Journal of Machine Learning Research 24 (2023) 1-32 Submitted 12/21; Revised 3/23; Published 4/23

Model-Based Causal Discovery for Zero-Inflated Count Data

Junsouk Choi JCHOI@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 98195-4322, USA

Yang Ni YNI@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 94720-1776, USA

Editor: Joris Mooij

Abstract

Zero-inflated count data arise in a wide range of scientific areas such as social science, biology, and genomics. Very few causal discovery approaches can adequately account for excessive zeros as well as various features of multivariate count data such as overdispersion. In this paper, we propose a new zero-inflated generalized hypergeometric directed acyclic graph (ZiG-DAG) model for inference of causal structure from purely observational zero-inflated count data. The proposed ZiG-DAGs exploit a broad family of generalized hypergeometric probability distributions and are useful for modeling various types of zero-inflated count data with great flexibility. In addition, ZiG-DAGs allow for both linear and nonlinear causal relationships. We prove that the causal structure is identifiable for the proposed ZiG-DAGs via a general proof technique for count data, which is applicable beyond the proposed model for investigating causal identifiability. Score-based algorithms are developed for causal structure learning. Extensive synthetic experiments as well as a real dataset with known ground truth demonstrate the superior performance of the proposed method against state-of-the-art alternative methods in discovering causal structure from observational zero-inflated count data. An application of reverse-engineering a gene regulatory network from a single-cell RNA-sequencing dataset illustrates the utility of ZiG-DAGs in practice.

Keywords: Bayesian network; Causal identifiability; Directed acyclic graph; Observational zero-inflated count data; Single-cell RNA-sequencing.

1. Introduction

Discovering causal structure of an unknown system is an important task in practically all areas of science. Knowing the causal structure is not only useful for predicting a system's behavior under external interventions, but also has implications for machine learning tasks such as covariate shift and transfer learning (Schölkopf et al., 2012). The most effective and principled way for causal discovery is to conduct controlled experiments. However, it is often expensive, unethical, or even impossible in certain fields such as genomics (Opgen-Rhein and Strimmer, 2007) and social sciences (Bollen, 1989). Hence, causal discovery approaches that can infer the unknown causal structures from purely observational data are often desired.

This paper considers causal discovery for purely observational zero-inflated count data. Observational zero-inflated count data are common across multiple disciplines, for instance, educational psychology (Fox, 2013), genomics (Kang et al., 2011), ecology (Barry and Welsh, 2002), behavior studies (Hua et al., 2014), and economics (Staub and Winkelmann, 2013). A specific application, by which we are motivated, is to reverse-engineer gene regulatory networks from single-cell RNA-sequencing (scRNA-seq) data. The scRNA-seq technology measures the abundance of mRNA within single cells, resulting in count data with excessive zeros because of technological limits in sequencing the low amounts of mRNA in individual cells. For causal structure learning from observational zero-inflated count data, we work under the framework of causal Bayesian networks (BNs), which have been widely used for representing causal relationships among variables via directed acylic graphs (DAGs).

Learning the structure of BNs is not trivial because the size of the space of possible graph structures grows super-exponentially in the number of variables. Furthermore, BNs may not be distinguishable from each other with observational data. Multiple DAGs can encode the same conditional independence assertions and in general, DAGs are identifiable only up to Markov equivalence class (MEC) in which all DAGs encode the same set of conditional independences (Heckerman et al., 1995). Therefore, in the past, many approaches have focused on identifying the MEC rather than individual DAGs (Spirtes et al., 2000; Chickering, 2002; Kalisch and Bühlman, 2007; Castelletti et al., 2018). For example, the well-known PC algorithm infers a set of conditional independencies and recovers a MEC that is compatible with the inferred conditional independencies (Spirtes et al., 2000). The GES algorithm performs greedy search over the space of MECs and obtains the best-scored MEC (Chickering, 2002). However, DAGs within the same MEC may have drastically different causal interpretations.

Since 2006, it has been shown that for some classes of BNs, the exact graph structure, not just the MEC, may be identifiable from observational data alone. For continuous variables, BNs are often represented by sparse additive noise models. Under this formulation, the underlying DAG is identifiable if the functional form of the additive noise model is linear with non-Gaussian noises (Shimizu et al., 2006; Wang and Drton, 2020) and if the functional form is nonlinear with mild regularity assumptions on the function-noise pair (Hoyer et al., 2008; Peters et al., 2011, 2014). Peters and Bühlmann (2014); Chen et al. (2019) have also shown that unique identification of DAG structure is possible under linear additive noise models with Gaussian noises having equal variances.

The vast majority of the existing works that establish identifiability theorems for BNs have focused on continuous variables; identifiability issues of BNs for count data are less studied. Park and Raskutti (2015) proposed linear Poisson BNs for observational count data and investigated the overdispersion scores to prove that the unique identification of the underlying DAG is possible. However, the applicability of Poisson BNs may be limited due to the restrictive assumption of Poisson distribution that the variance is equal to the mean. Park and Park (2019) generalized the idea of Poisson BNs to a family of generalized hypergeometric distributions that includes the Poisson distribution, the hyper-Poisson distribution, the negative binomial distribution, and many more. An identifiability theorem for the generalized hypergeometric BNs was established using the moment ratio scores. Although the generalized hypergeometric BNs are a quite general class of count BNs, they tend not to adequately model count data with excessive zeros.

There have been a few recent BNs that are fully identifiable for observational zero-inflated data. Using Hurdle conditional distributions, Yu et al. (2020) proposed fully identifiable BNs for zero-inflated Gaussian data. Recently, we (Choi et al., 2020) developed zero-inflated Poisson BNs for observational zero-inflated count data. We have shown, theoretically and empirically, that the underlying causal DAG can be identified from observational data alone. However, the zero-inflated Poisson BNs have the same limitation as Poisson BNs, that is, Poisson distribution is a restrictive distribution. In particular, the Poisson-based BNs do not adequately account for overdispersion, a common feature of count data. Hence it is desirable to further develop a more general class of count BNs that can account for a broad range of multivariate count data with excessive zeros.

In this paper, we introduce a fairly general class of count BNs for observational zero-inflated count data, termed zero-inflated generalized hypergeometric DAGs (ZiG-DAGs). We extend the zero-inflated Poisson BNs (Choi et al., 2020) to zero-inflated generalized hypergeometric models, which include many common count distributions. Therefore, the proposed ZiG-DAGs are capable of modeling various types of zero-inflated count data, for example, overdispersed zero-inflated count data. In addition, we allow for both linear and nonlinear causal relationships in order to flexibly capture real causality in practice whereas Choi et al. (2020) only considers linear causal relationships. Based on a new general proof technique, we prove that the proposed ZiG-DAG is uniquely identifiable, justifying its use for casual discovery. The general proof technique can be potentially used to check identifiability for other discrete BNs as well. The established identifiability theorems do not require the causal faithfulness assumption (Uhler et al., 2013) typically required by constraint-based algorithms. For the structure learning of ZiG-DAGs, we develop score-based algorithms: exhaustive search for small graphs and greedy search for moderate-to-large graphs. Specifically, we consider two different greedy search algorithms to deal with the local optima problem of greedy search. We empirically demonstrate that the proposed methods compare favorably against state-of-the-art alternatives. We also illustrate the utility of ZiG-DAGs in real-world problems using a scRNA-seq dataset.

The remainder of this paper is organized as follows. We set up necessary notations and definitions for BNs in Section 2.1 and we introduce the proposed ZiG-DAG models for observational zero-inflated count data in Section 2.2. Section 3 establishes identifiability theorems for the proposed ZiG-DAGs. In Section 4, we develop score-based algorithms for causal structure learning of ZiG-DAGs. We demonstrate the utility of our methods through synthetic data in Section 5 and real-world applications in Section 6. Section 7 provides our closing discussion.

2. Bayesian Networks for Observational Zero-inflated Count Data

2.1 Notation and Background

We start with some basic notations for DAGs and BNs. Let $\boldsymbol{X} = \{X_1, \dots, X_d\}$ denote a set of d random variables. A DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ consists of a set of nodes $\boldsymbol{V} = \{1, \dots, d\}$ corresponding to the variables $\boldsymbol{X}$ and a set of directed edges $\boldsymbol{E} \subset \boldsymbol{V} \times \boldsymbol{V}$ representing the causal relationships between the nodes $\boldsymbol{V}$ without cycles. If we have a directed edge

$(j,k) \in \boldsymbol{E}$ (or slightly abusing the notation, $j \to k \in \boldsymbol{E}$) for $j,k \in \boldsymbol{V}$, node j is called a parent of k and node k is called a child of j. We denote the set of parents of node j in $\mathcal{G}$ by $pa_{\mathcal{G}}(j)$ and the set of children of j in $\mathcal{G}$ by $ch_{\mathcal{G}}(j)$. Node k is said to be a descendant of node j if there exists a directed path $j = j_0 \to j_1 \to \cdots \to j_l = k$ and otherwise is said to be a non-descendant of j. We use $nd_{\mathcal{G}}(j)$ to denote the set of non-descendants of j (excluding j itself). A BN $\mathcal{B}$ for $\boldsymbol{X}$ is a pair $\mathcal{B} = (\mathcal{G}, p)$ with the joint distribution $p(\cdot)$ factorizing over $\mathcal{G}$ as follows:

$$p(X_1, \ldots, X_d) = \prod_{j=1}^{d} p(X_j | \boldsymbol{X}_{pa_{\mathcal{G}}(j)}), \tag{1}$$

where $\boldsymbol{X}_{pa_{\mathcal{G}}(j)} = \{X_k : k \in pa_{\mathcal{G}}(j)\}$ and $p(X_j|\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ is the conditional probability distribution of X_j given its parents. We say a joint distribution $p(\cdot)$ is (local) Markov with respect to a DAG $\mathcal{G}$ if each variable X_j is independent of its non-descendants $\boldsymbol{X}_{nd_{\mathcal{G}}(j)} = \{X_k : k \in nd_{\mathcal{G}}(j)\}$ given its parents $\boldsymbol{X}_{pa_{\mathcal{G}}(j)}$. The factorization in (1) is equivalent to the Markov property of $p(\cdot)$ (Verma and Pearl, 1990). In this paper, we make the causal Markov assumption – $p(\cdot)$ is Markov with respect to the causal DAG $\mathcal{G}$ – so that we can interpret $\mathcal{G}$ causally; in other words, each node is assumed to be independent of all its non-effects conditional on all its direct causes.

In general, the DAG $\mathcal{G}$ of a BN $\mathcal{B} = (\mathcal{G}, p)$ is not identifiable from the joint distribution $p(\cdot)$. Indeed, the joint distribution $p(\cdot)$ is Markov with respect to many different DAGs including all fully connected DAGs. Therefore, we have many possible BNs with different graph structures for the same joint distribution. To overcome this indeterminacy, one can make additional assumptions and obtain a restricted model for which the graph is identifiable from the joint distribution. A common assumption in the literature for learning BNs is faithfulness. A joint distribution $p(\cdot)$ is faithful with respect to a DAG $\mathcal{G}$ if the graph $\mathcal{G}$ encodes all the conditional independence constraints in the joint distribution $p(\cdot)$. If faithfulness is assumed, DAGs are identifiable up to MEC (Spirtes et al., 2000). Two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if the two DAGs encodes the same set of conditional independence constraints and a MEC is defined by a set of DAGs that are Markov equivalent. For example, despite the seemingly different graph structures, the DAGs in Figure 1(a)-(c) forms a MEC, which encodes the only conditional independence $X_1 \perp\!\!\!\perp X_2 | X_3$, whereas the DAG in Figure 1(d) encodes the marginal independence of X_1 and X_2 only and forms another MEC. Since both the Markov property and faithfulness only constrain conditional independencies in the joint distribution, we cannot distinguish DAGs in the same MEC, which impose the same set of conditional independence assertions. For instance, the well-known PC algorithm (Spirtes et al., 2000) and the GES algorithm (Chickering, 2002), under the faithfulness assumption, aim to find the best MEC rather than the best individual DAG.

