Optimal Backward Learning Approach to Gaussian Linear Structural Equation Models

Sunmin Oh   Donguk Shin   Youngmin Ahn   Gunwoong Park

*All authors have contributed equally.

Department of Statistics, Seoul National University

Contributions

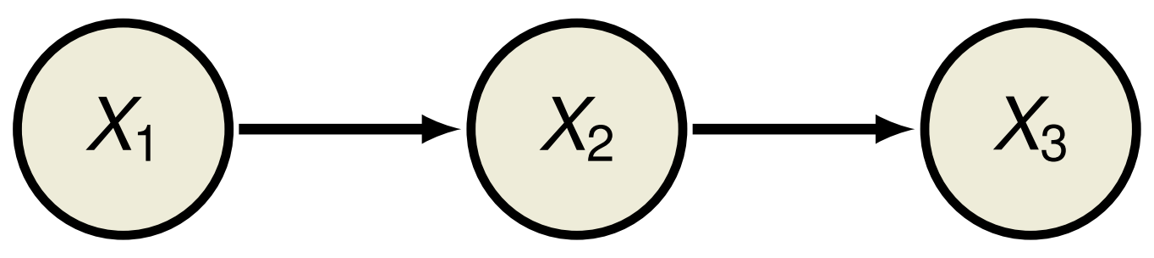
✓ **Provide the class of optimally recoverable Gaussian linear structural equation models (SEMs).**

✓ **Develop the first optimal backward learning approach for Gaussian linear SEMs using the best-subset selection.**

Preliminaries

Definitions & Notations

- A DAG $G = (V, E)$ comprises a set of nodes $V = \{1, 2, ..., p\}$ and a set of directed edges $E \subset V \times V$ with no directed cycles. $V = \{1, 2, 3\}$, $E = \{(1, 2), (2, 3)\}$.



- Ordering: $\pi = (\pi_1, \pi_2, \pi_3) = (1, 2, 3)$ represents directions of edges. Usually not unique.
- Parents (Pa): $X_1 \to X_2$ and $\mathrm{Pa}(2) = \{1\}$.
- Ancestors (An): $X_1 \to X_2 \to X_3$, $X_2 \to X_3$ and $\mathrm{An}(3) = \{1, 2\}$.
- d_{in}: Maximum indegree.
- d_m: Maximum degree of the moralized graph.

Gaussian Linear SEMs: A Gaussian linear SEM is a special type of DAG model in which the joint distribution is defined by the following linear equations: For all $j \in V$,

$$X_j = \sum_{k \in \mathrm{Pa}(j)} \beta_{k,j} X_k + \epsilon_j,$$

where $\beta_{k,j}$ is the linear weight of an edge from X_k to X_j, and $\epsilon = (\epsilon_j)_{j \in V}$ are independent but possibly non-identical Gaussian distributions, $N(0, \sigma_j^2)$.

Previous Works for Gaussian Linear SEMs

- Peters and Bühlmann (2014) show Gaussian linear SEMs with equal error variances are identifiable.
- Many learning algorithms have been developed. Particularly, Gao et al. (2022) develop an optimal forward learning algorithm for Gaussian linear SEM with equal error variances.

Optimally Recoverable Models by Backward Learning Approach

Necessity of the New Identifiability

♣ **To relax the equal error variance constraint on the class of identifiable Gaussian linear SEMs.**

❖ **To apply backward learning strategy rather than forward one in an optimal manner.**

Theorem: Identifiability Condition

For a Gaussian linear SEM with true ordering π, DAG G is uniquely identifiable if for any node $j = \pi_r \in V, \ell \in \mathrm{An}(j)$, and $T = V \setminus \{\pi_{r+1}, ..., \pi_p\}$,

$$\max_{\substack{T' \subset T \setminus \{\ell\} \\ |T'| = \min\{d_{in}+1, |T|-1\}}} \mathrm{Var}(X_\ell \mid X_{T'})^{-1} > \mathrm{Var}(X_j \mid X_{T \setminus \{j\}})^{-1}.$$

- This identifiable class of Gaussian linear SEMs, which can be optimally recoverable, is obtained by slightly modifying backward learning condition of Park(2020).
- In order to apply backward learning strategy, the size of conditioning set has to be adjusted as $d_{in} + 1$ while that of Gao et al. (2022) is d_{in}.
- In this setting, the true ordering π can be estimated from the last, using the best-subset selection, since the conditioning set of falsely ordered variable must include some non-ancestral variable.

