoveries, as well as false negatives. With regard to directionality, we employ the Structural Hamming Distance (SHD), defined as the minimal number of operations required to transform one DAG to the other, including edge insertions, deletions or flips, c.f., [24]. Note that a smaller SHD value indicates

two DAGs are closer to each other. To compute the SHD, we use the R-package `pcalg`. All the metrics are used on the estimation of adjacency matrix $\boldsymbol{A}$, since we focus on the reconstructability of DAG.

For tuning, PC algorithm and MMHC require one tuning parameter α controlling the significance level for independence tests, yet there is no practical tuning way via a separate tuning set. In this simulation, the significance level is fixed at 0.05. This choice of α seems sensible as the number of estimated edges is roughly the same as the number of edges in the true graph, as shown in the simulation. For ARGES, the first stage of neighborhood selection needs one tuning parameter corresponding to the LASSO penalization, we use the functions `huge.path()` in the R package `huge` to select the tuning range based on the data, and the function `huge.fit()` to select the tuning parameter. For the observational DAG method [27], τ is chosen from a set $\{0.1, 0.05, 0.01\}$, and the sparsity regularization parameter μ_1 is chosen so that the number of estimated edges roughly ranges from 0 to 100. For our methods, τ and μ_1 are selected similarly, and $\mu_2 = r \times \mu_1$ with the ratio r selected from $\{1, 2, 4, 8\}$. For each method, the optimal tuning parameters are obtained by maximizing the predicted log-likelihood (4) based on an independent tuning set of size 1000 over a set pre-specified grid points.

In simulations, we examine a sparse neighborhood graph and a sparse graph with non-sparse neighborhoods in Examples 1 and 2, respectively. A sparse neighborhood requires each node to have sparse links, but a sparse graph does not necessarily have sparse neighborhoods. To further investigate the operating characteristics of the methods, we consider two additional situations in Examples 3 and 4, in which the true graphs satisfy the variance constraint while other settings resemble Examples 1 and 2, respectively. Finally, to compare the performances of the methods in non-sparse situations, Examples 5 and 6 are added by increasing the sampling probabilities when generating the edges. More details are given in the example settings.

Example 1 (Sparse neighborhood)**.** A DAG with 50 nodes is generated with a random generation mechanism as described in [11]. First, the partial ordering of these nodes is randomly generated. Second, we sample edges according to a binomial distribution with probability 0.02, where the edges are assigned to a weight 0.5. The intervention matrix $\boldsymbol{B}$ is a diagonal matrix with diagonal values being 0.5s, describing a situation that each node in DAG is intervened by exactly one intervention covariate. Third, we set the error variances $(\sigma_1^2, \cdots \sigma_p^2)^\top$ to be a sequence from 1.5 to 1 with equally spaced points. Finally, we sample $\boldsymbol{X}$ from a p-dimensional normal distribution $N(0, \boldsymbol{I})$ and generate $\boldsymbol{Y}$ is generated according to (2).

Example 2 (Non-sparse neighborhood)**.** This example is modified from the previous example to generate a DAG of 50 nodes with a special structure of so-called "one-control-all", where all directed edges are connected from the first node to the other nodes with connection strength of 0.5. Evidently, the neighborhood of the first node is not sparse, but the overall graph remains sparse. The intervention matrix $\boldsymbol{B}$ and the error variances are the same as in Example 1.

TABLE 1

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 1. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(100,50)	Our	**0.17(0.03)**	**0.01(0.01)**	**10.60(2.32)**
(100,50)	Int	0.28(0.06)	0.02(0.02)	20.60(6.04)
(100,50)	Obs	0.39(0.20)	0.23(0.07)	39.10(22.90)
(100,50)	PC	0.29(0.06)	0.20(0.07)	21.10(5.20)
(100,50)	GES	0.62(0.05)	0.24(0.06)	64.50(10.28)
(100,50)	MMHC	0.23(0.08)	0.19(0.09)	22.60(6.08)
(100,50)	ARGES	0.50(0.00)	1.00(0.01)	50.00(0.00)
(200,50)	Our	**0.05(0.03)**	**0.00(0.00)**	**2.50(1.84)**
(200,50)	Int	0.14(0.04)	**0.00(0.01)**	8.20(2.74)
(200,50)	Obs	0.37(0.16)	0.13(0.06)	29.80(13.12)
(200,50)	PC	0.29(0.05)	0.11(0.06)	19.10(3.63)
(200,50)	GES	0.51(0.05)	0.15(0.08)	43.40(5.25)
(200,50)	MMHC	0.18(0.05)	0.09(0.05)	16.30(5.14)
(200,50)	ARGES	0.49(0.04)	0.73(0.15)	48.70(2.41)
(500,50)	Our	**0.01(0.02)**	**0.00(0.00)**	**0.50(0.97)**
(500,50)	Int	0.05(0.05)	**0.00(0.00)**	2.50(2.80)
(500,50)	Obs	0.18(0.10)	**0.00(0.01)**	11.70(7.60)
(500,50)	PC	0.28(0.05)	0.04(0.05)	19.00(3.89)
(500,50)	GES	0.45(0.06)	0.13(0.06)	35.50(8.40)
(500,50)	MMHC	0.19(0.07)	0.08(0.06)	13.90(5.53)
(500,50)	ARGES	0.25(0.06)	0.10(0.07)	18.40(4.40)

Example 3 (Sparse neighborhood with a causal model satisfying (3))**.** This example is modified from Example 1, in which the intervention matrix is diagonal, and the squared diagonals are generated from a sequence from 0.5 to 1 with equally spaced points so that the variance constraint is satisfied. The true θ is 2 in (3). Other settings remain the same as Example 1.

Example 4 (Non-sparse neighborhood with a causal model satisfying (3))**.** This example is modified from Example 2, in which the intervention matrix and $\theta = 2$ are set in the same fashion as in Example 3.

Example 5 (Non-sparse case)**.** This example is modified from Example 1, in which the sampling probability of the binomial distribution is increased to 0.1 when generating the edges, the rest are in the same fashion as in Example 1.

Example 6 (Non-sparse case with a causal model satisfying (3))**.** This example is modified from Example 3, in which the sampling probability of the binomial distribution is increased to 0.1 when generating the edges, the rest are in the same fashion as in Example 3.

As indicated in Tables 1–4, the proposed method with the variance constraint (2) performs favorably against the other methods. In Examples 1 and 3, it performs the best in all the cases in terms of all the three evaluation metrics. In

TABLE 2

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 2. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'N/A' means the method did not return a result after 24 hours. The best performers are in bold.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(100,50)	Our	**0.59(0.19)**	**0.05(0.05)**	81.40(29.54)
(100,50)	Int	0.63(0.13)	0.07(0.07)	91.90(35.95)
(100,50)	Obs	0.67(0.16)	0.21(0.09)	115.20(65.37)
(100,50)	PC	0.84(0.06)	0.90(0.04)	63.20(2.57)
(100,50)	GES	0.61(0.05)	0.17(0.06)	66.00(10.80)
(100,50)	MMHC	0.68(0.04)	0.80(0.02)	60.40(3.47)
(100,50)	ARGES	0.50(0.00)	0.96(0.04)	**49.00(0.00)**
(200,50)	Our	0.38(0.18)	0.01(0.02)	35.20(18.56)
(200,50)	Int	0.37(0.24)	**0.00(0.01)**	39.40(29.05)
(200,50)	Obs	0.29(0.25)	0.06(0.06)	36.60(51.85)
(200,50)	PC	0.78(0.07)	0.82(0.06)	63.10(5.11)
(200,50)	GES	0.44(0.05)	0.06(0.01)	32.80(7.15)
(200,50)	MMHC	0.60(0.05)	0.69(0.02)	57.30(4.22)
(200,50)	ARGES	**0.07(0.02)**	0.23(0.09)	**11.90(4.12)**
(500,50)	Our	**0.07(0.16)**	**0.00(0.01)**	**6.40(16.78)**
(500,50)	Int	0.29(0.10)	**0.00(0.00)**	20.90(8.27)
(500,50)	Obs	0.28(0.18)	**0.00(0.00)**	22.50(15.20)
(500,50)	PC	N/A	N/A	N/A
(500,50)	GES	0.38(0.13)	0.03(0.02)	25.90(20.59)
(500,50)	MMHC	N/A	N/A	N/A
(500,50)	ARGES	0.11(0.14)	0.04(0.01)	7.30(15.02)

Example 2, it achieves the top performances in terms of FDR and FNR when $n = 100$, in terms of FNR when $n = 200$ and performs the best in terms of all the three metrics when $n = 500$. In Example 4, it also achieves the top performances in terms of FDR when $n = 100$, in terms of FDR and SHD when $n = 200$ and in terms of FNR when $n = 500$. Interestingly, in Examples 2 and 4, PC algorithm and MMHC do not handle "non-sparse neighborhood" structures well as they fail to return results within 24 hours in the case of $n = 500$. Also, when n is small, ARGES performs poorly, almost identifying no directed edges, such as in the cases of $n = 100$ of Examples 1–4 and $n = 200$ of Examples 3 and 4. One explanation is that when n is small the tuning process of the neighborhood selection step tends to select a large penalty, resulting in a very sparse conditional independent graph (CIG).

