Bivariate Causal Discovery for Categorical Data via Classification with Optimal Label Permutation - Supplementary Material

Yang Ni
Department of Statistics
Texas A&M University
College Station, TX 77843
`yni@stat.tamu.edu`

1 Proof of Theorem 1

Consider $X \in \{1, \dots, S\}$ and $Y \in \{1, \dots, L\}$ where $S, L > 2$. Let Σ be the set of all permutations of size L and Π be the set of all permutations of size S. Let

$$\Theta = \{(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) \mid p_{X\to Y}(X, Y|\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) \equiv p_{Y\to X}(X, Y|\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \pi) \text{ for some } (\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \pi)\}$$

be the set of model parameters such that $p_{X\to Y}(X, Y|\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma)$ is not identifiable. Note that $\Theta \subset \mathbb{R}^{2S+L-3} \times \Sigma$. Let $\lambda(\cdot)$ be the $2S + L - 3$ dimensional Lebesgue measure and let $\mu(\cdot)$ be the counting measure. Define $m(\cdot)$ be the product measure of $\lambda(\cdot)$ and $\mu(\cdot)$, i.e., for any $A \subset \mathbb{R}^{2S+L-3}$ and $B \subset \Sigma$, $m(A, B) = \lambda(A) \times \mu(B)$. We will show that $m(\Theta) = 0$.

For any $\widetilde{\sigma} \in \Sigma$, let

$$\Theta_{\widetilde{\sigma}} = \{(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \widetilde{\sigma}) \mid p_{X\to Y}(X, Y|\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \widetilde{\sigma}) \equiv p_{Y\to X}(X, Y|\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \pi) \text{ for some } (\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \pi)\}.$$

Because $\Theta = \cup_{\widetilde{\sigma}\in\Sigma}\Theta_{\widetilde{\sigma}}$, we have

$$m(\Theta) \le \sum_{\widetilde{\sigma}\in\Sigma} m(\Theta_{\widetilde{\sigma}})$$

For any $\widetilde{\sigma} \in \Sigma$ and $\widetilde{\pi} \in \Pi$, let

$$\Theta_{\widetilde{\sigma},\widetilde{\pi}} = \{(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \widetilde{\sigma}) \mid p_{X\to Y}(X, Y|\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \widetilde{\sigma}) \equiv p_{Y\to X}(X, Y|\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \widetilde{\pi}) \text{ for some } (\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta})\}.$$

Since $\Theta_{\widetilde{\sigma}} = \cup_{\widetilde{\pi}\in\Pi}\Theta_{\widetilde{\sigma},\widetilde{\pi}}$, we have

$$m(\Theta_{\widetilde{\sigma}}) \le \sum_{\widetilde{\pi}\in\Pi} m(\Theta_{\widetilde{\sigma},\widetilde{\pi}})$$

Hence,

$$m(\Theta) \le \sum_{\widetilde{\sigma}\in\Sigma}\sum_{\widetilde{\pi}\in\Pi} m(\Theta_{\widetilde{\sigma},\widetilde{\pi}})$$

Because Σ and Π are finite sets, to show $m(\Theta) = 0$, we only need to show $m(\Theta_{\widetilde{\sigma},\widetilde{\pi}}) = 0$ for any $\widetilde{\sigma} \in \Sigma$ and $\widetilde{\pi} \in \Pi$.

We begin by equating, $p_{X\to Y}(X, Y|\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \widetilde{\sigma})$ and $p_{Y\to X}(X, Y|\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \widetilde{\pi})$,

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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

for all $X \in \{1, \ldots, S\}$ and $Y \in \{1, \ldots, L\}$. The set of solutions to these equations is exactly $\Theta_{\widetilde{\sigma},\widetilde{\pi}}$. The left-hand side of (1) is given by

