

High-Dimensional Poisson DAG Model Learning Using ℓ_1 -Regularized Regression

Gunwoong Park¹, Sion Park¹

¹ Department of Statistics, University of Seoul

Abstract

In this paper, we develop a new approach to learning high-dimensional Poisson directed acyclic graphical (DAG) models from only observational data without strong assumptions such as faithfulness and strong sparsity. A key component of our method is to decouple the ordering estimation or parent search where the problems can be efficiently addressed using ℓ_1 -regularized regression and the mean-variance relationship. We show that sample size $n = \Omega(d^2 \log^9 p)$ is sufficient for our polynomial time Mean-variance Ratio Scoring (MRS) algorithm to recover the true directed graph, where p is the number of nodes and d is the maximum indegree. We verify through simulations that our algorithm is statistically consistent in the high-dimensional $p > n$ setting, and performs well compared to state-of-the-art ODS, GES, and MMHC algorithms. We also demonstrate through multivariate real count data that our MRS algorithm is well-suited to estimating DAG models for multivariate count data in comparison to other methods used for discrete data.

1 INTRODUCTION

Directed acyclic graphical (DAG) models, also referred to as Bayesian networks, are popular probabilistic statistical models to analyze and visualize (functional) causal or directional dependence relationships among random variables.(see e.g., [1, 2, 3, 4]). However, learning DAG models from only observational data is a notoriously difficult problem due to non-identifiability and exponentially growing computational complexity. Prior works have addressed the question of identifiability for different classes of joint distribution $\mathbb{P}(G)$. [5, 6] show the Markov equivalence class (MEC) where graphs that belong to the same MEC have the same conditional independence relations. [7, 8, 9, 10] show that the underlying graph of a DAG model is recoverable up to the MEC under faithfulness or related assumptions that can be very restrictive [11].

Also well studied is how learning a DAG model is computationally non-trivial due to the super-exponentially growing size of the space of DAGs in the number of nodes [12]. Hence, it is NP-hard to search DAG space [13, 14], and many existing algorithms such as PC [7], Greedy Equivalence Search (GES) [8], Min-Max Hill Climbing (MMHC) [15] and greedy DAG search (GDS) [4], take greedy search methods that may not guarantee recovering the true MEC.

Recently, a number of fully identifiable classes of DAG models have been introduced [16, 17, 18, 4, 19, 20]. In addition, some of these models can be successfully learned from high-dimensional data by decomposing the DAG learning problem into ordering estimation and skeleton estimation [21, 19, 20]. The main reasoning is that if ordering is known or recoverable, learning a directed graphical model is as hard as

learning an undirected graphical model or Markov random field (MRF). [22, 23, 24, 25] show that sparse undirected graphs can be estimated via ℓ_1 -regularized regression in high-dimensional settings under suitable conditions.

In this paper, we focus on learning Poisson DAG models [19, 20] for multivariate count data in high-dimensional settings since large-scale multivariate *count data* frequently arises in many fields, such as high-throughput genomic sequencing data, spatial incidence data, sports science data, and disease incidence data. Like learning the Poisson undirected graphical model or MRF introduced in [25], where the sample bound is $\Omega(d_m^2 \log^3 p)$, it is not surprising that Poisson DAG models can be learned in high dimensional settings when the indegree of the graph d is bounded. [20] establishes the consistency of learning Poisson DAG models with the sample bound $n = \Omega(\max\{d_m^4 \log^{12} p, \log^{5+d} p\})$. This huge sample complexity difference between directed and undirected graphical models is induced mainly for three reasons: (i) nonexistence ordering, (ii) the known parametric functional form (the standard log link) for the dependencies, and (iii) the restrictive non-positive parameter space in Poisson MRFs (see details in [25]).

The main objective of this work is to propose a new milder identifiability assumption, and to develop a new polynomial time approach, called Mean-variance Ratio Scoring (MRS), for learning high-dimensional Poisson DAG models when the parametric functional form for the dependencies is known while the parameters are unbounded and unknown. We address the question of learning high-dimensional Poisson DAG models under the causal sufficiency assumption that all relevant variables have been observed. However, we do not assume the faithfulness assumption that might be restrictive [11].

The MRS algorithm combines the idea of the mean-variance Poisson relationship for recovering an ordering, and the sparsity-encouraging ℓ_1 -regularized regression in finding the parents of each node. We provide sufficient conditions and its sample complexity $n = \Omega(d^2 \log^9 p)$ under which the MRS algorithm recovers the Poisson DAG model with a high probability in the high-dimensional $p > n$ setting.

We demonstrate through simulations and a real baseball data application involving multivariate count data that our MRS algorithm performs better than state-of-the-art OverDispersion Scoring (ODS) [19], GES [8], MMHC [15] and Poisson MRF [25], on average, in terms of the both run-time and accuracy of recovering a graph structure and its MEC. In our simulation study, we consider both the extremely sparse ($d = 1$) and sparse ($d = 10$) high-dimensional settings. Our real data example involving MLB player statistics for 2003 season shows that our MRS algorithm is applicable to multivariate count data while the Poisson undirected graph has too many edges, and the MMHC algorithm tends to select very few edges when variables represent counts.

1.1 Our Contributions

We summarize the major contributions of the paper as follows:

- We introduce a milder fully identifiability condition for Poisson DAG models for multivariate count data.

- We develop the reliable and scalable lasso-based MRS algorithm which learns sparse high dimensional Poisson DAG models.
- We provide sufficient conditions and its sample complexity $n = \Omega(d^2 \log^9 p)$ under which the MRS algorithm recovers the Poisson DAG model. To the best of our knowledge, this is the only theoretical result that applies to the high-dimensional setting for the strongly correlated multivariate count data.

The remainder of this paper is structured as follows. Section 2.1 summarizes the necessary notations and problem settings, Section 2.2 discusses the Poisson DAG model and its new identifiability condition, and Section 2.3 provides a detailed comparison between Poisson DAG models and MRFs. In Section 3, we introduce our polynomial-time DAG learning algorithm, which we refer to as the Mean-variance Ratio Scoring (MRS). Section 3.1 discusses computational complexity of our algorithm, and Section 3.2 provides statistical guarantees for learning Poisson DAG models via the MRS algorithm. Section 4 empirically evaluates our methods, compared to state-of-the-art ODS, GES, and MMHC algorithms using synthetic data, and confirms that our algorithm is one of the few DAG-learning algorithms that performs well in terms of statistical and computational complexity in low and high-dimensional settings. Lastly, Section 5 compares our MRS algorithm to the Poisson MRF and MMHC algorithm by analyzing a real 2003 season MLB multivariate count data.

2 POISSON DAG MODELS

We first introduce some necessary notations and definitions for directed acyclic graph (DAG) models. Then, we give a detailed description of previous work on learning Poisson DAG models [19], and we propose a strictly milder identifiability condition. Lastly, we discuss how Poisson DAG models and MRFs [26, 25] are related.

2.1 Problem Set-up and Notations

A DAG $G = (V, E)$ consists of a set of nodes $V = \{1, 2, \dots, p\}$ and a set of directed edges $E \subseteq V \times V$ with no directed cycles. A directed edge from node j to k is denoted by (j, k) or $j \rightarrow k$. 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 $j \rightarrow \dots \rightarrow k$, 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* $\pi = (\pi_1, \dots, \pi_p)$ 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 the ordering and the skeleton that is the set of directed edges without their directions.

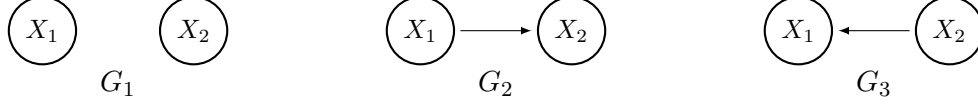


Figure 1: Bivariate directed acyclic graphs of G_1 , G_2 , and G_3 .

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 $P(G) = P(X_1, X_2, \dots, X_p)$. 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 $j \in V$, $\mathbb{P}(X_j | X_S)$ denotes the conditional distribution of a variable X_j given a random vector X_S . Then, a DAG model has the following factorization [27]:

$$\mathbb{P}(G) = \mathbb{P}(X_1, X_2, \dots, X_p) = \prod_{j=1}^p \mathbb{P}(X_j | X_{\text{Pa}(j)}), \quad (1)$$

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

We suppose that there are n independent and identically distributed samples $X^{1:n} := (X^{(i)})_{i=1}^n$ from a given graphical model where $X^{(i)} := X_{1:p}^{(i)} = (X_1^{(i)}, X_2^{(i)}, \dots, X_p^{(i)})$ is a p -variate random vector. The notation $\hat{\cdot}$ denotes an estimate based on samples $X^{1:n}$. We also accept the causal sufficiency assumption that all important variables have been observed.

2.2 Poisson DAG Model and its Identifiability

The definition of Poisson DAG models in [19] is that each conditional distribution given its parents $X_j | X_{\text{Pa}(j)}$ is Poisson such that

$$X_j | X_{\text{Pa}(j)} \sim \text{Poisson}(g_j(X_{\text{Pa}(j)})), \quad (2)$$

where for any arbitrary positive link function $g_j : \mathcal{X}_{\text{Pa}(j)} \rightarrow \mathbb{R}^+$. Hence using the factorization in Equation (1), the joint distribution is as follows:

$$f_G(X) = \prod_{j \in V} f_j(X_j | X_{\text{Pa}(j)}). \quad (3)$$

where f_j is the probability density function of Poisson.

Now, we briefly explain how Poisson DAG models are identifiable from only the distribution using the bivariate Poisson DAG 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 $X_2 | X_1 \sim \text{Poisson}(g_2(X_1))$; and $G_3 : X_2 \sim \text{Poisson}(\lambda_2)$ and $X_1 | X_2 \sim \text{Poisson}(g_1(X_2))$ for arbitrary positive functions $g_1, g_2 : \mathbb{N} \cup \{0\} \rightarrow \mathbb{R}^+$.

We exploit the mean-variance relationship of a Poisson distribution to find the true underlying graph. For G_1 , $\text{Var}(X_j) = \mathbb{E}(X_j)$ for all $j = 1, 2$. For G_2 , $\text{Var}(X_1) = \mathbb{E}(X_1)$, while

$$\text{Var}(X_2) = \mathbb{E}[\text{Var}(X_2 | X_1)] + \text{Var}[\mathbb{E}(X_2 | X_1)] = \mathbb{E}[\mathbb{E}(X_2 | X_1)] + \text{Var}[\mathbb{E}(X_2 | X_1)] > \mathbb{E}(X_2),$$

as long as $\text{Var}[\mathbb{E}(X_2 \mid X_1)] > 0$. For G_3 , $\text{Var}(X_1) > \mathbb{E}(X_1)$ and $\text{Var}(X_2) = \mathbb{E}(X_2)$ in the same manner.

This idea of a mean-variance relationship can easily apply to general p-variate Poisson DAG models, and hence the models are identifiable by testing overdispersion for the conditional distribution of each node conditioning on other variables, and equivalently testing whether the following moments ratio, $\mathbb{E}(X_j^2)/(\mathbb{E}(X_j) + \mathbb{E}(X_j)^2)$, is equal to 1 or greater than 1.

Theorem 2.1. Assume that for all $j \in V$, non-empty $Pa_0(j) \subset Pa(j)$, and $S_j \subset Nd(j) \setminus Pa_0(j)$,

$$\mathbb{E}(X_j^2) > \mathbb{E}(\mathbb{E}(X_j \mid X_{S_j}) + \mathbb{E}(X_j \mid X_{S_j})^2)$$

the Poisson DAG model is identifiable.

We include the proof in Section 3.2. This identifiability condition in Theorem 2.1 is equivalent to $\mathbb{E}(\text{Var}(\mathbb{E}(X_j \mid X_{Pa(j)}) \mid X_{S_j})) > 0$, meaning that any Poisson DAG model is identifiable if all parents of node j contribute to its variability. The identifiability condition is strictly milder than the previous identifiability condition $\text{Var}(\mathbb{E}(X_j \mid X_{Pa(j)}) \mid X_{S_j}) > 0$ in [19]. It allows for some $x_{Pa(j)} \in X_{Pa(j)}$ not to influence the variable X_j .

2.3 Comparison to Poisson MRF

In this section, we compare Poisson DAG models and MRFs where the conditional distributions of each node given its parents and neighbors are Poisson, respectively. To simplify the comparison, we consider the Poisson structural equation model (SEM) where the link functions g_j 's in Equation (2) are the standard log link function for Poisson generalized linear models (GLMs), i.e., $g_j(X_{Pa(j)}) = \exp(\theta_j + \sum_{k \in Pa(j)} \theta_{jk} X_k)$ where $(\theta_{jk})_{k \in Pa(j)}$ represents the non-zero linear weights. Using factorization (1), the joint distribution of the Poisson SEM can be written as:

$$f(X_1, X_2, \dots, X_p) = \exp\left(\sum_{j \in V} \theta_j X_j + \sum_{(k,j) \in E} \theta_{jk} X_j X_k - \sum_{j \in V} \log X_j! - \sum_{j \in V} e^{\theta_j + \sum_{k \in Pa(j)} \theta_{jk} X_k}\right). \quad (4)$$

This is a form similar to the joint distribution of Poisson MRFs in [25], where the joint distribution has the following form:

$$f(X_1, X_2, \dots, X_p) = \exp\left(\sum_{j \in V} \theta_j X_j + \sum_{(k,j) \in E} \theta_{jk} X_j X_k - \sum_{j \in V} \log X_j! - A(\theta)\right), \quad (5)$$

where $A(\theta)$ is the log of the normalization constant. The key difference between a Poisson SEM and a Poisson MRF is the normalization constant $A(\theta)$ in Equation (5), as opposed to the term $\sum_{j \in V} e^{\theta_j + \sum_{k \in Pa(j)} \theta_{jk} X_k}$ in Equation (4), which depends on variables.

[25] proves that Poisson MRF (5) is normalizable if and only if all (θ_{jk}) values are less than or equal to 0. This means Poisson MRFs only capture negative dependency relations. In addition, [25] addresses

the learning Poisson MRFs when the functional form of dependencies is $X_j \mid X_{V \setminus j} \sim \text{Poisson}(\exp(\theta_j + \sum_{k \in \mathcal{N}(j)} \theta_{jk} X_k))$ where $\mathcal{N}(j)$ denotes the neighbors of a node j in the graph.

