Identifiability of Generalized Hypergeometric Distribution (GHD) DAG Models

Gunwoong Park, Hyewon Park

University of Seoul, Statistics

2019-01-03

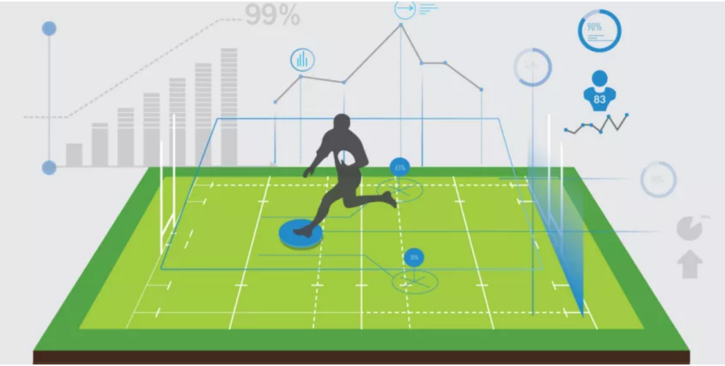
Outline

- Introduction
  - Directed graphical model
- Generalized Hypergeometric Distribution (GHD) DAG Models
  - Identifiability
- Algorithm
  - Statistical Guarantees
- Numerical Experiments

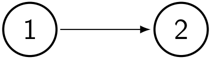
Introduction

Introduction: Sports data

Why do we learn the graphical models?

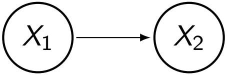


What is the directed graph?



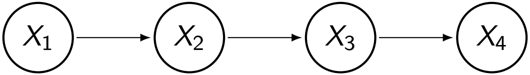
- $G = (V, E)$
- V : A set of nodes, e.g. $V = \{1, 2\}$.
- E : A set of directed edges, e.g. $E = \{(1, 2)\}$.

What is the directed graphical model?



- $X := (X_j)_{j \in V}$, a set of random variables, e.g. $X = \{X_1, X_2\}$.
- X_1 and X_2 are correlated with each other.

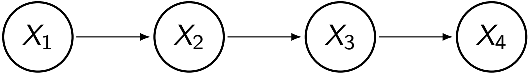
Notations for directed graphical model



- $X_{Pa(j)}$: A parents set of X_j, e.g. $X_{Pa(2)} = \{X_1\}$.
- $X_{De(j)}$: A set of all decendants of X_j, e.g. $X_{De(2)} = \{X_3, X_4\}$
- $X_{Nd(j)}$: A set of non-decendatns of X_j,
  e.g. $X_{Nd(2)} = \{X_1\}$, $X_{Nd(3)} = \{X_1, X_2\}$.

Generalized Hypergeometric Distribution (GHD) DAG Models

Directed Acyclic Graph (DAG)



- A directed graph with no directed cycles.
- DAG model has the factorization,

$$P(G) = P(X_1, X_2, \ldots, X_p) = \prod_{j=1}^{p} P(X_j \mid X_{\mathrm{Pa}(j)}).$$

- In this graph,

$$P(G) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3).$$

Probability generating function

- A power series representation of the probability mass function of the random variable.

$$G(s) = \mathbb{E}(s^x) = \sum_{x=0}^{\infty} p(x)s^x.$$

- Example: A Poisson random variable with rate parameter λ .

$$G(s) = \mathbb{E}(s^x) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} s^x = e^{\lambda(s-1)}.$$

Generalized hypergeometric function

$$ {}_pF_q[a_1, \ldots, a_p; b_1, \ldots, b_q; \theta] := \sum_{j \geq 0} \frac{\langle a_1 \rangle^j \cdots \langle a_p \rangle^j \, \theta^j}{\langle b_1 \rangle^j \cdots \langle b_q \rangle^j \, j!}. $$

- $\langle a \rangle^j = a(a+1)\cdots(a+j-1)$, the rising factorial, e.g. $\langle 2 \rangle^4 = 2(2+1)(2+2)(2+3) = 120$.
- $(a)_j = a(a-1)\cdots(a-j+1)$, the falling factorial, e.g. $(5)_3 = 5(5-1)(5-2) = 60$.
- $\langle a \rangle^0 = (a)_0 = 1$.