Overall, both the variance constraint (3) and intervention lead to improvements of the reconstruction accuracy across all the situations. Specifically, the proposed method has an average amount of improvement $(10.00, 10.50, 8.20, 6.80)$ in terms of SHD over the intervention method without the variance constraint when $n = 100$ in all the four examples, respectively. The amount of improvement becomes $(5.70, 4.20, 8.30, 6.70)$ when n increases to 200 and $(2.00, 14.50, 2.90, 19.00)$ when $n = 500$. The improvement remains noticeable in Examples 2 and 4

TABLE 3

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 3. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'NaN' in the FDR column means the ARGES method does not have any discoveries in all the replications.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(100,50)	Our	**0.11(0.05)**	**0.04(0.04)**	**7.60(3.41)**
(100,50)	Int	0.22(0.08)	0.05(0.04)	15.80(5.63)
(100,50)	Obs	0.44(0.08)	0.14(0.05)	36.70(9.48)
(100,50)	PC	0.30(0.05)	0.21(0.07)	22.80(4.54)
(100,50)	GES	0.61(0.05)	0.21(0.09)	63.30(7.97)
(100,50)	MMHC	0.27(0.09)	0.22(0.11)	26.50(8.61)
(100,50)	ARGES	NaN	1.00(0.00)	50.00(0.00)
(200,50)	Our	**0.04(0.06)**	**0.00(0.00)**	**2.30(3.40)**
(200,50)	Int	0.16(0.11)	**0.00(0.01)**	10.60(9.07)
(200,50)	Obs	0.25(0.06)	0.03(0.04)	16.80(4.92)
(200,50)	PC	0.27(0.06)	0.12(0.05)	18.00(4.14)
(200,50)	GES	0.53(0.04)	0.17(0.05)	44.70(6.43)
(200,50)	MMHC	0.22(0.07)	0.08(0.06)	18.10(5.34)
(200,50)	ARGES	0.50(0.00)	1.00(0.00)	50.00(0.00)
(500,50)	Our	**0.01(0.01)**	**0.00(0.00)**	**0.40(0.52)**
(500,50)	Int	0.06(0.02)	**0.00(0.00)**	3.30(1.34)
(500,50)	Obs	0.17(0.05)	0.02(0.02)	10.00(3.40)
(500,50)	PC	0.27(0.04)	0.05(0.03)	18.10(3.31)
(500,50)	GES	0.42(0.08)	0.10(0.06)	32.00(8.51)
(500,50)	MMHC	0.14(0.04)	0.03(0.03)	8.80(3.01)
(500,50)	ARGES	0.43(0.06)	0.46(0.16)	41.80(6.07)

even when n is large. Interestingly, a DAG's structure can be well-reconstructed even without the variance constraint, particularly when n is large, for example, $n = 500$ in Examples 1 and 3. Moreover, the proposed method has an average improvement of $(28.50, 33.80, 29.10, 40.20)$ in SHD over the observational method when $n = 100$, $(27.30, 1.40, 14.50, 60.60)$ when $n = 200$ and $(11.20, 16.10, 9.60, 21.70)$ when $n = 500$. The amount of improvement is significant across all the cases except the case of $n = 200$ in Example 2.

In Examples 2 and 4, with a non-sparse neighborhood structure, the estimation becomes more challenging. In such situations, we added the simulations with $n = 1000$, as shown in Tables 5–6, the proposed method with the variance constraint (2) outperforms all the competitors, with SHD very close to 0 in Example 2 and exactly 0 in Example 4. As a comparison, PC and MMHC fail to return results within 24 hours, GES and ARGES perform even worse than the cases with $n = 500$. For large sample sizes, GES and ARGES tend to have more false positives, with estimated graph structure failing to determine the directions of the edges between the first node and the rest.

As demonstrated in Tables 7–8, the benefits of the proposed methods are evident against the competitors in the non-sparse cases of Examples 5 and 6,

TABLE 4

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 4. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'N/A' means the method did not return a result after 24 hours. The best performers are in bold. 'NaN' in the FDR column means the ARGES method does not have any discoveries in all the replications.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(100,50)	Our	**0.65(0.09)**	0.16(0.09)	93.40(41.51)
(100,50)	Int	0.66(0.12)	**0.15(0.05)**	100.20(49.61)
(100,50)	Obs	0.76(0.02)	0.22(0.08)	133.60(20.30)
(100,50)	PC	0.87(0.05)	0.91(0.04)	65.70(3.86)
(100,50)	GES	0.63(0.04)	0.24(0.06)	68.70(9.83)
(100,50)	MMHC	0.68(0.03)	0.79(0.02)	60.20(2.70)
(100,50)	ARGES	NaN	1.00(0.00)	**49.00(0.00)**
(200,50)	Our	**0.27(0.15)**	0.20(0.03)	**21.30(14.74)**
(200,50)	Int	0.33(0.13)	**0.02(0.04)**	28.00(15.24)
(200,50)	Obs	0.61(0.09)	0.05(0.05)	81.90(29.97)
(200,50)	PC	0.82(0.09)	0.86(0.06)	63.60(5.72)
(200,50)	GES	0.46(0.04)	0.06(0.01)	34.60(7.18)
(200,50)	MMHC	0.61(0.04)	0.70(0.02)	57.30(3.71)
(200,50)	ARGES	0.50(0.00)	0.97(0.04)	49.00(0.00)
(500,50)	Our	0.14(0.05)	**0.00(0.00)**	8.10(2.88)
(500,50)	Int	0.29(0.24)	**0.00(0.01)**	27.10(23.13)
(500,50)	Obs	0.38(0.05)	**0.00(0.00)**	29.80(5.85)
(500,50)	PC	N/A	N/A	N/A
(500,50)	GES	0.35(0.06)	0.04(0.00)	17.90(4.95)
(500,50)	MMHC	N/A	N/A	N/A
(500,50)	ARGES	**0.05(0.02)**	0.05(0.02)	**2.80(0.92)**

TABLE 5

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 2. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'N/A' means the method did not return a result after 24 hours. The best performers are in bold.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(1000,50)	Our	**0.02(0.02)**	**0.00(0.00)**	**1.20(1.23)**
(1000,50)	Int	0.09(0.06)	**0.00(0.00)**	5.00(3.40)
(1000,50)	Obs	0.17(0.12)	**0.00(0.00)**	11.10(8.37)
(1000,50)	PC	N/A	N/A	N/A
(1000,50)	GES	0.58(0.01)	**0.00(0.00)**	58.80(2.20)
(1000,50)	MMHC	N/A	N/A	N/A
(1000,50)	ARGES	0.46(0.15)	0.00(0.01)	44.40(14.90)

TABLE 6

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 4. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'N/A' means the method did not return a result after 24 hours. The best performers are in bold.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(1000,50)	Our	**0.00(0.00)**	**0.00(0.00)**	**0.00(0.00)**
(1000,50)	Int	0.04(0.03)	0.00(0.01)	2.10(1.79)
(1000,50)	Obs	0.07(0.14)	0.00(0.01)	5.20(10.00)
(1000,50)	PC	N/A	N/A	N/A
(1000,50)	GES	0.57(0.01)	**0.00(0.00)**	58.00(2.36)
(1000,50)	MMHC	N/A	N/A	N/A
(1000,50)	ARGES	0.50(0.00)	**0.00(0.0)**	49.10(0.32)

TABLE 7

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 5. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'NaN' in the FDR column means the ARGES method does not have any discoveries in all the replications.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(200,50)	Our	0.34(0.11)	**0.02(0.21)**	**135.80(66.63)**
(200,50)	Int	0.47(0.14)	0.21(0.34)	211.20(74.60)
(200,50)	Obs	0.53(0.01)	0.04(0.02)	277.30(9.79)
(200,50)	PC	**0.27(0.07)**	0.78(0.03)	203.40(7.37)
(200,50)	GES	0.86(0.02)	0.63(0.06)	604.10(23.94)
(200,50)	MMHC	0.60(0.07)	0.90(0.02)	236.60(4.84)
(200,50)	ARGES	NaN	1.00(0.00)	250.00(0.00)
(500,50)	Our	**0.18(0.11)**	**0.01(0.02)**	**61.70(43.24)**
(500,50)	Int	0.36(0.28)	0.21(0.32)	175.20(198.45)
(500,50)	Obs	0.44(0.07)	0.11(0.21)	194.40(41.48)
(500,50)	PC	0.19(0.04)	0.71(0.02)	187.50(5.70)
(500,50)	GES	0.83(0.03)	0.64(0.05)	505.80(31.71)
(500,50)	MMHC	0.55(0.06)	0.87(0.02)	225.70(6.58)
(500,50)	ARGES	NaN	1.00(0.00)	250.00(0.00)
(1000,50)	Our	**0.02(0.01)**	0.03(0.04)	**10.60(10.51)**
(1000,50)	Int	0.09(0.06)	**0.03(0.02)**	29.20(19.47)
(1000,50)	Obs	0.10(0.11)	0.04(0.04)	39.30(40.73)
(1000,50)	PC	0.18(0.04)	0.69(0.02)	181.30(7.24)
(1000,50)	GES	0.80(0.02)	0.71(0.03)	394.40(15.90)
(1000,50)	MMHC	0.55(0.05)	0.85(0.02)	222.20(6.92)
(1000,50)	ARGES	NaN	1.00(0.00)	250.00(0.00)

TABLE 8

Averaged false discovery rate (FDR), false negative rate (FNR), and Structural Hamming Distance (SHD), as well as their standard errors for four competing methods based on 10 simulation replications in Example 6. Here "Our", "Int", "Obs", "PC", "GES", "MMHC", "ARGES" denote the proposed method subject to the variance constraint, the proposed method without the variance constraint, the proposed method without intervention, PC method, GES method, MMHC method and ARGES method, respectively. The best performers are in bold. 'NaN' in the FDR column means the ARGES method does not have any discoveries in all the replications.

