Filter, Rank, and Prune: Learning Linear Cyclic Gaussian Graphical Models

Soheun Yi[1] **Sanghack Lee**[*,2]

[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]Graduate School of Data Science, Seoul National University
[*]correspondence to: sanghack@snu.ac.kr

Abstract

Causal structures in the real world often exhibit cycles naturally due to equilibrium, homeostasis, or feedback. However, causal discovery from observational studies regarding cyclic models has not been investigated extensively because the underlying structure of a linear cyclic structural equation model (SEM) cannot be determined solely from observational data. Inspired by the Bayesian information Criterion (BIC), we construct a score function that assesses both accuracy and sparsity of the structure to determine which linear Gaussian SEM is the best when only observational data is given. Then, we formulate a causal discovery problem as an optimization problem of the measure and propose the **F**ilter, **R**ank, and **P**rune (FRP) method for solving it. We empirically demonstrate that our method outperforms competitive cyclic causal discovery baselines.

1 INTRODUCTION

Structural equation models (SEM) (Kaplan, 2008, Kline, 2023) are widely used in various fields such as biology (Sachs et al., 2005, Smith et al., 2011), climatology (Runge et al., 2019), and operations research (Barua et al., 2016, Chowdhury et al., 2022, Shah and Goldstein, 2006) to represent complex data structures and to perform inferences on them. For these areas where decisions are made to take action, it is essential to infer the structure of the underlying graphical model from which the data is generated. One key feature of SEM is that each equation in an SEM can be viewed as a causal mechanism, and, thus, it is naturally represented as a causal graph.

Causal discovery is crucial for these real-world applications. In certain situations, including experimental settings, one can obtain interventional data and exploit them to recover the structure of a graphical model (Brouillard et al., 2020, Hyttinen et al., 2013, Maathuis et al., 2009, Rothenhäusler et al., 2015). However, in many cases, researchers are prohibited from conducting interventions due to expensive costs, ethical considerations, or the inexistence of interventional machinery. In such cases, the system's causal structure should be inferred solely from observational data. Most existing methods for causal discovery from observational data assume that the underlying model can be represented as a directed acyclic graph (DAG) (Colombo and Maathuis, 2014, Drton and Maathuis, 2017, Lachapelle et al., 2019, Spirtes et al., 2000, Zheng et al., 2018) with no unmeasured confounders (i.e., causal sufficiency). However, this is not always the case. Systems with feedback loops, e.g., brain network modeling (Smith et al., 2011), innately involve cycles in their causal structures. Therefore, it is mandatory to incorporate cycles for the learned graphical model in certain situations.

Despite the delineated importance of causal discovery for cyclic graphical models, literature on this topic is considerably scarce. This is mainly due to the fact that the underlying directed graphs (DG) of linear SEMs cannot be determined from observational data if the model is cyclic. For a linear acyclic SEM, a topological order often enables the identification of its underlying graph. For instance, a linear acyclic SEM with homoscedastic exogenous noises can be fully recovered by minimizing the mean squared error (MSE) penalizing the number of edges, or equivalently ℓ^0 regularization; for the proof for this, a topological order of the model is essential (Loh and Bühlmann, 2014, Van de Geer and Bühlmann, 2013). As another example, Park (2020) and Raskutti and Uhler (2018) leveraged different assumptions about the topological order of an acyclic SEM to identify its underlying DAG. However, this is not the case for a linear cyclic SEM for which a topological order cannot be defined. Furthermore, regarding linear Gaussian SEMs, there are "equivalent" DGs by which an identical set of distributions can be explained (Ghassami et al., 2020),

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

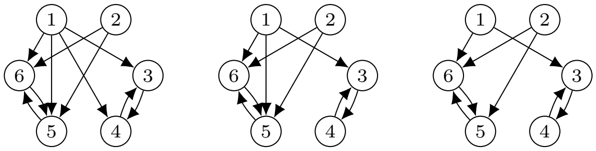

Figure 1: Directed graphs in a distribution equivalence class (Ghassami et al., 2020).

i.e., there may exist multiple DGs that can equally explain the observational data.

Therefore, we are destined to determine the "best" DG among those "equivalent" DGs. In light of "Occam's razor", we aim to find a DG with the fewest "causal connections", i.e., the smallest number of edges. It is the best among equivalent DGs in the sense that it provides the simplest explanation for observed data. Figure 1 shows equivalent DGs of 10, 9, and 8 edges on the left, center, and right, respectively; in this case, we would like to consider the DG on the right is the "best" among them. Motivated by the Bayesian information criterion (BIC) (Neath and Cavanaugh, 2012), which reflects the sparsity of a model, we present a novel mathematical formulation by offering a fresh perspective on a BIC-like score function employed within the structure learning method proposed by Ghassami et al. (2020).

We propose a method for solving this problem of finding the "best" DG given observational data. Unlike the case of acyclic DGs, designing a causal discovery method for possibly cyclic DGs based on continuous optimization is challenging since there is no explicit constraint to guide a continuous formulation of the problem. Hence, we opted for a combinatorial approach, which necessitates reducing the search space or a set of potential edges. We built such a procedure by exploiting a precision matrix since it completely characterizes the structure of the Gaussian distribution.

We summarize our contributions as follows. (1) Regarding linear cyclic Gaussian SEMs, we propose a novel measure to evaluate DGs and mathematically formulate a selection of the best DG among equivalent DGs. (2) We devise the **F**ilter, **R**ank, and **P**rune (FRP) method for solving this problem based on a solid theoretical understanding of structural coefficients and loss landscape. It efficiently and effectively eliminates spurious edges, reaching state-of-the-art performance.

1.1 Related Work

Methods for learning the structure of a linear cyclic SEM from observational data can be categorized into a few groups: constraint-based, score-based, and others using assumptions on external noises.

Constraint-based methods exploit a set of conditional independence (CI) presented in the observed data, relying on the faithfulness assumption, which states that CIs in the observed data reflect CIs (i.e., d-separation) in the underlying cyclic graph. In this approach, Richardson (1996) proposed a method that finds an equivalence class represented as a partial ancestral graph that is compatible with CIs in the observed data. On the other hand, Hyttinen et al. (2013) encoded CIs with Boolean variables, formulating a causal discovery as a Boolean satisfiability problem (SAT). There are also some methods that relax causal sufficiency allowing latent confounders. Forré and Mooij (2018) exploited σ-separation to permit unmeasured confounders.

Score-based methods design a score that reflects how well a graph can model the observational data, where optimizing it renders learning of an underlying DG. One typical approach of score-based methods is to employ the ℓ^1 penalty or LASSO (Tibshirani, 1996) to view a structure learning problem as an instance of continuous optimization: this approach has been investigated to devise structure learning methods in various settings. (Friedman et al., 2008, Meinshausen and Bühlmann, 2006) tailored LASSO to uncover the underlying undirected graph with a sparsity constraint. Zheng et al. (2018) reformulated a causal discovery of an SEM with DAG to a continuous optimization problem by configuring the acyclicity constraint as a continuous constraint. This method gave rise to several follow-up methods, including GOLEM (Ng et al., 2020), NOTEARS-TOPO (Deng et al., 2023). Sethuraman et al. (2023) designed a flow-based method to discover a possibly cyclic structure. Together with the likelihood loss, they employ ℓ^1 penalty similar to (Zheng et al., 2018) to impose sparsity on the rendered graph. Fitch (2019) proposed a method of learning a possibly cyclic DG from Gaussian observational data based on the LASSO while assuming the underlying structure to be a stationary Gaussian process, being fundamentally different from linear SEMs.

Other score-based methods incorporate improving the score in a discrete manner, which allows using a discontinuous penalty including the ℓ^0 penalty. Ghassami et al. (2020) proposed a score function that is the sum of likelihood loss and the ℓ^0 penalty to learn the causal structure of a linear Gaussian SEM up to a *quasi-equivalence* class they have defined. To elaborate in simple terms, they have defined two DGs are quasi-equivalent if the set of precision matrices that they can both generate has a non-zero Lebesgue measure. Améndola et al. (2020) utilized greedy search to discover cyclic simple mixed graphs, which permits bi-directed edges while restricting to at most one edge per pair of nodes.

A branch of methods relies on assumptions about the external noises. For acyclic SEMs, Shimizu et al. (2006) used independent component analysis (ICA) (Hyvärinen and Oja, 2000) to find an underlying DAG assuming non-Gaussian noises. Lacerda et al. (2008) takes a similar approach to recover DGs that are not necessarily acyclic from continuous observational data. Sanchez-Romero et al. (2019) exploited skewness assumption on external noise to construct a hybrid method; they find a skeleton (or undirected edges) of the underlying DG, and then direct edges based on skewness assumption or ICA. For a more extensive survey of the literature on the causal discovery, we refer the readers to (Glymour et al., 2019, Vowels et al., 2021).

2 PRELIMINARIES

We introduce notations essential to understanding our paper. Let $A_{i,:}$ and $A_{:,j}$ be the i-th row and j-th column of a matrix A, respectively. We denote by I_p an identity matrix in $\mathbb{R}^{p\times p}$. We write $\mathrm{Diag}(A)$ as a diagonal matrix whose diagonal elements correspond to the diagonal elements of $A \in \mathbb{R}^{p\times p}$.

Linear Gaussian SEM A *linear SEM* is a system with p observable variables, where one variable can be represented as a linear combination of other variables with independent noise added. When the noise is Gaussian, the system is called a *linear Gaussian SEM*. We formulate this system as follows:

$$X = XW + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Omega), \tag{1}$$

where $X \in \mathbb{R}^{1\times p}$ is a vector of observable variables, $W \in \mathbb{R}^{p\times p}$ is the weight matrix of the system, a positive diagonal matrix Ω is the covariance matrix of exogenous noise. We assume that the system has no self-loop, thus, $\mathrm{Diag}(W) = 0$. To represent multiple samples, we denote $\mathbf{X} \in \mathbb{R}^{n\times p}$ as a vertical stack of n ($\gg p$) samples from a linear SEM. Therefore, $\mathbf{X}$ satisfies

$$\mathbf{X} = \mathbf{X}W + \mathcal{E}, \quad \mathcal{E}_{i,:} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \Omega). \tag{2}$$

Underlying DG and Faithfulness Consider a linear SEM with the weight matrix W as in (1). Its underlying directed graph $\mathcal{G} = (V, E)$ is constructed from W of the SEM as follows:

$$V = \{1, \ldots, p\}, \quad E = \{(i,j) \mid W_{ij} \neq 0\}.$$

Let us denote the i-th variable of the linear SEM by X_i, the i-th element of X. A linear SEM is said to be *faithful* to DG $\mathcal{G}$ if

$$X_i \perp\!\!\!\perp X_j \mid X_{\mathbf{S}} \Rightarrow i \text{ and } j \text{ are } d\text{-separated given } \mathbf{S} \text{ in } \mathcal{G}$$

for all $i \neq j$, $\mathbf{S} \subseteq V \setminus \{i,j\}$ and $X_{\mathbf{S}} = (X_k)_{k\in\mathbf{S}}$. In the context of a linear "Gaussian" SEM, to determine conditional independence, we can measure partial correlation denoted by, e.g., $\mathrm{Corr}(X_i, X_j \mid X_{\mathbf{S}})$ between X_i and X_j given $X_{\mathbf{S}}$. To ensure the uniform convergence of the PC algorithm (Spirtes et al., 2000), Zhang and Spirtes (2002) proposed an assumption stronger than faithfulness, called λ-*strong-faithfulness*:

$$\begin{aligned}&|\mathrm{Corr}(X_i, X_j \mid X_{\mathbf{S}})| \leq \lambda \\ &\qquad \Rightarrow i \text{ and } j \text{ are } d\text{-separated given } \mathbf{S} \text{ in } \mathcal{G},\end{aligned} \tag{3}$$

where $\lambda > 0$ is a constant. However, λ-strong-faithfulness is too restrictive: λ-strong faithfulness is highly likely violated when, e.g., the weights are sampled from a uniform distribution (Uhler et al., 2013). Against this background, we introduce a weaker version of faithfulness, namely *the* λ-*edge-faithfulness*:

Definition 2.1 (λ-edge-faithfulness)**.** Consider a linear SEM with its underlying graph $\mathcal{G} = (V, E)$. We say the linear SEM is λ-*edge-faithful* to $\mathcal{G}$ if $|\mathrm{Corr}(X_i, X_j \mid X_{V\setminus\{i,j\}})| > \lambda$ holds for all $(i,j) \in E$.