In many applications of BNs, a specific family of distributions is assumed for the conditional distribution of each node given its parents. For example, we will assume that the conditional probability for each node comes from a zero-inflated count model. Even with such distributional assumptions, the DAG may still be non-identifiable due to the distribution equivalence. Two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are distribution equivalent if for every BN $\mathcal{B} = (\mathcal{G}, p)$ there exists a different BN $\mathcal{B}' = (\mathcal{G}', p')$ such that the joint distributions are identical, i.e., $p(X_1, \ldots, X_d) = p'(X_1, \ldots, X_d)$. For example, for Gaussian BNs or multinomial BNs, they

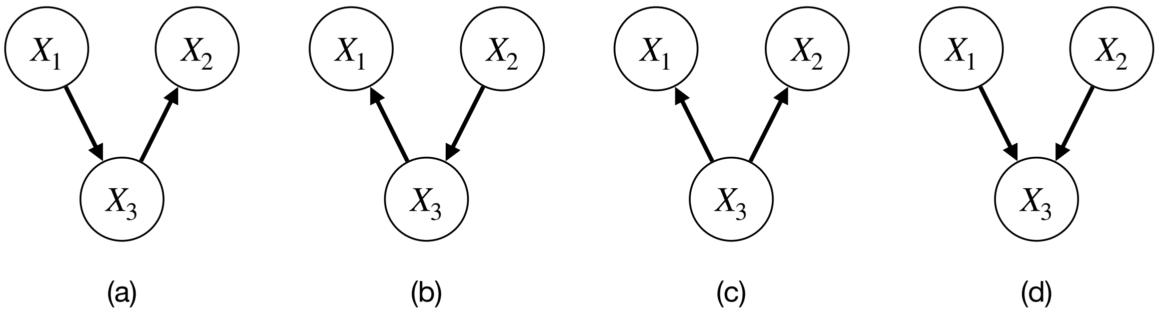

Figure 1: Examples of DAGs with three nodes. DAGs in (a)-(c) are Markov equivalent and form a Markov equivalence class that encodes $X_1 \perp\!\!\!\perp X_2 | X_3$. DAG in (d) forms another Markov equivalence class that encodes $X_1 \perp\!\!\!\perp X_2$.

are distribution equivalent if and only if they are Markov equivalent. Hence, we can identify only the MEC from the joint distribution for Gaussian and multinomial BNs. Hence, not surprisingly, if we assume that data are generated from one of the three DAGs in Figure 1(a)-(c), the best answer that we can achieve using Gaussian BNs or multinomial BNs is that one of them is the true model. This is unsatisfactory for many applications and it has been recently shown that there exist certain cases where we can overcome the issue of distribution equivalence and the graph structure is fully identifiable. The existing works often represent continuous BNs as sparse additive noise models and under this framework, the underlying DAG is identifiable if the functional form of the additive noise model is linear and the noises are non-Gaussian (Shimizu et al., 2006), if nonlinear functions are considered with very mild additional conditions (Hoyer et al., 2008; Peters et al., 2011), or if the functions are linear and the noises are Gaussian with equal variance (Peters and Bühlmann, 2014). However, most existing approaches focus on BNs for continuous data, and the identifiability of BNs for count data are much less studied (Park and Raskutti, 2015; Park and Park, 2019; Choi et al., 2020).

2.2 Zero-Inflated Generalized Hypergeometric Directed Acyclic Graphs

We consider a broad family of discrete distributions for count data. Kemp (1968a,b) defines a family of generalized hypergeometric probability distributions (GHPDs), which includes a lot of common probability distributions for count data and has many useful properties such as recurrence relationships for both their probabilities and their factorial moments. Let $(a)_k = a(a+1)\cdots(a+k-1)$ denote the ascending (rising) factorial with $(a)_0 = 1$. The generalized hypergeometric function is then defined as

$$_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; \lambda) = \sum_{i \geq 0} \frac{(a_1)_i \cdots (a_p)_i \lambda^i}{(b_1)_i \cdots (b_q)_i i!}.$$

Note that $a_1, \ldots, a_p$ are exchangeable and so are $b_1, \ldots, b_q$. A distribution is said to be a GHPD if its probability generating function can be written in the following form:

$$G(s; \boldsymbol{a}, \boldsymbol{b}, \lambda) = \frac{{}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; \lambda s)}{{}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; \lambda)}, \tag{2}$$

where $\boldsymbol{a} = (a_1, \ldots, a_p)$ and $\boldsymbol{b} = (b_1, \ldots, b_q)$. A large number of discrete distributions for count data belong to the class of GHPDs, for example, binomial, Poisson, negative binomial, hypergeometric, beta-binomial, and beta-negative binomial. Table 1 provides some examples of GHPDs with their probability generating functions (see also Kemp 1968a; Dacey 1972; Johnson et al. 2005).

Table 1: Examples of GHPDs and their probability generating functions

Distributions	Probability generating function	Parameters
Binomial	$\frac{{}_1F_0(-n;;-ps/(1-p))}{{}_1F_0(-n;;-p/(1-p))}$	$0 < p < 1$
Poisson	$\frac{{}_0F_0(;;\theta s)}{{}_0F_0(;;\theta)}$	$\theta > 0$
Hyper-Poisson	$\frac{{}_1F_1(1;\psi;\theta s)}{{}_1F_1(1;\psi;\theta)}$	$\psi > 0, \theta > 0$
Geometric	$\frac{{}_1F_0(1;;qs)}{{}_1F_0(1;;q)}$	$0 < q < 1$
Negative Binomial	$\frac{{}_1F_0(k;;qs)}{{}_1F_0(k;;q)}$	$k > 0, 0 < q < 1$
Hypergeometric	$\frac{{}_2F_1(-n,-Np;N-Np-n+1;s)}{{}_2F_1(-n,-Np;N-Np-n+1;1)}$	$n, N \in \mathbb{N}, 0 < p < 1$
Beta-Negative Binomial	$\frac{{}_2F_1(k,\ell;k+\ell+m;s)}{{}_2F_1(k,\ell;k+\ell+m;1)}$	$k, \ell, m \geq 0$
Extended Generalized Waring	$\frac{{}_2F_1(k,\ell;k+\ell+m;\theta s)}{{}_2F_1(k,\ell;k+\ell+m;\theta)}$	$k, \ell > 0, m \in \mathbb{R}, 0 < \theta < 1$

We define, by using the GHPDs, ZiG-DAGs for observational zero-inflated count data. In order to explicitly account for excessive zeros in count data, we adopt the zero-inflated model. We say a BN $\mathcal{B} = (\mathcal{G}, p)$ for random counts $\boldsymbol{X}$ is a ZiG-DAG if for each node $j \in \boldsymbol{V}$, the conditional distribution $p(X_j | \boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ of the factorization (1) has a probability generating function of the following form,

$$G_j\left(s; \boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right) = \pi_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right) + \left(1 - \pi_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right)\right) G\left(s; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right)\right), \tag{3}$$

where $G(s; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j)$ is a GHPD probability generating function defined by (2) with $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jp_j})$ and $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{jq_j})$. Here, $\pi_j(\cdot)$ and $\lambda_j(\cdot)$ are functions/mappings from $\mathcal{X}_{pa_{\mathcal{G}}(j)}$ to $\mathbb{R}$, which connect the parents $\boldsymbol{X}_{pa_{\mathcal{G}}(j)}$ of node j to its conditional distribution, where $\mathcal{X}_{pa_{\mathcal{G}}(j)} \subset \{0, 1, 2, \ldots\}^{|pa_{\mathcal{G}}(j)|}$. For a ZiG-DAG, the probability mass function of each conditional distribution is given by

$$\begin{aligned}
&Pr\left(X_j = x | \boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right) \\
&= \begin{cases} \pi_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right) + \left(1 - \pi_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right)\right) \text{GHP}\left(0; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right)\right) & \text{if } x = 0, \\ \left(1 - \pi_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right)\right) \text{GHP}\left(x; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j\left(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}\right)\right) & \text{if } x = 1, 2, \ldots, \end{cases}
\end{aligned}$$

where $\text{GHP}\left(x; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j\right)$ is the probability mass function of a GHPD of which the probability generating function is given by $G\left(s; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j\right)$. Particularly, $\pi_j(\cdot) \in (0,1)$ is the probability that extra zeros occur in addition to the zeros that arise from the GHPD, and $\lambda_j(\cdot)$ is the power parameter of GHPD that is closely related to its moments. For example, in the Poisson probability generating function $G(s; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j) = {}_0F_0(;;\lambda_j s)/{}_0F_0(;;\lambda_j)$, λ_j represents the mean of the Poisson distribution. As another example, in the negative binomial probability generating function $G(s; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j) = {}_1F_0(k;;\lambda_j s)/{}_1F_0(k;;\lambda_j)$, λ_j denotes the probability of "success", which can be reparametrized in terms of the first and second moments of the negative binomial distribution.

For $\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ and $\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$, we consider both linear and nonlinear functional forms. We use the logit function, $\text{logit}(\cdot)$, as link function for π_j, where $\text{logit}(\rho) = \log(\rho/(1-\rho))$ for $0 < \rho < 1$. Let $h_j(\cdot)$, $j = 1, \ldots, d$, denote any suitable link function for λ_j, which is assumed to be strictly increasing for invertibility. First, we define linear ZiG-DAGs by assuming $\text{logit}(\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}))$ and $h_j(\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}))$ vary linearly with the parents $\boldsymbol{X}_{pa_{\mathcal{G}}(j)}$ of node $j \in \boldsymbol{V}$.

Definition 1 (Linear ZiG-DAGs) *We say a BN $\mathcal{B} = (\mathcal{G}, p)$ is a linear ZiG-DAG if the joint distribution $p(\cdot)$ factorizes with respect to the DAG $\mathcal{G}$ as in* (1) *with each conditional distribution $p(X_j|\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ having a probability generating function* (3)*, where $\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ and $\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ are given by*

$$\begin{aligned}
\text{logit}\left(\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})\right) &= \sum_{k \in pa_{\mathcal{G}}(j)} \alpha_{jk} X_k + \delta_j, \\
h_j\left(\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})\right) &= \sum_{k \in pa_{\mathcal{G}}(j)} \beta_{jk} X_k + \gamma_j, \qquad (4)
\end{aligned}$$

with some strictly increasing functions h_j.

The zero-inflated Poisson BN in our recent work (Choi et al., 2020) is a special case of the proposed linear ZiG-DAG. Furthermore, in order to allow more flexible causal relationships, we propose nonlinear ZiG-DAGs by adopting the additive model framework. Particularly, for each X_j, we model $\text{logit}(\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}))$ and $h(\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)}))$ as the sum of nonlinear functions of X_k, $k \in pa_{\mathcal{G}}(j)$.

Definition 2 (Nonlinear ZiG-DAGs) *We say a BN $\mathcal{B} = (\mathcal{G}, p)$ is a nonlinear ZiG-DAG if the joint distribution p factorizes with respect to the DAG $\mathcal{G}$ as in* (1) *with each conditional distribution $p(X_j|\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ having a probability generating function* (3)*, where $\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$*

and $\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})$ *are given by*

$$\begin{aligned} \text{logit}\left(\pi_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})\right) &= \sum_{k \in pa_{\mathcal{G}}(j)} f_{jk}(X_k) + \mu_j, \\ h_j\left(\lambda_j(\boldsymbol{X}_{pa_{\mathcal{G}}(j)})\right) &= \sum_{k \in pa_{\mathcal{G}}(j)} g_{jk}(X_k) + \nu_j, \end{aligned} \tag{5}$$

with some strictly increasing functions h_j *and nonlinear functions* f_{jk} *and* g_{jk}.

Without loss of generality, we assume that $\mathrm{E}\left[f_{jk}(X_k)\right] = \mathrm{E}\left[g_{jk}(X_k)\right] = 0, \forall j \in \boldsymbol{V}, k \in pa_{\mathcal{G}}(j)$ because they can always otherwise be absorbed into the intercepts μ_j and ν_j. If zero-inflated random counts $\boldsymbol{X}$ follow either a linear ZiG-DAG or a nonlinear ZiG-DAG, they satisfy, by definition, the Markov property (conditional independencies) encoded in the underlying DAG. As mentioned earlier, BNs may not be identifiable due to Markov and distribution equivalence. In the next section, we will show that under the proposed ZiG-DAG models, the causal graph structure is identifiable from observational data alone.

3. Identifiability Theory

Recently, much effort has been directed to show that some assumptions on the conditional distribution of each node can impose non-independence constraints on the joint distribution so that the DAG of a BN is identifiable (Shimizu et al., 2006; Hoyer et al., 2008; Peters et al., 2014; Peters and Bühlmann, 2014). However, the existing literature mostly addresses the identifiability issue of BNs for continuous data, and there are much fewer identifiability results on BNs for count data. Park and Raskutti (2015); Park and Park (2019) developed BNs by using Poisson and the generalized hypergeometric family and showed that their causal orderings are identifiable. Our recent work (Choi et al., 2020) investigated the identifiability of the zero-inflated Poisson BNs. These three methods are special cases of the proposed ZiG-DAGs; however, none of their proof techniques is applicable in our setting. Therefore, before we state the main identifiability theories for both linear and nonlinear ZiG-DAGs, we provide a general framework to check the identifiability of discrete BNs. We provide a sufficient condition under which two discrete BNs with different DAGs must have different joint distributions. The proofs of the identifiability theorems for the proposed ZiG-DAGs are based on such a sufficient condition. Specifically, Proposition 4 formulates the sufficient condition in terms of probability generating functions for the conditional distribution of each node given its parents. As discrete distributions are often defined by the probability generating function, one can potentially use Proposition 4 to verify the identifiability of other discrete BNs. We first state two assumptions that our identifiability theories require:

Condition 3 *We assume (i) there exists no unmeasured confounder, and (ii) there is no selection bias.*

No unmeasured confounder (also known as causal sufficiency) and no selection bias are commonly adopted in the literature for causal structure learning (Chickering, 2002; Shimizu et al., 2006; Peters and Bühlmann, 2014; Maathuis et al., 2018). The BN factorization (1) does not hold if either or both assumptions in Condition 3 are violated.

For two discrete BNs $\mathcal{B} = (\mathcal{G}, p)$ and $\mathcal{B}^* = (\mathcal{G}^*, p^*)$, we denote $pa(j) = pa_{\mathcal{G}}(j)$, $pa^*(j) = pa_{\mathcal{G}^*}(j)$, $ch(j) = ch_{\mathcal{G}}(j)$, $ch^*(j) = ch_{\mathcal{G}^*}(j)$, and $nd(j) = nd_{\mathcal{G}}(j)$ for the ease of notation. We let $G_j(s; \boldsymbol{x}_{pa(j)})$ and $G_j^*(s; \boldsymbol{x}_{pa^*(j)})$ denote the probability generating functions for the conditional distributions $p(x_j|\boldsymbol{x}_{pa(j)})$ and $p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ of $\mathcal{B}$ and $\mathcal{B}^*$, respectively. Note that by definition, $p(x_j|\boldsymbol{x}_{pa(j)}) = G_j^{(x_j)}(0; \boldsymbol{x}_{pa(j)})$ where $G_j^{(x_j)}$ denotes the x_j-th derivative of G_j.

