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 Filter, Rank, and Prune (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.
We summarize our contributions as follows. (1) Regarding 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. Forró and Mooij (2018) exploited $\sigma$-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 $\ell^1$ 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), Meihausen and Bühmann (2006), and 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 $\ell^1$ 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 $\ell^0$ penalty. Ghassami et al. (2020) proposed a score function that is the sum of likelihood loss and the $\ell^0$ 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. (2000) 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 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. Forró and Mooij (2018) exploited $\sigma$-separation to permit unmeasured confounders.

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 Filter, Rank, and Prune (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
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 $\text{Diag}(A)$ 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:

$$X = XW + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Omega),$$  \hspace{1cm} (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 $\Omega$ is the covariance matrix of exogenous noise. We assume that the system has no self-loop, thus, $\text{Diag}(W) = 0$. To represent multiple samples, we denote $X \in \mathbb{R}^{n \times p}$ as a vertical stack of $n \gg p$ samples from a linear SEM. Therefore, $X$ satisfies

$$X = XW + \mathcal{E}, \quad \mathcal{E}_{i:} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Omega).$$  \hspace{1cm} (2)

**Underlying DG and Faithfulness** Consider a linear SEM with the weight matrix $W$ as in (1). Its underlying directed graph $G = (V, E)$ 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 $G$ if

$$X_i \perp X_j \mid X_S \Rightarrow i$ and $j$ are $d$-separated given $S$ in $G$$

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

$$|\text{Corr}(X_i, X_j \mid X_S)| \leq \lambda$$

$$\Rightarrow i$ and $j$ are $d$-separated given $S$ in $G$,$$  \hspace{1cm} (3)

where $\lambda > 0$ is a constant. However, $\lambda$-strong-faithfulness is too restrictive: $\lambda$-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 $\lambda$-edge-faithfulness:

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

This relaxes $\lambda$-strong-faithfulness (3) in two folds. First, $S$ 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 $i \rightarrow k \leftarrow j$. 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 recausalize 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 $Z$, the partial correlation of $X$ and $Y$ given $Z$ is defined

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

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

$\Sigma = (I_p - W)^{-1} \Omega (I_p - W)^{-1}$, that is,

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

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

$$\psi_{ij} = \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}, \quad \hat{\psi}_{ij} := \frac{\hat{\Theta}_{ij}}{\sqrt{\hat{\Theta}_{ii}\hat{\Theta}_{jj}}}. $$  \hspace{1cm} (5)

Given the precise relationship between $\psi_{ij}$ and $\Theta$, 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 $\text{KL}(P_1 || P_2)$ 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 $\Theta_1$ and $\Theta_2$, we write
\[
\text{KL}_{\mathcal{N}}(\Theta_1 \parallel \Theta_2) := \text{KL}(\mathcal{N}(0, \Theta_1^{-1}) \parallel \mathcal{N}(0, \Theta_2^{-1})) \\
= \frac{1}{2} \text{tr}(\Theta_1^{-1} - \Theta_2^{-1}) + \frac{1}{2} \log \det \Theta_1 - \frac{1}{2} \log \det \Theta_2 - \frac{p}{2}.
\]
Identifying a DG with its edge set $E$, we can simplify 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:
\[
\mathcal{L}(E, \Theta) := \min\{\text{KL}(\Theta \parallel (I_p - \bar{W})\Theta^{-1}(I_p - \bar{W})^\top) \mid \text{Supp}\bar{W} \subseteq E, \Theta \text{ positive diagonal}\},
\]
where $\Theta$ is the precision matrix of the true distribution and $\text{Supp}\bar{W} := \{(i, j) \mid \bar{W}_{ij} \neq 0\}$. 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 $\Theta$, calculating $\mathcal{L}(E, \Theta)$ is a mathematically challenging task as it involves solving a non-convex optimization problem:
\[
\text{minimize}_{\bar{W}, \Omega \in \mathbb{R}^{p \times p}} \text{KL}(\Theta \parallel (I_p - \bar{W})\Omega^{-1}(I_p - \bar{W})^\top), \\
\text{subject to } \text{Supp}\bar{W} \subseteq E, \Omega \text{ positive diagonal}.
\]
We can simplify it by denoting $Q := (I_p - \bar{W})\Omega^{-1/2}$, so that we have $QQ^\top = (I_p - \bar{W})\Omega^{-1}(I_p - \bar{W})^\top$. 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 $\Omega^{-1/2}$ is a positive diagonal matrix, the zero entries of $Q$ and $I_p - \bar{W}$ coincide. Therefore, we obtain the following problem:
\[
\text{minimize}_{Q \in \mathbb{R}^{p \times p}} \text{KL}(\Theta \parallel QQ^\top), \\
\text{subject to } Q_{ij} = 0 \text{ for all } i \neq j \text{ with } (i, j) \notin E.
\]
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 $\Theta$ 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 $\text{KL}(\Theta \parallel \cdot)$ is convex when defined on the set of positive definite matrices, $\text{KL}(\Theta \parallel QQ^\top)$ 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 $X \in \mathbb{R}^{n \times p}$ follows a Gaussian distribution
\[
X_i \sim \mathcal{N}(0, \Theta^{-1}),
\]
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 $\text{BIC} = -2 \log L + k \log n$, 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:
\[
\mathcal{L}_\mu(E, \Theta) := \mathcal{L}(E, \Theta) + \mu |E|.
\]

