Ordinal Causal Discovery: Supplementary Materials

Yang Ni[1] **Bani Mallick**[1]

[1]Department of Statistics, Texas A&M University, College Station, Texas, USA

1 PROOF OF THEOREM 1

We need the notion of *real analytic function*.

Definition (Real Analytic Function) *A real function is said to be analytic if it is infinitely differentiable and matches its Taylor series in a neighborhood of every point.*

Suppose $X \in \{1,\dots,S\}$ and $Y \in \{1,\dots,L\}$. Consider two competing causal models $p_{X\to Y}(X,Y|\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\gamma})$ and $p_{Y\to X}(Y,X|\boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\eta})$. We will show that these two causal models are in general not equivalent, i.e., $P_{X\to Y}(X=s,Y=\ell|\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\gamma}) \neq P_{Y\to X}(X=s,Y=\ell|\boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\eta})$ for some $s\in\{1,\dots,S\}$ and $\ell\in\{1,\dots,L\}$, where $S,L>2$. We prove it by contradiction. Suppose for any $s\in\{1,\dots,S\}$ and $\ell\in\{1,\dots,L\}$,

$$P_{X\to Y}(X=s,Y=\ell|\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\gamma}) = P_{Y\to X}(X=s,Y=\ell|\boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\eta}). \tag{1}$$

The left-hand side of (1) is given by

$$\begin{aligned}P_{X\to Y}(X=s,Y=\ell|\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\gamma}) &= P_{X\to Y}(Y=\ell|X=s,\boldsymbol{\beta},\boldsymbol{\gamma})P_{X\to Y}(X=s|\boldsymbol{\pi})\\ &= [P_{X\to Y}(Y\le\ell|X=s,\boldsymbol{\beta},\boldsymbol{\gamma}) - P_{X\to Y}(Y\le\ell-1|X=s,\boldsymbol{\beta},\boldsymbol{\gamma})]P_{X\to Y}(X=s|\boldsymbol{\pi})\\ &= [F(\gamma_\ell-\beta_s)-F(\gamma_{\ell-1}-\beta_s)]\pi_s,\end{aligned}$$

where $F(x)$ is the logistic link function $F(x)=\frac{e^x}{1+e^x}$ or the probit link function $F(x)=\Phi(x)$ with $\Phi(x)$ being the standard normal cumulative distribution function. Similarly, the right-hand side of (1) is given by

$$P_{Y\to X}(X=s,Y=\ell|\boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\eta}) = P_{Y\to X}(X=s|Y=\ell,\boldsymbol{\alpha},\boldsymbol{\eta})P_{Y\to X}(Y=\ell|\boldsymbol{\rho}) = [F(\eta_s-\alpha_\ell)-F(\eta_{s-1}-\alpha_\ell)]\rho_\ell.$$

Therefore, (1) leads to

$$[F(\gamma_\ell-\beta_s)-F(\gamma_{\ell-1}-\beta_s)]\pi_s = [F(\eta_s-\alpha_\ell)-F(\eta_{s-1}-\alpha_\ell)]\rho_\ell \tag{2}$$

Note that the right-hand side of (2) is a telescoping series in s. Hence, summing up both sides of (2) over s from 1 to S, we have

$$\begin{aligned}\sum_{s=1}^{S}[F(\gamma_\ell-\beta_s)-F(\gamma_{\ell-1}-\beta_s)]\pi_s &= [F(\eta_S-\alpha_\ell)-F(\eta_0-\alpha_\ell)]\rho_\ell\\ &= \rho_\ell.\end{aligned} \tag{3}$$

The last equation is because $\eta_S=\infty$ and $\eta_0=-\infty$ and hence $F(\eta_S-\alpha_\ell)=1$ and $F(\eta_0-\alpha_\ell)=0$. Plug (3) into (2),

$$[F(\gamma_\ell-\beta_s)-F(\gamma_{\ell-1}-\beta_s)]\pi_s = [F(\eta_s-\alpha_\ell)-F(\eta_{s-1}-\alpha_\ell)]\sum_{s=1}^{S}[F(\gamma_\ell-\beta_s)-F(\gamma_{\ell-1}-\beta_s)]\pi_s$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

and hence

$$\frac{[F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s} = F(\eta_s - \alpha_\ell) - F(\eta_{s-1} - \alpha_\ell) \tag{4}$$

Now, consider $s = 1$ in (4) and note $\eta_0 = -\infty$ and $\eta_1 = 0$,

$$\begin{aligned}\frac{[F(\gamma_\ell - \beta_1) - F(\gamma_{\ell-1} - \beta_1)]\pi_1}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s} &= F(\eta_1 - \alpha_\ell) - F(\eta_0 - \alpha_\ell) \\ &= F(-\alpha_\ell).\end{aligned}$$

