Identifiability of Generalized Hypergeometric Distribution (GHD) Directed Acyclic Graphical Models

Gunwoong Park
Department of Statistics, University of Seoul

Hyewon Park
Department of Statistics, University of Seoul

Abstract

We introduce a new class of identifiable DAG models where the conditional distribution of each node given its parents belongs to a family of generalized hypergeometric distributions (GHD). A family of generalized hypergeometric distributions includes a lot of discrete distributions such as the binomial, Beta-binomial, negative binomial, Poisson, hyper-Poisson, and many more. We prove that if the data drawn from the new class of DAG models, one can fully identify the graph structure. We further present a reliable and polynomial-time algorithm that recovers the graph from finitely many data. We show through theoretical results and numerical experiments that our algorithm is statistically consistent in high-dimensional settings ($p > n$) if the indegree of the graph is bounded, and out-performs state-of-the-art DAG learning algorithms.

1 INTRODUCTION

Probabilistic directed acyclic graphical (DAG) models or Bayesian networks provide a widely used framework for representing causal or directional dependence relationships among many variables. One of the fundamental problems associated with DAG models is learning a causal structure given samples from the joint distribution $P(G)$ over a set of nodes of a graph G.

Prior works have addressed the question of identifiability for different classes of joint distribution $P(G)$. Frydenberg (1990); Heckerman et al. (1995) show the Markov equivalence class (MEC) where graphs that belong to the same MEC have the same conditional independence relations. Chickering (2003); Spirtes et al. (2000); Tsamardinos and Aliferis (2003); Zhang and Spirtes (2016) show that the underlying graph of a DAG model is recoverable up to MEC under the faithfulness or some related conditions. However since many MECs contain more than one graph, a true graph cannot be determined.

Recently, many works show fully identifiable DAG models under stronger assumptions on $P(G)$. Peters and Bühlmann (2014) proves that Gaussian structural equation models with equal or known error variances are identifiable. In addition, Shimizu et al. (2006) shows that linear non-Gaussian models where each variable is determined by a linear function of its parents plus a non-Gaussian error term are identifiable. Hoyer et al. (2009); Mooij et al. (2009); Peters et al. (2012) relax the assumption of linearity and prove that nonlinear additive noise models where each variable is determined by a non-linear function of its parents plus an error term are identifiable under suitable regularity conditions. Instead of considering linear or additive noise models, Park and Raskutti (2015, 2017) introduce discrete DAG models where the conditional distribution of each node given its parents belongs to the exponential family of discrete distributions such as Poisson, binomial, and negative binomial. They prove that the discrete DAG models are identifiable as long as the variance is a quadratic function of the mean.

Learning DAG or causal discovery from *count data* is an important research problem because such count data are increasingly ubiquitous in big-data settings, including high-throughput genomic sequencing data, spatial incidence data, sports science data, and disease incidence data (Inouye et al. 2017). However as we discussed, most existing methods focus on the continuous or limited discrete DAG models. Hence it is important to model complex multivariate count data using a broader family of discrete distributions.

In this paper, we generalize the main idea in Park and Raskutti (2015, 2017) to a *family of generalized hypergeometric distributions (GHD)* that includes Poisson, hyper-Poisson, binomial, negative binomial, beta-

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

binomial, hypergeometric, inverse hypergeometric and many more (see more examples in Dacey 1972; Kemp 1968; Kemp and Kemp 1974 and Supplementary). We introduce a new class of identifiable DAG models where the conditional distribution of each node given its parents belongs to a family of GHDs. In addition, we prove that the class of GHD DAG models is identifiable from the joint distribution $P(G)$ using *convex* relationship between the mean and the r-th factorial moment for some positive integer r under the causal sufficiency assumption that all relevant variables have been observed. However we do not assume the faithfulness assumption that can be very restrictive (Uhler et al. 2013).

We also develop the reliable and scalable Moments Ratio Scoring (MRS) algorithm which learns any large-scale GHD DAG model. We provide computational complexity and statistical guarantees of our MRS algorithm to show that it has polynomial run-time and is consistent for learning GHD DAG models, even in the high-dimensional $p > n$ setting when the indegree of the graph d is bounded. We demonstrate through simulations and a real NBA data that our MRS algorithm performs better than state-of-the-art GES (Chickering 2003), MMHC (Tsamardinos et al. 2006), and ODS (Park and Raskutti 2015) algorithms in terms of both run-time and recovering a graph structure.

The remainder of this paper is structured as follows: Section 2.1 summarizes the necessary notation, Section 2.2 defines GHD DAG models and Section 2.3 proves that GHD DAG models are identifiable. In Section 3, we develop a polynomial-time algorithm for learning GHD DAG models and provide its theoretical guarantees and computational complexity in terms of the triple (n, p, d). Section 4 empirically evaluates our methods compared to GES, MMHC, and ODS algorithms on synthetic and real basketball data.

2 GHD DAG MODELS AND IDENTIFIABILITY

In this section, we first introduce some necessary notations and definitions for directed acyclic graph (DAG) models. Then we propose novel generalized hypergeometric distribution (GHD) DAG models. Lastly, we discuss their identifiability using a convex relation between the mean and r-th factorial moments.