Aligned with our motivation for constructing this measure, $\mathcal{L}_\mu(E, \Theta)$ is equivalent to the BIC up to scaling and translation when $\mu = \log n$ and $\Theta = (X^\top X/n)^{-1}$. 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 $X$ is a full-rank matrix so that $X^\top X$ is invertible. Therefore, we can compute the MLE of $\Theta$ by $\hat{\Theta} = (X^\top X/n)^{-1}$, which characterizes $\hat{\psi}_{ij}$ defined in (5). We can prove that $\hat{\psi}_{ij}$ 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).
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 $\psi_{ij}$ from observational data and invoke Theorem 3.2 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 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 $\psi_{ij}$ from observational data and invoke Theorem 3.2 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 $1 - 2 \exp(-cp^1/2)$, $E \subseteq \{(i, j) \mid |\psi_{ij}| > Cpn^{-1/4}\}$. Now, consider a procedure of removing $(i, j)$ that satisfies $|\hat{\psi}_{ij}| \leq Cpn^{-1/4}$ 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 $E_0 = \{(i, j) \mid |\hat{\psi}_{ij}| > Cpn^{-1/4}\}$ is a sensible choice. We note that setting the threshold $Cpn^{-1/4} = 0.1$ 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., $\Theta_{ij}$, instead of partial correlations $\psi_{ij}$ 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.
with the increase in only a small fraction of the empirical loss.

Prune Stage  Let $\hat{E}$ be the ordered set of edges sorted from the previous stage. We aim to prune edges as much as possible while maintaining the empirical score of the remaining edges. 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, especially the necessity of this process is supported in Section 5.2. The Prune stage is described in Algorithm 1. The FRP algorithm is described in Algorithm 2. It starts from excessive additional time costs. Finally, we apply FRP to a global optimum. Therefore, we run multiple instances (it has been observed that instances are enough) of FRP and their original versions to ensure a fair comparison. Next, we present ablation studies and the roles of hyperparameters on the performance.

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

1: function \text{RANK}(\hat{E}, \Theta)
2: \hspace{1cm} Solve (11) to obtain $Q^*$
3: \hspace{1cm} Sort $(i, j) \in \hat{E}$ by $|Q^*_ij|$ ascendingly
4: \hspace{1cm} return $\hat{E}$
5: function \text{PRUNE}(\hat{E}, \Theta, \epsilon_{\text{tol}})
6: \hspace{1cm} $T \leftarrow \Delta(\hat{E}, \Theta) + \epsilon_{\text{tol}}$
7: \hspace{1cm} $i \leftarrow \max\{i \geq 0 | \Delta(\hat{E}[i:], \Theta) \leq T\}$ \triangleright Binary search
8: \hspace{1cm} if $i \geq 1$ return $\hat{E}[i:]$
9: \hspace{1cm} for $i \leftarrow 1$ to $|\hat{E}| - 1$ do
10: \hspace{2cm} if $\Delta(\hat{E}[-i], \Theta) \leq T$
11: \hspace{2cm} \hspace{1cm} return $\hat{E}[-i]$ \hspace{1cm} \triangleright Single-edge removal phase
12: \hspace{1cm} return $\hat{E}$