(n, p)	Method	FDR(A)	FNR(A)	SHD
(200,50)	Our	0.39(0.10)	**0.04(0.03)**	**164.20(69.71)**
(200,50)	Int	0.50(0.06)	**0.04(0.02)**	243.60(50.96)
(200,50)	Obs	0.51(0.15)	0.16(0.21)	273.10(126.29)
(200,50)	PC	**0.26(0.08)**	0.78(0.02)	202.30(6.77)
(200,50)	GES	0.86(0.03)	0.63(0.06)	588.30(27.56)
(200,50)	MMHC	0.60(0.07)	0.90(0.02)	237.70(6.13)
(200,50)	ARGES	NaN	1.00(0.00)	250.00(0.00)
(500,50)	Our	**0.14(0.04)**	**0.00(0.01)**	**41.50(15.35)**
(500,50)	Int	0.27(0.16)	0.10(0.26)	104.80(83.12)
(500,50)	Obs	0.36(0.10)	0.02(0.02)	148.80(56.29)
(500,50)	PC	0.16(0.03)	0.69(0.02)	178.80(5.25)
(500,50)	GES	0.83(0.02)	0.65(0.04)	485.20(26.17)
(500,50)	MMHC	0.59(0.07)	0.88(0.02)	231.30(5.42)
(500,50)	ARGES	NaN	1.00(0.00)	250.00(0.00)
(1000,50)	Our	**0.02(0.01)**	**0.02(0.03)**	**8.60(6.50)**
(1000,50)	Int	0.10(0.04)	0.04(0.02)	31.00(14.91)
(1000,50)	Obs	0.13(0.06)	0.04(0.03)	47.40(20.37)
(1000,50)	PC	0.18(0.05)	0.68(0.02)	180.00(7.39)
(1000,50)	GES	0.80(0.04)	0.71(0.06)	382.80(23.86)
(1000,50)	MMHC	0.57(0.07)	0.85(0.03)	223.50(7.15)
(1000,50)	ARGES	NaN	1.00(0.00)	250.00(0.00)

where each node in the graph has 10 edges instead of 2 edges on average. The proposed method with the variance constraint continues to perform the best in all cases in the three metrics, except the case of $n = 200$ in which PC performs better in FDR. When the sample size increases from 200 to 500 and then to 1000, our methods exhibit a larger amount of improvement, as the SHD reduces from 135.80 to 61.70 and to 10.60 in Example 5 while from 164.20 to 41.50 and to 8.60 in Example 6. By comparison, ARGES fails to produce a meaningful graph in all cases, GES yields very large values of FDR, FNR, and SHD. Whereas PC and MMHC do perform better than GES, their performances do not improve much at all when the sample size increases.

In conclusion, intervention clearly has a positive impact on the reconstructability of a DAG through interventional covariates. With the variance constraint, the accuracy of reconstruction can be further enhanced.

5.2. Analysis of Alzheimer's disease dataset

This section applies our proposed method to analyze the Alzheimer's disease dataset [25], where 8560 gene expressions were collected for 176 Alzheimer's dis-

ease patients and 187 healthy participants. Our primary goal is to reconstruct a causal network of Alzheimer's disease-related genes with the help of intervention, and compare the DAG structures for the patient and control groups, for identifying regulatory gene-gene interactions that differentiate these two patient groups.

Biologically, transcription factors (TFs) are proteins controlling the transcription process. By regulating genes, TFs ensure that target genes have right expressions in a cell and during a biological organism. In fact, a TF binds to its target DNA sequence and thus can be mapped to specific genes. For our purpose, we use the database [23] to extract a list of human DNA binding TFs and map them to genes in the dataset, resulting in 1031 mapped TF genes. Then we treat TF genes as intervention covariates to facilitate reconstruction of causal relations encoded by the gene networks.

To identify TF gene expressions associated with Alzheimer's disease in our dataset, we examine the KEGG database [12]. There, 168 genes in the Alzheimer's disease pathway, among which the expressions of 99 genes are mapped to the pathway. Among the mapped genes, we perform a two-sample t-test to obtain the significant ones between the patient and control groups, resulting in 43 selected genes at the significance level 0.05.

After pre-processing, we apply the proposed method to both the groups to reconstruct two DAG networks for the disease and control groups with 176 and 187 subjects, involving $p = 43$ genes as causal variables and $W = 1031$ TFs as intervention variables, where the tuning parameters of the method are estimated by a five-fold cross validation. As shown in Figure 1, there are 29 and 33 estimated directed connections in the patient and control groups, respectively, with 11 shared common directed connections. Moreover, the two networks share some common sub-structures, particularly, we can find common directed connections from `NDUFAB1` to `ATP5A1`, `ATP5G3`, and `COX7A2`. However, different sub-structures are also revealed. In the patient group, several genes, including `SDHD`, `NDUFA1`, `NDUFAB1`, `ATP5G3`, `ATP5A1` and `ATP5C1` have directed connections to `SDHB`, whereas in the control group, only two of them are connected to `SDHB`. This suggests that the differences in the two DAG networks reflect the disparity in the gene regulatory relations between an Alzheimer's disease subject and a healthy subject.

Biologically, our estimated directed connections match the known biological pathway of Alzheimer's disease in the KEGG database. For example, the estimated directed connections between `NDU`-genes, `SDH`-genes, `ATP`-genes and `COX`-genes match the pathway of the electron transport chain in mitochondria. Also, the estimated connection from `CALM3` to `PPP3CB` matches the biological pathway from calcium-modulated protein(CaM) to protein phosphatase 3 catalytic subunit alpha(PP3CA).

5.3. Analysis of cytometry data

This section applies the proposed method to analyze the flow cytometry data in [19] to reconstruct causal relations between phosphorylated proteins and phos-

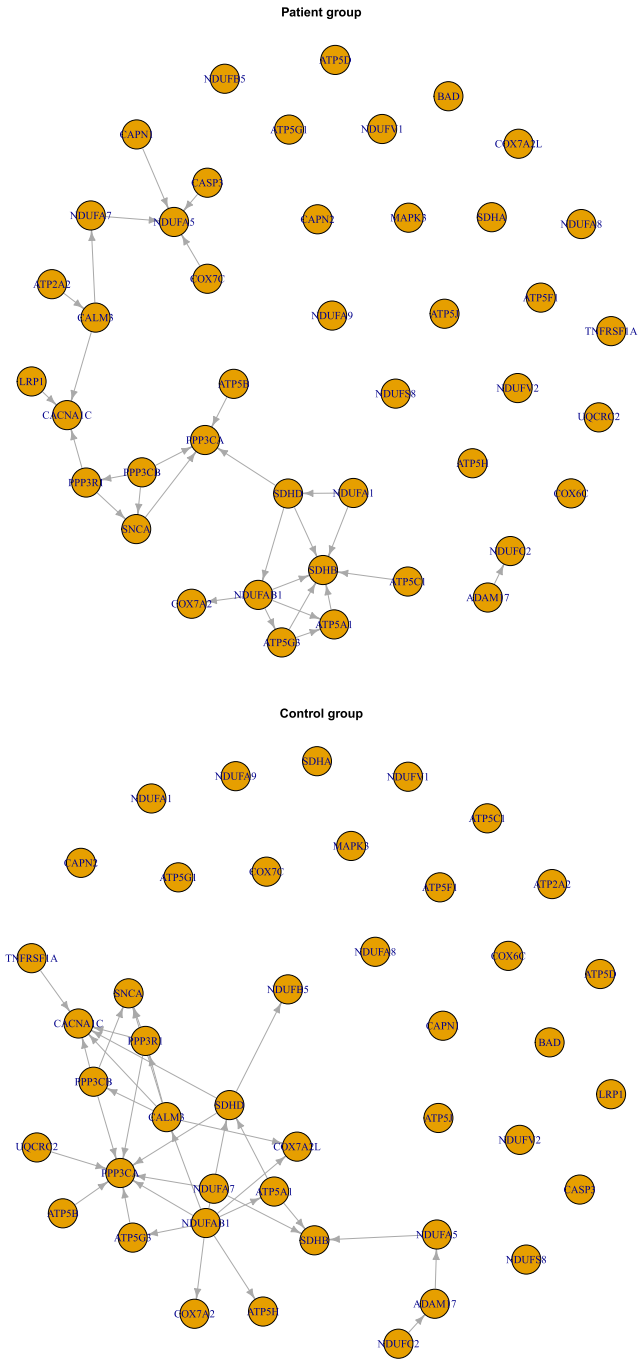

FIG 1. *Top: DAG network for Alzheimer's disease patient group. There are 38 directed connections. Bottom: DAG network for healthy participant group. There are 30 directed connections, 11 of which are shared by the DAG network of patient group.*

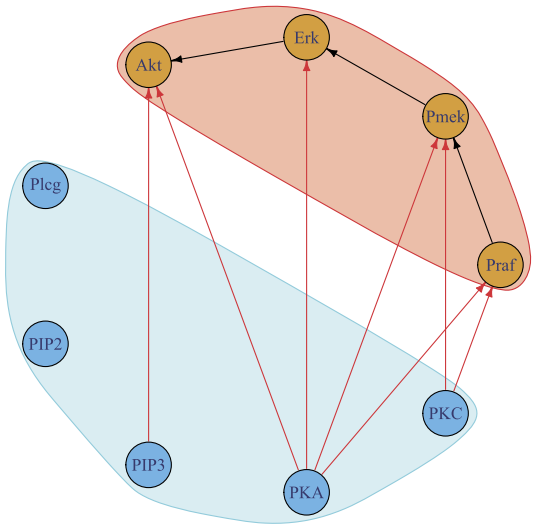

FIG 2. *Consensus intracellular signaling network as benchmark. Four nodes in brown are intervened by five intervention nodes marked in blue, with directed links indicating their intervention directions by solid red lines. Note that no links are present between intervention nodes.*

pholipids in human primary naive CD4+ T cells of the immune system. These cells were perturbed with molecular interventions to drive the ordering of connections in an intracellular signaling network. The original flow cytometry data contains $n = 7466$ cell measurements, each consisting of the amount of $p = 11$ proteins and phospholipids. The data is collected from nine different experiments in which different components in the network are intervened, either by stimulatory cues or inhibitory interventions. For our analysis, we use the continuous version of the original data [19]. Our objective is to reconstruct this network while the consensus network [19] is used as a benchmark for discovery.

