Journal of Machine Learning Research 18 (2018) 1-44 Submitted 4/17; Revised 12/17; Published 4/18

Learning Quadratic Variance Function (QVF) DAG Models via OverDispersion Scoring (ODS)

Gunwoong Park GWPARK23@UOS.AC.KR
Department of Statistics
University of Seoul
Seoul, 02592, South Korea

Garvesh Raskutti RASKUTTI@STAT.WISC.EDU
Department of Statistics
Department of Computer Science
Wisconsin Institute for Discovery, Optimization Group
University of Wisconsin
Madison, WI 53706, USA

Editor: Hui Zou

Abstract

Learning DAG or Bayesian network models is an important problem in multi-variate causal inference. However, a number of challenges arises in learning large-scale DAG models including model identifiability and computational complexity since the space of directed graphs is huge. In this paper, we address these issues in a number of steps for a broad class of DAG models where the noise or variance is signal-dependent. Firstly we introduce a new class of identifiable DAG models, where each node has a distribution where the variance is a quadratic function of the mean (QVF DAG models). Our QVF DAG models include many interesting classes of distributions such as Poisson, Binomial, Geometric, Exponential, Gamma and many other distributions in which the noise variance depends on the mean. We prove that this class of QVF DAG models is identifiable, and introduce a new algorithm, the OverDispersion Scoring (ODS) algorithm, for learning large-scale QVF DAG models. Our algorithm is based on firstly learning the moralized or undirected graphical model representation of the DAG to reduce the DAG search-space, and then exploiting the quadratic variance property to learn the ordering. We show through theoretical results and simulations that our algorithm is statistically consistent in the high-dimensional $p > n$ setting provided that the degree of the moralized graph is bounded and performs well compared to state-of-the-art DAG-learning algorithms. We also demonstrate through a real data example involving multi-variate count data, that our ODS algorithm is well-suited to estimating DAG models for count data in comparison to other methods used for discrete data.

Keywords: Bayesian Networks, Directed Acyclic Graph, Identifiability, Multi-variate Count Distribution, Overdispersion

1. Introduction

Probabilistic directed acyclic graphical (DAG) models or Bayesian networks provide a widely used framework for representing causal or directional dependence relationships amongst multiple variables. DAG models have applications in various areas including genomics,

neuroimaging, statistical physics, spatial statistics and many others (see e.g., Doya 2007; Friedman et al. 2000; Kephart and White 1991). One of the fundamental problems associated with DAG models or Bayesian networks is structure learning from observational data.

If the number of variables is large, a number of challenges arise that make learning large-scale DAG models extremely difficult even when variables have a natural causal or directional structure. These challenges include: (1) identifiability since inferring causal directions from only observational data is in general not possible in the absence of additional assumptions; (2) computational complexity since it is NP-hard to search over the space of DAGs (Chickering, 1996); (3) providing sample size guarantee in the setting where the number of nodes p is large. In this paper we develop a general framework and algorithm for learning large-scale DAG models that addresses these challenges in a number of steps: Firstly, we introduce a new class of provably identifiable DAG models where each node has a conditional distribution where the variance is a quadratic function of the mean, which we refer to as QVF (quadratic variance function) distributions; secondly, we introduce a general OverDispersion Scoring (ODS) algorithm for learning large-scale QVF DAG models; thirdly, we provide theoretical guarantees for our ODS algorithm which proves that our algorithm is consistent in the high-dimensional setting $p > n$ provided that the moralized graph of the DAG is sparse; and finally, we show through a simulation study that our ODS algorithm supports our theoretical result has favorable performance to a number of state-of-the-art algorithms for learning both low-dimensional and high-dimensional DAG models.

Our algorithm is based on combining two ideas: *overdispersion* and *moralization*. Overdispersion is a property of Poisson and other random variables where the variance depends on the mean and we use overdispersion to address the identifiability issue. While overdispersion is a known phenomena used and exploited in many applications (see e.g., Dean 1992; Zheng et al. 2006), overdispersion has never been exploited for learning DAG models aside from our prior work (Park and Raskutti, 2015) which focuses on Poisson DAG models. In this paper, we show that overdispersion applies much more broadly and is used to prove identifiability for a broad class of DAG models. To provide a scalable algorithm with statistical guarantees, even in the high-dimensional setting, we exploit the moralized graph, that is the undirected representation of the DAG. Learning the moralized graph allows us to exploit sparsity and considerably reduces the DAG search-space which has both computational and statistical benefits. Furthermore, moralization allows us to use existing scalable algorithms and theoretical guarantees for learning large-scale undirected graphical models (e.g., Friedman et al. 2009; Yang et al. 2012).

A number of approaches have been used to address the identifiability challenge by imposing additional assumptions. For example ICA-based methods for learning ordering requires independent noise and non-Gaussianity (see e.g., Shimizu et al. 2006), structural equation models with Gaussian noise with equal or known variances (Peters et al., 2012), and non-parametric structural equation models with independent noise (see e.g., Peters and Bühlmann 2013). These approaches are summarized elegantly in an information-theoretic framework in Janzing and Scholkopf (2010). Our approach is along similar lines in that we impose overdispersion as an additional assumption which induces asymmetry and guarantees identifiability. However by exploiting overdispersion, our approach applies when the noise distribution of each node depends on its mean whereas prior approaches apply when

the additive noise variance is independent of the mean. Additionally, we exploit graph sparsity which has also been exploited in prior work by Loh and Bühlmann (2014); Raskutti and Uhler (2013); van de Geer and Bühlmann (2013) for various DAG models with independent additive noise components. Furthermore, sparsity allows us to develop a tractable algorithm where we reduce the DAG space by learning the moralized graph, an idea which has been used in prior work in Tsamardinos and Aliferis (2003).

1.1 Our Contributions

We summarize the major contributions of the paper as follows:

- We introduce the class of QVF DAG models, that include many interesting classes of multi-variate distributions and provide conditions under which QVF DAG models are identifiable.
- Using QVF DAG models, we develop the reliable and scalable generalized ODS algorithm which learns any large-scale QVF DAG model. Our algorithm combines two key ideas, moralization and overdispersion. Moralization significantly reduces computational complexity by exploiting sparsity of the moralized graph, while overdispersion exploits properties of QVF DAG models to estimate the causal ordering. The generalized ODS algorithm adapts the algorithm developed in Park and Raskutti (2015) to general QVF DAG models whilst the algorithm in Park and Raskutti (2015) focuses exclusively on Poisson DAG models.
- We provide statistical guarantees to show that our ODS algorithm is consistent for learning QVF DAG models, even in the high-dimensional $p > n$ setting, provided that the degree of the moralized graph is bounded. To the best of our knowledge, this is the only theoretical result that applies to the high-dimensional setting when the variables at each node model counts.
- We demonstrate through simulation studies and a real data application involving multi-variate count data that our ODS algorithm performs favorably compared to the state-of-the-art GES and MMHC algorithms. In our simulation study, we consider both the low-dimensional and high-dimensional setting. Our real data example involving NBA player statistics for 2009/10 season shows that our ODS algorithm is applicable to multi-variate count data while the GES and MMHC algorithms tend to select very few edges when variables represent counts.

The remainder of the paper is organized as follows: In Section 2, we define QVF DAG models and prove identifiability for this class of models. In Section 3, we introduce our polynomial-time DAG learning algorithm which we refer to as the generalized OverDispersion Scoring (ODS). Statistical guarantees for learning QVF DAG models using our ODS algorithm are provided in Section 3.2, and we provide numerical experiments on both small DAGs and large-scale DAGs with node-size up to 5000 nodes in Section 4. Our theoretical guarantees in Section 3.2 prove that even in the setting where the number of nodes p is larger than the sample size n, it is possible to learn the DAG structure under the assumption that the degree d of the so-called moralized graph of the DAG is small. Our numerical experiments in Section 4 support the theoretical results and show that our algorithm performs

well compared to other state-of-the-art DAG learning methods. Our numerical experiments confirm that our algorithm is one of the few DAG-learning algorithms that performs well in terms of statistical and computational complexity in high-dimensional $p > n$ settings, provided that the degree of the moralized graph d is bounded. Finally in Section 5 we show that our ODS algorithm performs well in terms of the eye test compared to state-of-the-art algorithms for a multi-variate count data set that involves basketball statistics.

2. Quadratic Variance Function (QVF) DAG Models and Identifiability

A DAG $G = (V, E)$ consists of a set of nodes V and a set of directed edges $E \in V \times V$ where each $e \in E$ is an ordered pair of distinct nodes. Hence our graphs are *simple*, i.e., there are no directed cycles or multiple edges between any pair of nodes. A directed edge from node j to k is denoted by (j, k) or $j \to k$. The set of *parents* of node k denoted by $\text{pa}(k)$ consists of all nodes j such that $(j, k) \in E$. If there is a directed path $j \to \cdots \to k$, then k is called a *descendant* of j and j is an *ancestor* of k. The set $\text{de}(k)$ denotes the set of all descendants of node k. The *non-descendants* of node k are $\text{nd}(k) := V \setminus (\{k\} \cup \text{de}(k))$. An important property of DAGs is that there exists an (possibly non-unique) *ordering* π^* 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. Without loss of generality, we set $V = \{1, 2, \cdots, p\}$ and assume the true ordering is $\pi^* = (1, 2, \cdots, p)$.

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

$$f_G(X) = \prod_{j \in V} f_j(X_j \mid X_{\text{pa}(j)}), \tag{1}$$

where $f_j(X_j \mid X_{\text{pa}(j)})$ refers to the conditional distribution of a random variable X_j in terms of its parents $X_{\text{pa}(j)} := \{X_s : s \in \text{pa}(j)\}$.

A core concept in this paper is *identifiability* for a family of probability distributions defined by the DAG factorization provided above. Intuitively identifiability addresses the question of what assumption we make on the conditional distributions $f_j(X_j \mid X_{\text{pa}(j)})$ allows us to uniquely determine the structure of that DAG G given the joint PDF $f_G(X)$.

To define identifiability precisely, let $\mathcal{P}$ denote the set of *conditional distributions* $f_j(X_j \mid X_{\text{pa}(j)})$ for all $j \in V$. Further for a graph $G = (V, E)$, define the class of *joint distributions* with respect to graph G and class of distributions $\mathcal{P}$ by

$$\mathcal{F}(G; \mathcal{P}) := \{f_G(X) = \prod_{j \in V} f_j(X_j \mid X_{\text{pa}(j)}) \ ; \ \text{where } f_j(X_j \mid X_{\text{pa}(j)}) \in \mathcal{P} \ \forall \ j \in V\}.$$

Next let $\mathcal{G}_p$ denote the set of p-node directed acyclic graphs. Now we define identifiability for the class $\mathcal{P}$ over the space of DAGs $\mathcal{G}_p$.

Definition 1 (Identifiability) *A class of conditional distributions $\mathcal{P}$ is* identifiable *over $\mathcal{G}_p$ if $G \neq G'$ where $G, G' \in \mathcal{G}_p$, there exists no $f_G \in \mathcal{F}(G; \mathcal{P})$ and $f_{G'} \in \mathcal{F}(G'; \mathcal{P})$ such that $f_G = f_{G'}$.*

Prior work has addressed the question of identifiability for different classes of $\mathcal{P}$. For example ICA-based methods make the assumption that $\mathcal{P}$ are independent error with non-Gaussian components (Shimizu et al., 2006) and prove that this class is identifiable as well as $\mathcal{P}$ corresponding to a non-parametric model with additive independent noise (Peters and Bühlmann, 2013), $\mathcal{P}$ represents structural linear equation models with Gaussian errors with equal or known variances (Peters et al., 2012). On the other hand, if $\mathcal{P}$ represents structural linear equation models with Gaussian errors with general variance is not identifiable and only the Markov equivalence class of DAG models is identifiable (Heckerman et al., 1995).

The main results of our paper give another class of identifiable graphical models. In our setting, $\mathcal{P}$ is a setting where the variance is a linear function of the mean so we deal with signal-dependent noise or variance, and more importantly is applicable for discrete distributions. We define $\mathcal{P}$ more precisely in the next section.

2.1 Quadratic Variance Function (QVF) DAG Models

Now we define quadratic variance function (QVF) DAG models. For QVF DAG models each node has a conditional distribution $\mathcal{P}$ given its parents with the property that the variance is a quadratic function of the mean. More precisely,

Definition 2 (QVF DAG models) *Quadratic variance function (QVF) DAG models are DAG models where conditional distribution of each node given its parents satisfies the* quadratic variance function property*: for all $j \in V$, there exist $\beta_{j0}, \beta_{j1} \in \mathbb{R}$ such that*

$$Var(X_j \mid X_{pa(j)}) = \beta_{j0}\mathbb{E}(X_j \mid X_{pa(j)}) + \beta_{j1}\mathbb{E}(X_j \mid X_{pa(j)})^2. \tag{2}$$

To the best of our knowledge, quadratic variance function (QVF) probability distributions were first introduced in the context of natural parameter exponential families (NEF) (Morris, 1982) which include Poisson, Binomial, Negative Binomial and Gamma distributions. In the directed graphical model framework, each node distribution is influenced by its parents. Hence for natural exponential families with quadratic variance functions (NEF-QVF), we provide an explicit form of joint distributions for DAG models.

For NEF-QVF, the conditional distribution of each node given its parents takes the simple form:

$$P(X_j \mid X_{\mathrm{pa}(j)}) = \exp\left(\theta_{jj}X_j + \sum_{(k,j)\in E}\theta_{jk}X_kX_j - B_j(X_j) - A_j\left(\theta_{jj} + \sum_{(k,j)\in E}\theta_{jk}X_k\right)\right)$$

where $A_j(\cdot)$ is the log-partition function, $B_j(\cdot)$ is determined by a chosen exponential family, and $\theta_{jk} \in \mathbb{R}$ is a parameter corresponding to a node j. By the factorization property (1), the joint distribution of a NEF-QVF DAG model takes the following form:

$$P(X) = \exp\left(\sum_{j\in V}\theta_{jj}X_j + \sum_{(k,j)\in E}\theta_{jk}X_kX_j - \sum_{j\in V}B_j(X_j) - \sum_{j\in V}A_j\left(\theta_{jj} + \sum_{(k,j)\in E}\theta_{jk}X_k\right)\right). \tag{3}$$

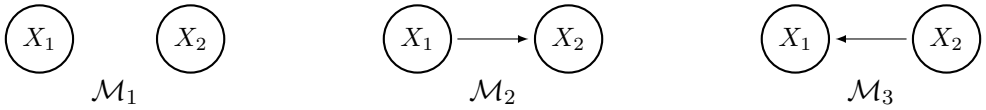


Figure 1: Directed graphical models of $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$

From Equation (3), we provide examples of classes of NEF-QVF DAG models. For Poisson DAG models studied in Park and Raskutti (2015) the log-partition function $A_j(\cdot) = \exp(\cdot)$, and $B_j(\cdot) = \log(\cdot!)$. Similarly, Binomial DAG models can be derived as an example of QVF DAG models where the conditional distribution for each node is binomial with known parameter N_j and the log-partition function $A_j(\cdot) = N_j \log(1 + \exp(\cdot))$, and $B_j(\cdot) = -\log \binom{N_j}{\cdot}$. Another interesting instance is Exponential DAG models where each node conditional distribution given its parents is Exponential. Then, $A_j(\cdot) = -\log(-\cdot)$ and $B_j(\cdot) = 0$.

Our framework also naturally extends to mixed DAG models, where the conditional distributions have different distributions which incorporates different data types. In addition, our models extend to nonlinear and nonparametric DAG models depending on the data as long as the node distribution $\mathcal{P}$ satisfies the QVF property in Equation (2). This means our model is identifiable without information on how a node and its parents are related. In Section 4, we will provide numerical experiments on Poisson and Binomial DAG models.

2.2 Identifiability of QVF DAG Models

In this section we prove that QVF DAG models are identifiable. To provide intuition, we prove identifiability for the two-node Poisson DAG model in Park and Raskutti (2015). Consider all three models illustrated in Figure 1: $\mathcal{M}_1 : X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, where X_1 and X_2 are independent; $\mathcal{M}_2 : X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \mid X_1 \sim \text{Poisson}(g_2(X_1))$; and $\mathcal{M}_3 : X_2 \sim \text{Poisson}(\lambda_2)$ and $X_1 \mid X_2 \sim \text{Poisson}(g_1(X_2))$ for arbitrary positive functions $g_1, g_2 : \mathbb{N} \cup \{0\} \to \mathbb{R}^+$. Our goal is to determine whether the underlying DAG model is $\mathcal{M}_1, \mathcal{M}_2$ or $\mathcal{M}_3$.

We exploit the equidispersion property that for a Poisson random variable X, $\text{Var}(X) = \mathbb{E}(X)$, while for a distribution which is conditionally Poisson, the marginal variance is overdispersed relative to the marginal expectation, $\text{Var}(X) > \mathbb{E}(X)$. Hence for $\mathcal{M}_1$, $\text{Var}(X_1) = \mathbb{E}(X_1)$ and $\text{Var}(X_2) = \mathbb{E}(X_2)$. For $\mathcal{M}_2$, $\text{Var}(X_1) = \mathbb{E}(X_1)$, while

$$\text{Var}(X_2) = \mathbb{E}(\text{Var}(X_2 \mid X_1)) + \text{Var}(\mathbb{E}(X_2 \mid X_1)) = \mathbb{E}(\mathbb{E}(X_2 \mid X_1)) + \text{Var}(g_2(X_1)) > \mathbb{E}(X_2),$$

as long as $\text{Var}(g_2(X_1)) > 0$. The first equality follows from the total variance decomposition and the second equality follows from the equidispersion property of Poisson distribution.

Similarly under $\mathcal{M}_3$, $\text{Var}(X_2) = \mathbb{E}(X_2)$ and $\text{Var}(X_1) > \mathbb{E}(X_1)$ as long as $\text{Var}(g_1(X_2)) > 0$. Hence we can distinguish models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ by testing whether the variance is greater than or equal to the expectation. With finite samples, the quantities $\mathbb{E}(\cdot)$ and $\text{Var}(\cdot)$ can be estimated from data and we describe this more precisely in Sections 3 and 3.2.

For general QVF DAG models, the variance for each node distribution is not necessarily equal to the mean. Hence we introduce a linear transformation $T_j(X_j)$ such that

Distribution	$\mathcal{P}$	β_0	β_1	ω
Binomial	$\text{Bin}(N, p)$	1	$-\frac{1}{N}$	$\frac{N}{N-\mu}$
Poisson	$\text{Poi}(\lambda)$	1	0	1
Geometric	$\text{Geo}(p)$	1	1	$\frac{1}{1+\mu}$
Negative Binomial	$\text{NB}(R, p)$	1	$\frac{1}{R}$	$\frac{R}{R+\mu}$
Exponential	$\text{Exp}(\lambda)$	0	1	$\frac{1}{\mu}$
Gamma	$\text{Gamma}(\alpha, \beta)$	0	$\frac{1}{\alpha}$	$\frac{\alpha}{\mu}$

Table 1: Examples of distributions for QVF DAG models with β_0, β_1 and ω where μ is its expectation

$\text{Var}(T_j(X_j) \mid X_{\text{pa}(j)}) = \mathbb{E}(T_j(X_j) \mid X_{\text{pa}(j)})$ in Proposition 3. This transformation enables us to use the notion of *overdispersion* for recovering QVF DAG models. We present examples of distributions $\mathcal{P}$ for QVF DAG models with the triple $(\beta_0, \beta_1, \omega)$ in Table 1.