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

where $\beta_s = \boldsymbol{X}^T \boldsymbol{\beta}$ for $X = s$. Similarly, the right-hand side of (1) is given by

$$\begin{aligned}
&P_{Y\to X}(X = s, Y = \ell | \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \widetilde{\pi}) \\
&= P_{Y\to X}(X = s | Y = \ell, \boldsymbol{\alpha}, \boldsymbol{\eta}, \widetilde{\pi}) P_{Y\to X}(Y = \ell | \boldsymbol{\rho}) \\
&= [P_{Y\to X}(X \le s | Y = \ell, \boldsymbol{\alpha}, \boldsymbol{\eta}, \widetilde{\pi}) - P_{Y\to X}(X \le s - 1 | Y = \ell, \boldsymbol{\alpha}, \boldsymbol{\eta}, \widetilde{\pi})] P_{Y\to X}(Y = \ell | \boldsymbol{\rho}) \\
&= [F(\eta_{\widetilde{\pi}(s)} - \alpha_\ell) - F(\eta_{\widetilde{\pi}(s)-1} - \alpha_\ell)]\rho_\ell,
\end{aligned}$$

where $\alpha_\ell = \boldsymbol{Y}^T \boldsymbol{\alpha}$ for $Y = \ell$. Therefore, (1) leads to

$$[F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s = [F(\eta_{\widetilde{\pi}(s)} - \alpha_\ell) - F(\eta_{\widetilde{\pi}(s)-1} - \alpha_\ell)]\rho_\ell \tag{2}$$

Summing up both sides of (2) over s from 1 to S, we have

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

The second equality is due to a simple reordering of the summands. The third equality is due to the cancellation from telescoping series in s. The last equality 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),

$$\begin{aligned}
&[F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s \\
&= [F(\eta_{\widetilde{\pi}(s)} - \alpha_\ell) - F(\eta_{\widetilde{\pi}(s)-1} - \alpha_\ell)] \sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s
\end{aligned}$$

and hence

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

Now, consider $s = \widetilde{\pi}^{-1}(1)$ in (4) and note $\eta_0 = -\infty$ and $\eta_1 = 0$,

$$\begin{aligned}
\frac{[F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_{\widetilde{\pi}^{-1}(1)}) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_{\widetilde{\pi}^{-1}(1)})]\omega_{\widetilde{\pi}^{-1}(1)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s} &= F(\eta_1 - \alpha_\ell) - F(\eta_0 - \alpha_\ell) \\
&= F(-\alpha_\ell).
\end{aligned}$$

Therefore,

$$\alpha_\ell = -F^{-1} \left\{ \frac{[F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_{\widetilde{\pi}^{-1}(1)}) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_{\widetilde{\pi}^{-1}(1)})]\omega_{\widetilde{\pi}^{-1}(1)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s} \right\} \tag{5}$$

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

$$\begin{aligned}
\eta_{j^*} = F^{-1} &\left\{ \frac{\sum_{j=1}^{j^*} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_{\widetilde{\pi}^{-1}(j)}) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_{\widetilde{\pi}^{-1}(j)})]\omega_{\widetilde{\pi}^{-1}(j)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s} \right\} \\
- F^{-1} &\left\{ \frac{[F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_{\widetilde{\pi}^{-1}(1)}) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_{\widetilde{\pi}^{-1}(1)})]\omega_{\widetilde{\pi}^{-1}(1)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s} \right\},
\end{aligned} \tag{6}$$

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

$$\begin{aligned}
&F^{-1} \left\{ \frac{\sum_{j=1}^{j^*} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_{\widetilde{\pi}^{-1}(j)}) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_{\widetilde{\pi}^{-1}(j)})]\omega_{\widetilde{\pi}^{-1}(j)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s} \right\} \\
&- F^{-1} \left\{ \frac{[F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_{\widetilde{\pi}^{-1}(1)}) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_{\widetilde{\pi}^{-1}(1)})]\omega_{\widetilde{\pi}^{-1}(1)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell)-1} - \beta_s)]\omega_s} \right\} \\
=&F^{-1} \left\{ \frac{\sum_{j=1}^{j^*} [F(\gamma_{\widetilde{\sigma}(\ell^*)} - \beta_{\widetilde{\pi}^{-1}(j)}) - F(\gamma_{\widetilde{\sigma}(\ell^*)-1} - \beta_{\widetilde{\pi}^{-1}(j)})]\omega_{\widetilde{\pi}^{-1}(j)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell^*)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell^*)-1} - \beta_s)]\omega_s} \right\} \\
&- F^{-1} \left\{ \frac{[F(\gamma_{\widetilde{\sigma}(\ell^*)} - \beta_{\widetilde{\pi}^{-1}(1)}) - F(\gamma_{\widetilde{\sigma}(\ell^*)-1} - \beta_{\widetilde{\pi}^{-1}(1)})]\omega_{\widetilde{\pi}^{-1}(1)}}{\sum_{s=1}^{S} [F(\gamma_{\widetilde{\sigma}(\ell^*)} - \beta_s) - F(\gamma_{\widetilde{\sigma}(\ell^*)-1} - \beta_s)]\omega_s} \right\}
\end{aligned} \tag{7}$$