While Poisson MRFs have strong restrictions on the functional form for dependencies and the parameter space, it can be successfully learned in the high-dimensional settings with less restrictive constraints of sparsity. [25] shows that Poisson MRFs can be recovered via ℓ_1 -regularized regression if $n = \Omega(d_m^2 \log^3 p)$, where d_m is the degree of the undirected graph. In contrast, [19, 20] show that Poisson DAG models can be learned via the ODS algorithm if $n = \Omega(\max\{d_m^4 \log^{12} p, \log^{5+d} p\})$ where d_m is obtained by the moralized graph and d is the maximum indegree of the graph. This big difference in the sample complexity primarily comes from the unknown functional form for the dependencies in Poisson DAG models. In the next section, we will show that a significant advantage can be achieved by assuming the parametric function for the dependencies in terms of recovering the graphs.

3 ALGORITHM

Here, we present our Mean-variance Ratio Scoring (MRS) algorithm for learning the Poisson SEM (4). Our algorithm alternates between an element-wise ordering search using the (conditional) mean-variance ratio, and a parent search using the ℓ_1 -regularized GLM. Hence, the algorithm chooses a node for the first element of the ordering, and then determines its parents. The algorithm iterates this procedure until the last element of the ordering and its parents are determined.

Without loss of generality, assume that $\pi = (1, 2, \dots, p)$ is the true ordering. Then Poisson SEMs (4) have the conditional distribution of X_j given that all variables before j in the ordering are reduced to the following Poisson GLM:

$$P(X_j \mid X_{1:(j-1)}) = \exp\left\{\theta_j X_j + \sum_{k \in 1:(j-1)} \theta_{jk} X_k X_j + \log X_j! - \exp\left(\theta_j + \sum_{k \in 1:(j-1)} \theta_{jk} X_k\right)\right\}, \quad (6)$$

where $\theta_{jk} \in \mathbb{R}$ represents the influence of node k on node j . For ease of notation, let $\theta(j)$ be a set of parameters related to Poisson GLM (6). Then $\theta(j) = (\theta_j, \theta_{\setminus j}) \in \mathbb{R} \times \mathbb{R}^{j-1}$ where $\theta_{\setminus j} = (\theta_{jk})_{k \in \{1, 2, \dots, j-1\}}$ is a zero-padded vector with non-zero entries if $k \in \text{Pa}(j)$.

Our MRS (Algorithm 1) involves learning the ordering by comparing mean-variance ratio scores of nodes using the following equations:

$$\hat{\mathcal{S}}_r(1, j) := \frac{\hat{\mathbb{E}}(X_j^2)}{\hat{\mathbb{E}}(X_j) + \hat{\mathbb{E}}(X_j)^2} \quad \text{and} \quad \hat{\mathcal{S}}_r(m, j) := \frac{\hat{\mathbb{E}}(X_j^2)}{\hat{\mathbb{E}}(\hat{\mathbb{E}}(X_j \mid X_{1:(m-1)}) + \hat{\mathbb{E}}(X_j \mid X_{1:(m-1)})^2)}, \quad (7)$$

where $\hat{\mathbb{E}}(X_j) = \frac{1}{n} \sum_{i=1}^n X_j^{(i)}$, and $\hat{\mathbb{E}}(X_j \mid X_S) = \exp(\hat{\theta}_j^S + \sum_{k \in S} \hat{\theta}_{jk}^S X_k)$ where $\hat{\theta}_S(j) = (\hat{\theta}_j^S, \hat{\theta}_{\setminus j}^S)$ is the solution of the following ℓ_1 -regularized GLM:

$$\hat{\theta}_S(j) := \arg \min \frac{1}{n} \sum_{i=1}^n \left[-X_j^{(i)} \left(\theta_j + \sum_{k \in S} \theta_{jk} X_k^{(i)} \right) + \exp\left(\theta_j + \sum_{k \in S} \theta_{jk} X_k^{(i)} \right) \right] + \lambda_j \sum_{k \in S} |\theta_{jk}|. \quad (8)$$

Algorithm 1: Mean-variance Ratio Scoring (MRS)

Input : n i.i.d. samples, $X^{1:n}$

Output: Estimated ordering $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_p)$ and an edge structure, $\hat{E} \in V \times V$

Set $\hat{\pi}_0 = \emptyset$;

for $m = \{1, 2, \dots, p\}$ **do**

 Set $S = \{\hat{\pi}_1, \dots, \hat{\pi}_{m-1}\}$;

for $j \in \{1, 2, \dots, p\} \setminus S$ **do**

 Estimate $\hat{\theta}_S(j)$ for ℓ_1 -regularized generalized linear model (8);

 Calculate scores $\hat{S}(m, j)$ using Equation (7);

end

 The m^{th} element of the ordering, $\hat{\pi}_m = \arg \min_j \hat{S}(m, j)$;

 The parents of the m^{th} element of the ordering, $\hat{\text{Pa}}(\hat{\pi}_m) = \{k \in S \mid \hat{\theta}_{\hat{\pi}_m k}^S \neq 0\}$;

end

Return: Estimate the edge set, $\hat{E} = \cup_{m \in V} \{(k, \hat{\pi}_m) \mid k \in \hat{\text{Pa}}(\hat{\pi}_m)\}$

This score represents the level of how well each node satisfies the identifiability condition in Theorem 2.1. Hence, the correct element of the ordering has a score of 1, otherwise strictly greater than 1 in population. The ordering is determined one node at a time by selecting the node with the smallest score.

The novelty of our algorithm is learning an ordering by testing which nodes have correct mean-variance relations according to the conditional distribution using the ℓ_1 -regularized GLM. By substituting the estimation of parameters $\theta(j)$ for an estimation of the conditional mean, we gain significant computational and statistical improvements compared to previous work in [19] where the method of moments are used for an estimation of the conditional mean. In principle, the computation of a conditional mean exponentially grows in the number of conditioning variables.

As we discussed, the problem of a learning directed graph structure is the same as the problem of a learning undirected graph structure if the ordering is known. Hence, given the estimated ordering, the parents of each node j can be learned via ℓ_1 -regularized GLMs (see details in [22, 23, 24, 25]). Therefore, we determine the estimated parents of a node j as $\hat{\text{Pa}}(j) := \{k \in S : \hat{\theta}_{jk}^S \neq 0\}$ where $S = \hat{\pi}_{1:(j-1)}$ and $\hat{\theta}_S(j)$ is the solution to Equation (8).

3.1 Computational Complexity

The computation complexity for the MRS algorithm involves the ℓ_1 -regularized GLM algorithm [28] where the worse-case complexity is $O(np)$ for a single ℓ_1 -regularized regression run. More precisely, the coordinate descent method updates each gradient in $O(p)$ operations. Hence, with d non-zero terms in the GLM, a complete cycle costs $O(pd)$ operations if no new variables become non-zero, and costs $O(np)$ for each new variable entered (see details in [29]). Since our algorithm has p iterations and there are $p - j + 1$ regressions

with $j - 1$ features for the j th iteration, the total worst-case complexity is $O(np^3)$.

The estimation of a Poisson MRF also involves a node-wise ℓ_1 -regularized GLM over all other variables, and hence the worse-case complexity is $O(np^2)$ if the coordinate descent method is exploited. The addition of estimation of ordering makes p times more computationally inefficient than the standard method for learning Poisson MRFs.

Learning a DAG model is NP-hard in general [13]. Hence, many state-of-the-art MEC and DAG learning algorithms, such as PC [7], GES [8], and MMHC [15], are inherently greedy search algorithms. In the numerical experiments in Section 4, we compare MRS to greedy hill-climbing search-based GES and MMHC algorithms in terms of run time, and show that MRS has a significantly better computational complexity.

3.2 Theoretical Guarantees

In this section, we provide theoretical guarantees on the MRS algorithm for learning Poisson SEMs (4). 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.

We begin by discussing the assumptions we impose on Poisson SEMs. Since we apply ℓ_1 -regularized regression for the parent selection, most assumptions are similar to those imposed in [23, 24, 25, 20] where ℓ_1 -regularized regression was used for graphical model learning.

Important quantities are the Hessian matrices of the negative conditional log-likelihood of a node j given some subsets of the nodes in the ordering, $S_j \in \{\emptyset, \{\pi_1\}, \{\pi_1, \pi_2\}, \dots, \{\pi_1, \dots, \pi_{j-1}\}\}$. Let $Q^{j, S_j} := \nabla^2 \ell_j^{S_j}(\theta_{S_j}^*(j); X^{1:n})$ where

$$\ell_j^{S_j}(\theta_{S_j}(j), X^{1:n}) := \frac{1}{n} \sum_{i=1}^n \left[-X_j^{(i)} \left(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k^{(i)} \right) + \exp \left(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k^{(i)} \right) \right], \quad (9)$$

$$\theta_{S_j}^*(j) := \arg \min_{\theta \in \Theta_{S_j}(j)} \mathbb{E} \left[-X_j \left(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k \right) + \exp \left(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k \right) \right], \quad (10)$$

where $\Theta_{S_j}(j) := \{\theta_{S_j}(j) \in \mathbb{R}^{|S_j|+1} : \theta_{jk}^{S_j} = 0 \text{ for } k \notin \text{Pa}(j)\}$.

For simplicity we let A_{SS} denote the $|S| \times |S|$ sub-matrix of the matrix A corresponding to variables X_S .

Assumption 3.1 (Dependence Assumption). For any $j \in V$ with $S_j \in \{\{\pi_1\}, \{\pi_1, \pi_2\}, \dots, \{\pi_1, \dots, \pi_{j-1}\}\}$ and non-empty $T_j = S_j \cap \text{Pa}(j)$, there exists positive constants ρ_{\min} and ρ_{\max} such that

$$\min_{j \in V} \lambda_{\min} \left(Q_{T_j, T_j}^{j, S_j} \right) \geq \rho_{\min}, \quad \text{and} \quad \max_{j \in V} \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^n X_{\text{Pa}(j)}^{(i)} (X_{\text{Pa}(j)}^{(i)})^T \right) \leq \rho_{\max},$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest eigenvalues of the matrix A , respectively.

Assumption 3.2 (Incoherence Assumption). For any $j \in V$ with $S_j \in \{\{\pi_1\}, \{\pi_1, \pi_2\}, \dots, \{\pi_1, \dots, \pi_{j-1}\}\}$ and non-empty $T_j = S_j \cap \text{Pa}(j)$, there exists a constant $\alpha \in (0, 1]$ such that

$$\max_{j, S_j} \max_{t \in T_j^c} \|Q_{tT_j}^{j, S_j} (Q_{T_j T_j}^{j, S_j})^{-1}\|_1 \leq 1 - \alpha.$$

Assumption 3.1 ensures that the parent variables $X_{\text{Pa}(j)}$ are not too dependent. In addition, Assumption 3.2 ensures that parent and non-parent variables are not highly correlated. These two assumptions are standard in all neighborhood regression approaches to variable selection involving ℓ_1 -regularized based methods, and these conditions have imposed in proper works for both high-dimensional regression and graphical model learning.

To ensure suitable concentration bounds hold, we require a boundedness assumption on the Hessian matrices Q^{j, S_j} to control the tail behavior of likelihood functions. Since the Hessian matrices are proportional to $\exp(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k)$, we require the following assumption.

Assumption 3.3 (Concentration Bound Assumption). For any $j \in V$ and $\theta_{S_j}^*(j) = (\theta_j^{S_j}, \theta_{jk}^{S_j})$ in Equation (10), there exists a constant $M_{\max} > 0$ such that

$$\max_{j, S_j} \mathbb{E}(e^{\exp(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k)}) < M_{\max}.$$

Assumption 3.3 implies that the maximum eigenvalues of the Hessian matrices are bounded with a high probability. In addition, this assumption leads to the boundedness on the moment generating function because for $S_j = \text{Pa}(j)$,

$$\max_{j \in V} \mathbb{E}(\exp(X_j)) = \max_{j \in V} \mathbb{E}(e^{(e-1)\exp(\theta_j^{S_j} + \sum_{k \in S_j} \theta_{jk}^{S_j} X_k)}) < (M_{\max})^{e-1}.$$

Thus, it can be understood that too large rate parameters are not allowed for all conditional distributions of nodes given their parents.

The prior work in [25] on learning Poisson MRFs also imposes technical conditions that control the tail behavior of $(X_j)_{j=1}^p$ and the Hessian matrices, which can be satisfied based on the restricted negative parameter space $\theta_{jk} < 0$. However for learning Poisson SEMs where there is no parameter restriction, the related assumption used in [25] is unrealistic.

Lastly, we require a stronger version of the (conditional) mean-variance ratio identifiability assumption in Theorem 2.1, because we move from the population to the finite samples.

Assumption 3.4. For all $j \in V$ and any non-empty $\text{Pa}_0(j) \subset \text{Pa}(j)$ where $S_j \subset \text{Nd}(j) \setminus \text{Pa}_0(j)$, there exists an $M_{\min} > 0$ such that

$$\mathbb{E}(X_j^2) > (1 + M_{\min})\mathbb{E}(\mathbb{E}(X_j \mid X_{S_j}) + \mathbb{E}(X_j \mid X_{S_j})^2).$$

Putting together Assumptions 3.1, 3.2, 3.3, and 3.4, we have the following main result that a Poisson SEM can be recovered via our MRS algorithm in high-dimensional settings. The theorem provides not only sufficient conditions, but also the probability that our method recovers the true graph structure.

Theorem 3.5. Consider a Poisson SEM (4) with parameter vector $(\theta(j))_{j \in V}$ and the maximum indegree of the graph d . Suppose that the regularization parameter (8) is chosen, such that

$$\frac{64\sqrt{2}(2-\alpha)}{\alpha} \frac{d \log^2 \eta}{\kappa_1(n, p)} \leq \lambda_j \leq \frac{\alpha}{16 \cdot 10^3(2-\alpha)} \frac{\rho_{\min}^2}{\rho_{\max}} \frac{1}{d \log^2 \eta},$$

for any $\alpha \in (0, 1]$, and $\kappa_1(n, p) \geq \frac{2048 \cdot 10^3 (2-\alpha)^2}{\alpha^2} \frac{\rho_{\max}}{\rho_{\min}^2}$. Suppose also that Assumptions 3.1, 3.2, 3.3 and 3.4 are satisfied and the values of the parameters in Equation (4) are sufficiently large such that $\min_{(j,k) \in E} |\theta_{jk}| \geq \frac{10}{\rho_{\min}} \sqrt{d} \lambda_j$. Then, for any $\epsilon > 0$, there exists a positive constant C_ϵ such that if the sample size is sufficiently large $n > C_\epsilon d^2 \log^9(\max\{n, p\})$, then the MRS algorithm uniquely recovers the graph with a high probability:

$$P(\hat{G} = G) \geq 1 - \epsilon.$$

Detailed proof is provided in Appendices B and C. Appendix B provides the error probability that the ℓ_1 -regularized regression recovers the true parents of each node given the true ordering, and Appendix C provides the error probability that the ℓ_1 -regularized regression recovers the ordering. The key technique for the proof is that the *primal-dual witness* method used in sparse regularized regression and related techniques [22, 23, 24, 25]. Theorem 3.5 intuitively makes sense because neighborhood selection via the ℓ_1 -regularized regression is a well-studied problem, and its bias can be controlled by choosing the appropriate regularization parameter λ_j . Hence, our mean-variance ratio scores can be sufficiently close to the true scores to recover the true ordering.

