# **Supplementary Material**

Gunwoong Park Department of Statistics University of Wisconsin-Madison Madison, WI 53706 parkg@stat.wisc.edu Garvesh Raskutti Department of Statistics Department of Computer Science Wisconsin Institute for Discovery, Optimization Group University of Wisconsin-Madison Madison, WI 53706 raskutti@cs.wisc.edu

## 1 **Proof of Theorem 3.1**

*Proof.* We prove it by induction that requires p steps to find a causal ordering that is consistent with the DAG. Without loss of generality, assume that one of the true causal ordering  $\pi^*$  is  $\{1, 2, ..., p\}$ . For ease of notation, let  $\mathcal{F}_s = \{X_1, X_2, \cdots, X_s\}$ . Let k = 1 be the first step:

 $\operatorname{Var}(X_j) = \mathbb{E}(\operatorname{Var}[X_j | \mathcal{F}_{j-1}]) + \operatorname{Var}(\mathbb{E}[X_j | \mathcal{F}_{j-1}]),$ 

where the outer expectation and variance is taken over  $X_1, X_2, ..., X_{j-1}$ . Since the conditional distribution  $X_j | \mathcal{F}_{j-1} \sim \text{Poisson}(g_j(X_{\text{Pa}(j)}))$ , we have  $\text{Var}[X_j | \mathcal{F}_{j-1}] = \mathbb{E}[X_j | \mathcal{F}_{j-1}] = g_j(X_{\text{Pa}(j)})$ . Hence,

$$Var(X_j) = \mathbb{E}(\mathbb{E}[X_j | \mathcal{F}_{j-1}]) + Var(g_j(X_{\mathbf{Pa}(j)}))$$
$$= \mathbb{E}(X_j) + Var(g_j(X_{\mathbf{Pa}(j)})),$$

yielding that

$$\operatorname{Var}(X_j) - \mathbb{E}(X_j) = \operatorname{Var}(g_j(X_{\operatorname{Pa}(j)})).$$

Clearly, if Pa(j) is empty, meaning the node is the first component of the causal ordering,  $Var(g_j(X_{Pa(j)})) = 0$ . Otherwise,  $Var(g_j(X_{Pa(j)})) > 0$  by the assumption. Hence for any node that can not be the first in the ordering,  $Var(X_j) - \mathbb{E}(X_j) > 0$ . Hence we pick any node  $X_k$  such that  $Var(X_k) - \mathbb{E}(X_k) = 0$  as being the first element of the causal ordering and  $X_1$  satisfies the above equation.

For k = m, assume  $X_1, X_2, ..., X_m$  is a valid causal ordering for the first m nodes. Now we consider

$$\operatorname{Var}(X_j|\mathcal{F}_m) = \mathbb{E}(\operatorname{Var}[X_j|\mathcal{F}_{j-1}]|\mathcal{F}_m) + \operatorname{Var}(\mathbb{E}[X_j|\mathcal{F}_{j-1}]|\mathcal{F}_m)$$

for j = m + 1, m + 2, ..., p, where the expectation and variance are taken over the variables  $X_1, X_2, ..., X_m$ . Again, for any j = m + 1, m + 2, ..., p, we have  $\operatorname{Var}[X_j | \mathcal{F}_{j-1}] = \mathbb{E}[X_j | \mathcal{F}_{j-1}] = g_j(X_{\operatorname{Pa}(j)})$ . Further, since  $X_1, X_2, ..., X_m$  is a valid causal ordering for the first m nodes,

$$\operatorname{Var}(X_j|\mathcal{F}_m) = \mathbb{E}(\mathbb{E}[X_j|\mathcal{F}_{j-1}]|\mathcal{F}_m) + \operatorname{Var}(\mathbb{E}(X_j|\mathcal{F}_{j-1})|\mathcal{F}_m) \\ = \mathbb{E}(X_j|\mathcal{F}_m) + \operatorname{Var}(g_j(X_{\mathbf{P}_{\mathbf{a}(j)}})|\mathcal{F}_m).$$