This relaxes λ-strong-faithfulness (3) in two folds. First, $\mathbf{S}$ does not need to be other than $V \setminus \{i,j\}$. Second, the new assumption only considers $(i,j) \in E$, the adjacent pairs while ignoring cases where i and j are endpoints of a collider $i \to k \leftarrow j$. It turns out that rates of the two assumptions being true are dramatically different on synthetic SEMs generated following the generating mechanism employed in the experiment section (12): we present numerical evidence in Table 1.

Precision Matrix and Partial Correlation We first recapitulate the definition of the partial correlation:

Definition 2.2 (Partial correlation (Kendall, 1946))**.** For two random scalar variables X, Y and possibly multi-dimensional random variable $\mathbf{Z}$, the partial correlation of X and Y given $\mathbf{Z}$ is defined

$$\mathrm{Corr}(X, Y \mid \mathbf{Z}) = \mathrm{Corr}(R_X, R_Y),$$

where R_X, R_Y is the linear regression residuals of X, Y with respect to Z and $\mathrm{Corr}(R_X, R_Y)$ is the correlation between R_X and R_Y.

The partial correlation is closely related to the *precision matrix* or the *inverse covariance matrix* of the linear SEM (1). They are simply the inverse of its covariance matrix $\Sigma = (I_p - W)^{-\top}\Omega(I_p - W)^{-1}$, that is,

$$\Theta := \Sigma^{-1} = (I_p - W)\Omega^{-1}(I_p - W)^{\top}. \tag{4}$$

where we assume $I_p - W$ to be invertible. Denoting $\psi_{ij} := \mathrm{Corr}(X_i, X_j \mid X_{V\setminus\{i,j\}})$, we remark on a representation of ψ_{ij} in Θ (Kendall, 1946) and its natural estimator:

$$\psi_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}, \qquad \widehat{\psi}_{ij} := -\frac{\widehat{\Theta}_{ij}}{\sqrt{\widehat{\Theta}_{ii}\widehat{\Theta}_{jj}}}. \tag{5}$$

Given the precise relationship between ψ_{ij} and Θ, we prefer to specify a Gaussian distribution by its precision matrix rather than its covariance matrix, i.e., $\mathcal{N}(0, \Theta^{-1})$ rather than $\mathcal{N}(0, \Sigma)$.

Evaluation on Accuracy of a DG Let $\mathrm{KL}(P_1 \| P_2)$ denote the KL-divergence between two probability distributions P_1 and P_2. To denote the KL-divergence between two Gaussian distributions characterized by precision matrices Θ_1 and Θ_2, we write

$$\begin{aligned}&\mathrm{KL}_{\mathcal{N}}(\Theta_1 \parallel \Theta_2)\\&:= \mathrm{KL}(\mathcal{N}(0,\Theta_1^{-1}) \parallel \mathcal{N}(0,\Theta_2^{-1}))\\&= \frac{1}{2}\mathrm{tr}(\Theta_1^{-1}\Theta_2) + \frac{1}{2}\log\det\Theta_1 - \frac{1}{2}\log\det\Theta_2 - \frac{p}{2}.\end{aligned}$$

Identifying a DG with its edge set E, we can measure the accuracy of the DG by calculating how close a set of distributions that E can represent and the true distribution are. Specifically, we employ the concept of KL-divergence to define the measure as follows:

$$\begin{aligned}&\mathcal{L}(E,\Theta)\\&:= \min\{\mathrm{KL}_{\mathcal{N}}(\Theta \parallel (I_p - \widehat{W})\widehat{\Omega}^{-1}(I_p - \widehat{W})^\top) \\&\qquad\qquad \mid \mathrm{Supp}_{\widehat{W}} \subseteq E,\, \widehat{\Omega} \text{ positive diagonal}\},\end{aligned} \tag{6}$$

where Θ is the precision matrix of the true distribution and $\mathrm{Supp}_{\widehat{W}} := \{(i,j) \mid \widehat{W}_{ij} \neq 0\}$. If $\mathcal{L}(E,\Theta)$ is smaller, it implies that the graph having an edge set E can represent the underlying model more accurately. In particular, E can "represent" the true distribution if $\mathcal{L}(E,\Theta) = 0$.

Given access to Θ, calculating $\mathcal{L}(E,\Theta)$ is a mathematically challenging task as it involves solving a non-convex optimization problem:

$$\begin{aligned}\underset{\widehat{W},\widehat{\Omega}\in\mathbb{R}^{p\times p}}{\text{minimize}}\quad & \mathrm{KL}_{\mathcal{N}}(\Theta \parallel (I_p - \widehat{W})\widehat{\Omega}^{-1}(I_p - \widehat{W})^\top),\\ \text{subject to}\quad & \mathrm{Supp}_{\widehat{W}} \subseteq E,\ \widehat{\Omega} \text{ positive diagonal.}\end{aligned}$$

We can simplify it by denoting $Q = (I_p - \widehat{W})\widehat{\Omega}^{-1/2}$, so that we have $QQ^\top = (I_p - \widehat{W})\widehat{\Omega}^{-1}(I_p - \widehat{W})^\top$. This formulation renders $Q_{ii} > 0$; with a slight relaxation to $Q_{ii} \geq 0$, we can dismiss this condition because negating i'th column of Q for each i with $Q_{ii} < 0$ will not change the value of $QQ^\top$. Since $\widehat{\Omega}^{-1/2}$ is a positive diagonal matrix, the zero entries of Q and $I_p - \widehat{W}$ coincide. Therefore, we obtain the following problem:

$$\begin{aligned}\underset{Q\in\mathbb{R}^{p\times p}}{\text{minimize}}\quad & \mathrm{KL}_{\mathcal{N}}(\Theta \parallel QQ^\top),\\ \text{subject to}\quad & Q_{ij} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin E.\end{aligned} \tag{7}$$

We remark on two analytical features of this problem. First, the term $QQ^\top$ in the objective function reminds the low-rank approximation of a matrix or Burer–Monteiro factorization (Burer and Monteiro, 2003). However, its approximation target Θ is full-rank rather than low-rank, so most theories regarding the literature of Burer–Monteiro factorization is hardly applicable to our problem. Additionally, while $\mathrm{KL}_{\mathcal{N}}(\Theta \parallel \cdot)$ is convex when defined on the set of positive definite matrices, $\mathrm{KL}_{\mathcal{N}}(\Theta \parallel QQ^\top)$ is not a convex function of Q. Therefore, we take a random initialization approach based on L-BFGS (Nocedal and Wright, 2006) to solve this non-convex problem. Further elaboration can be found in Appendix F.

3 PROBLEM FORMULATION

In this section, we formulate a causal discovery problem for a linear Gaussian SEM. We consider a linear Gaussian SEM with p observable variables and the exogenous noise distribution $\mathcal{N}(0,\Omega)$ as in (2). Observational data $\mathbf{X} \in \mathbb{R}^{n\times p}$ follows a Gaussian distribution

$$\mathbf{X}_{i,:} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\Theta^{-1}), \tag{8}$$

where the precision matrix $\Theta = (I_p - W)\Omega^{-1}(I_p - W)^\top$. We aim to find a sparse structure by which $\mathcal{N}(0,\Theta^{-1})$ can be explained. There can be DGs with a fewer number of edges than the true DG that explain $\mathcal{N}(0,\Theta^{-1})$ with the same level of accuracy as depicted in Figure 1. Hence, we are not targeting to recover edges of the true underlying DG, i.e., $\{(i,j) \mid W_{ij} \neq 0\}$.

Performance Measure Recall the measure $\mathcal{L}(E,\Theta)$ defined in (6), which serves to quantify the representability of the set of edges E with respect to the Gaussian distribution $\mathcal{N}(0,\Theta^{-1})$. As we put more elements in E, $\mathcal{L}(E,\Theta)$ monotonically decreases. This is a trade-off between the sparsity and the accuracy of the model. Therefore, we should consider both in measuring the quality of the estimated DG. In this regard, we inspect the Bayesian information criterion (BIC) (Neath and Cavanaugh, 2012), defined $\mathrm{BIC} = -2\log L + k\log n$, where n, k, and L are the number of samples, the number of parameters, and the maximized value of the likelihood function of the model, respectively. Informally, the term $-2\log L$ serves as a measure of how well the model can explain the data, corresponding to $\mathcal{L}(E,\Theta)$ in our context. Moreover, the term $k\log n$ represents the sparsity of the model, drawing an analogy to the number of edges $|E|$ in our specific problem. This gives rise to the following measure, which we aim to minimize:

$$\mathcal{L}_\mu(E,\Theta) := \mathcal{L}(E,\Theta) + \mu|E|. \tag{9}$$

Aligned with our motivation for constructing this measure, $\mathcal{L}_\mu(E,\Theta)$ is equivalent to the BIC up to scaling and translation when $\mu = \frac{\log n}{2n}$ and $\Theta = (\mathbf{X}^\top\mathbf{X}/n)^{-1}$. In this aspect, we call $\mathcal{L}_\mu(E,\Theta)$ as the BIC score. Indeed, it also equals a score function introduced by Ghassami et al. (2020) (see Appendix C for details). Thus, we aim to find an edge set E that minimizes $\mathcal{L}_\mu(E,\Theta)$.

Assumptions and Theoretical Guarantees We assume that $\mathbf{X}$ is a full-rank matrix so that $\mathbf{X}^\top\mathbf{X}$ is invertible. Therefore, we can compute the MLE of Θ by $\widehat{\Theta} = (\mathbf{X}^\top\mathbf{X}/n)^{-1}$, which characterizes $\widehat{\psi}_{ij}$ defined in (5). We can prove that $\widehat{\psi}_{ij}$ is consistent in the sense of Theorem 3.2 under Assumption 3.1, which is common in the literature of inverse covariance matrix estimation (Janková and van de Geer, 2015, Liu et al., 2012, Yuan, 2010).

