Supplementary for Identifiability of Generalized Hypergeometric Distribution (GHD) DAG Models

Gunwoong Park
Department of Statistics, University of Seoul

Hyewon Park
Department of Statistics, University of Seoul

1 Supplementary

1.1 Proof for Proposition 2.2

Proof. For any positive integer $r \geq 1$, Kemp (1968) shows that

$$\mathbb{E}((X_j)_r \mid X_{\mathrm{Pa}(j)}) \quad = \quad \theta^r \prod_{i=1}^{p} \frac{(a_i + r - 1)!}{(a_i - 1)!} \prod_{j=1}^{q} \frac{(b_j - 1)!}{(b_j + r - 1)!}. \tag{1}$$

Then, the expectation can be obtained when $r = 1$.

$$\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)}) \quad = \quad \theta \times \prod_{i=1}^{p} a_i \prod_{j=1}^{q} \frac{1}{b_j}.$$

By plugging this into Eqn. (1), we have

$$\mathbb{E}((X_j)_r \mid X_{\mathrm{Pa}(j)}) \quad = \quad \mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)})^r \prod_{i=1}^{p} \frac{(a_i + r - 1)!}{(a_i - 1)! a_i^r} \prod_{j=1}^{q} \frac{(b_j - 1)! b_j^r}{(b_j + r - 1)!}.$$

□

1.2 Proof for Theorem 2.4

Proof. Without loss of generality, we assume the true ordering is unique and $\pi = (\pi_1, ..., \pi_p)$. For notational convenience, we define $X_{1:j} = (X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_j})$ and $X_{1:0} = \emptyset$. We prove the identifiability of our GHD DAG models by mathematical induction.

As we discussed in the main body of the paper, for any node $j \in V \setminus \{\pi_1\}$, Prop. 2.2 and Assumption 2.3 ensure that

$$\frac{\mathbb{E}((X_j)_r)}{f_j^{(r)}(\mathbb{E}(X_j))} > 1, \text{ and } \quad \frac{\mathbb{E}((X_{\pi_1})_r)}{f_{\pi_1}^{(r)}(\mathbb{E}(X_{\pi_1}))} = \frac{f_{\pi_1}^{(r)}(\mathbb{E}(X_{\pi_1})))}{f_{\pi_1}^{(r)}(\mathbb{E}(X_{\pi_1}))} = 1.$$

Hence we can determine π_1 as the first element of the ordering.

For the $(m-1)^{th}$ element of the ordering, assume that the first $m-1$ elements of the ordering and their parents are correctly estimated. Now, we consider the m^{th} element of the ordering and its parents. Again Prop. 2.2 and Assumption 2.3 yield that for$j \in \{\pi_{m+1}, \pi_{m+2}, \cdots, \pi_p\}$,

$$\frac{\mathbb{E}((X_j)_r)}{\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{1:(m-1)})))} > \frac{\mathbb{E}(\mathbb{E}((X_j)_r \mid X_{1:(m-1)}))}{\mathbb{E}(\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)})) \mid X_{1:(m-1)}))} = 1.$$

In addition, it is clear that

$$\frac{\mathbb{E}(\mathbb{E}((X_{\pi_m})_r \mid X_{1:(m-1)}))}{\mathbb{E}(f_{\pi_m}^{(r)}(\mathbb{E}(X_{\pi_m} \mid X_{1:(m-1)})))} = \frac{\mathbb{E}(\mathbb{E}((X_{\pi_m})_r \mid X_{\mathrm{Pa}(\pi_m)}))}{\mathbb{E}(f_{\pi_m}^{(r)}(\mathbb{E}(X_{\pi_m} \mid X_{\mathrm{Pa}(\pi_m)})))} = \frac{\mathbb{E}((X_{\pi_m})_r)}{\mathbb{E}(f_{\pi_m}^{(r)}(\mathbb{E}(X_{\pi_m} \mid X_{\mathrm{Pa}(\pi_m)})))} = 1.$$

Hence we can estimate a valid m^{th} component of the ordering π_m and its parents by testing whether the r-th moments ratio is whether greater than or equal to 1. By the mathematical induction this completes the proof.

□

1.3 Proof for Theorem 3.2

Proof. We first reintroduce some necessary notations and definitions to make the proof concise. Without loss of generality, assume that the true ordering is unique and $\pi = (\pi_1, ..., \pi_p)$. In addition, we assume the true skeleton is provided. For ease of notation, we drop the r in the both r-th moments ratio scores and CMR function. Then, the element of the score can be written as:

$$\mathcal{S}(j,k)(X_{C_{jk}}) := \frac{\mathbb{E}(X_k^r \mid X_{C_{jk}})}{f_k(\mathbb{E}(X_k \mid X_{C_{jk}})) - \sum_{m=0}^{r-1} s(r,m)\mathbb{E}(X_k^m \mid X_{C_{jk}})}, \text{ and}$$
$$\widehat{S}(j,k)(X_{\widehat{C}_{jk}}) := \frac{\widehat{\mathbb{E}}(X_k^r \mid X_{\widehat{C}_{jk}})}{f_k(\widehat{\mathbb{E}}(X_k \mid X_{\widehat{C}_{jk}})) - \sum_{m=0}^{r-1} s(r,m)\widehat{\mathbb{E}}(X_k^m \mid X_{\widehat{C}_{jk}})}.$$