Hence, following on similar lines,

$$\operatorname{Var}(X_j | \mathcal{F}_m) - \mathbb{E}(X_j | \mathcal{F}_m) = \operatorname{Var}[g_j(X_{\operatorname{Pa}(j)}) | \mathcal{F}_m].$$

Hence if  $\operatorname{Pa}(j) \setminus \{1, 2, ..., m\}$  is empty,  $\operatorname{Var}(g_j(X_{\operatorname{Pa}(j)})|\mathcal{F}_m) = 0$  and  $\operatorname{Var}(X_j|\mathcal{F}_m) - \mathbb{E}(X_j|\mathcal{F}_m) = 0$ . Any such node can be next on the causal ordering and  $X_m$  holds the above property. On the other hand, for any node in which  $\operatorname{Pa}(j) \setminus \{1, 2, ..., m\}$  is non-empty  $\operatorname{Var}(X_j|\mathcal{F}_m) - \mathbb{E}(X_j|\mathcal{F}_m) > 0$  which excludes it from being next in the causal ordering. Hence  $X_1, X_2, ..., X_{m+1}$  is a valid causal ordering for the first m + 1 nodes. This completes the proof by induction.

## 2 Proof of Theorem 4.2

*Proof.* Let  $X^{(i)} = (X_1^{(i)}, ..., X_p^{(i)})$  be the i.i.d *n* samples from the given DAG model. Let  $\pi^*$  be a true causal ordering and  $\hat{\pi}$  be the estimated causal ordering. Without loss of generality, assume that the true causal ordering  $\pi^*$  is  $\{1, 2, ...p\}$ . For an arbitrary permutation or causal ordering  $\pi$ , let  $\pi_j$  represent its  $j^{th}$  element.

Let  $E_u$  denote the set of undirected edges corresponding to the *moralized* graph (i.e. the directed edges without directions and edges between nodes with common children). Recall the definitions  $\mathcal{N}(j) := \{k \in \{1, 2, ..., p\} \mid (j, k) \in E_u\}$  denote the neighborhood set of j in the moralized graph and  $K(j) = \{k \mid k \in \mathcal{N}(j-1) \cap \{j, ..., p\}\}$  denote a candidate set for  $\pi_j$  and  $C_{jk} = \mathcal{N}(k) \cap \{\pi_1, \pi_2, ..., \pi_{j-1}\}$  which is the intersection of the neighbors of k with  $\{1, 2, ..., j-1\}$ .

Recall that for ease of notation for any  $j \in \{1, 2, ..., p\}$ , and  $S \subset \{1, 2, ..., p\}$  let  $\mu_{j|S}$  and represent  $\mathbb{E}[X_j|X_S]$  and  $\sigma_{j|S}^2 = \operatorname{Var}(X_j|X_S)$ . Also, denote  $\mu_{j|S}(x_S)$  and represent  $\mathbb{E}[X_j|X_S = x_S]$  and  $\sigma_{j|S}^2(x_S) = \operatorname{Var}(X_j|X_S = x_S)$ . Let  $n_S(x_S) = \sum_{i=1}^n \mathbf{1}(X_S^{(i)} = x_S)$  and  $n_S = \sum_{x_S} n(x_S)\mathbf{1}(n(x_S) \ge c_0.n)$  for an arbitrary  $c_0 \in (0, 1)$ .

The overdispersion score of  $k \in K(j)$  for the  $j^{th}$  component of the causal ordering, defined in the second step of our ODS algorithm only considers elements of  $\mathcal{X}(\widehat{C}_{jk}) = \{x \in \{X_{\widehat{C}_{jk}}^{(1)}, X_{\widehat{C}_{jk}}^{(2)}, ..., X_{\widehat{C}_{jk}}^{(n)}\} \mid n(x) \geq c_0.n\}$  so we only count up elements that occur sufficiently frequently.