Assumption 3.1 (Bounded eigenvalues)**.** Consider the ground truth distribution $\mathcal{N}(0, \Theta^{-1})$ of p variables. Let $M_p > 0$ be a constant dependent on p. Then, we assume

$$1/M_p \leq \lambda_{\min}(\Theta) \leq \lambda_{\max}(\Theta) \leq M_p,$$

where $\lambda_{\min}(\Theta)$ and $\lambda_{\max}(\Theta)$ are the minimum and maximum eigenvalue of Θ, respectively.

Theorem 3.2. *Consider a Gaussian distribution of p variables as in* (2) *satisfying Assumption 3.1. Let ψ_{ij} and $\widehat{\psi}_{ij}$ be as in* (5)*. Then, for a sufficiently large n, $|\widehat{\psi}_{ij} - \psi_{ij}| \leq C_p n^{-1/4}$ holds with probability at least $1 - 2\exp(-cpn^{1/2})$ for all i, j, where $c > 0$ is an absolute constant and $C_p > 0$ depends only on p.*

Theorem 3.2 and the following Assumption 3.3, which regards the edge-faithfulness of the true distribution, lead to Theorem 3.4.

Assumption 3.3. Consider the ground truth distribution $\mathcal{N}(0, \Theta^{-1})$ of p variables, where each of them corresponds to $V = \{1, \ldots, p\}$. Let ϵ_p be a constant dependent on p. We assume there exists a set of edges E that minimizes $\mathcal{L}_\mu(E, \Theta)$ while ensuring the corresponding linear SEM being ϵ_p-edge-faithful to the graph (V, E).

Theorem 3.4. *Consider a Gaussian distribution $\mathcal{N}(0, \Theta^{-1})$ of p variables, satisfying Assumptions 3.1 and 3.3. Let $\widehat{\psi}_{ij}$ be as in* (5)*. Then, there exists a set of edges E that minimizes $\mathcal{L}_\mu(E, \Theta)$ and satisfies the following property:*

> *For a sufficiently large n, $E \subseteq \{(i,j) \,|\, |\widehat{\psi}_{ij}| > C_p n^{-1/4}\}$ holds with probability at least $1 - 2\exp(-cpn^{1/2})$, where $c > 0$ is an absolute constant and $C_p > 0$ depends only on p.*

We remark that Theorems 3.2 and 3.4 can be derived from Lemma 29 in (Loh and Bühlmann, 2014) and Remark 5.40 in (Vershynin, 2010). We defer proofs to Appendix D. In Section 4.1, we will demonstrate how we can use Theorem 3.4 to refine probable edges.

4 FILTER, RANK, AND PRUNE METHOD

We now propose a method, namely the *Filter, Rank, and Prune* (FRP) algorithm to solve the problem presented in the previous section. FRP can be summarized as follows: (1) We calculate $\widehat{\psi}_{ij}$ from observational data and invoke Theorem 3.4 for *filtering* out spurious edges. (2) Then, we execute an algorithm (Section 4.2) which repeats *ranking* the candidates and *pruning* unnecessary ones of the lowest ranks. If there are no edge candidates to prune, then the algorithm terminates.

4.1 Filter Stage via Partial Correlation

Recall the result of Theorem 3.4; with probability at least $1 - 2\exp(-cpn^{1/2})$, $E \subseteq \{(i,j) \,|\, |\widehat{\psi}_{ij}| > C_p n^{-1/4}\}$. Now, consider a procedure of removing (i,j) that satisfies $|\widehat{\psi}_{ij}| \leq C_p n^{-1/4}$ from edge candidates. If we can use a sufficiently large number of samples, this procedure will not exclude any true edges from edge candidates with arbitrarily high probability, as Theorem 3.4 indicates. Therefore, setting initial edge candidates by $\widehat{E}_0 = \{(i,j) \,|\, |\widehat{\psi}_{ij}| > C_p n^{-1/4}\}$ is a sensible choice. We note that setting the threshold $C_p n^{-1/4} = 0.1$ for $n = 1000$ showed good performance in our experiments.

It is worth comparing with a similar thresholding approach to learn the structure of a (undirected) graphical model. Loh and Bühlmann (2014) used, in Lemma 15, entries of the precision matrix, i.e., $\widehat{\Theta}_{ij}$, instead of partial correlations $\widehat{\psi}_{ij}$ to determine whether (i,j) belongs to the edge set. Since partial correlation is invariant upon the scale of exogenous noise, we do not need to estimate the variance of the noise, as opposed to using the precision matrix itself. Therefore, utilizing partial correlation has a practical advantage, not only providing a connection to relax strong-faithfulness as mentioned in Section 2.

4.2 Rank and Prune Stages

After obtaining the initial edge candidates $\widehat{E}_0$, we try to minimize $\mathcal{L}_\mu(E, \Theta)$. However, as we do not have access to Θ, we use $\widehat{\Theta}$ instead. We mathematically formulate our problem as follows:

$$\underset{\widehat{E}}{\text{minimize}} \quad \mathcal{L}_\mu(\widehat{E}, \widehat{\Theta}) \quad \text{subject to} \quad \widehat{E} \subseteq \widehat{E}_0. \tag{10}$$

We take an iterative approach to solve the problem. Each iteration consists of two stages: the *Rank* stage and the *Prune* stage. In the Rank stage, we "rank" the current edge candidates by solving a subproblem about whether we can remove an edge without causing much loss of accuracy. In the Prune stage, we "prune" unnecessary edges through a hybrid of binary and sequential searches utilizing the rank determined in the previous stage. We repeat these two stages until there are no more edge candidates to prune.

Rank Stage Let $\widehat{E}$ be a set of edges to be considered in this stage. We solve the following problem to rank the edges to prune some of them in the later stage:

$$\begin{aligned} \underset{Q \in \mathbb{R}^{p \times p}}{\text{minimize}} \quad & \mathrm{KL}_{\mathcal{N}}(\widehat{\Theta} \,\|\, QQ^\top) + \mathrm{reg}(Q), \\ \text{subject to} \quad & Q_{ij} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin \widehat{E}, \end{aligned} \tag{11}$$

where reg is a regularization term that induces sparsity in a solution. To prevent a bias created by the regularization term from being large, we adopt SCAD penalty (Fan and Li, 2001) (see Appendix F for details). Once we have obtained a solution $Q^\star$, we rank edges $(i,j) \in \widehat{E}$ by $|Q^\star_{ij}|$ in ascending order as demonstrated in RANK function in Algorithm 1. Notably, edges with low ranks would have small corresponding weights, allowing for their removal with the increase in only a small fraction of the empirical loss.

Prune Stage Let $\widehat{E}$ be the ordered set of edges sorted from the previous stage. We aim to prune edges as much as possible while keeping the empirical score of the remaining edges smaller than or equal to $T = \mathcal{L}_\mu(\widehat{E}, \widehat{\Theta}) + \epsilon_{\text{tol}}$, where ϵ_{tol} is a tolerance parameter. We first seek to prune edges with the lowest ranks. If removing the first edge is unacceptable, we seek the possibility of removing another edge without harming the accuracy seriously: the necessity of this process is supported in Section 5.2. The Prune stage is described in Algorithm 1 (0-based indexing), where $\widehat{E}[i:] = \{\widehat{E}_i, \widehat{E}_{i+1}, \dots, \widehat{E}_{|\widehat{E}|-1}\}$ and $\widehat{E}[-i]$ denotes a set of all elements of $\widehat{E}$ except the i-th element. We employ binary search for the second action taken in the PRUNE function.[1] Such technique is possible because $(\mathcal{L}(\widehat{E}[i:], \widehat{\Theta}))_{i=0}^{|\widehat{E}|-1}$ is sorted in an increasing order—removing more edges does not decrease the accuracy term of the BIC score, that is, $\mathcal{L}(E_1, \widehat{\Theta}) \leq \mathcal{L}(E_2, \widehat{\Theta})$ if $E_1 \supseteq E_2$. We choose $\epsilon_{\text{tol}} = \mu$, where μ is a predefined penalty constant in (9), as this choice renders $\mathcal{L}(\widehat{E}, \widehat{\Theta})$ to be the exact BIC score when $\widehat{\Theta} = \Theta$.

4.3 Complete Algorithm

The FRP algorithm is described in Algorithm 2. It starts with initializing edge candidates $\widehat{E}$, followed by iterations of the Rank and Prune stages until no further edges are pruned from $\widehat{E}$. We note that the Rank stage relies on solving a nonconvex problem (11), thus, the algorithm may not converge to a global optimum. Therefore, we run multiple (N_{init}) instances (it has been observed that 2 instances are enough to outperform other baselines in Section 5.1) of FRP and choose $\widehat{E}$ giving the smallest $\mathcal{L}_\mu(\widehat{E}, \widehat{\Theta})$ among all outputs. We run these instances in parallel, preventing the process from excessive additional time costs.

5 EXPERIMENTS

In this section, we provide the experimental results of FRP. We first measure the performance of FRP on a synthetic dataset with competitive baselines, which are revised from their original versions to ensure a fair comparison. Next, we present ablation studies and the roles of hyperparameters on the performance. Finally, we apply FRP to a real-world dataset. All experiments are conducted using two 24-core CPUs (Intel Xeon 6342 with a base frequency of 2.8 GHz). Our implementation of FRP is available at `https://github.com/soheunyi/frp`.

Algorithm 1 Rank and Prune Stages

```
function RANK($\widehat{E}, \widehat{\Theta}$)
    Solve (11) to obtain $Q^\star$
    Sort $(i, j) \in \widehat{E}$ by $|Q^\star_{ij}|$ ascendingly
    return $\widehat{E}$
function PRUNE($\widehat{E}, \widehat{\Theta}, \epsilon_{\text{tol}}$)
    $T \leftarrow \mathcal{L}(\widehat{E}, \widehat{\Theta}) + \epsilon_{\text{tol}}$
    $i \leftarrow \max\{i \geq 0 \,|\, \mathcal{L}(\widehat{E}[i:], \widehat{\Theta}) \leq T\}$    ▷ Binary search
    if $i \geq 1$ return $\widehat{E}[i:]$
    for $i \leftarrow 1$ to $|\widehat{E}| - 1$ do
        if $\mathcal{L}(\widehat{E}[-i], \widehat{\Theta}) \leq T$
            return $\widehat{E}[-i]$    ▷ Single-edge removal phase
    return $\widehat{E}$
```

Algorithm 2 The FRP Algorithm

```
function FRP($\mathbf{X}$, $\epsilon_{\text{tol}}$, $C_p$, $N_{\text{init}}$)
    $\widehat{\Theta} \leftarrow (\mathbf{X}^\top \mathbf{X}/n)^{-1}$
    $\widehat{\psi}_{ij} \leftarrow -\widehat{\Theta}_{ij} / \sqrt{\widehat{\Theta}_{ii}\widehat{\Theta}_{jj}}$ for all $i, j$
    for $k \leftarrow 1$ to $N_{\text{init}}$ do    ▷ Parallel execution
        $\widehat{E}_k \leftarrow \{(i,j) \,|\, |\widehat{\psi}_{ij}| > C_p n^{-1/4}\}$
        while True do
            $\widehat{E}_k \leftarrow$ RANK($\widehat{E}_k, \widehat{\Theta}$)
            $\widehat{E}'_k \leftarrow$ PRUNE($\widehat{E}_k, \widehat{\Theta}, \epsilon_{\text{tol}}$)
            if $\widehat{E}'_k = \widehat{E}_k$ then break
            $\widehat{E}_k \leftarrow \widehat{E}'_k$
    return argmin$_{\widehat{E}_k}$ $\mathcal{L}_\mu(\widehat{E}_k, \widehat{\Theta})$
```

5.1 Performance Evaluation

Baselines To demonstrate the performance of FRP, we include DGLEARN (Ghassami et al., 2020) to our baselines, which is designed to solve a similar problem by minimizing the similar score with FRP. In addition, we add the following baselines to make our comparison more comprehensive: NOTEARS (Zheng et al., 2018), GOLEM (Ng et al., 2020), and NODAGS-Flow (Sethuraman et al., 2023). This is in line with baseline choices made by Sethuraman et al. (2023), while LLC (Hyttinen et al., 2012) is excluded as it returns the zero weight matrix when only observational data is given. We used the implementation from `https://github.com/syanga/dglearn` for DGLEARN, `https://github.com/Genentech/nodags-flows` for NODAGS-Flow and GOLEM, and `https://github.com/xunzheng/notears` for NOTEARS.