♣ **The proposed condition is far weaker than the equal error variance assumption.**

The Need for Backward Learning Strategy

In the Gaussian linear SEM settings, a conditional variance can be obtained from the inverse covariance matrix. Hence, identifiable conditions depend on error variances as well as edge weights.

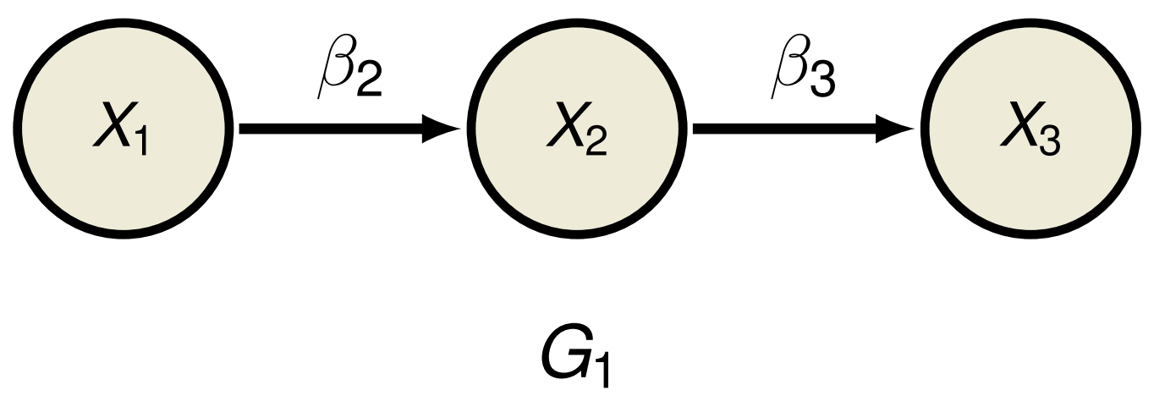


Figure: Three-node Gaussian linear SEMs where $X_1 = \epsilon_1, X_2 = \beta_2 X_1 + \epsilon_2, X_3 = \beta_3 X_2 + \epsilon_3$.

- Forward Learning Condition (Park, 2020)

$$\text{(A1)}\ \ \sigma_1^2 < \sigma_2^2 + \beta_2^2\sigma_1^2, \quad \text{(A2)}\ \ \sigma_2^2 < \sigma_3^2 + \beta_3^2\sigma_2^2, \quad \text{(A3)}\ \ \sigma_1^2 < \sigma_3^2 + \beta_3^2\sigma_2^2 + \beta_2^2\beta_3^2\sigma_1^2.$$

- Proposed Condition

$$\text{(B1)}\ \ \sigma_1^2 < \sigma_2^2 + \beta_2^2\sigma_1^2, \quad \text{(B2)}\ \ \sigma_2^2 < \sigma_3^2 + \beta_3^2\sigma_2^2, \quad \text{(B3)}\ \ \sigma_1^2\sigma_2^2 < \sigma_2^2\sigma_3^2 + \beta_2^2\sigma_1^2\sigma_3^2.$$

- The first two conditions are identical. However, the last condition is different:
  If $(\beta_1, \beta_2, \beta_3) = (0.3, 0.3, 0.3)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (10^2, 0.1^2, 1)$, then (B3) is satisfied, whereas (A3) is violated.