2.1 Problem Set-up and Notation

A DAG $G = (V, E)$ consists of a set of nodes $V = \{1, 2, \cdots, p\}$ and a set of directed edges $E \in V \times V$ with no directed cycles. A directed edge from node j to k is denoted by (j, k) or $j \to k$. The set of *parents* of node k denoted by $\text{Pa}(k)$ consists of all nodes j such that $(j, k) \in E$. If there is a directed path $j \to \cdots \to k$, then k is called a *descendant* of j and j is an *ancestor* of k. The set $\text{De}(k)$ denotes the set of all descendants of node k. The *non-descendants* of node k are $\text{Nd}(k) := V \setminus (\{k\} \cup \text{De}(k))$. An important property of DAGs is that there exists a (possibly non-unique) *ordering* $\pi = (\pi_1, ..., \pi_p)$ of a directed graph that represents directions of edges such that for every directed edge $(j, k) \in E$, j comes before k in the ordering. Hence learning a graph is equivalent to learning an ordering and skeleton that is a set of edges without their directions.

We consider a set of random variables $X := (X_j)_{j \in V}$ with a probability distribution taking values in probability space $\mathcal{X}_v$ over the nodes in G. Suppose that a random vector X has a joint probability density function $P(G) = P(X_1, X_2, ..., X_p)$. For any subset S of V, let $X_S := \{X_j : j \in S \subset V\}$ and $\mathcal{X}(S) := \times_{j \in S} \mathcal{X}_j$. For any node $j \in V$, $P(X_j \mid X_S)$ denotes the conditional distribution of a variable X_j given a random vector X_S. Then, a DAG model has the following factorization (Lauritzen 1996):

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

where $P(X_j \mid X_{\text{Pa}(j)})$ is the conditional distribution of a variable X_j given its parents $X_{\text{Pa}(j)}$.

We suppose that there are n i.i.d samples $X^{1:n} := (X^{(i)})_{i=1}^{n}$ drawn from a given DAG models where $X^{(i)} := (X_1^{(i)}, X_2^{(i)}, \cdots, X_p^{(i)})$ is a p-variate random vector. We use the notation $\widehat{\cdot}$ to denote an estimate based on samples $X^{1:n}$. In addition, we assume the causal sufficiency that all variables have been observed.

2.2 Generalized Hypergeometric Distribution (GHD) DAG models

We begin by introducing a family of generalized hypergeometric distributions (GHDs) defined by Kemp (1968). A family of GHDs includes a large number of discrete distributions and has a special form of probability generating functions expressed in terms of the generalized hypergeometric series. We borrow the notations and terminologies in Kemp and Kemp (1974) to explain detailed properties of a family of GHDs. Let $\langle a \rangle^j = a(a+1)\cdots(a+j-1)$ be the rising factorial, $(a)_j = a(a-1)\cdots(a-j+1)$ be the falling factorial, and $\langle a \rangle^0 = (a)_0 = 1$. In addition, generalized hypergeometric function is:

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

Kemp (1968); Kemp and Kemp (1974) show that GHDs have probability generating functions of the following form:

$$G(s \mid a, b) = \; _pF_q[a_1, ..., a_p; b_1, ..., b_q; \theta(s-1)].$$

This class of distributions includes a lot of discrete distributions such as the binomial, beta-binomial, Poisson, Poisson type, displaced Poisson, hyper-Poisson, logarithmic, and generalized log-series. We provide more examples with their probability generating functions in Supplementary (see also in Dacey 1972; Kemp 1968; Kemp and Kemp 1974).

Now we define the generalized hypergeometric distribution (GHD) DAG models:

Definition 2.1 (GHD DAG Models)**.** *The DAG models belong to generalized hypergeometric distribution (GHD) DAG models if the conditional distribution of each node given its parents belongs to a family of generalized hypergeometric distributions and the parameter depend only on its parents: For each $j \in V$, $X_j \mid X_{Pa(j)}$ has the following probability generating function*

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

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

A popular example of GHD DAG models is a Poisson DAG model in Park and Raskutti (2015) where a conditional distribution of each node $j \in V$ given its parents is Poisson and the rate parameter is an arbitrary positive function $\theta_j(X_{\mathrm{Pa}(j)})$. Unlike Poisson DAG models, GHD DAG models are hybrid models where the conditional distributions have various distributions which incorporate different data types. In addition, the exponential family of discrete distributions discussed in Park and Raskutti (2017) is also included in a family of GHDs. Hence, our class of DAG models is strictly broader than the previously studied identifiable DAG models for multivariate count data.

GHD DAG models have a lot of useful properties for identifying a graph structure. One of the useful properties is the recurrence relation involving factorial moments:

Proposition 2.2 (CMR Property)**.** *Consider a GHD DAG model. Then for any $j \in V$ and any integer $r = 2, 3, ...,$ there exists 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*

$$\mathbb{E}\big((X_j)_r \mid X_{Pa(j)}\big) = f_j^{(r)}\big(\mathbb{E}(X_j \mid X_{Pa(j)}); a(j), b(j)\big).$$

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

The detail of the proof is provided in Supplementary. Prop. 2.2 claims that the GHD DAG models always satisfy the *r-th constant moments ratio (CMR) property* that the r-th factorial moment is a function of the mean. The condition $\max X_j \geq r$ for $r \geq 2$ rules out DAG models with Bernoulli and multinomial distributions which are known to be non-identifiable (Heckerman et al. 1995). We will exploit the CMR property for model identifiability in the next section.