Synthtic Graphs and Data We conducted experiments on $p \in \{10, 15, 20\}$. For each p, we created 10 different Erdős–Rényi graphs (Erdős and Rényi, 1960) with numbers of edges that, in expectation, give 0.75, 0.5, 0.25, and 0.125 of non-diagonal entries of the precision matrix to be zero; we provide the details in Appendix G. When an underlying DG $\mathcal{G} = (V, E)$ is specified, we generate a random weight matrix W and the covariance matrix of exogenous noise Ω

[1] Binary search has been utilized by Lengerich et al. (2021) and Sethuraman et al. (2023) to threshold a possibly cyclic weight matrix to obtain a DAG.

by the following procedure:

$$W_{ij} \sim \begin{cases} \text{Uniform}([-1,-0.6] \cup [0.6,1]) & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Omega_{ij} \sim \begin{cases} \text{Uniform}([1,2]) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Observational data of size $n = 1000$ is sampled via $X = \mathcal{E}(I_p - W)^{-1}$ where each row of $\mathcal{E}$ is sampled from $\mathcal{N}(0, \Omega)$. To ensure Assumption 3.1 is satisfied, we restrict the true precision matrix to have eigenvalues in $[10^{-3}, 10^{3}]$ via the accept-reject approach.

Evaluation We evaluated FRP and the baselines on the BIC score $\mathcal{L}_\mu(\widehat{E}, \Theta)$, where $\widehat{E}$ is the estimated edge set and Θ is the ground truth precision matrix. There are some considerations regarding the baselines to make the comparison fair. To make FRP and DGLEARN minimize the same score, we set $\mu = \frac{\log n}{2n}$ for $\mathcal{L}_\mu(E, \widehat{\Theta})$ (9). We set a total timeout of DGLEARN to 3600 seconds by terminating each stage of the algorithm as in Appendix G. For NOTEARS, we evaluated performance varying ℓ^1-regularization parameter in $\{10^{-3}, 10^{-2}, 10^{-1}\}$ and reported the best performance for each synthetic graph. As we allow cycles, we *inactivated* the acyclicity constraint of GOLEM to prevent it from being misguided by the constraint. Also, NOTEARS, GOLEM, and NODAGS-Flow estimate a weight matrix and then apply thresholding just once to obtain an edge set. Since different thresholds can lead to different performances, we should evaluate the performance of these baselines with multiple thresholds. We set thresholds to be (1) set absolutely to 0.1, 0.2, 0.3, (2) set relatively to $1/16$, $1/8$, $1/4$ of the maximum absolute value of the estimated weight matrix. We report the best performance among these thresholds to make our baselines more competitive.

Empirical Results FRP shows the best performance in terms of the BIC score in most pairs of node and edge (p, e), as depicted in Figure 2. Indeed, FRP renders the best score in all tuples except for $(p, e) = (10, 9)$ (as shown in Table 2). As seen in the first graph of Figure 2, the performance gap between FRP and the others widens as the underlying graph has more edges. A similar trend appears when the number of nodes grows with fixed sparsity of the true precision matrix. These observations imply the robustness of FRP to the increase in the number of nodes and edges.

5.2 Ablation Studies

Necessity of Each Stage We investigated the effects on the performance induced by three components consisting of FRP: the Filter stage, the Rank stage, and the single-edge removal. Specifically, we conducted experiments with (1) disabling the Filter stage and feeding all edges to the initial edge candidate, (2) using a random order of edges instead of the order provided by RANK function, and (3) removing the single-edge removal phase. The results are depicted in Figure 3 (Full results are in Appendix H). The BIC score was observed to be worse in each case, suggesting the necessity of the three ingredients for better performance.

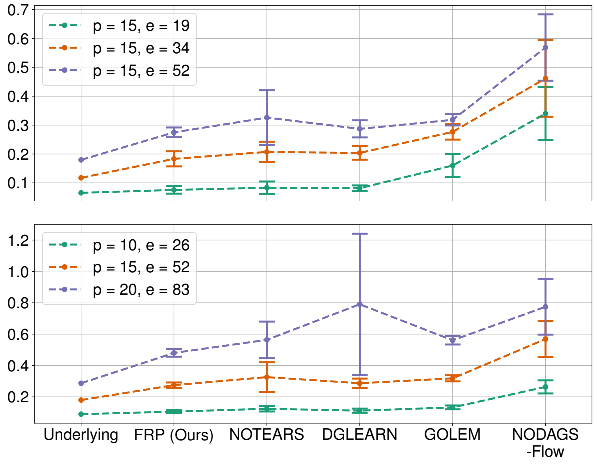

Figure 2: The BIC score between estimated graphs for FRP (ours), NOTEARS, DGLEARN, GOLEM, and NODAGS-Flow. Markers and half-width of the error bar indicate the mean and standard deviation, respectively. p and e are the numbers of nodes and edges in the underlying graph, and "Underlying" indicates μe, the BIC score of the underlying graph. See Table 2 for full results.

Hyperparameters We studied how hyperparameters ϵ_{tol} (tolerance parameter in the Prune stage), $C_p n^{-1/4}$ (the partial correlation threshold in the Filter stage), and N_{init} (the number of initializations in FRP) influence the performance. Results are presented in Figure 4, which we explain in the following.

$\boldsymbol{\epsilon_{\text{tol}}}$: In theory, $\epsilon_{\text{tol}} = \mu$ yields the best BIC score. While this choice shows decent performance in every case, there are some cases where $\epsilon_{\text{tol}} = 2\mu$ performs better. This is possible because the FRP is using $\mathcal{L}_\mu(\widehat{E}, \widehat{\Theta})$ which is an estimate but not the true BIC score.

$\boldsymbol{C_p n^{-1/4}}$: Using very small values for the threshold is similar to deactivating the Filter stage, thus aggravating the BIC score. Conversely, setting the threshold too high results in excessive filtration of the initial edge candidate, making the output incapable of accurately representing the data.

$\boldsymbol{N_{\text{init}}}$: Running more instances of the FRP improves the BIC score. While execution time gets longer as N_{init} increases, the growth rate is much slower compared to that of N_{init}.

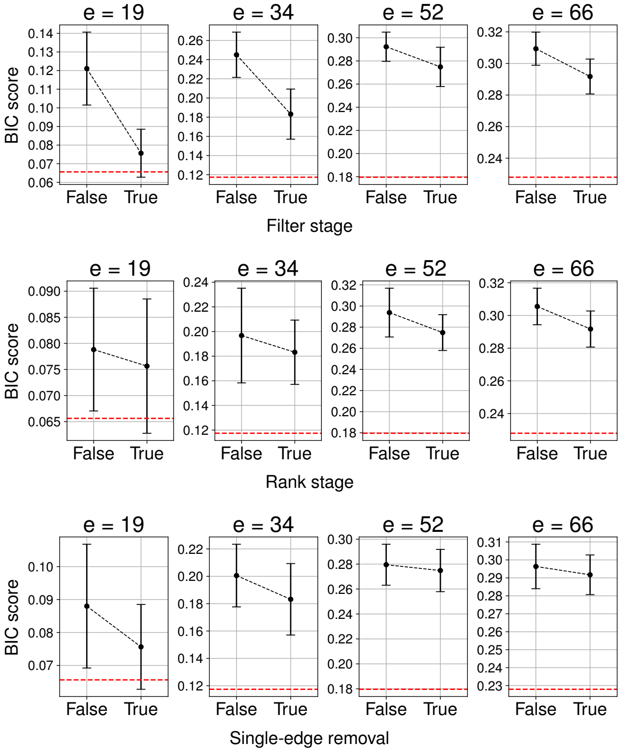

Figure 3: The BIC score between estimated graphs for FRP with (True) and without (False) the Filter stage, the Rank stage, and the one edge removal phase with $p = 15$. The red dashed lines indicate μe.

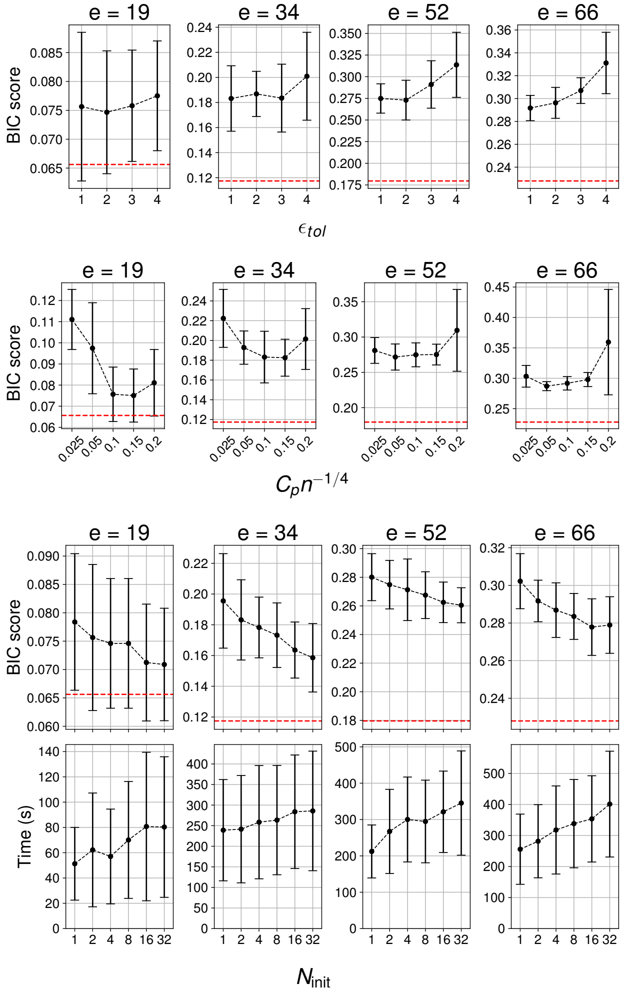

Figure 4: The BIC score/execution time between estimated graphs for FRP with different values of hyperparameters (ϵ_{tol}, $C_p n^{-1/4}$, and N_{init}) with $p = 15$.

d

5.3 Application to a Real World Dataset

We applied FRP to the resting state fMRI data collected by Shah et al. (2018). Although there is no evidence supporting that this data is generated from a linear Gaussian SEM, our results indicate a potential symmetry in the connectivity present in the left and right hemispheres, which aligns with the findings reported by Shah et al. (2018). We refer the readers Appendix E for a more detailed demonstration.

