Discovering causal structures in privacy-protected data: Frugality in anchored Gaussian DAG models

Junhyoung Chung

November 1, 2024

Seoul National University
Department of Statistics

Main contributions and outline

Main contributions

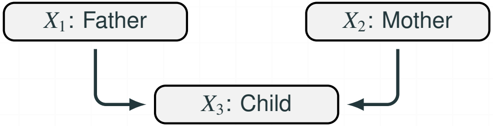
- Discover an identifiability condition for Gaussian linear SEMs with **post-randomized additive measurement error**.
- Develop a consistent algorithm that **captures an underlying true CPDAG**.

Outline

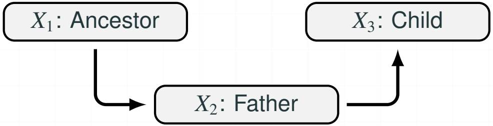
- Motivation
- Anchored DAG model
- Model identifiability
- Algorithm
- Numerical experiments
- Discussion

Directed Acyclic Graphical (DAG) model

- A DAG model is a useful tool to figure out relationships between variables.
- A DAG model is identifiable up to its MEC under the faithfulness assumption.
- Suppose that there are three variables of family gene information, $X_3 = f(X_1, X_2)$ (functional relationship):



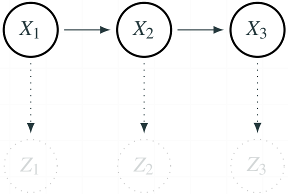
- $X_1 \perp\!\!\!\perp X_2, \quad X_1 \not\perp\!\!\!\perp X_2 \mid X_3,$
- $X_1 \not\perp\!\!\!\perp X_3, \quad X_1 \not\perp\!\!\!\perp X_3 \mid X_2,$
- $X_2 \not\perp\!\!\!\perp X_3, \quad X_2 \not\perp\!\!\!\perp X_3 \mid X_1.$



- $X_1 \not\perp\!\!\!\perp X_2, \quad X_1 \not\perp\!\!\!\perp X_2 \mid X_3,$
- $X_1 \not\perp\!\!\!\perp X_3, \quad X_1 \perp\!\!\!\perp X_3 \mid X_2,$
- $X_2 \not\perp\!\!\!\perp X_3, \quad X_2 \not\perp\!\!\!\perp X_3 \mid X_1.$

Anchored DAG model

- How to solve the problem of causal discovery with *measurement errors*?
- Estimating causal relationships directly from corrupted data may lead to incorrect inference.



Anchored graph

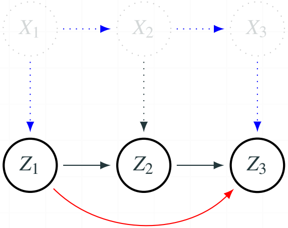
- $X_1 \not\perp X_2, \quad X_1 \not\perp X_3, \quad X_2 \not\perp X_3,$
  $X_1 \not\perp X_2 \mid X_3, \quad X_1 \perp\!\!\!\perp X_3 \mid X_2, \quad X_2 \not\perp X_3 \mid X_1.$

- $Z_1 \not\perp Z_2, \quad Z_1 \not\perp Z_3, \quad Z_2 \not\perp Z_3,$
  $Z_1 \not\perp Z_2 \mid Z_3, \quad Z_1 \not\perp Z_3 \mid Z_2, \quad Z_2 \not\perp Z_3 \mid Z_1.$

- X: Latent variables
- Z: Observed variables

Anchored DAG model

- How to solve the problem of causal discovery with *measurement errors*?
- Estimating causal relationships directly from corrupted data may lead to incorrect inference.