Proposition 4 *Let $\mathcal{B} = (\mathcal{G}, p)$ and $\mathcal{B}^* = (\mathcal{G}^*, p^*)$ be any two discrete BNs, where $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ and $\mathcal{G}^* = (\boldsymbol{V}, \boldsymbol{E}^*)$. Suppose that for every node $j \in \boldsymbol{V}$ for which*

$$\frac{G_j^{(x_j+1)}(0; \boldsymbol{x}_{pa(j)})}{G_j^{(x_j)}(0; \boldsymbol{x}_{pa(j)})} = \frac{G_j^{*(x_j+1)}(0; \boldsymbol{x}_{pa^*(j)})}{G_j^{*(x_j)}(0; \boldsymbol{x}_{pa^*(j)})} \prod_{k \in ch^*(j) \cap nd(j)} \frac{G_k^{*(x_k)}(0; \boldsymbol{x}_{pa^*(k) \setminus \{j\}}, x_j + 1)}{G_k^{*(x_k)}(0; \boldsymbol{x}_{pa^*(k) \setminus \{j\}}, x_j)} \tag{6}$$

holds for all possible $x_1, \ldots, x_d$, it is also true that $pa(j) = pa^*(j)$ and $ch^*(j) \cap nd(j) = \emptyset$ holds. Then, if the joint distributions of $\mathcal{B}$ and $\mathcal{B}^*$ are equivalent, i.e., $p = p^*$, we have $\boldsymbol{E} = \boldsymbol{E}^*$.

All proofs can be found in the appendices. The main idea behind the proof is to show that if the observational joint distributions p and p^* are identical, then the proposition condition (6) necessarily implies that $\mathcal{G}$ and $\mathcal{G}^*$ have to be identical. Given any topological ordering of the graph $\mathcal{G}$, we first show that the parent sets, in $\mathcal{G}$ and $\mathcal{G}^*$, of the last node in the ordering have to be identical if the joint distributions are the same. Then we use mathematical induction to show that this is also true for any node of $\boldsymbol{V}$ and therefore $\mathcal{G}$ and $\mathcal{G}^*$ have to be identical.

Sometimes, it is also of interest to identify the model parameters. When the graph structure is identifiable, the parameter identifiability simplifies to a question of whether parameters associated with the conditional distribution of each node are identifiable. Since Proposition 4 implies that the graph structure is already identifiable, if a conditional distribution of each node is uniquely determined by the associated parameters, then we necessarily have a one-to-one correspondence between the joint distribution of the BN and the set of all associated parameters, and hence parameters are identifiable.

Corollary 5 *Let $\Xi = \{\Xi_j\}_{j \in \boldsymbol{V}}$ and $\Xi^* = \{\Xi_j^*\}_{j \in \boldsymbol{V}}$ be sets of parameters that are associated with discrete BNs $\mathcal{B}$ and $\mathcal{B}^*$, respectively, where Ξ_j and Ξ_j^* denote the sets of parameters associated with the conditional distribution of the node j only. Suppose that for any $j \in \boldsymbol{V}$, the assumption in Proposition 4 holds and, furthermore, $\Xi_j = \Xi_j^*$ whenever $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$. Then, if the joint distributions of $\mathcal{B}$ and $\mathcal{B}^*$ are equivalent, i.e., $p = p^*$, we have $\Xi = \Xi^*$.*

Using Proposition 4 and Corollary 5, we prove identifiability of the underlying DAG and the associated parameters for both the linear ZiG-DAG and the nonlinear ZiG-DAG in Theorems 6 and 7.

Theorem 6 *Let $\mathcal{B} = (\mathcal{G}, p)$ be a linear ZiG-DAG. Assume Condition 3 holds. Then, if the variables are not binary, the graph $\mathcal{G}$ is identifiable from the joint distribution $p(\boldsymbol{X})$. For given (p_j, q_j) (which characterizes the generalized hypergeometric function ${}_{p_j}F_{q_j}$) and given h_j (the link function for λ_j), there is a unique set of parameters for the linear ZiG-DAG that induces the observed distribution $p(\boldsymbol{X})$.*

Theorem 7 *Let $\mathcal{B} = (\mathcal{G}, p)$ be a nonlinear ZiG-DAG. Assume Condition 3 holds. Then, if the variables are not binary, the graph $\mathcal{G}$ is identifiable from the joint distribution $p(\boldsymbol{X})$. For given (p_j, q_j) (which characterizes the generalized hypergeometric function ${}_{p_j}F_{q_j}$) and given h_j (the link function for λ_j), there is a unique set of parameters for the nonlinear ZiG-DAG that induces the observed distribution $p(\boldsymbol{X})$.*

In Theorems 6 and 7, the assumptions that p_j, q_j, h_j are given, which we make for parameter identifiability, indicates that the conditional distribution of each node in ZiG-DAGs should take a specific GHPD model among the family of GHPDs along with a specific link function. One example of such a combination would be the Poisson distribution with the log link function. Furthermore, the assumption excludes limiting cases of a given GHPD. For instance, if we use the negative binomial distribution for a node, we do not allow it to degenerate to a Poisson distribution, since the negative binomial distribution has $p_j = 1$, $q_j = 0$, while the Poisson distribution has $p_j = 0$, $q_j = 0$. This assumption seems reasonable since we have to decide which GHPD and link function to use in practice. In Sections 5 and 6, for the proposed ZiG-DAG, we consider the Poisson distribution, the hyper-Poisson distribution, and the negative binomial distribution with the log link function. With such choices of GHPD and link function, Theorems 6 and 7 state that both the causal structure and the model parameters for the proposed ZiG-DAGs are fully identifiable from the joint distribution.

Theorems 6 and 7 do not require faithfulness to prove that the exact graph structure is identifiable under the proposed ZiG-DAG models. While continuous BNs such as linear Gaussian BNs may have accidental cancellation of positive and negative causal effects and hence may become unfaithful, the proposed ZiG-DAGs do not allow such cancellation due to inherent asymmetry of count distributions. Faithfulness can be violated in an equilibrium-maintaining system such as a biological system (Andersen, 2013) and in datasets with limited sample size (Uhler et al., 2013). In such cases, therefore, causal discovery approaches that require the faithfulness assumption are not favorable. In our specific motivating application of reverse-engineering gene regulatory networks from scRNA-seq data, one should avoid the common practice of “Gaussianizing” raw scRNA-seq data because then one needs to additionally make the faithfulness assumption that may not be suitable in gene regulatory systems; instead, directly working with raw zero-inflated count data with the proposed ZiG-DAG does not suffer from this limitation.

4. Algorithms

In this section, we discuss algorithms for learning the causal structures of both the linear ZiG-DAGs and the nonlinear ZiG-DAGs. We will consider score-based approaches, which complement the Bayesian inference procedure developed in our recent work (Choi et al., 2020).

4.1 Structure Learning for Linear ZiG-DAGs

Suppose that we are given zero-inflated count data $\boldsymbol{x} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$ that are n independent realizations of $\boldsymbol{X}$ from a linear ZiG-DAG model $\mathcal{B} = (\mathcal{G}, p)$. For the linear ZiG-DAG, we denote the model parameters by $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{b}\}$ with $\boldsymbol{\alpha} = \{\alpha_{jk}\}_{(j,k)\in\boldsymbol{E}}$,

$\boldsymbol{\beta} = \{\beta_{jk}\}_{(j,k)\in \boldsymbol{E}}$, $\boldsymbol{\delta} = \{\delta_j\}_{j\in\boldsymbol{V}}, \boldsymbol{\gamma} = \{\gamma_j\}_{j\in\boldsymbol{V}}, \boldsymbol{a} = \{\boldsymbol{a}_j\}_{j\in\boldsymbol{V}}$, and $\boldsymbol{b} = \{\boldsymbol{b}_j\}_{j\in\boldsymbol{V}}$. We let $p(\cdot|\boldsymbol{\theta}, \mathcal{G})$ denote the joint distribution of the linear ZiG-DAG given the model parameters $\boldsymbol{\theta}$ and the DAG $\mathcal{G}$. We score each DAG by the Bayesian information criterion (BIC),

$$\mathrm{BIC}(\mathcal{G}|\boldsymbol{x}) = -2\sum_{i=1}^{n} \log p(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}, \mathcal{G}) + |\boldsymbol{\theta}| \log(n), \tag{7}$$

where $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate of the model parameters and $|\boldsymbol{\theta}|$ denotes the number of model parameters. As the individual DAG is identifiable for the proposed ZiG-DAGs, the consistency of the BIC ensures that the true DAG uniquely achieves the minimum BIC with probability converging to 1 as $n \to \infty$ (Claeskens et al., 2008). We take two strategies to minimize the BIC given by (7) with respect to the DAG $\mathcal{G}$: (1) exhaustive search and (2) greedy search.

Exhaustive Search For small graphs where the number of nodes d is small, the BIC can be minimized by computing the scores for all possible DAGs and find the DAG with the lowest score. This approach is exact and is useful for small d (say, $d \leq 4$). As the number of nodes d grows, however, this approach becomes computationally infeasible very quickly because the number of DAGs grows super-exponentially in d.

Greedy Search For larger graphs, exhaustive search is infeasible; we will use greedy search instead. Greedy search algorithms in the context of BN learning consider local moves from the current graph and makes the locally optimal choice at each iteration. We consider two strategies, hill climbing (HC) and tabu search (TS) algorithms.

The HC algorithm explores the neighborhood of the current DAG in the space of all possible DAGs. The neighborhood is defined using local moves. At each iteration, the algorithm scores all the DAGs that can be reached from the current graph by an edge addition, deletion, or reversal. The current DAG is then replaced by the DAG that provides the largest improvement, i.e., largest decrease in BIC in our case. We stop the algorithm if the improvement is no longer possible. We summarize the HC procedure in Algorithm 1. Although this algorithm finds a local optimal graph, there is no guarantee that the graph obtained by HC is a global optimum.

In order to avoid being trapped in local optima, the TS algorithm allows s additional local moves (edge addition, deletion, and reversal) when we reach a local optimal graph for which the score cannot be improved. These additional steps explore new territories around the local optimum even if they do not improve the score and may find new direction to arrive at a better structure. Note that the final solution should be the best DAG found anywhere during the search, not the DAG at which the algorithm stops. Furthermore, we keep a list (the tabu list) of all local moves that we have applied within the last t iterations. During the search over the neighborhood of the current DAG, our TS algorithm do not consider local modifications that reverse the local moves in the tabu list. For example, if we add an edge $j \to k$, we cannot delete the edge in the next t steps. This forces the search to explore new directions in the space of DAGs, instead of tweaking with the same parts of the current solution. Our TS algorithm is summarized in Algorithm 2.

Algorithm 1 Hill climbing

```
Input: data $\boldsymbol{x}$, initial DAG $\mathcal{G}_0$.
Compute BIC($\mathcal{G}_0|\boldsymbol{x}$) and set BIC$_{max}$ = BIC($\mathcal{G}_0|\boldsymbol{x}$).
Set $\mathcal{G}_{max} = \mathcal{G}_0$.
repeat
    Initialize $Improvement = false$.
    for all DAGs $\mathcal{G}'$ reachable from $\mathcal{G}_{max}$ by an edge addition, deletion, or reversal do
        Compute BIC($\mathcal{G}'|\boldsymbol{x}$).
        if BIC($\mathcal{G}'|\boldsymbol{x}$) < BIC$_{max}$ then
            Set $\mathcal{G}_{max} = \mathcal{G}'$ and BIC$_{max}$ = BIC($\mathcal{G}'|\boldsymbol{x}$).
            Set $Improvement = true$.
        end if
    end for
until $Improvment$ is $false$
Output: DAG $\mathcal{G}_{max}$.
```

Algorithm 2 Tabu search

```
Input: data $\boldsymbol{x}$, initial DAG $\mathcal{G}_0$, number of additionally allowed steps $s$, size of the tabu list $t$.
Compute BIC($\mathcal{G}_0|\boldsymbol{x}$) and set BIC$_{max}$ = BIC($\mathcal{G}_0|\boldsymbol{x}$).
Set $\mathcal{G}^* = \mathcal{G}_{max} = \mathcal{G}_0$.
Initialize $LastImprovement = 0$.
while $LastImprovement < s$ do
    Initialize BIC$^* = \infty$.
    for all DAGs $\mathcal{G}'$ reachable from $\mathcal{G}^*$ by an edge addition, deletion, or reversal do
        if $\mathcal{G}'$ does not reverse local moves in the tabu list, (i.e., in the last $t$ steps) then
            Compute BIC($\mathcal{G}'|\boldsymbol{x}$).
            if BIC($\mathcal{G}'|\boldsymbol{x}$) < BIC$^*$ then
                Set $\mathcal{G}^* = \mathcal{G}'$ and BIC$^*$ = BIC($\mathcal{G}'|\boldsymbol{x}$).
            end if
        end if
    end for
    if BIC$^*$ < BIC$_{max}$ then
        Set $\mathcal{G}_{max} = \mathcal{G}^*$ and BIC$_{max}$ = BIC$^*$.
        Set $LastImprovement = 0$.
    else
        Set $LastImprovement = LastImprovement + 1$.
    end if
end while
Output: DAG $\mathcal{G}_{max}$.
```

4.2 Structure Learning for Nonlinear ZiG-DAGs

While our identifiability theory for nonlinear ZiG-DAGs is general, for structure learning, we need to make specific choice of the nonlinear functions f_{jk} and g_{jk}. For example, one can expand f_{jk} and g_{jk} with the Fourier bases if the functional relationship is expected to be periodic. Similarly, if we expect that the relationship might show a very localized behavior, wavelets can be a good choice. In this paper, we employ spline basis expansion for f_{jk} and g_{jk}. Splines are popular in semiparametric function estimation because of the ease of their construction, their flexibility and accuracy to approximate a smooth function, and their interpretability through the representation by a compact set of basis functions and coefficients. Particularly, f_{jk} and g_{jk} are modeled by cubic B-splines,

$$f_{jk}(\cdot) = \sum_{l=1}^{M_f} \zeta_{jkl} B_{jkl}(\cdot) \quad \text{and} \quad g_{jk}(\cdot) = \sum_{l=1}^{M_g} \eta_{jkl} C_{jkl}(\cdot),$$

where $\{B_{jkl}(\cdot)\}_{l=1}^{M_f}$ and $\{C_{jkl}(\cdot)\}_{l=1}^{M_g}$ are cubic B-spline basis functions with some pre-specified knots. In summary, the nonlinear ZiG-DAG model is parameterized by spline coefficients ζ_{jkl}, η_{jkl} and the other node-specific model parameters $\mu_j, \nu_j, \boldsymbol{a}_j, \boldsymbol{b}_j$. The BIC for each DAG can be evaluated in the same way with (7) and we can use either exhaustive search or greedy search as in Section 4.1 for estimating the underlying graph for nonlinear ZiG-DAGs. The R implementation of the proposed method is available in the R package `ZiGDAG` (`https://github.com/junsoukchoi/ZiGDAG.git`).

