Bayesian causal inference in probit graphical models

김성민

서울대학교
통계학과, 베이즈통계 연구실

2021. 12. 10

CONTENTS

1. Material
2. Directed Acyclic Graph
3. DAG-Probit Models
4. Causal effect
5. Bayesian Inference
6. Result
7. Discussion

CONTENTS

1. Material
2. Directed Acyclic Graph
3. DAG-Probit Models
4. Causal effect
5. Bayesian Inference
6. Result
7. Discussion

Material

- Castelletti, F., & Consonni, G. (2021). Bayesian causal inference in probit graphical models. *Bayesian Analysis*, 1(1), 1-25.Castelletti and Consonni (2021)

CONTENTS

1 Material

2 Directed Acyclic Graph

3 DAG-Probit Models

4 Causal effect

5 Bayesian Inference

6 Result

7 Discussion

DAG

- Let $\mathcal{D} = (V, E)$ be a DAG.
  - V : Set of nodes
  - E : Set of edges $(E \subseteq V \times V)$
- $\mathcal{D}$ does not contain cycles.
- $(u, v) \equiv u \to v$
  - u is a parent of a v
  - $pa(v)$: The parent set of v in $\mathcal{D}$
  - $ch(u)$: The children set of u in $\mathcal{D}$

Gaussian DAG-models

- Consider a collection of random variables $(X_1, \ldots, X_q)$
- Assume that their joint pdf is Markov w.r.t. D.
  - $X_i \perp\!\!\!\perp X_{nd(i)\setminus pa(i)} | X_{pa(i)}$

$$f(x_1, \ldots, x_q) = \Pi_{j=1}^{q} f(x_j | x_{pa(j)})$$

- If the joint distribution is Gaussian with zero mean, we write

$$X_1, \ldots, X_q | \Omega \propto N_q(0, \Omega^{-1}),\ \Omega \in \mathcal{P}_{\mathcal{D}}$$

- The $\Omega = \Sigma^{-1}$ is precision matrix and the $\mathcal{P}_{\mathcal{D}}$ is the space of symmetric positive definite precision matrices which follows the Markov properties of the DAG $\mathcal{D}$.

Gaussian DAG-models

- Without loss of generality that $\mathcal{D}$ is in a parent ordering.
    - $i \rightarrow j$ implies $i > j$
- $\Omega_{ij} = 0$ iff $X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}}$
- We decompose the Ω by using *modified Cholesky decomposition*
    - $\Omega = LD^{-1}L^T$.
    - L is *lower-triangular* matrix. ($L_{ii} = 1 \quad \& \quad L_{ij} \neq 0$ iff $i \rightarrow j$)
    - $D = diag(\sigma^2)$

Gaussian DAG-models

- The joint PDF can be written as

$$f(x_1, \ldots, x_q | D, L) = \Pi_{j=1}^{q} dN(x_j | -L_{\prec j]}^T x_{pa(j)}, \sigma_j^2)$$

- $\prec j \succ = pa(j)$ and $\prec j] = pa(j) \times j$.
- $L_{\prec j]} = -\Sigma_{\prec j \succ} \Sigma_{\prec j]}$ and $\sigma_j^2 = \Sigma_{jj|pa(j)}$

CONTENTS

1. Material
2. Directed Acyclic Graph
3. DAG-Probit Models
4. Causal effect
5. Bayesian Inference
6. Result
7. Discussion

DAG-Probit Models

- This paper consider the binary response which is affected by the continuous variable.
- By introducing latent variable X_1, we can deal with the binary response Y.

$$Y = \begin{cases} 1, & \text{if } X_1 \in [\theta_0, +\infty), \\ 0, & \text{if } X_1 \in (-\infty, \theta_0]. \end{cases}$$

- Finally, if there are n independent samples $(y_i, x_{i,2}, \ldots, x_{i,q})$, the likelihood is as follows:

$$f(y, X|D, L, \theta_0) = \Pi_{i=1}^{n} f(x_{i,1}, \ldots, x_{i,q}|D, L) \cdot 1(\theta_{y_i-1} < x_{i,1} \leq \theta_{y_i})$$

$$\text{where} \quad \theta_{-1} = -\infty, \; \theta_1 = +\infty$$

CONTENTS

1. Material
2. Directed Acyclic Graph
3. DAG-Probit Models
4. Causal effect
5. Bayesian Inference
6. Result
7. Discussion

Causal effect

- Is Y the causal of the effect X?
- $f(X|Y=y)$ is not zero for CASE A and B.
- Does B indeed have the causal relation?