Proposition 3 *Let $X = (X_1, X_2, \cdots, X_p)$ be a random vector associated with a QVF DAG model with quadratic variance coefficients $(\beta_{j0}, \beta_{j1})_{j=1}^p$ specified in Equation (2). Then, there exists a transformation $T_j(X_j) = \omega_j X_j$ for any node $j \in V$ where $\omega_j = (\beta_{j0} + \beta_{j1}\mathbb{E}(X_j \mid X_{pa(j)}))^{-1}$ such that*

$$Var(T_j(X_j) \mid X_{pa(j)}) = \mathbb{E}(T_j(X_j) \mid X_{pa(j)}).$$

Proof For any node $j \in V$,

$$\begin{aligned} \text{Var}(\omega_j X_j \mid X_{\text{pa}(j)}) &= \omega_j^2 \text{Var}(X_j \mid X_{\text{pa}(j)}) \\ &\overset{(a)}{=} \omega_j^2(\beta_{j0}\mathbb{E}(X_j \mid X_{\text{pa}(j)}) + \beta_{j1}\mathbb{E}(X_j \mid X_{\text{pa}(j)})^2) \\ &\overset{(b)}{=} \omega_j \mathbb{E}(X_j \mid X_{\text{pa}(j)}) \\ &= \mathbb{E}(\omega_j X_j \mid X_{\text{pa}(j)}). \end{aligned}$$

Equality (a) follows from the quadratic variance property (2), and (b) follows from the definition of ω_j. ■

Now we extend to general p-variate QVF DAG models. The key idea to extending identifiability from the bivariate to multivariate scenario involves conditioning on parents of each node, and then testing overdispersion.

Assumption 4 *For all $j \in V$ and any $pa_0(j) \subset pa(j)$ where $pa_0(j) \neq \emptyset$ and $S \subset nd(j) \setminus pa_0(j)$:*

(a) $Var(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_S) > 0$, *and*

(b) $\beta_{j0} + \beta_{j1}\mathbb{E}(X_j \mid X_S) \neq 0$.

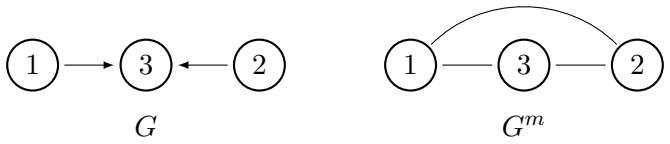


Figure 2: Moralized graph G^m for DAG G

Assumption 4(a) ensures that all parents of node j contribute to its variability, hence a conditional variance is bigger than the conditional expectation. Assumption 4(b) rules out the extremely skewed distributions. For example, Binomial distribution with a parameter N has $\beta_0 = 1$ and $\beta_1 = -\frac{1}{N}$. Then, the assumption is satisfied as long as the conditional expectation is less than N. Similarly Exponential distribution has $\beta_0 = 0$ and $\beta_1 = 1$, hence the assumption is satisfied as long as the conditional expectation is positive. Poisson distribution has $\beta_0 = 1$ and $\beta_1 = 0$, so the assumption is always satisfied.

Theorem 5 (Identifiability for p-variate QVF DAG models) *Consider the class of QVF DAG models* (1) *with quadratic variance coefficients* $(\beta_{j0}, \beta_{j1})_{j=1}^p$ (2)*. If for all* $j \in V$*,* $\beta_{j1} > -1$ *and Assumption 4 is satisfied, then the class of QVF DAG models is identifiable according to Def. 1.*

The proof is provided in Appendix A. Theorem 5 shows that any QVF DAG model is identifiable under the assumption that all parents of node j contribute to its variability. The condition $\beta_{j1} > -1$ rules out DAG models with Bernoulli and multinomial distributions which are known to be non-identifiable (Heckerman et al., 1995) with $\beta_{j1} = -1$.

3. OverDispersion Scoring (ODS) Algorithm

In this section, we present our generalized OverDispersion Scoring (ODS) algorithm. The ODS algorithm is introduced by Park and Raskutti (2015) for Poisson DAG models, however it does not cover other QVF distributions. In this paper, we generalize the ODS algorithm for recovering QVF DAG models. An important concept we need to introduce for the generalized ODS algorithm is the *moral* graph or undirected graphical model representation of a DAG (see e.g., Cowell et al. 1999). The moralized graph G^m for a DAG $G = (V, E)$ is an undirected graph where $G^m = (V, E^m)$ where E^m includes the edge set E for the DAG G with directions removed plus edges between any nodes that are parents of a common child. Figure 2 represents the moralized graph for a simple 3-node example where $E = \{(1,3),(2,3)\}$ for the DAG G. Since nodes 1 and 2 are parents with a common child 3, the additional edge $(1,2)$ arises, and therefore $E^m = \{(1,2),(1,3),(2,3)\}$. Finally, the *neighborhood* for a node j refers to the adjacent nodes to j in the moralized graph, and is denoted by $\mathcal{N}(j) := \{k \in V \mid (j,k) \text{ or } (k,j) \in E^m\}$.

Our generalized ODS algorithm (Algorithm 1) has three main steps: 1) estimate the moralized graph G^m for the DAG G; 2) estimate the ordering of the DAG G using overdispersion scoring based on the moralized graph from step 1); and 3) estimate the DAG structure, given the ordering from step 2). Although Steps 2) and 3) are sufficient to recover

Algorithm 1: Generalized OverDispersion Scoring (ODS)

Input : n i.i.d. samples from a QVF-DAG model
Output: Estimated ordering $\widehat{\pi} \in \mathbb{N}^p$ and an edge structure, $\widehat{E} \in V \times V$

Step 1: Estimate the undirected edges $\widehat{E}^m = \cup_{j \in V} \cup_{k \in \widehat{\mathcal{N}}(j)} (j,k)$ where $\widehat{\mathcal{N}}(j)$ is estimated neighborhood set of a node j in the moralized graph;
Step 2: Estimate the ordering using overdispersion scores;
for $j \in \{1, 2, \cdots, p\}$ **do**
  Calculate overdispersion scores $\widehat{\mathcal{S}}(1, j)$ using Equation (4);
end
The first element of the ordering $\widehat{\pi}_1 = \arg\min_j \widehat{\mathcal{S}}(1, j)$;
for $m = \{2, 3, \cdots, p-1\}$ **do**
  for $j \in \{1, 2, \cdots, p\} \setminus \{\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \widehat{\pi}_{m-1}\}$ **do**
    Find candidate parents set $\widehat{C}_{mj} = \widehat{\mathcal{N}}(j) \cap \{\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \widehat{\pi}_{m-1}\}$;
    Calculate overdispersion scores $\widehat{\mathcal{S}}(m, j)$ using Equation (5);
  end
  The m^{th} element of an ordering $\widehat{\pi}_m = \arg\min_j \widehat{\mathcal{S}}(m, j)$;
  Step 3: Estimate the directed edges toward $\widehat{\pi}_m$, denoted by $\widehat{D}_m$;
end
The last element of the ordering $\widehat{\pi}_p = \{1, 2, \cdots, p\} \setminus \{\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \widehat{\pi}_{p-1}\}$;
The directed edges toward $\widehat{\pi}_p$, denoted by $\widehat{D}_p = \{(j, \widehat{\pi}_p) \mid j \in \widehat{\mathcal{N}}(\widehat{\pi}_p)\}$;

Return: $\widehat{\pi} = (\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \widehat{\pi}_p)$, and $\widehat{E} = \cup_{m=\{2,3,\cdots,p\}} \widehat{D}_m$

DAG structures, Step 1) is performed because it reduces both computational and sample complexity by exploiting the sparsity of the moralized graph for the DAG.

The main purpose of Step 1) is to reduce the search-space by exploiting sparsity of the moralized graph. The moralized graph provides a *candidate parents* set for each node. Similar ideas of reducing the search-space by utilizing the moralized graph or different undirected graphs are applied in existing algorithms (e.g., Tsamardinos and Aliferis 2003; Friedman et al. 1999; Loh and Bühlmann 2014). The concept of candidate parents set exploits two properties; (i) the neighborhood of a node is a superset of its parents, and (ii) a node should appear later than its parents in the ordering. Hence, the candidate parents set for a given node j is the intersection of its neighborhood and elements of the ordering which appear before that node j, and is denoted by $C_{mj} := \mathcal{N}(j) \cap \{\pi_1, \pi_2, ..., \pi_{m-1}\}$ where m^{th} element of the ordering is j (i.e., $\pi_m = j$). The estimated candidate parents set is $\widehat{C}_{mj} := \widehat{\mathcal{N}}(j) \cap \{\widehat{\pi}_1, \widehat{\pi}_2, ..., \widehat{\pi}_{m-1}\}$ that is also specified in Algorithm 1.

This candidate parents set is used as a conditioning set for the overdispersion score in Step 2). In principle, the size of the conditioning set for an overdispersion score could be $p-1$ if the moralized graph is not used. Since Step 2) requires computation of a conditional mean and variance, both the computational complexity and sample complexity depend significantly on the number of variables we condition on as illustrated in Sections 3.1

and 3.2. Therefore by making the conditioning set for the overdispersion score of each node as small as possible, we gain significant computational and statistical improvements.

A number of choices are available for estimation of the moralized graph. Since the moralized graph is an undirected graph, standard undirected graph learning algorithms such as HITON (Aliferis et al., 2003) and MMPC algorithms (Tsamardinos and Aliferis, 2003) as well as ℓ_1-penalized likelihood regression for generalized linear models (GLM) (Friedman et al., 2009). In addition, standard DAG learning algorithms such as PC (Spirtes et al., 2000), GES (Chickering, 2003) and MMHC algorithms (Tsamardinos and Aliferis, 2003) can be applied to estimate the Markov equivalence class and then the moralized graph is generated from the Markov equivalence class.

Step 2) of the generalized ODS algorithm involves learning the ordering by comparing overdispersion scores of nodes using Equations (4) and (5). The basic idea is to determine which nodes are overdispersed based on the sample conditional mean and conditional variance after the transformation in Proposition 3. The ordering is determined one node at a time by selecting the node with the smallest overdispersion score which is representative of a node that is least likely to be overdispersed.

Regarding the overdispersion scores, suppose that there are n i.i.d. samples $X^{1:n} := (X^{(i)})_{i=1}^n$ where $X^{(i)} := (X_1^{(i)}, X_2^{(i)}, \cdots, X_p^{(i)})$ is a p-variate random vector drawn from an underlying QVF DAG model with quadratic variance coefficients $(\beta_{j0}, \beta_{j1})_{j=1}^p$. We use the notation $\widehat{\cdot}$ to denote an estimate based on $X^{1:n}$. In addition, we use $n(x_S) := \sum_{i=1}^n \mathbf{1}(X_S^{(i)} = x_S)$ for $x_S \in \mathcal{X}(S)$ to denote the conditional sample size, and $n_S := \sum_{x_S} n(x_S)\mathbf{1}(n(x_S) \geq c_0 \cdot n)$ for an arbitrary $c_0 \in (0,1)$ to denote a truncated conditional sample size. We discuss the choice of c_0 shortly.

More precisely the overdispersion scores in Step 2) of Algorithm 1 involves the following equations:

$$\widehat{\mathcal{S}}(1,j) \;:=\; \widehat{\omega}_j^2 \cdot \widehat{\mathrm{Var}}(X_j) - \widehat{\omega}_j \cdot \widehat{\mathbb{E}}(X_j) \;\text{ where }\; \widehat{\omega}_j := (\beta_{10} + \beta_{11}\widehat{\mathbb{E}}(X_j))^{-1}, \tag{4}$$

$$\widehat{\mathcal{S}}(m,j) \;:=\; \sum_{x \in \mathcal{X}_{\widehat{C}_{mj}}} \frac{n(x)}{n_{\widehat{C}_{mj}}} \Big[\widehat{\omega}_{mj}(x)^2 \widehat{\mathrm{Var}}(X_j \mid X_{\widehat{C}_{mj}} = x) - \widehat{\omega}_{mj}(x)\widehat{\mathbb{E}}(X_j \mid X_{\widehat{C}_{mj}} = x)\Big] \tag{5}$$

where $\widehat{\omega}_{mj}(x) := (\beta_{j0} + \beta_{j1}\widehat{\mathbb{E}}(X_j \mid X_{\widehat{C}_{mj}} = x))^{-1}$. $\widehat{C}_{mj}$ is the estimated candidate parents set of node j for the m^{th} element of the ordering and $\mathcal{X}_{\widehat{C}_{mj}} := \{x_{\widehat{C}_{mj}} \in \mathcal{X}(\widehat{C}_{mj}) : n(x_{\widehat{C}_{mj}}) \geq c_0 \cdot n\}$ to ensure we have enough samples for each element of an overdispersion score. c_0 is a tuning parameter of our algorithm that we specify in Theorem 14 and our numerical experiments. $\widehat{\omega}_{mj}(x)$ is an empirical version of the transformation in Proposition 3 assuming $\widehat{C}_{mj}$ is the parents of a node j. Since there are many conditional distributions, our overdispersion score is the weighted average of differences between conditional sample means and variances after the estimated transformation $\widehat{\omega}_{mj}(x)$. Then, the score is a measure of the level of overdispersion. As demonstrated in Section 2.2, the correct elements of an ordering achieve zero overdispersion score, otherwise positive in population.

Finding the set of parents of a node j boils down to selecting the parents out of all elements before a node j in the ordering. Hence given the estimated ordering from Step 2),

Step 3) can be reduced to p neighborhood selection problems which can be performed using ℓ_1-penalized likelihood regression for GLMs (Friedman et al., 2009) as well as standard DAG learning algorithms such as the PC (Spirtes et al., 2000), GES (Chickering, 2003), and MMHC algorithms (Tsamardinos and Aliferis, 2003).

3.1 Computational Complexity

For Steps 1) and 3) of the generalized ODS algorithm, we use off-the-shelf algorithms and the computational complexity depends on the choice of algorithm. For example, if we use the neighborhood selection ℓ_1-penalized likelihood regression for GLMs (Friedman et al., 2009) as is used in Yang et al. (2012), the worst-case complexity is $O(\min(n,p)np)$ for a single ℓ_1-penalized likelihood regression, but since there are p nodes, the total worst-case complexity is $O(\min(n,p)np^2)$. Similarly, if we use ℓ_1-penalized likelihood regression for Step 3) the worst-case complexity is also $O(\min(n,p)np^2)$ but maybe less if the degree d of the moralized graph is small.

For Step 2) where we estimate the ordering, there are $(p-1)$ iterations and each iteration has a number of overdispersion scores $\widehat{\mathcal{S}}(m,j)$ to be computed which is bounded by $O(p)$. Hence the total number of overdispersion scores that need to be computed is $O(p^2)$. Since the time for calculating each overdispersion score is proportional to the sample size n, the complexity is $O(np^2)$.

Hence, Step 1) is the main computational bottleneck of the generalized ODS algorithm. The addition of Step 2) which estimates the ordering does not significantly add to the computational bottleneck. Consequently, the generalized ODS algorithm, which is designed for learning DAGs is almost as computationally efficient as standard methods for learning undirected graphical models. As we show in numerical experiments, the ODS algorithm using ℓ_1-penalized likelihood regression for GLMs in both Steps 1) and 3) is faster than the state-of-the-art GES algorithm.

3.2 Statistical Guarantees

In this section, we provide theoretical guarantees for our generalized ODS algorithm. We provide sample complexity guarantees for the algorithm in the high-dimensional setting in three steps, by proving consistency of Steps 1), 2) and 3) in Sections 3.2.1, 3.2.2 and 3.2.3, respectively. All three main results are expressed in terms of the triple (n,p,d).

Although any off-the-shelf algorithms can be used in Steps 1) and 3), our theoretical guarantees focus on the case when we use the R package glmnet (Friedman et al., 2009) for neighborhood selection. We focus on glmnet since there exist provable theoretical guarantees for neighborhood selection for graphical model learning in the high-dimensional setting (see e.g., Yang et al. 2012; Ravikumar et al. 2010) and performs well in our simulation study. The glmnet package involves minimizing the ℓ_1-penalized generalized linear model loss.

Without loss of generality, assume that $(1,2,\cdots,p)$ is the true ordering and for ease of notation let $[\cdot]_k$ and $[\cdot]_S$ denotes parameter(s) corresponding to the variable X_k and random vector X_S, respectively. Suppose that $\theta^*_{D_j} \in \Theta_{D_j}$ denotes the solution of the following GLM problem where $\Theta_{D_j} := \{\theta \in \mathbb{R}^p : [\theta]_k = 0 \text{ for } k \notin \text{pa}(j)\}$.

$$\theta^*_{D_j} := \arg\min_{\theta \in \Theta_{D_j}} \mathbb{E}\left(-X_j([\theta]_j + \langle [\theta]_{\text{pa}(j)}, X_{\text{pa}(j)}\rangle) + A_j([\theta]_j + \langle [\theta]_{\text{pa}(j)}, X_{\text{pa}(j)}\rangle)\right), \quad (6)$$

where $A_j(\cdot)$ is the log-partition function determined by the GLM family (3), and $\langle \cdot, \cdot \rangle$ represents the inner product. In the special case where X_j has an NEF-QVF distribution (3) with log-partition function $A_j(\cdot)$, $\theta^*_{D_j}$ corresponds exactly to the set of true parameters, that is θ^*_{jk} is the coefficient $k \in \text{pa}(j)$ which represents the influence of node k on node j. However our results apply more generally and we do not require that X_j belongs to an NEF-QVF DAG model.

Similar definitions are required for parameters associated with the moralized graph G^m. Define $\theta^*_{M_j} \in \Theta_{M_j}$ as the solution of the following GLM problem for a node j over its neighbors where $\Theta_{M_j} := \{\theta \in \mathbb{R}^p : [\theta]_k = 0 \text{ for } k \notin \mathcal{N}(j)\}$.

$$\theta^*_{M_j} := \arg\min_{\theta \in \Theta_{M_j}} \mathbb{E}\left(-X_j([\theta]_j + \langle [\theta]_{\mathcal{N}(j)}, X_{\mathcal{N}(j)} \rangle) + A_j([\theta]_j + \langle [\theta]_{\mathcal{N}(j)}, X_{\mathcal{N}(j)} \rangle)\right). \quad (7)$$

We impose the following identifiability assumptions on $\theta^*_{D_j}$ and $\theta^*_{M_j}$ for ensuring each parents and each neighbor has non-zero influence on a node j, respectively.

Assumption 6 *(a) For any node $j \in V$ and $k \in pa(j)$,*

$$Cov(X_j, X_k) \neq Cov(X_k, \bigtriangledown A_j([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{pa(j)\setminus k}, X_{pa(j)\setminus j} \rangle)).$$

(b) For any node $j \in V$ and $k \in \mathcal{N}(j)$,

$$Cov(X_j, X_k) \neq Cov(X_k, \bigtriangledown A_j([\theta^*_{M_j}]_j + \langle [\theta^*_{M_j}]_{\mathcal{N}(j)\setminus k}, X_{\mathcal{N}(j)\setminus j} \rangle)).$$

Assumption 6 can be understood as a notion of restricted faithfulness only for neighbors and parents for each node. To provide intuition consider the special case of Gaussian DAG models. The log-partition function is $A_j(\eta) = \frac{\eta^2}{2}$, so that $\bigtriangledown A_j(\eta) = \eta$. Then, the condition boils down to $\text{Cov}(X_j, X_k) \neq \sum_{m \in \text{pa}(j)\setminus k} [\theta^*_{D_j}]_m \text{Cov}(X_k, X_m)$, meaning the directed path from X_k to X_j does not exactly cancel the sum of paths from other parents of X_k. For general exponential families, the right-hand side involves non-linear functions of the variables of X corresponding to sets of measure 0. Under Assumption 6, the following result holds.

Lemma 7 *(a) Under Assumption 6(a), for all $1 \leq j \leq p$, $supp(\theta^*_{D_j}) = pa(j)$.*

(b) Under Assumption 6(b), for all $1 \leq j \leq p$, $supp(\theta^*_{M_j}) = \mathcal{N}(j)$.