where C_{jk} is the candidate parents set and $s(r,k)$ is Stirling numbers of the first kind. Hence the r-th moments ratio score is

$$\widehat{\mathcal{S}}(j,k) := \sum_{x \in \mathcal{X}_{\widehat{C}_{jk}}} \frac{n(x)}{n_{\widehat{C}_{jk}}} \widehat{\mathcal{S}}_r(j,k)(x).$$

We define the following important events: For each node $j \in V$ and set $S \subset V \setminus (\mathrm{De}(j) \cup \{j\})$ and for any $\epsilon > 0$, let

$$\begin{aligned}
\zeta_1 &:= \left\{ \min_{j=1,...,p-1} \min_{k=j,...,p} \left| \mathcal{S}(j,\pi_k) - \widehat{S}(j,\pi_k) \right| > \frac{M_{\min}}{2} \right\} \\
\zeta_2 &:= \left\{ \max_{j \in V} \left| \widehat{\mathbb{E}}(X_j^r \mid X_S) - \mathbb{E}(X_j^r \mid X_S) \right| < \epsilon \right\} \\
\zeta_3 &:= \left\{ \max_{j \in V} \left| f_j\left(\widehat{\mathbb{E}}(X_j \mid X_S)\right) - f_j\left(\mathbb{E}(X_j \mid X_S)\right) \right| < \frac{\epsilon}{2} \right\} \\
\zeta_4 &:= \left\{ \max_{j \in V} \left| \left(\sum_{k=0}^{r-1} s(r,k)\widehat{\mathbb{E}}(X_j^k \mid X_S) - \sum_{k=0}^{r-1} s(r,k)\mathbb{E}(X_j^k \mid X_S) \right) \right| < \frac{\epsilon}{2} \right\} \\
\zeta_5 &:= \left\{ \max_{j \in V} \max_{i \in \{1,2,\cdots,n\}} |X_j^{(i)}| < 4 \log \eta \right\}. \qquad (2)
\end{aligned}$$

Here we use the method of moments estimators $\frac{1}{n}\sum_{i=1}^n (X_j^{(i)})^k$ as unbiased estimators for $\mathbb{E}(X_j^k)$ for all $1 \le k \le r$.

We prove that our algorithm recovers the ordering of any GHD DAG model in the high dimensional settings if the indgree is bounded. The probability that ordering is correctly estimated from our method can be written as

$$\begin{aligned}
& P\left(\widehat{\pi} = \pi\right) \\
= \; & P\left(\widehat{S}(1,\pi_1) < \min_{j=2,...,p} \widehat{S}(1,\pi_j), \widehat{S}(2,\pi_2) < \min_{j=3,...,p} \widehat{S}(1,\pi_j), ..., \widehat{S}(p-1,\pi_{p-1}) < \widehat{S}(p-1,\pi_p) \right) \\
= \; & P\left(\min_{j=1,...,p-1} \min_{k=j+1,...,p} \widehat{S}(j,\pi_k) - \widehat{S}(j,\pi_j) > 0 \right) \\
= \; & P\left(\min_{j=1,...,p-1} \min_{k=j+1,...,p} \left\{ (\mathcal{S}(j,\pi_k) - \mathcal{S}(j,\pi_j)) - \left(\mathcal{S}(j,\pi_k) - \widehat{S}(j,\pi_k)\right) + \left(\mathcal{S}(j,\pi_j) + \widehat{S}(j,\pi_j)\right) \right\} > 0 \right) \\
\ge \; & P\left(\min_{j=1,...,p-1} \min_{k=j+1,...,p} \{(\mathcal{S}(j,\pi_k) - \mathcal{S}(j,\pi_j))\} > M_{\min}, \; \textit{and} \min_{j=1,...,p-1} \min_{k=j,...,p} \left| \mathcal{S}(j,\pi_k) - \widehat{S}(j,\pi_k) \right| < \frac{M_{\min}}{2} \right).
\end{aligned}$$