Algorithm 2 The FRP Algorithm

1: function \text{FRP}(X, \epsilon_{\text{tol}}, C_p, N_{\text{init}})
2: \hspace{1cm} $\Theta \leftarrow (X' X/n)^{-1}$
3: \hspace{1cm} $\overline{\psi}_{ij} \leftarrow -\frac{\sqrt{\Theta_{ij}}}{\Theta_{jj}}$ for all $i, j$
4: \hspace{1cm} for $k \leftarrow 1$ to $N_{\text{init}}$ do \triangleright Parallel execution
5: \hspace{2cm} $E_k \leftarrow \{(i, j) | \overline{\psi}_{ij} > C_p n^{-1/4}\}$
6: \hspace{2cm} while True do
7: \hspace{3cm} $E_k \leftarrow \text{RANK}(E_k, \Theta)$
8: \hspace{3cm} $E_k' \leftarrow \text{PRUNE}(E_k, \Theta, \epsilon_{\text{tol}})$
9: \hspace{3cm} if $E_k' = E_k$ then break
10: \hspace{3cm} $E_k \leftarrow E_k'$
11: \hspace{1cm} return \argmin_{E_k} \mu(\hat{E}_k, \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/gentech/nodags-flows for NODAGS-Flow and GOLEM, and https://github.com/xunzheng/notears for NOTEARS.

Synthetic 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 C. When an underlying DG $G = (V, E)$ is specified, we generate a random weight matrix $W$ and the covariance matrix of exogenous noise $\Omega$.
by the following procedure:

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

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

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 evaluate 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 \(\mathcal{L}_\mu(\tilde{E}, \Theta)\), where \(\tilde{E}\) is the estimated edge set and \(\Theta\) 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 \(\mu = \frac{\log n}{2E}\) for \(\mathcal{L}_\mu(\tilde{E}, \Theta)\) \((9)\). 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 \(\ell_1\)-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.

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

\(\epsilon_{\text{tol}}\): 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 \(\mathcal{L}_\mu(\tilde{E}, \tilde{\Theta})\) which is an estimate but not the true BIC score.

\(C_p n^{-1/4}\): 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.

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

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


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 \cite{Richardson1996}). Let \(G = (V, E)\) be a graph with the set of nodes \(V\) and the set of edges \(E\). For \(X, Y \in V\) and \(Z \subseteq V \setminus \{X, Y\}\), \(X, Y\) are \(d\)-separated given \(Z\) 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 \(X_k \in Z\), \((X_{k-1}, X_k, X_{k+1})\) forms a collider, i.e., \(X_{k-1} \rightarrow X_k \leftarrow X_{k+1}\), 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 \(Z\).

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

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

\[
\begin{align*}
X_1 &= aX_2 + \epsilon_1, \\
X_2 &= bX_1 + \epsilon_2,
\end{align*}
\]

where \(\epsilon_1, \epsilon_2\) 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 \(\epsilon_1\) and \(\epsilon_2\) as follows:

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

(13)