For interventional data, we pre-process by selecting nine out of the eleven components, four of which form a DAG with three directed links while remaining five serve as intervention nodes; see Figure 2 for a display of an enlarge network with the nine nodes, known as consensus network, where links between the five intervention nodes are purposely removed for the intervention purpose. Note that the removal of these links does not affect our analysis, because each node is independent of its non-descendants given its direct parents by the local Markov property. Among the five intervention nodes, only three of them are informative with the other two `Plcg` and `PIP2` having no effects on the four nodes to be intervened.

We fit our proposed method with and without interventions. For tuning, we use one-tenth of the samples for training, and the rest for tuning. As displayed in Figures 3 and 4, the proposed method with intervention correctly identi-

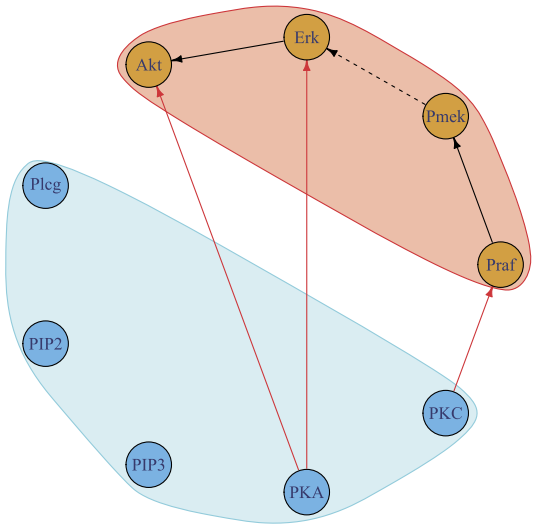

FIG 3. *Intracellular signaling network reconstructed by the proposed method with intervention. Two directed links in solid black are correctly identified, denoted by solid black lines, while the directed link from* `Pmek` *to* `Erk` *is missed, denoted by a dashed black line. Three intervention links are identified by our model (denoted by solid red lines), all of which are present in the consensus network.*

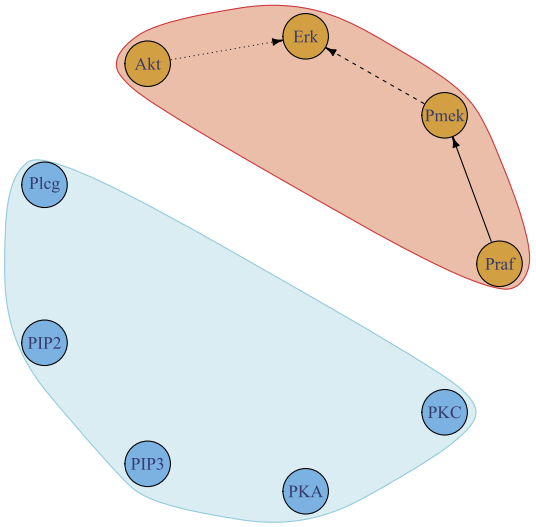

FIG 4. *Intracellular signaling network estimated by the model without intervention. One directed connection is correctly identified, denoted by the solid black line, while another connection is reversed, denoted by dotted black line, and the directed connection from* `Pmek` *to* `Erk` *is missed, denoted by dashed black line.*

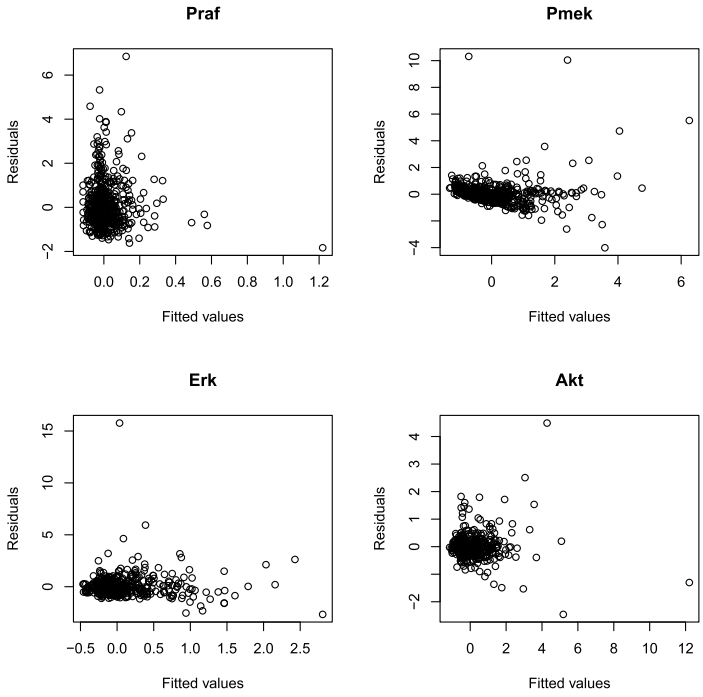

FIG 5. *Residual plots from our intervention model. Each subplot corresponds to one node in the DAG network.*

fies the directed link from `Erk` to `Akt`, while the observational method alters the direction oppositely. Both the methods enable to reconstruct the directed link from `Praf` to `Pmek`, whereas they both miss the link from `Pmek` to `Erk`. Consequently, the proposed method with intervention identifies more correct directed links than that without intervention. With respect to the intervention effects, our method identifies three intervention links in the network, which rule out the true non-informative intervention nodes `Plcg` and `PIP2`. `PIP3` is also non-informative in our estimation, and a possible reason is that `Akt` is already intervened by `PKA`, and thus the intervention effect of `PIP3` on `Akt` is unnecessary. Similarly, the intervention from `PKC` to `Praf` is identified by the proposed method, while both`PKA` and `PKC` have interventions on `Praf` in the consensus network. Finally, the proposed method yields a sparse intervention pattern.

One plausible explanation of missing the link from `Pmek` to `Erk` is that the linear causal model fails to capture nonlinear functional relations among cytometry measurements of proteins `Praf`, `Pmek`, `Erk`, and `Akt`, as evident from nonlinear patterns revealed by their residual plots of the structural equation model in Figure 5. Another possibility is that the five nodes on the bottom-left in Figure 3 are not originally designed as intervention nodes in the experiments.

6. Discussion

In this paper, a constrained maximum likelihood method is proposed to reconstruct the structure of a DAG using interventional data and efficient ADMM algorithms are developed to solve the optimization problems. In particular, a novel variance constraint is introduced and leverages the information in the interventional data to improve identifiability. Theories are established and it is shown that with the introduction of our variance constraint, the graphical structure is shown to be fully identifiable under some mild assumptions. The theoretical results are also demonstrated in our simulation results, as our proposed method performs well against competitors and can reconstruct the DAG more accurately than the observational method.

Appendix A: Technical Details

A.1. Computation details for solving (9)

For (9), after plugging in $\hat{\theta}$ and fix $\boldsymbol{D}$, we proceed in two steps. First, we relax nonconvex constraints (7) and (8) using a sequence of convex approximations involving $2^p + 1$ linear constraints, where each approximation is refined iteratively. Then we solve each subproblem by employing a constrained alternating direction method of multipliers to estimate. The underlying process iterates until convergence.

For convex relaxation of nonconvex constraints (7) and (8) in (9), we employ difference convex (DC) programming similar to [27]. In particular, we decompose J_τ into a difference of two convex functions: $\mathrm{J}_\tau(z) = S_1(z) - S_2(z) \equiv \min(\frac{|z|}{\tau}, 1) = \frac{|z|}{\tau} - \max(\frac{|z|}{\tau} - 1, 0)$. On this ground, we construct a sequence of convex approximating sets iteratively by replacing S_2 in the decomposition at iteration m by its affine majorization at iteration $m-1$. (9) becomes

$$\begin{aligned}
&\min\nolimits_{(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\lambda})} l(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\lambda}) \\
\text{subj to}\quad &\tfrac{1}{\tau}\textstyle\sum_{1\le j\ne l\le p}|A_{jl}|w_{jl}^{(m-1)} \le K_1 - \sum_{1\le j<l\le p}(1-w_{jl}^{(m-1)}), \\
&\tfrac{1}{\tau}\textstyle\sum_{1\le j\le p,1\le l\le W}|B_{jl}|v_{jl}^{(m-1)} \le K_2 - \sum_{1\le j\le p,1\le l\le W}(1-v_{jl}^{(k,m-1)}), \\
&\lambda_{js} + \tau I(l\ne s) - \lambda_{ls} \ge |A_{jl}|_1 w_{jl}^{(m-1)} + \tau(1-w_{jl}^{(m-1)}); \\
&j,l,s = 1,\dots,p, j\ne l, \\
&\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{D} = \hat{\theta}\boldsymbol{I},
\end{aligned} \tag{19}$$

where $w_{jl}^{(m-1)} = I(\|\hat{A}_{jl}^{(m-1)}\|_1 \le \tau)$, $v_{jl}^{(m-1)} = I(|\hat{B}_{jl}^{(m-1)}| \le \tau)$; $1 \le i,j \le p$, and $(\hat{\boldsymbol{A}}^{(m-1)}, \hat{\boldsymbol{B}}^{(m-1)})$ is the solution at iteration $m-1$.