By Assumption 3.1 (A1) $\mathcal{S}(j,\pi_k) > 1 + M_{\min}$, the above lower bound of the probability is reduced to

$$\begin{aligned}
P\left(\widehat{\pi}=\pi\right) &\geq 1 - P\left(\min_{j=1,\ldots,p-1}\min_{k=j,\ldots,p}\left|\mathcal{S}(j,\pi_k)-\widehat{S}(j,\pi_k)\right| > \frac{M_{\min}}{2}\right) \\
&= 1 - P(\zeta_1) \\
&= 1 - \{P(\zeta_1 \mid \zeta_2,\zeta_3,\zeta_4)P(\zeta_2,\zeta_3,\zeta_4) + P(\zeta_1 \mid (\zeta_2,\zeta_3,\zeta_4)^c)P((\zeta_2,\zeta_3,\zeta_4)^c)\} \\
&\geq 1 - \{P(\zeta_1 \mid \zeta_2,\zeta_3,\zeta_4) + P((\zeta_2,\zeta_3,\zeta_4)^c)\} \\
&= 1 - \{P(\zeta_1 \mid \zeta_2,\zeta_3,\zeta_4) + P((\zeta_2,\zeta_3,\zeta_4)^c \mid \zeta_5)P(\zeta_5) + P((\zeta_2,\zeta_3,\zeta_4)^c \mid \zeta_5^c)P(\zeta_5^c)\} \\
&\geq 1 - \{P(\zeta_1 \mid \zeta_2,\zeta_3,\zeta_4) + P((\zeta_2,\zeta_3,\zeta_4)^c \mid \zeta_5) + P(\zeta_5^c)\} \\
&\geq 1 - \Big\{ \underbrace{P(\zeta_1 \mid \zeta_2,\zeta_3,\zeta_4)}_{Prop1.1} + \underbrace{P(\zeta_2^c \mid \zeta_5) + P(\zeta_3^c \mid \zeta_5) + P(\zeta_4^c \mid \zeta_5)}_{Prop1.2} + \underbrace{P(\zeta_5^c)}_{Prop1.3} \Big\}.
\end{aligned}$$

Next we introduce the following three propositions to show that the above lower bound converges to 1. The first proposition proves that estimated score is accurate under some regularity conditions. For ease of notation, let

$$g_j(\widehat{\mathbb{E}}(X_j \mid X_{\widehat{C}_{jk}})) = f_k(\widehat{\mathbb{E}}(X_k \mid X_{\widehat{C}_{jk}})) - \sum_{m=0}^{r-1} s(r,m)\widehat{\mathbb{E}}(X_k^m \mid X_{\widehat{C}_{jk}}).$$

Proposition 1.1. *Given the sets ζ_2,ζ_3,ζ_4 in Egn. (2), $P(\zeta_1 \mid \zeta_2,\zeta_3,\zeta_4) = 0$ if one of the following conditions are satisfied for any $S \subset V \setminus (De(j) \cup \{j\})$:*

(i) $2\mathbb{E}(X_j^r \mid X_S) + (2 - M_{\min})g_j(\mathbb{E}(X_j \mid X_S)) \leq 0$ *or*

(ii) $\epsilon < \frac{M_{\min} g_j(\mathbb{E}(X_j|X_S))^2}{\left(2\mathbb{E}(X_j^r|X_S)+(2-M_{\min})g_j(\mathbb{E}(X_j|X_S))\right)}$.

The first condition (i) is satisfied if $M_{\min}$ in Assumption 3.1 (A1) is sufficiently large and the second condition (ii) is satisfied if ϵ is sufficiently small. This means that if the estimated r-th factorial moment is close to the true r-th factorial moment, then ζ_1 is not satisfied with probability 1. Hence we discuss the error bound for the r-th factorial moment estimator in the next.

The following propositions show the error bound for the higher order moment X_j^k for $1 \leq k \leq r$ given the set ζ_5, and therefore the error bound for the r-th factorial moment estimator:

Proposition 1.2. *For any node $j \in V$ and any set $S \subset V \setminus (De(j) \cup \{j\})$ and for any $\epsilon > 0$,*

(i) For ζ_2,

$$P(\zeta_2^c \mid \zeta_5) \leq 2 \cdot p \cdot exp\left\{-\frac{2N_{\min}\epsilon^2}{(4\log^2\eta)^r}\right\}.$$

(ii) For ζ_3, and $m \in (\mathbb{E}(X_j \mid X_S) - \epsilon/2, \mathbb{E}(X_j \mid X_S) + \epsilon/2)$,

$$P(\zeta_3^c \mid \zeta_5) \leq 2 \cdot p \cdot exp\left\{-\frac{N_{\min}\epsilon^2}{8(\max(f_j'(m)))^2\log^2\eta}\right\}.$$

(iii) For ζ_4,

$$P(\zeta_4^c \mid \zeta_5) \leq 2 \cdot p \cdot r \cdot exp\left\{-\frac{2N_{\min}\epsilon^2}{\max_{k\in\{1,\ldots,r-1\}} s(r,k)(4\log^2\eta)^r}\right\}.$$

where $N_{\min}$ is a predetermined minimum sample size in Assumption 3.1 (A3) and $s(r,k)$ is Stirling numbers of the first kind.