5. Experiments

We empirically evaluate the causal discovery performance of both linear and nonlinear ZiG-DAG models with synthetic data. We compare the proposed method with state-of-the-art BN learning algorithms for count data: the overdispersion scoring (ODS) algorithm for Poisson BNs (Park and Raskutti, 2015) and the moments ratio scoring (MRS) algorithm for generalized hypergeometric BNs (Park and Park, 2019). We also consider the ZiDAG for zero-inflated Gaussian data (Yu et al., 2020) with the $\log(x+1)$ transformation of the synthetic count data.

5.1 Linear ZiG-DAG

We first consider a linear ZiG-DAG, where the conditional distribution of each node has a probability generating function given by (3) with $G\left(s; \boldsymbol{a}_j, \boldsymbol{b}_j, \lambda_j\right) = {}_1F_1\left(1; \psi_j, \lambda_j s\right) / {}_1F_1\left(1; \psi_j, \lambda_j\right)$. That is, the conditional distribution of each node follows a zero-inflated hyper-Poisson, which is a quite flexible distribution as the hyper-Posson distribution allows for both overdispersion and underdisperion in count data. We sample data from the linear ZiG-DAG with different sample sizes $n \in \{250, 500, 1000, 2000\}$ and different numbers of nodes $d \in \{10, 25, 50, 100\}$. For each simulation setting, we set the causal DAG $\mathcal{G}$ by randomly generating a sparse DAG with d edges. Given the DAG, we generate coefficients $(\alpha_{jk}, \beta_{jk})$ in (4) from independent uniform distributions: $\alpha_{jk} \sim \mathrm{U}(0.5, 2)$ and $\beta_{jk} \sim \mathrm{U}(-2, -0.5)$ for $k \in pa_{\mathcal{G}}(j)$ and $j \in \boldsymbol{V}$. The intercepts δ_j and γ_j in (4) are chosen uniformly at random from $(-1.5, 1)$ and $(1, 1.5)$, respectively. The additional parameters ψ_j for the GHPD (hyper-

Poisson distribution) are sampled as $\log(\psi_j) \sim \mathrm{U}(-2, 2)$. These ranges are chosen so that the resulting observations are not all zeros or do not have extremely large values. Each simulation setting is repeated 50 times, and the simulated datasets have $\sim 50\%$ zeros.

For ZiG-DAG, we implement both HC and TS algorithms as introduced in Section 4. Since they are greedy, initial values can affect the outcome. We consider two ways of initialization: first, we start HC (HC0) and TS (TS0) at the empty graph; and second, we initialize HC (HC1) and TS (TS1) with the DAGs obtained by MRS, which are expected to be better than empty graphs. To assess the causal discovery performance of each method, we calculate the true positive rate (TPR), the false discovery rate (FDR), and the Mattews correlation coefficient (MCC) for selection of true directed edges. MCC is a balanced measure of binary classification that takes a value between -1 and 1 with 1 indicating perfect agreement between the true and estimated graphs (i.e., perfect selection), 0 indicating random guess, and -1 indicating total disagreement.

We summarize in Tables 2-3 the operating characteristics of each method for different combinations of the sample size n and the number of nodes d. For every simulation setting, the proposed methods consistently outperform ODS, MRS, and ZiDAG. Specifically, as the sample size increases, our greedy search algorithms find the causal structure more accurately as expected. Our approaches also show satisfactory performance for various graph sizes including moderately large graphs ($d = 100$). We make additional observations for difference between the HC and TS algorithms. When the greedy search algorithms starts at the empty graph, i.e., HC0 and TS0, the performance of TS is better than that of HC. However, if we consider HC1 and TS1 for which we provide more informative initial DAG, there is no statistically significant difference between HC1 and TS1 in most cases. In subsequent simulations, for simplicity, we leave out HC1, TS0 and TS1, and only consider HC0 to learn the proposed ZiG-DAGs from data.

Model Misspecification When the conditional distribution of each node in a ZiG-DAG model is misspecified, our identifiability theories do not guarantee that we can find the true DAG. Therefore, an important question is how well our algorithms recovers the true graph when misspecified distributions are used. We investigate this empirically. We choose the simulation scenario with $n = 1000$ and $d = 50$, and apply two different linear ZiG-DAG models. The first one is a linear ZiG-DAG (ZiG-DAG-HP) using the zero-inflated hyper-Poisson distribution as above. The second one is another linear ZiG-DAG (ZiG-DAG-NB) where the conditional distribution of each node is assumed to follow a zero-inflated negative binomial distribution. In ZiG-DAG-NB, every conditional distribution is misspecified, as the true data-generating model is ZiG-DAG-HP. The simulation results are shown in Figure 2. Both ZiG-DAG-HP and ZiG-DAG-NB are better than ODS, MRS, and ZiDAG. Although ZiG-DAG-NB is a misspecified model, its performance is still better than the alternative state-of-the-art approaches. This shows that the proposed ZiG-DAG is useful for learning the true causal structure even if the true conditional distributions are misspecified.

5.2 Nonlinear ZiG-DAG

We next assess the performance of the nonlinear ZiG-DAG models. We sample data from a nonlinear ZiG-DAG with $n = 500$ and $d = 10$, where the conditional distribution of each node is again assumed to be a zero-inflated hyper-Poisson. We randomly choose the true

Table 2: Linear ZiG-DAG. Average operating characteristics over 50 simulations for different sample sizes $n \in \{250, 500, 1000, 2000\}$ with $d = 50$. The standard error for each statistic is given within parentheses.

		Sample size, n			
Method	Measure	250	500	1000	2000
HC0	TPR	0.728 (0.009)	0.844 (0.009)	0.924 (0.008)	0.946 (0.007)
	FDR	0.387 (0.009)	0.226 (0.009)	0.125 (0.009)	0.083 (0.009)
	MCC	0.660 (0.008)	0.804 (0.009)	0.897 (0.009)	0.930 (0.008)
HC1	TPR	0.802 (0.009)	0.892 (0.007)	0.958 (0.005)	0.971 (0.004)
	FDR	0.354 (0.008)	0.203 (0.008)	0.101 (0.008)	0.074 (0.007)
	MCC	0.713 (0.008)	0.840 (0.007)	0.926 (0.007)	0.947 (0.005)
TS0	TPR	0.739 (0.009)	0.862 (0.009)	0.932 (0.007)	0.956 (0.006)
	FDR	0.391 (0.009)	0.222 (0.009)	0.121 (0.009)	0.074 (0.008)
	MCC	0.663 (0.009)	0.815 (0.008)	0.903 (0.008)	0.939 (0.007)
TS1	TPR	0.798 (0.009)	0.889 (0.008)	0.953 (0.006)	0.971 (0.004)
	FDR	0.369 (0.009)	0.214 (0.009)	0.110 (0.008)	0.073 (0.007)
	MCC	0.702 (0.009)	0.832 (0.008)	0.919 (0.007)	0.948 (0.005)
ODS	TPR	0.418 (0.006)	0.454 (0.004)	0.474 (0.005)	0.474 (0.003)
	FDR	0.710 (0.005)	0.726 (0.004)	0.753 (0.004)	0.776 (0.003)
	MCC	0.331 (0.005)	0.335 (0.004)	0.323 (0.004)	0.305 (0.003)
MRS	TPR	0.662 (0.006)	0.755 (0.004)	0.809 (0.004)	0.816 (0.003)
	FDR	0.467 (0.006)	0.454 (0.005)	0.464 (0.004)	0.505 (0.003)
	MCC	0.585 (0.006)	0.633 (0.004)	0.650 (0.004)	0.626 (0.003)
ZiDAG	TPR	0.619 (0.010)	0.710 (0.009)	0.756 (0.007)	0.778 (0.007)
	FDR	0.291 (0.010)	0.243 (0.009)	0.243 (0.008)	0.252 (0.008)
	MCC	0.656 (0.010)	0.727 (0.009)	0.751 (0.008)	0.758 (0.008)

nonlinear functions f_{jk} and g_{jk} in (5) from three candidates, respectively:

$$f_1(z) = \frac{1}{2} z (z - 3) \, , \; f_2(z) = \sin (z) \, , \; f_3(z) = \exp \left(\frac{1}{2} z - 1 \right) ,$$

and

$$g_1(z) = -\frac{1}{2} \left(z - \frac{3}{2} \right)^2 , \; g_2(z) = \cos (z) \, , \; g_3(z) = -\frac{1}{2} \log (z + 1) \, .$$

The intercepts μ_j, ν_j in (5) and the additional parameters ψ_j for the hyper-Poisson are generated as in Section 5.1: $\mu_j \sim \mathrm{U}(-1.5, -1), \nu_j \sim \mathrm{U}(1, 1.5)$, and $\log(\psi_j) \sim \mathrm{U}(-2, 2)$.

Table 3: Linear ZiG-DAG. Average operating characteristics over 50 simulations for different numbers of nodes $d \in \{10, 25, 50, 100\}$ with $n = 1000$. The standard error for each statistic is given within parentheses.

		Number of nodes, d			
Method	Measure	10	25	50	100
	TPR	0.948 (0.012)	0.891 (0.013)	0.924 (0.008)	0.864 (0.007)
HC0	FDR	0.067 (0.015)	0.166 (0.019)	0.125 (0.009)	0.255 (0.008)
	MCC	0.932 (0.015)	0.855 (0.017)	0.897 (0.009)	0.800 (0.007)
	TPR	0.912 (0.014)	0.977 (0.005)	0.958 (0.005)	0.870 (0.006)
HC1	FDR	0.141 (0.021)	0.060 (0.009)	0.101 (0.008)	0.269 (0.008)
	MCC	0.869 (0.020)	0.956 (0.007)	0.926 (0.007)	0.795 (0.007)
	TPR	0.964 (0.010)	0.925 (0.011)	0.932 (0.007)	0.869 (0.006)
TS0	FDR	0.051 (0.013)	0.130 (0.017)	0.121 (0.009)	0.254 (0.008)
	MCC	0.951 (0.013)	0.892 (0.015)	0.903 (0.008)	0.803 (0.007)
	TPR	0.932 (0.014)	0.969 (0.005)	0.953 (0.006)	0.874 (0.007)
TS1	FDR	0.103 (0.020)	0.081 (0.009)	0.110 (0.008)	0.267 (0.009)
	MCC	0.902 (0.019)	0.941 (0.007)	0.919 (0.007)	0.798 (0.008)
	TPR	0.386 (0.010)	0.419 (0.008)	0.474 (0.005)	0.543 (0.004)
ODS	FDR	0.677 (0.009)	0.775 (0.006)	0.753 (0.004)	0.761 (0.003)
	MCC	0.262 (0.010)	0.265 (0.007)	0.323 (0.004)	0.350 (0.003)
	TPR	0.742 (0.010)	0.874 (0.009)	0.809 (0.004)	0.733 (0.004)
MRS	FDR	0.331 (0.011)	0.423 (0.009)	0.464 (0.004)	0.623 (0.002)
	MCC	0.664 (0.010)	0.695 (0.009)	0.650 (0.004)	0.519 (0.003)
	TPR	0.640 (0.017)	0.700 (0.012)	0.756 (0.007)	0.794 (0.006)
ZiDAG	FDR	0.295 (0.016)	0.368 (0.016)	0.243 (0.008)	0.254 (0.007)
	MCC	0.632 (0.018)	0.649 (0.015)	0.751 (0.008)	0.768 (0.006)

For learning the nonlinear ZiG-DAG, we use $M_f = M_g = 4$ spline basis with a knot being placed at the 50% quantile of the data. We also consider the linear ZiG-DAG for comparison. Additionally, since ZiDAG allows for both linear and nonlinear causal relationships, in this simulation study, we use ZiDAG with nonlinear implementation.