Now consider a regularization version of (19), with a slack variable $\boldsymbol{\xi}$ added to the inequality constraint, yielding (16).

A.2. Analytic updating expressions for ADMM in (20)

At ADMM iteration step $s+1$, the updating formula are

$$\begin{aligned}
\boldsymbol{A}^{(s+1)} &= \operatorname{argmin}_{\boldsymbol{A}} L_\rho(\boldsymbol{A}, \boldsymbol{C}^{(s)}, \boldsymbol{B}^{(s)}, \boldsymbol{\lambda}^{(s)}, \boldsymbol{\xi}^{(s)}, \boldsymbol{y}^{(s)}, \boldsymbol{U}^{(s)}, \boldsymbol{V}^{(s)}),\\
\boldsymbol{C}^{(s+1)} &= \operatorname{argmin}_{\boldsymbol{C}} L_\rho(\boldsymbol{A}^{(s+1)}, \boldsymbol{C}, \boldsymbol{B}^{(s)}, \boldsymbol{\lambda}^{(s)}, \boldsymbol{\xi}^{(s)}, \boldsymbol{y}^{(s)}, \boldsymbol{U}^{(s)}, \boldsymbol{V}^{(s)})\\
&\quad \text{subj to } \ \boldsymbol{C}\boldsymbol{C}^\top + \hat{\boldsymbol{D}} = \sigma^2 \boldsymbol{I},\\
\boldsymbol{B}^{(s+1)} &= \operatorname{argmin}_{\boldsymbol{B}} L_\rho(\boldsymbol{A}^{(s+1)}, \boldsymbol{C}^{(s+1)}, \boldsymbol{B}, \boldsymbol{\lambda}^{(s)}, \boldsymbol{\xi}^{(s)}, \boldsymbol{y}^{(s)}, \boldsymbol{U}^{(s)}, \boldsymbol{V}^{(s)}),\\
\boldsymbol{\lambda}^{(s+1)} &= \operatorname{argmin}_{\boldsymbol{\lambda}} L_\rho(\boldsymbol{A}^{(s+1)}, \boldsymbol{C}^{(s+1)}, \boldsymbol{B}^{(s+1)}, \boldsymbol{\lambda}, \boldsymbol{\xi}^{(s)}, \boldsymbol{y}^{(s)}, \boldsymbol{U}^{(s)}, \boldsymbol{V}^{(s)}),\\
\boldsymbol{\xi}^{(s+1)} &= \operatorname{argmin}_{\boldsymbol{\xi}} L_\rho(\boldsymbol{A}^{(s+1)}, \boldsymbol{C}^{(s+1)}, \boldsymbol{B}^{(s+1)}, \boldsymbol{\lambda}^{(s+1)}, \boldsymbol{\xi}, \boldsymbol{y}^{(s)}, \boldsymbol{U}^{(s)}, \boldsymbol{V}^{(s)}),\\
&\quad \text{subj to } \ \xi_{ijk} \geq 0; i, j, k = 1, \ldots, p, j \neq k,\\
y_{ijk}^{(s+1)} &= y_{ijk}^{(s)} + \big(|F_{ij}^{(s+1)}| + \xi_{ijk}^{(s+1)} - \tau\lambda_{ik}^{(s+1)} - \tau I(j \neq k) + \tau\lambda_{jk}^{(s+1)}\big),\\
U_{jl}^{(s+1)} &= U^{(s)} + \big(A_{jl}^{(s+1)} - F_{jl}^{(s+1)}\big),\ 1 \leq j, l \leq p,\\
U_{j,l+p}^{(s+1)} &= U^{(s)} + \big(B_{jl}^{(s+1)} - F_{j,l+p}^{(s+1)}\big),\ 1 \leq j \leq p, 1 \leq l \leq W,\\
V_{jl}^{(s+1)} &= V_{jl}^{(s)} + (C_{jl}^{(s+1)} - B_{jl}^{(s+1)}),\ 1 \leq j \leq p, 1 \leq l \leq W.
\end{aligned} \tag{20}$$

A.2.1. $\boldsymbol{A}$-step and $\boldsymbol{B}$-step

For simplicity, denote $\boldsymbol{H} = (\boldsymbol{A}, \boldsymbol{B})$ be the concatenation of adjacency matrix and intervention matrix, let $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{X})$ be the concatenated data matrices. For each row of $\boldsymbol{H}$, the optimization problem is summarized as $\min_{\boldsymbol{H}_{j,j^-}} \frac{1}{2}||\boldsymbol{z}_j - \boldsymbol{Z}_{j^-}\boldsymbol{H}_{j,j^-}^\top||^2 + \frac{\rho}{2}\|\boldsymbol{H}_{j,j^-} - \boldsymbol{F}_{j,j^-}^{(s)} + \boldsymbol{U}_{j,j^-}^{(s)}\|_F^2$, where $\boldsymbol{H}_{j,j^-}$ is the jth row of $\boldsymbol{H}$ with H_{jj} excluded, $\boldsymbol{Z}_{j^-}$ is $\boldsymbol{Z}$ with its jth column removed and $\boldsymbol{x}_j$ is the jth column of $\boldsymbol{X}$. The minimizer is the solution to $\big(\mathbf{Z}_{j^-}^\top\mathbf{Z}_{j^-} + \rho\boldsymbol{I}\big)\boldsymbol{H}_{j,j^-} = \mathbf{Z}_{j^-}^\top\boldsymbol{z}_j + \rho(\boldsymbol{F}_{j,j^-}^{(s)} - \boldsymbol{U}_{j,j^-}^{(s)})$, where the factorization of $\mathbf{Z}_{j^-}^\top\mathbf{Z}_{j^-} + \rho\boldsymbol{I}$ can be cached to speed up subsequent updates.

A.2.2. $\boldsymbol{C}$-step

By Lemma 1, the updating formula for $\boldsymbol{C}$ is $\boldsymbol{C}^{(s+1)} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{P}\boldsymbol{O}^\top$, where $\boldsymbol{O}$ and $\boldsymbol{P}$ are obtained from singular value decomposition of $(\boldsymbol{B}^{(s)} - \boldsymbol{V}^{(s)})^\top\boldsymbol{\Lambda}^{1/2} = \boldsymbol{O}\boldsymbol{E}\boldsymbol{P}^\top$.

A.2.3. $\boldsymbol{F}$-step

$\boldsymbol{F}$-step updates two parts of the matrix, one is the adjacency matrix, one is the intervention matrix.

Part I: adjacency matrix For $i = 1, \cdots, p$ and $j = 1, \cdots, p$, we solve the following problem:

$$\min_F \ \mu_1 \sum_{i,j} |F_{ij}| w_{ij}^{m-1} + \frac{\rho}{2} \sum_{ijk} (|F_{ij}| w_{ij}^{m-1} + \tau(1 - w_{ij}^{m-1}) - L_{ijk}^s) + \frac{\rho}{2} \sum_{i,j} (A_{ij}^{s+1} - F_{ij} + U_{ij}^s)^2,$$

where $L_{ijk}^s = \lambda_{ik}^s + \tau I(j \neq k) - \lambda_{jk}^s - \xi_{ijk}^s - y_{ijk}^s$.

We can solve F_{ij} elementwise:

$$F_{ij}^{s+1} = \begin{cases} S(\dfrac{A_{ij}^{s+1} + U_{ij}^s}{1+p}, \dfrac{\mu_1 - p\rho \sum_k L_{ijk}^s}{\rho(1+p)}) & \text{if} \quad w_{ij}^{m-1} = 1, \\ A_{ij}^{s+1} + U_{ij}^s & \text{if} \quad w_{ij}^{m-1} = 0, \end{cases} \text{ where}$$

$$S(b, \lambda) = \begin{cases} b - 0.5\lambda & \text{if } b > 0.5\lambda, \\ b + 0.5\lambda & \text{if } b < -0.5\lambda, \\ 0 & \text{otherwise,} \end{cases} \text{ is the soft-thresholding operator.}$$

Part II: intervention matrix For $i = 1, \cdots, p$ and $j = p+1, \cdots, p+W$, we solve the following problem:

$$\min_F \ \mu_2 \sum_{i,j} |F_{ij}| v_{ij}^{m-1} + \frac{\rho}{2} \sum_{i,j} (A_{ij}^{s+1} - F_{ij} + U_{ij}^s)^2 + \frac{\rho}{2} \sum_{i,j} (C_{i,j-p}^s - F_{ij} + Z_{i,j-p}^s)^2.$$

We can solve F_{ij} elementwise:

$$F_{ij}^{s+1} = \begin{cases} S(\frac{1}{2}(A_{ij}^{s+1} + U_{ij}^s + C_{i,j-p}^s + Z_{i,j-p}^s), \ \frac{\mu_2}{2\rho}) & \text{if} \quad v_{ij}^{m-1} = 1, \\ \frac{1}{2}(A_{ij}^{s+1} + U_{ij}^s + C_{i,j-p}^s + Z_{i,j-p}^s) & \text{if} \quad v_{ij}^{m-1} = 0. \end{cases}$$