Using the parameters $(\theta^*_{M_j})_{j=1}^p$ and $(\theta^*_{D_j})_{j=1}^p$ and their relationships to $\text{pa}(j)$ and $\mathcal{N}(j)$ respectively, we provide consistency guarantees for Steps 1) and 3) respectively.

3.2.1 Step 1): Recovery of the Moralized Graph via ℓ_1-penalized Likelihood Regression for GLMs

We first focus on the theoretical guarantee for recovering the moralized graph G^m. As we mentioned earlier, we approach this problem by solving an empirical version of the ℓ_1-penalized likelihood regression. Given n i.i.d. samples $X^{1:n} = (X^{(i)})_{i=1}^n$ where $X^{(i)} =$

$(X_1^{(i)}, X_2^{(i)} \cdots, X_p^{(i)})$ is a p-variate random vector drawn from the underlying DAG model, we define the conditional negative log-likelihood for a variable X_j:

$$\ell_j(\theta; X^{1:n}) := \frac{1}{n}\sum_{i=1}^{n}\Big(-X_j^{(i)}([\theta]_j + \langle [\theta]_{V\setminus j}, X_{V\setminus j}^{(i)}\rangle) + A_j([\theta]_j + \langle [\theta]_{V\setminus j}, X_{V\setminus j}^{(i)}\rangle)\Big) \tag{8}$$

where $\theta \in \mathbb{R}^p$ and $A_j(\cdot)$ is the log-partition function determined based on the chosen GLM family (3).

We analyze the ℓ_1-penalized log-likelihood for each node $j \in V$:

$$\hat{\theta}_{M_j} := \arg\min_{\theta\in\mathbb{R}^p} \ell_j(\theta; X^{1:n}) + \lambda_n \|[\theta]_{V\setminus j}\|_1 \tag{9}$$

where $\lambda_n > 0$ is the regularization parameter. Based on $\hat{\theta}_{M_j}$, the estimated neighborhood of node j is $\widehat{\mathcal{N}}(j) := \{k \in V \setminus j : [\hat{\theta}_M]_k \neq 0\}$. Based on Lemma 7, $\text{supp}(\theta^*_{M_j}) = \mathcal{N}(j)$ where $\theta^*_{M_j}$ is defined by (7). Hence if for each j, $\hat{\theta}_{M_j}$ in (9) is sufficiently close to $\theta^*_{M_j}$, we conclude that $\widehat{\mathcal{N}}(j) = \mathcal{N}(j)$.

We begin by discussing the assumptions we impose on the DAG G. Since we apply the neighborhood selection strategy in Steps 1) and 3), we will present assumptions for both steps here. Most of the assumptions are similar to those imposed in Yang et al. (2012) where neighborhood selection is used for graphical model learning. Important quantities are the Hessian matrices of the negative conditional log-likelihood of a variable X_j given either the rest of the nodes $Q^{M_j} = \bigtriangledown^2 \ell_j(\theta^*_{M_j}; X^{1:n})$, and the nodes before j in the ordering $Q^{D_j} = \bigtriangledown^2 \ell_j^D(\theta^*_{D_j}; X^{1:n})$ which we discuss in Section 3.2.3. Let A_{SS} be the $|S| \times |S|$ submatrix of the matrix A_j corresponding to variables X_S.

Assumption 8 (Dependence assumption) *There exists a constant $\rho_{\min} > 0$ such that*

$$\min_{j\in V} \min\Big(\lambda_{\min}(Q^{M_j}_{\mathcal{N}(j)\mathcal{N}(j)}), \lambda_{\min}(Q^{D_j}_{pa(j)pa(j)})\Big) \geq \rho_{\min}.$$

Moreover, there exists a constant $\rho_{\max} < \infty$ such that

$$\max_{j\in V}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} X^{(i)}_{\mathcal{N}(j)}(X^{(i)}_{\mathcal{N}(j)})^T\right)\right) \leq \rho_{\max}$$

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

Assumption 9 (Incoherence assumption) *There exists a constant $\alpha \in (0, 1]$ such that*

$$\max_{j\in V} \max\left(\max_{t\in\mathcal{N}(j)^c} \|Q^{M_j}_{t\mathcal{N}(j)}(Q^{M_j}_{\mathcal{N}(j)\mathcal{N}(j)})^{-1}\|_1, \max_{t'\in pa(j)^c} \|Q^{D_j}_{t'pa(j)}(Q^{D_j}_{pa(j)pa(j)})^{-1}\|_1\right) \leq 1-\alpha.$$

The dependence assumption 8 can be interpreted as ensuring that the variables in both $\mathcal{N}(j)$ and $\text{pa}(j)$ are not too dependent. In addition, the incoherence assumption 9 ensures that variables that are not in the set of true variables are not highly correlated with

variables in the true variable set. These two assumptions are standard in all neighborhood regression approaches for variable selection involving ℓ_1-based methods and these conditions have imposed in proper work both for high-dimensional regression and graphical model learning (Yang et al., 2012; Meinshausen and Bühlmann, 2006; Wainwright et al., 2006; Ravikumar et al., 2011).

To ensure suitable concentration bounds hold, we impose two further technical assumptions. Firstly we require a boundedness assumption on the moment generating function to control the tail behavior.

Assumption 10 (Concentration bound assumption) *There exists a constant $M > 0$ such that*

$$\max_{j \in V} \mathbb{E}(exp(|X_j|)) < M.$$

We also require conditions on the first and third derivatives on the log-partition functions $A_j(.)$ for $1 \leq j \leq p$ in Equations (8) and (10). Let $A_j'(.)$ and $A_j'''(.)$ are the first and third derivatives of $A_j(.)$ respectively.

Assumption 11 (Log-partition assumption) *For the log-partition functions $A_j(\cdot)$ in Equation (8) or (10), there exist constants κ_1 and κ_2 such that $\max_{j \in V}\{|A_j'(a)|, |A_j'''(a)|\} \leq n^{\kappa_2}$ for $a \in [0, \kappa_1 \max\{\log(n), \log(p)\})$, $\kappa_1 \geq 6\max(\|\theta^*_{M_j}\|_1, \|\theta^*_{D_j}\|_1)$ and $\kappa_2 \in [0, 1/4]$.*

Prior work in Yang et al. (2012); Ravikumar et al. (2011); Jalali et al. (2011) impose similar technical conditions that control the tail behavior of $(X_j)_{j=1}^p$. It is important to note that there exist many distributions and associated parameters that satisfy these assumptions. For example the Binomial, Multinomial or Exponential distributions, the log-partition assumption 11 is satisfied with $\kappa_2 = 0$ because the log-partition function $A_j(\cdot)$ is bounded. For the Poisson distribution which has one of the steepest log-partition function, $A_j(\cdot) = \exp(\cdot)$. Hence, in order to satisfy Assumption 11, we require $\|\theta^*_{M_j}\|_1 \leq \frac{\log n}{48 \log p}$ with $\kappa_2 = \frac{1}{8}$.

Putting together Assumptions 8 9, 10, and 11, we have the following main result that the moralized graph can be recovered via ℓ_1-penalized likelihood regression for GLMs in high-dimensional settings.

Theorem 12 (Learning the moralized graph) *Consider the DAG model (1) satisfying the QVF property (2) and d is the maximum degree of the moralized graph. Suppose that Assumptions 6(b), 8, 9, 10 and 11 are satisfied. Assume $\hat{\theta}_{M_j}$ is any solution to the optimization problem (9) and $\frac{9\log^2(\max\{n,p\})}{n^a} \leq \lambda_n \leq \frac{\rho^2_{\min}}{30 n^{\kappa_2}\log(\max\{n,p\})d\rho_{\max}}$ for some $a \in (2\kappa_2, 1/2)$, and $\min_{j \in V}\min_{t \in \mathcal{N}(j)}|[\theta^*_M]_t| \geq \frac{10}{\rho_{\min}}\sqrt{d}\lambda_n$. Then for any constant $\epsilon > 0$, there exists a positive constant C_ϵ such that if $n \geq C_\epsilon(d\log^3\max\{n,p\})^{\frac{1}{a-\kappa_2}}$,*

$$\mathbb{P}(supp(\hat{\theta}_{M_j}) = \mathcal{N}(j)) \geq 1 - \epsilon,$$

for all $j \in V$.

We defer the proof to Appendix C. The key technique for the proof is that standard *primal-dual witness* method used in Wainwright et al. (2006); Ravikumar et al. (2011); Jalali et al. (2011); and Yang et al. (2012). Theorem 12 shows that the moralized graph G^m can be recovered via ℓ_1-penalized likelihood regression if sample size $n = \Omega((d\log^3(\max\{n,p\}))^{\frac{1}{a-\kappa_2}})$ with high probability.

3.2.2 Step 2): Recovering the Ordering using OverDispersion Scores

In this section, we provide theoretical guarantees for recovering the ordering for the DAG G via our generalized ODS algorithm. The first required condition is a stronger version of the identifiability assumption (Assumption 4) since we move from the population distribution to the finite sample setting.

Assumption 13 *For all $j \in V$ and any $pa_0(j) \subset pa(j)$ where $pa_0(j) \neq \emptyset$ and $S \subset nd(j) \setminus pa_0(j)$:*

(a) There exists an $M_{\min} > 0$ such that $Var(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_S) > M_{\min}$.

(b) There exists an $\omega_{\min} > 0$ such that $|\beta_{j0} + \beta_{j1}\mathbb{E}(X_j \mid X_S)| > \omega_{\min}$.

Assumption 10 is required since the overdispersion score is sensitive to the accuracy of the sample conditional mean and conditional variance. Since the true ordering π^* may not be unique, we use $\mathcal{E}(\pi^*)$ to denote the set of all the orderings that are consistent with the true DAG G.

Theorem 14 (Recovery of the ordering) *Consider the DAG model (1) satisfying the QVF property (2) with co-efficients $(\beta_{j0}, \beta_{j1})_{j=1}^p$ and d is the maximum degree of the moralized graph. Suppose that $\beta_{j1} > -1$ for all $j \in V$, and the structure of the moralized graph G^m is known. Suppose also that Assumptions 10 and 13 are satisfied. Then for any $\epsilon > 0$ and $c_0 \geq \log^d(\max\{n,p\})$, there exists a positive constant K_ϵ such that for $n \geq K_\epsilon \log^{5+d}(\max\{n,p\})$,*

$$P(\widehat{\pi} \in \mathcal{E}(\pi^*)) \geq 1 - \epsilon.$$

The detail of the proof is provided in Appendix D. The proof is novel and involves the combination of the transformation and overdispersion property exploited in Theorem 5. Intuitively, the estimated overdispersion scores $\widehat{\mathcal{S}}(m,j)$ converge to the true overdispersion scores $\mathcal{S}(m,j)$ as the sample size n increases which is where we exploit Assumption 10. This allows us to recover a true ordering for the DAG G. Assuming the moralized graph G^m is known is essential to exploiting the degree condition on the moralized graph and emphasizes the importance of Step 1) and Theorem 12.

Theorem 14 claims that if the triple (n,d,p) satisfies $n = \Omega(\log^{5+d} p)$, our generalized ODS algorithm correctly estimates the true ordering. Therefore if the moralized graph is sparse (i.e., $d = \Omega(\log p)$), our generalized ODS algorithm recovers the true casual ordering in the high-dimensional settings. Note that if the moralized graph is not sparse and $d = \Omega(p)$, the generalized ODS algorithm requires an extremely large sample size. Prior work on DAG learning algorithms in the high-dimensional setting has been based on learning the Markov equivalence class in settings with additive independent noise (see e.g., Loh and Bühlmann 2014; van de Geer and Bühlmann 2013).

3.2.3 Step 3): Recovery of the DAG via ℓ_1-penalized Likelihood Regression

Similar to Step 1), we provide a theoretical guarantee for Step 3) using ℓ_1-penalized likelihood regression where we estimate the parents of each node $\text{pa}(j)$. Importantly, we assume that Step 2) of the ODS algorithm has occurred and using Theorem 14, a true ordering has been learned. Recall that we impose the assumption that the true ordering is $\pi^* = (1, 2, \cdots, p)$. Then, we estimate the parents of a node j over the possible parents $\{1, 2, \cdots, j-1\}$.

For notational convenience, we use $X_{1:j} = (X_1, X_2, \cdots, X_j)$. Then for any variable X_j, the conditional negative log-likelihood for a given GLM is as follows:

$$\ell_j^D(\theta; X^{1:n}) := \frac{1}{n}\sum_{i=1}^{n}\Big(-X_j^{(i)}([\theta]_j + \langle [\theta]_{1:j-1}, X_{1:j-1}^{(i)}\rangle) + A_j([\theta]_j + \langle [\theta]_{1:j-1}, X_{1:j-1}^{(i)}\rangle)\Big) \quad (10)$$

where $\theta \in \mathbb{R}^j$, and $A_j(\cdot)$ is the log-partition function determined by a chosen GLM family.

We solve the negative conditional log-likelihood with ℓ_1 norm penalty for each variable X_j:

$$\hat{\theta}_{D_j} := \arg\min_{\theta \in \mathbb{R}^j} \ell_j^D(\theta; x) + \lambda_n^D \|[\theta]_{1:j-1}\|_1. \quad (11)$$

Recall that under Assumption 6(a), Lemma 7(a) shows that $\text{supp}(\theta^*_{D_j}) = \text{pa}(j)$. Hence if the solution of Equation (11) for each node $j \in V$ is close to $\theta^*_{D_j}$ in Equation (6), ℓ_1-penalized likelihood regression successfully recovers the parents of node j.

Theorem 15 (Learning DAG structure) *Consider the DAG model* (1) *satisfying the QVF property* (2) *and d is the maximum degree of the moralized graph. Suppose that Assumptions 6(a), 8, 9, 10 and 11 are satisfied. Assume $\hat{\theta}_{D_j}$ is any solution to the optimization problem* (11) *and* $\frac{9\log^2(\max\{n,p\})}{n^a} \leq \lambda_n^D \leq \frac{\rho_{\min}^2}{30n^{\kappa_2}\log(\max\{n,p\})d\rho_{\max}}$ *for some $a \in (2\kappa_2, 1/2)$, and* $\min_{j\in V}\min_{t\in\mathcal{N}(j)}|[\theta^*_D]_t| \geq \frac{10}{\rho_{\min}}\sqrt{d}\lambda_n$. *Then for any $\epsilon > 0$, there exists a positive constant C_ϵ such that if* $n \geq C_\epsilon(d\log^3(\max\{n,p\}))^{\frac{1}{a-\kappa_2}}$,

$$\mathbb{P}(\textit{supp}(\hat{\theta}_{D_j}) = \textit{pa}(j)) \geq 1 - \epsilon,$$

for all $j \in V$.

The details of the proof are provided in Appendix E. The proof technique is again based on the primal-dual technique as is used for the proof of Theorem 12. Theorem 15 shows that ℓ_1-penalized likelihood regression successfully recovers the structure of G if the sample size is $n = \Omega((d\log^3(\max\{n,p\}))^{\frac{1}{a-\kappa_2}})$ given the true ordering. Note once again that we exploit the sparsity d of the moralized graph.

So far, we have provided sample complexity guarantees for all three steps of the generalized ODS algorithm. Combining Theorems 12, 14, and 15, we reach our final main result that the generalized ODS algorithm successfully recovers the true structure of a QVF DAG with high probability. Furthermore if G is sparse (i.e., $d = \Omega(\log p)$), the generalized ODS algorithm recovers the structure of QVF DAG models in the high-dimensional setting.

Corollary 16 (Learning QVF DAG models) *Consider the DAG model* (1) *satisfying the QVF property* (2) *and d is the maximum degree of the moralized graph. Suppose that Assumptions 6, 8, 9, 10 and 11 are satisfied and all other conditions of Theorems 12, 14, and 15 are satisfied and $\widehat{G}$ is the output of the ODS algorithm. Then for any $\epsilon > 0$, there exists a positive constant C_ϵ such that if $n \geq C_\epsilon \max(d \log^3(\max\{n, p\}))^{\frac{1}{a-\kappa_2}}, \log^{5+d} p)$,*

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

Concretely, we apply Corollary 16 to popular examples for our class of QVF DAG models. As we discussed earlier, Poisson DAG models have $(\beta_{j0}, \beta_{j1}) = (1, 0)$, the steepest log-partition function $A_j(\cdot) = \exp(\cdot)$, and $\kappa_2 = \frac{1}{8}$ if $\|\theta^*_{M_j}\|_1 \leq \frac{\log n}{48 \log(\max\{n,p\})}$. Then, our generalized ODS algorithm recovers Poisson DAG models with high probability if $n = \Omega(\max\{(d \log^3 p)^4, \log^{5+d} p\})$ and $a = \frac{3}{8}$. Binomial DAG models have $(\beta_{0j}, \beta_{1j}) = (0, -\frac{1}{N})$ where N is a binomial distribution parameter, the log-partition function $A_j(\cdot) = N \log(1 + \exp(\cdot))$, $\kappa_2 = 0$. Then, the generalized ODS algorithm recover Binomial DAG models with high probability if $n = \Omega(\max\{(d \log^3 p)^3, \log^{5+d} p\})$ and $a = \frac{1}{3}$.

4. Simulation Experiments

In this section, we support our theoretical guarantees with numerical experiments and show that our generalized ODS algorithm 1 performs favorably compared to state-of-the-art DAG learning algorithms when applied to QVF DAG models. In order to validate Theorems 12, 14, and 15, we conduct a simulation study using 50 realizations of p-node Poisson and Binomial DAG models (3). That is, the conditional distribution for each node given its parents is either Poisson and Binomial. For all our simulation results, we generate DAG models (see Figure 3) that ensure a unique ordering $\pi^* = (1, 2, \cdots, p)$ with edges randomly generated while respecting the desired maximum number of parents constraints for the DAG. In our experiments, we always set the number of parents to two (the number of neighbors of each node is at least three, and therefore $d \in [3, p-1]$).

The set of parameters (θ_{jk}) for our GLM DAG models (3) encodes the DAG structure as follows: if there is no directed edge from node k to j, $\theta_{jk} = 0$, otherwise $\theta_{jk} \neq 0$. Non-zero parameters $\theta_{jk} \in E$ were generated uniformly at random in the range $\theta_{jk} \in [-1, -0.5]$ for Poisson DAG models and $\theta_{jk} \in [0.5, 1]$ for Binomial DAG models. In addition, we fixed parameters $N_1, N_2, \cdots, N_p = 4$ for Binomial DAG models. These parameter values were chosen to ensure Assumptions 10 and 11 are satisfied and most importantly, the count values do not blow up. Lastly, we set the thresholding constant for computing the ODS score to $c_0 = 0.005$ although any value below 0.01 seems to work well in practice. We consider more general parameter choices but for brevity, focus on these parameter settings.