6 DISCUSSION

We presented a novel method, FRP, for learning a linear cyclic Gaussian SEM from observational data. FRP outperformed competitive baselines in terms of the BIC score. We note that our assumption of $\mathbf{X}$ being a full-rank matrix might be invalidated in a high-dimensional setting, where the number of variables p is much larger than the sample size n. In this case, one might replace the MLE estimate of the precision matrix with other methods introduced in the literature of inverse covariance matrix estimation, including the graphical LASSO (Meinshausen and Bühlmann, 2006) to take advantage of sparsity in the underlying graph. Such variation does not affect the FRP but Section 4.1, which does not matter since we can establish a result analogous to Theorem 3.4 as long as the inverse covariance matrix estimation method is consistent.

FRP has some room for improvement. Regarding the filter stage, an edge that violates the edge-faithfulness assumption can be excluded from initial edge candidates, and would not be considered in the following stages. It is possible that the output DG would lose accuracy or add more edges to compensate for this loss of accuracy. Furthermore, the prune stage does not add or exchange edges, so the algorithm might get stuck into a local optimum. To mitigate this issue, transformations between DGs in the same equivalence class proposed by Ghassami et al. (2020) might allow FRP to escape from a local optimum, improving its performance.

Acknowledgments

This work was partly supported by IITP (2022-0-00953-PICA/30%) and NRF (RS-2023-00211904/70%) grant funded by the Korean government (MSIT).

References

Carlos Améndola, Philipp Dettling, Mathias Drton, Federica Onori, and Jun Wu. Structure learning for cyclic linear causal models. In *Conference on Uncertainty in Artificial Intelligence*, pages 999–1008. PMLR, 2020.

Shubharthi Barua, Xiaodan Gao, Hans Pasman, and M. Sam Mannan. Bayesian network based dynamic operational risk assessment. *Journal of Loss Prevention in the Process Industries*, 41:399–410, 2016.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33, 2020.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Soumyadeb Chowdhury, Oscar Rodriguez-Espindola, Prasanta Dey, and Pawan Budhwar. Blockchain technology adoption for managing risks in operations and supply chain management: evidence from the uk. *Annals of Operations Research*, pages 1–36, 2022.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

Chang Deng, Kevin Bello, Bryon Aragam, and Pradeep Kumar Ravikumar. Optimizing NOTEARS objectives via topological swaps. In *Proceedings of the 40th International Conference on Machine Learning*, pages 7563–7595. PMLR, 2023.

Mathias Drton and Marloes H. Maathuis. Structure Learning in Graphical Modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393, 2017.

Paul Erdős and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae*, 6:290–297, 1960.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

Katherine Fitch. Learning directed graphical models from Gaussian data. *arXiv preprint arXiv:1906.08050*, 2019.

Patrick Forré and Joris M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *Conference on Uncertainty in Artificial Intelligence*, pages 269–278, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

AmirEmad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, pages 3494–3504. PMLR, 2020.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, 2019.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

Antti Hyttinen, Patrik O. Hoyer, Frederick Ederhardt, and Matti Järvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Conference on Uncertainty in Artificial Intelligence*, pages 301–310. AUAI Press, 2013.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Jana Janková and Sara van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9:1205–1229, 2015.

David Kaplan. *Structural equation modeling: Foundations and extensions*, volume 10. SAGE publications, 2008.

Maurice George Kendall. *The advanced theory of statistics.*, volume II. Charles Griffin and Co., Ltd., London, 1946.

Rex B. Kline. *Principles and practice of structural equation modeling*. Guilford publications, 2023.

Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Conference on Uncertainty in Artificial Intelligence*, page 366–374, 2008.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2019.

Ben Lengerich, Caleb Ellington, Bryon Aragam, Eric P. Xing, and Manolis Kellis. NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters. *arXiv preprint arXiv:2111.01104*, 2021.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula grphical models. *The Annals of Statistics*, 40(4): 2293–2326, 2012.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15 (1):3065–3105, 2014.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A): 3133–3164, 2009.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Andrew A. Neath and Joseph E. Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the Role of Sparsity and DAG Constraints for Learning Linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, pages 17943–17954. Curran Associates, Inc., 2020.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *The Journal of Machine Learning Research*, 21(1):2896–2929, 2020.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461, 1996.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1): 2553, 2019.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Ruben Sanchez-Romero, Joseph D. Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

Muralikrishnna G. Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Huetter. NODAGS-Flow: Nonlinear Cyclic Causal Structure Learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 6371–6387. PMLR, 2023.

Preya Shah, Danielle S. Bassett, Laura E.M. Wisse, John A. Detre, Joel M. Stein, Paul A. Yushkevich, Russell T. Shinohara, John B. Pluta, Elijah Valenciano, Molly Daffner, David A. Wolk, Mark A. Elliott, Brian Litt, Kathryn A. Davis, and Sandhitsu R. Das. Mapping the structural and functional network architecture of the medial temporal lobe using 7T MRI. *Human Brain Mapping*, 39(2): 851–865, 2018.

Rachna Shah and Susan Meyer Goldstein. Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, 24(2):148–169, 2006.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7(10):2003–2030, 2006.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, and Matthew Webster. Network modelling methods for FMRI. *NeuroImage*, 54:875–891, 2011.

Peter Spirtes, Clark Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. Springer, 2000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

Sara Van de Geer and Peter Bühlmann. ℓ^0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Matthew J. Vowels, Necati C. Camgoz, and Richard Bowden. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*, 2021.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 632–639, 2002.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A OMITTED DEFINITIONS

Definition A.1 (d-separation (Richardson, 1996)). Let $\mathcal{G} = (V, E)$ be a graph with the set of nodes V and the set of edges E. For $X, Y \in V$ and $\mathbf{Z} \subseteq V \setminus \{X, Y\}$, X, Y are d-separated given $\mathbf{Z}$ if and only if there does not exist an acyclic undirected path $(X = X_0, X_1, \ldots, X_n = Y)$ between X and Y satisfying

- for any $1 \leq k \leq n-1$ with $X_k \in \mathbf{Z}$, (X_{k-1}, X_k, X_{k+1}) forms a collider, i.e., $X_{k-1} \to X_k \leftarrow X_{k+1}$, and
- for any $1 \leq k \leq n-1$ with (X_{k-1}, X_k, X_{k+1}) forming a collider, there exists a descendent of X_k that is a member of $\mathbf{Z}$.

B IMPOSSIBILITY OF DISCOVERING LINEAR CYCLIC SEMS VIA MINIMIZING THE LEAST SQUARE ERROR

Consider the following linear cyclic SEM represented by a 2-cycle:

$$\begin{aligned} X_1 &= aX_2 + \epsilon_1, \\ X_2 &= bX_1 + \epsilon_2, \\ \epsilon_1, \epsilon_2 &\sim \mathcal{N}(0, 1), \end{aligned}$$

where ϵ_1, ϵ_2 are independent and a, b are nonzero constants satisfying $ab \neq 1$. Given this, we can compute X_1 and X_2 in terms of ϵ_1 and ϵ_2 as follows:

$$X_1 = \frac{1}{1-ab}(\epsilon_1 + a\epsilon_2), \quad X_2 = \frac{1}{1-ab}(b\epsilon_1 + \epsilon_2). \tag{13}$$

Now, consider the least square loss function of weights $(\widehat{a}, \widehat{b})$:

$$\mathrm{LS}(\widehat{a}, \widehat{b}) = \mathbb{E}[(X_1 - \widehat{a}X_2)^2 + (X_2 - \widehat{b}X_1)^2].$$

Using (13), we can calculate LS in terms of ϵ_1 and ϵ_2 as follows:

$$\begin{aligned} \mathrm{LS}(\widehat{a}, \widehat{b}) &= \frac{1}{(1-ab)^2}\mathbb{E}[(\epsilon_1 + a\epsilon_2 - \widehat{a}(b\epsilon_1 + \epsilon_2))^2 + (b\epsilon_1 + \epsilon_2 - \widehat{b}(\epsilon_1 + a\epsilon_2))^2] \\ &= \frac{1}{(1-ab)^2}\Big[((1-\widehat{a}b)^2 + (b-\widehat{b})^2)\mathbb{E}[\epsilon_1^2] + ((a-\widehat{a})^2 + (1-a\widehat{b})^2)\mathbb{E}[\epsilon_2^2] \\ &\qquad + 2((1-\widehat{a}b)(a-\widehat{a}) + (b-\widehat{b})(1-a\widehat{b}))\mathbb{E}[\epsilon_1\epsilon_2]\Big] \\ &= \frac{1}{(1-ab)^2}\left[((1-\widehat{a}b)^2 + (a-\widehat{a})^2) + ((b-\widehat{b})^2 + (1-a\widehat{b})^2)\right]. \end{aligned}$$

Therefore, the minimizer of LS is given by

$$\widehat{a} = \frac{a+b}{1+b^2}, \quad \widehat{b} = \frac{a+b}{a^2+1},$$

which, in general, are not equal to a and b. In particular, if $a + b = 0$, then $\widehat{a} = \widehat{b} = 0$, thus failing to recover the true causal directions.