Proposition 1.3. *Under Assumption 3.1 (A2),*

$$P(\zeta_5^c) \leq \frac{V_1}{\eta^2}.$$

Hence for any $\epsilon \in \left(0, \left|\frac{M_{\min} g_j(\mathbb{E}(X_j|X_S))^2}{(2\mathbb{E}((X_j)_r|X_S)+(2-M_{\min})g_j(\mathbb{E}(X_j|X_S)))}\right|\right)$, the MRS algorithm recovers the true ordering at least of

$$\begin{aligned} P\left(\widehat{\pi} = \pi\right) &\geq 1 - \Big\{ \underbrace{P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4)}_{Prop1.1} + \underbrace{P(\zeta_2^c \mid \zeta_5) + P(\zeta_3^c \mid \zeta_5) + P(\zeta_4^c \mid \zeta_5)}_{Prop1.2} + \underbrace{P(\zeta_5^c)}_{Prop1.3} \Big\}. \\ &= 1 - 2 \cdot p \cdot \exp\left\{ -\frac{2N_{\min}\epsilon^2}{(4\log^2\eta)^r} \right\} - 2 \cdot p \cdot \exp\left\{ -\frac{N_{\min}\epsilon^2}{8(\max(f_j'(m)))^2 \log^2\eta} \right\} \\ &\quad -2 \cdot p \cdot r \cdot \exp\left\{ -\frac{2N_{\min}\epsilon^2}{\max_{k\in\{1,\ldots,r-1\}} s(r,k)(4\log^2\eta)^r} \right\} - \frac{V_1}{\eta^2}. \end{aligned}$$

This result clams that if $N_{\min} = O(\log^{2r}(\eta)\log(p))$, our algorithm correctly estimate the ordering of the graph.

Lastly, we show the relationship between the sample size n and $N_{\min}$ to satisfy Assumption 3.1 (A3). Suppose that d is the maximum number of parents of a node. Then the maximum size of the candidate parents set is d. The scenario is that a conditioning set has two possible cases. If there is only one element for the conditioning set, there is no difference between the conditional and marginal distributions. In the best case, $n = 2N_{\min}$ when the there are two conditional distributions $|\mathcal{X}_C| = 2$. Hence if $n = O((\log^{2r}(\eta)(\log(p) + \log(r)))$, our algorithm works in the high dimensional settings. In the worst case given ζ_5, the sample size is $n = (4\log(\eta)^d - 2)(N_{\min} - 1) + 2N_{\min} = 4\log(\eta)^d(N_{\min} - 1) + 2$ where the number of all elements of $\{x \in \mathcal{X}_C \mid \sum_i^n \mathbf{1}(X_C^{(i)} = x) \geq N_{\min}\}$ is two and all other elements of $\mathcal{X}_C$ has $N_{\min} - 1$ repetitions. In this worst case, if $n = O(\log(\eta)^{(2r+d)}(\log(p) + \log(r)))$ our algorithm correctly recovers the ordering with high probability.

□

1.3.1 Proof for Proposition 1.1

Proof. For ease of notation, let $\eta = \max\{n, p\}$ and the r-th moments ratio score:

$$\widehat{\mathcal{S}}(j,k) := \sum_{x \in \mathcal{X}_{\widehat{C}_{jk}}} \frac{n(x)}{n_{\widehat{C}_{jk}}} \widehat{\mathcal{S}}_r(j,k)(x).$$

In addition, let

$$g_k(\widehat{\mathbb{E}}(X_k \mid X_{\widehat{C}_{jk}})) = f_k(\widehat{\mathbb{E}}(X_k \mid X_{\widehat{C}_{jk}})) - \sum_{m=0}^{r-1} s(r,m)\widehat{\mathbb{E}}(X_k^m \mid X_{\widehat{C}_{jk}}).$$

For any $j \in V$, $k \in \{\pi_j, ..., \pi_p\}$ and $x \in \mathcal{X}_{\widehat{C}_{jk}}$, we have

$$\begin{aligned} & P\left(|\widehat{S}(j,k)(x) - \mathcal{S}(j,k)(x)| > \frac{M_{min}}{2} \Big| \zeta_2, \zeta_3, \zeta_4 \right) \\ = \; & P\left(\left| \frac{\widehat{\mathbb{E}}(X_k^r \mid x)}{g_k(\widehat{\mathbb{E}}(X_k \mid x))} - \frac{\mathbb{E}(X_k^r \mid x))}{g_k(\mathbb{E}(X_k \mid x))} \right| > \frac{M_{min}}{2} \Big| \zeta_2, \zeta_3, \zeta_4 \right) \\ \leq \; & P\left(\frac{\mathbb{E}(X_k^r \mid x) + \epsilon}{g_k(\mathbb{E}(X_k \mid x)) - \epsilon} - \frac{\mathbb{E}(X_k^r \mid x))}{g_k(\mathbb{E}(X_k \mid x))} > \frac{M_{min}}{2} \; or \; \frac{\mathbb{E}(X_k^r \mid x))}{g_k(\mathbb{E}(X_k \mid x))} - \frac{\mathbb{E}(X_k^r \mid x) - \epsilon}{g_k(\mathbb{E}(X_k \mid x)) + \epsilon} > \frac{M_{min}}{2} \right) \\ = \; & P\left(\frac{\epsilon(g_k(\mathbb{E}(X_k \mid x)) + \mathbb{E}(X_k^r \mid x))}{g_k(\mathbb{E}(X_k \mid x))(g_k(\mathbb{E}(X_k \mid x)) - \epsilon)} > \frac{M_{\min}}{2} \; or \; \frac{\epsilon(g_k(\mathbb{E}(X_k \mid x)) + \mathbb{E}(X_k^r \mid x))}{g_k(\mathbb{E}(X_k \mid x))(g_k(\mathbb{E}(X_k \mid x)) + \epsilon)} > \frac{M_{\min}}{2} \right) \\ = \; & P\left(M_{\min} g_k(\mathbb{E}(X_k \mid x))^2 < \epsilon\left(2\mathbb{E}(X_k^r \mid x) + (2 - M_{\min}) g_k(\mathbb{E}(X_k \mid x))\right)\right). \end{aligned}$$