Theorem 3.5 claims that if $n = \Omega(d^2 \log^9 p)$, our MRS algorithm recovers an underlying graph with a high probability. Hence, our MRS algorithm works in a high-dimensional setting, provided that the indegree of a graph d is bounded. This sample bound result shows that our method has much more relaxed constraints on the sparsity of the graph than the previous work in [20], where the sample bound is $n = \Omega(\max\{(d_m \log^3 p)^4, \log^{5+d} p\})$. Moreover, it also shows that learning Poisson DAG models requires more samples than the learning Poisson MRFs in [25], where the sample bound is $n = \Omega(d_m^2 \log^3 p)$ due to the existence of the ordering and the unrestricted parameter space.

4 Numerical Experiments

In this section, we provide simulation results to support our main theoretical results of Theorem 3.5 and the computational complexity in Section 3.1: (i) the MRS algorithm recovers the ordering and edges more accurately as sample size increases; (ii) the required sample size $n = \Omega(d^2 \log^9 p)$ depends on the number of nodes p and the complexity of the graph d ; (iii) the MRS algorithm accurately learns the graphs in high-dimensional settings ($p > n$); and (iv) the computational complexity is $O(np^3)$ at worst. We also show that the MRS algorithm performs favorably compared to the ODS [19], GES [8], and MMHC [15] algorithms. Lastly, we investigate how sensitive our MRS algorithm is to deviations from the assumption about the link functions by using the identity link function in Equation (3).

4.1 Random Poisson SEMs

We conducted simulations using 200 realizations of p -node Poisson SEMs (4) with the randomly generated underlying DAG structure while respecting the indegree constraints $d \in \{1, 5, 10\}$. A graph with $d = 1$ is a special case where there is no v-structure, and therefore, the corresponding MEC is completely undirected. The set of non-zero parameters $\theta_j, \theta_{jk} \in \mathbb{R}$ in Equation (4) was generated uniformly at random in the range $\theta_j \in [1, 3]$, $\theta_{jk} \in [-1.5, -0.5] \cup [0.5, 1.5]$ for $d = 1$, and $\theta_{jk} \in [-1, -0.1] \cup [0.1, 1]$ for $d = 5, 10$, which helps the generated values of samples to avoid either all zeros or from going beyond the maximum possible value of the R program ($> 10^{309}$). Nevertheless, if some samples were beyond the maximum possible value, we regenerated the parameters and samples.

The MRS and ODS algorithms were implemented using ℓ_1 -regularized likelihood where we used five-fold cross validation to choose the regularization parameters. Where mean squared error was within two standard error of the minimum mean squared error, we chose the minimum value for the mean-variance ratio scores and the largest value for parent selection. That was because a less biased estimator is preferred for the score calculation, and we preferred a sparse graph containing only legitimate edges. We acknowledge that the level of sparsity can be adjusted according to the importance of precision or recall.

In Fig. 2, we compare the MRS algorithm to state-of-the-art ODS, GES and MMHC algorithms for graph node size $p = \{20, 200\}$, varying sample size $n \in \{25, 50, \dots, 250\}$ for $d = 1$ and $n = \{100, 200, \dots, 1000\}$ for $d = 10$, and provide two results: (i) the average precision ($\frac{\# \text{ of correctly estimated edges}}{\# \text{ of estimated edges}}$); (ii) the average recall ($\frac{\# \text{ of correctly estimated edges}}{\# \text{ of true edges}}$). As discussed, the both GES and MMHC algorithms only recover the partial graph by leaving some arrows undirected. Therefore, we also provide average precision and recall for the estimated MECs in Fig. 3. Lastly, we provide an oracle, where the true parents of each node are used, while the ordering is estimated via ℓ_1 -regularized GLM. Hence, we can see where the errors come from between the ordering estimation or parent selection. We considered more parameters (θ_{jk}, n, p, d), but for brevity, we focus on these settings.

As we can see in Fig. 2, the MRS algorithm more accurately recovers the true directed edges as sample size increases. In addition, the MRS algorithm is more precise for small sparse graphs than for large-scale or dense graphs, given the same sample size. Hence it confirms that the MRS algorithm is consistent, and the sample bound $n = \Omega(d^2 \log^9 p)$ depends on p and d .

The MRS algorithm significantly outperforms state-of-the-art GES and MMHC algorithms in terms of both precision and recall, on average, except for cases $p = 20, d = 1, n \leq 50$. It is worth noting that the GES and MMHC algorithms are not consistent, because the recall for any tree graph must be zero in population, whereas the recall from GES and MMHC increases as sample size increases. Hence, we can conclude that the GES and MMHC algorithms find correct directed edges by finding incorrect v-structures. It is an expected result because the comparison methods only works with a non-faithful distribution, which rarely arises in finite sample settings [11].

Fig. 2 shows that the MRS and ODS algorithms have similar performance in identifying directed edges

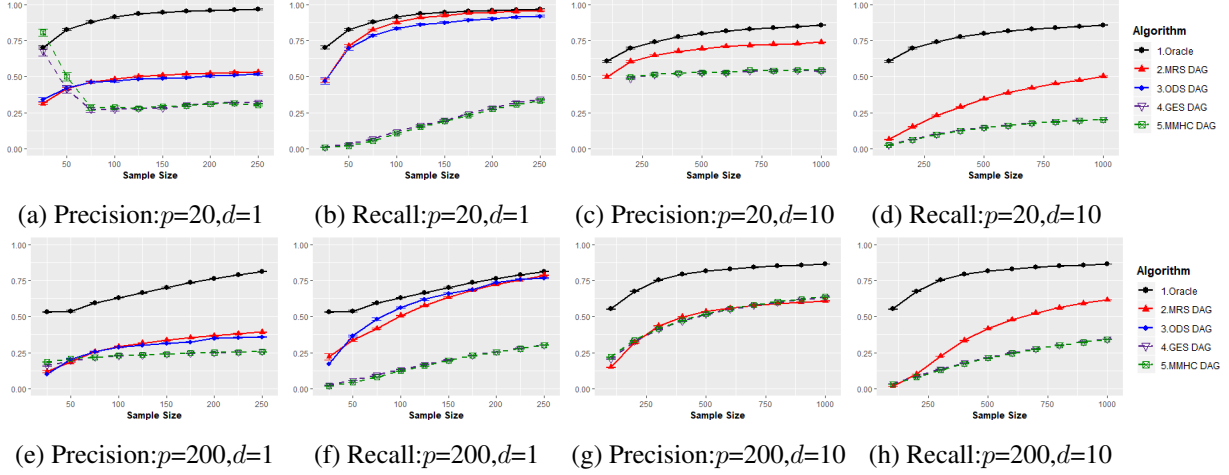


Figure 2: Comparison of the MRS algorithm to the oracle, ODS, GES and MMHC algorithms in terms of precision and recall for Poisson SEMs with $p \in \{20, 200\}$ and $d \in \{1, 10\}$.

n	100	200	300	400	500	600	700	800	900	1000
p = 20	199	175	107	64	1	0	0	0	0	0
p = 50	200	200	200	199	192	179	151	140	99	86

Table 1: Number of failures in ODS algorithm implementations from among 200 sets of samples for different node sizes $p \in \{20, 50\}$, and sample sizes $n \in \{100, 200, \dots, 1000\}$, when the indegree is $d = 5$.

when the indegree is a small $d = 1$. It makes sense because the ODS algorithm recovers any Poisson DAG models if the moralized graph is sparse. In other words, the accuracy of the ODS algorithm may be poor for the non-sparse graph. Moreover, the ODS algorithm often fails to be implemented due to a lack of samples for the estimation of conditional variance, that is, $\sum_{i=1}^n \mathbf{1}(X_S^{(i)} = x) < 2$ for all $x \in \mathcal{X}_S$. Table 1 shows the number of failures in the ODS algorithm implementations for node size $p \in \{20, 50\}$ and sample size $n \in \{100, 200, \dots, 1000\}$ when the indegree is $d = 5$, and the degree of the moralized graph is at most $d_m = p - 1$. It empirically confirms that the ODS algorithm requires a huge number of samples to be implemented when a true graph is not sparse. Hence, we do not apply the ODS algorithm for the graphs with $d = 10$. It is consistent with our main result that our method can learn the non-sparse Poisson SEMs while the ODS algorithm might not.

Fig. 3 shows the analogous results for the recovery of MECs, in which the MRS and all comparison algorithms consistently learn the true MECs. The performance of the MRS algorithm gets better as sample size increases or node size decreases. In addition, we can see that the MRS algorithm still recovers the MEC of the Poisson SEM better on average than the comparison methods. However, it must be pointed out that our MRS algorithm applies to Poisson SEMs (4), while the ODS algorithm accurately learns sparse Poisson DAG models where arbitrary link functions are allowed. In addition, the GES and MMHC algorithms apply

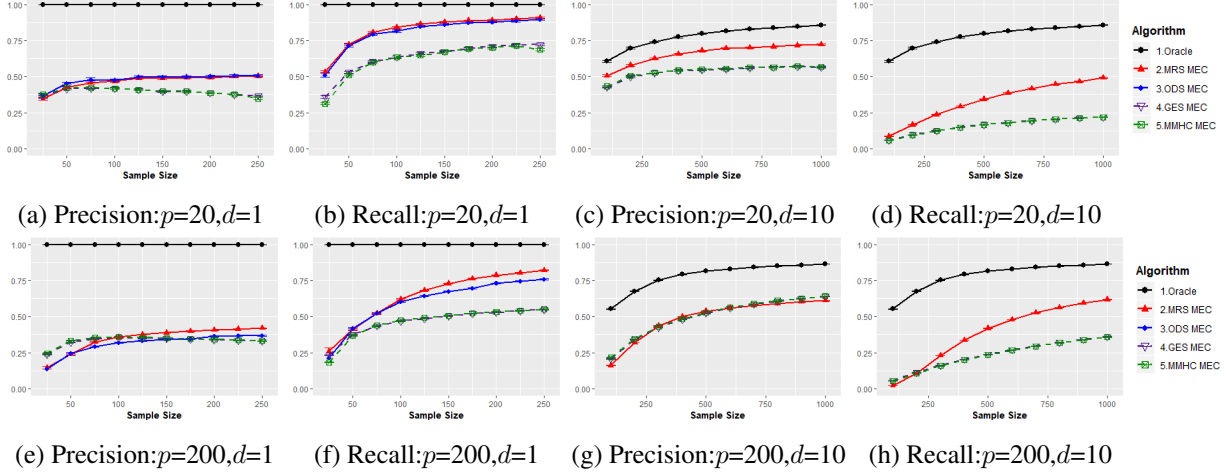


Figure 3: Comparison of the MRS algorithm to the oracle, ODS, GES, and MMHC algorithms in terms of the precision and recall for the MECs of Poisson SEMs with $p \in \{20, 100\}$ and $d \in \{1, 10\}$.

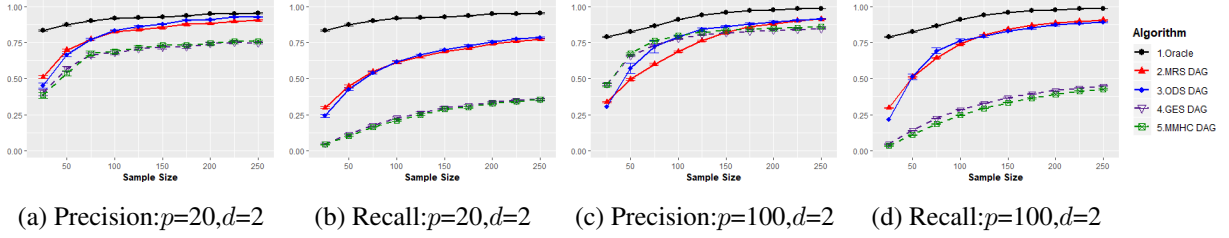


Figure 4: Comparison of the MRS algorithm to the oracle, ODS, GES and MMHC algorithms in terms of the precision and recall for Poisson DAG models with $p \in \{20, 100\}$, $d = 2$, and the identity link function.

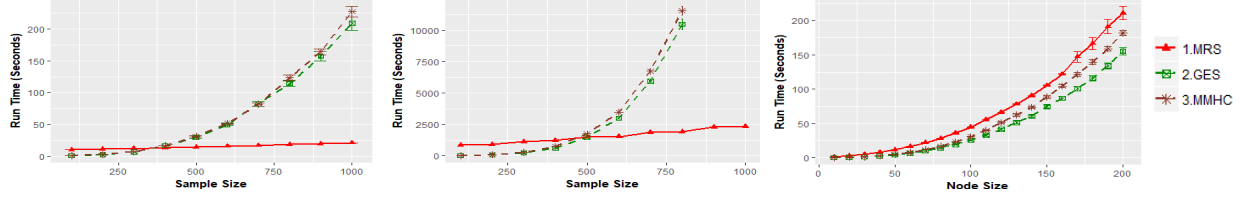
to more general classes of DAG models.

4.2 Random Poisson DAG Models

When the data are generated by a random Poisson DAG model (2) where g_j is not the standard log link function, our MRS algorithm is not guaranteed to estimate the true directed acyclic graph and its ordering. Hence, an important question is how sensitive our method is to deviations from the link assumption. In this section, we empirically investigate this question.

We generated the 200 samples with the same procedure specified in Section 4.1, but with the indegree constraint $d = 2$, and except that identity link function $g_j(\eta) = \eta$ and the range of parameters was $\theta_{jk} \in [-1.5, -0.5] \cup [0.5, 1.5]$. We note that the link function must be positive, but we allow the negative value of θ_{jk} by randomly choosing $\theta_j \in [1, 10]$. If any Poisson rate parameter is negative, we regenerated the parameters.