Therefore,

$$\alpha_\ell = -F^{-1}\left\{\frac{[F(\gamma_\ell - \beta_1) - F(\gamma_{\ell-1} - \beta_1)]\pi_1}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}\right\} \tag{5}$$

Sequentially plug (5) into (4) for $s^* = 2, \ldots, S-1$ (note that one can at least plug in once for $s^* = 2$ because $S > 2$),

$$\eta_{s^*} = F^{-1}\left\{\frac{\sum_{s=1}^{s^*} [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}\right\} - F^{-1}\left\{\frac{[F(\gamma_\ell - \beta_1) - F(\gamma_{\ell-1} - \beta_1)]\pi_1}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}\right\}, \tag{6}$$

Because the left-hand side of (6) is independent of ℓ whereas the right-hand side of (6) depends on ℓ, we have,

$$\begin{aligned}&F^{-1}\left\{\frac{\sum_{s=1}^{s^*} [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}\right\} - F^{-1}\left\{\frac{[F(\gamma_\ell - \beta_1) - F(\gamma_{\ell-1} - \beta_1)]\pi_1}{\sum_{s=1}^S [F(\gamma_\ell - \beta_s) - F(\gamma_{\ell-1} - \beta_s)]\pi_s}\right\} \\ -&F^{-1}\left\{\frac{\sum_{s=1}^{s^*} [F(\gamma_{\ell^*} - \beta_s) - F(\gamma_{\ell^*-1} - \beta_s)]\pi_s}{\sum_{s=1}^S [F(\gamma_{\ell^*} - \beta_s) - F(\gamma_{\ell^*-1} - \beta_s)]\pi_s}\right\} + F^{-1}\left\{\frac{[F(\gamma_{\ell^*} - \beta_1) - F(\gamma_{\ell^*-1} - \beta_1)]\pi_1}{\sum_{s=1}^S [F(\gamma_{\ell^*} - \beta_s) - F(\gamma_{\ell^*-1} - \beta_s)]\pi_s}\right\} = 0,\end{aligned} \tag{7}$$

for any $\ell \neq \ell^*$. The link function F is an analytic function: (i) the logistic link $F(x) = \frac{e^x}{1+e^x}$ is a composition of elementary functions and hence is analytic; and (ii) the probit link $F(x) = \Phi(x)$ is analytic because the error function erf() is analytic. Since $F'(x)$ is nowhere zero in either case, $F^{-1}(x)$ is analytic. Since the left-hand side of (7) is a composition of F, F^{-1}, sums, products, and reciprocals of $\gamma_\ell, \gamma_{\ell^*}, \gamma_{\ell-1}, \gamma_{\ell^*-1}, \beta_1, \ldots, \beta_S, \pi_1, \ldots, \pi_S$, it is an analytic function [Krantz and Parks, 2002] and therefore its zero set must have Lebesgue measure zero [Mityagin, 2015]. In summary, we have proven that the two causal models are not equivalent for almost all $(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ with respect to the Lebesgue measure. Note that although our proof is for logistic or probit link, it is generalizable to other link functions as long as they are analytic functions and their derivatives are nowhere zero.

2 ADDITIONAL EXPERIMENT RESULTS

2.1 SYNTHETIC DATA

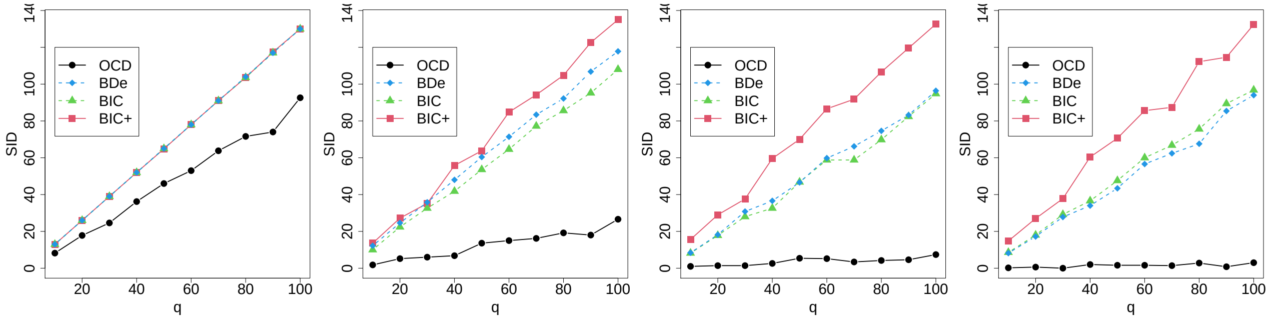
Number of Categories $L = 3$ We investigate scenarios where the number of categories $L = 3$. The data are generated as in the main text ($n = 500, q = 10$) except that the number of categories is now set to $L = 3$. Six scenarios with different levels of signal strength are considered, $\sigma = 0.25, 0.5, 0.75, 1, 1.25, 1.5$. We report the SHD of OCD, BIC+, BIC, and BDe in Table 1, which shows that OCD significantly outperforms competing methods.