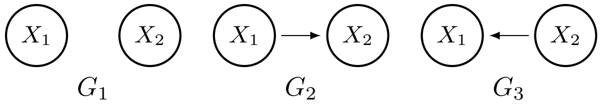


Figure 1: Bivariate DAGs of G_1, G_2 and G_3

2.3 Identifiability

In this section we prove that GHD DAG models are identifiable. To provide intuition, we show identifiability for the bivariate Poisson DAG model discussed in Park and Raskutti (2015). Consider all possible graphical models illustrated in Fig. 1: $G_1 : X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, where X_1 and X_2 are independent; $G_2 : X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \mid X_1 \sim \text{Poisson}(\theta_2(X_1))$; and $G_3 : X_2 \sim \text{Poisson}(\lambda_2)$ and $X_1 \mid X_2 \sim \text{Poisson}(\theta_1(X_2))$ for arbitrary positive functions $\theta_1, \theta_2 : \mathbb{N} \cup \{0\} \rightarrow \mathbb{R}^+$. Our goal is to determine whether the underlying graph is G_1, G_2 or G_3 from the probability distribution $P(G)$.

We exploit the CMR property for Poisson, $\mathbb{E}((X_j)_r) = \mathbb{E}(X_j)^r$ for any positive integer $r \in \{2, 3, ...\}$. For G_1, $\mathbb{E}((X_1)_r) = \mathbb{E}(X_1)^r$ and $\mathbb{E}((X_2)_r) = \mathbb{E}(X_2)^r$. For G_2, $\mathbb{E}((X_1)_r) = \mathbb{E}(X_1)^r$, while

$$\begin{aligned}\mathbb{E}((X_2)_r) &= \mathbb{E}(\mathbb{E}((X_2)_r \mid X_1)) = \mathbb{E}(\mathbb{E}(X_2 \mid X_1)^r) \\ &> \mathbb{E}(\mathbb{E}(X_2 \mid X_1))^r = \mathbb{E}(X_2)^r,\end{aligned}$$

as long as $\mathbb{E}(X_2 \mid X_1)$ is not a constant. The inequality follows from the Jensen's inequality.

Similarly for G_3, $\mathbb{E}((X_2)_r) = \mathbb{E}(X_2)^r$ and $\mathbb{E}((X_1)_r) > \mathbb{E}(X_1)^r$ as long as $\mathbb{E}(X_1 \mid X_2)$ is not a constant. Hence we can distinguish graphs 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.

Now we state the identifiability condition for the general case of p-variate GHD DAG models:

Assumption 2.3 (Identifiability Condition)**.** *For a given GHD DAG model, the conditional distribution of each node given its parents is known. In other words, the r-th factorial CMR functions $(f_j^{(r)}(x; a(j), b(j)))_{j \in V}$ are known. Moreover, for any node $j \in V$, $\mathbb{E}(X_j \mid X_{Pa(j)})$ is non-degenerated.*

Prop. 2.2 and Assumption 2.3 enable us to use the following property: for any node $j \in V$, $\mathbb{E}((X_j)_r) =$

$\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{\text{Pa}(j)}); a(j), b(j)))$, while for any non-empty $\text{Pa}_0(j) \subset \text{Pa}(j)$ and $S_j \subset \text{Nd}(j) \setminus \text{Pa}_0(j)$,

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

because the CMR function is strictly convex.

We state the first main result that general p-variate GHD DAG models are identifiable:

Theorem 2.4 (Identifiability)**.** *Under Assumption 2.3, the class of GHD DAG models is identifiable.*

We defer the proof in Supplementary. The key idea of the identifiability is to search a smallest conditioning set S_j for each node j such that the moments ratio $\mathbb{E}((X_j)_r)/\mathbb{E}(f_j^{(r)}(\mathbb{E}(X_j \mid X_{S_j}))) = 1$. Thm. 2.4 claims that the assumption on nodes distributions is sufficient to uniquely identify GHD DAG models. In other words, the well-known assumptions such as faithfulness, non-linear causal relation, non-Gaussian additive noise assumptions are not necessary (Hoyer et al. 2009; Mooij et al. 2009; Peters and Bühlmann 2014; Peters et al. 2012; Shimizu et al. 2006).

Thm. 2.4 implies that Poisson DAG models are identifiable even when the form of rate parameter functions θ_j are unknown because the model assumes all node (conditional) distributions are Poisson. Thm. 2.4 also claims that hybrid DAG models, in which the distributions of nodes are different, are identifiable as long as the distributions are known while the forms of parameter functions are unknown. In Section 4, we provide numerical experiments on Poisson and hybrid DAG models to support Thm. 2.4.