In Fig. 4, we compare the MRS to state-of-the art ODS, GES and MMHC algorithms for varying sample



(a) Varying n for $p = 100, d = 5$ (b) Varying n for $p = 500, d = 5$ (c) Varying p for $n = 500, d = 5$

Figure 5: Comparison of the MRS algorithm to the GES and MMHC algorithms in terms of the running time with respect to node size p and sample size n

size $n \in \{25, 50, \dots, 250\}$, and node size $p \in \{20, 100\}$. Fig. 4 shows that the MRS algorithm consistently recovers the true graph, and hence, we can see that the MRS algorithm is not so sensitive to deviations from the link assumption. Comparing it to the ODS algorithm, the MRS algorithm shows slightly worse performance because the ODS algorithm is designed to learn general Poisson DAG models with any type of link functions. However, we can see that the MRS algorithm still performs better than the greedy search-based methods in both average precision and recall.

4.3 Computational Complexity

Fig. 5 compares the run-time of the MRS, GES, and MMHC algorithms for learning Poisson SEMs with indegree $d = 5$ by varying sample size $n \in \{100, 200, \dots, 1000\}$ with fixed node size $p \in \{100, 500\}$, and varying node size $p \in \{10, 20, \dots, 200\}$ with fixed sample size $n = 500$. Fig. 5 supports the worst case computational complexity $O(np^3)$ discussed in Section 3.1. In addition, it shows that the MRS algorithm is significantly faster than the greedy search-based GES and MMHC algorithms when a sample size is large. The greedy search-based comparison algorithms are ironically not suitable for large-scale graphs due to the huge computational cost. For $p = 500$ and $n = 1000$, the expected run time of the algorithms is about half a day.

5 Real Multivariate Count Data: MLB Statistics

We now apply the MRS algorithm and state-of-the-art ODS and MMHC algorithms to a simple data set that involves multivariate count data that models baseball statistics for Major League Baseball (MLB) players during the 2003 season. To the best of our knowledge, our MRS algorithm is the only algorithm that provides a reliable and scalable approach to non-extreme sparse DAG learning with multivariate count data although it is under strong assumptions. In particular, other approaches, such as PC, MMHC, and approaches based on conditional independence testing, suffer severely from the fact that we are dealing with discrete variables where the number of discrete states is potentially large, or infinite, and represents counts. In addition ODS

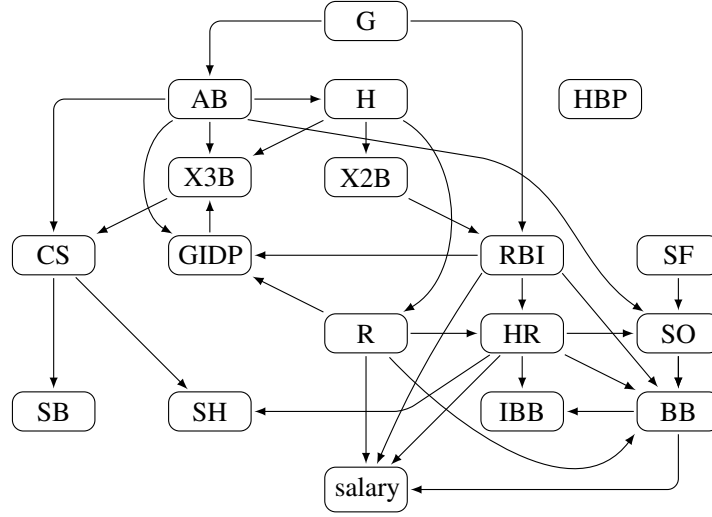


Figure 6: MLB player statistics directed graph estimated by the MRS algorithm for Poisson DAG models.

algorithm cannot deal with a non-sparse graph such as a graph including a hub node. Lastly, the Poisson MRF may provide an extremely complicated graph because it connects all pairs of nodes having a common child like a moralized graph.

Our original data set consists of 800 MLB player salary and batting statistics from the 2003 season (see R package Lahman [30] for detailed information). The data set contains 23 covariates: Salary, Number of: Games Played (G), At Bats (AB), Runs (R), Hits (H), Doubles (X2B), Triples (X3B), Home Runs (HR), Runs Batted In (RBI), Stolen Bases (SB), times Caught Stealing (CS), Bases on Balls (BB), Strikeouts (SO), Intentional Walks (IBB), times Hit by Pitch (HBP), Sacrifice Hits (SH), Sacrifice Flies (SF), and times Grounded into Double Plays (GIDP), plus Player ID, Year ID, Stint, Team ID, and League ID. However, we eliminated Player ID, Year ID, Stint, Team ID, and League ID because our focus is to find the directional or causal relationships between salary and batting statistics. In addition, we only considered players in the top 25% in terms of the number of games played, because the baseball statistics relationships from players who played only a few games could be uncertain. Therefore, the data set we considered contained 18 variables and 200 observations.

We assumed each node to a conditional distribution given its parents is Poisson because most MLB statistics, except for salary, reflect the number of successes or attempts that were counted during the season. Hence, we applied the MRS algorithm for Poisson DAG models with leave-one-out cross validation to choose the tuning parameters, and we chose the largest value where mean squared error is within 2.5 standard error of the minimum mean squared error, because we prefer a sparse graph containing only legitimate edges.

Fig. 6 shows the directed graph estimated by our MRS algorithm. The estimated graph reveals clear causal/directional relationships between batting statistics. This makes sense, because players with larger

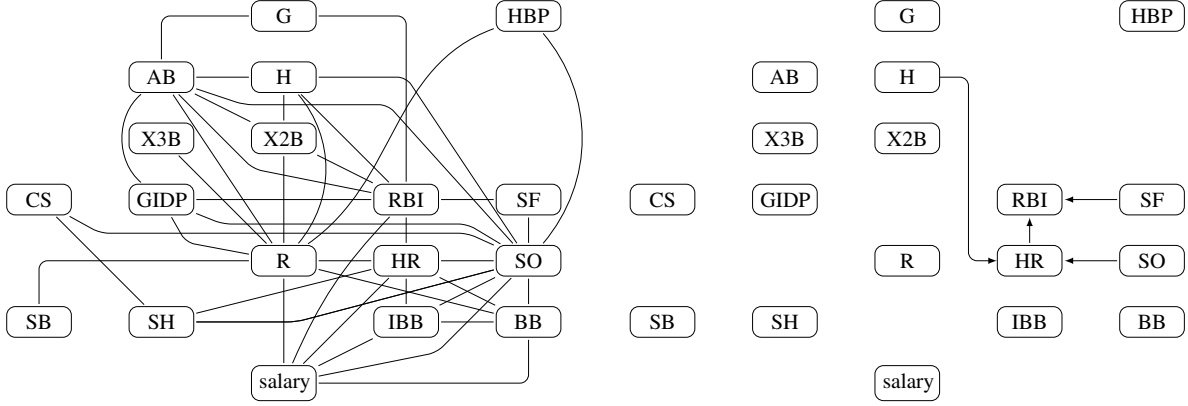


Figure 7: MLB player statistics undirected graph estimated by ℓ_1 -penalized likelihood regression (left) and a directed acyclic graph estimated by the MMHC algorithm (right).

numbers of HR, BB, RBI, and/or R have a better salary. The more games played, or the more batting chances, the higher H, BB, SO, RBI, and other statistics. Moreover, the higher the total number of hits, the more X2Bs, X3Bs, Rs and the fewer SOs. Players with more home runs and base on balls get intentional walks more frequently. Lastly, the more stolen bases are attempted, the more they are caught stealing, because there is no success without failure.

We acknowledge that our proposed DAG model returns many errors due to restrictive assumptions that are not completely satisfied by the real data. However, the benefit is best seen by comparing MRS to other DAG learning approaches and an undirected graphical model for multivariate count data. In particular, we applied Poisson undirected graphical models [25] in which ℓ_1 -regularized Poisson regressions are applied. We provide the estimated undirected graph with the largest tuning parameter where mean squared of error is within 2.5 standard error of the minimum mean squared error. The estimated undirected graph in Fig. 7 (left side) shows that a lot of nodes are connected by edges, that many edges are unexplainable, and that some legitimate edges are missing (e.g., [H, X3B], [SB, CS] are not connected), because the Poisson undirected graphical model only permits negative conditional relationships, whereas most variables are positively correlated. Hence, it may not be useful to understand the relationships between MLB statistics.

We also compared the MMHC algorithm. As discussed, the MMHC algorithm does not guarantee finding a complete directed graph, and prefers a sparser graph when the faithfulness assumption is violated, which often arises in finite sample settings [11]. Hence, the estimated directed graph in Fig. 7 (right side) is extremely sparse, with only four directed edges: [H, HR], [SO, HR], [HR, RBI], and [SF, RBI]. Lastly, ODS algorithm failed to be implemented as expected because of some hub nodes such as the number of games, at bats, and runs batted in.

Since our method is the first identifiability result for the strongly correlated count data to the best of

our knowledge, our method better identifies the directional/causal relationships between MLB statistics. However, we acknowledge that, like most other DAG-learning approaches, very strong assumptions are required for reliable recovery.

References

- [1] J. O. Kephart and S. R. White, “Directed-graph epidemiological models of computer viruses,” in *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on.* IEEE, 1991, pp. 343–359.
- [2] N. Friedman, M. Linial, I. Nachman, and D. Pe’er, “Using bayesian networks to analyze expression data,” *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.
- [3] K. Doya, *Bayesian brain: Probabilistic approaches to neural coding.* MIT press, 2007.
- [4] J. Peters and P. Bühlmann, “Identifiability of gaussian structural equation models with equal error variances,” *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.
- [5] M. Frydenberg, “The chain graph markov property,” *Scandinavian Journal of Statistics*, pp. 333–353, 1990.
- [6] D. Heckerman, D. Geiger, and D. M. Chickering, “Learning Bayesian networks: The combination of knowledge and statistical data,” *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.
- [7] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search.* MIT press, 2000.
- [8] D. M. Chickering, “Optimal structure identification with greedy search,” *The Journal of Machine Learning Research*, vol. 3, pp. 507–554, 2003.
- [9] I. Tsamardinos and C. F. Aliferis, “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, 2003.
- [10] J. Zhang and P. Spirtes, “The three faces of faithfulness,” *Synthese*, vol. 193, no. 4, pp. 1011–1027, 2016.
- [11] C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu, “Geometry of the faithfulness assumption in causal inference,” *The Annals of Statistics*, pp. 436–463, 2013.
- [12] F. Harary, “New directions in the theory of graphs,” MICHIGAN UNIV ANN ARBOR DEPT OF MATHEMATICS, Tech. Rep., 1973.
- [13] D. M. Chickering, D. Geiger, D. Heckerman *et al.*, “Learning bayesian networks is np-hard,” Citeseer, Tech. Rep., 1994.
- [14] D. M. Chickering, “Learning bayesian networks is np-complete,” in *Learning from data.* Springer, 1996, pp. 121–130.

- [15] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, “The max-min hill-climbing bayesian network structure learning algorithm,” *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.
- [16] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, “A linear non-Gaussian acyclic model for causal discovery,” *The Journal of Machine Learning Research*, vol. 7, pp. 2003–2030, 2006.
- [17] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” in *Advances in neural information processing systems*, 2009, pp. 689–696.
- [18] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf, “Identifiability of causal graphs using functional models.” Corvallis, OR, USA: AUAI Press, Jul. 2011, pp. 589–598.
- [19] G. Park and G. Raskutti, “Learning large-scale poisson dag models based on overdispersion scoring,” in *Advances in Neural Information Processing Systems*, 2015, pp. 631–639.
- [20] —, “Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods),” *Journal of Machine Learning Research*, vol. 18, no. 224, pp. 1–44, 2018.
- [21] P. Bühlmann, J. Peters, J. Ernest *et al.*, “Cam: Causal additive models, high-dimensional order search and penalized regression,” *The Annals of Statistics*, vol. 42, no. 6, pp. 2526–2556, 2014.
- [22] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the lasso,” *The Annals of Statistics*, pp. 1436–1462, 2006.
- [23] M. J. Wainwright, J. D. Lafferty, and P. K. Ravikumar, “High-dimensional graphical model selection using ℓ_1 -regularized logistic regression,” in *Advances in neural information processing systems*, 2006, pp. 1465–1472.
- [24] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu *et al.*, “High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence,” *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.
- [25] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, “Graphical models via univariate exponential family distributions,” *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3813–3847, 2015.
- [26] E. Yang, P. K. Ravikumar, G. I. Allen, and Z. Liu, “On poisson graphical models,” in *Advances in Neural Information Processing Systems*, 2013, pp. 1718–1726.
- [27] S. L. Lauritzen, *Graphical models*. Oxford University Press, 1996.
- [28] J. Friedman, T. Hastie, and R. Tibshirani, “glmnet: Lasso and elastic-net regularized generalized linear models,” *R package version*, vol. 1, no. 4, 2009.

- [29] —, “Regularization paths for generalized linear models via coordinate descent,” *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.
- [30] M. Friendly, *Lahman: Sean 'Lahman' Baseball Database*, 2017, r package version 6.0-0. [Online]. Available: <https://CRAN.R-project.org/package=Lahman>

A Proof for Theorem 2.1

Proof. This proof follows the same strategies as [20]. Without loss of generality, we assume the true ordering is unique, and $\pi = (\pi_1, \dots, \pi_p)$. For simplicity, we define $X_{1:j} = (X_{\pi_1}, X_{\pi_2}, \dots, X_{\pi_j})$ and $X_{1:0} = \emptyset$. In addition we define a mean-variance related function, $f(\mu) = \mu + \mu^2$.

We now prove identifiability of Poisson DAG models using mathematical induction:

Step (1) For the first step π_1 , we have $\mathbb{E}(X_{\pi_1}^2) = f(\mathbb{E}(X_{\pi_1}))$, while for any node $j \in V \setminus \{\pi_1\}$:

$$\mathbb{E}(X_j^2) = \mathbb{E}(\mathbb{E}(X_j^2 \mid X_{\text{Pa}(j)})) = \mathbb{E}(f(\mathbb{E}(X_j \mid X_{\text{Pa}(j)}))) > f(\mathbb{E}(X_j)),$$

by Jensen's inequality, and the equality never holds unless $\mathbb{E}(X_j \mid X_{\text{Pa}(j)})$ is constant (a degenerate random variable), which violates the identifiability assumption in Theorem 2.1. Hence, we can determine π_1 as the first element of the causal ordering.

Step (m-1) For the $(m-1)^{\text{th}}$ element of the ordering, assume that the first $m-1$ elements of the ordering and their parents are correctly estimated.