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

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

Using (13), we can calculate LS in terms of \(\epsilon_1\) and \(\epsilon_2\) as follows:

\[
LS(\tilde{a}, \tilde{b}) = \frac{1}{(1-ab)^2}E[(\epsilon_1 + a\epsilon_2 - \tilde{a}(b\epsilon_1 + \epsilon_2))^2 + (b\epsilon_1 + \epsilon_2 - \tilde{b}(\epsilon_1 + a\epsilon_2))^2]
\]

\[
= \frac{1}{(1-ab)^2}\left[ (1-\tilde{a})^2 + (b - \tilde{b})^2)E[\epsilon_1^2] + ((a - \tilde{a})^2 + (1-\tilde{b})^2)E[\epsilon_2^2] + 2((1-\tilde{a})(a - \tilde{a}) + (b - \tilde{b})(1-a\tilde{b}))E[\epsilon_1\epsilon_2]\right]
\]

\[
= \frac{1}{(1-ab)^2}\left[ (1-\tilde{a})^2 + (a - \tilde{a})^2 + ((b - \tilde{b})^2 + (1-\tilde{b})^2)\right].
\]

Therefore, the minimizer of LS is given by

\[
\hat{\tilde{a}} = \frac{a + b}{1+b^2}, \quad \hat{\tilde{b}} = \frac{a + b}{a^2 + 1},
\]

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

C THE BIC AND OUR PERFORMANCE MEASURE

Consider fitting observational data \(X \in \mathbb{R}^{n \times p}\) 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 \(\Omega\), 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:

\[
\ell(\Theta) = -\frac{np}{2}\log 2\pi + \frac{n}{2}\log \det \Theta - \frac{1}{2} \sum_{i=1}^{n} X_i\Theta X_i^\top
\]
\[
\begin{align*}
&= -\frac{np}{2} \log 2\pi + \frac{n}{2} \log \det \Theta - \frac{1}{2} \text{tr}(X^T X \Theta) \\
&= -\frac{np}{2} \log 2\pi + \frac{n}{2} \log \det \Theta - \frac{n}{2} \text{tr}(\Theta^{-1} \Theta),
\end{align*}
\]

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

\[
\text{BIC} = -2 \max_{W, \Omega} \ell(\Theta) + (p + |E|) \log n
\]

\[
= -2 \max_{W, \Omega} \left\{ -\frac{np}{2} \log 2\pi + \frac{n}{2} \log \det \Theta - \frac{n}{2} \text{tr}(\Theta^{-1} \Theta) \right\} + (p + |E|) \log n
\]

\[
= 2n \left[ \min_{W, \Omega} \left\{ -\frac{1}{2} \log \det \Theta + \frac{1}{2} \text{tr}(\Theta^{-1} \Theta) \right\} + \frac{\log n}{2n} |E| \right] + np \log 2\pi + p \log n
\]

\[
= 2n \left[ \min_{W, \Omega} \{ \text{KL}_N(\hat{\Theta} \parallel \Theta) \} + \frac{\log n}{2n} |E| + \frac{p}{2} - \frac{1}{2} \log \det \Theta \right] + np \log 2\pi + p \log n \tag{14}
\]

\[
= 2n \mathcal{L}_\mu(E, \hat{\Theta}) - n \log \det \Theta + (n + n \log 2\pi + \log n)p, \tag{15}
\]

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

\[
\mathcal{L}_\mu(E, \hat{\Theta}) = \mathcal{L}(E, \hat{\Theta}) + \mu|E| = \min_{W, \Omega} \{ \text{KL}_N(\hat{\Theta} \parallel \Theta) \} + \mu|E|,
\]

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

Furthermore, the similar holds for \( \mathcal{L}_\mu(E, \hat{\Theta}) \) and the score function of a DG introduced in \([15] \) of \( \text{Ghassami et al.} \). The score function is defined as

\[
\tilde{\mathcal{L}}(X; W, \Omega) = \min_{W, \Omega} \left\{ -n \log \det(I_p - W) + \sum_{i=1}^p \frac{n}{2} \log \Omega_{ii} + \frac{\|X_{:,i} - XW_{:,i}\|^2}{2\Omega_{ii}} \right\} + \frac{\log n}{2} \|W\|_0
\]

in Equation (3) of \([15] \) of \( \text{Ghassami et al.} \). Observe that

\[
- n \log(\det(I_p - W)) + \sum_{i=1}^p \frac{n}{2} \log(\Omega_{ii})
\]