Simple calculations yield that the above upper bound is zero if either

(i) $2\mathbb{E}(X_k^r \mid x) + (2 - M_{\min}) g_k(\mathbb{E}(X_k \mid x)) \leq 0$ or

(ii) $\epsilon < \frac{M_{\min} g_k(\mathbb{E}(X_k|x))^2}{\left(2\mathbb{E}(X_k^r|x) + (2-M_{\min}) g_k(\mathbb{E}(X_k|x))\right)}$.

□

1.3.2 Proof for Proposition 1.2

Since the proof for Prop. 1.2 (i) - (iii) are analogous, we provide the proof for (iii) and then we provide the proof for (ii).

Proof. Using Hoeffding's inequality given ζ_5, for $1 \leq k \leq r$ and any $\epsilon > 0$,

$$P\left(\left|\widehat{\mathbb{E}}(X_j^k \mid X_S) - \mathbb{E}(X_j^k \mid X_S)\right| > \epsilon\right) \leq 2 \cdot p \cdot \exp\left\{-\frac{N_{\min}\epsilon^2}{8\log^{2k}\eta}\right\}.$$

Hence, given ζ_5,

$$\begin{aligned}
& P\left(\left|\sum_{k=0}^{r-1} s(r,k)\widehat{\mathbb{E}}(X_j^k \mid X_S) - \sum_{k=0}^{r-1} s(r,k)\mathbb{E}(X_j^k \mid X_S)\right| > \epsilon \mid \zeta_5\right) \\
\leq & \sum_{k=1}^{r-1} P\left(\left|\widehat{\mathbb{E}}(X_j^k \mid X_S) - \mathbb{E}(X_j^k \mid X_S)\right| > \frac{\epsilon}{s(r,k)} \mid \zeta_5\right) \\
\leq & \sum_{k=1}^{r-1} 2 \cdot p \cdot \exp\left\{-\frac{N_{\min}\epsilon^2}{8s(r,k)\log^{2k}\eta}\right\} \\
\leq & \; 2 \cdot p \cdot r \cdot \exp\left\{-\frac{N_{\min}\epsilon^2}{8\max_k s(r,k)\log^{2r}\eta}\right\}.
\end{aligned}$$

□

Now we provide the proof for (ii).

Proof. By Mean value theorem, we obtain

$$f_j(\widehat{\mathbb{E}}(X_j \mid X_S)) - f_j(\mathbb{E}(X_j \mid X_S)) = f_j'(\bar{m})\left(\widehat{\mathbb{E}}(X_j \mid X_S) - \mathbb{E}(X_j \mid X_S)\right).$$

where f_j' is the first derivative of f_j and $\bar{m}$ is some point between $\widehat{\mathbb{E}}(X_j \mid X_S)$ and $\mathbb{E}(X_j \mid X_S)$.

Given the $|\widehat{\mathbb{E}}(X_j) - \mathbb{E}(X_j)| < \epsilon/2$ from Prop. 1.2(iii), we obtain

$$f_j(\widehat{\mathbb{E}}(X_j \mid X_S)) - f_j(\mathbb{E}(X_j \mid X_S)) = \max_m f_j'(m)\left(\widehat{\mathbb{E}}(X_j \mid X_S) - \mathbb{E}(X_j \mid X_S)\right).$$

