Castelletti and Consonni (2020): Bayesian inference of causal effects from observational data in Gaussian graphical models

Hyoin An, Yuxuan Xin, and Haozhen Yu

The Ohio State University

April 23, 2021

Introduction

- The Bayesian approach using Directed Acyclic Graphs (DAGs) is one of the most popular approaches in causal studies.
- The true underlying DAG is not identifiable from observational data alone.
- Instead, we can aim to estimate a Markov equivalence class of graphs, represented by an **Essential graph** (EG; Anderson et al., 1997)
- Castelletti and Consonni (2020) propose a Bayesian method which can (1) learn the structure of the equivalence class of graphs and (2) infer causality in a way the uncertainty could be estimated.

Bayesian networks in Causal inference

- Let $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ be a DAG, where $\mathcal{V} = \{1, ..., q\}$ is a set of vertices and $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ is a set of directed edges.
- $pa(\nu) = pa_{\mathcal{D}}(\nu)$: a parent set of ν in $\mathcal{D}$.
- $fa(\nu) = fa_{\mathcal{D}}(\nu) = \nu \cup pa(\nu)$: the family of ν in $\mathcal{D}$.
- Consider $\boldsymbol{X} = (X_1, ..., X_q)$ that respect $\mathcal{D}$. Then, we let $f(\cdot)$ denote the joint probability function of $(X_1, ..., X_q)$
- Assume $f(\cdot)$ has the Markov property of the DAG, then the observational (or preintervention) distribution is:

$$f(x_1, ..., x_q) = \prod_{j=1}^{q} f(x_j | \boldsymbol{x}_{pa(j)}).$$

Bayesian networks in Causal inference

The postintervention distribution of $(X_1, ..., X_q)$ is:

$$f(x_1, ..., x_q \mid \mathrm{do}(X_i = \widetilde{x}_i)) = \begin{cases} \prod_{j=1, j\neq i}^{q} f(x_j | \boldsymbol{x}_{pa(j)})|_{x_i = \widetilde{x}_i} & \text{if } x_i = \widetilde{x}_i \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{do}(X_i = \widetilde{x}_i)$ is an intervention that fixes X_i to $\widetilde{x}_i$. (Pearl, 2000)

The marginal postintervention distribution of $Y = X_1$ is:

$$f(y \mid \mathrm{do}(X_i = \widetilde{x}_i)) = \int f(y \mid \widetilde{x}_i, \boldsymbol{x}_{pa(i)}) f(\boldsymbol{x}_{pa(i)}) d\boldsymbol{x}_{pa(i)}.$$

Then, the **causal effect** of $\mathrm{do}(X_i = \widetilde{x}_i)$ is denoted by γ_i, and defined as:

$$\gamma_i = \frac{\partial}{\partial x} \mathbb{E}(Y \mid \mathrm{do}(X_i = \widetilde{x}_i))|_{x_i = \widetilde{x}_i}.$$

Bayesian networks in Causal inference

Castelletti and Consonni (2020) restrict their attention to the case of **Gaussian graphical models**:

$$\boldsymbol{X}|\Sigma \sim \mathcal{N}_q(\mathbf{0}, \Sigma),$$

where Σ is a covariance matrix Markov with respect to $\mathcal{D}$ and $\Omega = \Sigma^{-1}$ is a precision matrix.

Then, the causal effect of $\text{do}(X_i = \widetilde{x}_i)$ becomes:

$$\gamma_i = \left[[\Sigma_{Y,fa(i)}](\Sigma_{fa(i),fa(i)})^{-1}\right]_1,$$

where subscript 1 corresponds to the first entry of the vector.

Bayesian Inference of DAG Model Parameters

Cholesky Parameterization and DAG-Wishart Priors

- Assume $x|\Sigma \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- Re-parameterize it through the structural equation model

$$L^\top x = \epsilon$$

where L is lower triangular matrix of coefficients and

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$$

where $\mathbf{D} = diag(\sigma^2)$ and $\Sigma = L^{-\top} D L^{-1} = \Omega^{-1}$.

Bayesian Inference of DAG Model Parameters

Cholesky Parameterization and DAG-Wishart Priors

- Convert precision matrix Ω by node-parameters $\theta_j = (D_{jj}, L_{\prec j]})$, j=1,...,q.

$$D_{jj} = \Sigma_{jj|pa(j)}, \;\; L_{\prec j]} = -\Sigma_{\prec j \succ} \Sigma_{\prec j]}$$

- General DAG-Wishart Distribution on (D, L) has density $(a_j = n + 2q - 2j + 2)$:

$$p(D, L) \propto exp[-\frac{1}{2} tr(LD^{-1}L^{\top})U] \prod_{j=1}^{q} D_{jj}^{-a_j/2}$$

$$D_{jj} \sim I - Ga(\frac{a_j}{2} - \frac{|pa(j)|}{2} - 1, \frac{1}{2} U_{jj|\prec j \succ});$$

$$L_{\prec j]} | D_{jj} \sim \mathcal{N}_{pa(j)}(-U_{\prec j \succ}^{-1} U_{\prec j]}, D_{jj} U_{\prec j \succ}^{-1}).$$