We report in Table 4 the simulation results based on 50 repetitions. Overall, the nonlinear ZiG-DAG outperforms the other approaches including the linear ZiG-DAG. Especially, the nonlinear ZiG-DAG results in extremely low FDR compared to the other competitors.

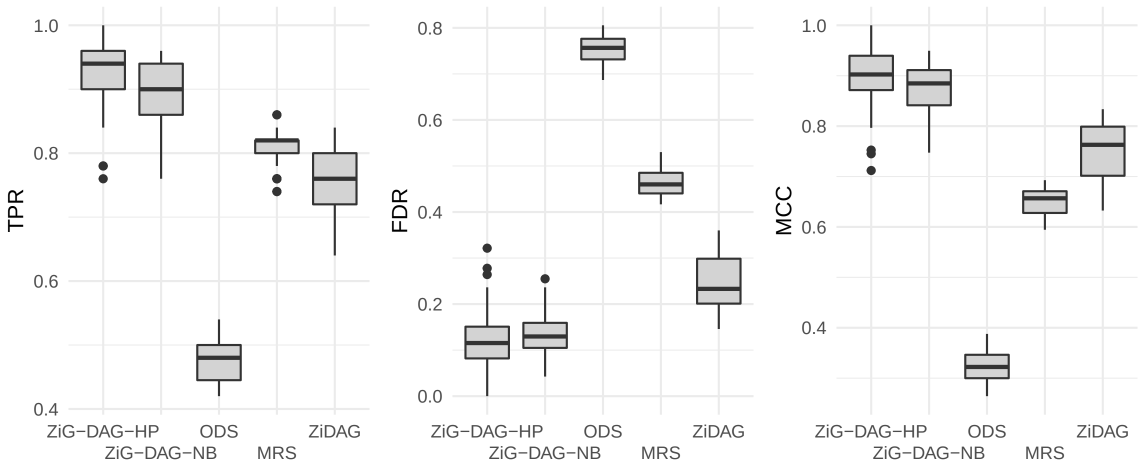

Figure 2: Box plots of operating characteristics for ZiG-DAG-HP, ZiG-DAG-NB, ODS, MRS, and ZiDAG applied to synthetic datasets generated from a linear ZiG-DAG with $n = 1000$ and $d = 50$.

Not surprisingly, in this nonlinear simulation setting, ZiDAG gives better results than the linear ZiG-DAG.

Table 4: Nonlinear ZiG-DAG. Average operating characteristics over 50 simulations with $n = 500$ and $d = 10$. The standard error for each statistic is given within parentheses.

	Nonlinear ZiG-DAG	Linear ZiG-DAG	ODS	MRS	ZiDAG
TPR	0.622 (0.014)	0.662 (0.018)	0.614 (0.008)	0.588 (0.020)	0.568 (0.013)
FDR	0.179 (0.018)	0.377 (0.017)	0.417 (0.010)	0.405 (0.022)	0.249 (0.014)
MCC	0.684 (0.017)	0.596 (0.020)	0.546 (0.007)	0.540 (0.023)	0.616 (0.014)

5.3 Non-zero-inflation

Although the proposed ZiG-DAG models are primarily developed to deal with excessive zeros in count data, they are also applicable and robust to count data generated from non-zero-inflated distributions. We perform additional simulations to support this claim. We generate data from a negative binomial BN, which does not include any zero-inflation components. The parameters for the negative binomial BN are sampled uniformly at random in a similar way to Section 5.1. The resulting data have ∼26% zeros, which are much less than the zero-inflated case. As in Section 5.1, we consider ZiG-DAG-HP and ZiG-DAG-NB that assume a zero-inflated hyper-Poisson distribution and a zero-inflated negative binomial distribution for the conditional distribution of each node, respectively. Furthermore, we consider two distinctive MRS algorithms that learn DAGs for hyper-Poisson BNs (MRS-HP) and negative binomial BNs (MRS-NB). Since MRS-NB requires an input of the dispersion parameter of the negative binomial distribution, we provide it with the true dispersion parameter value.

The simulation results are shown in Table 5. Even though the data are not zero-inflated, our approaches, ZiG-DAG-HP and ZiG-DAG-NB, generally show better performance than the alternative methods (ODS, MRS-HP, MRS-NB, and ZiDAG). Even though MRS-NB uses the correct distributional model and the true dispersion parameter, it shows worse performance than our methods as well as ZiDAG with respect to FDR and MCC. This might be because the performance of the MRS algorithm highly relies on the choice of external methods for the skeleton estimation. In our experiments, MRS utilizes the R package MXM to estimate the skeleton of DAG, which might provide unreliable skeleton estimates in this simulation setting.

Table 5: Non-zero-inflation. Average operating characteristics over 50 simulations for a negative binomial BN with $n = 500$ and $d = 50$. The standard error for each statistic is given within parentheses.

	ZiG-DAG-HP	ZiG-DAG-NB	ODS	MRS-HP	MRS-NB	ZiDAG
TPR	0.692 (0.009)	0.774 (0.007)	0.306 (0.004)	0.637 (0.008)	0.714 (0.008)	0.582 (0.008)
FDR	0.342 (0.010)	0.334 (0.008)	0.658 (0.005)	0.514 (0.009)	0.455 (0.009)	0.267 (0.010)
MCC	0.668 (0.009)	0.712 (0.007)	0.310 (0.004)	0.546 (0.008)	0.615 (0.008)	0.647 (0.009)

5.4 Latent Confounders

Recall that Theorems 6 and 7 in Section 3 assume causal sufficiency (Condition 3), that is, there exist no latent confounders. Although the causal sufficiency assumption is common in the causal literature, in real applications, it is difficult to check whether an unmeasured latent confounder exists, and there is always a possibility that we do not observe some variables of interest. Therefore, we test how sensitive our method is to the existence of latent confounders. We consider two true causal DAGs in Figure 3 that have three nodes, X_1, X_2, X_3. Given each causal graph, we generate zero-inflated count data from a linear ZiG-DAGs and treat X_3 as an unmeasured confounder (i.e., hide it from the algorithms).

The graph in Figure 3(a) assumes a casual effect of X_2 on X_1, which is confounded by X_3. For the simulation truth corresponding to Figure 3(a), we assume that the conditional distribution of each node is a zero-inflated hyper-Poisson similarly to Section 5.1. We denote $\boldsymbol{c} = (\alpha_{13}, \beta_{13}) = (\alpha_{23}, \beta_{23})$ and $\boldsymbol{d} = (\alpha_{12}, \beta_{12})$, and set $\delta_j = -1, \gamma_j = 0$ and $\psi_j = 5$. We consider different levels of confounding effects $\boldsymbol{c} = \sigma \times (-0.8, 0.8)$ where $\sigma \in \{0, 0.1, 0.2, \ldots, 1\}$, while fixing the causal effect $\boldsymbol{d} = (0.8, -0.8)$. For each level of confounding effect, we simulate 50 datasets with sample size $n = 250$. Figure 4(a) plots the average accuracy (ACC) over 50 repeat simulations of ZiG-DAG for identifying the true causal direction $X_2 \to X_1$. We also consider MRS and ZiDAG as benchmarks. Our approach finds the true causal direction quite well across the confounding levels, while ZiDAG becomes worse when the confounding effect is relatively large ($\sigma > 0.5$). MRS does not work well in this case.

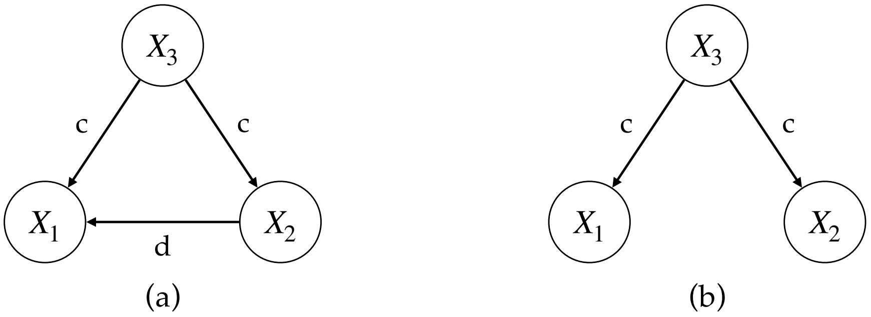

Figure 3: Two different confounding scenarios with a confounder X_3.

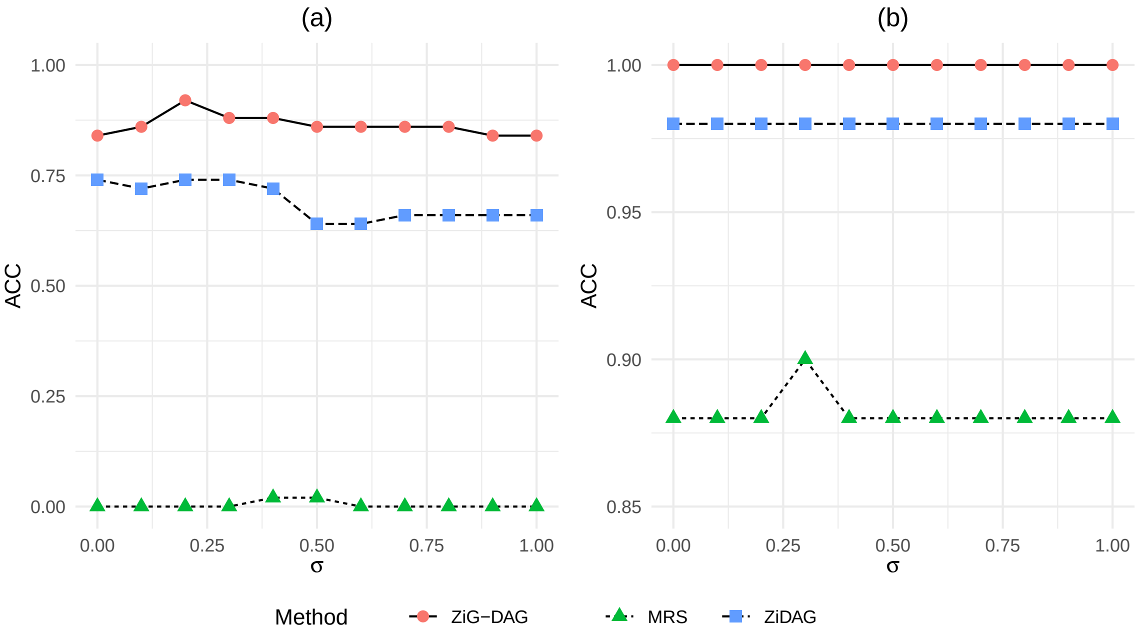

Figure 4: Plots of average ACC of ZiG-DAG, MRS, and ZiDAG against different levels of confounding effect σ under the confounding scenarios of (a) Figure 3(a) and (b) Figure 3(b).

Next, we consider the DAG in Figure 3(b). There is no causal effect between X_1 and X_2 whereas the confounding effect by X_3 is still present. Therefore, we set $\boldsymbol{d} = \mathbf{0}$; otherwise the same simulation truth with Figure 3(a) is used. We consider the same confounding effects with Figure 3(a). Figure 4(b) displays the resulting ACCs of ZiG-DAG, MRS, and ZiG-DAG, again averaged over 50 repeat simulations. In the range of the confounding level being considered, ZiG-DAG does not add any spurious causal relation between X_1 and X_2. In summary, the empirical results in Figure 4(a)-(b) indicate that the proposed ZiG-DAG is relatively robust to the presence of hidden confounders.

6. Real Data Analyses

We illustrate the utility of the proposed ZiG-DAG by performing two analyses of a scRNA-seq dataset (Li et al., 2017) that consists of 561 cells from 11 primary colorectal cancer (CRC) tumors and matched normal mucosa.

6.1 Real-Data Validation with Known Causal Relationships

Using the real scRNA-seq data and known causal relationships in the biological literature, we validate the causal identifiability of ZiG-DAG and compare it to other state-of-the-art alternatives. First, from the TRRUST database (Han et al., 2018), we extract a list of literature-curated pairs of transcription factor and its target. This list establishes a biological ground truth of cause-and-effect relationships with the transcription factors being causes and the targets being effects. We extract from our scRNA-seq data the pairs of genes on the list for which the maximum information coefficient (Reshef et al., 2011), a measure of linear and nonlinear correlations between two variables, is greater than 0.5. This results in 47 pairs for validation.

We apply the proposed ZiG-DAG to each pair of genes. Specifically, we use a nonlinear ZiG-DAG where the conditional distribution of each node is a zero-inflated hyper-Poisson distribution. For comparison, we apply MRS and ZiDAG to the same dataset. We calculate the accuracy of identifying true causal relationships for ZiG-DAG, MRS, and ZiDAG, and the results are 60%, 51%, and 53%, respectively. Out of a total of 47 pairs, the proposed ZiG-DAG correctly identifies 28 causal relationships. This indicates that the proposed method is capable of finding true causal relationships in real data: the p-value for a binomial test is 0.0002 when compared to random guesses. Furthermore, among the three count BNs, ZiG-DAG has the highest accuracy.

6.2 Reverse Engineering of Gene Regulatory Network

In this section we aim to reconstruct a gene regulatory network for $d = 26$ genes from the TGF-β signaling pathway, which has been shown as the most activated signaling pathway in the analysis of Li et al. (2017). Before reconstructing the gene regulatory network, we filter cell doublets and multiplets using an `R` package for single cell genomics, Seurat (Hao et al., 2021), and retain 472 cells, which contain $\sim 40\%$ zeros. To estimate the gene regulatory network, we use the HC algorithm for the nonlinear ZiG-DAG that assumes a zero-inflated hyper-Poisson distribution as the conditional distribution of each node. We initialize the algorithm with the DAG obtained by MRS, which shows promising performance in the experiments of Section 5.1.