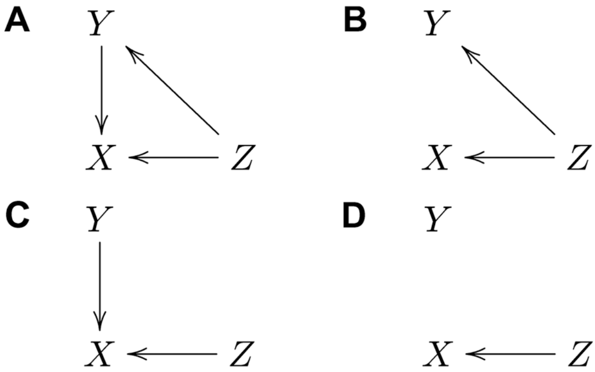


Figure: Causal effect (Chicharro and Ledberg (2012))

Causal effect

- A deterministic intervention on variable X_s, is denoted by $do(X_s = \tilde{x})$ (X_s is fixed as $\tilde{x}$)
- The *post-intervention* density of X can be factorized as follows:

$$f(x_1, \ldots, x_q | do(X_s = \tilde{x})) = \begin{cases} \prod_{j=1, j \neq s}^{q} f(x_j | \mathbf{x_{pa(j)}})|_{x_s = \tilde{x}}, & \text{if } x_s = \tilde{x}, \\ 0, & \textit{otherwise} \end{cases}$$

Post-intervention density of X_1

- Under the assumption $(X_1, \ldots, X_q)|\Sigma \sim N_q(0, \Sigma)$, the *post-intervention* of X_1 can be written as,

$$f(x_1|do(X_s = \tilde{x})), \Sigma) = \int f(x_1|\tilde{x}, \mathbf{x_{pa(s)}}, \Sigma) \cdot f(\mathbf{x_{pa(s)}}|\Sigma) d\mathbf{x_{pa(s)}}$$

$$= \int dN(x_1|\gamma_s \tilde{x} + \gamma^T \mathbf{x_{pa(s)}}, \delta_1^2) \cdot dN(\mathbf{x_{pa(s)}}|0, \Sigma_{pa(s),pa(s)}) d\mathbf{x_{pa(s)}}$$

Post-intervention density of X_1

- By some additional calculation on the upper equation, we can write the pdf of the *post-intervention* of X_1 as follows:

$$f(x_1|do(X_s=\tilde{x})),\Sigma) = dN(x_1|\gamma_s\tilde{x}, \frac{\delta_1^2}{1-(\gamma^T\mathbf{T}^{-1}\gamma/\delta_1^2)})$$

$$\delta_1^2 = \Sigma_{1|fa(i)},\ \ fa(i) = pa(i)\cup i$$

$$(\gamma_s,\gamma^T)^T = \Sigma_{1,fa(i)}(\Sigma_{fa(i),fa(i)})^{-1}$$

$$\mathbf{T} = (\Sigma_{pa(i),pa(i)})^{-1} + \frac{1}{\delta_1^2}\gamma\gamma^T$$

- And we can obtain the $\mathbb{E}(Y|do(X_s=\tilde{x})),\Sigma,\theta_0)$.