C THE BIC AND OUR PERFORMANCE MEASURE

Consider fitting observational data $\mathbf{X} \in \mathbb{R}^{n \times p}$ with a linear Gaussian SEM with an edge set E as in (1). Denote its weight and covariance matrix of exogenous noises by W and Ω, respectively. Then, the log-likelihood function can be represented by the precision matrix $\Theta = (I_p - W)\Omega^{-1}(I_p - W)^\top$ as follows:

$$\ell(\Theta) = -\frac{np}{2}\log 2\pi + \frac{n}{2}\log\det\Theta - \frac{1}{2}\sum_{i=1}^{n}\mathbf{X}_i\Theta\mathbf{X}_i^\top$$

$$
\begin{aligned}
&= -\frac{np}{2}\log 2\pi + \frac{n}{2}\log\det\Theta - \frac{1}{2}\mathrm{tr}(\mathbf{X}^\top\mathbf{X}\Theta) \\
&= -\frac{np}{2}\log 2\pi + \frac{n}{2}\log\det\Theta - \frac{n}{2}\mathrm{tr}(\widehat{\Theta}^{-1}\Theta),
\end{aligned}
$$

where $\widehat{\Theta} = (\mathbf{X}^\top\mathbf{X}/n)^{-1}$. Note that $W_{ij} = 0$ for $(i,j) \notin E$ and Ω is a positive diagonal matrix, so there are $p + |E|$ free parameters in this model in total. Hence, the BIC of this model is

$$
\begin{aligned}
\mathrm{BIC} &= -2\max_{W,\Omega}\ell(\Theta) + (p + |E|)\log n \\
&= -2\max_{W,\Omega}\left\{-\frac{np}{2}\log 2\pi + \frac{n}{2}\log\det\Theta - \frac{n}{2}\mathrm{tr}(\widehat{\Theta}^{-1}\Theta)\right\} + (p+|E|)\log n \\
&= 2n\left[\min_{W,\Omega}\left\{-\frac{1}{2}\log\det\Theta + \frac{1}{2}\mathrm{tr}(\widehat{\Theta}^{-1}\Theta)\right\} + \frac{\log n}{2n}|E|\right] + np\log 2\pi + p\log n \\
&= 2n\left[\min_{W,\Omega}\left\{\mathrm{KL}_{\mathcal{N}}(\widehat{\Theta} \parallel \Theta)\right\} + \frac{\log n}{2n}|E| + \frac{p}{2} - \frac{1}{2}\log\det\widehat{\Theta}\right] + np\log 2\pi + p\log n \qquad (14) \\
&= 2n\mathcal{L}_\mu(E,\widehat{\Theta}) - n\log\det\widehat{\Theta} + (n + n\log 2\pi + \log n)p, \qquad (15)
\end{aligned}
$$

where $\mu = \frac{\log n}{2n}$ is the penalty coefficient. For (15), we used the definition of $\mathcal{L}$ and $\mathcal{L}_\mu$ given in (6) and (9) to obtain

$$
\mathcal{L}_\mu(E,\widehat{\Theta}) = \mathcal{L}(E,\widehat{\Theta}) + \mu|E| = \min_{W,\Omega}\{\mathrm{KL}_{\mathcal{N}}(\widehat{\Theta} \parallel \Theta)\} + \mu|E|,
$$

given $\Theta = (I_p - W)\Omega^{-1}(I_p - W)^\top$. This concludes that the BIC is equivalent to $\mathcal{L}_\mu(E,\widehat{\Theta})$ up to scaling and tranlation, where $\mu = \frac{\log n}{2n}$ and $\widehat{\Theta} = (\mathbf{X}^\top\mathbf{X}/n)^{-1}$.

Furthermore, the similar holds for $\mathcal{L}_\mu(E,\widehat{\Theta})$ and the score function of a DG introduced in Ghassami et al. (2020). The score function is defined as

$$
\widetilde{\mathcal{L}}(\mathbf{X}; W, \Omega) = \min_{W,\Omega}\left\{-n\log\det(I_p - W) + \sum_{i=1}^{p}\left(\frac{n}{2}\log\Omega_{ii} + \frac{\|\mathbf{X}_{:,i} - \mathbf{X}W_{:,i}\|^2}{2\Omega_{ii}}\right)\right\} + \frac{\log n}{2}\|W\|_0
$$

in Equation (3) of Ghassami et al. (2020). Observe that

$$
\begin{aligned}
&-n\log(\det(I_p - W)) + \sum_{i=1}^{p}\frac{n}{2}\log(\Omega_{ii}) \\
&= -\frac{n}{2}\left(\log\det(I_p - W) + \log\det(I_p - W)^\top - \log\det\Omega\right) \\
&= -\frac{n}{2}\log\det((I_p - W)\Omega^{-1}(I_p - W)^\top) \\
&= -\frac{n}{2}\log\det\Theta,
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{i=1}^{p}\frac{1}{2\Omega_{ii}}\|\mathbf{X}_{:,i} - \mathbf{X}W_{:,i}\|^2 \\
&= \frac{1}{2}\sum_{i=1}^{p}(\Omega^{-1})_{ii}(\mathbf{X}(I_p - W)_{:,i})^\top\mathbf{X}(I_p - W)_{:,i} \\
&= \frac{1}{2}\sum_{i=1}^{p}(\Omega^{-1})_{ii}(I_p - W)_{:,i}^\top\mathbf{X}^\top\mathbf{X}(I_p - W)_{:,i} \\
&= \frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{p}\mathbf{X}^\top\mathbf{X}(I_p - W)_{:,i}(\Omega^{-1})_{ii}(I_p - W)_{:,i}^\top\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\left(\sum_{i=1}^{p}(I_p - W)_{:,i}(\Omega^{-1})_{ii}(I_p - W)_{:,i}^\top\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}(I_p - W)\Omega^{-1}(I_p - W)^\top\right) \\
&= \frac{n}{2}\mathrm{tr}(\widehat{\Theta}^{-1}\Theta).
\end{aligned}
$$

Assume W_{ij} is nonzero for $(i,j) \in E$, i.e., $\|W\|_0 = |E|$. Then, we obtain

$$
\begin{aligned}
&\widetilde{\mathcal{L}}(\mathbf{X}; W, \Omega) \\
&= \min_{W,\Omega}\left\{-\frac{n}{2}\log\det\Theta + \frac{n}{2}\mathrm{tr}(\widehat{\Theta}^{-1}\Theta)\right\} + \frac{\log n}{2}|E| \\
&= n\left(\min_{W,\Omega}\left\{-\frac{1}{2}\log\det\Theta + \frac{1}{2}\mathrm{tr}(\widehat{\Theta}^{-1}\Theta) + \frac{1}{2}\log\det\widehat{\Theta} - \frac{p}{2}\right\} + \frac{\log n}{2n}|E|\right) + \frac{n}{2}(p - \log\det\widehat{\Theta}) \\
&= n\mathcal{L}_\mu(E, \widehat{\Theta}) + \frac{n}{2}(p - \log\det\widehat{\Theta}).
\end{aligned}
$$

D DEFERRED PROOFS OF THEOREMS 3.2 AND 3.4

In this appendix, we prove Theorems 3.2 and 3.4 under Assumptions 3.1 and 3.3. To assist understanding, we recall relevant definition, assumptions, and theorems:

Definition 2.1 (λ-edge-faithfulness)**.** Consider a linear SEM with its underlying graph $\mathcal{G} = (V, E)$. We say the linear SEM is λ-*edge-faithful* to $\mathcal{G}$ if $|\mathrm{Corr}(X_i, X_j \mid X_{V\setminus\{i,j\}})| > \lambda$ holds for all $(i,j) \in E$.

Assumption 3.3. Consider the ground truth distribution $\mathcal{N}(0, \Theta^{-1})$ of p variables, where each of them corresponds to $V = \{1, \dots, p\}$. Let ϵ_p be a constant dependent on p. We assume there exists a set of edges E that minimizes $\mathcal{L}_\mu(E, \Theta)$ while ensuring the corresponding linear SEM being ϵ_p-edge-faithful to the graph (V, E).

Assumption 3.1 (Bounded eigenvalues)**.** Consider the ground truth distribution $\mathcal{N}(0, \Theta^{-1})$ of p variables. Let $M_p > 0$ be a constant dependent on p. Then, we assume

$$
1/M_p \le \lambda_{\min}(\Theta) \le \lambda_{\max}(\Theta) \le M_p,
$$

where $\lambda_{\min}(\Theta)$ and $\lambda_{\max}(\Theta)$ are the minimum and maximum eigenvalue of Θ, respectively.

Theorem 3.2. *Consider a Gaussian distribution of p variables as in* (2) *satisfying Assumption 3.1. Let ψ_{ij} and $\widehat{\psi}_{ij}$ be as in* (5)*. Then, for a sufficiently large n, $|\widehat{\psi}_{ij} - \psi_{ij}| \le C_p n^{-1/4}$ holds with probability at least $1 - 2\exp(-cpn^{1/2})$ for all i, j, where $c > 0$ is an absolute constant and $C_p > 0$ depends only on p.*

Theorem 3.4. *Consider a Gaussian distribution $\mathcal{N}(0, \Theta^{-1})$ of p variables, satisfying Assumptions 3.1 and 3.3. Let $\widehat{\psi}_{ij}$ be as in* (5)*. Then, there exists a set of edges E that minimizes $\mathcal{L}_\mu(E, \Theta)$ and satisfies the following property:*

> *For a sufficiently large n, $E \subseteq \{(i,j) \mid |\widehat{\psi}_{ij}| > C_p n^{-1/4}\}$ holds with probability at least $1 - 2\exp(-cpn^{1/2})$, where $c > 0$ is an absolute constant and $C_p > 0$ depends only on p.*

Since Θ is positive definite, and the diagonal entries of a positive definite matrix cannot be smaller than any of its eigenvalues, we have

$$
\Theta_{ii} \ge \lambda_{\min}(\Theta) \ge 1/M_p. \tag{16}
$$

Furthermore, Assumption 3.1 gives bounds to $\|\Theta\|_2$ and $\|\Theta^{-1}\|_2$ as follows:

Lemma D.1. *Assume a linear SEM with p variables equipped with the precision matrix $\Theta \in \mathbb{R}^{p\times p}$ satisfies Assumption 3.1. Then, $\|\Theta\|_2 \le \sqrt{p}M_p$ and $\|\Theta^{-1}\|_2 \le \sqrt{p}M_p$.*

Proof to Lemma D.1. Let $\lambda_1 \ge \cdots \ge \lambda_p$ be eigenvalues of Θ. Let $\Theta = U\Lambda U^\top$ be an eigenvalue decomposition of Θ, where $\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_p)$ and $U^\top U = UU^\top = I_p$. Then, we have

$$
\|\Theta\|_2^2 = \mathrm{tr}(\Theta^2) = \mathrm{tr}(U\Lambda^2U^\top) = \mathrm{tr}(\Lambda^2U^\top U) = \mathrm{tr}(\Lambda^2) = \sum_{i=1}^p \lambda_i^2 \le pM_p^2,
$$

thus, $\|\Theta\|_2 \le \sqrt{p}M_p$. Similarly, we have $\|\Theta^{-1}\|_2 \le \sqrt{p}M_p$. □

We now prove Theorem 3.2.

Proof to Theorem 3.2. Let Σ and $\Theta = \Sigma^{-1}$ be the covariance and precision matrix of the SEM, respectively. We refer to (Vershynin, 2010, Remark 5.40.); setting $t = p^{1/2}n^{1/4}$ (where n, N are defined as in (Vershynin, 2010, Remark 5.40.)) renders

$$\left\| \frac{1}{n} X^\top X - \Sigma \right\|_2 \leq \max(\delta, \delta^2) \|\Sigma\|_2$$

holds with probability at least $1 - 2\exp(-cpn^{1/2})$ where $\delta = (1 + Dn^{-1/4})p^{1/2}n^{-1/4}$ (D and c are constants independent of n and p).