Anchored graph

- X: Latent variables
- Z: Observed variables

- $X_1 \not\perp\!\!\!\perp X_2, \quad X_1 \not\perp\!\!\!\perp X_3, \quad X_2 \not\perp\!\!\!\perp X_3,$
  $X_1 \not\perp\!\!\!\perp X_2 \mid X_3, \quad X_1 \perp\!\!\!\perp X_3 \mid X_2, \quad X_2 \not\perp\!\!\!\perp X_3 \mid X_1.$

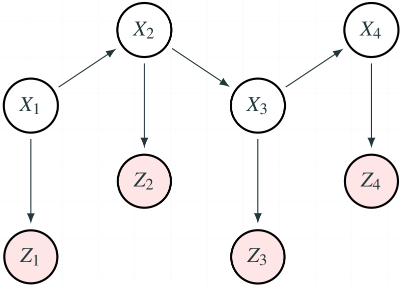
- $Z_1 \not\perp\!\!\!\perp Z_2, \quad Z_1 \not\perp\!\!\!\perp Z_3, \quad Z_2 \not\perp\!\!\!\perp Z_3,$
  $Z_1 \not\perp\!\!\!\perp Z_2 \mid Z_3, \quad Z_1 \not\perp\!\!\!\perp Z_3 \mid Z_2, \quad Z_2 \not\perp\!\!\!\perp Z_3 \mid Z_1.$

Frugality property: Graph theory

Frugality property using graph theory

Consider a p-variate anchored DAG.

- If a pair of latent nodes is d-connected, the corresponding pair of anchored nodes is also d-connected by any set of anchored nodes.



- An active path between X_1 and X_4 is *blocked* by X_2 or X_4.
- An active path between Z_1 and Z_4 *cannot be blocked* by Z_2 or Z_3.

Frugality property: Probability theory

Theorem: Frugality property

Consider a DAG model $(G, P(X))$ and its corresponding anchored DAG model $(G_{an}, P(X, X'))$, where X is a vector of latent variables and $X' = F(X)$ is any function of latent variables in which $X'_j = F_j(X_j)$ for all $j \in V$. Suppose that $P(X, X')$ is faithful to G_{an}. Then, for any $P(X, X')$ and $G' \in \mathcal{G}_{fr}(P(X'))$,

- the skeleton of G' is a supergraph of the skeleton of G.
- $|G| = |G'|$ if and only if $\mathcal{M}(G) = \mathcal{M}(G')$.

- In short, the true graph is *always sparser* than the corresponding corrupted graph in terms of d-connections.

Anchored Gaussian DAG model

- Anchored Gaussian DAG model: For $j \in \{1, 2, ..., p\}$,

$$Z_j = f_j(X_j), \quad \text{where} \quad X_j \sim N\left(0, \sigma_j^2\right).$$

- To establish its identifiability, it is assumed for each observed variable to be
  - a linear function of the corresponding latent variable and a measurement error with known variance (Zhang et al., 2017)

$$Z_j = X_j + E_j, \quad \text{where} \quad E_j \sim N\left(0, s_j^2\right).$$

  - any function of the latent variable with known moment relationships between the latent variables and the observed variables (Saeed et al., 2020).

$$Z_j = f_j(X_j), \quad \text{where } f_j \text{ is known possibly stochastic function.}$$

Post-randomized additive measurement error model

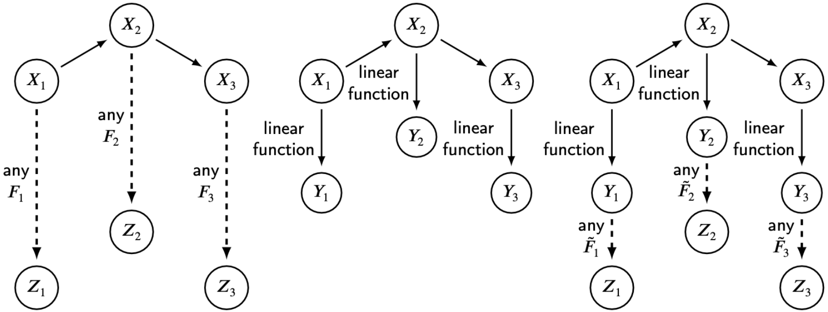


Figure 1: Three types of anchored models: an anchored DAG model (left), an additive measurement error model (middle), and a post-randomized additive measurement error model (right).