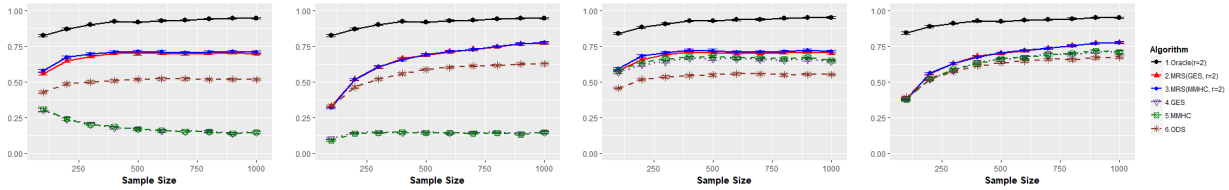
for $m \in (\mathbb{E}(X_j \mid X_S) - \epsilon/2, \mathbb{E}(X_j \mid X_S) + \epsilon/2)$. Again applying Hoeffding's inequality given ζ_5, for any $\epsilon > 0$,

$$\begin{aligned}
P\left(\min_{j\in V} f_j\left(\widehat{\mathbb{E}}(X_j \mid X_S)\right) - f_j\left(\mathbb{E}(X_j)\right) > \epsilon \mid \zeta_5\right) & \leq p \cdot \min_{j\in V} P\left(\left(\widehat{\mathbb{E}}(X_j \mid X_S) - \mathbb{E}(X_j \mid X_S)\right) > \frac{\epsilon}{\max_m f_j'(m)} \mid \zeta_5\right) \\
& \leq 2 \cdot p \cdot \exp\left\{-\frac{N_{\min}\epsilon^2}{8(\max_m f_j'(m))^2\log^2\eta}\right\}.
\end{aligned}$$

□

Distributions	p.g.f. $G(s)$	Parameters
Poisson	${}_0F_0[;;\lambda(s-1)]$	$\lambda > 0$
Hyper-Poisson (Bardwell and Crow)	${}_1F_1[1;b;\lambda(s-1)]$	$\lambda > 0$
Negative Binomial	${}_1F_0[k;;p(s-1)]$	$k, p > 0$
Poisson Beta	${}_1F_1[a;a+b;\lambda(s-1)]$	$a, b, \lambda > 0$
Negative Binomial Beta	${}_2F_1[k,a;a+b;\lambda(s-1)]$	$k, a, b, \lambda > 0$
STERRED Geometric	${}_2F_1[1,1;2;q(s-1)/(1-q)]$	$1 > q > 0$
Shifted UNSTERRED Poisson	${}_1F_1[2;1;\lambda(s-1)]$	$1 \geq \lambda > 0$

Table 1: Examples of hypergeometric distributions and their probability generating functions $G(s)$



(a) DAG Precision:$p = 20$ (b) DAG Recall:$p = 20$ (c) MEC Precision:$p = 20$ (d) MEC Recall:$p = 20$

Figure 1: Comparison of our MRS algorithms using GES and MMHC algorithms in Step 1) and $r = 2$ to the ODS, GES, MMHC algorithms in terms of precision and recall for Poisson and Hybrid DAG models with $p = 20$.

1.3.3 Proof for Proposition 1.3

Proof. The proof is directly from the concentration bound:

$$\begin{aligned}
P(\zeta_5^c) &= P\left(\min_{j\in V}\min_{i\in\{1,2,\dots,n\}}\left|X_j^{(i)}\right| > 4\log\eta\right)\\
&\leq n\cdot p\cdot P\left(P\left|X_j^{(i)}\right| > 4\log\eta\right)\\
&\leq n\cdot p\cdot\frac{\mathbb{E}(\exp(X_j^{(i)}))}{\eta^4}\\
&\leq n\cdot p\cdot\frac{\mathbb{E}(\mathbb{E}(\exp(X_j^{(i)})\mid X_{\mathrm{Pa}(j)}))}{\eta^4}\\
&\overset{(a)}{\leq} n\cdot p\cdot\frac{V_1}{\eta^4}\\
&\leq \frac{V_1}{\eta^2}.
\end{aligned}$$

Inequality (a) is from Assumption 3.1 (A2). □

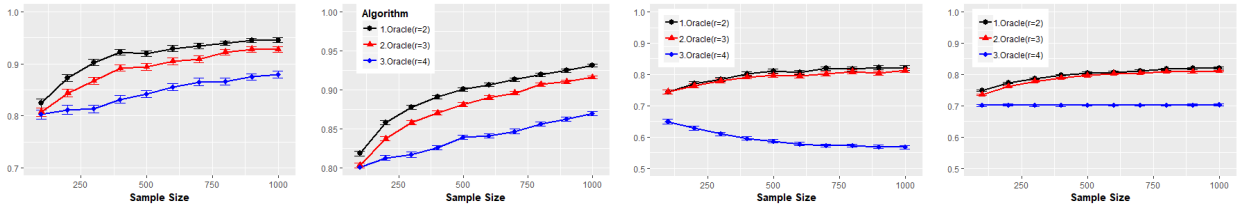
1.4 Examples of Hypergeometric Distributions

We provide examples of hypergeometric distributions and their probability generating functions in Table 1.