3 ALGORITHM

In this section, we present our Moments Ratio Scoring (MRS) algorithm for learning GHD DAG models. Our MRS algorithm has two main steps: 1) identifying the skeleton (i.e., edges without their directions) using existing skeleton learning algorithms; and 2) estimating the ordering of the DAG using moments ratio scores, and assign the directions to the estimated skeleton based on the estimated ordering.

Although GHD DAG models can be recovered only using the r-th CMR property according to Thm. 2.4, our algorithm exploits the skeleton to reduce the search space of DAGs. From the idea of constraining the search, our algorithm achieves computational and statistical improvements. More precisely, Step 1) provides *candidate parents* set for each node. The concept of candidate parents set exploits two properties; (i) the neighborhood of a node j in the graph denoted by $\mathcal{N}(j) := \{k \in V \mid (j,k) \text{ or } (k,j) \in E\}$ is a superset of its parents, and (ii) a node j should appear later than its parents in the ordering. Hence, the candidate parents set for a node j is the intersection of its neighborhood and elements of the ordering which appear before that node j, and is denoted by $C_{mj} := \mathcal{N}(j) \cap \{\pi_1, \pi_2, ..., \pi_{m-1}\}$ where m^{th} element of the ordering is j (i.e., $\pi_m = j$). The estimated candidate parents set is $\widehat{C}_{mj} := \widehat{\mathcal{N}}(j) \cap \{\widehat{\pi}_1, \widehat{\pi}_2, ..., \widehat{\pi}_{m-1}\}$ that is specified in Alg.1

This candidate parents set is used as a conditioning set for a moments ratio score in Step 2). If the idea of candidate parents set is not applied, the size of the conditioning set for a moments ratio score could be $p-1$. Since Step 2) computes conditional moments, the sample complexity depends significantly on the number of variables we condition on as illustrated in Section 3.2. Therefore by making the conditioning set for a moments ratio score of each node as small as possible, we gain huge statistical improvements.

The idea of reducing the search space of DAGs has been studied in many sparse candidate algorithms (Zhang and Hyvärinen 2009; Hyvärinen and Smith 2013). Hence for Step 1) of our algorithm, any off-the-shelf candidate parents set learning algorithms can be applied such as MMPC (Tsamardinos and Aliferis 2003). Moreover, any standard MEC learning algorithms such as PC, GES, and MMHC can be exploited because MEC provides the skeleton of a graph (Verma and Pearl 1992). In Section 4, we provide the simulation results of the MRS algorithm where GES and MMHC algorithms are applied in Step 1).

Step 2) of the MRS algorithm involves learning the ordering by comparing moments ratio scores of nodes using Eqn. (1). The ordering is determined one node at a time by selecting the node with the smallest moments ratio score because the correct element of the ordering has the score 1, otherwise strictly greater than 1 in population.

Regarding the moments ratio scores, the score can be exploited for recovering the ordering only if the CMR property holds, which implies that the score should not be zero. Even if the zero value score is impossible in population, zero value scores often arise for a low count data such that all samples are less than r. Hence in order to avoid zero value scores due to a sample r-th factorial moment (i.e., $\widehat{\mathbb{E}}((X)_r) = 0$), we use an alternative ratio $\mathbb{E}(X^r)/\big(f^{(r)}(\mathbb{E}(X)) - \sum_{k=0}^{r-1} s(r,k)\mathbb{E}(X^k)\big)$ where $s(r,k)$ is Stirling numbers of the first kind. This alternative ratio score comes from $(x)_r = \sum_{k=0}^{r} s(r,k)x^k$, therefore $\mathbb{E}(X^r) = f^{(r)}(\mathbb{E}(X)) - \sum_{k=0}^{r-1} s(r,k)\mathbb{E}(X^k)$.

Hence the moments ratio scores in Step 2) of Alg.1 involve the following equations:

Algorithm 1: Moments Ratio Scoring

```
Input  : $n$ i.i.d. samples, $X^{1:n}$
Output : Estimated ordering $\widehat{\pi}$ and an edge
         structure, $\widehat{E} \in V \times V$
Step 1: Estimate the skeleton of the graph $\widehat{\mathcal{N}}$ ;
Step 2: Estimate an ordering of the graph using
  r-th moments ratio scores;
Set $\widehat{\pi}_0 = \emptyset$;
for $m = \{1, 2, \cdots, p-1\}$ do
    for $j \in \{1, 2, \cdots, p\} \setminus \{\widehat{\pi}_1, \cdots, \widehat{\pi}_{m-1}\}$ do
        Find candidate parents set
          $\widehat{C}_{mj} = \widehat{\mathcal{N}}(j) \cap \{\widehat{\pi}_1, \cdots, \widehat{\pi}_{m-1}\}$;
        Calculate r-th moments ratio scores
          $\widehat{\mathcal{S}}_r(m, j)$ using (1);
    end
    The $m^{th}$ element of the ordering
      $\widehat{\pi}_m = \arg\min_j \widehat{\mathcal{S}}(m, j)$;
end
The last element of the ordering
  $\widehat{\pi}_p = \{1, 2, \cdots, p\} \setminus \{\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \widehat{\pi}_{p-1}\}$;
Return : Estimate the edge sets:
  $\widehat{E} = \cup_{m \in V}\{(k, \widehat{\pi}_m) \mid k \in \widehat{\mathcal{N}}(\widehat{\pi}_m) \cap (\widehat{\pi}_1, ..., \widehat{\pi}_{m-1})\}$
```