Higher-Dimensional Synthetic Data As shown in Figure 1, for all tested signal strength $\sigma \in \{0.25, 0.5, 0.75, 1\}$ and number of nodes $q = 10, \ldots, 100$, SHD and SID of OCD are uniformly better than the competing methods and in general, OCD is quite stable as q increases when the signal strength is at least moderate $\sigma \geq 0.5$ whereas the competing methods quickly deteriorate with q regardless of the signal strength.

The CPU times of OCD in the synthetic data are shown in Figure 2, which appear to scale linearly in n and L, and quadratically in q.

	Signal Strength σ					
	0.25	0.5	0.75	1	1.25	1.5
OCD	5.2	1.6	1	0.8	0.2	0.2
BIC+	6.4	5	3.6	3.8	3.8	3.6
BIC	7	5.8	4.6	4	3.2	3.8
BDe	7	6.8	6.2	5.2	5	4.6

Table 1: Structural hamming distance between the true graph and the estimated graphs from OCD, BIC+, BIC, and BDe. The data are generated as in the main text with different levels of signal strength except that the number of categories is set to $L = 3$.



(a) SID in q ($\sigma = 0.25$)

(b) SID in q ($\sigma = 0.5$)

(c) SID in q ($\sigma = 0.75$)

(d) SID in q ($\sigma = 1$)

Figure 1: SID for OCD, BDe, BIC, and BIC+ as functions of q in the synthetic ordinal data with the sample size fixed at $n = 500$ and different signal strength $\sigma \in \{0.25, 0.5, 0.75, 1\}$.

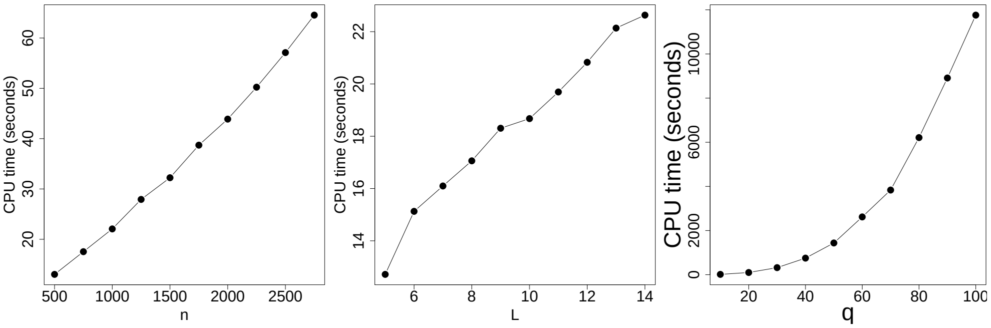
Synthetic Data with Denser Graphs We consider a scenario with $n = 500$ observations, $q = 50$ nodes, and $L = 5$ categories. We randomly generate graphs with 25, 50, and 100 edges (Figure 3). We report the Structural hamming distance (SHD) between the true graph and the estimated graphs from OCD, BIC+, BIC, and BDe in Table 2. We find very minor decrease in performance of OCD whereas the competing methods perform substantially worse and deteriorate much faster.

2.2 REAL DATA

Sensitivity to Small Added Noise to CEP Data Following the idea in Mooij et al. 2016, we test the sensitivity of the best performing causal discovery methods (OCD with $L = 15$, SLOPE, and bQCD) to small added noises. Specifically, we add independent centered Gaussian noise to X and Y with standard deviation $\tau \in \{10^{-8}, 10^{-7}, \ldots, 10^{-1}\}$. We repeat the simulation of noises 5 times under each noise level and the average ACC and AUC are shown in Figure 4. All methods have stable performance for $\tau = 10^{-8} - 10^{-5}$. The performance of SLOPE starts deteriorating at $\tau = 10^{-4}$ whereas OCD and bQCD are much more robust (significant drop in ACC at $\tau = 10^{-1}$). The robustness of OCD is expected because small added noise will not significantly affect data discretization.