1.5 Simulations

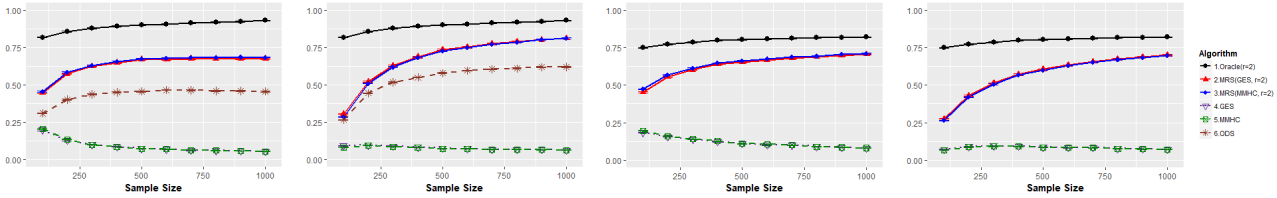
In this section, we provide various simulation results by comparing the MRS algorithm to state-of-the-art ODS, GES and MMHC algorithms in terms of recovering MECs and DAGs.

Not surprisingly, the MRS algorithm performs better than ODS, GES and MMHC algorithms in terms of recovering MECs since the ordering is well estimated by the MRS algorithm. Fig. 1 empirically confirms this in Poisson and Hybrid DAG models with $p = 20$.



(a) Poisson: $p = 20$ (b) Poisson: $p = 100$ (c) Hybrid: $p = 20$ (d) Hybrid: $p = 100$

Figure 2: Comparison of the MRS algorithms using different values of $r = 2, 3, 4$ for the scores in terms of recovering the ordering of Poisson and Hybrid DAG models given the true skeletons.



(a) Poisson: Precision (b) Poisson: Recall (c) Hybrid: Precision (d) Hybrid: Recall

Figure 3: Comparison of our MRS algorithms using GES and MMHC algorithms in Step 1) and $r = 2$ to the ODS, GES, MMHC algorithms in terms of recovering Poisson and Hybrid DAG models with $p = 100$.

To authenticate the validation of Thm. 3.2, we again plot the average precision ($\frac{\#\text{ of correctly estimated edges}}{\#\text{ of estimated edges}}$) as a function of sample size ($n \in \{100, 200, ..., 1000\}$) for different node sizes ($p = \{20, 100\}$) given the true skeleton. As explained with large-scale graphs, Fig. 2 supports our main idea: (i) our algorithm recovers the ordering more accurately as sample size increases; (ii) our algorithm can recover the ordering in high dimensional settings; and (iii) the required sample size $n = \Omega(\log^{2r+d}(\max(n, p))\log(p))$ depends on the choice r because our algorithm with $r = 2$ performs significantly better than our algorithms with $r = 3, 4$. For Hybrid DAG models with $r = 4$, the accuracy seems poor because Binomial with $N = 3$ cannot satisfy Assumption 3.1 (A1) with $r = 4$ i.e., $\mathbb{E}((X_j)_4) = 0$. However the overall precision is significantly better than 0.5 which is the precision of the random graph with the true skeleton.

In Fig. 3, we compare the MRS algorithm where $r = 2$ for the score, and GES and MMHC algorithms are applied in Step 1) to state-of-the art ODS, GES and MMHC algorithms by providing two results as a function of sample size $n \in \{100, 200, ..., 1000\}$ for fixed node size $p = 100$: (i) the average precision ($\frac{\#\text{ of correctly estimated edges}}{\#\text{ of estimated edges}}$); (ii) the average recall ($\frac{\#\text{ of correctly estimated edges}}{\#\text{ of ture edges}}$). We also provide an oracle where the true skeleton is used while the ordering is estimated via the moments ratio scores.

As we explained with DAG models with $p \in \{200, 500\}$ in the main body of the paper, Fig. 3 also provides that the MRS algorithm is more accurate than state-of-the-art ODS, GES and MMHC algorithms in both precision and recall.

1.6 Real Multivariate Count Data: 2009/2010 NBA Player Statistics

The original data set contains 24 covariates: player name, team name, player's position, total minutes played, total number of field goals made, field goals attempted, threes made, threes attempted, free throws made, free throws attempted, offensive rebounds, rebounds, assists, steals, turnovers, blocks, personal fouls, disqualifications, technicals fouls, ejections, flagrant fouls, games started and total points. We eliminated player name, team name, number of games played, and player's position, because our focus is to find the directional or causal relationships between statistics. We also eliminated ejections and flagrant fouls because both did not occur in our data set. Therefore the data set we consider contains 18 discrete variables.

Fig. 4 (left) shows that the magnitude of NBA statistics are significantly different, and hence we expect our MRS algorithm would be more accurate than the comparison ODS algorithm. Moreover, Fig. 4 (right) shows that all 18 variables are positively correlated. This makes sense because the total minutes played is likely to be positively correlated with other statistics, and some statistics have causal or directional relationships (e.g., the