To validate Theorems 12 and 14, we plot the proportion (out of 50) of simulations in which our generalized ODS algorithm recovers the correct ordering to validate π^* in Figure 4. We plot the accuracy rates in recovering the true ordering $\mathbf{1}(\hat{\pi} = \pi^*)$ as a function of the sample size ($n \in \{100, 500, 1000, 2500, 5000, 10000\}$) for different node sizes ($p = 10$ for (a) and (c), and $p = 100$ for (b) and (d)) and different distributions (Poisson for (a) and (b) and Binomial for (c) and (d)). In each sub-figure, two different choices for off-the-shelf algorithms for Step 1) are used; (i) ℓ_1 penalized likelihood regression (Friedman et al., 2009) where we chose the regularization parameter $\lambda = \frac{0.75}{\log(\max\{n,p\})}$ for Poisson DAG models and

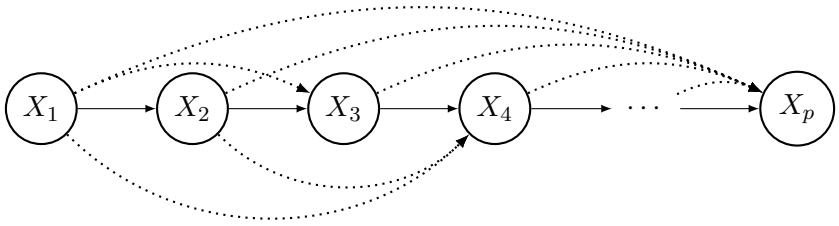
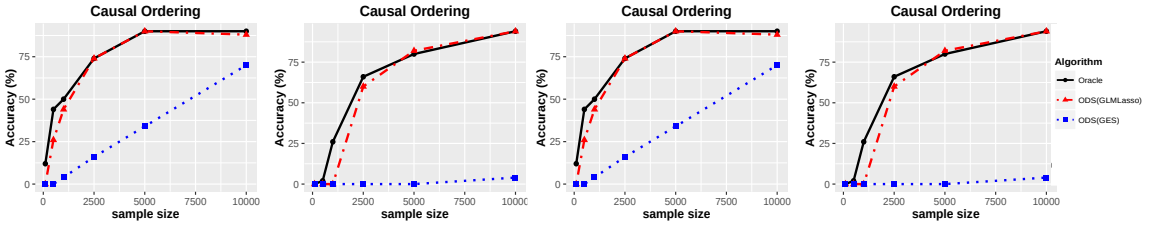


Figure 3: Structure of the DAG we used in numerical experiments. Solid directed edges are always present and dotted directed edges are randomly chosen based on the given number of parents of each node constraints



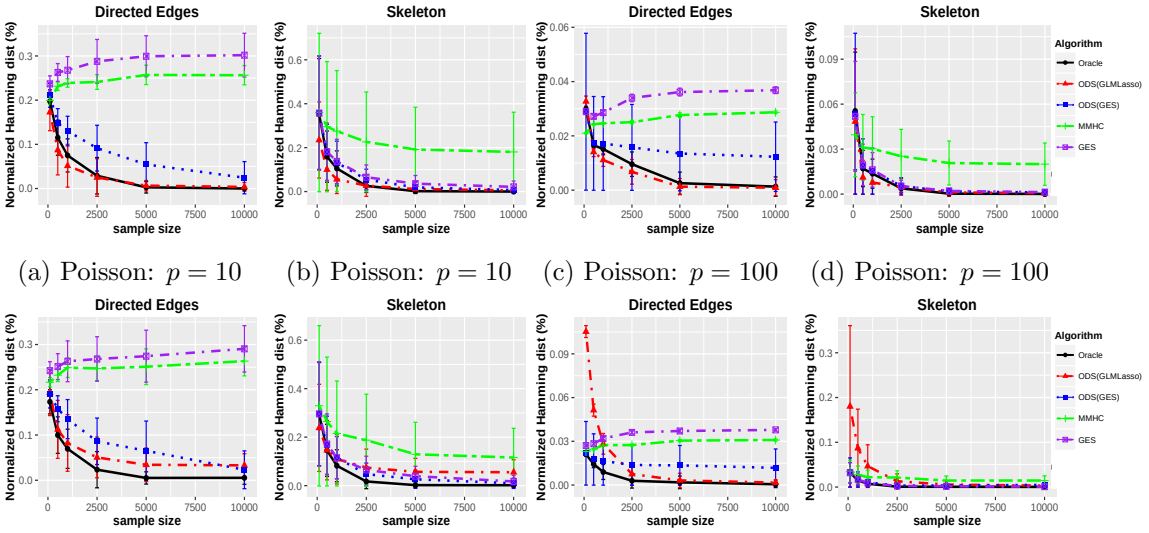
(a) Poisson: $p = 10$ (b) Poisson: $p = 100$ (c) Binomial: $p = 10$ (d) Binomial: $p = 100$

Figure 4: Probability of recovering the ordering of a DAG via our generalized ODS algorithm using two different algorithms (ℓ_1-penalized likelihood regression and GES algorithm) in Step 1)

$\lambda = \frac{.10}{\log(\max\{n,p\})}$ for Binomial DAG models; and (ii) the GES algorithm (Chickering, 2003) is applied for Step 1) where we used the mBDe (modified Bayesian Dirichlet equivalent, Heckerman et al. 1995) score and then the moralized graph is generated by moralizing the estimated DAG.

Figure 4 shows that our generalized ODS algorithm recovers the true ordering π^* well if the sample size is large, which supports our theoretical results. In addition, we can see that the ℓ_1-penalized based generalized ODS algorithm seems to perform substantially better than the GES-based ODS algorithm. Furthermore, since ℓ_1-penalized likelihood regression is the only algorithm that scales to the high-dimensional setting ($p \geq 1000$), we used ℓ_1-penalized likelihood regression in Steps 1) and 3) of the generalized ODS algorithm for large-scale DAG models.

Figure 5 provides a comparison of how accurately our generalized ODS algorithm performs in terms of recovering the full DAG model. We use two comparison metrics related to how many edges and directions are incorrect. First, we measured the Hamming distance between the skeleton (edges without directions) of the true DAG and the estimated DAG in (a), (c), (e) and (g). In addition, we measured the Hamming distance between the estimated and true DAG models (with directions) in (b), (d), (f), and (h). We normalized the Hamming distances by dividing by the maximum number of errors $\binom{p}{2}$ for the skeleton and



(a) Poisson: $p = 10$ (b) Poisson: $p = 10$ (c) Poisson: $p = 100$ (d) Poisson: $p = 100$

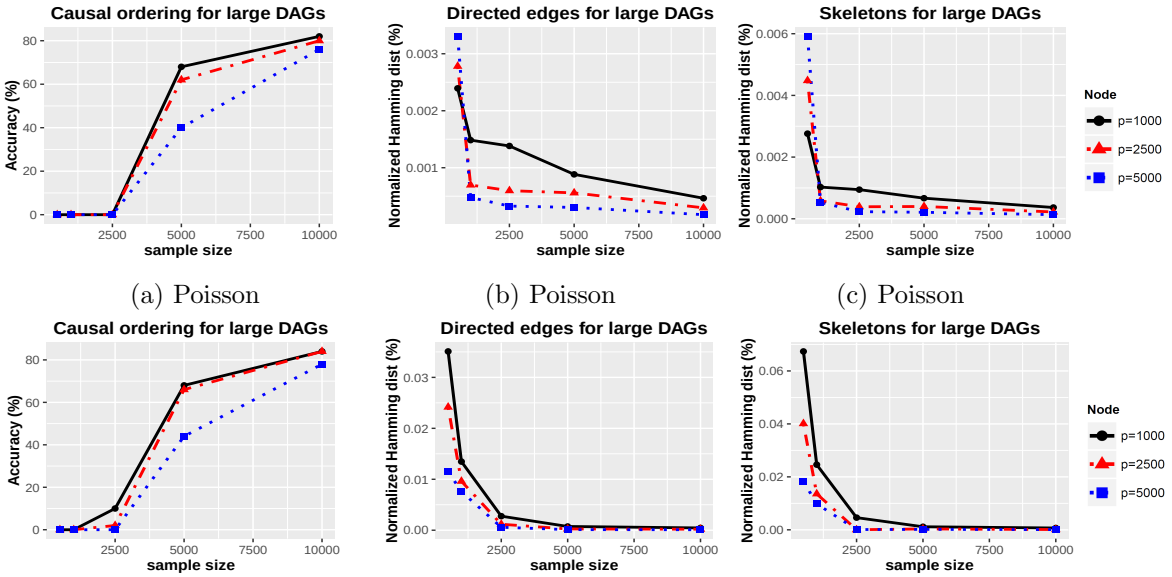
(e) Binomial: $p = 10$ (f) Binomial: $p = 10$ (g) Binomial: $p = 100$ (h) Binomial: $p = 100$

Figure 5: Comparison of the generalized ODS algorithms using ℓ_1-penalized likelihood regression (in Steps 1) and 3)) and the GES algorithm (in Steps 1) and 3)) to two state-of-the-art DAG learning algorithms (the MMHC and the GES algorithms) in terms of Hamming distance to skeletons and directed edges of Poisson and Binomial DAG models.

$p(p-1)$ for the full DAG respectively meaning the maximum normalized distance is 1. We compare to two state-of-the-art directed graphical model learning algorithms, the MMHC and GES algorithms for both Poisson and Binomial DAG models. Similar to learning the ordering, we used two generalized ODS algorithms exploiting ℓ_1-penalization in both Steps 1) and 3) and the GES algorithm in both Steps 1) and 3). We considered small-scale DAG models with $p = 10$ in (a), (b), (e) and (f), and $p = 100$ in (c), (d), (g) and (h).

As we see in Figure 5, the ODS algorithms significantly out-perform state-of-the-art MMHC and GES algorithms in terms of directed edges and skeleton. For small sample sizes, the generalized ODS algorithms have poor performance because they fail to recover the ordering, however we can see that the GES-based generalized ODS algorithm always performs better than the GES algorithm. This is because the generalized ODS algorithm adds directional information to the estimated skeleton via the GES algorithm, and hence the GES-based generalized ODS algorithm cannot be worse than the GES algorithm in terms of recovering both directed edges and skeleton. Furthermore Figure 5 shows that as sample size increases, our generalized ODS algorithms recovers the true directed edges and the skeleton for the DAG more accurately than state-of-the-art methods, which is consistent with our theoretical results.

Next we consider the performance for large-scale DAG models to show that the ODS algorithm works in the high-dimensional setting. In all experiments, we used the ℓ_1-penalized likelihood regression for GLMs in Steps 1) and 3) for the generalized ODS algorithm since it is the only graph-learning algorithm that scales. Figure 6 plots the statistical performance



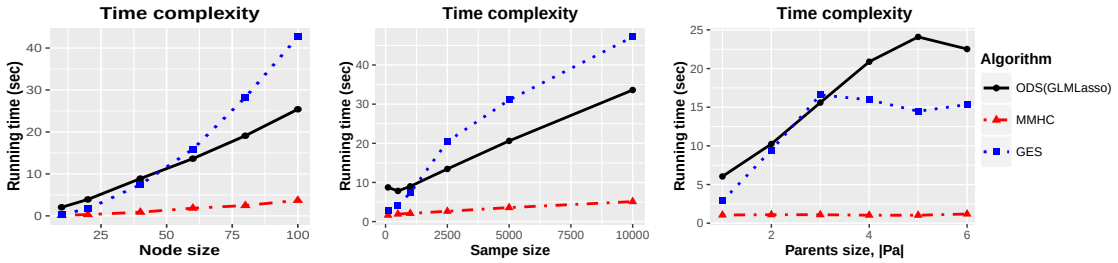
(a) Poisson (b) Poisson (c) Poisson

(d) Binomial (e) Binomial (f) Binomial

Figure 6: Performance of the generalized ODS algorithm using ℓ_1-penalized likelihood regression in both Steps 1) and 3) for large-scale DAG models with the node size $p = \{1000, 2500, 5000\}$

of the generalized ODS algorithm for large-scale Poisson DAGs in (a), (b), and (c) and Binomial DAGs in (d), (e), and (f). Furthermore, (a) and (d) represent the accuracy rates of the recovering the ordering, (b) and (e) show the normalized Hamming distance to the true skeleton, and (c) and (f) show the normalized Hamming distance for the true edge set of the DAG. Accuracies vary as a function of sample size ($n \in \{500, 1000, 2500, 5000, 10000\}$) for each node size ($p = \{1000, 2500, 5000\}$). Similar to small-scale DAG models, Figure 6 shows that the generalized ODS algorithm recovers the ordering and the skeleton of the DAG in the high-dimensional settings.

In Figure 7, we compared the run-time of the generalized ODS algorithms using ℓ_1-penalized likelihood regression for GLMs in Steps 1) and 3) to the run-time of the MMHC and the GES algorithms. We measured the run-time for Poisson DAG models by varying (a) node size $p \in \{10, 20, 40, 60, 80, 100\}$ with fixed sample size $n = 10000$ and exactly two parents of each node, (b) sample size $n \in \{100, 500, 1000, 2500, 5000, 10000\}$ with the fixed node size $p = 100$ and two parents of each node, and (c) the number of parents of each node $\in \{1, 2, 3, 4, 5, 6\}$ with the fixed sample size $n = 10000$ and node size $p = 20$. The results of (a) and (b) show that the generalized ODS algorithm is not always slower than the GES algorithm. In addition, (c) also shows that the run-time of the generalized ODS algorithm depends significantly on the number of parents for each node. Figure 7 shows that the generalized ODS algorithm is significantly slower than the MMHC algorithm, however this is because the MMHC algorithm often stops earlier before they reach the true DAG (see Figure 5).



(a) Poisson: $n = 10^4, d \geq 3$ (b) Poisson: $p = 100, d \geq 3$ (c) Poisson: $n = 10^4, p = 20$

Figure 7: Comparison of the generalized ODS algorithms using ℓ_1-penalized likelihood regression in Steps 1) and 3) to two standard DAG learning algorithms (the MMHC and the GES algorithms) in terms of running time with respect to (a) node size p, (b) sample size n, and (c) number of parents of each node

5. Real Multi-variate Count Data: 2009/2010 NBA Player Statistics

In terms of real data applications, one of the advantages of our ODS algorithm is that it provides a scalable approach for learning DAG models when variables are counts. In particular other approaches such as GES, 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 this section, we demonstrate this advantage using a simple data set that involves multi-variate count data which models basketball statistics for NBA players during the 2009/10 season. To the best of our knowledge, our ODS algorithm is the only algorithm that provides a reliable and scalable approach for DAG learning with multi-variate count data, albeit under strong assumptions.

Our data set consists of 441 NBA player statistics from season 2009/2010 (see R package SportsAnalytics for detailed information). The original data set contains 24 covariates: player name, team name, players position (PG, SG, SF, PF or C), total minutes played, total number of field goals made, field goals attempted, threes made, threes attempted, free throws made, free throws attempted, offensive rebounds, rebounds, assists, steals, turnovers, blocks, personal fouls, disqualifications, technicals fouls, ejections, flagrant fouls, games started and total points. We eliminated player name, team name, number of games played, and players position, because our focus is to find the directional or causal relationships between statistics. We also eliminated ejections and flagrant fouls because both did not occur in our data set. Therefore the data set we consider contains 18 variables.

As we see in Figure 8 (left), all 18 variables are positively correlated. This makes sense because the total minutes played is likely to be positively correlated with other statistics, and some statistics have causal relationships (e.g., the more shooting attempt implies the more shooting made). The box plots in Figure 8 (right) show that the NBA statistics are significantly different depending on the player position. This is also makes sense because each position takes a different role. C and PF are expected to play near the baseline, hence they have more rebounds, blocks, and fouls. PG is expected to pass a ball and play far

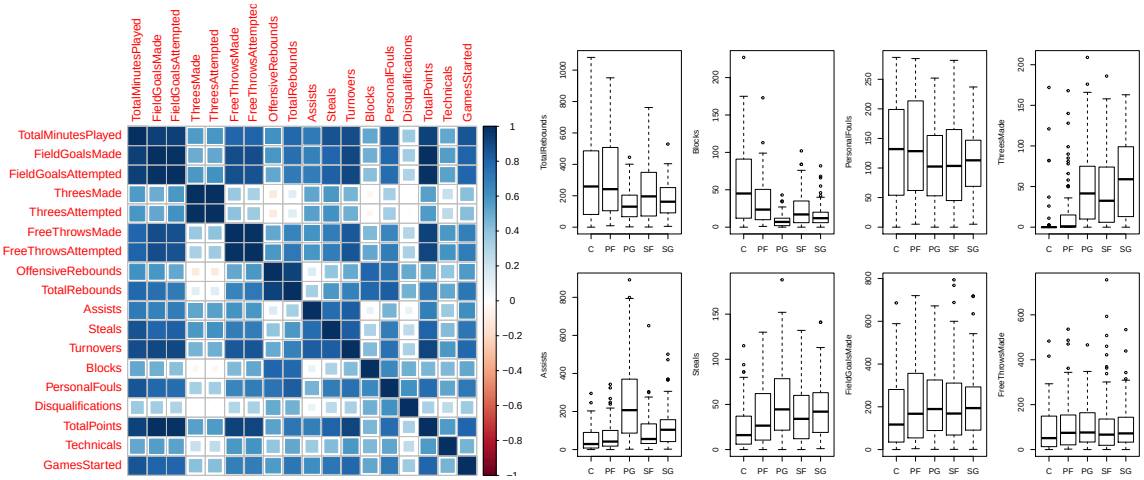


Figure 8: Correlation Plots for NBA statistics (left). Blue represents a high correlation and white represents a small correlation. Box plots for some NBA statistics depending on positions (right). Box plots consider the total number of rebounds, blocks, personal fouls, assists, steals, three points made, field goals made, and free throws made.

from the basket, hence PG has more steals, assists, turnover and the number of three points made. Hence the directed graph for each position may not be the same. For example, the directed graph for the position C and PF may not have an edge between total points and three points made because those positions players usually make a very small number of three points made while the directed graph for other positions may have an edge between total points and three points made. We also plotted DAG models for different positions, but for ease of presentation we combined all positions.

We assumed each node conditional distribution given its parents is Poisson because most of NBA statistics we consider are the number of successes or attempts counted in the season. Hence we applied the ODS algorithm 1 for Poisson DAG models where ℓ_1-penalized likelihood regression is used in Steps 1) and 3). We used leave-one-out cross validation to choose the tuning parameters, and chose the largest value where mean squared error is within 1 standard error of the minimum mean squared of error because we prefer a sparse graph containing only legitimate edges.

Figure 9 (left) shows the directed graph estimated by our method. The estimated graph reveals clear causal/directional relationships between statistics. A large number of shootings attempted implies a large number of shootings made that implies large total points. Moreover, a large number of rebounds implies a large number of offensive rebounds, and a large number of fouls implies more frequent disqualifications. Lastly, the more total minutes played, the more number of games started, total points and other statistics.

We also find the two clusters related to positions; (i) C and PF related nodes (blocks, offensive rebounds, rebounds, personal fouls, technical fouls, and disqualification) (ii) PG related nodes (steals, assists, turnover, and the number of three points attempts and made). Within the clusters, the nodes are highly connected although there may be no causal or

Eliminated Edges 1	Assist → ThreesMade, Turnovers → FreeThrowsMade, Disqualification → GamesStarted, Technicals → Blocks, Steal → ThreesMade, Steal → TotalPoints,
Eliminated Edge 2	TotalPoints → ThreesMade
Added Edge	Steals → TotalMinutesPlayed

Table 2: The differences between the estimated DAGs in Figure 9.

directional relationships. It can be understood that position variable is a latent variables, and if the position variable is considered in the graph, some false directed edges may be eliminated. However, we do not add the position variable in the graph because Multinomial distribution does not belong to the class of QVF distribution.

There are many unexplainable edges in Figure 9 (left) due to the assumptions made which are not completely satisfied by the real data. In order to obtain a sparser graph with legitimate edges, we applied the ODS algorithm with the same procedures except that we chose a larger tuning parameter where mean squared of error is within 2.5 standard error of the minimum mean squared of error.

Figure 9 (right) shows the estimated directed graph using large tuning parameters. Compared to Figure 9 (left), the estimated DAG has fewer edges as expected. Specifically, the estimated DAG in Figure 9 (right) excludes unrealistic edges (Eliminated Edges 1 in Table. 2). However the estimated DAG also loses a legitimate edge (Eliminated Edge 2 in Table. 2) because C and PF have fewer number of three points made. Lastly, the estimated DAG includes explainable additional edge (Added Edge in Table. 2) because Step 1) of the ODS algorithm reduces the search-space of DAGs well, and improves the accuracy of the graph structure learning.