Generalized Hypergeometric Distributions (GHD)

- A family of GHDs has a special form of probability generating functions expressed in terms of the generalized hypergeometric series.

$$
\begin{aligned}
G(s) &= {}_pF_q[a_1, \ldots, a_p; b_1, \ldots, b_q; \theta(s-1)] \\
&= \sum_{j \geq 0} \frac{\langle a_1 \rangle^j \cdots \langle a_p \rangle^j}{\langle b_1 \rangle^j \cdots \langle b_q \rangle^j} \frac{(\theta(s-1))^j}{j!}
\end{aligned}
$$

- Example: Poisson distribution

$$
G(s) = e^{\lambda(s-1)} = \sum_{j \geq 0} \frac{(\lambda(s-1))^j}{j!} = {}_0F_0[;;\lambda(s-1)]
$$

Examples of the hypergeometric distribution

Distributions	p.g.f. $G(s)$	Parameters
Poisson	${}_0F_0[;;\lambda(s-1)]$	$\lambda > 0$
Hyper-Poisson	${}_1F_1[1;b;\lambda(s-1)]$	$\lambda > 0$
Binomial	${}_1F_0[-N;;-p(s-1)]$	$N, p > 0$
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	${}_1F_1[2;1;\lambda(s-1)]$	$1 \geq \lambda > 0$

GHD DAG models

Definition: GHD DAG Models

For each $j \in V$, $X_j \mid X_{\mathrm{Pa}(j)}$ has the following probability generating function

$$G(s; a(j), b(j)) = {}_{p_j}F_{q_j}[a(j); b(j); \theta(X_{\mathrm{Pa}(j)})(s-1)]$$

where $a(j) = (a_{j1}, \ldots a_{jp_j})$, $b(j) = (b_{j1}, \ldots b_{jq_j})$, and $\theta : \mathcal{X}_{\mathrm{Pa}(j)} \to \mathbb{R}$.

- The conditional distribution of each node given its parents belongs to a family of GHDs.
- The parameter depends only on its parents.

Proposition of GHD DAG models

Proposition: Constant Moments Ratio (CMR) Property

Consider a GHD DAG model. Then for any $j \in V$ and any integer $r = 2, 3, \ldots,$ there exist a r-th factorial constant moments ratio (CMR) function

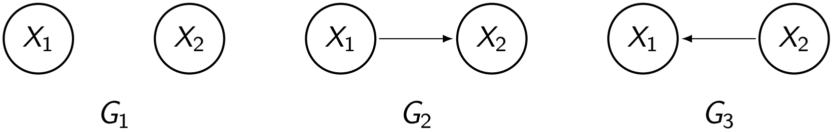
$f_j^{(r)}(x; a(j), b(j)) = x^r \prod_{i=1}^{p_j} \left(\frac{(a_{ji}+r-1)_r}{a_{ji}^r} \right) \prod_{k=1}^{q_j} \left(\frac{b_{jk}^r}{(b_{jk}+r-1)_r} \right)$ such that

$$\begin{aligned}
\mathbb{E}\big((X_j)_r \mid X_{\mathrm{Pa}(j)}\big) &= \mathbb{E}\big(X_j(X_j-1)\cdots(X_j-r+1) \mid X_{\mathrm{Pa}(j)}\big) \\
&= f_j^{(r)}\big(\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)}); a(j), b(j)\big),
\end{aligned}$$

as long as $\max X_j \geq r$.

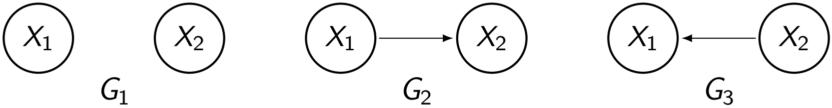
GHD DAG models: Identifiability

Is it possible to recover a graph from distribution?



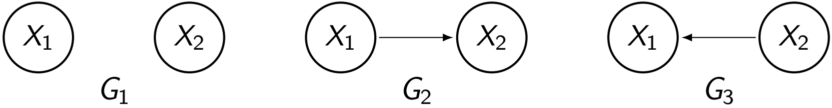
- We can distinguish G_2 and G_3 from G_1.
- How can we know the direction of the edge?
- We exploit the CMR property for model identifiability.

Example : The CMR property for Poisson



- $G_1 : X_1 \sim Poisson(\lambda_1),\ X_2 \sim Poisson(\lambda_2),\ X_1 \perp\!\!\!\perp X_2$.
- $G_2 : X_1 \sim Poisson(\lambda_1),\ X_2 \mid X_1 \sim Poisson(\theta_2(X_1))$.
- $G_3 : X_2 \sim Poisson(\lambda_2),\ X_1 \mid X_2 \sim Poisson(\theta_1(X_2))$.
- $\theta_1, \theta_2 : \mathbb{N} \cup \{0\} \to \mathbb{R}^+$, arbitrary positive functions.

Example : The CMR property for Poisson



- We exploit the CMR property for Poisson, $\mathbb{E}((X_j)_r) = \mathbb{E}(X_j)^r$ for any positive integer $r \in \{2, 3, \ldots\}$.
- For G_1, $\mathbb{E}((X_1)_r) = \mathbb{E}(X_1)^r$ and $\mathbb{E}((X_2)_r) = \mathbb{E}(X_2)^r$.