A.2.4. $\boldsymbol{\lambda}$*-step and* ξ*-step*

$(\boldsymbol{\lambda}^{s+1}, \xi^{s+1})$ is updated by $\boldsymbol{\lambda}^{s+1} = \boldsymbol{M}_{p\times p} \boldsymbol{W}_{p\times p}^{s+1}$,
where

$$\boldsymbol{M}_{p\times p} = \frac{1}{\tau} \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & \frac{2}{p} & \frac{1}{p} & \dots & \frac{1}{p} \\ 1 & \frac{1}{p} & \frac{2}{p} & \dots & \frac{1}{p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \frac{1}{p} & \dots & \frac{1}{p} & \frac{2}{p} \end{pmatrix}$$

$$\begin{aligned} W_{1j}^{s+1} &= 1, \\ W_{ik}^{s+1} &= \tfrac{1}{2}(\tau + \textstyle\sum_j (|B_{kj}^{s+1}| w_{ij}^{(m-1)} + \tau(1 - w_{ij}^{(m-1)}) + \xi_{ijk}^{s+1} + y_{ijk}^s) \\ &\quad - \textstyle\sum_j (|B_{kj}^{s+1}| w_{ij}^{(m-1)} + \tau(1 - w_{ij}^{(m-1)}) + \xi_{jik}^{s+1} + y_{jik}^s)); \quad i \neq k, \end{aligned}$$

$$W_{kk}^{s+1} = \tfrac{1}{2}(-(p-1)\tau + \textstyle\sum_j(|B_{kj}^{s+1}|w_{ij}^{(m-1)} + \tau(1-w_{ij}^{(m-1)}) + \xi_{kjk}^{s+1} + y_{kjk}^{s}) - \sum_j(|B_{kj}^{s+1}|w_{ij}^{(m-1)} + \tau(1-w_{ij}^{(m-1)}) + \xi_{jkk}^{s+1} + y_{jkk}^{s})),$$

for $i, j, k = 1, \ldots, p$.

$$\xi_{ijk}^{s+1} := \max(0, (\tau\lambda_{ik}^{s} + \tau I(j \neq k) - \tau\lambda_{jk}^{s} - |B_{ij}^{s+1}| - y_{ijk}^{s})); i, j, k = 1, \ldots, p.$$

A.3. Computation details for estimating $\boldsymbol{D_0}$

A good estimate of $\boldsymbol{D}_0$ is obtained by solving (9) without the variance constraint (3). First, we introduce a re-parametrization to exploit the convexity. Let $\boldsymbol{R} = \boldsymbol{D}^{-1/2}$, $\boldsymbol{\Phi} = \boldsymbol{D}^{-1/2}\boldsymbol{A}$ and $\boldsymbol{\Psi} = \boldsymbol{D}^{-1/2}\boldsymbol{B}$, then (2) becomes $\boldsymbol{Y} = \boldsymbol{R}^{-1}\boldsymbol{\Phi}\boldsymbol{Y} + \boldsymbol{R}^{-1}\boldsymbol{\Psi}\boldsymbol{X} + \boldsymbol{R}^{-1}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I})$ and $\boldsymbol{\Phi} = (\phi_{jk})_{p\times p}$, $\boldsymbol{\Psi} = (\psi_{jw})_{p\times W}$ are scaled versions of adjacency matrix $\boldsymbol{A}$ and intervention matrix $\boldsymbol{B}$ and $\boldsymbol{R} = (r_1, \cdots, r_p)$. Under the new parametrization, the likelihood (4) becomes $l(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{R}) = \sum_{j=1}^{p}\left[-n\log r_j + \frac{1}{2}\sum_{i=1}^{n}\left(r_j y_{ij} - \sum_{k\neq j}\phi_{jk}y_{ik} - \sum_{w=1}^{W}\psi_{jw}x_{iw}\right)^2\right]$, an easy check will confirm that $l(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{R})$ is convex in $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{R})$. Then (9) without variance constraint (3) is written as

$$\begin{aligned} &\min_{(\boldsymbol{\Phi},\boldsymbol{\Psi},\boldsymbol{R})}\ l(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{R}) = \\ &\textstyle\sum_{j=1}^{p}\left[-n\log r_j + \frac{1}{2}\sum_{i=1}^{n}\left(r_j y_{ij} - \sum_{k\neq j}\phi_{jk}y_{ik} - \sum_{w=1}^{W}\psi_{jw}x_{iw}\right)^2\right], \\ &\text{subj to}\quad \textstyle\sum_{1\leq j<l\leq p} \mathrm{J}_\tau(\phi_{jl}) \leq K_1, \quad \sum_{1\leq j\leq p, 1\leq l\leq W} \mathrm{J}_\tau(\psi_{jl}) \leq K_2, \\ &\textstyle\sum_{j_1=j_{L+1}:1\leq k\leq L} \mathrm{J}_\tau(\phi_{j_k j_{k+1}}) \leq L-1;\ \text{any } (j_1, \ldots, J_L), L = 2, \ldots, p. \end{aligned} \tag{21}$$

To solve (21), following the computation strategy illustrated in Appendix 7.1 and 7.2, after transformation, in the mth DC step, we solve

$$\begin{aligned} &\min_{(\boldsymbol{\Phi},\boldsymbol{\Psi},\boldsymbol{F},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{U},\boldsymbol{R})} L_\rho(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{F}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{U}, \boldsymbol{R}) = l(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{R}) \\ &+\tfrac{\mu_1}{\tau}\textstyle\sum_{1\leq j\neq l\leq p}|F_{jl}|w_{jl}^{(m-1)} + \tfrac{\mu_2}{\tau}\sum_{1\leq j\leq p, 1\leq l\leq W}|\psi_{jl}|v_{jl}^{(m-1)} \\ &+\textstyle\sum_{1\leq s\leq p}\sum_{1\leq j\neq l\leq p}\frac{\rho}{2}\Big(|F_{jl}|w_{jl}^{(m-1)} + \tau(1-w_{jl}^{(m-1)}) + \xi_{jls} - \lambda_{jl} \\ &\qquad -\tau I(l\neq s) + \lambda_{ls} + y_{jls}\Big)^2 \\ &+\tfrac{\rho}{2}\textstyle\sum_{1\leq j,l\leq p}(\phi_{jl} - F_{jl} + U_{jl})^2 \\ &+\tfrac{\rho}{2}\textstyle\sum_{1\leq j\leq p, 1\leq l\leq W}(\psi_{jl} - F_{j,l+p} + U_{j,l+p})^2, \end{aligned} \tag{22}$$

which is solved over blocks $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{F}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{U}, \boldsymbol{R})$. The updating formula are similar to those in Appendix 7.2, with one exception, which is $\boldsymbol{R}$-step, which is illustrated in the following section.

After solving (21), estimates $(\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{R}})$ are obtained, then $\boldsymbol{D}_0 = \hat{\boldsymbol{R}}^{-2}$.

A.3.1. $\boldsymbol{R}$*-step*

Denote $\boldsymbol{T} = (\boldsymbol{\Phi}, \boldsymbol{\Psi})$ be the concatenation of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, let $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{X})$ be the concatenated data matrices we solve the minimization problem: $\min_{r_j} -n\log r_j +$

$\frac{1}{2}\left|\left|r_j \boldsymbol{z}_j - \boldsymbol{Z}_{j^-}\boldsymbol{T}_{jj^-}^T\right|\right|^2$. The derivative equation becomes $-\frac{n}{r_j} + \sum_{i=1}^n (r_j z_{ij} - \sum_{k\neq j} T_{jk} z_{ik}) z_{ij} = 0$, yielding a solution $r_j = \frac{b+\sqrt{b^2+4na}}{2a}$, where $a = \sum_{i=1}^n z_{ij}^2$ and $b = \sum_{i=1}^n z_{ij} \sum_{k\neq j} T_{jk} z_{ik}$.

A.4. Technical proofs

Proof of Lemma 2. Let $\boldsymbol{\epsilon}' = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\epsilon}$, under (3), (2) reduces to

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{\epsilon}', \tag{23}$$

where $\boldsymbol{\epsilon}' \sim N(0, \theta\boldsymbol{I})$. In (23), $\boldsymbol{\Omega} = (\boldsymbol{I} - \boldsymbol{A})^\top(\boldsymbol{I} - \boldsymbol{A})/\theta$. Given $\boldsymbol{A}$ and the distribution of $\boldsymbol{Y}$, the value of θ is unique. This completes the proof. □

Proof of Theorem 1. Let $\boldsymbol{\epsilon}' = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\epsilon}$. Note that $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma}_X)$. Hence $\boldsymbol{\epsilon}' \sim N(0, \theta\boldsymbol{I})$, which is independent of any specific value of $\boldsymbol{B}$. Now (2) is written as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{\epsilon}'. \tag{24}$$

By (3), $\boldsymbol{\epsilon}' \sim N(0, \theta\boldsymbol{I})$, (24) reduces to an observational model with equal error variance. By Theorem 1 of [18], $\boldsymbol{A}$ is identifiable from the distribution of $\boldsymbol{Y}$, which in turn is identifiable from the joint distribution $(\boldsymbol{Y}, \boldsymbol{X})$ that is proportional to the distribution of $\boldsymbol{Y}$. This completes the proof. □

Proof of Lemma 1. The update step for $\boldsymbol{C}$ in ADMM is

$$\boldsymbol{C}^{(k+1)} = \text{argmin}_{\boldsymbol{C}} \ ||\boldsymbol{C} - \boldsymbol{B}^{(k)} + \boldsymbol{V}^{(k)}||_F^2, \ \text{s.t. } \boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Lambda}, \tag{25}$$

(25) is equivalent to $\max_{\boldsymbol{C}} \ \text{Tr}\ [\boldsymbol{C}^\top(\boldsymbol{B}^{(k)} - \boldsymbol{V}^{(k)})]$, s.t. $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Lambda}$, we can transform it into a standard form by introducing $\boldsymbol{Q} = \boldsymbol{C}^\top\boldsymbol{\Lambda}^{-1/2}$ and $\boldsymbol{S} = (\boldsymbol{B}^{(k)} - \boldsymbol{V}^{(k+1)})^\top\boldsymbol{\Lambda}^{1/2}$. Then (25) becomes

$$\max_{\boldsymbol{Q}} \ \text{Tr}\ [\boldsymbol{Q}^\top\boldsymbol{S}], \ \text{s.t. } \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}. \tag{26}$$