- Post-randomized additive measurement error model: For $j \in \{1, 2, ..., p\}$,

  $Z_j = f_j(X_j + E_j)$, where $E_j \sim N\left(0, s_j^2\right)$ and f_j is known possibly stochastic function.
- We allow the variance of E_j to be *unknown*.

Examples of post-randomized additive measurement error model

- Gaussian additive noise models: For $j \in \{1, 2, ..., p\}$,

$$Z_j = f_j(X_j + E_j) = X_j + E_j + \tilde{E}_j, \quad \text{where } E_j \sim N(0, s_j^2) \text{ and } \tilde{E}_j \sim N(0, \eta_j^2).$$

  - η_j^2 should be known, whereas we don't need the information of s_j^2.

- Dropout models: For $j \in \{1, 2, ..., p\}$,

$$Z_j = f_j(X_j + E_j) = \begin{cases} X_j + E_j & \text{with probability } \gamma_j, \\ 0 & \text{with probability } 1 - \gamma_j. \end{cases}, \quad \text{where} \quad E_j \sim N(0, s_j^2).$$

  - $\mathbb{E}(X_j) = \mathbb{E}(Z_j)/\gamma_j$, $\mathbb{E}(X_j^2) = \mathbb{E}(Z_j^2)/\gamma_j - \eta_j^2$, and $\mathbb{E}(X_j X_k) = \mathbb{E}(Z_j Z_k)/\gamma_j \gamma_k$.
  - γ_j should be known, but s_j^2 remains unknown.

Main result

Identifiability

the post-randomized additive measurement error models with unknown measurement error variance are identifiable up to MEC if

- the true graph meets the faithfulness assumption for its probability distribution,
- it is known how the covariance matrix of the latent variables with additive measurement error Cov(Y) is derived from the observed distribution, such that Cov$(Y) = \mathcal{T}($Cov$(Z))$, and
- the frugality assumption is satisfied.

Anchored PC algorithm

PC algorithm for learning anchored Gaussian DAG models

- **Input:** Covariance matrix for observed variables $\mathrm{Cov}(Z)$, and transformation $\mathcal{T}$ such that $\mathrm{Cov}(X) + \eta I_p = \mathcal{T}(\mathrm{Cov}(Z))$
- **Output:** Complete Partial DAG (CPDAG), $\widehat{G}_{cp}$

Step 1: Compute the covariance matrix for latent variables with measurement errors $\mathrm{Cov}(Y) = \mathcal{T}(\mathrm{Cov}(Z))$

Step 2: Set EtaSet $\subset (\Lambda_{\min}(\mathrm{Cov}(Y)), 0]$ for measurement error variances

For $\eta' \in$ EtaSet

Step 3-1: Calculate the partial correlations of X from $\Sigma_{\eta'} = \mathrm{Cov}(Y) - \eta' I_p$

Step 3-2: Find the C.I. relations

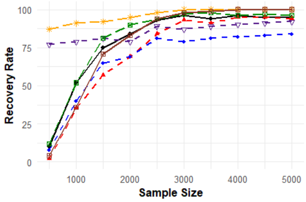
Step 3-3: Estimate a CPDAG, $\widehat{G}_{\eta'}$, using the *PC algorithm* based on the C.I. relations

Determine the most frugal $\widehat{G}_{\hat{\eta}}$ as $\widehat{G}_{cp}$ where $\hat{\eta} = \arg\min_{\eta'} |\widehat{G}_{\eta'}|$