Step (m) Now, we consider the m^{th} element of the causal ordering and its parents. It is clear that π_m achieves $\mathbb{E}(X_{\pi_m}^2) = \mathbb{E}(f(\mathbb{E}(X_{\pi_m} \mid X_{1:(m-1)})))$. However, for $j \in \{\pi_{m+1}, \dots, \pi_p\}$,

$$\begin{aligned} \mathbb{E}(X_j^2 \mid X_{1:(m-1)}) &= \mathbb{E}(\mathbb{E}(X_j^2 \mid X_{\text{Pa}(j)}) \mid X_{1:(m-1)}) \\ &= \mathbb{E}(f(\mathbb{E}(X_j \mid X_{\text{Pa}(j)}) \mid X_{1:(m-1)})) \geq f(\mathbb{E}(X_j \mid X_{1:(m-1)})). \end{aligned}$$

by Jensen's inequality, and the equality cannot hold unless $\mathbb{E}(X_j \mid X_{\text{Pa}(j)})$ is a function of $X_{1:(m-1)}$ that is equivalent to $\text{Var}(\mathbb{E}(X_j \mid X_{\text{Pa}(j)}) \mid X_{1:(m-1)}) = 0$ that is the identifiability condition in [19].

However, in this paper, we assume $\mathbb{E}(\text{Var}(\mathbb{E}(X_j \mid X_{\text{Pa}(j)}) \mid X_{1:(m-1)})) > 0$, which allows $\text{Var}(\mathbb{E}(X_j \mid X_{\text{Pa}(j)}) \mid X_{1:(m-1)}) = 0$ for some $X_{1:(m-1)} \in \mathcal{X}_{1:(m-1)}$. By taking the expectation on both sides, we still obtain $\mathbb{E}(X_j^2) > \mathbb{E}(f(\mathbb{E}(X_j \mid X_{1:(m-1)})))$ under our identifiability condition. Hence, we can estimate a true m^{th} component of the ordering π_m .

In terms of the parent search, it is clear that by conditional independence relations naturally encoded by factorization (1) $\mathbb{E}(X_{\pi_m}^2) = \mathbb{E}(f(\mathbb{E}(X_{\pi_m} \mid X_{1:(m-1)}))) = \mathbb{E}(f(\mathbb{E}(X_{\pi_m} \mid X_{\text{Pa}(\pi_m)})))$. Hence we can also choose the minimum conditioning set from among $X_{1:(m-1)}$ as the parents of π_m such that the above moments relation holds. By mathematical induction, this completes the proof. \square

B Proof for Theorem 3.5: Parents Recovery

Proof. We provide the proof for Theorem 3.5 using the *primal-dual witness method* that is also used many other works [22, 23, 24, 25]. In this proof, we show in Appendix B, the error probability for the recovery of the parents of a node π_j from among all the nodes given the partial ordering $(\pi_1, \pi_2, \dots, \pi_{j-1})$ via

ℓ_1 -regularized regression. In Appendix C, the error bounds for the recovery of the ordering both via ℓ_1 -regularized regression.

Without loss of generality, let the true ordering be $\pi = (1, 2, \dots, p)$, and hence, $\pi_{1:j} = (\pi_1, \pi_2, \dots, \pi_j) = (1, 2, \dots, j)$. For ease of notation, $[\cdot]_k$ and $[\cdot]_S$ denote parameters corresponding to variable X_k and random vector X_S , respectively. In order to make the arguments easier to understand, we restate the negative log likelihood (9) and related arguments.

First, we define a new parameter vector $\theta_{S_j} \in \mathbb{R}^{|S_j|}$ without parameter θ_j corresponding to the node j since the node j is not penalized in regression problem (8). Then, the conditional negative log-likelihood of the GLM for X_j given X_{S_j} can be written as:

$$\ell_j^{S_j}(\theta_{S_j}; X^{1:n}) := \frac{1}{n} \sum_{i=1}^n \left(-X_j^{(i)} \langle \theta_{S_j}, X_{S_j}^{(i)} \rangle + \exp(\langle \theta_{S_j}, X_{S_j}^{(i)} \rangle) \right), \quad (11)$$

where $\langle \cdot, \cdot \rangle$ is an inner product.

We also define $\theta_{S_j}^* \in \Theta_{S_j}$ for Equation (10), which denotes the true solution of the following GLM problem, where $\Theta_{S_j} := \{\theta_{S_j} \in \mathbb{R}^{|S_j|} : [\theta_{S_j}]_k = 0 \text{ for } k \notin \text{Pa}(j)\}$:

$$\theta_{S_j}^* := \arg \min_{\theta \in \Theta_{S_j}} \mathbb{E} \left(-X_j(\langle \theta, X_{S_j} \rangle) + \exp(\langle \theta, X_{S_j} \rangle) \right). \quad (12)$$

The main goal of the proof is to find the unique minimizer of the following convex problem:

$$\hat{\theta}_{S_j} := \arg \min_{\theta \in \mathbb{R}^{|S_j|}} \mathcal{L}_j(\theta, \lambda_j) = \arg \min_{\theta \in \mathbb{R}^{|S_j|}} \{ \ell_j^{S_j}(\theta; X^{1:n}) + \lambda_j \|\theta\|_1 \}. \quad (13)$$

By setting the *sub-differential* to 0, $\hat{\theta}_{S_j}$ satisfies the following condition:

$$\nabla_{\theta} \mathcal{L}_j^{S_j}(\hat{\theta}_{S_j}, \lambda_j) = \nabla_{\theta} \ell_j^{S_j}(\hat{\theta}_{S_j}; X^{1:n}) + \lambda_j \hat{Z}_j^{S_j} = 0 \quad (14)$$

where $\hat{Z}_j^{S_j} \in \mathbb{R}^{|S_j|}$ and $[\hat{Z}_j^{S_j}]_t = \text{sign}([\hat{\theta}_{S_j}]_t)$ if $t \in S_j \cap \text{Pa}(j)$, otherwise $[\hat{Z}_j^{S_j}]_t < 1$.

Lemma B.1 directly follows from prior works [25], where each node's conditional distribution is in the form of a generalized linear model. For notational convenience, let $T_j = S_j \cap \text{Pa}(j)$.

Lemma B.1 (Uniqueness of Solution, Lemma 8 in [25]). *Suppose that $|[\hat{Z}_j^{S_j}]_t| < 1$ for $t \notin T_j$. Then, the solution $\hat{\theta}_{S_j}$ of Equation (13) satisfies $[\hat{\theta}_{S_j}]_t = 0$ for all $t \notin T_j$. Furthermore, if the sub-matrix of Hessian matrix $Q_{T_j T_j}^{S_j}$ is invertible, then $\hat{\theta}_{S_j}$ is unique.*

The remainder of the proof is to show $|[\hat{Z}_j^{S_j}]_t| < 1$ for all $t \notin T_j$. Note that the restricted solution in Equation (21) is $(\tilde{\theta}_{S_j}, \tilde{Z}_j^{S_j})$ and the unrestricted solution in Equation (13) is $(\hat{\theta}_{S_j}, \hat{Z}_j^{S_j})$. Equation (14) with the dual solution can be represented by

$$\nabla^2 \ell_j^{S_j}(\theta_{S_j}^*; X^{1:n})(\tilde{\theta}_{S_j} - \theta_{S_j}^*) = -\lambda_j \tilde{Z}_j^{S_j} - W_j^{S_j} + R_j^{S_j} \quad (15)$$

where

(a) $W_j^{S_j}$ is the sample score function:

$$W_j^{S_j} := -\nabla \ell_j(\theta_{S_j}^*; X^{1:n}). \quad (16)$$

(b) $R_j^{S_j} = (R_{jk}^{S_j})_{k \in S_j}$ and $R_{jk}^{S_j}$ is the remainder term by applying the coordinate-wise mean value theorem:

$$R_{jk}^{S_j} := [\nabla^2 \ell_j^{S_j}(\theta_{S_j}^*; X^{1:n}) - \nabla^2 \ell_j^{S_j}(\bar{\theta}_{S_j}; X^{1:n})]_k^T (\tilde{\theta}_{S_j} - \theta_{S_j}^*). \quad (17)$$

Here $\bar{\theta}_{S_j}$ is a vector on the line between $\tilde{\theta}_{S_j}$ and $\theta_{S_j}^*$, and $[\cdot]_k^T$ is the row of a matrix corresponding to variable X_k .

Then, the following proposition provides a sufficient condition to control $\tilde{Z}_j^{S_j}$.

Proposition B.2. *If $\max(\|W_j^{S_j}\|_\infty, \|R_j^{S_j}\|_\infty) \leq \frac{\lambda_j \alpha}{4(2-\alpha)}$, then $|\tilde{Z}_j^{S_j}|_t| < 1$ for all $t \notin T_j$.*

Next, we introduce the following three lemmas under Assumptions 3.1, 3.2, and 3.3 to show that conditions in Proposition B.2 hold. For ease of notation, let $\eta = \max\{n, p\}$, $M = \max\{M_{\min}, (M_{\min})^{e-1}\}$, $\tilde{\theta}_S = [\tilde{\theta}_{S_j}]_{T_j}$, $\tilde{Z}_S = [\tilde{Z}_j^{S_j}]_{T_j}$, $\tilde{\theta}_{S^c} = [\tilde{\theta}_{S_j}]_{S_j \setminus T_j}$, and $\tilde{Z}_{S^c} = [\tilde{Z}_j^{S_j}]_{S_j \setminus T_j}$.

Lemma B.3. *For any $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:j-1}\}$ and $\lambda_j \geq \frac{64\sqrt{2}(2-\alpha)}{\alpha} \frac{\log^2 \eta}{\kappa_1(n,p)}$ for some $\alpha \in (0, 1]$,*

$$P\left(\frac{\|W_j^{S_j}\|_\infty}{\lambda_j} \leq \frac{\alpha}{4(2-\alpha)}\right) \geq 1 - 2d \cdot \exp\left(-\frac{n}{\kappa_1(n,p)^2}\right) - M \cdot \eta^{-2}.$$

where $\kappa_1(n, p)$ is an arbitrary function of n and p .

Lemma B.4. *Suppose that for all $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:j-1}\}$, $\|W_j^{S_j}\|_\infty \leq \frac{\lambda_j}{4}$. Then, for $\lambda_j \leq \frac{\rho_{\min}^2}{160\rho_{\max}d \log^2 \eta}$,*

$$P\left(\|\tilde{\theta}_S - \theta_S^*\|_2 \leq \frac{5}{\rho_{\min}} \sqrt{d} \lambda_j\right) \geq 1 - M \cdot \eta^{-2}.$$

Lemma B.5. *Suppose that for all $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:j-1}\}$, $\|W_j^{S_j}\|_\infty \leq \frac{\lambda_j}{4}$. Then, for $\lambda_j \leq \frac{\alpha \rho_{\min}^2}{16000(2-\alpha)\rho_{\max}d \log^2 \eta}$ and $\alpha \in (0, 1]$,*

$$P\left(\frac{\|R_j^{S_j}\|_\infty}{\lambda_j} \leq \frac{\alpha}{4(2-\alpha)}\right) \geq 1 - M \cdot \eta^{-2}.$$

The rest of the proof is straightforward using Lemmas B.3, B.4, and B.5. Consider the choice of regularization parameter $\lambda_{j0} = \frac{128(2-\alpha)}{\alpha} \frac{\log^2 \eta}{\kappa_1(n,p)}$, where $\kappa_1(n, p) \geq \frac{2048 \cdot 10^3 (2-\alpha)^2}{\alpha^2} \frac{\rho_{\max}}{\rho_{\min}^2} d \log^4 \eta$ ensuring that $\frac{64\sqrt{2}(2-\alpha)}{\alpha} \frac{\log^2 \eta}{\kappa_1(n,p)} \leq \lambda_{j0} \leq \frac{\alpha \rho_{\min}^2}{16 \cdot 10^3 (2-\alpha) \rho_{\max} d \log^2 \eta}$ for any $\alpha \in (0, 1]$. Hence, if we set $\kappa_1(n, p) = C_{\max} d \log^4 \eta$ where $C_{\max} = \frac{2048 \cdot 10^3 (2-\alpha)^2}{\alpha^2} \frac{\rho_{\max}}{\rho_{\min}^2}$, then all conditions for Lemma B.3, B.4, and B.5 are satisfied. Therefore,

$$\|\tilde{Z}_{S^c}\|_\infty \leq (1-\alpha) + (2-\alpha) \left[\frac{\|W_j^{S_j}\|_\infty}{\lambda_j} + \frac{\|R_j^{S_j}\|_\infty}{\lambda_j} \right] \leq (1-\alpha) + \frac{\alpha}{4} + \frac{\alpha}{4} < 1, \quad (18)$$

with a probability of at least $1 - 2d \cdot \exp\left(-\frac{n}{(C_{\max} d \log^4 \eta)^2}\right) - 3M\eta^{-2}$.

Proposition B.6. Suppose that, for any $j \in V$, partial ordering (π_1, \dots, π_j) is correctly estimated. If $\min_{t \in S} [\theta_S^*]_t \geq \frac{10}{\rho_{\min}} \sqrt{d} \lambda_j$ for all $j \in V$,

$$\text{supp}(\hat{\theta}_{S_j}) = \text{Pa}(j).$$

Proposition B.6 guarantees that ℓ_1 -regularized likelihood regression recovers the parents for each node with a high probability. Since there are p regression problems, for any $\epsilon > 0$, there exists a positive constant $C_\epsilon > 0$ such that if $n \geq C_\epsilon(d^2 \log^9 \eta)$,

$$P(\hat{G} = G) \geq 1 - 2d \cdot \exp\left(\left(1 - \frac{C_\epsilon}{C_{\max}^2}\right) \log \eta\right) - 3M\eta^{-1} \geq 1 - \epsilon.$$

□

C Proof for Theorem 3.5: Ordering Recovery

Proof. We begin by reintroducing some necessary notations and definitions to make the proof concise. Without loss of generality, assume that the true ordering is unique and $\pi = (\pi_1, \dots, \pi_p) = (1, 2, \dots, p)$. For notational convenience, we define $X_{1:j} = (X_{\pi_1}, X_{\pi_2}, \dots, X_{\pi_j}) = (X_1, X_2, \dots, X_j)$ and $X_{1:0} = \emptyset$. We restate the mean-variance ratio scores for a node k and the j th element of the ordering:

$$\mathcal{S}(j, k) := \frac{\mathbb{E}(X_k^2)}{\mathbb{E}(f(\mathbb{E}(X_k | X_{1:(j-1)})))} \quad \text{and} \quad \hat{\mathcal{S}}(j, k) := \frac{\hat{\mathbb{E}}(X_k^2)}{\hat{\mathbb{E}}(f(\hat{\mathbb{E}}(X_k | X_{1:(j-1)})))},$$

where $f(\mu) := \mu + \mu^2$, $\mathbb{E}(X_k | X_{S_k}) = \exp(\theta_k^* + \sum_{t \in S_k} \theta_{kt}^* X_t)$, and $\hat{\mathbb{E}}(X_k | X_{S_k}) = \exp(\hat{\theta}_k + \sum_{t \in S_k} \hat{\theta}_{kt} X_t)$ where $\theta_{S_k}^* = (\theta_k^*, \theta_{kt}^*)$ and $\hat{\theta}_{S_k} = (\hat{\theta}_k, \hat{\theta}_{kt})$ are the solutions of the problem (10) and of the ℓ_1 -regularized GLM (8), respectively. In addition, we use the unbiased method-of-moment estimator for a marginal expectation, $\hat{\mathbb{E}}(X_k^2) = \frac{1}{n} \sum_{i=1}^n (X_k^{(i)})^2$.