❖ **Backward learning strategy can recover some Gaussian linear SEMs which forward learning strategy can not recover.**

Optimal Backward Learning (OptBGSM) Algorithm

OptBGSM Algorithm

```
Input  : $n$ i.i.d samples $X^{1:n}$, threshold $\eta$, and the maximum indegree bound $\Delta_{in}$
Output: Estimated graph structure, $\widehat{G} = (V, \widehat{E})$
Set $S = V$
for $r = \{p, p-1, ..., 2\}$ do
  Step 1) Determine the $r$-th element of the ordering and its candidate parent set. For any $j \in S$,
      $\widehat{C}_j := \arg\min_{C_j \subset S \setminus \{j\},\ |C_j| = \min\{\Delta_{in}+1, r-1\}} \widehat{\mathrm{Var}}(X_j \mid X_{C_j}), \quad \hat{\pi}_r := \arg\max_{j \in S} \widehat{\mathrm{Var}}(X_j \mid X_{\widehat{C}_j})$
  Step 2) Determine the parent set of $\hat{\pi}_r$ and its size.
      $\hat{q} := \arg\max_{q \le \min\{\Delta_{in}, r-1\}} \left( \min_{D_{q-1} \subset \widehat{C}_{\hat{\pi}_r},\ |D_{q-1}| = q-1} \widehat{\mathrm{Var}}(X_{\hat{\pi}_r} \mid X_{D_{q-1}}) - \min_{D_q \subset \widehat{C}_{\hat{\pi}_r},\ |D_q| = q} \widehat{\mathrm{Var}}(X_{\hat{\pi}_r} \mid X_{D_q}) > \eta \right)$
      $\widehat{\mathrm{Pa}}(\hat{\pi}_r) := \arg\min_{D_{\hat{q}} \subset \widehat{C}_{\hat{\pi}_r}, |D_{\hat{q}}| = \hat{q}} \widehat{\mathrm{Var}}(X_{\hat{\pi}_r} \mid X_{D_{\hat{q}}})$
  Update $S = S \setminus \{\hat{\pi}_r\}$
end
Return: Estimated edge set, $\widehat{E} = \cup_{r \in \{p, p-1, ..., 2\}} \{(k, \widehat{\pi}_r) : k \in \widehat{\mathrm{Pa}}(\widehat{\pi}_r)\}$
```

- In parent estimation, OptBGSM applies backward elimination while Gao et al. (2022) apply forward selection.
- Backward elimination is favorable particularly when the indegree of each node is different.

Conditional Variance Estimator: For any node $j \in V$ and $T \subset V \setminus \{j\}$,

$$\widehat{\mathrm{Var}}(X_j \mid X_T) = \widehat{\Sigma}_{j,j} - \widehat{\Sigma}_{j,T}\widehat{\Sigma}_{T,T}^{-1}\widehat{\Sigma}_{T,j},$$

where $\widehat{\Sigma}_{A,B}$ is the $|A| \times |B|$ submatrix of the sample covariance matrix $\widehat{\Sigma}$ corresponding to X_A and X_B.

Theorem: Consistency of the OptBGSM Algorithm

Under regularity conditions, there exist positive $A_1 > 0$ and $A_2 > 0$ such that

$$P(\widehat{G} = G) \ge 1 - A_1 (p/d_{in})^{d_{in}} \exp(-A_2 n).$$

Lemma: Lower Bound of Sample Complexity for Arbitrary Estimator

For any $0 < \delta < 1/2$, there are positive $B_1 > 0$ and $B_2 > 0$ such that for any estimator $\widehat{G}$,

$$n \le (1 - 2\delta) B_1 d_{in} \log(p/d_{in}) \implies \sup_{F \in \mathcal{F}_{p,d_{in}}} P(\widehat{G} \ne G(F)) \ge \delta - \frac{B_2}{p d_{in} \log(p/d_{in})},$$

where $\mathcal{F}_{p,d_{in}}$ is a class of p-node Gaussian linear SEMs with d_{in}, which satisfying regularity conditions.

Corollary: Optimality of the OptBGSM Algorithm

The proposed algorithm is optimal in sample complexity $n \asymp d_{in} \log(p/d_{in})$ under regularity conditions.