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

$$\widehat{\mathcal{S}}_r(m, j) := \sum_{x \in \mathcal{X}_{\widehat{C}_{mj}}} \frac{n_{\widehat{C}_{mj}}(x)}{n_{\widehat{C}_{mj}}} \widehat{\mathcal{S}}_r(m, j)(x)$$

$$\widehat{\mathcal{S}}_r(m, j)(x) := \frac{\hat{\mu}^r_{j|\widehat{C}_{mj}}(x)}{f_j^{(r)}(\hat{\mu}^1_{j|\widehat{C}_{mj}}(x)) - \sum_{k=0}^{r-1} s(r,k)\hat{\mu}^k_{j|\widehat{C}_{mj}}(x)}$$

where $\widehat{C}_{mj}$ is the estimated candidate parents set of node j for the m^{th} element of the ordering and $\hat{\mu}^k_{j|S}(x_S) := \widehat{\mathbb{E}}(X_j^k \mid X_S = x_S)$. In addition, $n(x_S) := \sum_{i=1}^n \mathbf{1}(X_S^{(i)} = x_S)$ if $n(x_S) \geq N_{\min}$ otherwise 0, that refers to the truncated conditional sample size for x_S, and $n_S := \sum_{x_S} n(x_S)$ refers to the total truncated conditional sample size for variables X_S. Lastly, we use the method of moments estimators $\widehat{\mathbb{E}}(X_j^k) = \frac{1}{n}\sum_{i=1}^n((X_j^{(i)})^k)$ as unbiased estimators.

Since there are many conditional distributions, our moments ratio score is the weighted average of the levels of how well each distribution satisfies the r-th CMR property. The score only contains the conditional expectations with $n(x_S) \geq N_{\min}$ for better accuracy because the accuracy of the estimation of a conditional expectation $\widehat{E}(X_j \mid x_S)$ relies on the sample size.

Finally, a directed graph is estimated combining the estimated skeleton from Step 1) and the estimated ordering from Step 2) that is $\widehat{E} := \cup_{j \in V}\{(k, \widehat{\pi}_j) \mid k \in \widehat{\mathcal{N}}(\widehat{\pi}_j) \cap (\widehat{\pi}_1, \widehat{\pi}_2, ..., \widehat{\pi}_{j-1})\}$.

3.1 Computational Complexity

The MRS algorithm uses any skeleton learning algorithms with known computational complexity for Step 1). Hence we first focus on our novel Step 2) of the MRS algorithm. In Step 2), there are $(p-1)$ iterations and each iteration has a number of moments ratio scores to be computed which is bounded by $O(p)$. Hence the total number of scores to be calculated is $O(p^2)$. The computation time of each score is proportional to the sample size n, the complexity is $O(np^2)$.

The total computational complexity of the MRS algorithm depends on the choice of the algorithm in Step 1). Since learning a DAG model is NP-hard (Chickering et al. 1994), many state-of-the-art DAG learning algorithms such as PC (Spirtes et al. 2000), GES (Chickering 2003), MMHC (Tsamardinos et al. 2006), and GDS (Peters and Bühlmann 2014) are inherently heuristic algorithms. Although these algorithms take greedy search strategies, the computational complexities of greedy search based GES and MMHC algorithms are empirically $O(n^2p^2)$. In addition, PC algorithm runs in the worst case in exponential time. Hence, Step 2) may not the main computational bottleneck of the MRS algorithm. In Section 4, we compare the MRS to GES algorithm in terms of log run-time, and show that the addition of estimation of ordering does not significantly add to the computational bottleneck.

3.2 Statistical Guarantees

The MRS algorithm exploits well-studied existing algorithms for Step 1). Hence, we focus on theoretical guarantees for Step 2) of the MRS algorithm given that the skeleton is correctly estimated in Step 1). The main result is expressed in terms of the triple (n, p, d) where n is a sample size, p is a graph node size, and d is the indegree of a graph. Lastly, we discuss the sufficient conditions for recovering the graph via the MRS algorithm according to the chosen skeleton learning algorithm for Step 1).

We begin by discussing three required conditions that the MRS algorithm recovers the ordering of a graph.

Assumption 3.1. *Consider the class of GHD DAG models with r-th factorial CMR function $f_j^{(r)}$ specified in Prop. 2.2. For all $j \in V$, any non-empty $Pa_0(j) \subset Pa(j)$, and $S_j \subset Nd(j) \setminus Pa_0(j)$,*

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

$$\mu_{j|S_j} / \big(f_j^{(r)}(\mu_{j|S_j}) - \sum s(r,k)\mu^k_{j|S_j}\big) > 1 + M_{\min}$$

(A2) there exists a positive constant V_1 such that

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

(A3) there are some elements $x_{S_j} \in \mathcal{X}_{S_j}$ such that $\sum_{i=1}^{n} \mathbf{1}(X_{S_j}^{(i)} = x_{S_j}) \geq N_{\min}$ where $N_{\min} > 0$ is the predefined minimum sample size in the MRS algorithm.