Figure 5 displays the estimated gene regulatory network for $d = 26$ genes of the TGF-β signaling pathway. In total, 26 directed edges are found by our nonlinear ZiG-DAG. Some of the estimated gene regulations are consistent with known regulatory relationships in the existing biological literature. For example, the proposed model finds gene regulations involving SMAD proteins, which are main signal transducers for receptors of the TGF-β superfamily. Specifically, `SMAD2` affects mRNA profiles of `ZFYVE9` (Runyan et al., 2009) and `RBX1` regulates the `SMAD4` protein stability (Inoue and Imamura, 2008). Moreover, the

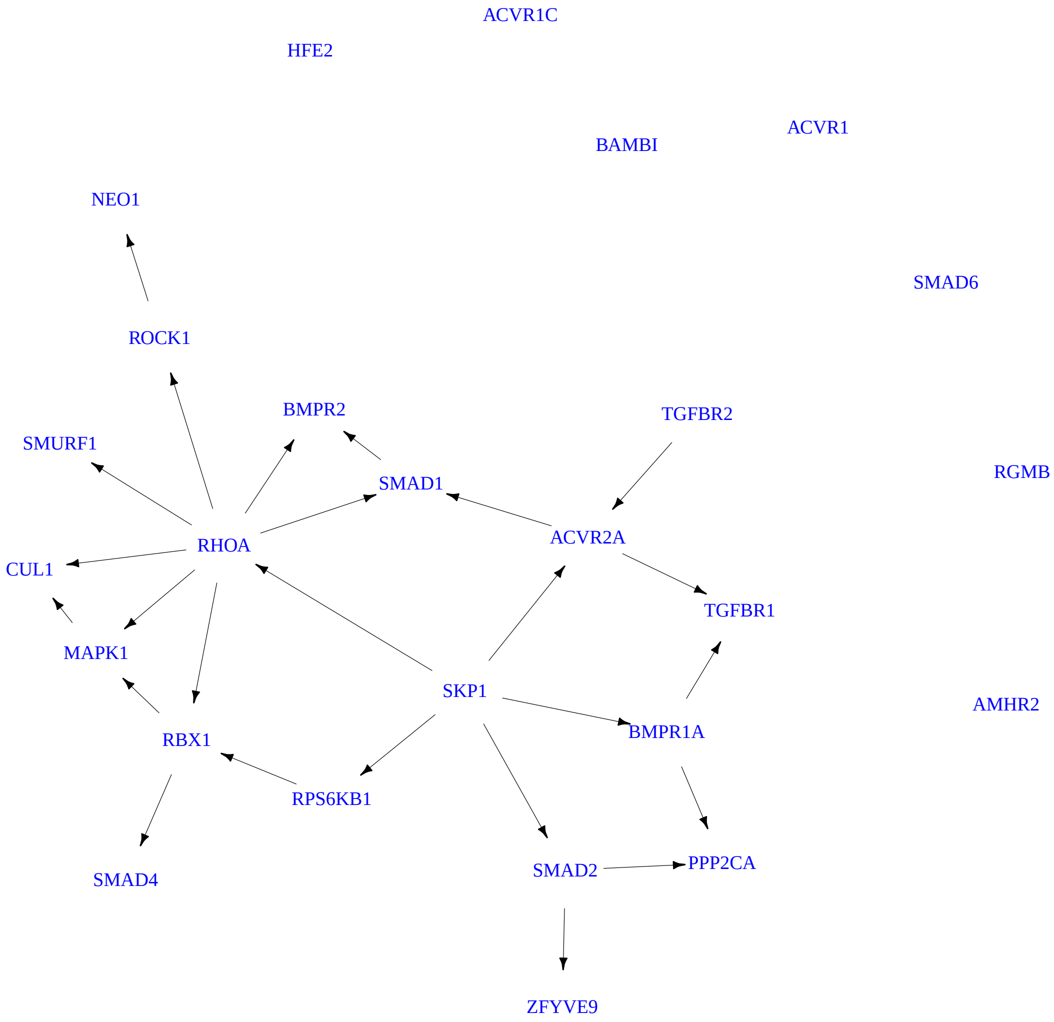

Figure 5: The estimated gene regulatory network for $d = 26$ genes of the TGF-β signaling pathway using the nonlinear ZiG-DAGs.

estimated network also confirms the fact that `ROCK` is a well-known downstream effector of `RHOA`.

Furthermore, we can find 2 hub genes in the estimated network: `RHOA` and `SKP1` with out-degrees of 7 and 5. Hub genes are of particular importance because they are often involved in essential regulatory relationships. In fact, the importance of our hub genes in TGF-β signaling has been supported by the existing literature. `RHOA` is a small GTPase of the RHO family, whose inactivation plays key roles in colorectal cancer progression/metastasis by interacting with many members of TGF-β signaling pathway (Rodrigues et al., 2014; Dopeso et al., 2018). `SKP1` belongs to the SCF complex, which is a RING-type E3 ubiquitin ligase that participates in the degradation of a wide variety of proteins that regulates TGF-β signaling (Inoue and Imamura, 2008).

7. Discussion

We have proposed a novel BN model, ZiG-DAG, to infer causal relationships in observational zero-inflated count data. ZiG-DAGs are built upon a fairly general class of count distributions, namely generalized hypergeometric probability distributions, and therefore can account for various types of zero-inflated count data including overdispersed or underdispersed zero-inflated count data. We have also considered not only linear causal relationships but also nonlinear relationships. The identifiability theory for the proposed ZiG-DAGs has been established using a general proof technique, which can potentially be used to show identifiability of other discrete BN models. The proposed ZiG-DAG models are paired with two structure learning procedures, exhaustive search and greedy search. Through extensive numerical experiments and real data analysis, we have empirically validated the identifiability theory for ZiG-DAGs and have shown its superior performance against state-of-the-art alternatives.

There are a few future research directions that can be taken. First, the proposed approach can be extended for modeling interventional zero-inflated count data. This may be done by modifying the likelihood according to the *do*-calculus framework of Pearl (2009). The second direction is to establish an identifiability theory in the presence of latent confounders. Although we have empirically shown in Section 5.4 that ZiG-DAG is relatively robust against confounding, we do not yet have theoretical support of it; the proofs of our identifiability theorems are not directly applicable as they rely on the factorization (1), which requires causal sufficiency. Third, the acyclicity of BNs may be restrictive in applications where the underlying systems have feedback loops, for example, genetic systems. We may relax this acyclicity restriction by using directed cyclic graphs.

Acknowledgments

Ni's research is partially supported by NSF DMS-1918851, NSF DMS-2112943, and NIH 1R01GM148974-01.

Appendix A. Proof of Propostion 4

First, we state a lemma on the conditions under which conditional distributions of the same node are identical in two discrete BNs, which is needed for the proof of Proposition 4.

Lemma 8 *Suppose that $\mathcal{B} = (\mathcal{G}, p)$ and $\mathcal{B}^* = (\mathcal{G}^*, p^*)$ are any two discrete BNs, and $(1, 2, \dots, d)$ is the topological ordering of the DAG $\mathcal{G}$ of $\mathcal{B}$. If for $j' \in \boldsymbol{V}$, we have $pa(j') = pa^*(j')$, $ch^*(j') \cap nd(j') = \emptyset$, and*

$$\prod_{j=1}^{j'} p(x_j|\boldsymbol{x}_{pa(j)}) = \prod_{j=1}^{j'} p^*(x_j|\boldsymbol{x}_{pa^*(j)}), \tag{8}$$

then the conditional distribution of node j' is the same in $\mathcal{B}$ and $\mathcal{B}^*$, i.e., $p(x_{j'}|\boldsymbol{x}_{pa(j')}) = p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')})$.

Proof Since $(1, \dots, d)$ is the topological ordering of $\mathcal{G}$ and $ch^*(j') \cap nd(j') = \emptyset$, node j' cannot be a parent of nodes $\{1, 2, \dots, j'-1\}$ in both $\mathcal{B}$ and $\mathcal{B}^*$, and hence in (8), $p(x_j|\boldsymbol{x}_{pa(j)})$ and $p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ for $j = 1, \dots, j'-1$ are functions of all the variables $x_1, \dots, x_d$ but $x_{j'}$. The only terms in (8) that depends on $x_{j'}$ are $p(x_{j'}|\boldsymbol{x}_{pa(j')})$ and $p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')})$. Furthermore, both $p(x_{j'}|\boldsymbol{x}_{pa(j')})$ and $p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')})$ are functions of $x_{j'}$ and $\boldsymbol{x}_{pa^\cap(j')}$, where we denote $pa^\cap(j) = pa(j') = pa^*(j')$. Let $x_1, \dots, x_{j'-1}, x_{j'+1}, \dots, x_d$ be fixed. Then, (8) is simplified as

$$Cp(x_{j'}|\boldsymbol{x}_{pa(j')}) = C^* p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')}) \tag{9}$$

where $C = \prod_{j=1}^{j'-1} p(x_j|\boldsymbol{x}_{pa(j)})$ and $C^* = \prod_{j=1}^{j'-1} p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ are constants not depending on $x_{j'}$. If we sum up (9) over all possible $x_{j'}$, we have

$$C \sum_{x_{j'}} p(x_{j'}|\boldsymbol{x}_{pa(j')}) = C^* \sum_{x_{j'}} p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')}).$$

Due to the fact that $\sum_{x_{j'}} p(x_{j'}|\boldsymbol{x}_{pa(j')}) = 1$ and $\sum_{x_{j'}} p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')}) = 1$, we obtain $C = C^*$, and since $x_1, \dots, x_d$ are arbitrary, it follows that $p(x_{j'}|\boldsymbol{x}_{pa(j')}) = p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')})$ for any possible $x_{j'}$ and $\boldsymbol{x}_{pa^\cap(j)}$. ■

We now provide the proof of Proposition 4. We show that if the joint distributions p and p^* are equivalent, then the identity of the causal structures $\mathcal{G}$ and $\mathcal{G}^*$ automatically follows from the proposition assumption.

Proof We assume that the joint distributions of $\mathcal{B}$ and $\mathcal{B}^*$ are the same, i.e.,

$$\prod_{j=1}^{d} p(x_j|\boldsymbol{x}_{pa(j)}) = \prod_{j=1}^{d} p^*(x_j|\boldsymbol{x}_{pa^*(j)}) \tag{10}$$

for all possible values of $x_1, \dots, x_d$. Without loss of generality, assume that $(1, \dots, d)$ is the topological ordering of the DAG $\mathcal{G}$ of $\mathcal{B}$, i.e., the nodes are labeled such that there is no directed edge in $\mathcal{G}$ from later nodes to earlier nodes. Such orderings must exist (although not necessarily unique) because of the acyclicity of DAGs. We then show by mathematical induction that $pa(j) = pa^*(j)$ for all $j = 1, 2, \ldots, d$ (hence $\boldsymbol{E} = \boldsymbol{E}^*$), which contradicts our assumption that $\boldsymbol{E} \neq \boldsymbol{E}^*$.

We begin with the last node d that has no child in the graph $\mathcal{G}$. Taking the ratio of (10) at $x_d + 1$ and x_d, we obtain

$$\frac{p(x_d + 1|\boldsymbol{x}_{pa(d)})}{p(x_d|\boldsymbol{x}_{pa(d)})} = \frac{p^*(x_d + 1|\boldsymbol{x}_{pa^*(d)})}{p^*(x_d|\boldsymbol{x}_{pa^*(d)})} \prod_{k \in ch^*(d)} \frac{p^*(x_k|\boldsymbol{x}_{pa^*(k)\setminus\{d\}}, x_d + 1)}{p^*(x_k|\boldsymbol{x}_{pa^*(k)\setminus\{d\}}, x_d)}$$

for all $x_1, \ldots, x_d$. Note that this implicitly assumes that the conditional distributions $p(x_j|\boldsymbol{x}_{pa(j)})$ and $p^*(x_j|\boldsymbol{x}_{pa^*(j)})$, $j = 1, \ldots, p$, are positive over their supports. The above ratio can be rewritten using probability generating functions as follows:

$$\frac{G_d^{(x_d+1)}(0; \boldsymbol{x}_{pa(d)})}{G_d^{(x_d)}(0; \boldsymbol{x}_{pa(d)})} = \frac{G_d^{*(x_d+1)}(0; \boldsymbol{x}_{pa^*(d)})}{G_d^{*(x_d)}(0; \boldsymbol{x}_{pa^*(d)})} \prod_{k \in ch^*(d)} \frac{G_k^{*(x_k)}(0; \boldsymbol{x}_{pa^*(k)\setminus\{d\}}, x_d + 1)}{G_k^{*(x_k)}(0; \boldsymbol{x}_{pa^*(k)\setminus\{d\}}, x_d)}.$$

The proposition assumption then indicates $pa(d) = pa^*(d)$ and $ch^*(d) = ch^*(d) \cap nd(d) = \emptyset$. Moreover, Lemma 8 implies that $p(x_d|\boldsymbol{x}_{pa(d)}) = p^*(x_d|\boldsymbol{x}_{pa^*(d)})$.