Comparison to Previous Works

Algorithm	Sample complexity	Error distribution	Mutual incoherence	Computational complexity
OptBGSM	$\Omega(d_{in} \log(p/d_{in}))$	heterogeneous	Not required	$O(p^2 \min\{n, p^{d_{in}+1} d_{in}^3\})$
OptFGSM	$\Omega(d_{in} \log(p/d_{in}))$	identical	Not required	$O(p^2 \min\{n, p^{d_{in}+1} d_{in}^3\})$
BHLSM	$\Omega(d_m^2 \log p)$	heterogeneous	Required	$O(n(p^3 + p^2 d_m^2))$
LISTEN	$\Omega(d_m^4 \log p)$	heterogeneous	Not required	$O(d_m p \log p + d_m^2 p)$
TD	$\Omega(d_{in}^2 \log p)$	identical	Required	larger than $O(d_{in}^3 \binom{p}{d_{in}})$

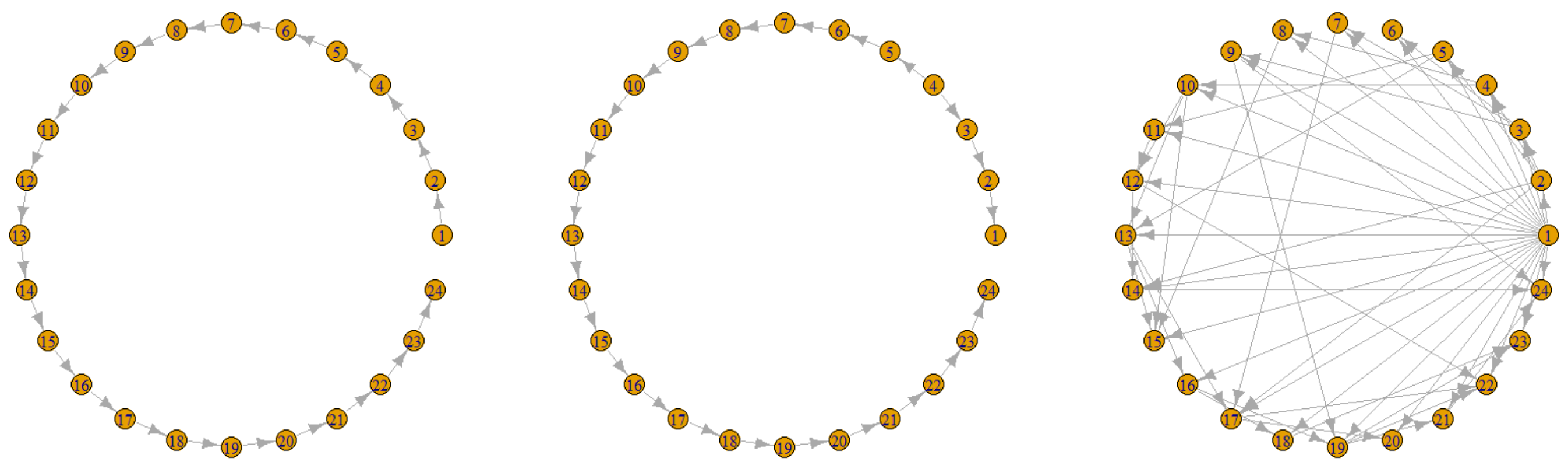
- OptFGSM (Gao et al., 2022), BHLSM (Park et al., 2021), LISTEN (Ghoshal and Honorio, 2018), TD (Chen et al., 2019) algorithms.

Hourly Soil Temperature Data (Seoul, South Korea)

- Source: Korea Meteorological Administration (https://data.kma.go.kr/cmmn/main.do).
- # of samples: 92 (from August 1 to October 31, 2022).
- # of variables: 24 hour variables (1-24).
- The soil temperature at each time is a function of the preceding hour's temperature and other weather factors such as humidity, precipitation, and the time the sun rises and sets:

$$X_j = \beta_{j-1,j} X_{j-1} + \epsilon_j \quad \text{for any } j \in \{2, ..., 24\}.$$

- Because of the seasonality effect and variability in other weather factors, the error variances are considerably large, which can cause violation of identifiability condition.



(a) True Graph   (b) OptBGSM   (c) OptFGSM

Figure: The hourly soil temperature DAG estimated by OptBGSM and OptFGSM algorithms.

Algorithm	Falsely detected	Missed	Reversed
OptBGSM	None	None	3 edges
OptFGSM	55 edges	20 edges	None

- Forward learning identifiability condition is severely violated owing to different error variances.
- However, OptBGSM successfully detects temporal relationships.