\[
= - \frac{n}{2} \left( \log \det(I_p - W) + \log \det(I_p - W)^T - \log \det \Omega \right)
\]

\[
= - \frac{n}{2} \log \det((I_p - W)\Omega^{-1}(I_p - W)^T)
\]

\[
= - \frac{n}{2} \log \det(\Theta),
\]

and

\[
\sum_{i=1}^p \frac{1}{2\Omega_{ii}} \|X_{:,i} - XW_{:,i}\|^2
\]

\[
= \frac{1}{2} \sum_{i=1}^p (\Omega^{-1})_{ii}(X(I_p - W)_{:,i})^T X(I_p - W)_{:,i}
\]

\[
= \frac{1}{2} \sum_{i=1}^p (\Omega^{-1})_{ii} X(I_p - W)_{:,i}^T X(I_p - W)_{:,i}
\]

\[
= \frac{1}{2} \text{tr} \left( \sum_{i=1}^p X^T X(I_p - W)_{:,i} (\Omega^{-1})_{ii} (I_p - W)_{:,i}^T \right)
\]

\[
= \frac{1}{2} \text{tr} \left( X^T X \left( \sum_{i=1}^p (I_p - W)_{:,i} (\Omega^{-1})_{ii} (I_p - W)_{:,i}^T \right) \right)
\]
\[ \frac{1}{2} \text{tr} \left( X^T (I_p - W) \Omega^{-1} (I_p - W)^T \right) = \frac{n}{2} \text{tr}(\hat{\Theta}^{-1} \Theta). \]

Assume \( W_{ij} \) is nonzero for \((i, j) \in E\), i.e., \( \|W\|_0 = |E| \). Then, we obtain
\[
\tilde{L}(X; W, \Omega) = \min_{W, \Omega} \left\{ -\frac{n}{2} \log \det \Theta + \frac{n}{2} \text{tr}(\hat{\Theta}^{-1} \Theta) \right\} + \log n \frac{|E|}{2}
\]
\[
= n \left( \min_{W, \Omega} \left\{ -\frac{1}{2} \log \det \Theta + \frac{1}{2} \text{tr}(\hat{\Theta}^{-1} \Theta) + \frac{1}{2} \log \det \hat{\Theta} - \frac{p}{2} \right\} + \log n \frac{|E|}{2} \right) + \frac{n}{2} (p - \log \det \hat{\Theta})
\]
\[= n \mathcal{L}_\mu(E, \hat{\Theta}) + \frac{n}{2} (p - \log \det \hat{\Theta}). \]

### 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 (\( \lambda \)-edge-faithfulness).** Consider a linear SEM with its underlying graph \( G = (V, E) \). We say the linear SEM is \( \lambda \)-edge-faithful to \( G \) if \( |\text{Corr}(X_i, X_j | X_{V \setminus \{i, j\}})| > \lambda \) 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 \( V = \{1, \ldots, p\} \). Let \( \epsilon_p \) 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 \( \epsilon_p \)-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
\[ 1/M_p \leq \lambda_{\min}(\Theta) \leq \lambda_{\max}(\Theta) \leq M_p, \]
where \( \lambda_{\min}(\Theta) \) and \( \lambda_{\max}(\Theta) \) are the minimum and maximum eigenvalue of \( \Theta \), respectively.

**Theorem 3.2.** Consider a Gaussian distribution of \( p \) variables as in (2) satisfying Assumption 3.1. Let \( \psi_{ij} \) and \( \hat{\psi}_{ij} \) be as in (5). Then, for a sufficiently large \( n \), \( |\hat{\psi}_{ij} - \psi_{ij}| \leq C_p n^{-1/4} \) holds with probability at least \( 1 - 2 \exp(-cpn^{1/2}) \) for all \( i, j \), where \( c \) 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 \( \hat{\psi}_{ij} \) be as in (5). Then, there exists a set of edges \( E \) that minimizes \( \mathcal{L}_\mu(E, \Theta) \) and satisfies the following property:

\[ \text{For a sufficiently large } n, E \subseteq \{(i, j) \mid |\hat{\psi}_{ij}| > C_p n^{-1/4}\} \text{ holds with probability at least } 1 - 2 \exp(-cpn^{1/2}), \]

where \( c > 0 \) is an absolute constant and \( C_p > 0 \) depends only on \( p \).