Now assume that for any $j = j'+1, \ldots, d$, it holds that $pa(j) = pa^*(j)$, $ch^*(j) \cap nd(j) = \emptyset$ and $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$. We then show that it also holds for $j = j'$. First, observe that the ratio of (10) at $x_{j'} + 1$ and $x_{j'}$ is given by

$$\begin{aligned} &\frac{p(x_{j'} + 1|\boldsymbol{x}_{pa(j')})}{p(x_{j'}|\boldsymbol{x}_{pa(j')})} \prod_{k \in ch(j')} \frac{p(x_k|\boldsymbol{x}_{pa(k)\setminus\{j'\}}, x_{j'} + 1)}{p(x_k|\boldsymbol{x}_{pa(k)\setminus\{j'\}}, x_{j'})} \\ &= \frac{p^*(x_{j'} + 1|\boldsymbol{x}_{pa^*(j')})}{p^*(x_{j'}|\boldsymbol{x}_{pa^*(j')})} \prod_{k \in ch^*(j')} \frac{p^*(x_k|\boldsymbol{x}_{pa^*(k)\setminus\{j'\}}, x_{j'} + 1)}{p^*(x_k|\boldsymbol{x}_{pa^*(k)\setminus\{j'\}}, x_{j'})}. \end{aligned} \tag{11}$$

Note that the induction assumption implies $ch(j') = ch^*(j') \setminus nd(j')$ and

$$\frac{p(x_k|\boldsymbol{x}_{pa(k)\setminus\{j'\}}, x_{j'} + 1)}{p(x_k|\boldsymbol{x}_{pa(k)\setminus\{j'\}}, x_{j'})} = \frac{p^*(x_k|\boldsymbol{x}_{pa^*(k)\setminus\{j'\}}, x_{j'} + 1)}{p^*(x_k|\boldsymbol{x}_{pa^*(k)\setminus\{j'\}}, x_{j'})}$$

for $k \in ch(j') = ch^*(j') \setminus nd(j')$. Therefore, we can simplify (11) into a similar form of the case $j = d$:

$$\frac{G_{j'}^{(x_{j'}+1)}(0; \boldsymbol{x}_{pa(j')})}{G_{j'}^{(x_{j'})}(0; \boldsymbol{x}_{pa(j')})} = \frac{G_{j'}^{*\,(x_{j'}+1)}(0; \boldsymbol{x}_{pa^*(j')})}{G_{j'}^{*\,(x_{j'})}(0; \boldsymbol{x}_{pa^*(j')})} \prod_{k \in ch^*(j') \cap nd(j')} \frac{G_k^{*(x_k)}(0; \boldsymbol{x}_{pa^*(k)\setminus\{j'\}}, x_{j'} + 1)}{G_k^{*(x_k)}(0; \boldsymbol{x}_{pa^*(k)\setminus\{j'\}}, x_{j'})}.$$

It again follows from the proposition assumption and Lemma 8 that $pa(j') = pa^*(j')$, $ch^*(j') \cap nd(j') = \emptyset$, and $p(x_{j'}|\boldsymbol{x}_{pa(j')}) = p^*(x_{j'}|\boldsymbol{x}_{pa'(j')})$, which completes the proof. ■

Appendix B. Proof of Corollary 5

Proof Due to Proposition 4 and its assumption, the equivalence of the joint distributions of $\mathcal{B}$ and $\mathcal{B}^*$ (i.e., $p = p^*$) implies $\boldsymbol{E} = \boldsymbol{E}^*$, which in turn implies $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ for any $j \in \boldsymbol{V}$. Then by the corollary assumption, $\boldsymbol{\Xi}_j = \boldsymbol{\Xi}_j^*$ for any $j \in \boldsymbol{V}$. Hence, $\boldsymbol{\Xi} = \boldsymbol{\Xi}^*$. ■

Appendix C. Proof of Theorem 6

Proof We use Proposition 4 to prove that the DAG is identifiable for the linear ZiG-DAGs. Furthermore, we use Corollary 5 to show that the model parameters are also identifiable up to permutations of $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$, given that (p_j, q_j), which characterize the generalized hypergeometric function, and the link function h_j are fixed for each $j \in \boldsymbol{V}$. Let $\mathcal{B}$ and $\mathcal{B}^*$ be two arbitrary linear ZiG-DAGs and we show that they satisfy the sufficient condition of Proposition 4. Let $j \in \boldsymbol{V}$ be a node such that the identity (6) holds. We show that $pa(j) = pa^*(j)$ and $ch^*(j) \cap nd(j) = \emptyset$. We use the superscript $*$ to indicate parameters that define the linear ZiG-DAG $\mathcal{B}^*$.

First, we show $ch^*(j) \cap nd(j) = \emptyset$. Suppose by way of contradiction that $ch^*(j) \cap nd(j) \neq \emptyset$, and let $k' \in ch^*(j) \cap nd(j)$ such that $ch^*(k') \cap ch^*(j) \cap nd(j) = \emptyset$; such k' always exists due to the acyclicity of $\mathcal{G}^*$. For (6), let $x_j = 0$ while fixing x_l for $l \in \boldsymbol{V} \setminus \{j, k'\}$. Then, if $k' \notin pa(j)$, it is simplified to

$$C_1 = \begin{cases} C_2 \frac{1+\exp(\alpha^*_{k'j} + \tilde{\delta}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(\beta^*_{k'j} + \tilde{\gamma}^*_{k'}))}{1+\exp(\tilde{\delta}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(\tilde{\gamma}^*_{k'}))} & \text{for} \quad x_{k'} = 0 \\ C_2 r^{x_{k'}} & \text{for} \quad x_{k'} \neq 0, \end{cases}$$

and if $k' \in pa(j)$, it takes the following form:

$$\begin{aligned} & \frac{C_3 H_j(\beta_{jk'} x_{k'} + \tilde{\gamma}_j)}{1 + \exp(\alpha_{jk'} x_{k'} + \tilde{\delta}_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(\beta_{jk'} x_{k'} + \tilde{\gamma}_j))} \\ & = \begin{cases} C_2 \frac{1+\exp(\alpha^*_{k'j} + \delta^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(\beta^*_{k'j} + \gamma^*_{k'})))}{1+\exp(\delta^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(\gamma^*_{k'})))} & \text{for} \quad x_{k'} = 0 \\ C_2 r^{x_{k'}} & \text{for} \quad x_{k'} \neq 0, \end{cases} \end{aligned}$$

where $H_j = h_j^{-1}$, $H^*_{k'} = (h^*_{k'})^{-1}$, $\tilde{\delta}_j = \sum_{l \in pa(j) \setminus \{k'\}} \alpha_{jl} x_l + \delta_j$, $\tilde{\gamma}_j = \sum_{l \in pa(j) \setminus \{k'\}} \beta_{jl} x_l + \gamma_j$, $\tilde{\delta}^*_{k'} = \sum_{l \in pa^*(k') \setminus \{j\}} \alpha^*_{k'l} x_l + \delta^*_{k'}$, $\tilde{\gamma}^*_{k'} = \sum_{l \in pa^*(k') \setminus \{j\}} \beta^*_{k'l} x_l + \gamma^*_{k'}$, and

$$r = \frac{H^*_{k'}(\beta^*_{k'j} + \tilde{\gamma}^*_{k'})}{H^*_{k'}(\tilde{\gamma}^*_{k'})}.$$

Here, C_1, C_2, C_3 are some constants not depending on $x_{k'}$. The above identities are well defined since each conditional distribution (or equivalently, the derivatives of the probability generating function) should be positive over the entire support in our definition of the linear ZiG-DAG. Since $x_{k'}$ is not a binary variable (i.e., it takes integers beyond $\{0, 1\}$), taking the ratio of each of the above equations at $x_{k'}+1$ and $x_{k'}$, we observe that if $k' \notin pa(j)$,

$$r = \frac{1+\exp(\alpha^*_{k'j}+\tilde{\delta}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'};\boldsymbol{b}^*_{k'};H^*_{k'}(\beta^*_{k'j}+\tilde{\gamma}^*_{k'}))}{1+\exp(\tilde{\delta}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'};\boldsymbol{b}^*_{k'};H^*_{k'}(\tilde{\gamma}^*_{k'}))} = 1;$$

and if $k' \in pa(j)$,

$$\begin{aligned} r =& \frac{1+\exp(\alpha^*_{k'j}+\tilde{\delta}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'};\boldsymbol{b}^*_{k'};H^*_{k'}(\beta^*_{k'j}+\tilde{\gamma}^*_{k'}))}{1+\exp(\tilde{\delta}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'};\boldsymbol{b}^*_{k'};H^*_{k'}(\tilde{\gamma}^*_{k'}))} \\ & \times \frac{H_j(\beta_{jk'}+\tilde{\gamma}_j)\left\{1+\exp(\tilde{\delta}_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\tilde{\gamma}_j))\right\}}{H_j(\tilde{\gamma}_j)\left\{1+\exp(\alpha_{jk'}+\tilde{\delta}_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\beta_{jk'}+\tilde{\gamma}_j))\right\}} \end{aligned}$$

and

$$r = \frac{H_j\left(\beta_{jk'}(x_{k'}+1)+\tilde{\gamma}_j\right)\left\{1+\exp(\alpha_{jk'}x_{k'}+\tilde{\delta}_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\beta_{jk'}x_{k'}+\tilde{\gamma}_j))\right\}}{H_j\left(\beta_{jk'}x_{k'}+\tilde{\gamma}_j\right)\left\{1+\exp(\alpha_{jk'}(x_{k'}+1)+\tilde{\delta}_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\beta_{jk'}(x_{k'}+1)+\tilde{\gamma}_j))\right\}}$$

for any possible positive value of $x_{k'}$ in the support of $X_{k'}$. Since these equations hold for all possible values of $x_l, l \in \boldsymbol{V} \setminus \{j,k'\}$, in both cases, we have that $r=1$ and $\alpha^*_{k'j}=\beta^*_{k'j}=0$. This indicates $k' \notin ch^*(j)$, which contradicts the assumption that $k' \in ch^*(j) \cap nd(j) (\neq \emptyset)$.

We now show $pa(j) = pa^*(j)$. Given the above result, $ch^*(j) \cap nd(j) = \emptyset$, we can simplify (6) as

$$\begin{aligned} &\frac{(a_{j1}+x_j)\cdots(a_{jp_j}+x_j)H_j(\sum_{k\in pa(j)}\beta_{jk}x_k+\gamma_j)}{(b_{j1}+x_j)\cdots(b_{jq_j}+x_j)(x_j+1)} \\ &= \frac{(a^*_{j1}+x_j)\cdots(a^*_{jp^*_j}+x_j)H^*_j(\sum_{k\in pa^*(j)}\beta^*_{jk}x_k+\gamma^*_j)}{(b^*_{j1}+x_j)\cdots(b^*_{jq^*_j}+x_j)(x_j+1)}. \end{aligned} \tag{12}$$

for a positive integer x_j. For $l \in pa(j) \backslash pa^*(j)$, taking the ratio of (12) at $x_l = 1$ and $x_l = 0$, we obtain

$$\frac{H_j(\sum_{k\in pa(j)\backslash\{l\}}\beta_{jk}x_k+\beta_{jl}+\gamma_j)}{H_j(\sum_{k\in pa(j)\backslash\{l\}}\beta_{jk}x_k+\gamma_j)} = 1,$$

which leads to $\beta_{jl}=0$. Next, if $x_j=0$, (6) is simplified as

$$\begin{aligned} &\frac{a_{j1}\cdots a_{jp_j}H_j(\sum_{k\in pa(j)}\beta_{jk}x_k+\gamma_j)}{b_{j1}\cdots b_{jq_j}\{1+\exp(\sum_{k\in pa(j)}\alpha_{jk}x_k+\delta_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\sum_{k\in pa(j)\cap pa^*(j)}\beta_{jk}x_k+\gamma_j))\}} \\ &= \frac{a^*_{j1}\cdots a^*_{jp^*_j}H^*_j(\sum_{k\in pa^*(j)}\beta^*_{jk}x_k+\gamma^*_j)}{b^*_{j1}\cdots b^*_{jq^*_j}\{1+\exp(\sum_{k\in pa^*(j)}\alpha^*_{jk}x_k+\delta^*_j)_{p^*_j}F_{q^*_j}(\boldsymbol{a}^*_j;\boldsymbol{b}^*_j;H^*_j(\sum_{k\in pa(j)\cap pa^*(j)}\beta^*_{jk}x_k+\gamma^*_j))\}}. \end{aligned} \tag{13}$$

Again, take the ratio of (13) at $x_l=1$ and $x_l=0$ for $l \in pa(j) \setminus pa^*(j)$. We have

$$\frac{1+\exp(\sum_{k\in pa(j)\backslash\{l\}}\alpha_{jk}x_k+\alpha_{jl}+\delta_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\sum_{k\in pa(j)\cap pa^*(j)}\beta_{jk}x_k+\gamma_j))}{1+\exp(\sum_{k\in pa(j)\backslash\{l\}}\alpha_{jk}x_k+\delta_j)_{p_j}F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\sum_{k\in pa(j)\cap pa^*(j)}\beta_{jk}x_k+\gamma_j))} = 1,$$

and therefore $\alpha_{jl} = 0$. The finding that $\alpha_{jl} = \beta_{jl} = 0$ for $l \in pa(j) \setminus pa^*(j)$ implies that $pa(j) \setminus pa^*(j) = \emptyset$. Similarly, we get $pa^*(j) \setminus pa(j) = \emptyset$, and hence $pa(j) = pa^*(j)$.