According to the ODS algorithm, the truncated sample conditional expectation and variance of  $X_j$  given  $X_S = x$  for  $j \in \{1, 2, ...p\}$  and any subset  $S \subset \{1, 2, ...p\} \setminus \{j\}$  be following: for  $x \in \mathcal{X}(S)$ ,

$$\hat{\mu}_{j|S}(x) = \frac{1}{n_S(x)} \sum_{i=1}^n X_j^{(i)} \mathbf{1}(X_S^{(i)} = x)$$
$$\hat{\sigma}_{j|S}^2(x) = \frac{1}{n_S(x) - 1} \sum_{i=1}^n (X_j^{(i)} - \hat{\mu}_{j|S}(x))^2 \mathbf{1}(X_S^{(i)} = x)$$

The overdispersion score of  $k \in K(j)$  for the  $j^{th}$  element of the causal ordering is for  $x \in \mathcal{X}(C_{jk})$ ,

$$\widehat{s}_{jk}(x) = \widehat{\sigma}_{k|\widehat{C}_{jk}}^{\mathbb{Z}}(x) - \widehat{\mu}_{k|\widehat{C}_{jk}}(x)$$
$$\widehat{s}_{jk} = \widehat{\mathbb{E}}_{\widehat{C}_{jk}}(\widehat{s}_{jk}(x)) = \sum_{x \in \mathcal{X}(jk)} \frac{n_{\widehat{C}_{jk}}(x)}{n_{\widehat{C}_{jk}}} \widehat{s}_{jk}(x).$$

And the correct overdispersion score is

$$s_{jk}^* = \mathbb{E}_{C_{jk}}[\sigma_{k|C_{jk}}^2 - \mu_{k|C_{jk}}] = \mathbb{E}_{C_{jk}}[\operatorname{Var}(g_k(\operatorname{Pa}(k))|C_{jk})]$$

Let us define some events for the proof and d denote the maximum degree of the moralized graph. For any  $j \in \{1, 2, ..., p\}$  and  $k \in K(j)$ ,

$$\xi_1 = \{ \max_{j,k} |\hat{s}_{jk} - s_{jk}^*| < m/2 \}$$
  
$$\xi_2 = \{ \max_k \max_{i=1,\dots,n} X_k^{(i)} < n^{\frac{1}{5+d}} \}$$

We prove it by induction that requires p steps to recover a causal ordering that is consistent with the Poisson DAG. Without loss of generality, assume that the true causal ordering  $\pi^*$  is  $\{1, 2, ..., p\}$ . For the first step j = 1, a set of candidate element of  $\pi_1$  is  $K(1) = \{1, 2, ..., p\}$  and a candidate parent set of each node  $C_{1k} = \emptyset$  for all  $k \in K(1)$ .

$$\begin{split} P(\widehat{\pi}_{1} \neq \pi_{1}^{*}) &= P\left(\text{exists at least one } k \in K(1) \setminus \{1\} \text{ s.t. } \widehat{s}_{11} > \widehat{s}_{1k}\right) \\ &\leq |K(1)| \max_{k \in K(1) \setminus \{1\}} \left\{ P\left(s_{11}^{*} + \frac{m}{2} > s_{1k}^{*} - \frac{m}{2}|\xi_{1}\right) + P(\xi_{1}^{c}|\xi_{2}) + P(\xi_{2}^{c}) \right\} \\ &\leq p \max_{k \in K(1) \setminus \{1\}} \left\{ P\left(m > s_{1k}^{*}|\xi_{1}\right) + P(\xi_{1}^{c}|\xi_{2}) + P(\xi_{2}^{c}) \right\} \end{split}$$

By Assumption (A1),  $s_{1k}^* > m$  and we will represent some Propositions that respectively control  $P(\xi_1^c|\xi_2)$  and  $P(\xi_2^c)$ .

For the j-1 step, assume  $(\hat{\pi}_1, \hat{\pi}_2, ..., \hat{\pi}_{j-1})$  is a valid ordering for the first j-1 nodes. Note that with the correct  $\mathcal{N}(j)$ ,  $\hat{C}_{jk} = C_{jk}$ . Now, we consider  $\pi_j^*$ . The probability of a false recovery of  $\pi_j^*$  given the true undirected edges of the moralized graph and the true causal ordering before j is following:

$$\begin{aligned} & \mathsf{P}(\widehat{\pi}_{j} \neq \pi_{j}^{*} | \widehat{\pi}_{1} = \pi_{1}^{*}, ..., \widehat{\pi}_{j-1} = \pi_{j-1}^{*}) \\ &= P\left(\text{exists at least one } k \in K(j) \setminus \{j\} \text{ s.t. } \widehat{s}_{jj} > \widehat{s}_{jk}\right) \\ &\leq |K(j)| \max_{k \in K(j) \setminus \{j\}} \left\{ P\left(\widehat{s}_{jj} + m/2 > s_{jk}^{*} - m/2 | \xi_{1}\right) + P(\xi_{1}^{c} | \xi_{2}) + P(\xi_{2}^{c}) \right\} \\ &\leq |K(j)| \max_{k \in K(j) \setminus \{j\}} \left\{ P\left(m > s_{jk}^{*} | \xi_{1}\right) + P(\xi_{1}^{c} | \xi_{2}) + P(\xi_{2}^{c}) \right\} \end{aligned}$$

By Assumption (A1),  $s_{jk}^* > m$  and we represent some Propositions that respectively control  $P(\xi_1^c|\xi_2)$  and  $P(\xi_2^c)$ . Furthermore we also show a condition on  $c_0$ .

**Proposition 2.1.** For all  $j \in \{1, 2, ..., p\}, k \in K(j)$ ,  $c_0 \leq n^{-\frac{d}{5+d}}$  given  $\xi_2$  is a sufficient that a candidate parents set  $\mathcal{X}(C_{jk})$  is not empty

#### **Proposition 2.2.**

$$P(\xi_1^c|\xi_2) \le 2p^2 n^{\frac{d}{5+d}} \left\{ \exp\left(-\frac{m^2 n^{1/(5+d)}}{18}\right) + \exp\left(-\frac{m^2 n^{1/(5+d)}}{9}\right) + \exp\left(-\frac{m^2 n^{3/(5+d)}}{9}\right) \right\},$$

where m is the constant in Assumption (A1).

#### **Proposition 2.3.**

$$P(\xi_2^c) \le npMexp(-n^{1/(5+d)}\log 2)$$

where M is the constant in Assumption (A2).

Hence for any  $j \in \{1, 2, ... p\}$  with  $c_0 = n^{-\frac{d}{5+d}}$ ,

$$\begin{aligned} & \mathsf{P}(\widehat{\pi}_{j} \neq \pi_{j}^{*} | \widehat{\pi}_{1} = \pi_{1}^{*}, ..., \widehat{\pi}_{j-1} = \pi_{j-1}^{*}) \\ & \leq p \max_{k \in K(j) \setminus \{j\}} \left\{ P\left(m > s_{jk}^{*} | \xi_{1}\right) + P(\xi_{1}^{c} | \xi_{2}) + P(\xi_{2}^{c}) \right\} \\ & \leq 2p^{3} n^{\frac{d}{5+d}} \left\{ \exp\left(-\frac{m^{2} n^{1/(5+d)}}{18}\right) + \exp\left(-\frac{m^{2} n^{1/(5+d)}}{9}\right) + \exp\left(-\frac{m^{2} n^{3/(5+d)}}{9}\right) \right\} \\ & + np^{2} M \exp\left(-n^{1/(5+d)} \log 2\right) \end{aligned}$$
(1)

By using the above probability bound (1),

$$\begin{split} & P(\widehat{\pi} \neq \pi^*) \stackrel{(E_1)}{\leq} P(\widehat{\pi}_1 \neq \pi_1^*) + \ldots + P(\widehat{\pi}_{p-1} \neq \pi_{p-1}^* | \widehat{\pi}_1 = \pi_1^*, \ldots, \widehat{\pi}_{p-2} = \pi_{p-2}^*) \\ & \stackrel{(E_2)}{\leq} 2p^4 n^{\frac{d}{5+d}} \{ \exp\big( -\frac{m^2 n^{1/(5+d)}}{18} \big) + \exp\big( -\frac{m^2 n^{1/(5+d)}}{9} \big) + \exp\big( -\frac{m^2 n^{3/(5+d)}}{9} \big) \} \\ & \quad + np^3 M \exp\big( -n^{1/(5+d)} \log 2 \big) \end{split}$$

The first inequality  $(E_1)$  is followed from  $P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B|A^c)P(A^c) \le P(A) + P(B|A^c)$  for some events A, B. And  $(E_2)$  is directly from (1).