Since \( \Theta \) is positive definite, and the diagonal entries of a positive definite matrix cannot be smaller than any of its eigenvalues, we have
\[ \Theta_{ii} \geq \lambda_{\min}(\Theta) \geq 1/M_p. \]  

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, \( \|\Theta\|_2 \leq \sqrt{p} M_p \) and \( \|\Theta^{-1}\|_2 \leq \sqrt{p} M_p \).

**Proof to Lemma D.1.** Let \( \lambda_1 \geq \cdots \geq \lambda_p \) be eigenvalues of \( \Theta \). Let \( \Theta = U \Lambda U^T \) be an eigenvalue decomposition of \( \Theta \), where \( \Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_p) \) and \( U^T U = U U^T = I_p \). Then, we have
\[ \|\Theta\|_2^2 = \text{tr}(\Theta^2) = \text{tr}(U \Lambda^2 U^T) = \text{tr}(\Lambda^2) = \sum_{i=1}^p \lambda_i^2 \leq p M_p^2, \]
thus, \( \|\Theta\|_2 \leq \sqrt{p} M_p \). Similarly, we have \( \|\Theta^{-1}\|_2 \leq \sqrt{p} M_p \). \(\square\)
We now prove Theorem 3.2.

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

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

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

Now, take \( n \) sufficiently large to satisfy

\[
pM^2 \max(\delta, \rho^2) \leq \frac{1}{2},
\]

to obtain

\[
\|\Sigma\|_2\|\Sigma^{-1}\|_2 \max(\delta, \rho^2) = \|\Theta\|_2\|\Theta^{-1}\|_2 \max(\delta, \rho^2) \leq (\sqrt{p}Mp)^2 \max(\delta, \rho^2) \leq \frac{1}{2},
\]

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

\[
\|\hat{\Theta} - \Theta\|_2 = \left\| \left( \frac{1}{n}X^\top X \right)^{-1} - \Sigma^{-1} \right\|_2
\]

\[
\leq 2\|\Sigma^{-1}\|_2\|\Sigma\|_2 \max(\delta, \rho^2)
\]

\[
= 2\|\Theta\|_2\|\Theta^{-1}\|_2 \max(\delta, \rho^2)
\]

\[
\leq 2p^{3/2}M^3 \max(\delta, \rho^2).
\]

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

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

\[
\|\hat{\Theta} - \Theta\|_2 \leq 2p^{3/2}M^3 \max(\delta, \rho^2) < \min\left( \frac{1}{2M_p}, \sqrt{p}M_p \right),
\]

so that invoking (16) gives

\[
\hat{\Theta}_{ii} \geq \Theta_{ii} - \|\hat{\Theta} - \Theta\|_2 \geq \frac{1}{M_p} - \|\hat{\Theta} - \Theta\|_2 > \frac{1}{2M_p}
\]

for all \( i \) and (17) gives

\[
|\hat{\Theta}_{ij}| \leq \|\hat{\Theta}\|_2 \leq \|\Theta\|_2 + \|\hat{\Theta} - \Theta\|_2 < 2\sqrt{p}M_p
\]

for any \( i, j \). Now, see for an example \( i = 1 \) and \( j = 2 \),

\[
\hat{\psi}_{12} - \psi_{12} = \frac{\hat{\Theta}_{12}}{\sqrt{\hat{\Theta}_{11}\hat{\Theta}_{22}}} - \frac{\Theta_{12}}{\sqrt{\Theta_{11}\Theta_{22}}}
\]

\[
= \frac{\hat{\Theta}_{12}}{\sqrt{\hat{\Theta}_{11}}} \left( \frac{1}{\sqrt{\Theta_{22}}} - \frac{1}{\sqrt{\hat{\Theta}_{22}}} \right) + \frac{\hat{\Theta}_{12}}{\sqrt{\Theta_{11}}} \left( \frac{1}{\sqrt{\Theta_{22}}} - \frac{1}{\sqrt{\hat{\Theta}_{22}}} \right) + \frac{1}{\sqrt{\Theta_{11}\Theta_{22}}} (\Theta_{12} - \hat{\Theta}_{12}).
\]