Now, take n sufficiently large to satisfy

$$pM_p^2 \max(\delta, \delta^2) \leq \frac{1}{2},$$

to obtain

$$\|\Sigma\|_2 \|\Sigma^{-1}\|_2 \max(\delta, \delta^2) = \|\Theta\|_2 \|\Theta^{-1}\|_2 \max(\delta, \delta^2) \leq (\sqrt{p}M_p)^2 \max(\delta, \delta^2) \leq \frac{1}{2},$$

using Lemma D.1 and the bound M_p introduced in Assumption 3.1. Denote $\widehat{\Theta} = (\frac{1}{n}\mathbf{X}^\top \mathbf{X})^{-1}$. Using a machinery similar to that in the proof of Lemma 29 of (Loh and Bühlmann, 2014), we have

$$\begin{aligned} \|\widehat{\Theta} - \Theta\|_2 &= \left\| \left(\frac{1}{n}\mathbf{X}^\top \mathbf{X} \right)^{-1} - \Sigma^{-1} \right\|_2 \\ &\leq 2\|\Sigma^{-1}\|_2^2 \|\Sigma\|_2 \max(\delta, \delta^2) \\ &= 2\|\Theta\|_2^2 \|\Theta^{-1}\|_2 \max(\delta, \delta^2) \\ &\leq 2p^{3/2} M_p^3 \max(\delta, \delta^2). \end{aligned} \tag{17}$$

with probability at least $1 - 2\exp(-cpn^{1/2})$.

Now assume this event is the case. Take n to be sufficiently large to satisfy

$$\|\widehat{\Theta} - \Theta\|_2 \leq 2p^{3/2} M_p^3 \max(\delta, \delta^2) < \min\left(\frac{1}{2M_p}, \sqrt{p}M_p \right),$$

so that invoking (16) gives

$$\widehat{\Theta}_{ii} \geq \Theta_{ii} - \|\widehat{\Theta} - \Theta\|_2 \geq \frac{1}{M_p} - \|\widehat{\Theta} - \Theta\|_2 > \frac{1}{2M_p} \tag{18}$$

for all i and (17) gives

$$|\widehat{\Theta}_{ij}| \leq \|\widehat{\Theta}\|_2 \leq \|\Theta\|_2 + \|\widehat{\Theta} - \Theta\|_2 < 2\sqrt{p}M_p \tag{19}$$

for any i, j. Now, see for an example $i = 1$ and $j = 2$,

$$\begin{aligned} \widehat{\psi}_{12} - \psi_{12} &= \frac{\Theta_{12}}{\sqrt{\Theta_{11}\Theta_{22}}} - \frac{\widehat{\Theta}_{12}}{\sqrt{\widehat{\Theta}_{11}\widehat{\Theta}_{22}}} \\ &= \underbrace{\frac{\widehat{\Theta}_{12}}{\sqrt{\widehat{\Theta}_{11}}} \left(\frac{1}{\sqrt{\Theta_{22}}} - \frac{1}{\sqrt{\widehat{\Theta}_{22}}} \right)}_{(1)} + \underbrace{\frac{\widehat{\Theta}_{12}}{\sqrt{\Theta_{22}}} \left(\frac{1}{\sqrt{\Theta_{11}}} - \frac{1}{\sqrt{\widehat{\Theta}_{11}}} \right)}_{(2)} + \underbrace{\frac{1}{\sqrt{\Theta_{11}\Theta_{22}}} (\Theta_{12} - \widehat{\Theta}_{12})}_{(3)}. \end{aligned}$$

We exploit (16), (18), and (19) to bound each term. For (1),

$$\begin{aligned} \left| \frac{\widehat{\Theta}_{12}}{\sqrt{\widehat{\Theta}_{11}}} \left(\frac{1}{\sqrt{\Theta_{22}}} - \frac{1}{\sqrt{\widehat{\Theta}_{22}}} \right) \right| &= \left| \frac{\widehat{\Theta}_{12}}{\sqrt{\widehat{\Theta}_{11}}} \left(\frac{\widehat{\Theta}_{22} - \Theta_{22}}{\sqrt{\widehat{\Theta}_{22}\Theta_{22}} \left(\sqrt{\widehat{\Theta}_{22}} + \sqrt{\Theta_{22}} \right)} \right) \right| \\ &\leq 4\sqrt{p} M_p^3 \|\widehat{\Theta} - \Theta\|_2. \end{aligned}$$

and similar for (2),

$$\left| \frac{\widehat{\Theta}_{12}}{\sqrt{\Theta_{22}}} \left(\frac{1}{\sqrt{\Theta_{11}}} - \frac{1}{\sqrt{\widehat{\Theta}_{11}}} \right) \right| \leq 4\sqrt{p} M_p^3 \|\widehat{\Theta} - \Theta\|_2.$$

Finally, for (3),

$$\left| \frac{1}{\sqrt{\Theta_{11}\Theta_{22}}} (\Theta_{12} - \widehat{\Theta}_{12}) \right| \leq M_p \|\widehat{\Theta} - \Theta\|_2.$$

Putting it all together, we have

$$\begin{aligned} |\widehat{\psi}_{12} - \psi_{12}| &\leq \left(8\sqrt{p} M_p^3 + M_p\right) \|\widehat{\Theta} - \Theta\|_2 \\ &\leq 2p^{3/2} \left(8\sqrt{p} M_p^2 + 1\right) M_p^4 \max(\delta, \delta^2). \end{aligned}$$

This holds simultaneously for pairs other than $i = 1$ and $j = 2$, as long as the event (17) holds with probability at least $1 - 2\exp(-cpn^{1/2})$. By taking n sufficiently large, we have the desired result. □

Now we prove Theorem 3.4 to conclude this section.

Proof to Theorem 3.4. Let c and C_p be as in Theorem 3.2, so we have

$$|\widehat{\psi}_{ij} - \psi_{ij}| \leq C_p n^{-1/4}$$

with probability at least $1 - 2\exp(-cpn^{1/2})$. Now, let E be a set of edges that satisfies Assumption 3.3. Taking n sufficiently large to have $\epsilon_p > 2C_p n^{-1/4}$ leads to

$$\begin{aligned} (i, j) \in E &\Rightarrow |\psi_{ij}| > \epsilon_p > 2C_p n^{-1/4} \\ &\Rightarrow |\widehat{\psi}_{ij}| > C_p n^{-1/4}. \end{aligned}$$

□

E EXPERIMENTAL DETAILS FOR SECTION 5.3

We have used functional MRI dataset, publicly available at `https://github.com/shahpreya/MTLnet` (Shah et al., 2018), We have confirmed that the GitHub repository containing is licensed under the GNU General Public License v3.0. However, despite our best efforts, we were unable to ascertain licensing terms that specifically apply to the dataset. If there are any concerns or inquiries related to this matter, we will make every effort to address them to the best of our abilities.

The MTL dataset consists of the resting state fMRI data of 24 healthy adults. As we do not have a deep understanding of this domain, we basically followed a similar procedure taken in (Shah et al., 2018) regarding data selection. Each hemisphere were segemented into 10 subregions (CA1, CA2, DG, CA3, TAIL, SUB, ERC, BA35, BA36, PHC). We refer the readers to (Shah et al., 2018) for more accurate and detailed information.

We applied FRP separately for the left and right hemispheres for each subject. Then, we calculated the occurrence of connection, i.e., the existence of any edges between two nodes for each pair of nodes. The result is depicted in Figure 5.

Additionally, we sorted 10 regions by the total number of connections across all subjects. The results for the left hemisphere are as follows, in descending order: DG, CA1, SUB, ERC, TAIL, CA3, BA35, CA2, PHC, and BA36. For the right hemisphere, the order is ERC, SUB, DG, CA1, BA36, BA35, TAIL, CA2, PHC, CA3. Remarkably, we emphasize that ERC, SUB, CA1, and DG are the top four regions with the highest number of connections in both hemispheres. This finding aligns with Shah et al. (2018), which reported that CA1, DG, and SUB serve as functional hubs.

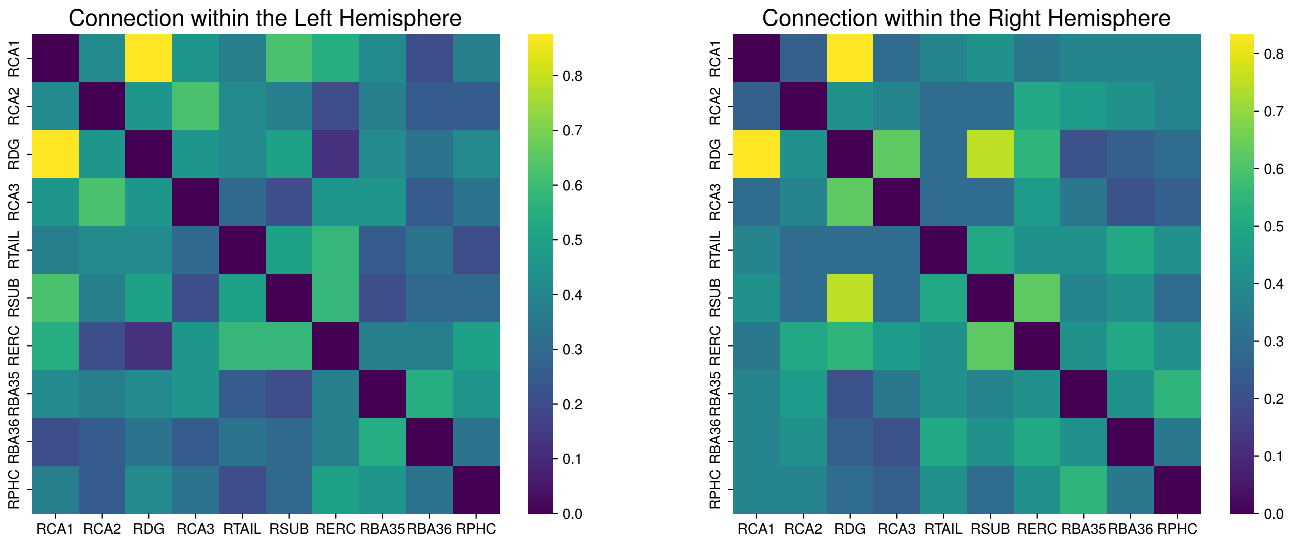

Figure 5: For the functional MRI data from 24 subjects, we applied FRP separately to the left and right hemispheres to discover connections within the brain regions.

F DETAILS FOR SOLVING OPTIMIZATION PROBLEMS

Random Initialization Approach Since the optimization problems we are solving (including (11) and (7)) are non-convex problems, we used a random initialization approach to solve them similar to approaches employed by Ghassami et al. (2020). For each problem, we ran L-BFGS-B at maximum `max_iters` $= 50$ times, with each component of initial points independently sampled from the standard normal distribution. If the function value at the end of each run does not get smaller than the best function value obtained so far minus `tol` $= 10^{-6}$ for `patience` $= 10$ consecutive runs, we stopped the optimization process.