Hence, there exists some positive constants  $C_1, C_2, C_3 > 0$  such that

$$P(\hat{\pi} \neq \pi^*) \le C_1 \exp\left(-C_2 n^{1/(5+d)} + C_3 \log \max\{p, n\}\right)$$

		I	

### 2.0.1 Proof of Proposition 2.1

*Proof.* Let  $|X_S|$  denote the cardinality of a set  $\{X_S^{(1)}, X_S^{(2)}, ..., X_S^{(n)}\}$  and  $|\mathcal{X}(S)|$  denote the cardinality of a set  $\mathcal{X}(S)$ . In worst case where  $|\mathcal{X}(S)| = 1$ , for all  $x \in \{X_S^{(1)}, X_S^{(2)}, ..., X_S^{(n)}\}$ ,  $n_S(x) = c_0.n - 1$  except for only one component  $y \in \mathcal{X}(S)$ . In this case, the sample size  $n = n_S(y) + (|X_S| - 1)(c_0.n - 1)$ . A simple calculation yields that

$$n_S(y) = n - (|X_S| - 1)(c_0 \cdot n - 1) = n - c_0 \cdot n |X_S| + c_0 \cdot n + |X_S| - 1.$$

Hence  $c_0.n \leq n_S(y)$  is equivalent to  $c_0 \leq \frac{n+|X_S|-1}{n.|X_S|}$ . Since  $\frac{1}{|X_S|} \leq \frac{n+|X_S|-1}{n|X_S|}$ , if  $c_0 \leq \frac{1}{|X_S|}$  there exists at least one component  $y \in \mathcal{X}(S)$ . In addition under the event  $\xi_2$ ,  $|X_S| \leq n^{\frac{d}{5+d}}$  which is all possible combinations. Hence if  $c_0 \leq n^{-\frac{d}{5+d}}$ ,  $|\mathcal{X}(S)| \neq 0$ .

#### 2.0.2 **Proof of Proposition 2.2**

*Proof.* This problem is reduced to the consistency rate of a sample conditional mean and conditional variance. For ease of notation, let  $n_{jk} = n_{C_{jk}}$  and  $n_{jk}(x) = n_{C_{jk}}(x)$ . Suppose that  $c_0 = n^{-\frac{d}{5+d}}$ . Then for any  $j \in \{1, 2, ..., p\}$  and  $k \in K(j)$ ,