We define the following necessary events: For each node $j \in V$, $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:(j-1)}\}$ and any $\epsilon_1 > 0$;

$$\begin{aligned} \xi_1 &:= \left\{ \max_{j \in V} \max_{i \in \{1, 2, \dots, n\}} |X_j^{(i)}| < 4 \log \eta \right\} \\ \zeta_1 &:= \left\{ \max_{j=1, \dots, p-1} \max_{k=j, \dots, p} |\mathcal{S}(j, \pi_k) - \hat{\mathcal{S}}(j, \pi_k)| > \frac{M_{\min}}{2} \right\} \\ \zeta_2 &:= \left\{ \max_{j \in V} |\hat{\mathbb{E}}(X_j^2) - \mathbb{E}(X_j^2)| < \epsilon_1 \right\} \\ \zeta_3 &:= \left\{ \max_{j \in V} \left| \hat{\mathbb{E}}\left(f\left(\hat{\mathbb{E}}(X_j | X_{S_j})\right)\right) - \mathbb{E}\left(f\left(\hat{\mathbb{E}}(X_j | X_{S_j})\right)\right) \right| < \epsilon_1 \right\} \\ \zeta_4 &:= \left\{ \max_{j \in V} \left| \mathbb{E}\left(f\left(\hat{\mathbb{E}}(X_j | X_{S_j})\right)\right) - \mathbb{E}\left(f\left(\mathbb{E}(X_j | X_{S_j})\right)\right) \right| < \epsilon_1 \right\}, \end{aligned}$$

where $f(\mu) := \mu + \mu^2$.

We begin by proving that our algorithm recovers the ordering of a Poisson SEM in the high-dimensional setting. The probability that ordering is correctly estimated from our method can be written as

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

The first term in the above probability is always satisfied because $\mathcal{S}(j, \pi_k) - \mathcal{S}(j, \pi_j) > (1 + M_{\min}) - 1 = M_{\min}$ from Assumption 3.4. Hence, the lower bound of the probability that ordering is correctly estimated using our method is reduced to

$$\begin{aligned}
P(\hat{\pi} = \pi) &\geq P\left(\max_{j=1, \dots, p-1} \max_{k=j+1, \dots, p} \left| \mathcal{S}(j, \pi_k) - \widehat{\mathcal{S}}(j, \pi_k) \right| < \frac{M_{\min}}{2}\right) \\
&= 1 - P(\zeta_1) \\
&= 1 - P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4) P(\zeta_2, \zeta_3, \zeta_4) - P(\zeta_1 \mid (\zeta_2, \zeta_3, \zeta_4)^c) P((\zeta_2, \zeta_3, \zeta_4)^c) \\
&\geq 1 - P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4) - P((\zeta_2, \zeta_3, \zeta_4)^c) \\
&= 1 - P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4) - P((\zeta_2, \zeta_3, \zeta_4)^c \mid \xi_1) P(\xi_1) - P((\zeta_2, \zeta_3, \zeta_4)^c \mid \xi_1^c) P(\xi_1^c) \\
&\geq 1 - P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4) - P((\zeta_2, \zeta_3, \zeta_4)^c \mid \xi_1) - P(\xi_1^c) \\
&\geq 1 - \underbrace{P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4)}_{\text{Lem C.1}} - \underbrace{P(\zeta_2^c \mid \xi_1) - P(\zeta_3^c \mid \xi_1) - P(\zeta_4^c \mid \xi_1)}_{\text{Lem C.2}} - \underbrace{P(\xi_1^c)}_{\text{Prop D.2}}. \tag{19}
\end{aligned}$$

Next, we introduce the following two lemmas to show the lower bound of the probability in (19) as a function of the triple (n, p, d) :

Lemma C.1. *Given the sets $\zeta_2, \zeta_3, \zeta_4$ and under Assumption 3.4, $P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4) = 0$ if for some small ϵ_1 such that for any non-empty $Pa_0(j) \subset Pa(j)$ and $S_j \subset \pi_{1:(j-1)} \setminus Pa_0(j)$,*

$$\epsilon_1 < \min \left\{ \frac{\mathbb{E}(X_j^2) M_{\min}}{2(M_{\min} + 3)(M_{\min} + 1)}, \frac{M_{\min}}{6} \frac{\mathbb{E}(f(\mathbb{E}(X_j \mid X_{S_j})))^2}{\mathbb{E}(X_j^2)} \right\},$$

where $f(\mu) = \mu + \mu^2$.

The condition in Lemma C.1 implies that if ϵ_1 is sufficiently small, the estimated score is close to the true score value.

The second lemma shows the error bound for the consistency of the estimators.

Lemma C.2. For any $\epsilon_1 > 0$ and

(i) For ζ_2 , $P(\zeta_2^c \mid \xi_1) \leq 1 - 2 \cdot p \cdot \exp \left\{ -\frac{2n\epsilon_1^2}{(4\log^2 \eta)^2} \right\}.$

(ii) For ζ_3 , there exist some positive constants C_{\max} and D_{\max} such that

$$P(\zeta_3^c \mid \xi_1) \leq 1 - 2 \cdot p \cdot d \cdot \exp \left(-\frac{n}{C_{\max}^2 d^2 \log^8 \eta} \right) - M \cdot \eta^{-2} - 2 \cdot p \cdot \exp \left\{ -\frac{n\epsilon_1^2}{D_{\max} \log^4 \eta} \right\}.$$

(iii) For ζ_4 , $P(\zeta_4^c \mid \xi_1) = 0.$

Therefore, we complete the proof: our method recovers the true ordering at least of

$$P(\hat{\pi} = \pi) \geq 1 - C_1 p \cdot d \cdot \exp \left(-C_2 \frac{n\epsilon_1^2}{d^2 \log^8 \eta} \right) - C_3 \cdot \eta^{-2}.$$

for some positive constants C_1, C_2 , and C_3 . □

D Important Propositions

We begin by introducing important propositions to control the tail behavior for the distribution of each node, which are required to prove the lemmas.

Proposition D.1. For any node $j \in V$,

$$\mathbb{E}(\exp(X_j)) < M.$$

where $M = \max\{M_{\max}, (M_{\max})^{e-1}\}.$

Proof. Using moment generating function of Poisson and Assumption 3.3, we have

$$\mathbb{E}(\exp(X_j)) = \mathbb{E}(\mathbb{E}(\exp(X_j) \mid X_{\text{Pa}(j)})) = \mathbb{E}(e^{(e-1)\exp(\theta_j^{\text{Pa}(j)} + \sum_{k \in \text{Pa}(j)} \theta_{jk}^{\text{Pa}(j)} X_k)}) < (M_{\max})^{e-1}.$$

□

Proposition D.2. Under Assumption 3.3, $P(\xi_1^c) \leq M \cdot \eta^{-2}.$

Proof. Recall that $X_j \in \mathbb{Z}^+ \cup \{0\}$. Applying the union bound and the Chernoff bound,

$$P(\xi_1^c) \leq n \cdot p \cdot \max_{j \in V} \max_{i \in \{1, \dots, n\}} P(X_j^{(i)} > 4 \log \eta) \leq \eta^{-2} \max_{i,j} \mathbb{E}[\exp(X_j^{(i)})].$$

By Assumption 3.3 and Proposition D.1, we obtain $\max_{i,j} \mathbb{E}(\exp(X_j^{(i)})) < M$, which completes the proof. □

Proposition D.3. For any $j \in V$ and $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:(j-1)}\}$, the solution $\theta_{S_j}^*$ in Equation (12) satisfies

$$P\left(\exp(\langle \theta_{S_j}^*, X_{S_j} \rangle) > 4 \log \eta\right) \leq M_{\max} \cdot \eta^{-4}.$$

Proof. From Assumption 3.3, we have

$$P\left(\exp(\langle \theta_{S_j}^*, X_{S_j} \rangle) > 4 \log \eta\right) \leq \frac{\mathbb{E}(\exp(\exp(\langle \theta_{S_j}^*, X_{S_j} \rangle)))}{\eta^4} \leq M_{\max} \cdot \eta^{-4}. \quad (20)$$

□

Proposition D.4. For given $j \in V$ and $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:(j-1)}\}$, the solution $\hat{\theta}_{S_j}$ in Equation (13) satisfies

$$P\left(\frac{1}{n} \sum_{i=1}^n \exp(\langle \hat{\theta}_{S_j}, X_{S_j}^{(i)} \rangle) \geq 4 \log \eta\right) \leq M \eta^{-2}.$$

Proof. By the first-order optimality condition of $\mathcal{L}_j^{S_j}(\theta_{S_j}, X^{1:n})$ in Equation (13), we have

$$\begin{aligned} \sum_{i=1}^n X_j^{(i)} &= \sum_{i=1}^n \exp(\langle \hat{\theta}_{S_j}, X_{S_j}^{(i)} \rangle) \\ \sum_{i=1}^n X_j^{(i)} X_k^{(i)} &= \sum_{i=1}^n \exp(\langle \hat{\theta}_{S_j}, X_{S_j}^{(i)} \rangle X_k^{(i)}) + \lambda_j \text{sign}([\hat{\theta}_{S_j}]_k). \end{aligned}$$

From Proposition D.2, we have

$$\begin{aligned} P\left(\frac{1}{n} \sum_{i=1}^n \exp(\langle \hat{\theta}_{S_j}, X_{S_j}^{(i)} \rangle) \geq 4 \log \eta\right) &= P\left(\frac{1}{n} \sum_{i=1}^n X_j^{(i)} \geq 4 \log \eta\right) \\ &\leq P\left(\frac{1}{n} \sum_{i=1}^n X_j^{(i)} \geq 4 \log \eta \mid \xi_1\right) + P(\xi_1^c) \\ &\leq M \eta^{-2}. \end{aligned}$$

□

E Proof for Propositions

E.1 Proof for Proposition B.2

Proof. We note that $\tilde{\theta}_{S^c} = (0, 0, \dots, 0)^T \in \mathbb{R}^{|S^c|}$ in our primal-dual construction. To improve readability, we let $\theta_S = [\theta_{S_j}]_{T_j}$, $\theta_{S^c} = [\theta_{S_j}]_{S_j \setminus T_j}$, and $A_S = [A_j^{S_j}]_{T_j}$ and $A_{S^c} = [A_j^{S_j}]_{S_j \setminus T_j}$ for a given $T_j = S_j \cap \text{Pa}(j)$. With these notations, W_S and R_S are sub-vectors of $W_j^{S_j}$ and $R_j^{S_j}$ corresponding to variables X_S , respectively.

We can restate condition (15) in block form as follows:

$$\begin{aligned} Q_{S^c S}[\tilde{\theta}_S - \theta_S^*] &= W_{S^c} - \lambda_j \tilde{Z}_{S^c} + R_{S^c}, \\ Q_{SS}[\tilde{\theta}_S - \theta_S^*] &= W_S - \lambda_j \tilde{Z}_S + R_S. \end{aligned}$$

Since Q_{SS} is invertible, the above equations can be rewritten as

$$Q_{S^c S} Q_{SS}^{-1} [W_S - \lambda_j \tilde{Z}_S - R_S] = W_{S^c} - \lambda_j \tilde{Z}_{S^c} - R_{S^c}.$$

Therefore,

$$[W_{S^c} - R_{S^c}] - Q_{S^c S} Q_{SS}^{-1} [W_S - R_S] + \lambda_j Q_{S^c S} Q_{SS}^{-1} \tilde{Z}_S = \lambda_j \tilde{Z}_{S^c}.$$

Taking the ℓ_∞ norm of both sides yields

$$\|\tilde{Z}_{S^c}\|_\infty \leq \|Q_{S^c S} Q_{SS}^{-1}\|_\infty \left[\frac{\|W_S\|_\infty}{\lambda_j} + \frac{\|R_S\|_\infty}{\lambda_j} + 1 \right] + \frac{\|W_{S^c}\|_\infty}{\lambda_j} + \frac{\|R_{S^c}\|_\infty}{\lambda_j}.$$

Recalling Assumption (3.2), we obtain $\|Q_{S^c S} Q_{SS}^{-1}\|_\infty \leq (1 - \alpha)$, and hence, we have

$$\begin{aligned} \|\tilde{Z}_{S^c}\|_\infty &\leq (1 - \alpha) \left[\frac{\|W_S\|_\infty}{\lambda_j} + \frac{\|R_S\|_\infty}{\lambda_j} + 1 \right] + \frac{\|W_{S^c}\|_\infty}{\lambda_j} + \frac{\|R_{S^c}\|_\infty}{\lambda_j} \\ &\leq (1 - \alpha) + (2 - \alpha) \left[\frac{\|W_j^{S_j}\|_\infty}{\lambda_j} + \frac{\|R_j^{S_j}\|_\infty}{\lambda_j} \right]. \end{aligned}$$

If both $\|W_j^{S_j}\|_\infty$ and $\|R_j^{S_j}\|_\infty$ are less than $\frac{\lambda_j \alpha}{4(2-\alpha)}$, as assumed, then

$$\|\tilde{Z}_{S^c}\|_\infty \leq (1 - \alpha) + \frac{\alpha}{2} < 1.$$

□

E.2 Proof for Proposition B.6

Proof. To prove the support of $\hat{\theta}_S$ is not strictly subset the true support X_S , it is sufficient to show that the maximum bias is bounded:

$$\|\hat{\theta}_S - \theta_S^*\|_\infty \leq \frac{\min_{t \in S} [\theta_S^*]_t}{2}.$$

From Lemma B.4, we have, with a high probability,

$$\|\hat{\theta}_S - \theta_S^*\|_\infty \leq \|\hat{\theta}_S - \theta_S^*\|_2 \leq \frac{5}{\rho_{\min}} \sqrt{d} \lambda_j.$$

Therefore, if $\min_{t \in S} [\theta_S^*]_t \geq \frac{10}{\rho_{\min}} \sqrt{d} \lambda_j$,

$$\|\hat{\theta}_S - \theta_S^*\|_\infty \leq \frac{\min_{t \in S} [\theta_S^*]_t}{2}.$$

□

F Proof for Lemmas

F.1 Proof for Lemma B.1

Proof. This lemma can be proved by the same manner developed for the special cases [23, 24]. In addition, this proof is directly from Lemma 8 in [25]. And, we restate the proof in our framework. The main idea of the proof is the *primal-dual-witness* method which asserts that there is a solution to the dual problem $\tilde{\theta}_{S_j} = \hat{\theta}_{S_j}$ if the following Karush-Kuhn-Tucker (KKT) conditions are satisfied.