A Regularization Term in the Rank Stage For solving the problem stated as (11), we utilized the SCAD penalty (Fan and Li, 2001), precisely defined as

$$\mathrm{reg}(Q) = \sum_{1 \le i \ne j \le p} \mathrm{SCAD}(Q_{ij}; \lambda, \gamma),$$

where SCAD is defined as

$$\mathrm{SCAD}(x; \lambda, \gamma) = \begin{cases} \lambda |x| & \text{if } |x| \le \lambda, \\ \dfrac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma - 1)} & \text{if } \lambda < |x| \le \gamma\lambda, \\ \dfrac{(\gamma+1)\lambda^2}{2} & \text{if } |x| > \gamma\lambda, \end{cases}$$

where $\lambda > 0$ and $\gamma > 2$ are hyperparameters. We set $\lambda = \mu$, where μ is a penalty term for an edge regarding (9), and $\gamma = 3.7$ as introduced in Fan and Li (2001) for all experiments. For implementation, we used `optimize.minimize` of `SciPy` (Harris et al., 2020) based on `NumPy` (Virtanen et al., 2020).

G DETAILED EXPERIMENTAL SETTINGS

Calculating Sparsity of Precision Matrix by Number of Edges Assume we create an Erdős-Rényi graph with p nodes and e edges. That is, we randomly select e edges among all possible $p(p-1)$ directed edges. Let W and $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ denote the weighted adjacency matrix and the covariance matrix of the exogenous noise of the determined linear Gaussian

SEM, respectively. Then, the precision matrix Θ of the SEM is given by

$$\Theta_{ij} = -\sigma_i^{-2} W_{ji} - \sigma_j^{-2} W_{ij} + \sum_{k \neq i,j} \sigma_k^{-2} W_{ik} W_{jk}$$

for all $i \neq j$. Given that $\sigma_i^{-2} > 0$ for all i and W_{ij} are sampled from the uniform distribution, we obtain an event $[\Theta_{ij} = 0]$ is equivalent to $[W_{ij} = W_{ji} = 0] \cap [W_{ik} W_{jk} = 0 \text{ for all } k \neq i, j]$ except a probability zero event. Note that $\mathbb{P}(W_{ij} = 0) = 1 - \frac{e}{p(p-1)}$, the probability of (i, j) not being an edge in the graph. Since all events $[W_{ij} = 0]$ are independent from each other, we have

$$\mathbb{P}(W_{ij} = W_{ji} = 0) = \left(1 - \frac{e}{p(p-1)}\right)^2, \quad \mathbb{P}(W_{ik} W_{jk} = 0) = 1 - \left(\frac{e}{p(p-1)}\right)^2$$

for any distinct i, j, and k. Therefore, we can calculate the expected number of zero entries in Θ as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \neq j} \mathbb{1}(\Theta_{ij} = 0)\right] &= \sum_{i \neq j} \mathbb{P}(\Theta_{ij} = 0) \\
&= \sum_{i \neq j} \mathbb{P}(W_{ij} = W_{ji} = 0) \mathbb{P}(W_{ik} W_{jk} = 0 \text{ for all } k \neq i, j) \\
&= \sum_{i \neq j} \mathbb{P}(W_{ij} = W_{ji} = 0) \prod_{k \neq i,j} \mathbb{P}(W_{ik} W_{jk} = 0) \\
&= p(p-1) \left(1 - \frac{e}{p(p-1)}\right)^2 \left(1 - \left(\frac{e}{p(p-1)}\right)^2\right)^{p-2}.
\end{aligned}$$

From this, if the expected number of zero entries in the precision matrix is given, then the number of edges e can be computed.

Experimental Settings of Section 5.1 Regarding DGLEARN, we followed the default choice hyperparameters by Ghassami et al. (2020) as introduced in examples uploaded at `https://github.com/syanga/dglearn`. We set `tabu_length` = `tabu_patience` = 4 for the `tabu_search` stage, and `max_path_len` = 6 for the `virtual_refine` stage. Timeouts of the `tabu_search`, `hill_climbing`, and `reduce_support` steps of DGLEARN are set to be 1800, 900, 900 seconds, respectively. For GOLEM, we set the ℓ^1 regularizer to $\log(n)/2n$ to ensure it minimizes the score having a similar scale to FRP and DGLEARN. For NODAGS-Flow, we follow the default hyperparameter choice provided in the implementation by Sethuraman et al. (2023), while increasing `epoch` from 10 to 500.

H DETAILED EXPERIMENTAL RESULTS

Table 1: Comparison between rates with which the λ-edge-faithfulness / λ-strong-faithfulness holds. The assumptions are checked for 1000 ground truths generated following the procedure in (12), and $\lambda = 0.01$. Success rate of the strong-faithfulness are upper bounded by checking d-separation for two nodes with given a set of zero, one, or two nodes. In every case, λ-strong-faithfulness holds with a lower rate than λ-edge-faithfulness.

p	e	λ-edge-faithfulness	λ-strong-faithfulness
10	9	0.972	≤ 0.774
10	33	0.591	0.0
20	31	0.888	0.0
20	105	0.078	0.0

Table 2: BIC scores for FRP (ours), DGLEARN (Ghassami et al., 2020), GOLEM (Ng et al., 2020), NODAGS-Flow (Sethuraman et al., 2023), and NOTEARS (Zheng et al., 2018). Boldface indicates the best score for each case. FRP outperforms the baselines in all cases except for $(p, e) = (10, 9)$.

		FRP	GOLEM	DGLEARN	NODAGS-Flow	NOTEARS
p	e	mean(std)	mean(std)	mean(std)	mean(std)	mean(std)
10	9	0.035(0.005)	0.102(0.069)	**0.032(0.002)**	0.193(0.044)	0.036(0.006)
	17	**0.075(0.010)**	0.111(0.013)	0.085(0.013)	0.221(0.053)	0.080(0.020)
	26	**0.106(0.009)**	0.133(0.012)	0.112(0.013)	0.264(0.042)	0.124(0.017)
	33	**0.124(0.010)**	0.136(0.006)	0.125(0.006)	0.238(0.039)	0.136(0.014)
15	19	**0.076(0.013)**	0.160(0.040)	0.082(0.010)	0.340(0.092)	0.083(0.021)
	34	**0.183(0.026)**	0.277(0.027)	0.204(0.023)	0.461(0.132)	0.207(0.035)
	52	**0.275(0.017)**	0.314(0.014)	0.287(0.030)	0.569(0.115)	0.326(0.095)
	66	**0.292(0.011)**	0.333(0.036)	0.413(0.195)	0.568(0.089)	0.317(0.017)
20	31	**0.143(0.015)**	0.368(0.034)	0.178(0.044)	0.564(0.086)	0.227(0.203)
	55	**0.360(0.040)**	0.529(0.045)	0.640(0.348)	0.864(0.208)	0.600(0.520)
	83	**0.480(0.024)**	0.561(0.027)	0.790(0.450)	0.775(0.178)	0.564(0.116)
	105	**0.515(0.020)**	0.559(0.014)	0.898(0.573)	0.790(0.176)	0.567(0.034)

Table 3: Execution times for FRP (ours), DGLEARN (Ghassami et al., 2020), GOLEM (Ng et al., 2020), NODAGS-Flow (Sethuraman et al., 2023), and NOTEARS (Zheng et al., 2018). Boldface indicates the best case for each case.

		FRP	GOLEM	DGLEARN	NODAGS-Flow	NOTEARS
p	e	mean(std)	mean(std)	mean(std)	mean(std)	mean(std)
10	9	4.080(2.425)	235.718(21.784)	2.441(2.215)	300.620(11.658)	**0.740**(**0.391**)
	17	25.452(22.152)	246.797(13.409)	48.423(44.516)	285.956(22.957)	**3.820**(**2.814**)
	26	56.521(38.110)	229.183(18.110)	229.297(189.930)	279.808(7.663)	**6.316**(**5.275**)
	33	46.259(27.379)	231.528(18.211)	172.567(199.396)	136.662(84.006)	**1.801**(**1.880**)
15	19	62.225(45.000)	89.369(3.748)	35.498(58.764)	110.855(13.949)	**1.890**(**1.255**)
	34	241.533(130.406)	95.752(3.598)	946.452(548.439)	135.051(20.481)	**3.791**(**2.674**)
	52	267.319(115.728)	95.142(5.135)	2026.537(935.796)	105.219(0.122)	**7.353**(**4.930**)
	66	281.527(117.877)	109.579(36.123)	2248.411(808.805)	105.241(0.325)	**6.691**(**4.222**)
20	31	172.274(83.908)	240.816(18.407)	303.569(274.758)	131.622(0.077)	**7.830**(**9.267**)
	55	775.805(400.416)	231.985(23.899)	2632.588(388.483)	131.606(0.080)	**17.495**(**11.540**)
	83	575.942(399.972)	237.677(30.874)	2763.594(111.471)	177.985(97.752)	**8.119**(**8.956**)
	105	580.293(374.534)	235.148(16.390)	2688.290(471.796)	410.854(93.712)	**9.945**(**8.215**)

Table 4: The BIC scores of FRP with and without the Filter stage.

		With Filter stage	*Without* Filter stage
p	e	mean(std)	mean(std)
10	9	**0.035**(**0.005**)	0.054(0.01)
	17	**0.075**(**0.01**)	0.094(0.015)
	26	**0.106**(**0.009**)	0.121(0.007)
	33	**0.124**(**0.01**)	0.133(0.009)
15	19	**0.076**(**0.013**)	0.121(0.02)
	34	**0.183**(**0.026**)	0.245(0.024)
	52	**0.275**(**0.017**)	0.292(0.013)
	66	**0.292**(**0.011**)	0.309(0.01)
20	31	**0.143**(**0.015**)	0.221(0.05)
	55	**0.36**(**0.04**)	0.473(0.032)
	83	**0.48**(**0.024**)	0.523(0.013)
	105	**0.515**(**0.02**)	0.54(0.016)

Table 5: The BIC scores of FRP with and without the Rank stage.

		With Rank stage	*Without* Rank stage
p	e	mean(std)	mean(std)
10	9	**0.035(0.005)**	0.035(0.005)
	17	0.075(0.01)	**0.075(0.014)**
	26	**0.106(0.009)**	0.107(0.011)
	33	**0.124(0.01)**	0.124(0.007)
15	19	**0.076(0.013)**	0.079(0.012)
	34	**0.183(0.026)**	0.197(0.038)
	52	**0.275(0.017)**	0.294(0.023)
	66	**0.292(0.011)**	0.306(0.011)
20	31	0.143(0.015)	**0.136(0.008)**
	55	**0.36(0.04)**	0.4(0.058)
	83	**0.48(0.024)**	0.523(0.033)
	105	**0.515(0.02)**	0.554(0.035)

Table 6: The BIC scores of FRP with and without the single-edge removal phase.

		With the single-edge removal	*Without* the single-edge removal
p	e	mean(std)	mean(std)
10	9	**0.035(0.005)**	0.036(0.007)
	17	**0.075(0.01)**	0.076(0.011)
	26	**0.106(0.009)**	0.108(0.009)
	33	**0.124(0.01)**	0.125(0.011)
15	19	**0.076(0.013)**	0.088(0.019)
	34	**0.183(0.026)**	0.201(0.023)
	52	**0.275(0.017)**	0.28(0.016)
	66	**0.292(0.011)**	0.296(0.012)
20	31	**0.143(0.015)**	0.175(0.025)
	55	**0.36(0.04)**	0.39(0.034)
	83	**0.48(0.024)**	0.503(0.02)
	105	**0.515(0.02)**	0.525(0.023)