The first condition is a stronger version of Assumption 2.3 since we move from the population to the finite sample setting. The second assumption is to control the tail behavior of the conditional distribution of each variable given its parents. It enables to control the accuracy of moments ratio scores (1) in high dimensional settings ($p > n$). The last assumption ensures that the score can be calculated.

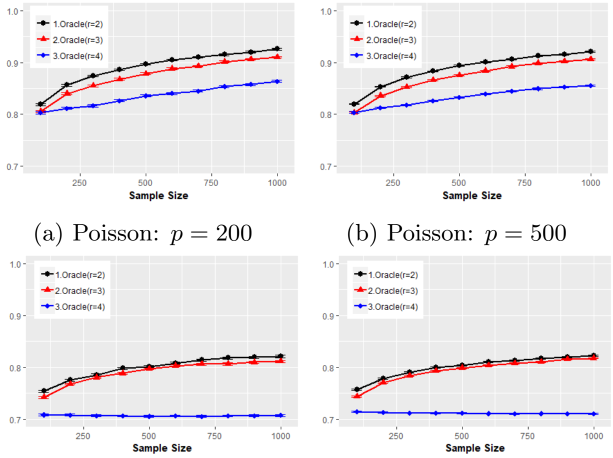
We now state the second main result under Assumption 3.1. Since the true ordering π is possibly not unique, we use $\mathcal{E}(\pi)$ to denote the set of all the orderings that are consistent with the DAG.

Theorem 3.2 (Recovery of the ordering)**.** *Consider a GHD DAG model where the conditional distribution of each node given its parents is known. Suppose that the skeleton of the graph is provided, the maximum indegree of the graph is d, and Assumptions 3.1(A1)-(A3) are satisfied. Then there exists 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: $P(\widehat{\pi} \in \mathcal{E}(\pi)) \geq 1 - \epsilon$.*

The detail of the proof is provided in Supplementary. Intuitively, it makes sense because the method of moment estimator converges to the true moment as sample size n increases. This allows the algorithm to recover a true ordering for the DAG G consistently.

Thm. 3.2 claims that if the sample size $n = \Omega(\log^{2r+d}(\max(n,p))\log(p))$, our MRS algorithm accurately estimates a true ordering with high probability. Hence our MRS algorithm works in high-dimensional settings ($p > n$) provided that the indegree of the graph d is bounded. This theoretical result is also consistent with learning Poisson DAG models shown in Park and Raskutti (2015) where if $n = \Omega(\log^{4+d}(\max(n,p))\log(p))$ their algorithm recovers the ordering well. Since Park and Raskutti (2015) uses the variance (the second order moments $r = 2$), both algorithms are expected to have the same performance of recovering graphs.

However the MRS algorithm performs better than the ODS algorithm in general because the moments difference the ODS algorithm exploits is proportional to magnitude of the conditional mean while the moments ratio is not. For a simple Poisson DAG $X_1 \to X_2$, $\mathbb{E}((X_2)_2) - \mathbb{E}(X_2)^2 = \mathrm{Var}(E(X_2 \mid X_1))$. Hence if $\mathbb{E}(X_2 \mid X_1) \approx 0$, the score in ODS is inevitably close to 0, while the score in MRS, $\mathbb{E}((X_2)_2)/\mathbb{E}(X_2)^2 = 1 + \mathrm{Var}(\mathbb{E}(X_2 \mid X_1))/\mathbb{E}(X_2)^2$ is not necessarily close to 1. Hence, Assumption 3.1(A1) is much milder than the related assumption for the ODS algorithm.



(a) Poisson: $p = 200$ (b) Poisson: $p = 500$

(c) Hybrid: $p = 200$ (d) Hybrid: $p = 500$

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.

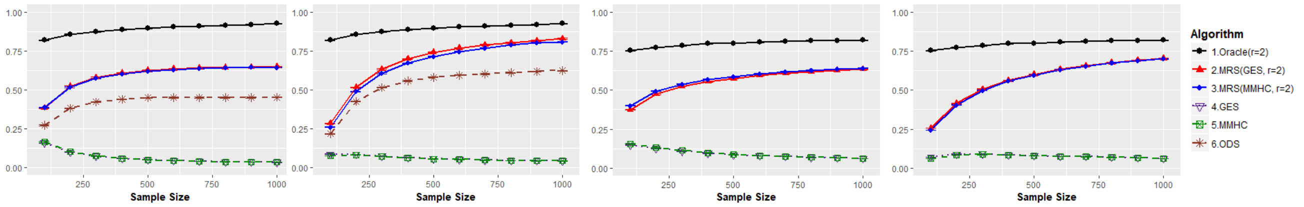
Now we discuss the sufficient conditions for recovering the true graph via the MRS algorithm according to the choice of the algorithm in Step 1). The PC, GES, and MMHC algorithms require the Markov, faithfulness, and causal sufficiency or related assumptions to recover the skeleton of a graph. Moreover GES, MMHC algorithms are greedy search based algorithms that are not guaranteed to recover the true skeleton of a graph. Therefore, the MRS algorithm may require strong assumptions or large sample size to recover the true graph based on the choice of the algorithm in Step 1). Although these assumptions can be very restrictive, we show through the simulations that MRS recovers the directed edges well even in high dimensional settings.