Example : The CMR property for Poisson

$$X_1 \qquad X_2 \qquad\qquad X_1 \longrightarrow X_2 \qquad\qquad X_1 \longleftarrow X_2$$

$$G_1 \qquad\qquad\qquad G_2 \qquad\qquad\qquad G_3$$

- For G_2, $\mathbb{E}((X_1)_r) = \mathbb{E}(X_1)^r$, while

$$\begin{aligned}\mathbb{E}((X_2)_r) &= \mathbb{E}(\mathbb{E}((X_2)_r \mid X_1)) = \mathbb{E}(\mathbb{E}(X_2 \mid X_1)^r) \\ &> \mathbb{E}(\mathbb{E}(X_2 \mid X_1))^r = \mathbb{E}(X_2)^r, \qquad \text{(by Jensen's inequality)}\end{aligned}$$

  as long as $\mathbb{E}(X_2 \mid X_1)$ is not a constant.

- For G_3, $\mathbb{E}((X_2)_r) = \mathbb{E}(X_2)^r$ and $\mathbb{E}((X_1)_r) > \mathbb{E}(X_1)^r$.
- **Now we can distinguish G_1, G_2 and G_3 by testing whether a moments ratio $\mathbb{E}((X_j)_r)/\mathbb{E}(X_j)^r$ is greater than or equal to 1.**

GHD DAG models: Identifiability

Identifiability

Assumptions: Identifiability Conditions

- For a given GHD DAG model, the conditional distribution of each node given its parents is known.
- For any node $j \in V$, $\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)})$ is non-degenerated.

Theorem: Identifiability

- Under the identifiability conditions, the class of GHD DAG models is identifiable.

GHD DAG models: Identifiability

- For any node $j \in V, \mathbb{E}((X_j)_r) = \mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)}); a(j), b(j)))$
- For any non-empty $\mathrm{Pa}_0(j) \subset \mathrm{Pa}(j)$ and $S_j \subset \mathrm{Nd}(j) \setminus \mathrm{Pa}_0(j)$,

$$\begin{aligned}\mathbb{E}((X_j)_r) &= \mathbb{E}(\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)}); a(j), b(j)) \mid X_{S_j})) \\ &> \mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{S_j}); a(j), b(j))),\end{aligned}$$

  because the CMR function is strictly convex.
- **To search a smallest conditioning set S_j for each node j such that the moments ratio** $\dfrac{\mathbb{E}((X_j)_r)}{\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{S_j})))} = 1.$

Algorithm

Algorithm

The Moments Ratio Scoring (MRS) algorithm consist of two steps.

1. Estimate the skeleton of graph.
    - Any sceleton learning algorithms.
      Examples) GES, MMHC.
2. Find an ordering.
    - Using moments ratio scores.
    - Find an element which has the smallest momet ratio score.

Moments ratio score

- Moments ratio:

$$\frac{\mathbb{E}((X_j)_r)}{\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{S_j})))}.$$

  - If all samples are less than r, the moments ratio is 0.

- Moments ratio score :

$$\widehat{\mathcal{S}}_r(j) := \frac{\widehat{\mathbb{E}}(X_j^r)}{f_j^{(r)}(\widehat{\mathbb{E}}(X_j)) - \sum_{k=0}^{r-1} s(r,k)\widehat{\mathbb{E}}(X_j^k)},$$

where $s(r,k)$ is Stirling numbers of the first kind.

Algorithm: Statistical guarantees

Assumptions

For all $j \in V$, any non-empty $\mathrm{Pa}_0(j) \subset \mathrm{Pa}(j)$, and $S_j \subset \mathrm{Nd}(j) \backslash \mathrm{Pa}_0(j)$,

(A1) There exists a positive constant $M_{\min} > 0$ such that

$$\frac{\widehat{\mathbb{E}}(X_j^r \mid X_{S_j})}{f_j^{(r)}(\widehat{\mathbb{E}}(X_j \mid X_{S_j})) - \sum_{k=0}^{r-1} s(r,k)\widehat{\mathbb{E}}(X_j^k \mid X_{S_j})} > 1 + M_{\min}.$$

(A2) There exist a positive constant V_1 such that

$$\mathbb{E}(\exp(X_j) \mid X_{\mathrm{Pa}(j)}) < V_1.$$

Algorithm: Statistical guarantees

Theorem: Recovery of the ordering

In regularity conditions, there exist constant $C_\epsilon > 0$ for any $\epsilon > 0$ such that if sample size is sufficiently large $n > C_\epsilon \log^{2r+d}(\max(n, p))(\log(p) + \log(r))$, the MRS algorithm with the r-th moments ratio scores recovers the ordering with high probability: $P(\widehat{\pi} \in \mathcal{E}(\pi)) \geq 1 - \epsilon$.