- (a) We define $\tilde{\theta}_{S_j} \in \Theta_{S_j}$, where $\Theta_{S_j} = \{\theta \in \mathbb{R}^{|S_j|} : \theta_{S^c} = 0\}$ is the solution to the following optimization problem:

$$\tilde{\theta}_{S_j} := \arg \min_{\theta \in \Theta_{S_j}} \mathcal{L}_j^{S_j}(\theta, \lambda_j) = \arg \min_{\theta \in \Theta_{S_j}} \{\ell_j^{S_j}(\theta; X^{1:n}) + \lambda_j \|\theta\|_1\}. \quad (21)$$

- (b) Define $\tilde{Z}_j^{S_j}$ to be a sub-differential for the regularizer $\|\cdot\|_1$ evaluated at $\tilde{\theta}_{S_j}$. For any $t \in S_j \cap \text{Pa}(j)$, $[\tilde{Z}_j^{S_j}]_t = \text{sign}([\tilde{\theta}_{S_j}]_t)$.
- (c) For any $t \notin S_j \cap \text{Pa}(j)$, $|[\tilde{Z}_j^{S_j}]_t| < 1$.

If conditions (a) to (c) are satisfied, $\tilde{\theta}_{S_j} = \hat{\theta}_{S_j}$ meaning that the solution to unrestricted problem (13) is the same as the solution to restricted problem (21) (See [24] for details).

In addition, if the sub-matrix of the Hessian $Q_{S_j}^{S_j}$ is invertible, restricted problem (21) is strictly convex, and hence, $\tilde{\theta}_{S_j}$ is unique. \square

F.2 Proof for Lemma B.3

Proof. In order to improve readability, we omit the superscript S_j if it is understood (i.e., $W_j = W_j^{S_j}$). Each entry of the sample score function W_j in Equation (16) has the form $W_{jt} = \frac{1}{n} \sum_{i=1}^n W_{jt}^{(i)}$ for any $t \in S = \text{Pa}(j) \cap S_j$. In addition, $W_{jt} = 0$ for all $t \notin S$, since $[\theta_{S_j}^*]_t = 0$ by the construction of $\theta_{S_j}^*$ in (11).

Hence simple calculation yields that, for any $t \in S$ and $i \in \{1, 2, \dots, n\}$,

$$W_{jt}^{(i)} = X_t^{(i)} X_j^{(i)} - \exp(\langle \theta_S^*, X_S^{(i)} \rangle) X_t^{(i)},$$

and $(|W_{jt}^{(i)}|)_{i=1}^n$ has mean 0 by using the first-order optimality condition, $\mathbb{E}(X_j) = \mathbb{E}(\exp(\langle \theta_S^*, X_S \rangle))$.

Now, we show that $W_{jt}^{(i)}$ is bounded with a high probability given ξ_1 by using Hoeffding's inequality. The first term is bounded $\max_i X_t^{(i)} X_j^{(i)} < 16 \log^2 \eta$, conditioning on ξ_1 . And the second term is also bounded above $16 \log^2 \eta$ with a high probability of at least $1 - M_{\max} \eta^{-4}$ under Proposition D.3. Therefore, $|W_{jt}^{(i)}|$ is bounded by $32 \log^2 \eta$.

Applying the union bound and Hoeffding's inequality, we have

$$P(\|W_j\|_\infty > \delta, \xi_1) \leq d \cdot \max_{t \in S} P(|W_{jt}| > \delta, \xi_1) \leq 2d \cdot \exp\left(-\frac{2n\delta^2}{32^2 \log^4 \eta}\right).$$

Suppose that $\delta = \frac{\lambda_j \alpha}{4(2-\alpha)}$ and $\lambda_j \geq \frac{4(2-\alpha)}{\alpha} \frac{32 \log^2 \eta}{\sqrt{2} \kappa_1(n, p)}$. Then,

$$P\left(\frac{\|W_j\|_\infty}{\lambda_j} > \frac{\alpha}{4(2-\alpha)}, \xi_1\right) \leq 2d \cdot \exp\left(-\frac{\alpha^2}{16(2-\alpha)^2} \frac{2n\lambda_j^2}{32^2 \log^4 \eta}\right) \leq 2d \cdot \exp\left(\frac{-n}{\kappa_1(n, p)^2}\right). \quad (22)$$

Therefore, we complete the proof:

$$\begin{aligned} P\left(\frac{\|W_j\|_\infty}{\lambda_j} > \frac{\alpha}{4(2-\alpha)}\right) &\leq P\left(\frac{\|W_j\|_\infty}{\lambda_j} > \frac{\alpha}{4(2-\alpha)}, \xi_1\right) + P(\xi_1^c) \\ &\leq 2d \cdot \exp\left(\frac{-n}{\kappa_1(n, p)^2}\right) + M \cdot \eta^{-2}. \end{aligned}$$

□

F.3 Proof for Lemma B.4

Proof. In order to establish error bound $\|\tilde{\theta}_S - \theta_S^*\| \leq B$ for some radius B , several works [22, 23, 24, 25, 20] already proved that it suffices to show $F(u_S) > 0$ for all $u_S := \tilde{\theta}_S - \theta_S^*$ such that $\|u_S\|_2 = B$ where

$$F(a) := \ell_j(\theta_S^* + a; X^{1:n}) - \ell_j(\theta_S^*; X^{1:n}) + \lambda_j(\|\theta_S^* + a\|_1 - \|\theta_S^*\|_1). \quad (23)$$

More specifically, since u_S is the minimizer of F and $F(0) = 0$ by the construction of Equation (23), $F(u_S) \leq 0$. Note that F is convex, and therefore we have $F(u_S) < 0$.

Next we claim that $\|u_S\|_2 \leq B$. In fact, if u_S lies outside the ball of radius B , then there exists $v \in (0, 1)$ such that the convex combination $v \cdot u_S + (1 - v) \cdot 0$ would lie on the boundary of the ball. However it contradicts the assumed strict positivity of F on the boundary because, by convexity,

$$F(v \cdot u_S + (1 - v) \cdot 0) \leq v \cdot F(u_S) + (1 - v) \cdot 0 \leq 0. \quad (24)$$

Thus it suffices to establish strict positivity of F on the boundary of the ball with radius $B := M_B \lambda_j \sqrt{d}$ where $M_B > 0$ is a parameter to be chosen later in the proof. Let $u_S \in \mathbb{R}^{|S|}$ be an arbitrary vector with $\|u_S\|_2 = B$. By the Taylor series expansion of F in (23),

$$F(u_S) = (W_S)^T u_S + u_S^T [\nabla^2 \ell_j(\theta_S^* + v u_S; X^{1:n})] u_S + \lambda_j(\|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1), \quad (25)$$

for some $v \in [0, 1]$.

The first term in Equation (25) has the following bound: applying $\|W_S\|_\infty \leq \frac{\lambda_j}{4}$ by assumption and $\|u_S\|_1 \leq \sqrt{d} \|u_S\|_2 \leq \sqrt{d} \cdot B$,

$$|(W_S)^T u_S| \leq \|W_S\|_\infty \|u_S\|_1 \leq \|W_S\|_\infty \sqrt{d} \|u_S\|_2 \leq (\lambda_j \sqrt{d})^2 \frac{M_B}{4}.$$

The third term in Equation (25) has the following bound: Applying the triangle inequality,

$$\lambda_j(\|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1) \geq -\lambda_j \|u_S\|_1 \geq -\lambda_j \sqrt{d} \|u_S\|_2 = -M_B (\lambda_j \sqrt{d})^2.$$

Now we show the bound for the second term using the minimum eigenvalue of a matrix $\nabla^2 \ell_j(\theta_S^* + vu_S)$:

$$\begin{aligned}
q^* &:= \lambda_{\min}(\nabla^2 \ell_j(\theta_S^* + vu_S)) \\
&\geq \min_{v \in [0,1]} \lambda_{\min}(\nabla^2 \ell_j(\theta_S^* + vu_S)) \\
&\geq \lambda_{\min}(\nabla^2 \ell_j(\theta_S^*)) - \max_{v \in [0,1]} \left\| \frac{1}{n} \sum_{i=1}^n \exp(\langle \theta_S^* + vu_S, X_S^{(i)} \rangle) u_S^T X_S^{(i)} X_S^{(i)} (X_S^{(i)})^T \right\|_2 \\
&\geq \rho_{\min} - \max_{v \in [0,1]} \max_{y: \|y\|_2=1} \frac{1}{n} \sum_{i=1}^n \exp(\langle \theta_S^* + vu_S, X_S^{(i)} \rangle) \cdot (y^T X_S^{(i)})^2 \cdot |u_S^T X_S^{(i)}|. \tag{26}
\end{aligned}$$

We first show the bound of the first term in Equation (26): Note that $\theta_S^* + vu_S$ is a linear (convex) combination of θ_S^* and $\tilde{\theta}_S$. Hence, by Propositions D.3 and D.4, we obtain

$$P\left(\frac{1}{n} \sum_{i=1}^n \exp(\langle \theta_S^* + vu_S, X_S^{(i)} \rangle) \geq 4 \log \eta\right) \leq M\eta^{-2}.$$

Now, we bound the second term in Equation (26): Recall that $\|X_S^{(i)}\|_\infty \leq 4 \log \eta$ for all i conditioning on ξ_1 . Recall $[u_S]_t = 0$ for $t \notin S$ by the primal-dual construction of (15). Applying $\|u_S\|_1 \leq \sqrt{d}\|u_S\|_2 \leq \sqrt{d} \cdot B$,

$$|u_S^T X_S^{(i)}| \leq 4 \log(\eta) \sqrt{d} \|u_S\|_2 \leq 4 \log \eta \cdot M_B \lambda_j d.$$

Lastly, it is clear that $\max_{y: \|y\|_2=1} (y^T X_S^{(i)})^2 \leq \rho_{\max}$ by the definition of the maximum eigenvalue and Assumption 3.1. Together with the above bounds, we obtain

$$P(q^* \leq \rho_{\min} - 16M_B \rho_{\max} d \lambda_j \log^2 \eta) \leq M\eta^{-2}.$$

For $\lambda_j \leq \frac{\rho_{\min}}{32M_B \rho_{\max} d \log^2 \eta}$, we have $q^* \geq \frac{\rho_{\min}}{2}$ with a high probability. Therefore,

$$F(u) \geq (\lambda_j \sqrt{n})^2 \left\{ -\frac{1}{4} M_B + \frac{\rho_{\min}}{2} M_B^2 - M_B \right\},$$

which is strictly positive for $M_B = \frac{5}{\rho_{\min}}$. Therefore, for $\lambda_j \leq \frac{\rho_{\min}^2}{160 \rho_{\max} d \log^2 \eta}$,

$$P\left(\|\tilde{\theta}_S - \theta_S^*\|_2 \geq \frac{5}{\rho_{\min}} \sqrt{d} \lambda_j\right) \leq M\eta^{-2}.$$

□

F.4 Proof for Lemma B.5

Proof. To improve readability, we use $R_S = [R_j^{S_j}]_S$ where $S = \text{Pa}(j) \cap S_j$. Then, each entry of $R_j^{S_j}$ in Equation (17) has the form $R_{jk} = \frac{1}{n} \sum_{i=1}^n R_{jk}^{(i)}$ for any $k \in S_j$, and it can be expressed as

$$\begin{aligned}
R_{jk} &= \frac{1}{n} \sum_{i=1}^n [\nabla^2 \ell_j(\theta_{S_j}^*; X^{1:n}) - \nabla^2 \ell_j(\tilde{\theta}_{S_j}; X^{1:n})]_k^T (\tilde{\theta}_{S_j} - \theta_{S_j}^*) \\
&= \frac{1}{n} \sum_{i=1}^n \left[\exp(\langle \theta_{S_j}^*, X_S^{(i)} \rangle) - \exp(\langle \tilde{\theta}_{S_j}, X_S^{(i)} \rangle) \right] \left[X_S^{(i)} (X_S^{(i)})^T \right]_k^T (\tilde{\theta}_{S_j} - \theta_{S_j}^*)
\end{aligned}$$

for $\bar{\theta}_S$, which is a point on the line between $\tilde{\theta}_S$ and θ_S^* (i.e., $\bar{\theta}_S^{(t)} = v \cdot \tilde{\theta}_S + (1-v) \cdot \theta_S^*$ for some $v \in [0, 1]$). The second equality holds because $\theta_{S^c}^* = \tilde{\theta}_{S^c} = (0, 0, \dots, 0) \in \mathbb{R}^{|S^c|}$.

Applying the mean value theorem again, we have

$$R_{jk} = \frac{1}{n} \sum_{i=1}^n \left\{ \exp \left(\left\langle \bar{\theta}_S, X_S^{(i)} \right\rangle \right) X_k^{(i)} \right\} \left\{ v(\tilde{\theta}_S - \theta_S^*)^T X_S^{(i)} (X_S^{(i)})^T (\tilde{\theta}_S - \theta_S^*) \right\}$$

for $\bar{\theta}_{S_j}$ which is a point on the line between $\bar{\theta}_{S_j}$ and $\theta_{S_j}^*$ (i.e., $\bar{\theta}_{S_j} = v \cdot \bar{\theta}_{S_j} + (1-v) \cdot \theta_{S_j}^*$ for $v \in [0, 1]$).