4 NUMERICAL EXPERIMENTS

In this section, we support our theoretical results in Thm. 3.2 and computational complexity in Section 3.1 with synthetic and real basketball data. In addition, we show that our algorithm performs favorably compared to the ODS, GES, and MMHC algorithms in terms of recovering the directed graphs.

4.1 Synthetic Data

Simulation Settings: We conduct two sets of simulation study using 150 realizations of p-node random GHD DAG models with the indegree constraints $d = 2$: (1) Poisson DAG models where the conditional distribution of each node given its parents is Poisson; and



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

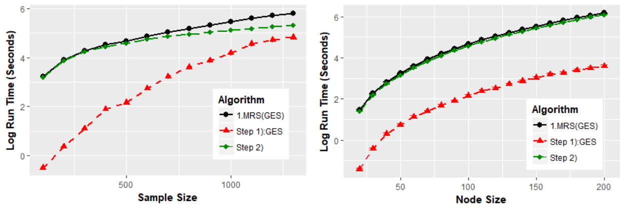
Figure 3: Comparison of our MRS algorithms using GES and MMHC algorithms in Step 1) and $r = 2$ to the ODS, GES, MMHC algorithms in terms of recovering Poisson and Hybrid DAG models with $p = 200$.

(2) Hybrid DAG models where the conditional distributions are sequentially Poisson, Binomial with $N = 3$, hyper-Poisson with $b = 2$, and Binomial with $N = 3$.

We set the (hyper) Poisson rate parameter $\theta_j(\text{Pa}(j)) = \exp(\theta_j + \sum_{k \in \text{Pa}(j)} \theta_{jk} X_k)$ and the binomial probability $p_j(\text{Pa}(j)) = \text{logit}^{-1}(\theta_j + \sum_{k \in \text{Pa}(j)} \theta_{jk} X_k)$. The set of non-zero parameters $\theta_{jk} \in \mathbb{R}$ were generated uniformly at random in the range $\theta_{jk} \in [-1.75, -0.25] \cup [0.25, 1.75]$ and $\theta_j \in [1, 3]$ for Poisson, and $\theta_{jk} \in [-1.2, -0.2]$ and $\theta_j \in [1, 3]$ for Hybrid DAG models. These ranges help the generated values of samples to avoid either all zeros (constant) or too large ($> 10^{309}$). However if some samples are all zeros or too large, we regenerate parameters and samples. We also set the $r \in \{2, 3, 4\}$ and $N_{\min} = 1$ for computing the r-th moments ratio scores. More simulation results with different settings are provided in Supplementary.

Simulation Results: In order to authenticate the validation of Thm. 3.2, we 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 ($p = \{200, 500\}$) given the true skeleton. Fig. 2 provides a comparison of how accurately our MRS algorithm performs in terms of recovering the orderings of the GHD DAG models. Fig. 2 supports our main theoretical results in Thm. 3.2: (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 precision seems not to increase as sample size increases. It makes sense because Binomial with $N = 3$ cannot satisfy the CMR property 2.2 and Assumption 3.1 (A1) with $r = 4$ i.e., $\mathbb{E}((X_j)_4) = 0$. However the precision 0.7 is significantly better than 0.5 which is the precision of the graph with a random ordering.

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 $p = 200$: (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.



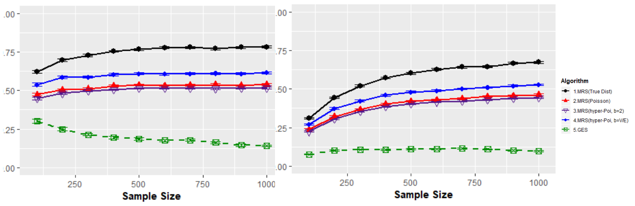
(a) Varying n (b) Varying p

Figure 4: Log run-time of the MRS algorithm using GES algorithm in Step 1) for learning Poisson DAG models with respect to (a) $n \in \{100, 200, ..., 1300\}$ with $p = 100$, and (b) $p \in \{10, 20, ..., 200\}$ with $n = 500$.

As we see in Fig. 3, the MRS algorithm accurately recovers the true directed edges as sample size increases. However since the skeleton estimation is not perfect, we can see the performances of our MRS algorithms using GES and MMHC in Step 1) are significantly worse than the oracle.

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. It makes sense because the moments ratio scores the MRS algorithm exploits are less sensitive to the magnitude of the moments than the score the ODS algorithm uses as discussed in Section 3.2, and because the GES and MMHC algorithms recover up to the MEC by leaving some arrows undirected. However it must be pointed out that our MRS algorithm apply to GHD DAG models while GES and MMHC apply to general classes of DAG models.