We acknowledge that our estimated DAG model makes many errors due to the restrictive assumption. However the benefit is best seen by comparing to other DAG learning approaches and an undirected graphical model. In particular, we applied Poisson undirected graphical models (Yang et al., 2013) which is the same procedure of Step 1) of our algorithm. The estimated undirected graph in Figure 10 (left) shows that a lot of nodes are connected by edges, and many edges are not explainable because the Poisson undirected graphical model only permits negative conditional relationships while all 18 variables are positively correlated. Hence it is not useful to understand the relationships between NBA statistics. We provide the estimated undirected graph with larger tuning parameter where mean squared of error is within 2.5 standard error of the minimum mean squared of error.

We also compare to the GES and MMHC algorithms. In particular, the estimated graphs in Figure 10 (right) are the same and both algorithms use the Bayesian Dirichlet score for count data which prefers a sparse graph when the positivity assumption is violated (i.e., $\widehat{P}(X_j = x_j \mid \mathrm{pa}(X_j)) \approx 0$). Since all statistics have high cardinality, which means each variable has almost no repeats in its data range, the positivity assumption is not satisfied. Hence the estimated directed graphs are extremely sparse which have a single directed edge between technical fouls and disqualification.

Since our method is the first identifiability result for the count data to the best of our knowledge, our method more reliably recovers the directional/causal relationships between

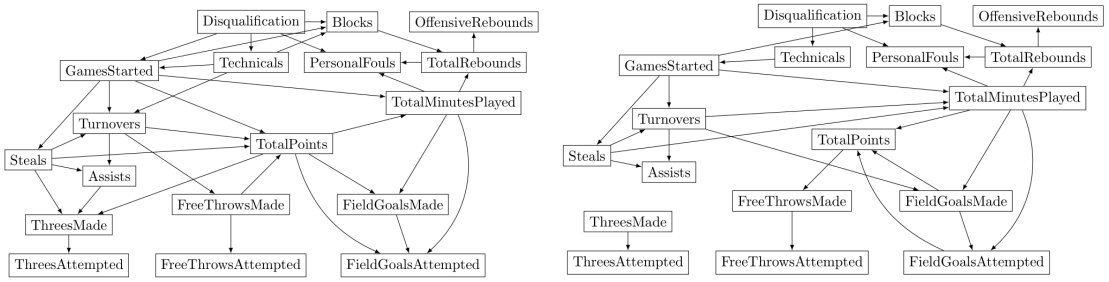


Figure 9: NBA players statistics directed graph estimated by the ODS algorithm for Poisson DAG models using ℓ_1-penalized likelihood regression in Steps 1) and 3) with small tuning parameters (left) and large tuning parameters (right).

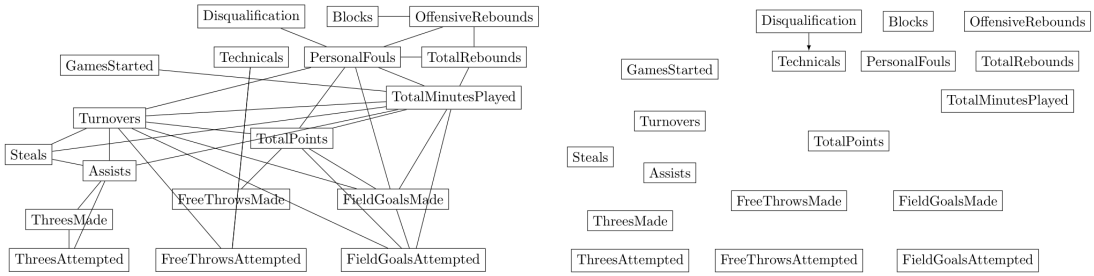


Figure 10: NBA players statistics undirected graph estimated by ℓ_1-penalized likelihood regression (left) and directed acyclic graph estimated by GES and MMHC algorithms (right).

NBA statistics. However we acknowledge that like most other DAG-learning approaches, very strong assumptions are required fore reliable recovery.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF DMS-1407028).

Appendix A. Proof for Theorem 5

Proof Without loss of generality, we assume the ordering is $\pi^* = (1, 2, \cdots, p)$. For notational convenience, we define $X_{1:j} = \{X_1, X_2, \cdots, X_j\}$ and $X_{1:0} = \emptyset$. For $m \in \{1, 2, ..., p-1\}$ and $j \in \{m, m+1, \cdots, p\}$, let $\omega_{jm} = (\beta_0 + \beta_1 \mathbb{E}(X_j \mid X_{1:m-1}))^{-1}$ and $\omega_{j1} = (\beta_0 + \beta_1 \mathbb{E}(X_j))^{-1}$. Recall that the overdispersion score of node j for m^{th} element of the ordering is Equation (5):

$$\mathcal{S}(m, j) = \omega_{jm}^2 \mathrm{Var}(X_j \mid X_{1:m-1}) - \omega_{jm} \mathbb{E}(X_j \mid X_{1:m-1}).$$

We now prove identifiability of our class of DAG models by induction. For the first element of the ordering ($m = 1$),

$$\begin{aligned}
\mathcal{S}(1, j) &= \omega_{j1}^2 \mathrm{Var}(X_j) - \omega_{j1} \mathbb{E}(X_j) \\
&\overset{(a)}{=} \omega_{j1}^2 \big\{ \mathrm{Var}(\mathbb{E}(X_j \mid X_{\mathrm{pa}(j)})) + \mathbb{E}(\mathrm{Var}(X_j \mid X_{\mathrm{pa}(j)})) - \omega_{j1}^{-1} \mathbb{E}(X_j) \big\} \\
&\overset{(b)}{=} \omega_{j1}^2 \big\{ \mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)})) + \mathbb{E}(\beta_0 \mathbb{E}(X_j \mid X_{\mathrm{pa}(j)}) + \beta_1 \mathbb{E}(X_j \mid X_{\mathrm{pa}(j)})^2) \\
&\qquad - (\beta_0 + \beta_1 \mathbb{E}(X_j)) \mathbb{E}(X_j) \big\} \\
&= \omega_{j1}^2 \big\{ \mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)})) + \beta_1 \mathbb{E}(\mathbb{E}(X_j \mid X_{\mathrm{pa}(j)})^2) - \beta_1 \mathbb{E}(X_j)^2 \big\} \\
&= \omega_{j1}^2 (1 + \beta_1) \mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)})).
\end{aligned}$$

(a) follows from the variance decomposition formula $\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathbb{E}(Y \mid X))$ for some random variables X and Y. In addition (b) follows from the quadratic variance property (2) of our class of distributions and the definition of ω_{j1}. Note that the score of the first element of the ordering is $\mathcal{S}(1, 1) = 0$ because $\mathrm{Var}(E(X_1)) = 0$, and other scores are strictly positive $\mathcal{S}(j, 1) > 0$ by the assumption $\beta_1 > -1$. Therefore 1 is the first element of the ordering.

For the $(m-1)^{th}$ element of the ordering, assume that the first $m-1$ elements of the ordering are correctly estimated. Now, we consider the m^{th} element of the ordering. Then, for $j \in \{m, m+1, \cdots, p\}$,

$$\begin{aligned}
\mathcal{S}(m, j) &= \omega_{jm}^2 \mathrm{Var}(X_j \mid X_{1:m-1}) - \omega_{jm} \mathbb{E}(X_j \mid X_{1:m-1}) \\
&\overset{(a)}{=} \omega_{jm}^2 \big\{ \mathrm{Var}(\mathbb{E}(X_j \mid X_{\mathrm{pa}(j)}) \mid X_{1:m-1}) + \mathbb{E}(\mathrm{Var}(X_j \mid X_{\mathrm{pa}(j)}) \mid X_{1:m-1}) - \omega_{jm}^{-1} \mathbb{E}(X_j \mid X_{1:m-1}) \big\} \\
&\overset{(b)}{=} \omega_{jm}^2 \big\{ \mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_{1:m-1}) + \mathbb{E}(\beta_0 \mathbb{E}(X_j \mid X_{\mathrm{pa}(j)} \mid X_{1:m-1}) \\
&\qquad + \beta_1 \mathbb{E}(X_j \mid X_{\mathrm{pa}(j)} \mid X_{1:m-1})^2) - (\beta_0 + \beta_1 \mathbb{E}(X_j \mid X_{1:m-1})) \mathbb{E}(X_j \mid X_{1:m-1}) \big\} \\
&= \omega_{jm}^2 \big\{ \mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_{1:m-1}) + \beta_1 \mathbb{E}(\mathbb{E}(X_j \mid X_{\mathrm{pa}(j)})^2 \mid X_{1:m-1}) - \beta_1 \mathbb{E}(X_j \mid X_{1:m-1})^2 \big\} \\
&= \omega_{jm}^2 (1 + \beta_1) \mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_{1:m-1}).
\end{aligned}$$

Again (a) follows from the variance decomposition formula, and (b) follows from the quadratic variance property (2) of our class of distributions and the definition of ω_{jm}. If $\mathrm{pa}(j) \backslash \{1, 2, \cdots, m-1\}$ is empty, $\mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_{1:m-1}) = 0$, and hence $\mathcal{S}(m, m) = 0$. On the other hand, for any node j in which $\mathrm{pa}(j) \backslash \{1, 2, \cdots, m-1\}$ is non-empty, $\mathcal{S}(m, j) > 0$ by the assumption $\beta_{j1} > -1$, which excludes it from being next in the ordering. Therefore, we can estimate a valid m^{th} component of the ordering, $\widehat{\pi}_m = m$. By induction this completes the proof. ■

Appendix B. Proof for Lemma 7

Proof We begin with part (a). By the construction $\theta^*_{D_j}$ in Equation (6), $[\theta^*_{D_j}]_k = 0$ for any node $k \notin \mathrm{pa}(j)$. Hence, it is sufficient to show that for any $k \in \mathrm{pa}(j)$, $[\theta^*_{D_j}]_k \neq 0$. Assume for the sake of contradiction that $[\theta^*_{D_j}]_k = 0$. Applying the first order optimality condition to Equation (6), we have

$$\begin{aligned} \mathbb{E}(X_j) &= \mathbb{E}(A'_j([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)}, X_{\mathrm{pa}(j)} \rangle)) \\ \mathbb{E}(X_j X_k) &= \mathbb{E}(A'_j([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)}, X_{\mathrm{pa}(j)} \rangle) X_k). \end{aligned} \tag{12}$$

By the definition of the covariance, we obtain

$$\begin{aligned} \mathbb{E}(X_j X_k) &= \mathrm{Cov}(A'([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)}, X_{\mathrm{pa}(j)} \rangle), X_k), \\ &\quad + \mathbb{E}(A'([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)}, X_{\mathrm{pa}(j)} \rangle)) \mathbb{E}(X_k). \end{aligned}$$

By Equation (12),

$$\mathbb{E}(X_j X_k) = \mathrm{Cov}(A'([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)}, X_{\mathrm{pa}(j)} \rangle), X_k) + \mathbb{E}(X_j)\mathbb{E}(X_k).$$

Therefore,

$$\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}(A'([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)}, X_{\mathrm{pa}(j)} \rangle), X_k).$$

By Assumption 6 (a), we have $[\theta^*_{D_j}]_k = 0$, and

$$\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}(D'([\theta^*_{D_j}]_j + \langle [\theta^*_{D_j}]_{\mathrm{pa}(j)\setminus k}, X_{\mathrm{pa}(j)\setminus j} \rangle), X_k),$$

which is a contradiction by our earlier assumption. Therefore $[\theta^*_{D_j}]_k \neq 0$. Furthermore since $k \in \mathrm{pa}(j)$ is arbitrary, the proof is complete. The proof for part (b) follows exactly the same line of reasoning. ■

Appendix C. Proof for Theorem 12

In this section, we provide the proof for Theorem 12 using the *primal-dual witness method* that also used many works (see e.g., Yang et al. 2012; Meinshausen and Bühlmann 2006; Wainwright et al. 2006; Ravikumar et al. 2011). We begin by introducing propositions to control the tail behavior for the distribution of each node:

Proposition 17 *Define*

$$\xi_1 := \{\max_{j \in V} \max_{i \in \{1, \cdots, n\}} |X_j^{(i)}| < 4 \log(\eta)\}.$$

Under Assumption 10, $P(\xi_1^c) \leq M \cdot \eta^{-2}$.

Proposition 18 *Suppose that X is a random vector according to the DAG model* (1)*, and Assumption 10 is satisfied. Then, for any vector $u \in \mathbb{R}^p$ such that $\|u\|_1 \leq c'$, for any positive constant δ,*

$$P(|\langle u, X \rangle| \geq \delta \log \eta) \leq M \cdot p \cdot \eta^{-\delta/c'}. \tag{13}$$

Using these concentration results, we show that ℓ_1-penalized regression recovers the neighborhood for a fixed node $j \in V$ with high probability. For ease of notation, we define a new parameter $\theta \in \mathbb{R}^{p-1}$ without the node j since the node j is not penalized in regression problem (9). Then, the conditional negative log-likelihood of the GLM (8) is:

$$\ell_j(\theta; X^{1:n}) := \frac{1}{n}\sum_{i=1}^{n}\left(-X_j^{(i)}\langle\theta, X_{V\setminus j}^{(i)}\rangle + A_j(\langle\theta, X_{V\setminus j}^{(i)}\rangle)\right).$$

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

$$\widehat{\theta}_{M_j} := \arg\min_{\theta\in\mathbb{R}^{p-1}} \mathcal{L}_j(\theta, \lambda_n) = \arg\min_{\theta\in\mathbb{R}^{p-1}} \{\ell_j(\theta; X^{1:n}) + \lambda_n\|\theta\|_1\}. \tag{14}$$

By setting the *sub-differential* to 0, $\widehat{\theta}_{M_j}$ must satisfy the following condition:

$$\bigtriangledown_\theta \mathcal{L}_j(\widehat{\theta}_{M_j}, \lambda_n) = \bigtriangledown_\theta \ell_j(\widehat{\theta}_{M_j}; X^{1:n}) + \lambda_n \widehat{Z} = 0 \tag{15}$$

where $\widehat{Z} \in \mathbb{R}^{p-1}$ and $\widehat{Z}_t = \text{sign}([\widehat{\theta}_{M_j}]_t)$ if $t \in \mathcal{N}(j)$, otherwise $|\widehat{Z}_t| < 1$.

The following Lemma 19 directly follows from prior works in Ravikumar et al. (2010) and Yang et al. (2012) where each node conditional distribution is in the form of a generalized linear model. For notational convenience, let $S = \mathcal{N}(j)$.

Lemma 19 *Suppose that $|\widehat{Z}_t| < 1$ for $t \notin S$. Then, the solution $\widehat{\theta}_{M_j}$ of (14) satisfies $[\widehat{\theta}_{M_j}]_t = 0$ for $t \notin S$. Furthermore, if the sub-matrix of the Hessian matrix $Q_{SS}^{M_j}$ is invertible, then $\widehat{\theta}_{M_j}$ is unique.*

The remainder of the proof is to show $|\widetilde{Z}_t| < 1$ for all $t \notin S$. Note that the restricted solution in Equation (19) is $(\widetilde{\theta}_{M_j}, \widetilde{Z})$. Equation (15) with the dual solution can be represented by

$$\bigtriangledown^2 \ell_j(\theta^*_{M_j}; X^{1:n})(\widetilde{\theta}_{M_j} - \theta^*_{M_j}) = -\lambda_n \widetilde{Z} - W_j^n + R_j^n$$

where:

(a) W_j^n is the sample score function.

$$W_j^n := -\bigtriangledown \ell_j(\theta^*_{M_j}; X^{1:n}). \tag{16}$$

(b) $R_j^n = (R_{j1}^n, R_{j2}^n, \cdots, R_{jp-1}^n)$ and R_{jk}^n is the remainder term by applying the coordinate-wise mean value theorem.

$$R_{jk}^n := [\bigtriangledown^2 \ell_j(\theta^*_{M_j}; X^{1:n}) - \bigtriangledown^2 \ell_j(\bar{\theta}^{(k)}_{M_j}; X^{1:n})]_k^T (\widetilde{\theta}^{(k)}_{M_j} - \theta^*_{M_j}). \tag{17}$$

Here $\bar{\theta}^{(k)}_{M_j}$ is a vector on the line between $\widetilde{\theta}$ and $\theta^*_{M_j}$ and $[\cdot]_k^T$ is the k^{th} row of a matrix.

Then, the following proposition provides a sufficient condition to control $\widetilde{Z}$.

Proposition 20 *Suppose that* $\max(\|W_j^n\|_\infty, \|R_j^n\|_\infty) \leq \frac{\lambda_n \alpha}{4(2-\alpha)}$. *Then $|\widetilde{Z}_t| < 1$ for all $t \notin S$.*

Next we introduce the following three lemmas to show that conditions in Proposition 20 hold. For ease of notation, let $\eta = \max\{n, p\}$ and $\widetilde{\theta}_S = [\widetilde{\theta}_{M_j}]_S$ and $\widetilde{\theta}_{S^c} = [\widetilde{\theta}_{M_j}]_{S^c}$. Suppose that Assumptions 8, 9, 10, and 11 are satisfied.

Lemma 21 *Suppose that* $\lambda_n \geq \frac{16\max\{n^{\kappa_2}\log\eta, \log^2\eta)\}}{n^a}$ *for some* $a \in \mathbb{R}$. *Then,*

$$P\left(\frac{\|W_j^n\|_\infty}{\lambda_n} \leq \frac{\alpha}{4(2-\alpha)}\right) \geq 1 - 2d \cdot exp(-\frac{\alpha^2}{8(2-\alpha)^2} \cdot n^{1-2a}) - M \cdot \eta^{-2}.$$

Lemma 22 *Suppose that* $\|W_j^n\|_\infty \leq \frac{\lambda_n}{4}$. *For* $\lambda_n \leq \frac{1}{40}\frac{\rho_{\min}^2}{\rho_{\max}}\frac{1}{n^{\kappa_2} d\log\eta}$,

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

Lemma 23 *Suppose that* $\|W_j^n\|_\infty \leq \frac{\lambda_n}{4}$. *For* $\lambda_n \leq \frac{\alpha}{400(2-\alpha)}\frac{\rho_{\min}^2}{\rho_{\max}}\frac{1}{n^{\kappa_2} d\log\eta}$,

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

The rest of the proof is straightforward using Lemmas 21, 22, and 23. Consider the choice of regularization parameter $\lambda_n = \frac{16\max\{n^{\kappa_2}\log\eta, \log^2\eta\}}{n^a}$ for a constant $a \in (2\kappa_2, 1/2)$ where κ_2 is determined by Assumption 11. Then, the condition for Lemma 21 is satisfied, and therefore $\|W_n\|_\infty \leq \frac{\lambda_n}{4}$. Moreover, the conditions for Lemmas 22 and 23 are satisfied for $n \geq C'\max\{(d\log^2\eta)^{\frac{1}{a-2\kappa_2}}, (d\log^3\eta)^{\frac{1}{a-\kappa_2}}\}$ for some positive constant C'. Then,

$$\|\widetilde{Z}_{S^c}\|_\infty \leq (1-\alpha) + (2-\alpha)\left[\frac{\|W_j^n\|_\infty}{\lambda_n} + \frac{\|R_j^n\|_\infty}{\lambda_n}\right] \leq (1-\alpha) + \frac{\alpha}{4} + \frac{\alpha}{4} < 1, \qquad (18)$$

with probability of at least $1 - C_1 d\exp(-C_2 n^{1-2a}) - C_3\eta^{-2}$ for positive constants C_1, C_2 and C_3.