Bayesian Inference of DAG Model Parameters

- Geiger and Heckerman(2002) propose a method to construct parameter priors for comparison of DAG-models.
- **Prior modularity:** Given two distinct DAG models with same parents for vertex j, prior for node parameter θ_j must be the same under both models.

$$p(\theta_j|\mathcal{D}_h) = p(\theta_j|\mathcal{D}_k) if\, pa_{\mathcal{D}_h}(j) = pa_{\mathcal{D}_k}(j)$$

- **Global parameter independence:** For every DAG model $\mathcal{D}$, the parameters θ_j should be a priori independent.

$$p(\theta|\mathcal{D}) = \prod_{j=1}^{q} p(\theta_j|\mathcal{D})$$

- All parameter priors are completely determined by a unique prior on the parameters of any of the equivalent complete DAGs.

Bayesian Inference of DAG Model Parameters

Objective Bayes Analysis

- Consider a random sample $x_1, ..., x_n$, where $x_i = (x_{i,1}, ..., x_{i,q})^\top$, and $x_i|\Omega \sim \mathcal{N}_q(0, \Omega^{-1})$.
- Assume a non-informative prior $p^N(\Omega) \propto |\Omega|^{-1}$.
- The likelihood function for Ω is

$$f(X|\Omega) \propto |\Omega|^{n/2} exp[-\frac{1}{2} tr(\Omega X^\top X)]$$

- Posterior Distribution is

$$\Omega|X \sim \mathcal{W}_q(n + q - 1, X^\top X),$$

which induces a complete DAG Wishart posterior on (D, L).

Bayesian Inference of DAG Model Parameters

Make inference on Σ and derive inference on γ_i.

- **Cholesky Parameterization of DAG**
  Convert Σ matrix to parameters D and L by structural equation model. Construct reasonable priors for (D, L), based on $G\&H(2002)$.
- **Bayesian Inference Procedure**
  1. Generate posterior draws of Cholesky parameters (D, L) for each target node.
  2. Obtain posterior draws of Σ.
  3. Obtain posterior draws of causal effect parameter γ_i for any **given DAG model**.

Bayesian Inference on Causal Effects

- We can only calculate casual effects based on a specific DAG model.
- When only observational data is available, however, the true DAG $\mathcal{D}$ cannot be distinguished from its equivalent class, and they can be represented by their essential graph.
- Therefore, we will estimate the causal effect of the essential graph (EG) which will be denoted as $\mathcal{G}$.

Bayesian Inference on Causal Effects

The uncertainty concerns the structure of data-generating mechanism

- Castelletti et al.(2018) develop the Objective Bayes (OB) methodology for addressing the uncertainty concerns the EG distribution.
- Assign a prior $p(\mathcal{G})$ to $\mathcal{G}$ by imposing a Bernoulli-beta distribution independently to each element of the adjacency matrix of the skeleton of $\mathcal{G}$, $\mathcal{G}^u$,

$$\mathcal{G}^u_{(j)} \sim Ber(\pi), \quad j = 1, ..., \frac{q(q-1)}{2},$$

$$\pi \sim Beta(a, b).$$

- The posterior distribution of $\mathcal{G}$ given the data is

$$p(\mathcal{G} \mid \boldsymbol{X}) = \frac{m_{\mathcal{G}}(\boldsymbol{X})p(\mathcal{G})}{\sum_{\mathcal{G} \in \mathcal{S}_q} m_{\mathcal{G}}(\boldsymbol{X})p(\mathcal{G})}.$$

Bayesian Inference on Causal Effects

The uncertainty refers to the size of the true causal effect

- Given an equivalent class of DAGs represented by their EG $\mathcal{G}$,let $\{\gamma_l(\mathcal{G}); l = 1, ..., L_{\mathcal{G}}\}$ be the collection of $L_{\mathcal{G}}$ distinct causal effects of X_i on Y.
- An overall measure of γ conditional on $\mathcal{G}$ is

$$\gamma_{\mathrm{avg}}(\mathcal{G}) = \frac{1}{L_{\mathcal{G}}} \sum_{l=1}^{L_{\mathcal{G}}} \gamma_l(\mathcal{G}).$$

- An estimate of $\gamma_{\mathrm{avg}}(\mathcal{G})$ is the corresponding posterior conditional expectation

$$\bar{\gamma}_{\mathsf{avg}}\left(\mathcal{G}; \boldsymbol{X}\right) = \frac{1}{L_{\mathcal{G}}} \sum_{l=1}^{L_{\mathcal{G}}} \mathbb{E}\left\{\gamma_l(\mathcal{G}) \mid \boldsymbol{X}, \mathcal{G}\right\}.$$