- d : The maximum indegree of the graph.
- $\mathcal{E}(\pi)$: The set of all the orderings that are consistent with the DAG G.
- The MRS algorithm accurately estimates a true ordering with high probability, if $n = \Omega(\log^{2r+d}(\max(n, p)) \log(p))$.

Numerical Experiments

Simulation settings

- Two sets of simulation study using 150 realizations of p-node random GHD DAG models.
  - Poisson DAG models: The conditional distribution of each node given its parents is Poisson.
  - Hybrid DAG models: The conditional distributions are sequentially Poisson, Binomial with $N = 3$, hyper-Poisson with $b = 2$, and Binomial with $N = 3$.
- $d = 2$.

Link functions and parameters

- (Hyper) Poisson prameter :

$$\theta_j(\mathsf{Pa}(j)) = \exp(\theta_j + \sum_{k \in \mathsf{Pa}(j)} \theta_{jk} X_k),$$

$$\theta_j \in [1, 3], \theta_{jk} \in [-1.75, -0.25] \cup [0.25, 1.75].$$

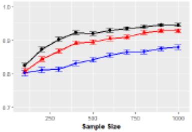
- Binomial probability :

$$p_j(\mathsf{Pa}(j)) = \mathrm{logit}^{-1}(\theta_j + \sum_{k \in \mathsf{Pa}(j)} \theta_{jk} X_k),$$

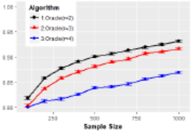
$$\theta_j \in [1, 3], \theta_{jk} \in [-1.2, -0.2].$$

Numerical experiments: Known skeleton

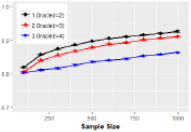
Using different values of $r \in \{2, 3, 4\}$.



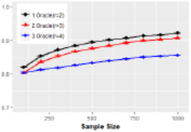
(a) Poisson: $p = 20$



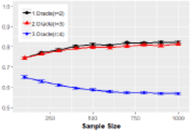
(b) Poisson: $p = 100$



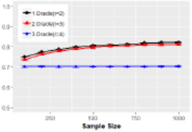
(c) Poisson: $p = 200$



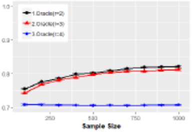
(d) Poisson: $p = 500$



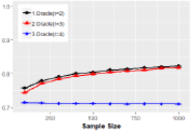
(e) Hybrid: $p = 20$



(f) Hybrid: $p = 100$



(g) Hybrid: $p = 200$

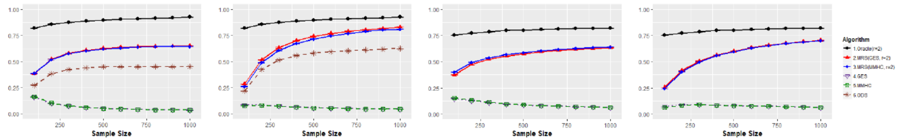


(h) Hybrid: $p = 500$

- The MRS algorithm with $r = 2$ performs better than the MRS algorithms with $r = \{3, 4\}$.

Numerical experiments: Unknown skeleton

Using GES, MMHC algorithms in Step 1, r = 2, p = 200.

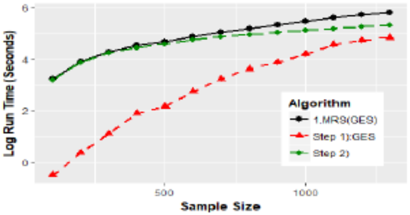


(a) Poisson: Precision (b) Poisson: Recall (c) Hybrid: Precision (d) Hybrid: Recall

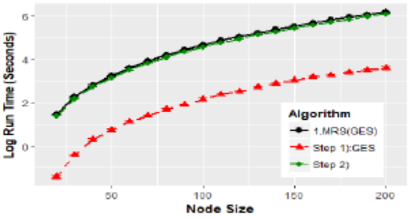
- The MRS algorithm accurately recovers the true directed edges as sample size increases.

Computational complexity

Using GES algorithm in Step 1.



(a) Varying n



(b) Varying p

- (a): $n \in \{100, 200, \ldots 1300\}$, $p = 100$.
- (b): $p \in \{10, 20, \ldots 200\}$, $n = 500$.
- Time complexcity of step 1 is $O(n^2p^2)$.
- Time complexcity of step 2 of MRS algorithm is $O(np^2)$.

Summary

- GHD DAG models.
  - The conditional distribution of each node given its parents belongs to a family of GHDs.
  - The prameter depend only on its parents.
- We can find the ordering of the DAG using moments ratio score.
- The MRS algorithm
  - recovers the ordering more accurately as sample size increases.
  - can recover the ordering in high dimensional settings.
  - The MRS algorithm with $r = 2$ performs better than the MRS algorithms with $r = \{3, 4\}$.

Thank you!