To prove sign consistency, it is sufficient to show that $\|\widehat{\theta}_{M_j} - \theta_{M_j}^*\|_\infty \leq \frac{\|\theta_{M_j}^*\|_{\min}}{2}$. By Lemma 22, we have $\|\widehat{\theta}_{M_j} - \theta_{M_j}^*\|_\infty \leq \|\widehat{\theta}_{M_j} - \theta_{M_j}^*\|_2 \leq \frac{5}{\lambda_{\min}}\sqrt{d}\ \lambda_n \leq \frac{\|\theta_{M_j}^*\|_{\min}}{2}$ as long as $\|\theta_{M_j}^*\|_{\min} \geq \frac{10}{\lambda_{\min}}\sqrt{d}\ \lambda_n$.

Lemma 7(b) guarantees that ℓ_1-penalized likelihood regression recovers the true neighborhood for each node with high probability. Because we have p likelihood regression problems, if $n \geq C'(d\log^2\eta)^{\frac{1}{a-2\kappa_2}})$, it follows that:

$$P(\widehat{G^m} = G^m) \geq 1 - C_1 d \cdot p \cdot \exp(-C_2 n^{1-2a}) - C_3\eta^{-1}.$$

C.1 Proof for Proposition 17

Proof Applying the union bound and the Chernoff bound,

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

By Assumption 10, we obtain $\max_{i,j}\mathbb{E}(\exp(|X_j|^{(i)})) < M$, which completes the proof. ■

C.2 Proof for Proposition 18

Proof We exploit Hölder's inequality $\langle u, X\rangle \leq \|u\|_1 \max_{j\in V}|X_j|$. Therefore, we have

$$P(|\langle u, X\rangle)| \geq \delta \log \eta) \leq P(\max_{j\in V}|X_j| \geq \frac{\delta}{\|u\|_1}\log\eta).$$

Using the union bound, we have

$$P(\max_{j\in V}|X_j| \geq \frac{\delta}{\|u\|_1}\log\eta) \leq p\cdot \max_{j\in V} P(|X_j| \geq \frac{\delta}{\|u\|_1}\log\eta).$$

Applying the Chernoff bounding technique and Assumption 10 $\max_j \mathbb{E}(\exp(|X_j|) < M$, we obtain

$$p\cdot \max_{j\in V} P(|X_j| \geq \frac{\delta}{\|u\|_1}\log\eta) \leq M\cdot p\cdot \eta^{-\frac{\delta}{\|u\|_1}}.$$

By the assumption $\|u\|_1 \leq c'$, we compete the proof. ■

C.3 Proof for Proposition 20

Proof Since $\widetilde{\theta}_{S^c} = (0, 0, ..., 0) \in \mathbb{R}^{|S^c|}$ in our primal-dual construction, we can re-state condition (15) in block form as follows. For notational simplicity, $Q := Q^{M_j}$.

$$\begin{aligned} Q_{S^cS}[\widetilde{\theta}_S - \theta_S] &= W^n_{S^c} - \lambda_n \widetilde{Z}_{S^c} + R^n_{S^c},. \\ Q_{SS}[\widetilde{\theta}_S - \theta^*_S] &= W^n_S - \lambda_n \widetilde{Z}_S + R^n_S, \end{aligned}$$

where W^n_S and R^n_S are sub-vectors of W^n_j and R^n_j indexed by S, respectively.

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

$$Q_{S^cS}Q^{-1}_{SS}[W^n_S - \lambda_n\widetilde{Z}_S - R^n_S] = W^n_{S^c} - \lambda_n\widetilde{Z}_{S^c} - R^n_{S^c}.$$

Therefore

$$[W^n_{S^c} - R^n_{S^c}] - Q_{S^cS}Q^{-1}_{SS}[W^n_S - R^n_S] + \lambda_n Q_{S^cS}Q^{-1}_{SS}\widetilde{Z}_S = \lambda_n \widetilde{Z}_{S^c}.$$

Taking the ℓ_∞ norm of both sides yields

$$\|\widetilde{Z}_{S^c}\|_\infty \leq |||Q_{S^cS}Q^{-1}_{SS}|||_\infty \left[\frac{\|W^n_S\|_\infty}{\lambda_n} + \frac{\|R^n_S\|_\infty}{\lambda_n} + 1\right] + \frac{\|W^n_{S^c}\|_\infty}{\lambda_n} + \frac{\|R^n_{S^c}\|_\infty}{\lambda_n}.$$

Recalling Assumption (9), we obtain $|||Q_{S^cS}Q^{-1}_{SS}|||_\infty \leq (1-\alpha)$, hence we have

$$\begin{aligned} \|\widetilde{Z}_{S^c}\|_\infty &\leq (1-\alpha)\left[\frac{\|W^n_S\|_\infty}{\lambda_n} + \frac{\|R^n_S\|_\infty}{\lambda_n} + 1\right] + \frac{\|W^n_{S^c}\|_\infty}{\lambda_n} + \frac{\|R^n_{S^c}\|_\infty}{\lambda_n} \\ &\leq (1-\alpha) + (2-\alpha)\left[\frac{\|W^n_j\|_\infty}{\lambda_n} + \frac{\|R^n\|_\infty}{\lambda_n}\right]. \end{aligned}$$

If $\|W^n_j\|_\infty$ and $\|R^n_j\|_\infty \leq \frac{\lambda_n\alpha}{4(2-\alpha)}$ as assumed,

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

■

C.4 Proof for Lemma 19

Proof The main idea of the proof is the *primal-dual-witness* method which asserts that there is a solution to the dual problem $\widetilde{\theta}_{M_j} = \widehat{\theta}_{M_j}$ if the following KKT conditions are satisfied:

(a) We define $\widetilde{\theta}_{M_j} \in \Theta_{M_j}$ where $\Theta_{M_j} = \{\theta \in \mathbb{R}^{p-1} : \theta_{S^c} = 0\}$ as the solution to the following optimization problem.

$$\widetilde{\theta}_{M_j} := \arg\min_{\theta \in \Theta_{M_j}} \mathcal{L}(\theta, \lambda_n) = \arg\min_{\theta \in \Theta_{M_j}} \{\ell_j(\theta; X^{1:n}) + \lambda_n \|\theta\|_1\}. \tag{19}$$

(b) Define $\widetilde{Z}$ to be a sub-differential for the regularizer $\|\cdot\|_1$ evaluated at $\widetilde{\theta}_{M_j}$. For any $t \in S$, $\widetilde{Z}_t = \text{sign}([\widetilde{\theta}_{M_j}]_t)$.

(c) For any $t \notin S$, $|\widetilde{Z}_t| < 1$.

If conditions (a), (b), and (c) are satisfied, $\widehat{\theta}_{M_j} = \widetilde{\theta}_{M_j}$, meaning that the solution of the unrestricted problem (14) is the same as the solution of the restricted problem (19). Conditions (a), (b) and (c) suffice to obtain a pair $(\widetilde{\theta}_{M_j}, \widetilde{Z})$ that satisfies the optimality condition (15), but do not guarantee that $\widetilde{Z}$ is an element of the sub-differential $\|\widetilde{\theta}_{M_j}\|_1$ (see details in Ravikumar et al. 2010, 2011). Since the sub-matrix of the Hessian $Q^{M_j}_{SS}$ is invertible, the restricted problem (19) is strictly convex, $\widetilde{\theta}_{M_j}$ is unique. ■

C.5 Proof for Lemma 21

Proof Each entry of the sample score function W^n_j in Equation (16) has the form $W^n_{jt} = \frac{1}{n}\sum_{i=1}^n W^{(i)}_{jt}$ for any $t \in S$. In addition, $W^n_{jt} = 0$ for all $t \notin S$ since $[\theta^*_{M_j}]_t = 0$ by the construction of $\theta^*_{M_j}$ in Equation (7). For any $t \in S$ and $i \in \{1, 2, \cdots, n\}$, $W^{(i)}_{jt} = X^{(i)}_t X^{(i)}_j - A'_j(\langle \theta^*_S, X^{(i)}_S \rangle) X^{(i)}_t$ are independent and have mean 0.

Now, we show that $(|W^{(i)}_{jt}|)_{i=1}^n$ are bounded with high probability given the following event ξ_1 using Hoeffding's inequality. Event ξ_1 is defined as follows:

$$\xi_1 \quad := \quad \left\{ \max_{j \in V} \max_{i \in \{1, \cdots, n\}} |X^{(i)}_j| < 4 \log \eta \right\}.$$

Conditioning on ξ_1, it follows that $\langle \theta^*_S, X^{(i)}_S \rangle < 4 \log(\eta) \cdot \|\theta^*_S\|_1$, Assumption 11 is satisfied. Hence $\max_i |A'_j(\langle \theta^*_S, X^{(i)}_S \rangle)| \leq n^{\kappa_2}$. Furthermore given ξ_1, $\max_i X^{(i)}_t X^{(i)}_j < 16 \log^2 \eta$. Therefore there exists a constant $C_{\max}(\eta, \kappa_2) := 16 \max\{n^{\kappa_2} \log \eta, \log^2 \eta\}$ such that $\max_{i,j,t} |W^{(i)}_{jt}| \leq C_{\max}(\eta, \kappa_2)$.

Recall that d is the maximum degree of the moralized graph, therefore $|S| \leq d$. Applying the union bound,

$$P(\|W^n_j\|_\infty > \delta, \xi_1) \leq d \cdot \max_{t \in S} P(|W^n_{jt}| > \delta, \xi_1).$$

Using Hoeffding's inequality,

$$d \cdot \max_{t \in S} P(|W_{jt}^n| > \delta, \xi_1) \leq 2d \cdot \exp(-\frac{2n\delta^2}{C_{\max}(\eta, \kappa_2)^2}).$$

Suppose that $\delta = \frac{\lambda_n \alpha}{4(2-\alpha)}$ and $\lambda_n \geq \frac{C_{\max}(\eta,\kappa_2)}{n^a}$ for some $a \in [0, 1/2)$. Then

$$\begin{aligned} P(\frac{\|W_j^n\|_\infty}{\lambda_n} > \frac{\alpha}{4(2-\alpha)}, \xi_1) &\leq 2d \cdot \exp\Big(- \frac{\alpha^2}{8(2-\alpha)^2} \frac{n\lambda_n^2}{C_{\max}(\eta, \kappa_2)^2} \Big) \\ &\leq 2d \cdot \exp\Big(- \frac{\alpha^2}{8(2-\alpha)^2} n^{1-2a} \Big). \end{aligned} \tag{20}$$

Since $P(A) = P(A \cap B) + P(A \cap B^c) \leq P(A \cap B) + P(B^c)$,

$$P(\frac{\|W_j^n\|_\infty}{\lambda_n} > \frac{\alpha}{4(2-\alpha)}) \quad \leq \quad P(\frac{\|W_j^n\|_\infty}{\lambda_n} > \frac{\alpha}{4(2-\alpha)}, \xi_1) + P(\xi_1^c).$$

Then, the probability bound in Equation (20) and Proposition 17 $P(\xi_1^c) \leq M \cdot \eta^{-2}$ directly implies that

$$P(\frac{\|W_j^n\|_\infty}{\lambda_n} > \frac{\alpha}{4(2-\alpha)}) \leq 2d \cdot \exp\Big(- \frac{\alpha^2}{8(2-\alpha)^2} n^{1-2a} \Big) + M \cdot \eta^{-2}.$$

■

C.6 Proof for Lemma 22

Proof In order to establish the error bound $\|\widetilde{\theta}_S - \theta_S^*\| \leq B$ for some radius B, several works (Yang et al., 2012; Ravikumar et al., 2010, 2011) already proved that it suffices to show $F(u_S) > 0$ for all $u_S := \widetilde{\theta}_S - \theta_S^*$ such that $\|u_S\|_2 = B$ where

$$F(a) := \ell_j(\theta_S^* + a; X^{1:n}) - \ell_j(\theta_S^*; X^{1:n}) + \lambda_n(\|\theta_S^* + a\|_1 - \|\theta_S^*\|_1). \tag{21}$$

More specifically, since $u_S = \widetilde{\theta}_S - \theta_S^*$ is the minimizer of F and $F(0) = 0$ by the construction of Equation (21), $F(u_S) \leq 0$. 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 the convex combination $v \cdot u_S + (1-v) \cdot 0$ would lie on the boundary of the ball, for an appropriately chosen $v \in (0, 1)$. By convexity,

$$F(v \cdot u_S + (1-v) \cdot 0) \leq v \cdot F(u_S) + (1-v) \cdot 0 \leq 0 \tag{22}$$

contradicting the assumed strict positivity of F on the boundary.

Thus it suffices to establish strict positivity of F on the boundary of the ball with radius $B := M_1 \lambda_n \sqrt{d}$ where $M_1 > 0$ 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 (21),

$$F(u_S) = (W_S^n)^T u_S + u_S^T [\bigtriangledown^2 \ell_j(\theta_M^* + v u_S; x)] u_S + \lambda_n(\|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1), \tag{23}$$

for some $v \in [0,1]$. Since $\|W_S^n\|_\infty \leq \frac{\lambda_n}{4}$ by assumption and $\|u_S\|_1 \leq \sqrt{d}\|u_S\|_2 \leq \sqrt{d} \cdot B$, the first term in Equation (23) has the following bound:

$$|(W_S^n)^T u_S| \leq \|W_S^n\|_\infty \|u_S\|_1 \leq \|W_S^n\|_\infty \sqrt{d}\|u_S\|_2 \leq (\lambda_n\sqrt{d})^2 \frac{M_1}{4}.$$

Applying the triangle inequality to the last part of Equation (23), we have the following bound.

$$\lambda_n(\|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1) \geq -\lambda_n\|u_S\|_1 \geq -\lambda_n\sqrt{d}\|u_S\|_2 = -M_1(\lambda_n\sqrt{d})^2.$$

Next we bound $\lambda_{\min}\left(\bigtriangledown^2 \ell_j(\theta_S^* + vu_S)\right)$ where $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of a matrix:

$$\begin{aligned}
q^* &:= \lambda_{\min}\left(\bigtriangledown^2 \ell_j(\theta_S^* + vu_S)\right) \\
&\geq \min_{v\in[0,1]} \lambda_{\min}\left(\bigtriangledown^2 \ell_j(\theta_S^* + vu_S)\right) \\
&\geq \lambda_{\min}\left(\bigtriangledown^2 \ell_j(\theta_S^*)\right) - \max_{v\in[0,1]} \|\frac{1}{n}\sum_{i=1}^n A_j'''(\langle \theta_S^* + vu_S, X_S\rangle) u_S^T X_S^{(i)} X_S^{(i)} (X_S^{(i)})^T\|_2 \\
&\geq \rho_{\min} - \max_{v\in[0,1]} \max_{y:\|y\|_2=1} \frac{1}{n}\sum_{i=1}^n |A_j'''(\langle \theta_S^* + vu_S, X_S\rangle)| \cdot |u_S^T X_S^{(i)}| \cdot (y^T X_S^{(i)})^2. \qquad (24)
\end{aligned}$$

Next we define the event ξ_2 in order to bound $A_j'''(\langle \theta_S^* + vu_S, X_S\rangle)$.

$$\xi_2 := \{\max_{i\in\{1,\cdots,n\}} \langle \theta_S^* + vu_S, X_S^{(i)}\rangle < \kappa_1 \log \eta\}.$$

On ξ_2, Assumption 11 is satisfied and

$$A_j'''(\langle \theta_S^* + vu_S, X_S\rangle) \leq n^{\kappa_2}. \qquad (25)$$

In addition, we bound the second term in Equation (24). Recall that $\|X_S^{(i)}\|_\infty \leq 4\log\eta$ for all $i \in \{1, 2, \cdots, n\}$ on ξ_1. Since $\|u_S\|_1 \leq \sqrt{d}\|u_S\|_2 \leq \sqrt{d} \cdot B$,

$$|u_S^T X_S^{(i)}| \leq 4\log(\eta)\sqrt{d}\|u_S\|_2 \leq 4\log(\eta) \cdot M_1\lambda_n d. \qquad (26)$$

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 8. Together with the bounds of Equations (25) and (26) on the events ξ_1 and ξ_2,

$$q^* \leq \rho_{\min} - 4n^{\kappa_2}\log(\eta) \cdot M_1\lambda_n d\ \rho_{\max}.$$

For $\lambda_n \leq \frac{\rho_{\min}}{8n^{\kappa_2}\log(\eta)M_1 d\rho_{\max}}$, we have $q^* \leq \frac{\rho_{\min}}{2}$. Therefore,

$$F(u) \geq (\lambda_n\sqrt{n})^2\Big\{-\frac{1}{4}M_1 + \frac{\rho_{\min}}{2}M_1^2 - M_1\Big\},$$

which is strictly positive for $M_1 = \frac{5}{\rho_{\min}}$. Therefore for $\lambda_n \leq \frac{\rho_{\min}^2}{40n^{\kappa_2}\log(\eta)d\rho_{\max}}$ given ξ_1 and ξ_2,

$$\|\widetilde{\theta}_S - \theta_S^*\|_2 \leq \frac{5}{\rho_{\min}}\sqrt{d}\lambda_n.$$

Since $P(A) = P(A \cap B \cap C) + P(A \cap (B \cap C)^c) \leq P(A \cap B \cap C) + P(B^c) + P(C^c)$,

$$P\left(\|\widetilde{\theta}_S - \theta^*_S\|_2 > \frac{5}{\rho_{\min}}\sqrt{d}\lambda_n\right) \leq P\left(\|\widetilde{\theta}_S - \theta^*_S\|_2 > \frac{5}{\rho_{\min}}\sqrt{d}\lambda_n, \xi_1, \xi_2\right) + P(\xi_1^c) + P(\xi_2^c).$$

Here the probability of ξ_2^c is upped bounded as follows.

$$\begin{aligned} P(\xi_2^c) &\overset{(a)}{\leq} n \max_i P(\langle \theta^*_{M_j} + v u_S, X_S^{(i)} \rangle > \kappa_1 \log \eta) \\ &\overset{(b)}{\leq} n \cdot M \cdot \eta^{-\frac{\kappa_1}{2\|\theta^*_{M_j}\|_1}} \\ &\overset{(c)}{\leq} M \cdot \eta^{-2}. \end{aligned}$$

(a) follows from the union bound, and (b) follows from Proposition 18, and $\|u_S\|_1 \leq \sqrt{d}\|u_S\|_2 \leq dM_1\lambda_n \leq \|\theta^*_{M_j}\|_1$ and $\min_{j\in V}\min_{t\in S}|[\theta^*_M]_t| \geq \frac{10}{\rho_{\min}}\sqrt{d}\lambda_n$. Lastly (c) follows from Assumption 11 that $\kappa_1 \geq 6\|\theta^*_{M_j}\|_1$.

In addition the probability bound of ξ_1^c is provided in Proposition 17. Therefore

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

■

C.7 Proof for Lemma 23

Proof According to Equation (17), R^n_{jt} for any $t \in S$ can be expressed as

$$\begin{aligned} R^n_{jt} &= \frac{1}{n}\sum_{i=1}^n [\bigtriangledown^2 \ell_j(\theta^*_{M_j}; X^{1:n}) - \bigtriangledown^2 \ell_j(\bar{\theta}^{(t)}_{M_j}; X^{1:n})]^T_t (\widetilde{\theta} - \theta^*_{M_j}) \\ &= \frac{1}{n}\sum_{i=1}^n [A''_j(\langle \theta^*_S, X^{(i)}_{V\setminus j}\rangle) - A''_j(\langle \bar{\theta}^{(t)}_{M_j}, X^{(i)}_{V\setminus j}\rangle)][X^{(i)}_{V\setminus j}(X^{(i)}_{V\setminus j})^T]^T_t (\widetilde{\theta} - \theta^*_{M_j}) \end{aligned}$$

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

By the mean value theorem,

$$R^t_{jt} = \frac{1}{n}\sum_{i=1}^n \left\{A'''_j(\langle \bar{\bar{\theta}}^{(t)}_{M_j}, X^{(i)}_{V\setminus j}\rangle) X^{(i)}_t\right\}\left\{v(\widetilde{\theta}_{M_j} - \theta^*_{M_j})^T X^{(i)}_{V\setminus j}(X^{(i)}_{V\setminus j})^T(\widetilde{\theta}_{M_j} - \theta^*_{M_j})\right\}$$

for $\bar{\bar{\theta}}^{(t)}_{M_j}$ which is a point on the line between $\bar{\theta}^{(t)}_{M_j}$ and $\theta^*_{M_j}$.