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

\[
\left| \frac{\hat{\Theta}_{12}}{\sqrt{\hat{\Theta}_{11}}} \left( \frac{1}{\sqrt{\Theta_{22}}} - \frac{1}{\sqrt{\hat{\Theta}_{22}}} \right) \right| = \left| \frac{\hat{\Theta}_{12}}{\sqrt{\hat{\Theta}_{11}}} \left( \frac{\hat{\Theta}_{22} - \Theta_{22}}{\sqrt{\hat{\Theta}_{22}\Theta_{22}} (\sqrt{\hat{\Theta}_{22}} + \sqrt{\Theta_{22}})} \right) \right|
\]

\[
\leq 4\sqrt{M^3} \sqrt{\Theta_{11}} \|\Theta - \Theta\|_2.
\]
Soheun Yi, Sanghack Lee

and similar for (2),

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

Finally, for (3),

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

Putting it all together, we have

$$|\hat{\psi}_{12} - \psi_{12}| \leq (8\sqrt{p}M_p^3 + M_p) \| \hat{\Theta} - \Theta \|_2 \leq 2p^{3/2} (8\sqrt{p}M_p^2 + 1) M_p^4 \max(\delta, \delta^2).$$

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. \hfill \Box  

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

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

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

$$(i, j) \in E \Rightarrow |\hat{\psi}_{ij}| > \epsilon_p > 2C_p n^{-1/4} \Rightarrow |\hat{\psi}_{ij}| > C_p n^{-1/4}.$$ \hfill \Box  

**E EXPERIMENTAL DETAILS FOR SECTION 5.3**

We have used functional MRI dataset, publicly available at [https://github.com/shahpreya/MTLnet](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 segmented 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, BA36, PHC. 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

\[
\text{reg}(Q) = \sum_{1 \leq i \neq j \leq p} \text{SCAD}(Q_{ij}; \lambda, \gamma),
\]

where SCAD is defined as

\[
\text{SCAD}(x; \lambda, \gamma) = \begin{cases} 
\lambda |x| & \text{if } |x| \leq \lambda, \\
2\gamma \lambda |x| - x^2 - \lambda^2 & \text{if } \lambda < |x| \leq \gamma \lambda, \\
(\gamma + 1)\lambda^2 & \text{if } |x| > \gamma \lambda,
\end{cases}
\]

where \( \lambda > 0 \) and \( \gamma > 2 \) are hyperparameters. We set \( \lambda = \mu \), where \( \mu \) 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 \texttt{optimize.minimize} of \texttt{SciPy} (Harris et al., 2020) based on \texttt{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 \( \Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) \) denote the weighted adjacency matrix and the covariance matrix of the exogenous noise of the determined linear Gaussian
SEM, respectively. Then, the precision matrix $\Theta$ 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 $[W_{ij} = W_{ji} = 0] \cap [W_{ik}W_{jk} = 0$ for all $k \neq i, j]$ except a probability zero event. Note that $P(W_{ij} = 0) = 1 - \frac{e}{p(p-1)}$, 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

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

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

$$\mathbb{E} \left[ \sum_{i \neq j} 1(\Theta_{ij} = 0) \right] = \sum_{i \neq j} P(\Theta_{ij} = 0)$$
$$= \sum_{i \neq j} P(W_{ij} = W_{ji} = 0)P(W_{ik}W_{jk} = 0$ for all $k \neq i, j)$
$$= \sum_{i \neq j} P(W_{ij} = W_{ji} = 0) \prod_{k \neq i, j} 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}.$$

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](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 $\ell^1$ 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.
## DETAILED EXPERIMENTAL RESULTS

Table 1: Comparison between rates with which the \( \lambda \)-edge-faithfulness / \( \lambda \)-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, \( \lambda \)-strong-faithfulness holds with a lower rate than \( \lambda \)-edge-faithfulness.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( e )</th>
<th>( \lambda )-edge-faithfulness</th>
<th>( \lambda )-strong-faithfulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
<td>0.972</td>
<td>( \leq 0.774 )</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
<td>0.591</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>31</td>
<td>0.888</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>105</td>
<td>0.078</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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 \( (p, e) = (10, 9) \).