To solve (26), we can do a singular value decomposition $\boldsymbol{S} = \boldsymbol{O}\boldsymbol{E}\boldsymbol{P}^\top$, where $\boldsymbol{O} \in \mathbb{R}^{w\times p}$, $\boldsymbol{E}, \boldsymbol{P} \in \mathbb{R}^{p\times p}$, $\boldsymbol{O}^\top\boldsymbol{O} = \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$. Then the cost function in (26) can be written in the form $\text{Tr}\ [\boldsymbol{Q}^\top\boldsymbol{O}\boldsymbol{E}\boldsymbol{P}^\top] = \text{Tr}\ [(\boldsymbol{Q}\boldsymbol{P})^\top\boldsymbol{O}\boldsymbol{E}] = \sum_{j=1}^p [\boldsymbol{M}^\top\boldsymbol{O}]_{jj} E_{jj}$, where $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{P} \in \mathbb{R}^{w\times p}$ and $[\boldsymbol{M}^\top\boldsymbol{O}]_{jj}$ denotes the jth diagonal element of the cross product $\boldsymbol{M}^\top\boldsymbol{O}$. Note that the constraint $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$ is equivalent to $\boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{I}$ since $\boldsymbol{P}$ is an orthogonal matrix. Since $E_{jj} \geq 0$, if we can maximize each of $[\boldsymbol{M}^\top\boldsymbol{O}]_{jj}; j = 1, \cdots, p$ at the same time, then the problem is solved.

Since $[\boldsymbol{M}^\top\boldsymbol{O}]_{jj} = \sum_{i=1}^W M_{ij}O_{ij}$, by the Cauchy-Schwartz inequality and the fact that $\boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{O}^\top\boldsymbol{O} = \boldsymbol{I}$, we have $\sum_{i=1}^W M_{ij}O_{ij} \leq \sqrt{\sum_{i=1}^W M_{ij}^2 \sum_{i=1}^W O_{ij}^2} = \sqrt{1\times 1} = 1$, where the equality holds if and only if $M_{ij} = O_{ij}$ for $i = 1, 2, \cdots, W$. If we maximize each $[\boldsymbol{M}^\top\boldsymbol{O}]_{jj}; j = 1, \cdots, p$ at the same time, we need $\boldsymbol{M} = \boldsymbol{O}$, which leads to $\boldsymbol{Q} = \boldsymbol{O}\boldsymbol{P}^\top$. Then the solution of (25) is $\boldsymbol{C}^{(k+1)} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{P}\boldsymbol{O}^\top$. This completes the proof. □

Proof of Theorem 2. When we ignore the intervention covariates $\boldsymbol{X}$ and pull them into the error terms, under the variance constraint (3), (2) reduces to (24) with equal error variances, under which $\boldsymbol{A}$ is identifiable from $\boldsymbol{\Omega}$. Note that $\boldsymbol{\Omega}^{obs} = (\boldsymbol{I} - \boldsymbol{A}^{obs})^\top(\boldsymbol{I} - \boldsymbol{A}^{obs})/\theta$ is a function of $(\boldsymbol{A}^{obs}, \theta)$, it follows that $P(\hat{\theta} \neq \hat{\theta}_O) \leq P(\hat{\boldsymbol{\Omega}}^{obs} \neq \hat{\boldsymbol{\Omega}}_O^{obs})$. The rest of the proof follows from Theorem 3 of [27]. This completes the proof. □

Proof of Theorem 3. First, we define a complexity measure for the size of a space $\mathcal{F}$. The bracketing Hellinger metric entropy of $\mathcal{F}$, denoted by $H(\cdot, \mathcal{F})$, is the logarithm of the cardinality of the u-bracketing of $\mathcal{F}$ of the smallest size. That is, for a bracket covering $S(\epsilon, m) = \{f_1^l, f_1^u, \cdots, f_m^l, f_m^u\} \subset \mathcal{L}_2$ satisfying $\max_{1\leq j\leq m} \|f_j^u - f_j^l\|_2 \leq \epsilon$ and for any $f \in \mathcal{F}$, there exists a j such that $f_j^l \leq f \leq f_j^u$, *a.e.* P, then $H(u, \mathcal{F}) = \log(\min\{m : S(u, m)\})$, where $\|f\|_2 = \int f^2(z)d\mu$, with μ the dominating measure.

Denote $E^\tau = \{(i, j) : |A_{ij} \geq \tau\}$. When $K = |E^0|$, $\sum_{1\leq i\neq j\leq p} J_\tau(A_{ij}) \leq |E^0|$, so $|\hat{E}^\tau| \leq |E^0|$. If $\hat{E}^\tau = E^0$, then $\sum_{1\leq i\neq j\leq p} |A_{ij}|I(|A_{ij}| < \tau) = 0$, then $\hat{\boldsymbol{A}} = \hat{\boldsymbol{A}}_O$. Therefore, it suffices to prove the case when $\hat{E}^\tau \neq E^0$.

Define $\boldsymbol{\Omega}_{E^\tau} = (\boldsymbol{I} - \boldsymbol{A}_{E^\tau})^\top(\boldsymbol{I} - \boldsymbol{A}_{E^\tau})/\theta_{E^\tau}$ for any $E^\tau \subset \{(i, j) : 1 \leq i \neq j \leq p\}$. We can partition E^τ as $E^\tau = (E^\tau \setminus E^0) \cap (E \cup E^0)$. Let $B_{kj} = \{\boldsymbol{\Omega}_{E^\tau} : E^\tau \neq E^0, |E^\tau \cap E^0| = k, |E^\tau \setminus E^0| = j, (d_1(|E^0| - k)C_{\min}(\boldsymbol{\Omega}^0) - d_3 q\tau^{d_2}) \leq h^2(\boldsymbol{\Omega}_{E^\tau}, \boldsymbol{\Omega}^0)\}; k = 0, \ldots, |E^0| - 1, j = 1, \ldots, |E^0| - k$. Then B_{kj} has $\binom{|E^0|}{k}\binom{p(p-1)-|E^0|}{j}$ different elements E^τ's of sizes $|E^\tau \cap E^0| = k, |E^\tau \setminus E^0| = j$. By definition $\{\boldsymbol{\Omega}_{E^\tau} : E^\tau \neq E^0, |E^\tau \neq E^0| \leq |E^0|, C_{\min}(\boldsymbol{\Omega}^0) \leq h^2(\boldsymbol{\Omega}_{E^\tau}, \boldsymbol{\Omega}^0)\} \subset \cup_{k=0}^{|E^0|-1} \cup_{j=1}^{|E^0|-k} B_{kj}$. Let $L(\boldsymbol{\Omega}) = \log f(\boldsymbol{\Omega}, y, x)$ where $f(\boldsymbol{\Omega}, y, x) = f(\boldsymbol{\Omega}, y|x)f_X(x)$ is the joint density. Then

$$
\begin{aligned}
& P(\hat{G} \neq G^0) \\
\leq\ & P(\hat{\boldsymbol{\Omega}} \neq \hat{\boldsymbol{\Omega}}_O) \\
\leq\ & P^*\Big(\sup_{\boldsymbol{\Omega}_{E^\tau}: E^\tau \neq E^0, |E^\tau| \leq |E^0|} (L(\boldsymbol{\Omega}_{E^\tau}) - L(\hat{\boldsymbol{\Omega}}_O)) \geq 0\Big) \\
\leq\ & P^*\Big(\sup_{\boldsymbol{\Omega}_{E^\tau}: E^\tau \neq E^0, |E^\tau| \leq |E^0|} (L(\boldsymbol{\Omega}_{E^\tau}) - L(\boldsymbol{\Omega}^0)) \geq 0\Big) \\
\leq\ & \sum_{E^\tau \subset \{(i,j):1\leq i<j\leq p\}: E^\tau \neq E^0, |E^\tau| \leq |E^0|} P^*\Big(\sup_{\boldsymbol{\Omega}_{E^\tau} \in B_{kj}} (L(\boldsymbol{\Omega}_{E^\tau}) - L(\hat{\boldsymbol{\Omega}}_O)) \geq 0\Big) \\
\equiv\ & I,
\end{aligned}
$$

where P^* is the outer measure.