$$\begin{split} &P(\xi_{1}^{c}|\xi_{2}) \leq p^{2} \max_{j,k} P(|\hat{s}_{jk} - s_{jk}^{*}| > \frac{m}{2}|\xi_{2}) \\ &\leq p^{2} \max_{j,k} P(\sum_{x \in \mathcal{X}(C_{jk})} \frac{n_{jk}(x)}{n_{jk}} |\hat{s}_{jk}(x) - s_{jk}^{*}(x)| > \frac{m}{2}|\xi_{2}) \\ & \stackrel{(E_{1})}{\leq} p^{2} \max_{j,k} \sum_{x \in \mathcal{X}(C_{jk})} P(|\hat{s}_{jk}(x) - s_{jk}^{*}(x)| > \frac{m}{2} \frac{n_{jk}}{n_{jk}(x)}|\xi_{2}) \\ & \stackrel{(E_{2})}{\leq} p^{2} \max_{j,k} |\mathcal{X}(C_{jk})| \max_{x \in \mathcal{X}(C_{jk})} P(|\hat{s}_{jk}(x) - s_{jk}^{*}(x)| > \frac{m}{2}|\xi_{2}) \\ & \stackrel{(E_{3})}{\leq} p^{2} n^{\frac{d}{5+d}} \max_{j,k,x} P(|(\hat{\sigma}_{k|C_{jk}}^{2}(x) - \hat{\mu}_{k|C_{jk}}(x)) - (\sigma_{k|C_{jk}}^{2}(x) - \mu_{k|C_{jk}}(x))| > \frac{m}{2}|\xi_{2}) \\ &\leq p^{2} n^{\frac{d}{5+d}} \max_{j,k,x} \left\{ P(|\hat{\sigma}_{k|C_{jk}}^{2}(x) - \hat{\sigma}_{j|C_{jk}}^{2}(x)| > \frac{m}{3}|\xi_{2}) + P(|\hat{\mu}_{k|C_{jk}}(x) - \mu_{k|C_{jk}}(x)| > \frac{m}{6}|\xi_{2}) \right\} \\ & \stackrel{(E_{4})}{\leq} 2p^{2} n^{\frac{d}{5+d}} \max_{j,k,x} \left\{ \exp\left(-\frac{m^{2}n_{jk}(x)}{18n^{4/(5+d)}}\right) + \exp\left(-\frac{m^{2}n_{jk}(x)}{9n^{4/(5+d)}}\right) + \exp\left(-\frac{m^{2}n_{jk}(x)}{9n^{2/(5+d)}}\right) \right\} \\ & \stackrel{(E_{5})}{\leq} 2p^{2} n^{\frac{d}{5+d}} \max_{j,k,x} \left\{ \exp\left(-\frac{m^{2}n^{1/(5+d)}}{18}\right) + \exp\left(-\frac{m^{2}n^{1/(5+d)}}{9}\right) + \exp\left(-\frac{m^{2}n^{3/(5+d)}}{9}\right) \right\} \\ &= 2p^{2} n^{\frac{d}{5+d}} \left\{ \exp\left(-\frac{m^{2}n^{1/(5+d)}}{18}\right) + \exp\left(-\frac{m^{2}n^{1/(5+d)}}{9}\right) + \exp\left(-\frac{m^{2}n^{3/(5+d)}}{9}\right) \right\}. \end{split}$$

 $\begin{array}{l} (E_1) \text{ is followed from that } P(\sum_i \omega_i X_i > \delta) \leq \sum_i P(X_i > \delta/\omega_i) \text{, and } (E_2) \text{ is from } \frac{n_{jk}(x)}{n_{jk}} < 1. \\ \text{Since } n_{jk}(x) \geq c_0.n \text{ for all } x \in \mathcal{X}(C_{jk}), |\mathcal{X}(C_{jk})| \leq 1/c_0 \text{ hence } (E_3) \text{ and } (E_5) \text{ hold. Moreover, } \\ (E_4) \text{ is followed from the Hoeffding's inequality (Theorem 2 [1]) since samples are independent and bounded above } n^{1/(5+d)} \text{ given } \xi_2. \end{array}$ 

## 2.0.3 Proof of Proposition 2.3

*Proof.* For any  $j \in \{1, 2, ..., p\}$ , the conditional distribution of  $X_j$  given  $X_{pa(j)}$  is Poisson with rate parameter  $g_j(Pa(j))$ . Hence for  $k \in K(j)$ ,

$$\begin{split} P(\xi_2^c) &= P(\max_{k \in K(j)} \max_{i=1,...,n} X_k^{(i)} > n^{1/(5+d)}) \\ &\stackrel{(E_1)}{\leq} np \max_{k \in K(j)} \max_{i=1,...,n} P(|X_k^{(i)}| > n^{1/(5+d)}) \\ &\stackrel{(E_2)}{\leq} np \max_{k \in K(j)} \max_{i=1,...,n} \mathbb{E} \mathsf{pa}_{(k)} \left[ \exp\left(-n^{1/(5+d)} \log 2 + g_k(\mathsf{pa}(k))\right) \right] \\ &\stackrel{(E_3)}{\leq} np \max_{k \in K(j)} \max_{i=1,...,n} M \exp(-n^{1/(5+d)} \log 2) \\ &= np M \exp\left(-n^{1/(5+d)} \log 2\right). \end{split}$$

 $(E_1)$  is followed from the union bound and |K(j)| < p.  $(E_2)$  is from the moment generating function of Poisson distribution with  $t = \log 2$ . And,  $(E_3)$  is from Assumption (A2).

## References

[1] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.