By Proposition 17, $\max_{i,j}|X^{(i)}_j| \leq 4\log\eta$ given ξ_1. Furthermore in Section C.6, we showed that $A'''_j(\langle \bar{\bar{\theta}}^{(t)}_{M_j}, X_{M\setminus j}\rangle) \leq n^{\kappa_2}$ given ξ_2 . Therefore, on ξ_1 and ξ_2, it follows that:

$$|R^n_{jt}| \leq 4n^{\kappa_2}\log(\eta)\rho_{\max}\|\widetilde{\theta} - \theta^*_M\|_2^2.$$

We showed that $\|\widetilde{\theta} - \theta^*_M\|_2 \leq \frac{5}{\rho_{\min}}\sqrt{d}\lambda_n$ for $\lambda_n \leq \frac{\alpha}{400(2-\alpha)}\frac{\rho^2_{\min}}{\rho_{\max}}\frac{1}{dn^{\kappa_2}\log(\eta)}$ given ξ_1 and ξ_2 in the proof of Lemma 22. Therefore we obtain

$$\|R^n\|_\infty \leq \frac{100\rho_{\max}}{\rho^2_{\min}} d\ n^{\kappa_2}\log(\eta)\ \lambda_n^2 \leq \frac{\alpha\lambda_n}{4(2-\alpha)}.$$

Since $P(A) = P(A \cap B \cap C) + P(A \cap (B \cap C)^c) \leq P(A \cap B \cap C) + P(B^c) + P(C^c)$,

$$P\left(\|R^n\|_\infty > \frac{\alpha\lambda_n}{4(2-\alpha)}\right) \leq P\left(\|R^n\|_\infty > \frac{\alpha\lambda_n}{4(2-\alpha)}, \xi_1, \xi_2\right) + P(\xi_1^c) + P(\xi_2^c).$$

Putting the probability bounds for ξ_1^c and ξ_2^c specified in Proposition 17 and Section C.6 together, we have

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

■

Appendix D. Proof for Theorem 14

Proof Without loss of generality, assume that the true ordering is $\pi^* = (1, 2, \cdots, p)$. Let $T_j(X_j) := \omega_j X_j$ where $\omega_j = (\beta_0 + \beta_1 \mathbb{E}(X_j \mid X_{\text{pa}(j)}))^{-1}$ (specified in Proposition 3). For any node $j \in V$ and $S \subset V \setminus \{j\}$, let $\mu_{j|S}$ and $\sigma^2_{j|S}$ represent $\mathbb{E}(T_j(X_j) \mid X_S)$ and $\text{Var}(T_j(X_j) \mid X_S)$ respectively. For realizations x_S, let $\mu_{j|S}(x_S)$ and $\sigma^2_{j|S}(x_S)$ denote $\mathbb{E}(T_j(X_j) \mid X_S = x_S)$ and $\text{Var}(T_j(X_j) \mid X_S = x_S)$, respectively. Let $n(x_S) = \sum_{i=1}^n \mathbf{1}(X_S^{(i)} = x_S)$ denote the total conditional sample size, and $n_S = \sum_{x_S} n(x_S)\mathbf{1}(n(x_S) \geq c_0 \cdot n)$ for an arbitrary $c_0 \in (0, 1)$ to denote the truncated conditional sample size.

Let E^m denote the set of undirected edges corresponding to the *moralized* graph. Recall the definitions $\mathcal{N}(j) = \{k \in V : (j, k) \text{ or } (k, j) \in E^m\}$ denote the neighborhood set of node j in the moralized graph, $C_{jk} = \mathcal{N}(k) \cap \{\pi_1, \pi_2, \cdots, \pi_{j-1}\}$. Since we assume the structure of the moralized graph is provided, $\widehat{C}_{jk} = C_{jk}$. Hence C_{jk} is used instead of an estimated set $\widehat{C}_{jk}$.

The overdispersion score of node $k \in V \setminus \{\pi_1, ..., \pi_{j-1}\}$ for the j^{th} component of the ordering π_j only depends on $\mathcal{X}(C_{jk}) = \{x \in \{X^{(1)}_{C_{jk}}, X^{(2)}_{C_{jk}}, \cdots, X^{(n)}_{C_{jk}}\} : n(x) \geq c_0 \cdot n\}$, so we only count up elements that occur sufficiently frequently.

According to the generalized ODS algorithm, the truncated sample conditional mean and variance of $T_j(X_j)$ given $X_S = x_S$ are:

$$\begin{aligned}
\widehat{\mu}_{j|S}(x_S) &:= \frac{1}{n_S(x_S)} \sum_{i=1}^n T_j(X_j^{(i)})\mathbf{1}(X_S^{(i)} = x_S), \\
\widehat{\sigma}^2_{j|S}(x_S) &:= \frac{1}{n_S(x_S) - 1} \sum_{i=1}^n (T_j(X_j^{(i)}) - \widehat{\mu}_{j|S}(x_S))^2\mathbf{1}(X_S^{(i)} = x_S).
\end{aligned}$$

We rewrite the overdispersion score (5) of node $k \in V \setminus \{\pi_1, ..., \pi_{j-1}\}$ for π_j as follows:

$$\begin{aligned}
\widehat{\mathcal{S}}(1,k) &:= \left[\left(\frac{\widehat{\sigma}_k}{\beta_0 + \beta_1\widehat{\mu}_k}\right)^2 - \frac{\widehat{\mu}_k}{\beta_0 + \beta_1\widehat{\mu}_k}\right], \\
\widehat{\mathcal{S}}(m,k) &:= \sum_{x \in \mathcal{X}(C_{mk})} \frac{n(x)}{n_{C_{mk}}} \left[\left(\frac{\widehat{\sigma}_{k|C_{mk}}(x)}{\beta_0 + \beta_1\widehat{\mu}_{k|C_{mk}}(x)}\right)^2 - \frac{\widehat{\mu}_{k|C_{mk}}(x)}{\beta_0 + \beta_1\widehat{\mu}_{k|C_{mk}}(x)}\right].
\end{aligned}$$

For notational convenience, let each entry of the overdispersion score $\widehat{\mathcal{S}}(m,k)$ for $x \in \mathcal{X}(C_{mk})$ be defined as:

$$\widehat{\mathcal{S}}(m,k)(x) := \left(\frac{\widehat{\sigma}_{k|C_{mk}}(x)}{\beta_0 + \beta_1\widehat{\mu}_{k|C_{mk}}(x)}\right)^2 - \frac{\widehat{\mu}_{k|C_{mk}}(x)}{\beta_0 + \beta_1\widehat{\mu}_{k|C_{mk}}(x)}. \tag{27}$$

The true overdispersion scores are:

$$\begin{aligned}
\mathcal{S}^*(1,k) &:= \left[\left(\frac{\sigma_k}{\beta_0 + \beta_1\mu_k}\right)^2 - \frac{\mu_k}{\beta_0 + \beta_1\mu_k}\right], \\
\mathcal{S}^*(m,k) &:= \sum_{x \in \mathcal{X}(C_{mk})} \frac{n(x)}{n_{C_{mk}}} \left[\left(\frac{\sigma_{k|C_{mk}}(x)}{\beta_0 + \beta_1\mu_{k|C_{mk}}(x)}\right)^2 - \frac{\mu_{k|C_{mk}}(x)}{\beta_0 + \beta_1\mu_{k|C_{mk}}(x)}\right], \\
\mathcal{S}^*(m,k)(x) &:= \left(\frac{\sigma_{k|C_{mk}}(x)}{\beta_0 + \beta_1\mu_{k|C_{mk}}(x)}\right)^2 - \frac{\mu_{k|C_{mk}}(x)}{\beta_0 + \beta_1\mu_{k|C_{mk}}(x)} \quad \text{for } x \in \mathcal{X}(C_{mk}).
\end{aligned}$$

Next we introduce Proposition 24 which ensures the each component of the true overdispersion score $\mathcal{S}^*(m,k)(x)$ for $k \neq \pi_m$ is bounded away from $m_{\min} > 0$.

Proposition 24 *For all $j \in V$, $pa_0(j) \subset pa(j)$, $pa_0(j) \neq \emptyset$ and $S \subset nd(j) \setminus pa_0(j)$, there exists $m_{\min} > 0$ such that*

$$Var(T_j(X_j) \mid X_S) - \mathbb{E}(T_j(X_j) \mid X_S) > m_{\min}.$$

Now we define the following two events: For any $j \in V$, $k \in V \setminus \{\pi_1, ..., \pi_{j-1}\}$ and $m \in \{1, 2, ..., p-1\}$

$$\begin{aligned}
\xi_1 &:= \{\max_j \max_{i \in \{1,2,\cdots,n\}} |X_j^{(i)}| < 4\log\eta\} \\
\xi_3 &:= \{\max_{m,k} |\widehat{\mathcal{S}}(m,k) - \mathcal{S}^*(m,k)| < \frac{m_{\min}}{2}\}.
\end{aligned}$$

Then,

$$\begin{aligned}
P(\widehat{\pi} \neq \pi^*) &\overset{(a)}{\leq} P(\widehat{\pi} \neq \pi^*, \xi_3) + P(\xi_3^c, \xi_1) + P(\xi_1^c) \\
&\overset{(b)}{\leq} P(\widehat{\pi}_1 \neq \pi_1^*, \xi_3) + P(\widehat{\pi}_2 \neq \pi_2^*, \xi_3 \mid \widehat{\pi}_1 = \pi_1^*) + \cdots \\
&\quad + P(\widehat{\pi}_p \neq \pi_p^*, \xi_3 \mid \widehat{\pi}_1 = \pi_1^*, \cdots, \widehat{\pi}_{p-1} = \pi_{p-1}^*) + P(\xi_3^c, \xi_1) + P(\xi_1^c).
\end{aligned} \tag{28}$$

(a) follows from $P(A) \leq P(A \cap B) + P(B^c)$, and (b) follows from the induction and the fact $P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B \mid A^c)P(A^c) \leq P(A) + P(B \mid A^c)$.

We prove the probability bound (28) by induction. For the first step ($m = 1$), overdispersion scores of π_1 in Equation (4) are used where a set of candidate element of π_1 is $\{1, 2, \cdots, p\}$. Then,

$$\begin{aligned} P(\widehat{\pi}_1 \neq \pi_1^*, \xi_3) &= P\left(\exists k' \in V \setminus \{\pi_1^*\} \text{ such that } \widehat{\mathcal{S}}(1, \pi_1^*) > \widehat{\mathcal{S}}(1, k'), \xi_3\right) \\ &\overset{(a)}{\leq} (p-1) \max_{k' \in V \setminus \{\pi_1^*\}} P\left(\mathcal{S}^*(1, \pi_1^*) + \frac{m_{\min}}{2} > \mathcal{S}^*(1, k') - \frac{m_{\min}}{2}, \xi_3\right) \\ &\overset{(b)}{=} (p-1) \max_{k' \in V \setminus \{\pi_1^*\}} P\left(m_{\min} > \mathcal{S}^*(1, k), \xi_3\right) \\ &\overset{(c)}{=} 0. \end{aligned}$$

(a) follows from the union bound and the definition of ξ_3. (b) follows from that $\mathcal{S}^*(1, \pi_1^*) = 0$ by the property of the transformation $T_j(\cdot)$ specified in Proposition 3, and (c) follows from Proposition 24.

For the $m = (j-1)^{th}$ step, assume that the first $j-1$ elements of the estimated ordering are correct $(\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \widehat{\pi}_{j-1}) = (\pi_1^*, \cdots, \pi_{j-1}^*)$. Then for the $m = j^{th}$ step, we consider the probability of a false recovery of π_j^* given $(\pi_1^*, \cdots, \pi_{j-1}^*)$. Using the same argument as the first step, the following result is straightforward.

$$\begin{aligned} P(\widehat{\pi}_j \neq \pi_j^*, \xi_3 \mid \pi_1^*, \cdots, \pi_{j-1}^*) &= P\left(\exists k \in V \setminus \{\pi_j^*\} \text{ such that } \widehat{\mathcal{S}}(j, \pi_j^*) > \widehat{\mathcal{S}}(j, k), \xi_3\right) \\ &\overset{(a)}{\leq} p \max_{k \in V \setminus \{\pi_j^*\}} P\left(\mathcal{S}^*(j, \pi_j^*) + \frac{m_{\min}}{2} > \mathcal{S}^*(j, k) - \frac{m_{\min}}{2}, \xi_3\right) \\ &\overset{(b)}{=} p \max_{k \in V \setminus \{\pi_j^*\}} P\left(m_{\min} > \mathcal{S}^*(j, k), \xi_3\right) \\ &\overset{(c)}{=} 0. \end{aligned}$$

Therefore, for any $j \in V$,

$$P(\widehat{\pi}_j \neq \pi_j^*, \xi_3 \mid \widehat{\pi}_1 = \pi_1^*, \cdots, \widehat{\pi}_{j-1} = \pi_{j-1}^*) = 0.$$

Then, the probability bound (28) is reduced to $P(\widehat{\pi} \neq \pi^*) \leq P(\xi_3^c, \xi_1) + P(\xi_1^c)$. Note that $P(\xi_1^c) \leq M \cdot \eta^{-2}$ by Proposition 17. The following lemma provides the upper bound of $P(\xi_3^c, \xi_1)$.

Lemma 25 *There exist positive constants C_1 and C_2 such that*

$$P(\xi_3^c, \xi_1) \leq C_1 p^2 c_0^{-1} exp\left(-C_2 \frac{c_0 \cdot n}{\log^4 \eta}\right).$$

where c_0 is the sample cut-off parameter.

Lastly, we define a condition on the sample cut-off parameter c_0. Intuitively if c_0 is too small, the estimated overdispersion scores may be biased due to the lack of samples.

In contrast, if c_0 is too large, all components of the conditioning set C_{mk} may not have enough samples size ($> c_0 \cdot n$), and therefore overdispersion scores cannot be calculated. The following proposition provides a maximum value of c_0 ensuring that overdispersion scores exist.

Proposition 26 *On the event ξ_1, if $c_0 \leq (3\log(\eta))^{-d}$ then the conditioning set C_{mk} has at least $c_0 \cdot n$ samples.*

The combination of Lemma 25 and Proposition 26 imply that for some C_1 and C_2

$$P(\xi_3^c, \xi_1) \leq C_1 p^2 \log^d(\eta) \exp\left(-C_2 \frac{n}{(\log(\eta))^{4+d}}\right).$$

Therefore,

$$P(\widehat{\pi} \neq \pi^*) \leq C_1 p^2 \log^d(\eta) \exp\left(-C_2 \frac{n}{\log^{4+d}\eta}\right) + \frac{M}{\eta^2}.$$

■

D.1 Proof for Proposition 24

Proof In the proof of the identifiability theorem in Appendix A, we obtain

$$\mathrm{Var}(T_j(X_j) \mid X_S) - \mathbb{E}(T_j(X_j) \mid X_S) = \frac{(1+\beta_1)\mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_S)}{(\beta_0 + \beta_1 \mathbb{E}(X_j \mid X_S))^2}.$$

By Assumption 13, $\mathrm{Var}(\mathbb{E}(X_j \mid X_{pa(j)}) \mid X_S) > M_{\min}$ and $|\beta_{j0} + \beta_{j1}\mathbb{E}(X_j \mid X_S)| > \omega_{\min}$. Then,

$$\mathrm{Var}(T_j(X_j) \mid X_S) - \mathbb{E}(T_j(X_j) \mid X_S) \geq \frac{(1+\beta_1)M_{\min}}{\omega_{\min}^2}.$$

Since $\beta_1 > -1$, the proof is complete. ■

D.2 Proof for Proposition 26

Proof Let $|X_S|$ denote the cardinality of a set $\{X_S^{(1)}, X_S^{(2)}, \cdots, X_S^{(n)}\}$ and $|\mathcal{X}(S)|$ denote the cardinality of the truncated set $\mathcal{X}(S) := \{x \in \{X_S^{(1)}, X_S^{(2)}, \cdots, X_S^{(n)}\} : n(x) \geq c_0 \cdot n\}$.

If $|\mathcal{X}(S)| = 1$, for all $x \in \{X_S^{(1)}, X_S^{(2)}, \cdots, X_S^{(n)}\}$, $n_S(x) = c_0 \cdot n - 1$ except for a single $z \in \mathcal{X}(S)$ where $n_S(z) \geq c_0.n$. In this case, the total sample size $n = n_S(z) + (|X_S| - 1)(c_0 \cdot n - 1)$. Hence

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

Since $c_0 \cdot n \leq n_S(z)$,

$$c_0 \leq \frac{n + |X_S| - 1}{n \cdot |X_S|}.$$

Note that $\frac{1}{|X_S|} \leq \frac{n+|X_S|-1}{n\cdot|X_S|}$ and $|X_j^{(i)}| \leq 4\log(\eta)$ for all $j \in V$ and $i \in \{1, 2, \cdots, n\}$ given ξ_1. Then the maximum cardinality of X_S is $(4\log(\eta))^{|S|}$. Hence if $c_0 \leq (4\log(\eta))^{-|S|}$ there exists a $z \in \mathcal{X}(S)$.