Note that $\bar{\theta}_{S_j}$ is a linear (convex) combination of $\theta_{S_j}^*$ and $\tilde{\theta}_S$. Hence from Propositions D.3 and D.4, we obtain

$$P \left(\frac{1}{n} \sum_{i=1}^n \exp \left(\left\langle \bar{\theta}_{S_j}, X_{S_j}^{(i)} \right\rangle \right) \geq 4 \log \eta \right) \leq M \eta^{-2}.$$

Therefore, we have $|R_{jk}| \leq 16 \rho_{\max} \log^2 \eta \|\tilde{\theta}_S - \theta_S^*\|_2^2$ with a high probability.

In Section F.3, we showed that $\|\tilde{\theta}_S - \theta_S^*\|_2 \leq \frac{5}{\rho_{\min}} \sqrt{d} \lambda_j$ for $\lambda_j \leq \frac{\rho_{\min}^2}{160 \rho_{\max} d \log^2 \eta}$. Therefore, if $\lambda_j \leq \frac{\rho_{\min}^2}{25 \cdot 160 \rho_{\max} d \log^2 \eta} \frac{\alpha}{4(2-\alpha)}$, we obtain

$$P \left(\|R_j\|_{\infty} \geq \frac{\alpha}{4(2-\alpha)} \lambda_j \right) \leq P \left(\|R_j\|_{\infty} \geq \frac{4000 \rho_{\max}}{\rho_{\min}^2} d \log^2 \eta \lambda_j^2 \right) \leq M \eta^{-2}.$$

□

F.5 Proof for Lemma C.1

Proof. Conditioning on the sets ζ_2, ζ_3 , and ζ_4 , we provide the following results for different two cases:

(i) For any $j \in \{1, 2, \dots, p-1\}$, and $X_S = X_{1:(j-1)}$, we have $\frac{\mathbb{E}(X_j^2)}{\mathbb{E}(f(\mathbb{E}(X_j | X_S)))} = 1$. Therefore, for $k = \pi_j$, we have the following probability bound:

$$\begin{aligned} & P \left(|\hat{\mathcal{S}}(j, k) - \mathcal{S}(j, k)| < \frac{M_{\min}}{2} \middle| \zeta_2, \zeta_3, \zeta_4 \right) \\ &= P \left(\left| \frac{\hat{\mathbb{E}}(X_k^2)}{\hat{\mathbb{E}}(f(\hat{\mathbb{E}}(X_k | X_S)))} - \frac{\mathbb{E}(X_k^2)}{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))} \right| < \frac{M_{\min}}{2} \middle| \zeta_2, \zeta_3, \zeta_4 \right) \\ &\geq P \left(\frac{\mathbb{E}(X_k^2) + \epsilon_1}{\mathbb{E}(f(\mathbb{E}(X_k | X_S))) - 2\epsilon_1} - \frac{\mathbb{E}(X_k^2)}{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))} < \frac{M_{\min}}{2} \text{ and } \right. \\ &\quad \left. \frac{\mathbb{E}(X_k^2)}{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))} - \frac{\mathbb{E}(X_k^2) - \epsilon_1}{\mathbb{E}(f(\mathbb{E}(X_k | X_S))) + 2\epsilon_1} < \frac{M_{\min}}{2} \right) \\ &\geq P \left(\epsilon_1 < \frac{\mathbb{E}(f(\mathbb{E}(X_k | X_S))) M_{\min}}{2(M_{\min} + 3)} \right) \\ &\geq P \left(\epsilon_1 < \frac{\mathbb{E}(X_k^2) M_{\min}}{2(M_{\min} + 3)(M_{\min} + 1)} \right). \end{aligned}$$

(ii) For $j \in \{1, 2, \dots, p-1\}$, $k \in \{\pi_{j+1}, \dots, \pi_p\}$ having parent π_j , and $X_S = X_{1:(j-1)}$, we have $\mathbb{E}(X_k^2) > (1 + M_{\min})\mathbb{E}(f(\mathbb{E}(X_k | X_S)))$ by Assumption 3.4. In addition, some elementary but complicated computations yield

$$\begin{aligned}
& P\left(\left|\widehat{\mathcal{S}}(j, k) - \mathcal{S}(j, k)\right| < \frac{M_{\min}}{2} \mid \zeta_2, \zeta_3, \zeta_4\right) \\
& \geq P\left(\epsilon_1 < \frac{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))^2 M_{\min}}{4\mathbb{E}(X_k^2) + 2\mathbb{E}(f(\mathbb{E}(X_k | X_S))) + 2\mathbb{E}(f(\mathbb{E}(X_k | X_S)))M_{\min}}\right) \\
& \geq P\left(\epsilon_1 < \frac{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))^2 M_{\min}(1 + M_{\min})}{4\mathbb{E}(X_k^2)(1 + M_{\min}) + 2\mathbb{E}(X_k^2) + 2M_{\min}\mathbb{E}(X_k^2)}\right) \\
& \geq P\left(\epsilon_1 < \frac{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))^2 M_{\min}(1 + M_{\min})}{6(1 + M_{\min})\mathbb{E}(X_k^2)}\right) \\
& \geq P\left(\epsilon_1 < \frac{M_{\min}}{6} \frac{\mathbb{E}(f(\mathbb{E}(X_k | X_S)))^2}{\mathbb{E}(X_k^2)}\right).
\end{aligned}$$

Therefore $P(\zeta_1 \mid \zeta_2, \zeta_3, \zeta_4) = 0$ if ϵ_1 is sufficiently small enough. For any node j , any set $S_j \in \{\pi_1, \pi_{1:2}, \dots, \pi_{1:(j-1)}\}$, and $k \in \{\pi_j, \pi_{j+1}, \dots, \pi_p\}$,

$$\epsilon_1 < \min \left\{ \frac{\mathbb{E}(X_k^2)M_{\min}}{2(M_{\min} + 3)(M_{\min} + 1)}, \frac{M_{\min}}{6} \frac{\mathbb{E}(f(\mathbb{E}(X_k | X_{S_j})))^2}{\mathbb{E}(X_k^2)} \right\}.$$

□

F.6 Proof for Lemma C.2

The proof for Lemma C.2 is closely related to the proof in Appendix B. Hence, for brevity, we do not present the details of the proof already shown in Appendix B.

$$(i) \ P(\zeta_2^c \mid \xi_1) \leq 2p \cdot \exp \left\{ -\frac{n\epsilon_1^2}{8\log^4 \eta} \right\}.$$

Proof. Using Hoeffding's inequality given ξ_1 , for any $\epsilon > 0$ and $j \in V$,

$$P\left(\left|\widehat{\mathbb{E}}(X_j^2) - \mathbb{E}(X_j^2)\right| > \epsilon_1 \mid \xi_1\right) \leq 2 \cdot \exp \left\{ -\frac{n\epsilon_1^2}{8\log^4 \eta} \right\}. \tag{27}$$

Hence conditioning on ξ_1 , we have

$$P\left(\max_{j \in V} \left|\widehat{\mathbb{E}}(X_j^2) - \mathbb{E}(X_j^2)\right| > \epsilon_1 \mid \xi_1\right) \leq 2p \cdot \exp \left\{ -\frac{n\epsilon_1^2}{8\log^4 \eta} \right\}.$$

□

$$(ii) \ P(\zeta_3^c \mid \xi_1) \leq 2pd \cdot \exp(-2n^{1-2a}) + 2p \cdot \exp \left\{ -\frac{2n\epsilon_1^2}{f(4\kappa_2(\eta)\log \eta)} \right\}.$$

Proof. We restate the condition in the set ζ_3 as

$$\left| \frac{1}{n} \sum_{i=1}^n f(\widehat{\mathbb{E}}(X_j | X_{S_j})) - \mathbb{E}(f(\mathbb{E}(X_j | X_{S_j}))) \right| < \epsilon_1.$$

In order to apply Hoeffding's inequality, we first show the bound for $\widetilde{E}(X_j | X_{S_j})$. Recall that $[\theta^*]_{S^c}$ and $[\widetilde{\theta}]_{S^c} = (0, 0, \dots, 0) \in \mathbb{R}^{|S^c|}$ in our primal-dual construction, and $|S| \leq d$. In Appendix F.3, we showed that $\|\widetilde{\theta}_S - \theta_S^*\|_2 \leq \frac{5}{\rho_{\min}} \sqrt{d} \lambda_j$ for $\lambda_j \leq \frac{\rho_{\min}^2}{160 \rho_{\max} d \log^2 \eta}$ with a high probability. Therefore given ξ_1 ,

$$\begin{aligned} \exp(\langle \widehat{\theta}_{S_j}, X_{S_j} \rangle) &= \exp(\langle \widehat{\theta}_{S_j} - \theta_{S_j}^*, X_{S_j} \rangle) \cdot \exp(\langle \theta_{S_j}^*, X_{S_j} \rangle) \\ &\leq \exp(\|\widehat{\theta}_S - \theta_S^*\|_2 \|X_S\|_2) \cdot \exp(\langle \theta_S^*, X_S \rangle) \\ &\leq \exp\left\{ \frac{5d \lambda_j}{\rho_{\min}} \|X_S\|_\infty \right\} \cdot \exp(\langle \theta_S^*, X_S \rangle) \\ &\leq \exp\left\{ \frac{20d \lambda_j}{\rho_{\min}} \log(\eta) \right\} \cdot \exp(\langle \theta_S^*, X_S \rangle) \\ &\leq \exp\left\{ \frac{\rho_{\min}}{8 \rho_{\max} \log \eta} \right\} \cdot \exp(\langle \theta_S^*, X_S \rangle). \end{aligned}$$

From Proposition D.4, for all $j \in V$,

$$\exp(\langle \widetilde{\theta}_{S_j}, X_{S_j} \rangle) \leq \exp\left\{ \frac{\rho_{\min}}{8 \rho_{\max} \log \eta} \right\} \cdot 4 \log \eta \leq 4 \exp\left\{ \frac{\rho_{\min}}{8 \rho_{\max}} \right\} \log \eta$$

with a high probability of at least $1 - M\eta^{-2}$.

Therefore,

$$f(\widehat{\mathbb{E}}(X_j | X_{S_j})) \leq 16 \cdot \exp\left\{ \frac{\rho_{\min}}{4 \rho_{\max}} \right\} \log^2 \eta + 4 \cdot \exp\left\{ \frac{\rho_{\min}}{8 \rho_{\max}} \right\} \log \eta.$$

Hence there exists a positive constant $D_1 > 0$ such that

$$f(\widehat{\mathbb{E}}(X_j | X_{S_j})) \leq D_1 \log^2 \eta.$$

Applying Hoeffding's inequality conditioned on ξ_1 , for any $\epsilon_1 > 0$ and any $j \in V$,

$$P\left(\left|\widehat{\mathbb{E}}(f(\widetilde{E}(X_j | X_{S_j}))) - \mathbb{E}(f(\widetilde{E}(X_j | X_{S_j})))\right| > \epsilon_1\right) \leq 2 \cdot \exp\left\{-\frac{2n\epsilon_1^2}{D_1^2 \log^4 \eta}\right\}. \quad (28)$$

Hence, there exist some constants $D_{\max} > 0$ such that

$$P\left(\max_{j \in V} \zeta_3^c\right) \leq 1 - 2p \cdot d \cdot \exp\left(-\frac{n}{\kappa_1(n, p)^2}\right) - M \cdot \eta^{-2} - 2p \cdot \exp\left\{-\frac{n\epsilon_1^2}{D_{\max} \log^4 \eta}\right\}.$$

As we showed in Appendix B, if we set $\kappa_1(n, p) = C_{\max} d \log^4 \eta$ where $C_{\max} = \frac{2048 \cdot 10^3 (2-\alpha)^2}{\alpha^2} \frac{\rho_{\max}}{\rho_{\min}^2}$,

$$P \left(\max_{j \in V} \zeta_3^c \right) \leq 1 - 2p \cdot d \cdot \exp \left(- \frac{n}{C_{\max}^2 d^2 \log^8 \eta} \right) - M \cdot \eta^{-2} - 2p \cdot \exp \left\{ - \frac{n \epsilon_1^2}{D_{\max} \log^4 \eta} \right\}.$$

□

(iii) $P(\zeta_4^c \mid \xi_1) = 0$.

Proof. We restate the condition in the set ζ_4 as

$$\left| \mathbb{E} \left(f(\mathbb{E}(X_j \mid X_{S_j})) - f(\widehat{\mathbb{E}}(X_j \mid X_{S_j})) \right) \right| < \epsilon_1.$$

By the mean-value theorem, for some $v \in [0, 1]$,

$$\begin{aligned} & f(\widehat{\mathbb{E}}(X_j \mid X_{S_j})) - f(\mathbb{E}(X_j \mid X_{S_j})) \\ &= f'(v\widehat{\mathbb{E}}(X_j \mid X_{S_j}) + (1-v)\mathbb{E}(X_j \mid X_{S_j}))(\widehat{\mathbb{E}}(X_j \mid X_{S_j}) - \mathbb{E}(X_j \mid X_{S_j})) \\ &= 2(v\widehat{\mathbb{E}}(X_j \mid X_{S_j}) + (1-v)\mathbb{E}(X_j \mid X_{S_j}) + 1/2)(\widehat{\mathbb{E}}(X_j \mid X_{S_j}) - \mathbb{E}(X_j \mid X_{S_j})). \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{E}(f(\widehat{\mathbb{E}}(X_j \mid X_{S_j})) - f(\mathbb{E}(X_j \mid X_{S_j}))) \\ &= f'(v\widehat{\mathbb{E}}(X_j \mid X_{S_j}) + (1-v)\mathbb{E}(X_j \mid X_{S_j}))(\widehat{\mathbb{E}}(X_j \mid X_{S_j}) - \mathbb{E}(X_j \mid X_{S_j})) \\ &= 2(v\widehat{\mathbb{E}}(X_j \mid X_{S_j}) + (1-v)\mathbb{E}(X_j \mid X_{S_j}) + 1/2)(\widehat{\mathbb{E}}(X_j \mid X_{S_j}) - \mathbb{E}(X_j \mid X_{S_j})) \\ &\leq \max |2(v\widehat{\mathbb{E}}(X_j \mid X_{S_j}) + (1-v)\mathbb{E}(X_j \mid X_{S_j}) + 1/2)| \cdot \mathbb{E}(\widehat{\mathbb{E}}(X_j \mid X_{S_j}) - \mathbb{E}(X_j \mid X_{S_j})) \\ &= 0 \end{aligned}$$

In the same manner, $\mathbb{E}(f(\mathbb{E}(X_j \mid X_{S_j})) - f(\widehat{\mathbb{E}}(X_j \mid X_{S_j}))) \leq 0$. This completes the proof.

□