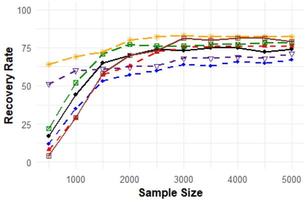
Numerical experiments

- 100 realizations for Gaussian additive measurement error models were randomly generated.
- True graphs were generated at random while respecting the pre-determined maximum indegree $d_{in} \in \{1, 2, 3\}$.
- The set of non-zero parameters $\beta_{j,k} \in \mathbb{R}$ was uniformly generated within the range $\beta_{j,k} \in (-0.8, -0.2) \cup (0.2, 0.8)$.
- Noise variances σ_j^2 were randomly chosen within the range $[0.5, 2]$, and we set the measurement error variance η^2 to 0.25.
- We compared **Anchored-SP and Frugal-PC algorithms** to state-of-the-art algorithms: ACI, PC, and MMHC.

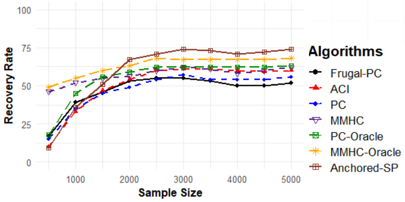
Numerical experiments



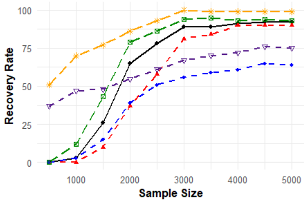
(a) $p = 5, d_{in} = 1$



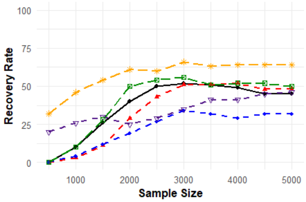
(b) $p = 5, d_{in} = 2$



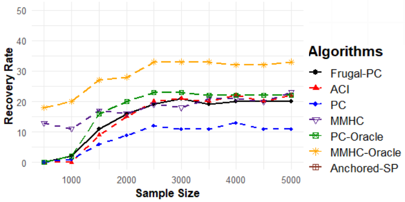
(c) $p = 5, d_{in} = 3$



(d) $p = 10, d_{in} = 1$



(e) $p = 10, d_{in} = 2$



(f) $p = 10, d_{in} = 3$

Summary and future works

- **Considered model**: Anchored Gaussian DAG models with post-randomized additive measurement error with unknown variance.
- **Contributions:**
  - Propose the frugality assumption aiding in true graph structure identification under unknown measurement error variance.
  - Develop a constraint-based structure learning algorithm, validated for consistency and effectiveness through extensive numerical experiments.
- **Future Works:**
  - Relax the Gaussianity assumption.
  - Recover a DAG rather than MEC.

Reference

- Anandkumar, A., Hsu, D., Javanmard, A., & Kakade, S. (2013). Learning linear Bayesian networks with latent variables. *International Conference on Machine Learning* (pp. 249-257). PMLR.
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., ... & Regev, A. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell, 167(7)*, 1853-1866.
- Halpern, Y., Horng, S., & Sontag, D. (2015). Anchored discrete factor analysis. *arXiv preprint arXiv:1511.03299*.
- Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis, 88(2)*, 365-411.
- Park, G. (2020). Identifiability of additive noise models using conditional variances. *The Journal of Machine Learning Research, 21*(1), 2896-2929.
- Saeed, B., Belyaeva, A., Wang, Y., & Uhler, C. (2020). Anchored causal inference in the presence of measurement error. *Conference on uncertainty in artificial intelligence* (pp. 619-628). PMLR.
- Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., & Glymour, C. (2017). Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*.

Appendix

Sparest permutation algorithm

- Consider all possible DAGs that satisfy the Markov condition and choose the one with the fewest edges.
- This approach becomes *impractical* as the number of potential DAGs increases super-exponentially with the number of nodes.
- To address this issue, computationally feasible algorithms, such as the PC algorithm, must be employed.
- However, adopting such algorithms requires certain trade-offs, including additional conditions for their application.