Bayesian Inference on Causal Effects

Two strategies to deal with the unknown $\mathcal{G}$

- OB-MA: Combining the idea of Bayesian Model Averaging (Hoeting et al.1999) and the posterior distribution of $\mathcal{G}$,

$$\bar{\gamma}_{OB-MA}(\boldsymbol{X}) = \sum_{\mathcal{G}_k} \mathbb{E}\left\{\gamma_{\text{avg}}\left(\mathcal{G}_k\right) \mid \boldsymbol{X}, \mathcal{G}_k\right\} p\left(\mathcal{G}_k \mid \boldsymbol{X}\right).$$

- OB-MED: Construct the projected median probability graph, denoting as $\mathcal{G}^*$. This leads to a set of distinct causal effects $\{\gamma_l(\mathcal{G}^*); l = 1, ..., L_{\mathcal{G}^*}\}$.

$$\bar{\gamma}_{OB-MED}(\boldsymbol{X}) = \bar{\gamma}_{\text{avg}}\left(\mathcal{G}^*; \boldsymbol{X}\right).$$

A Simulation Study: Set up

For a simulation study, we generated 40 DAGs for each scenario:

- The number of nodes $q = 5, 10, 20$
- The sample sizes $n = 50, 100, 200$
- The probability of edge inclusion $p_{edge} = 0.1$

and under each $\mathcal{D}$, n iid observations are generated using:

$$X_{i,j} = \mu_j + \sum_{k \in \mathrm{pa}_D(j)} \beta_{k,j} X_{i,k} + \varepsilon_{i,j},$$

where $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1)$, $\mu_j = 0$, $\beta_{k,j} \overset{iid}{\sim} \mathcal{U}(1,2)$, and $i, j = 1, ..., q$.

Additionally,

- A target node $i \in \{2, ..., q\}$ is randomly selected
- 40 true (average) causal effects are calculated.

A Simulation Study: Set up

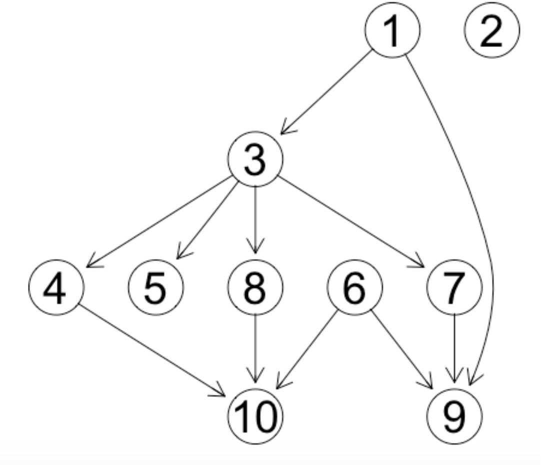

Figure: An Example of a DAG

A Simulation Study: Results

Structural Hamming Distance (SHD): a measure evaluating the distance between estimated and true graph

	n = 50	n = 100	n = 200
OBMED	6.25 (2.82)	4.70 (2.54)	4.53 (2.32)
PC0.1	6.15 (2.65)	5.68 (1.46)	5.52 (1.75)
PC0.05	5.68 (2.96)	5.03 (1.44)	5.10 (1.88)
PC0.01	5.50 (1.88)	4.55 (1.74)	4.56 (1.97)

Table: Mean and Standard deviations of SHD ($q = 10$)

- PC algorithm is suggested by Maathuis et al. (2009) and the number in the name after PC is significance level α.

Absolute-value distance: the difference between true causal effect and estimated causal effect at the targeted node: $d_M(i) = |\bar{\gamma}_M^{(i)} - \bar{\gamma}_{true}^{(i)}|$ for generic method M.

A Simulation Study: Results

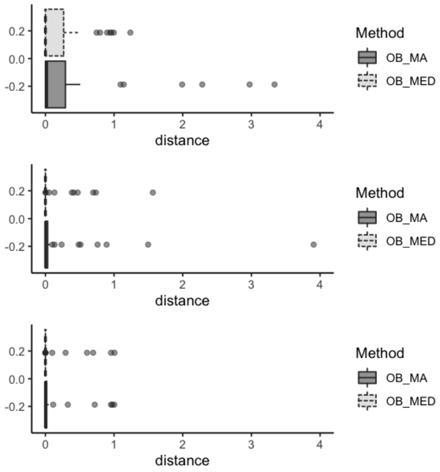

Figure: Absolute-value distance for $q = 10$, $n = 50, 100, 200$

Some Findings and Conclusion

- The proposed method to estimate the graph and causal effects considers the model uncertainty.
- The computation runs pretty slowly compared to the PC algorithm due to the MCMC sampling procedure.
- Their methods seem to estimate
  $(+)$ the existence of the causal relationship
  $(-)$ the magnitude of the causal effect, especially with the small n.
- Estimation of γ_i depends on the inverse calculation of the matrix.
- In our simulation study, the inverse matrix was unstable and that might explain the few cases where the distance is pretty large in the plots.

References

Castelletti, F., & Consonni, G. (2020)
Bayesian inference of causal effects from observational data in Gaussian graphical models
Biometrics, 277(1), 136-149.

Maathuis, M.H., Kalisch, M. & Bühlmann, P. (2009)
Estimating high-dimensional intervention effects from observational data
The Annals of Statistics, 37(6A), 3133–3164.