Computational Complexity: To validate the computational complexity discussed in Section 3.1, we show the log run-time of Step 1) and Step 2) of the MRS algorithm in Fig. 4 where the GES is applied for Step 1). We measured the run-time for learning Poisson DAG models by varying (a) $n \in \{100, 200, ..., 1300\}$ with the



(a) Hybrid: Precision (b) Hybrid: Recall

Figure 5: Comparison of the MRS algorithms with the different assumed node conditional distribution and the GES algorithm in terms of recovering Hybrid DAG models with $p = 20$.

fixed node size $p = 100$, and (b) $p \in \{10, 20, ..., 200\}$ with the fixed sample size $n = 500$. As we see in Fig. 4, the time complexity of Step 1) is $O(n^2p^2)$, and that of Step 2) of the MRS algorithm is $O(np^2)$. Hence we confirm the addition of estimation of ordering does not significantly add to the computational bottleneck.

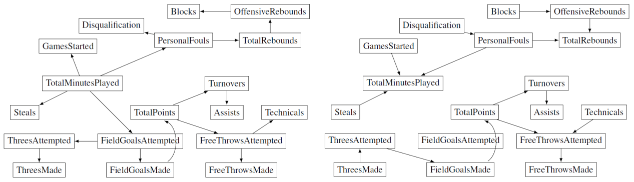
Deviations from True Distributions: When the data are generated by a GHD DAG model where the conditional distribution of each node given its parents is unknown, our algorithm is not guaranteed to estimate the true graph and its ordering. Therefore, an important question is how well the MRS algorithm recovers graphs when incorrect distributions are used. In this section, we heuristically investigate this question.

We use the same setting of the data generation for Hybrid GHD DAG models with the node size $p = 20$. We consider (i) the true (conditional) distributions, and assume all nodes (conditional) distributions are either (ii) Poisson; (iii) hyper-Poisson with $b = 2$; or (iv) hyper-Poisson with $b = \widehat{\text{Var}}(X)/\widehat{\mathbb{E}}(X)$ that is an estimator for the hyper-Poisson parameter b. We compare the MRS and GES algorithms by varying sample size $n \in \{100, 200, ..., 1000\}$ in Fig. 5.

Fig. 5 shows that the MRS algorithms recover the true graph better as sample size increases although there is no theoretical guarantees. It shows that the MRS algorithm enables to learn a part of ordering even if the true (conditional) distributions are unknown as long as there are sufficient samples.

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

We demonstrate the advantages of our graphical models for count-valued data by learning 441 NBA player statistics from season 2009/2010 (see R package SportsAnalytics for detailed information). We consider 18 discrete variables: 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, games started and total points. We provide the procedure of data preprocessing and the detailed summary of data in Supplementary.



(a) DAG from MRS (b) DAG from ODS

Figure 6: NBA players statistics DAG estimated by MRS (left) and DAG estimated by ODS (right).

The MRS and ODS algorithms are applied where GES algorithm is used in Step 1). We assume that the conditional distribution of each node given its parents is hyper-Poisson because most of NBA statistics we consider are the number of successes or attempts counted in the season. We emphasize that our method requires a known conditional distribution assumption to recover the true graph. However since we do not have prior node distribution information, we set $b_j = \widehat{\text{Var}}(X_j)/\widehat{\mathbb{E}}(X_j)$ as we used in simulations that enables the MRS algorithm successfully recovers the directed edges.

Fig. 6 shows the estimated directed graphs using the MRS and ODS algorithms. There are 8 distinct directed edges in the estimated DAG from the MRS algorithm while the estimated DAG from the ODS algorithm has opposite directions: TotalMinutesPlayed → PersonalFouls, Steals, and GamesStarted, ThreeAttempted → ThreeMade, TotalRebounds → OffensiveRebounds, OffensiveRebounds → Blocks, FreeThrowsAttempted → Technicals, and PersonalFouls → Disqualification. The connections between rebounds and blocks, and shooting attempted and technicals do not makes sense in both directions, and hence they might be incorrectly estimated edges in Step 1).

However the remaining 6 directed edges are better explainable 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. It is consistent to our main point that MRS algorithm provides more legitimate directed edges than the ODS algorithm by allowing a broader class of count distributions.

5 Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1C1B5085420).

References

Chickering, D. M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554.

Chickering, D. M., Geiger, D., Heckerman, D., et al. (1994). Learning bayesian networks is np-hard. Technical report, Citeseer.

Dacey, M. F. (1972). A family of discrete probability distributions defined by the generalized hypergeometric series. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 243–250.

Friedman, N., Nachman, I., and Peér, D. (1999). Learning bayesian network structure from massive datasets: the sparse candidate algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*, pages 333–353.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152.

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398.

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.

Kemp, A. W. and Kemp, C. (1974). A family of discrete distributions defined via their factorial moments. *Communications in Statistics-Theory and Methods*, 3(12):1187–1196.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pages 745–752. ACM.

Park, G. and Raskutti, G. (2015). Learning large-scale poisson dag models based on overdispersion scoring. In *Advances in Neural Information Processing Systems*, pages 631–639.

Park, G. and Raskutti, G. (2017). Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *arXiv preprint arXiv:1704.08783*.

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Scholkopf, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2012). Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Tsamardinos, I. and Aliferis, C. F. (2003). Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the ninth international workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann Publishers: Key West, FL, USA.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463.

Verma, T. and Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in Artificial Intelligence, 1992*, pages 323–330. Elsevier.

Zhang, J. and Spirtes, P. (2016). The three faces of faithfulness. *Synthese*, 193(4):1011–1027.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press.