$$\begin{aligned}\mathbb{E}(Y|do(X_s=\tilde{x})),\Sigma,\theta_0) &= P(Y=1|do(X_s=\tilde{x})),\Sigma,\theta_0)\\ &= P(X_1\geq\theta_0|do(X_s=\tilde{x})),\Sigma,\theta_0)\\ &= 1-\Phi\left(\frac{\theta_0-\gamma_s\tilde{x}}{\sqrt{\delta_1^2/(1-(\gamma^T\mathbf{T}^{-1}\gamma/\delta_1^2))}}\right)\end{aligned}$$

CONTENTS

1. Material
2. Directed Acyclic Graph
3. DAG-Probit Models
4. Causal effect
5. **Bayesian Inference**
6. Result
7. Discussion

Prior on the Cholesky parameters

- The standard conjugate prior of precision matrix (Ω) is the Wishart distribution($\mathcal{W}_q(a, U)$, where U is symmetric positive definite). We can induced the prior on the Cholesky parameter $((\sigma_j^2, L_{\prec j]}))$ for each node.

$$
\begin{aligned}
\sigma_j^2 &\sim InvGamma(\frac{a_j}{2} - \frac{|pa(j)|}{2} - 1, \frac{1}{2} U_{jj|pa(j)}) \\
L_{\prec j]} | \sigma_j^2 &\sim N_{|pa(j)|}(-U_{pa(j)}^{-1} U_{\prec j]}, \sigma_j^2 U_{pa(j)}^{-1})
\end{aligned}
$$

where $|A|$ is the number of elements of A and $a_j = a + q - 2j + 3$.

Prior on the Cholesky parameters

- Further the standard choice for the hyperparameter U is gI_q, then the prior can written as follows:

$$\sigma_j^2 \sim InvGamma(\frac{a_j}{2} - \frac{|pa(j)|}{2} - 1, \frac{1}{2}g)$$
$$L_{\prec j]}|\sigma_j^2 \sim N_{|pa(j)|}(0, \frac{1}{g}\sigma_j^2 I_{pa(j)})$$

- Then $p(\Omega|\mathcal{D}) = p(D, L|\mathcal{D}) = \Pi_{j=1}^{q} p(\sigma_j^2, L_{\prec j]}|\mathcal{D})$.

Prior on DAG space

- We set the probability of the existance of each edge is π. It means that the prior on given DAG $\mathcal{D}$ is as follows:

$$p(\mathcal{D}) = \pi^{|\mathcal{D}|}(1-\pi)^{\frac{q(q-1)}{2}-|\mathcal{D}|}$$

where $|\mathcal{D}|$ denotes the number of edges in the DAG $\mathcal{D}$.

Posterior distribution

- By the setting the likelihood of X and prior for $(\Omega, \theta_0, \mathcal{D})$, we can obtain the posterior of the model.

$$p(D, L, \mathcal{D}, \theta_0, X_1 | y, X_{-1}) \propto f(y, X | D, L, \mathcal{D}, \theta_0) p(D, L | \mathcal{D}) p(\mathcal{D})$$

CONTENTS

1. Material
2. Directed Acyclic Graph
3. DAG-Probit Models
4. Causal effect
5. Bayesian Inference
6. **Result**
7. Discussion

Probability of edge inclusion

- By the MCMC on the posterior distribution, the output is the collection of $\{\mathcal{D}^{(t)}\}_{t=1}^{T}$.
- We can compute the posterior probabilities of edge inclusion, i.e.

$$\hat{p}_{u\to v}(y, X_2, \ldots, X_q) \equiv \hat{p}_{u\to v} = \frac{1}{T}\sum_{t=1}^{T} 1_{u\to v}\{\mathcal{D}^{(t)}\}$$

.