for any $\ell, \ell^* \in \{1, \ldots, L\}$ and $\ell \neq \ell^*$ (note that one can always find $\ell \neq \ell^*$ because $L > 2$). Since the link function F is assumed to be a real analytic function (recall a real function is said to be analytic if it is infinitely differentiable and matches its Taylor series in a neighborhood of every point), and $F'(x)$ is assumed to be nowhere zero, $F^{-1}(x)$ is analytic. Since the left-hand side of (7) is a composition of F, F^{-1}, sums, products, and reciprocals of $\gamma_{\widetilde{\sigma}(\ell)}, \gamma_{\widetilde{\sigma}(\ell^*)}, \gamma_{\widetilde{\sigma}(\ell)-1}, \gamma_{\widetilde{\sigma}(\ell^*)-1}, \beta_1, \ldots, \beta_S, \omega_1, \ldots, \omega_S$, it is an analytic function [Krantz and Parks, 2002] and therefore its zero set must have Lebesgue measure zero [Mityagin, 2015]. In other words, we have proven $m(\Theta_{\widetilde{\sigma},\widetilde{\pi}}) = 0$, which completes the proof.

2 Proof of Theorem 2

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n observations from $P_{X\to Y}(X, Y)$. Consider the average difference between the log-likelihood of $M_0 : X \to Y$ and that of $M_1 : Y \to X$,

$$\begin{aligned}
&\frac{1}{n}\left[\sum_{i=1}^{n} \log P_{X\to Y}(X = x_i, Y = y_i) - \sum_{i=1}^{n} \log P_{Y\to X}(X = x_i, Y = y_i)\right] \\
=&\frac{1}{n}\sum_{i=1}^{n} \log \frac{P_{X\to Y}(X = x_i, Y = y_i)}{P_{Y\to X}(X = x_i, Y = y_i)} \\
\to&E_{(X,Y)\sim P_{X\to Y}(X,Y)} \log \frac{P_{X\to Y}(X, Y)}{P_{Y\to X}(X, Y)} \quad \text{as} \quad n \to \infty \quad \text{(Law of large numbers)} \\
=&KL(P_{X\to Y}(X, Y) || P_{Y\to X}(X, Y)) > 0
\end{aligned}$$

The last inequality is because KL divergence is nonnegative and it is strictly positive here because $P_{X\to Y}(X, Y) \neq P_{Y\to X}(X, Y)$ from Theorem 1.

3 Additional Simulation Results

3.1 Ablation Study

We performed an ablation study under Simulation Scenario 1 with $n = 1{,}000$ to demonstrate the importance of learning the category ordering. Specifically, we fixed the permutation at different label orderings. The results are presented in Table 1 where Kendall's Tau quantifies the correlation between the fixed label permutation and the true label permutation. As expected, the causal identification accuracy increases as the Kendall's Tau approaches 1. The performance of the COLP method with unknown permutation is close to that under the fixed ordering with Kendall's Tau=0.8. This ablation study stresses the importance of having label permutation as a parameter because otherwise the inference can be very wrong if the permutation is fixed to a bad ordering.

Kendall's Tau	0	0.2	0.4	0.6	0.8	1
Accuracy	0.17	0.33	0.61	0.82	0.96	1

Table 1: Ablation study.

3.2 Varying Number of Categories

We investigated how estimation of causal direction and label permutation vary as the number of categories increases under Simulation Scenario 1 with $n = 1,000$. We considered 10 different numbers of categories from 3 to 12. The results are reported in Table 2 where the second row is the accuracy of causal identification and the third row is the Kendall's Tau measuring the correlation between the estimated and true label permutations. For both metrics, a value close to 1 indicates good performance. We find that the performance of COLP is relatively stable with respect to the number of categories, especially for causal identification.

Number of Categories	3	4	5	6	7	8	9	10	11	12
Accuracy	0.86	0.89	0.91	0.91	0.88	0.91	0.9	0.9	0.9	0.88
Kendall's Tau	0.94	0.92	0.93	0.92	0.88	0.9	0.88	0.87	0.87	0.84

Table 2: Varying number of categories.

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.