For I, we apply Theorem 1 of [26] to bound each term in the sum. We verify the entropy condition (3.1) there for the bracketing entropy over B_{kj}. Let $\boldsymbol{\Pi}$ denote the covariance matrix of the joint distribution of $(\boldsymbol{Y}, \boldsymbol{X})$ and $\boldsymbol{\Gamma} = \boldsymbol{\Pi}^{-1}$, then $\boldsymbol{\Omega}$ is the upper $p \times p$ diagonal block of $\boldsymbol{\Gamma}$. Let $\mathcal{F}_{kj} = \{f^{1/2}(\boldsymbol{\Gamma}, \cdot, \cdot) : \boldsymbol{\Omega} \in B_{kj}\}$ be the class of square-root densities. Define $\boldsymbol{\Delta} = \tilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$ and let $\lambda_1, \ldots, \lambda_{p+W}$ be the eigenvalues of $\sqrt{\boldsymbol{\Pi}}\boldsymbol{\Delta}\sqrt{\boldsymbol{\Pi}}$, $z = (y^\top, x^\top)^\top$. Then $\max_{1\leq i\leq p+W} \lambda_i \leq \lambda_{\max}(\boldsymbol{\Delta}) \times \lambda_{\max}(\boldsymbol{\Pi}) \leq c(p + W)\|\boldsymbol{\Pi}\|_{\max}$ following Prob.III.6.14 [1]. By Lemma 6.5 of [28],

it can be shown

$$\begin{aligned}|\log f(\tilde{\mathbf{\Gamma}}, y, x) - \log f(\mathbf{\Gamma}, y, x)| &\leq \max_{1\leq i\leq p+W} \lambda_i(D(z)+p+W) \\ &\leq c(p+W)\|\mathbf{\Delta}\|_{\max}(D(z)+p+W),\end{aligned}$$

where $\lambda_{\max}$ denotes the largest eigenvalue, $D(z) = \lambda_{\max}(\mathbf{\Gamma})\ \mathrm{tr}(zz^\top) \leq M_4 \times \mathrm{tr}(zz^\top)$. Note that $E\ \mathrm{tr}(zz^\top) \leq c(p+W)$ for some constant $c > 0$. By Assumption A.2, $\lambda^2_{\max}(\sqrt{\mathbf{\Pi}}) \leq 1/M_3$. Then

$$\begin{aligned}I' &\equiv \int \sup_{\tilde{\mathbf{\Gamma}}\in B_\delta(\mathbf{\Gamma})} (f^{1/2}(\tilde{\mathbf{\Gamma}}, y, x) - f^{1/2}(\mathbf{\Gamma}, y, x))^2 d\mu \\ &\leq \sup_{\tilde{\mathbf{\Gamma}}\in B_\delta(\mathbf{\Gamma})} c^{'}(p+W)||\mathbf{\Delta}||^2_{\max} E(D(z)+p+W)^2,\end{aligned}$$

for some constant $c^{'} > 0$.

Then for some positive constant Q, $I' \leq Q(p+W)^4\delta^2$. By Lemma 1 of [15], it suffices to bound the entropy of B_{ij}. There are $|E|$ nonzero entries of $\boldsymbol{A}$ with $\binom{p(p-1)}{|E|}$ possible locations. By [13], for $u \geq \epsilon^2$,

$$\begin{aligned}H(u, \mathcal{F}_{ij}) &\leq c_0(\log \binom{p(p-1)}{|E|} + |E|\log(\frac{\min(M_2^{1/2}, 1)}{u})) \\ &\leq c_0(|E|\log(e\frac{p(p-1)}{|E|}) + |E|\log(\frac{\min(M_2^{1/2}, 1)}{u})) \\ &\leq c_0(|E|\log p \log(1/u))\end{aligned}$$

Then $\epsilon = \epsilon_{n,p,|E^0|} = \min(1, (2c_0)^{1/2}c_4^{-1}\log(2^{1/2}/c_3)\log p\ (|E^0|/n)^{1/2})$ satisfies

$$\sup_{0\leq|E^\tau|\leq|E^0|} \int_{2^{-8}\epsilon^2}^{2^{1/2}\epsilon} H(t/c_3, \mathcal{F}_{ij})dt \leq (2|E^0|)^{1/2}\epsilon\log(2^{1/2}/c_3) \leq c_4 n^{1/2}\epsilon^2. \quad (27)$$

for some constants $c_3, c_4 > 0$, say $c_3 = 10, c_4 = \frac{(2/3)^{5/2}}{512}$. By Assumption B, $C_{\min}(\mathbf{\Omega}^0) \geq \epsilon_{n,p,|E^0|}$ implies (27), provided that $2d_0^{-1} > (2c_0)^{1/2}c_4^{-1}\log(2^{1/2}/c_3)$. Using the facts about binomial coefficients: $\sum_{j=1}^{|E^0|-k}\binom{p(p-1)-|E^0|}{j} \leq (p(p-1)-|E^0|+1)^{|E^0|-k}$ and $\binom{|E^0|}{k} \leq |E^0|^k$, we have, by Theorem 1 of [26], that for a constant $c_2 > 0$, say $c_2 = \frac{4}{27}\frac{1}{1926}$,

$$\begin{aligned}I &\leq \sum_{k=0}^{|E^0|-1}\sum_{j=1}^{|E^0|-k} P^*\big(\sup_{\mathbf{\Omega}_{E^\tau}\in B_{kj}} (L(\mathbf{\Omega}_{E^\tau}) - L(\hat{\mathbf{\Omega}}_O)) \geq 0\big) \\ &\leq 4\sum_{k=0}^{|E^0|-1}\binom{|E^0|}{k}\exp\big(-c_2 n(d_1 C_{\min}(\mathbf{\Omega}^0) - d_3 q\tau^{d_2})\big)\sum_{j=1}^{|E^0|-k}\binom{p(p-1)-|E^0|}{j}\end{aligned}$$

$$\leq 4\sum_{i=1}^{|E^0|}\exp\big(-i((c_2d_1/2)C_{\min}(\mathbf{\Omega}^0)-\log(p(p-1)-|E^0|+1)-\log|E^0|)\big)$$
$$\leq R\big(\exp(-(c_2d_1/2)C_{\min}(\mathbf{\Omega}^0)-\log(p(p-1)-|E^0|+1)-\log|E^0|)\big),$$

provided that $\tau \leq C_{\min}(\mathbf{\Omega}^0)M_1/4p$, where $R(x) = x/(1-x)$. Moreover, since $I \leq 1$ and $\log(p(p-1)-|E^0|+1)+\log|E^0| \leq 2\log((p(p-1)+1)/2)$,

$$\begin{aligned} I &\leq 5\exp(-(c_2d_1/2)nC_{\min}(\mathbf{\Omega}^0)+2\log((p(p-1)+1)/2)) \\ &\leq \exp(-c_2nC_{\min}+2\log(p(p-1)+1)+3). \end{aligned}$$

Under Assumption B, $P(\hat{G} \neq G^0) \to 0$ as $n, p, |E^0| \to \infty$. For parameter estimation,

$$\begin{aligned} Eh^2(\hat{\mathbf{\Omega}}, \mathbf{\Omega}^0) &\leq E[h^2(\hat{\mathbf{\Omega}}_O, \mathbf{\Omega}^0)I(\hat{\mathbf{\Omega}} = \hat{\mathbf{\Omega}}_O)] + P(\hat{\mathbf{\Omega}} \neq \hat{\mathbf{\Omega}}_O) \\ &\leq (1+o(1))Eh^2(\hat{\mathbf{\Omega}}_O, \mathbf{\Omega}^0), \end{aligned}$$

then $\frac{Eh^2(\hat{\mathbf{\Omega}},\mathbf{\Omega}^0)}{Eh^2(\hat{\mathbf{\Omega}}_O,\mathbf{\Omega}^0)} \to 1$ as $n, p, |E^0| \to \infty$.

This completes the proof. □

References

[1] Bhatia, R. (2013). *Matrix analysis* **169**. Springer Science & Business Media. MR1477662

[2] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.

[3] Chickering, D. M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research* **3** 507–554. MR1991085

[4] Eaton, D. and Murphy, K. P. (2007). Exact Bayesian structure learning from uncertain interventions. In *International Conference on Artificial Intelligence and Statistics* 107–114.

[5] Edwards, D. (2000). *Introduction to graphical modelling*. Springer Verlag. MR1880319

[6] Ellis, B. and Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association* **103** 778–789. MR2524009

[7] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science Signalling* **303** 799.

[8] Fu, F. and Zhou, Q. (2013). Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent. *Journal of the American Statistical Association* **108** 288–300. MR3174620

[9] Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research* **13** 2409–2464. MR2973606

[10] Huang, S., Li, J., Ye, J., Fleisher, A., Chen, K., Wu, T. and Reiman, E. (2013). A Sparse Structure Learning Algorithm for Gaussian Bayesian Network Identification from High-Dimensional Data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **35** 1328–1342.

[11] Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research* **8** 613–636.

[12] Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28** 27–30.

[13] Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk* **14** 3–86. MR0112032

[14] Nandy, P., Hauser, A., Maathuis, M. H. et al. (2018). High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics* **46** 3151–3183. MR3851768

[15] Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *The Annals of Probability* **15** 897–919. MR0893905

[16] Pearl, J. (2003). Statistics and causal inference: A review. *Test* **12** 281–345. MR2044313

[17] Peng, S., Shen, X. and Pan, W. (2019). intdag: Reconstruction of a Directed Acyclic Graph with Interventions R package version 1.0.1.

[18] Peters, J. and Bühlmann, P. (2013). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika, first published online*, doi: 10.1093/biomet/ast043. MR3180667

[19] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523.

[20] Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232. MR2949354

[21] Shen, X., Pan, W., Zhu, Y. and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics* 1–26. MR3105798

[22] Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, prediction, and search* **81**. The MIT Press. MR1815675

[23] Tripathi, S., Christie, K. R., Balakrishnan, R., Huntley, R., Hill, D. P., Thommesen, L., Blake, J. A., Kuiper, M. and Lægreid, A. (2013). Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database* **2013**.

[24] Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* **65** 31–78.

[25] Webster, J. A., Gibbs, J. R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M. et al. (2009). Genetic control of human brain transcript expression in Alzheimer

disease. *The American Journal of Human Genetics* **84** 445–458.

[26] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* **23** 339–362. MR1332570

[27] Yuan, Y., Shen, X., Pan, W. and Wang, Z. (2018). Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika* **106** 109–125. MR3912386

[28] Zhou, S., Shen, X., Wolfe, D. A. et al. (1998). Local asymptotics for regression splines and confidence regions. *The annals of statistics* **26** 1760–1782. MR1673277