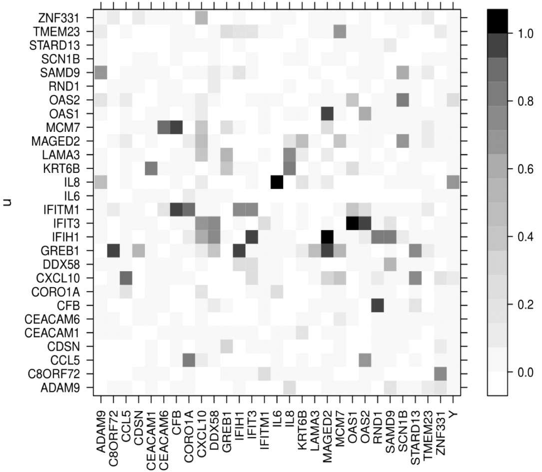
Bayesian Model Averaging

- The causal effect: $\beta_s^{(t)}(\tilde{x}) = \mathbb{E}(Y|do(X_s = \tilde{x})), \Sigma^{\mathcal{D}^{(t)}}, \theta_0^{(t)})$
- An overall summary of the causal effect of $do(X_s = \tilde{x})$ on Y can be computed as,

$$\hat{\beta}_s^{\text{BMA}}(\tilde{x}) = \frac{1}{T}\sum_{t=1}^{T} \beta_s^{(t)}(\tilde{x}).$$

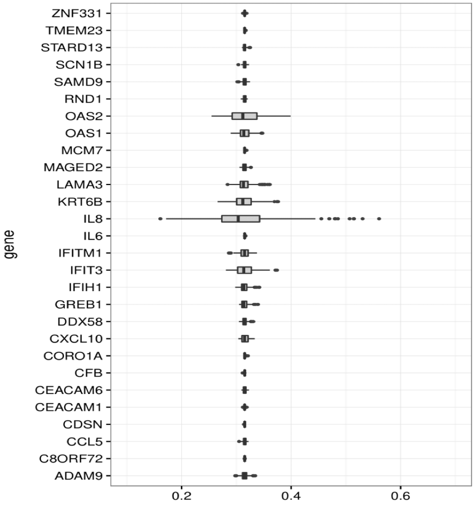
Analysis of gene expressions

- Heat map with estimated marginal posterior probabilities of edge inclusion for each edge $u \to v$



Analysis of gene expressions

- Box-plots of BMA estimate of causal effect.



CONTENTS

1 Material

2 Directed Acyclic Graph

3 DAG-Probit Models

4 Causal effect

5 Bayesian Inference

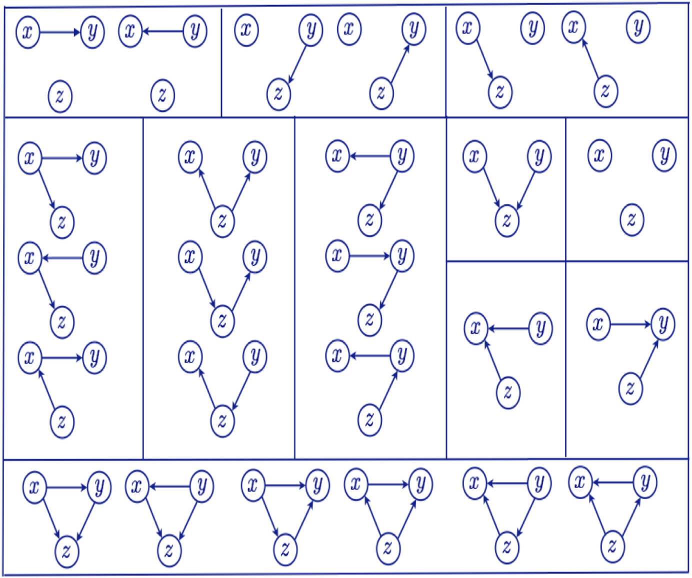
6 Result

7 Discussion

Discussion

- In the Gaussian setting, DAGs encoding the same conditional independencies are not distinguishable using observational data.
- Markov equivalent DAGs : DAGs encoding the same conditional independencies

Markov equivalence classes



References I

Castelletti, F. and Consonni, G. (2021). Bayesian causal inference in probit graphical models, *Bayesian Analysis* **1**(1): 1–25.

Chicharro, D. and Ledberg, A. (2012). When two become one: the limits of causality analysis of brain dynamics, *PLoS One* **7**(3): e32466.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery., *Int. J. Data Sci. Anal.* **3**(2): 81–91.