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

<table>
<thead>
<tr>
<th>p</th>
<th>e</th>
<th>FRP mean(std)</th>
<th>GOLEM mean(std)</th>
<th>DGLEARN mean(std)</th>
<th>NODAGS-Flow mean(std)</th>
<th>NOTEARS mean(std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
<td>4.080(2.425)</td>
<td>235.718(21.784)</td>
<td>2.441(2.215)</td>
<td>300.620(11.658)</td>
<td>0.740(0.391)</td>
</tr>
<tr>
<td>17</td>
<td>25.452(22.152)</td>
<td>246.797(13.409)</td>
<td>48.423(44.516)</td>
<td>285.956(22.957)</td>
<td>3.820(2.814)</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>56.521(38.110)</td>
<td>229.183(18.110)</td>
<td>229.297(189.930)</td>
<td>279.808(7.663)</td>
<td>6.316(5.275)</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>46.259(27.379)</td>
<td>231.528(18.211)</td>
<td>172.567(199.396)</td>
<td>136.662(84.006)</td>
<td>1.801(1.880)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>34</td>
<td>241.533(130.406)</td>
<td>95.752(3.598)</td>
<td>946.452(548.439)</td>
<td>135.051(20.481)</td>
<td>3.791(2.674)</td>
</tr>
<tr>
<td>52</td>
<td>267.319(115.728)</td>
<td>95.142(5.135)</td>
<td>2026.537(935.796)</td>
<td>105.219(0.122)</td>
<td>7.353(4.930)</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>281.527(117.877)</td>
<td>109.570(36.123)</td>
<td>2248.411(808.805)</td>
<td>105.241(0.325)</td>
<td>6.691(4.222)</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>p</th>
<th>e</th>
<th>With Filter stage mean(std)</th>
<th>Without Filter stage mean(std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
<td>0.035(0.005)</td>
<td>0.054(0.01)</td>
</tr>
<tr>
<td>17</td>
<td>25.452(22.152)</td>
<td>0.075(0.01)</td>
<td>0.094(0.015)</td>
</tr>
<tr>
<td>26</td>
<td>56.521(38.110)</td>
<td>0.106(0.009)</td>
<td>0.121(0.007)</td>
</tr>
<tr>
<td>33</td>
<td>46.259(27.379)</td>
<td>0.124(0.01)</td>
<td>0.133(0.009)</td>
</tr>
<tr>
<td>15</td>
<td>34</td>
<td>0.076(0.013)</td>
<td>0.121(0.02)</td>
</tr>
<tr>
<td>52</td>
<td>267.319(115.728)</td>
<td>0.183(0.026)</td>
<td>0.245(0.024)</td>
</tr>
<tr>
<td>66</td>
<td>281.527(117.877)</td>
<td>0.275(0.017)</td>
<td>0.292(0.013)</td>
</tr>
<tr>
<td>20</td>
<td>31</td>
<td>0.143(0.015)</td>
<td>0.221(0.05)</td>
</tr>
<tr>
<td>55</td>
<td>575.942(399.972)</td>
<td>0.36(0.04)</td>
<td>0.473(0.032)</td>
</tr>
<tr>
<td>83</td>
<td>835.805(400.416)</td>
<td>0.48(0.024)</td>
<td>0.523(0.013)</td>
</tr>
<tr>
<td>105</td>
<td>580.293(374.534)</td>
<td>0.515(0.02)</td>
<td>0.54(0.016)</td>
</tr>
</tbody>
</table>
Table 5: The BIC scores of FRP with and without the Rank stage.

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

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

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