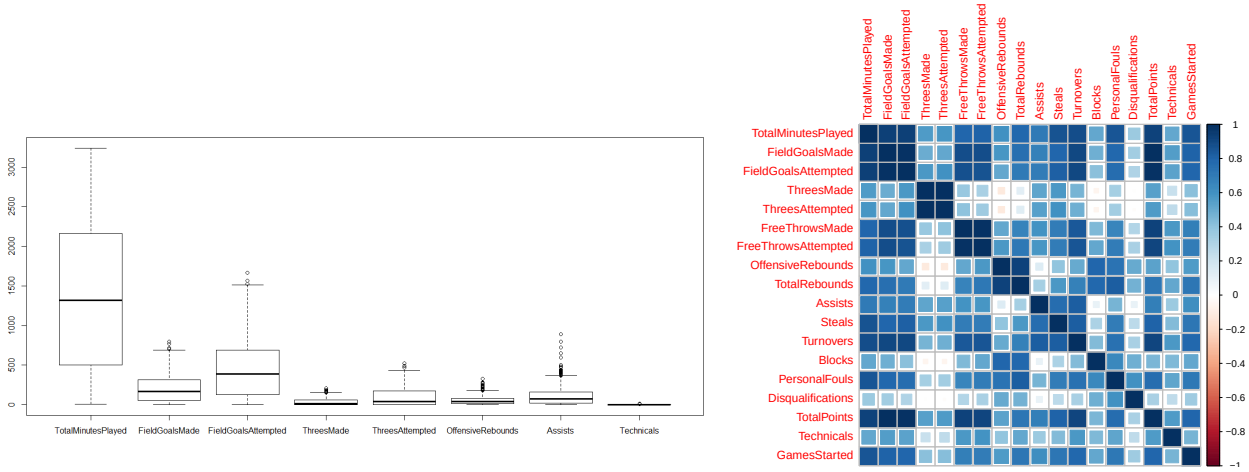
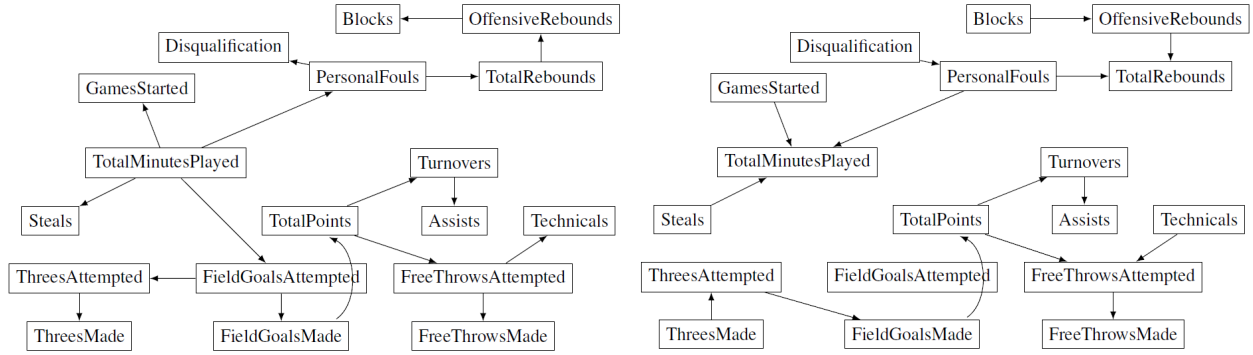


Figure 4: Box plots for some NBA statistics depending on positions (left). Box plots consider the total minutes played, total number of field goals made and attempted, threes made and attempted, offensive rebounds, assists, and technical fouls. Correlation Plots for NBA statistics (right). Blue represents a high correlation and white represents a small correlation.



(a) DAG from MRS

(b) DAG from ODS

Figure 5: NBA players statistics DAG estimated by MRS (left) and DAG estimated by ODS (right).

Explanable edges	TotalMinutesPlayed → PersonalFouls, Steals and GamesStarted, ThreeAttempted → ThreeMade, TotalRebounds → OffensiveRebounds, PersonalFouls → Disqualification
Unexplanable edges	OffensiveRebounds → Blocks, FreeThrowsAttempted → Techincals

Table 2: The set of directed edges in the estimated DAG via the MRS algorithm while the estimated DAG via the ODS algorithm has opposite directions.

more shooting attempt implies the more shooting made).

Here we shortly explain the summary of the data. All basketball statistics have significantly different levels of frequencies because some statistics such as the number of steals, blocks, and technical fouls are close to zero in general while the number of field goals attempted, free throws attempted, threes attempted are large. For example, the averages of the number of field goals attempted and technical fouls are 455.7 and 1.6, and their standard deviations are 373.3 and 2.6, respectively.

Fig. 5 shows the estimated directed graphs using the MRS and ODS algorithms. Explainable edges in Table. 2 shows the directed edges in the estimated DAG from the MRS algorithm while the estimated DAG from the

ODS algorithm has opposite directions. This set of directed edges is more acceptable because the total minutes played would be a reason for other statistics, and a large number of shooting attempted would lead to the more shootings made. Unexplainable edges in Table. 2 shows the set of unaccountable edges in terms of causal or directional relationships regardless of directions. Hence they are introduced by Step 1) estimation of the skeleton.

References

Kemp, A. W. (1968). A wide class of discrete distributions and the associated differential equations. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 401–410.