	# of Edges		
	25	50	100
OCD	0	0	1
BIC+	13	-	-
BIC	25	48	92
BDe	23	40	75

Table 2: Structural hamming distance between the true graph and the estimated graphs from OCD, BIC+, BIC, and BDe. The true graphs are generated randomly with 25, 50, and 100 edges. BIC+ is not applicable for 50 and 100 edges as it takes 150GB of memory.

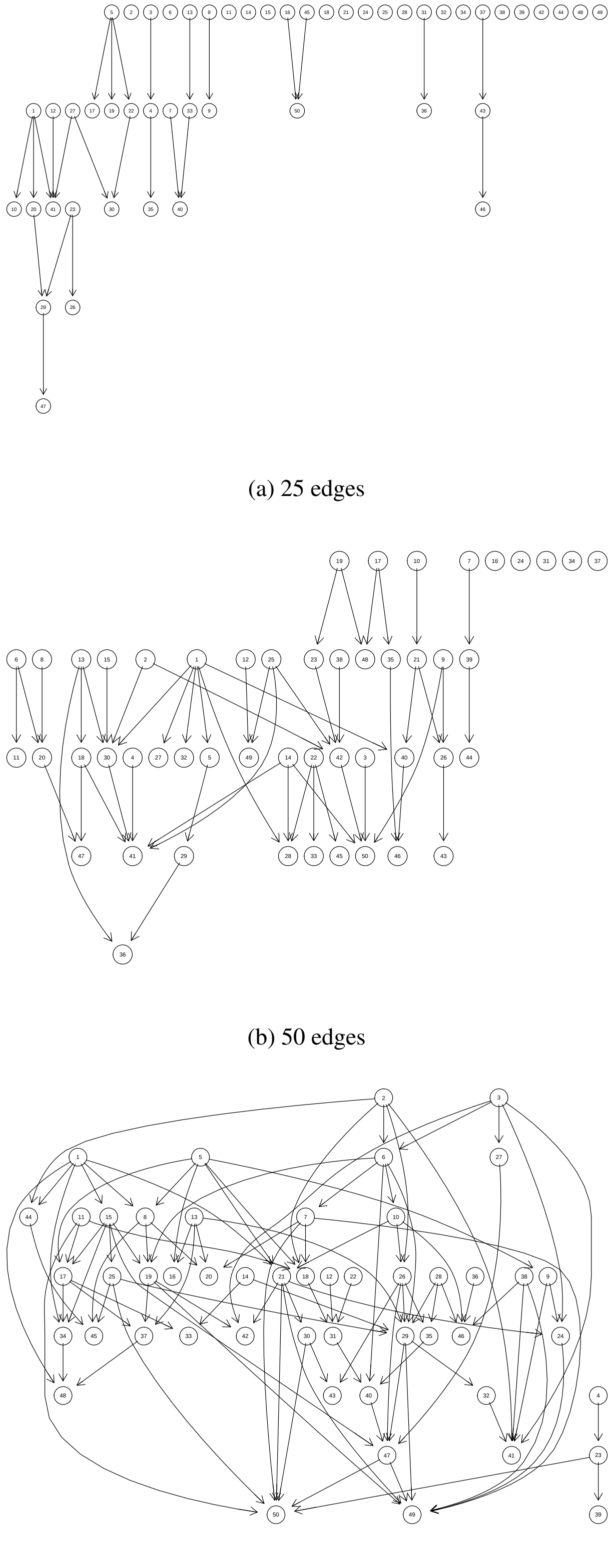


(a) CPU time in n

(b) CPU time in L

(c) CPU time in q

Figure 2: CPU times of OCD as functions of n, L, and q in the synthetic ordinal data.

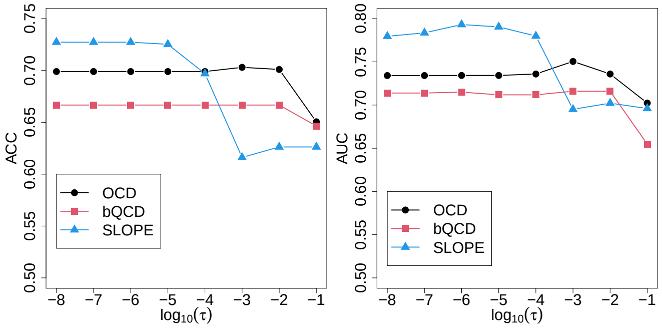


(a) 25 edges

(b) 50 edges

(c) 100 edges

Figure 3: Simulation truth of denser graphs.



(a) ACC

(b) AUC

Figure 4: Sensitivity to small added noise for the CEP data.

References

Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.