Recall that the size of a candidate parents set C_{mk} is bounded by the maximum degree of the moralized graph d. Therefore if $c_0 \leq 4\log(\eta)^{-d}$, there exists at least one $z \in \mathcal{X}(C_{mk})$. ■

D.3 Proof for Lemma 25

Proof For ease of notation, let $n_{mk} = n_{C_{mk}}$ and $n_{mk}(x) = n_{C_{mk}}(x)$ for $x \in \mathcal{X}(C_{mk})$. Using the union bound, for $m \in \{1, 2, ..., p-1\}$ and $k \in V \setminus \{\pi_1, ..., \pi_{j-1}\}$

$$\begin{aligned} P(\xi_3^c, \xi_1) &= P(\max_{m,k} |\widehat{\mathcal{S}}(m,k) - \mathcal{S}^*(m,k)| > \frac{m_{\min}}{2}, \xi_1) \\ &\leq p^2 \max_{m,k} P(|\widehat{\mathcal{S}}(m,k) - \mathcal{S}^*(m,k)| > \frac{m_{\min}}{2}, \xi_1). \end{aligned}$$

Since overdispersion scores have an additive form,

$$P(|\widehat{\mathcal{S}}(m,k) - \mathcal{S}^*(m,k)| > \frac{m_{\min}}{2}, \xi_1) \leq P(\sum_{x \in \mathcal{X}(C_{mk})} \frac{n_{mk}(x)}{n_{mk}} |\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1).$$

Applying $P(\sum_i Y_i > \delta) \leq \sum_i P(Y_i > \omega_i \delta)$ for any $\delta \in \mathbb{R}$ and $\omega_i \in \mathbb{R}^+$ such that $\sum_i \omega_i = 1$, we have

$$\begin{aligned} &P(\sum_{x \in \mathcal{X}(C_{mk})} \frac{n_{mk}(x)}{n_{mk}} |\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1) \\ &\qquad \leq \sum_{x \in \mathcal{X}(C_{mk})} P(|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1). \end{aligned}$$

Applying the union bound,

$$\begin{aligned} &\sum_{x \in \mathcal{X}(C_{mk})} P(|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1) \\ &\qquad \leq |\mathcal{X}(C_{mk})| \max_{x \in \mathcal{X}(C_{mk})} P(|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}(m,k)^*(x)| > \frac{m_{\min}}{2}, \xi_1). \end{aligned}$$

Since we only consider $x \in \mathcal{X}(C_{mk})$, it follows that $n_{mk}(x) \geq c_0 \cdot n$. Further since the total truncated sample size is less than total sample size, $c_0 \cdot n \cdot |\mathcal{X}(C_{mk})| \leq n$, and therefore the cardinality of C_{mk} is at most c_0^{-1}. Hence

$$\begin{aligned} &|\mathcal{X}(C_{mk})| \max_{x \in \mathcal{X}(C_{mk})} P(|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1) \\ &\qquad \leq c_0^{-1} \max_{x \in \mathcal{X}(C_{mk})} P(|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1). \end{aligned}$$

Since the overdispersion score is the difference between the conditional mean and conditional variance, the remainder of the proof is reduced to finding the sample complexity for

the sample conditional mean and variance. Suppose that $\epsilon := \widehat{\mu}_{k|C_{mk}}(x) - \mu_{k|C_{mk}}(x)$ and $\kappa \cdot \epsilon := \widehat{\sigma}^2_{k|C_{mk}}(x) - \sigma^2_{k|C_{mk}}(x)$ for some $\kappa \in \mathbb{R}$. By the definition of the overdispersion scores in Equation (27), we have

$$
\begin{aligned}
&\{\epsilon : |\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}\} \\
&\subset \left\{ \epsilon : \left| \left(\frac{\sigma_{k|C_{mk}}(x) + \kappa\epsilon}{\beta_0 + \beta_1 \mu_{k|C_{mk}}(x) + \epsilon} \right)^2 - \frac{\mu_{k|C_{mk}}(x) + \epsilon}{\beta_0 + \beta_1 \mu_{k|C_{mk}}(x) + \epsilon} \right. \right. \\
&\qquad \left. \left. - \left(\frac{\sigma_{k|C_{mk}}(x)}{\beta_0 + \beta_1 \mu_{k|C_{mk}}(x)} \right)^2 - \frac{\mu_{k|C_{mk}}(x)}{\beta_0 + \beta_1 \mu_{k|C_{mk}}(x)} \right| > \frac{m_{\min}}{2} \right\} \\
&= \{\epsilon : \epsilon \in (\epsilon_1, \epsilon_2) \cup (\epsilon_3, \epsilon_4)\}.
\end{aligned}
$$

where $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ are constants that depend on $\mu, \sigma^2, \beta_0, \beta_1, m_{\min}$, and κ and are constructed as follows:

$$
\begin{aligned}
\zeta_1(\mu, \sigma^2, \beta_0, \beta_1, m_{\min}, \kappa) &:= \beta_0^3(1 + \beta_1 m_{\min}) - \beta_1^4 m_{\min} \mu^3 + 2\beta_1^2 \mu^2 \kappa \sigma^2 - 2\beta_1^2 \mu \sigma^4 \\
&\quad + \beta_0^2(-2\beta_1 \mu - 3\beta_1^2 m_{\min} \mu + 2\kappa\sigma^2) - \beta_0 \beta_1 \big\{ \beta_1 \mu^2 + 3\beta_1^2 m_{\min} \mu^2 + 2\sigma^2(-2\kappa\mu + \sigma^2) \big\}, \\
\zeta_2(\mu, \sigma^2, \beta_0, \beta_1, m_{\min}, \kappa) &:= (\beta_0 + \beta_1 \mu)^2 \Big[\beta_0^4(1 + 2\kappa\mu) + 2\beta_1^2(\kappa\mu - \sigma^2)^2(\beta_1^2 \mu^2 m_{\min} + 2\sigma^4) \\
&\quad + 4\beta_0 \beta_1 (\kappa\mu - \sigma^2) \big\{ \beta_1^2 \mu m_{\min}(2\kappa\mu - \sigma^2) + \beta_1 \mu \sigma^2 - 2\kappa\sigma^2 \big\} \\
&\quad + 2\beta_0^3 \big\{ -2\kappa\sigma^2 + \beta_1(\mu + 4 m_{\min} \kappa^2 \mu - 2 m_{\min} \kappa \sigma^2) \big\} \\
&\quad + \beta_0^2 \big\{ 4\kappa^2\sigma^4 + 4\beta_1 \sigma^2(-2\kappa\mu + \sigma^2) + \beta_1^2(\mu^2 + 12 m_{\min} \kappa^2 \mu^2 - 12 m_{\min} \mu \kappa \sigma^2 + 2 m_{\min} \sigma^4) \big\} \Big], \\
\zeta_3(\mu, \sigma^2, \beta_0, \beta_1, m_{\min}, \kappa) &:= \beta_0^2(-2\kappa^2 + 2\beta_1 + \beta_1^2 m_{\min}) + 2\beta_0 \beta \mu(\beta_1 + \beta_1^2 m_{\min} - \kappa^2) \\
&\quad + \beta_1^2(\beta_1^2 m_{\min} \mu^2 + 2\sigma^4 - 2\kappa^2 \mu^2).
\end{aligned}
$$

Given $\zeta_1, \zeta_2, \zeta_3$,

$$
\begin{aligned}
\epsilon_1' &= \frac{\zeta_1(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, m_{\min}, \kappa) + \sqrt{\zeta_2(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, m_{\min}, \kappa)}}{\zeta_3(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, m_{\min}, \kappa)}, \\
\epsilon_2' &= \frac{-\zeta_1(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, m_{\min}, \kappa) + \sqrt{\zeta_2(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, m_{\min}, \kappa)}}{\zeta_3(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, m_{\min}, \kappa)}, \\
\epsilon_3' &= \frac{\zeta_1(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, -m_{\min}, \kappa) + \sqrt{\zeta_2(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, -m_{\min}, \kappa)}}{\zeta_3(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, -m_{\min}, \kappa)}, \\
\epsilon_4' &= \frac{-\zeta_1(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, -m_{\min}, \kappa) + \sqrt{\zeta_2(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, -m_{\min}, \kappa)}}{\zeta_3(\mu_{k|C_{mk}}(x), \sigma^2_{k|C_{mk}}(x), \beta_0, \beta_1, -m_{\min}, \kappa)}.
\end{aligned}
$$

Let $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ be the ordered values of $(\epsilon_1', \epsilon_2', \epsilon_3', \epsilon_4')$ from smallest to largest. Since $m_{\min} > 0$ it follows that $\epsilon_1, \epsilon_2 < 0$ and $\epsilon_3, \epsilon_4 > 0$.

For ease of notation, $\epsilon_{\min} = \min\{|\epsilon_2|, |\epsilon_3|\}$. Then,

$$\{\epsilon : |\widehat{\mathcal{S}}(j,k)(x) - \mathcal{S}^*(j,k)(x)| > \frac{m_{\min}}{2}\} \subset (-\infty, -\epsilon_{\min}) \cup (\epsilon_{\min}, \infty).$$

Hence

$$P\{|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}\}$$
$$\leq P\left(|\widehat{\mu}_{k|C_{mk}}(x) - \mu_{k|C_{mk}}(x)| > \epsilon_{\min}\right) + P\left(|\widehat{\sigma}^2_{k|C_{mk}}(x) - \sigma^2_{k|C_{mk}}(x)| > \kappa\epsilon_{\min}\right).$$

On ξ_1, $\max_{i,j} |X_j^{(i)}| \leq 4\log(\eta)$. Furthermore recall that $n_{mk}(x) \geq c_0 \cdot n$. By applying Hoeffding's inequality,

$$P(|\widehat{\mu}_{k|C_{mk}}(x) - \mu_{k|C_{mk}}(x)| > \epsilon_{\min}, \xi_1) \leq 2\exp\left(-\frac{\epsilon^2_{\min} c_0.n}{8\log^2\eta}\right).$$

Note that sample variance can be decomposed as follows:

$$\frac{1}{n-1}\left(\sum_i^n X_i^2 - \frac{1}{n}(\sum_i^n X_i)^2\right) = \frac{1}{n}\sum_i^n X_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j} X_i X_j.$$

Using Hoeffding's inequality for the decomposed sample variance,

$$P(|\widehat{\sigma}^2_{k|C_{mk}}(x) - \sigma^2_{k|C_{mk}}(x)| > |\kappa| \cdot \epsilon_{\min}, \xi_1) \leq 2\exp\left(-\frac{\kappa^2\epsilon^2_{\min} c_0 \cdot n}{128\log^4\eta}\right) + 2\exp\left(-\frac{\kappa^2\epsilon^2_{\min} c_0 \cdot n}{256\log^4\eta}\right).$$

Therefore,

$$P\{|\widehat{\mathcal{S}}(m,k)(x) - \mathcal{S}^*(m,k)(x)| > \frac{m_{\min}}{2}, \xi_1\}$$
$$\leq 2\left(\exp\left(-\frac{\epsilon^2_{\min} c_0.n}{8\log^2\eta}\right) + \exp\left(-\frac{\kappa^2\epsilon^2_{\min} c_0 \cdot n}{128\log^4\eta}\right) + \exp\left(-\frac{\kappa^2\epsilon^2_{\min} c_0 \cdot n}{256\log^4\eta}\right)\right).$$

This completes the proof since there exist constants C_1 and C_2 such that

$$P(\xi_3^c, \xi_1) \leq C_1 p^2 c_0^{-1} \exp\left(-C_2 \frac{c_0 \cdot n}{\log^4\eta}\right).$$

■

Appendix E. Proof for Theorem 15

Proof Once again we use the *primal-dual witness* method used in the the proof for Theorem 12. The only difference is the conditioning set. In this proof, the conditioning set is all elements of the ordering before node j rather than j is $V \setminus \{j\}$. Without loss of generality, we assume the true ordering is $\pi^* = (1, 2, \cdots, p)$. Then the conditioning set is $\{1, 2, \cdots, j-1\}$.

For ease of notation, we define the parameter $\theta \in \mathbb{R}^{j-1}$ since the node j is not penalized in (11). Then, the conditional negative log-likelihood of a GLM (10) for X_j given $X_{1:j-1}$ is:

$$\ell_j^D(\theta; X^{1:n}) = \frac{1}{n}\sum_{i=1}^{n}\left(-X_j^{(i)}\langle\theta, X_{1:j-1}^{(i)}\rangle + A_j(\langle\theta, X_{1:j-1}^{(i)}\rangle)\right).$$

Recall that for any node $j \in V$:

$$\widehat{\theta}_{D_j} := \arg\min_{\theta\in\mathbb{R}^{j-1}} \mathcal{L}_j^D(\theta, \lambda_n^D) = \arg\min_{\theta\in\mathbb{R}^{j-1}} \{\ell_j^D(\theta; X^{1:n}) + \lambda_n^D\|\theta\|_1\}.$$

Using the *sub-differential*, $\widehat{\theta}_{D_j}$ should satisfy the following condition. For notational simplicity, let $S = \mathrm{pa}(j)$ for node $j \in V$.

$$\bigtriangledown_\theta \mathcal{L}_j^D(\widehat{\theta}_{D_j}, \lambda_n^D) = \bigtriangledown_\theta \ell_j^D(\widehat{\theta}_{D_j}; X^{1:n}) + \lambda_n^D \widehat{Z} = 0 \tag{29}$$

where $\widehat{Z} \in \mathbb{R}^{j-1}$ and $\widehat{Z}_t = \mathrm{sign}([\widehat{\theta}_{D_j}]_t)$ if a node $t \in S$, otherwise $|\widehat{Z}_t| < 1$.

By Lemma 19, it is sufficient the show that $|\widehat{Z}_t| < 1$ for all $t \in S$. We note that the restricted solution is $(\widetilde{\theta}_{D_j}, \widetilde{Z})$. Equation (29) with the dual solution $(\widetilde{\theta}_{D_j}, \widetilde{Z})$ can be represented as $\bigtriangledown^2\ell_j^D(\theta^*_{D_j}; X^{1:n})(\widetilde{\theta}_{D_j} - \theta^*_{D_j}) = -\lambda_n^D\widetilde{Z} - W_{Dj}^n + R_{Dj}^n$ by using the mean value theorem where:

(a) W_{Dj}^n is the sample score function,

$$W_{Dj}^n := -\bigtriangledown \ell_j^D(\theta^*_{D_j}; X^{1:n}). \tag{30}$$

(b) $R_{Dj}^n = (R_{Dj1}^n, R_{Dj2}^n, \cdots, R_{Djj-1}^n)$ and R_{Djk}^n is the remainder term by applying coordinate-wise mean value theorem,

$$R_{Djk}^n := [\bigtriangledown^2\ell_j^D(\theta^*_{D_j}; X^{1:n}) - \bigtriangledown^2\ell_j^D(\bar{\theta}^{(k)}_{D_j}; X^{1:n})]_k^T(\widetilde{\theta}^{(k)}_{D_j} - \theta^*_{D_j}) \tag{31}$$

where $\bar{\theta}^{(j)}_{D_j}$ is a vector on the line between $\widetilde{\theta}_{D_j}$ and $\theta^*_{D_j}$ and $[\cdot]_k^T$ is the k^{th} row of a matrix.

Similar to Proposition 20, the following corollary provides a sufficient condition to control $\widetilde{Z}$.

Corollary 27 *Suppose that* $\max(\|W_{Dj}^n\|_\infty, \|R_{Dj}^n\|_\infty) \leq \frac{\lambda_n\alpha}{4(2-\alpha)}$. *Then,* $|\widetilde{Z}_t| < 1$ *for all* $t \notin pa(j)$.

Now we introduce the following three corollaries, to verify that the conditions in Proposition 27 hold, and the deviation $\widetilde{\theta}_{M_j} - \theta^*_{D_j}$ is sufficiently small to conclude $\widehat{\mathrm{pa}}(j) = \mathrm{pa}(j)$ with high probability. For ease of notation, let $\eta = \max\{n, p\}$ and For notational convenience, we use $\widetilde{\theta}_S = [\widetilde{\theta}_{D_j}]_S$ and $\widetilde{\theta}_{S^c} = [\widetilde{\theta}_{D_j}]_{S^c}$. Suppose that Assumptions 8, 9, 10, and 11 are satisfied.

Corollary 28 *Suppose that* $\lambda_n^D \geq \frac{16\max\{n^{\kappa_2}\log\eta, \log^2\eta\}}{n^a}$ *for some* $a \in \mathbb{R}$. *Then,*

$$P(\frac{\|W_{Dj}^n\|_\infty}{\lambda_n^D} \leq \frac{\alpha}{4(2-\alpha)}) \geq 1 - 2d \cdot exp(-\frac{\alpha^2}{8(2-\alpha)^2} \cdot n^{1-2a}) - M \cdot \eta^{-2}.$$

Corollary 29 *Suppose that* $\|W_{Dj}^n\|_\infty \leq \frac{\lambda_n^D}{4}$. *For* $\lambda_n^D \leq \frac{\rho_{\min}^2}{40\rho_{\max}} \frac{1}{n^{\kappa_2}\log\eta d}$,

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

Corollary 30 *Suppose that* $\|W_{Dj}^n\|_\infty \leq \frac{\lambda_n^D}{4}$. *For* $\lambda_n^D \leq \frac{\alpha}{400(2-\alpha)} \frac{\rho_{\min}^2}{\rho_{\max}} \frac{1}{n^{\kappa_2} d\log\eta}$,

$$P\left(\|R_{Dj}^n\|_\infty \leq \frac{\alpha\lambda_n^D}{4(2-\alpha)}\right) \geq 1 - 2M \cdot \eta^{-2}.$$

Consider the choice of regularization parameter $\lambda_n^D = \frac{16\max\{n^{\kappa_2}\log\eta, \log^2\eta\}}{n^a}$ where $a \in (2\kappa_2, 1/2)$. Then, the condition for Corollary 28 is satisfied, and therefore $\|W_{Dj}^n\|_\infty \leq \frac{\lambda_n^D}{4}$. Moreover, the conditions for Corollaries 29 and 30 are satisfied for a sufficiently large sample size $n \geq D'\max\{(d\log^2\eta)^{\frac{1}{a-2\kappa_2}}, (d\log^3\eta)^{\frac{1}{a-\kappa_2}}\}$ for a positive constant D'. Therefore, there exist some positive constants D_1, D_2 and D_3 such that

$$\|\widetilde{Z}_{S^c}\|_\infty \leq (1-\alpha) + (2-\alpha)\left[\frac{\|W_{Dj}^n\|_\infty}{\lambda_n^D} + \frac{\|R_{Dj}^n\|_\infty}{\lambda_n^D}\right] \leq (1-\alpha) + \frac{\alpha}{4} + \frac{\alpha}{4} < 1, \qquad (32)$$

with probability of at least $1 - D_1 d\exp(-D_2 n^{1-2a}) - D_3\eta^{-2}$.

For sign consistency, it is sufficient to show that $\|\widehat{\theta}_{D_j} - \theta_{D_j}^*\|_\infty \leq \frac{\|\theta_{D_j}^*\|_{\min}}{2}$. By Corollary 29, we have $\|\widehat{\theta}_{D_j} - \theta_{D_j}^*\|_\infty \leq \|\widehat{\theta}_{D_j} - \theta_{D_j}^*\|_2 \leq \frac{5}{\lambda_{\min}}\sqrt{d}\ \lambda_n^D \leq \frac{\|\theta_{D_j}^*\|_{\min}}{2}$ as long as $\|\theta_{D_j}^*\|_{\min} \geq \frac{10}{\lambda_{\min}}\sqrt{d}\ \lambda_n^D$.

Lastly, Lemma 7(a) guarantees that ℓ_1-penalized likelihood regression recovers the parent set for each node with high probability. Because we have p regression problems if $n \geq D'\max\{(d\log^2\eta)^{\frac{1}{a-2\kappa_2}}, (d\log^3\eta)^{\frac{1}{a-\kappa_2}}\}$, the full DAG model is recovered with high probability:

$$P(\widehat{G} = G) \geq 1 - D_1 d \cdot p \cdot \exp(-D_2 n^{1-2a}) - D_3\eta^{-1}.$$

■

References

Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. HITON: a novel Markov Blanket algorithm for optimal variable selection. In *AMIA Annual Symposium Proceedings*, volume 2003, page 21. American Medical Informatics Association, 2003.

David Maxwell Chickering. Learning Bayesian networks is NP-complete. In *Learning from data*, pages 121–130. Springer, 1996.

David Maxwell Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2003.

Robert G. Cowell, Phillip A. Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

Charmaine B Dean. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992.

Kenji Doya. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press, 2007.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

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

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

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

Ali Jalali, Pradeep D Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 378–387, 2011.

Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Trans. on Infor. Theory*, 56(10):5168–5194, 2010.

Jeffrey O Kephart and Steve R White. Directed-graph epidemiological models of computer viruses. In *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*, pages 343–359. IEEE, 1991.

Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Carl N Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80, 1982.

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

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, page ast043, 2013.

Jonas Peters, Joris Mooij, Dominik Janzing, et al. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graphs based on sparsest permutations. *arXiv preprint arXiv:1307.0366*, 2013.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Ioannis Tsamardinos and Constantin 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.

S. van de Geer and P. Bühlmann. Penalized maximum likelihood estimation for sparse directed acyclic graphs. *Annals of Statistics*, 41:536–567, 2013.

Martin J Wainwright, John D Lafferty, and Pradeep K Ravikumar. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472, 2006.

Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. On graphical models via univariate exponential family distributions. *arXiv preprint arXiv:1301.4183*, 2013.

Tian Zheng, Matthew J Salganik, and Andrew Gelman. How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association*, 101(474):409–423, 2006.