Next, we show that the model parameters of the linear ZiG-DAGs are also identifiable (up to permutations of $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$), under the assumption that p_j, q_j, and h_j are fixed (i.e., $p_j = p_j^*$, $q_j = q_j^*$, and $h_j = h_j^*$). We already know $pa(j) = pa^*(j)$ and thus we denote $pa^{\cap}(j) = pa(j) = pa^*(j)$. According to Corollary 5, it suffices to show that $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ implies that $\alpha_{jk} = \alpha_{jk}^*$, $\beta_{jk} = \beta_{jk}^*$ for $k \in pa^{\cap}(j)$, $\delta_j = \delta_j^*$, $\gamma_j = \gamma_j^*$, and $\boldsymbol{a}_j$ and $\boldsymbol{a}_j^*$, as well as $\boldsymbol{b}_j$ and $\boldsymbol{b}_j^*$, are equivalent up to a permutation. If we take the ratio of $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ at $x_j + 1$ and x_j, we observe that

$$\begin{aligned}&\frac{(a_{j1} + x_j) \cdots (a_{jp_j} + x_j) H_j(\sum_{k \in pa^{\cap}(j)} \beta_{jk} x_k + \gamma_j)}{(b_{j1} + x_j) \cdots (b_{jq_j} + x_j)(x_j + 1)} \\ &= \frac{(a_{j1}^* + x_j) \cdots (a_{jp_j}^* + x_j) H_j(\sum_{k \in pa^{\cap}(j)} \beta_{jk}^* x_k + \gamma_j^*)}{(b_{j1}^* + x_j) \cdots (b_{jq_j}^* + x_j)(x_j + 1)}\end{aligned} \tag{14}$$

for $x_j \neq 0$, and

$$\begin{aligned}&\frac{a_{j1} \cdots a_{jp_j} H_j(\sum_{k \in pa^{\cap}(j)} \beta_{jk} x_k + \gamma_j)}{b_{j1} \cdots b_{jq_j} \{1 + \exp(\sum_{k \in pa^{\cap}(j)} \alpha_{jk} x_k + \delta_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(\sum_{k \in pa^{\cap}(j)} \beta_{jk} x_k + \gamma_j))\}} \\ &= \frac{a_{j1}^* \cdots a_{jp_j}^* H_j(\sum_{k \in pa^{\cap}(j)} \beta_{jk}^* x_k + \gamma_j^*)}{b_{j1}^* \cdots b_{jq_j}^* \{1 + \exp(\sum_{k \in pa^{\cap}(j)} \alpha_{jk}^* x_k + \delta_j^*)_{p_j} F_{q_j}(\boldsymbol{a}_j^*; \boldsymbol{b}_j^*; H_j(\sum_{k \in pa^{\cap}(j)} \beta_{jk}^* x_k + \gamma_j^*))\}}\end{aligned} \tag{15}$$

for $x_j = 0$. Since (14) and (15) hold for any possible value of x_k for $k \in pa^{\cap}(j)$, we must have $\alpha_{jk} = \alpha_{jk}^*$, $\beta_{jk} = \beta_{jk}^*$ for $k \in pa^{\cap}(j)$, $\delta_j = \delta_j^*$, and $\gamma_j = \gamma_j^*$. It is also easy to see that $\boldsymbol{a}_j$ and $\boldsymbol{a}_j^*$, as well as $\boldsymbol{b}_j$ and $\boldsymbol{b}_j^*$, are equivalent up to a permutation. ■

Appendix D. Proof of Theorem 7

Proof To show that the graph structure of the nonlinear ZiG-DAGs is identifiable, we show $pa(j) = pa^*(j)$ and $ch^*(j) \cap nd(j) = \emptyset$, assuming the identity (6) holds for node $j \in \boldsymbol{V}$ for two arbitrary nonlinear ZiG-DAGs $\mathcal{B}$ and $\mathcal{B}^*$. Additionally, in order to establish the parameter identifiability of the nonlinear ZiG-DAGs, we show that equivalence of the parameters follows $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$, under the assumption that for each $j \in \boldsymbol{V}$, p_j, q_j, and h_j are fixed. We use the superscript * to indicate parameters that define the nonlinear ZiG-DAG $\mathcal{B}^*$.

We first show $ch^*(j) \cap nd(j) = \emptyset$. Suppose on the contrary that $ch^*(j) \cap nd(j) \neq \emptyset$. Consider $k' \in ch^*(j) \cap nd(j)$ such that $ch^*(k') \cap ch^*(j) \cap nd(j) = \emptyset$, as in the proof of Theorem 6. Under the nonlinear ZiG-DAGs, letting $x_j = 0$ while fixing x_l for all $l \in \boldsymbol{V} \setminus \{j, k'\}$, we can simplify (6) as

$$C_1 = \begin{cases} C_2 \frac{1 + \exp(f_{k'j}^*(1) + \tilde{\mu}_{k'}^*)_{p_{k'}^*} F_{p_{k'}^*}(\boldsymbol{a}_{k'}^*; \boldsymbol{b}_{k'}^*; H_{k'}^*(g_{k'j}^*(1) + \tilde{\nu}_{k'}^*))}{1 + \exp(f_{k'j}^*(0) + \tilde{\mu}_{k'}^*)_{p_{k'}^*} F_{p_{k'}^*}(\boldsymbol{a}_{k'}^*; \boldsymbol{b}_{k'}^*; H_{k'}^*(g_{k'j}^*(0) + \tilde{\nu}_{k'}^*))} & \text{for} \quad x_{k'} = 0 \\ C_2 r^{x_{k'}} & \text{for} \quad x_{k'} \neq 0, \end{cases}$$

if $k' \notin pa(j)$, and

$$\frac{C_3 H_j(g_{jk'}(x_{k'}) + \tilde{\nu}_j)}{1 + \exp(f_{jk'}(x_{k'}) + \tilde{\mu}_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(g_{jk'}(x_{k'}) + \tilde{\nu}_j))}$$
$$= \begin{cases} C_2 \frac{1+\exp(f^*_{k'j}(1)+\tilde{\mu}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'};\boldsymbol{b}^*_{k'};H^*_{k'}(g^*_{k'j}(1)+\tilde{\nu}^*_{k'}))}{1+\exp(f^*_{k'j}(0)+\tilde{\mu}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'};\boldsymbol{b}^*_{k'};H^*_{k'}(g^*_{k'j}(0)+\tilde{\nu}^*_{k'}))} & \text{for} \quad x_{k'} = 0 \\ C_2 r^{x_{k'}} & \text{for} \quad x_{k'} \neq 0, \end{cases}$$

if $k' \in pa(j)$, where C_1, C_2, C_3 are some constants, $H_j = h_j^{-1}$, $H^*_{k'} = (h^*_{k'})^{-1}$, $\tilde{\mu}_j = \sum_{l \in pa(j)\setminus\{k'\}} f_{jl}(x_l) + \mu_j$, $\tilde{\nu}_j = \sum_{l \in pa(j)\setminus\{k'\}} g_{jl}(x_l) + \nu_j$, $\tilde{\mu}^*_{k'} = \sum_{l \in pa^*(k')\setminus\{j\}} f^*_{k'l}(x_l) + \mu^*_{k'}$, $\tilde{\nu}^*_{k'} = \sum_{l \in pa^*(k')\setminus\{j\}} g^*_{k'l}(x_l) + \nu^*_{k'}$, and $r = \frac{H^*_{k'}(g^*_{k'j}(1)+\tilde{\nu}^*_{k'})}{H^*_{k'}(g^*_{k'j}(0)+\tilde{\nu}^*_{k'})}$. The above identities are well defined, because in our definition of the nonlinear ZiG-DAG, each conditional distribution (or equivalently, the derivatives of the probability generating function) should be positive over the entire support.

Because $x_{k'}$ is not binary (i.e., it takes integers beyond $\{0, 1\}$), if we take the ratio of the first equation above at $x_{k'} + 1$ and $x_{k'}$, we obtain that if $k' \notin pa(j)$,

$$r = \frac{1 + \exp(f^*_{k'j}(1) + \tilde{\mu}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(g^*_{k'j}(1) + \tilde{\nu}^*_{k'}))))}{1 + \exp(f^*_{k'j}(0) + \tilde{\mu}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(g^*_{k'j}(0) + \tilde{\nu}^*_{k'}))))} = 1.$$

If $k' \in pa(j)$, we take the ratio of the second equation and obtain

$$r = \frac{1 + \exp(f^*_{k'j}(1) + \tilde{\mu}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(g^*_{k'j}(1) + \tilde{\nu}^*_{k'}))))}{1 + \exp(f^*_{k'j}(0) + \tilde{\mu}^*_{k'})_{p^*_{k'}} F_{p^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H^*_{k'}(g^*_{k'j}(0) + \tilde{\nu}^*_{k'}))))}$$
$$\times \frac{H_j(g_{jk'}(1) + \tilde{\nu}_j) \left\{1 + \exp(f_{jk'}(0) + \tilde{\mu}_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(g_{jk'}(0) + \tilde{\nu}_j))\right\}}{H_j(g_{jk'}(0) + \tilde{\nu}_j) \left\{1 + \exp(f_{jk'}(1) + \tilde{\mu}_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(g_{jk'}(1) + \tilde{\nu}_j))\right\}}$$

and

$$r = \frac{H_j(g_{jk'}(x_{k'} + 1) + \tilde{\nu}_j) \left\{1 + \exp(f_{jk'}(x_{k'}) + \tilde{\mu}_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(g_{jk'}(x_{k'}) + \tilde{\nu}_j))\right\}}{H_j(g_{jk'}(x_{k'}) + \tilde{\nu}_j) \left\{1 + \exp(f_{jk'}(x_{k'} + 1) + \tilde{\mu}_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(g_{jk'}(x_{k'} + 1) + \tilde{\nu}_j))\right\}}$$

for any possible positive value of $x_{k'}$ in the support of $X_{k'}$. Similar to the proof of Theorem 6, we necessarily have in both cases that $r = 1$ as well as $f^*_{k'j}(0) = f^*_{k'j}(1)$, $g^*_{k'j}(0) = g^*_{k'j}(1)$. Now, we consider the ratio of (6) at $x_{k'} = 1$ and $x_{k'} = 0$ with arbitrary $x_j \neq 0$. Observe that

$$\frac{1 + \exp(f^*_{k'j}(x_j) + \tilde{\mu}^*_{k'})_{p^*_{k'}} F_{q^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H_{k'}(g^*_{k'j}(x_j) + \tilde{\nu}^*_{k'})}{1 + \exp(f^*_{k'j}(x_j + 1) + \tilde{\mu}^*_{k'})_{p^*_{k'}} F_{q^*_{k'}}(\boldsymbol{a}^*_{k'}; \boldsymbol{b}^*_{k'}; H_{k'}(g^*_{k'j}(x_j + 1) + \tilde{\nu}^*_{k'})} = 1$$

holds for any $x_j \neq 0$, indicating $f^*_{k'j}(x_j) = f^*_{k'j}(x_j + 1)$ and $g^*_{k'j}(x_j) = g^*_{k'j}(x_j + 1)$ for all possible $x_j \neq 0$. All together, we have $f^*_{k'j}(x_j) = f^*_{k'j}(x_j + 1)$ and $g^*_{k'j}(x_j) = g^*_{k'j}(x_j + 1)$ for any possible value of x_j. Because we have assumed $\mathrm{E}\left[f^*_{jk}(X_k)\right] = \mathrm{E}\left[g^*_{jk}(X_k)\right] = 0$ for all j, k, where X_k are count random variables, it implies that $f^*_{k'j}(x_j) = g^*_{k'j}(x_j) = 0$ for any value of x_j in its support, and hence $k' \notin ch^*(j)$. This contradicts the assumption that $k' \in ch^*(j) \cap nd(j) (\neq \emptyset)$.

Next, we show $pa(j) = pa^*(j)$. Let $l \in pa(j) \setminus pa^*(j)$. Since $ch^*(j) \cap nd(j) = \emptyset$, by taking the ratio of (6) at $x_l + 1$ and x_l, we have

$$\frac{r_j(x_j; x_l + 1, \boldsymbol{x}_{pa(j)\setminus\{l\}})}{r_j(x_j; x_l, \boldsymbol{x}_{pa(j)\setminus\{l\}})} = 1$$

for all x_j and $\boldsymbol{x}_{pa(j)}$, where

$$r_j(x_j; \boldsymbol{x}_{pa(j)}) = \begin{cases} \frac{a_{j1}\cdots a_{jp_j} H_j(\sum_{k\in pa(j)} g_{jk}(x_k)+\nu_j)}{b_{j1}\cdots b_{jq_j}\{1+\exp(\sum_{k\in pa(j)} f_{jk}(x_k)+\mu_j)_{p_j} F_{q_j}(\boldsymbol{a}_j;\boldsymbol{b}_j;H_j(\sum_{k\in pa(j)} g_{jk}(x_k)+\nu_j))\}} & \text{if} \quad x_j = 0 \\ \frac{(a_{j1}+x_j)\cdots(a_{jp_j}+x_j)H_j(\sum_{k\in pa(j)} g_{jk}(x_k)+\nu_j)}{(b_{j1}+x_j)\cdots(b_{jq_j}+x_j)(x_j+1)} & \text{if} \quad x_j \neq 0. \end{cases}$$

It easily follows that $f_{jl}(x_l+1) = f_{jl}(x_l)$ and $g_{jl}(x_l+1) = g_{jl}(x_l)$ for all possible values of x_l, and combining this again with the assumption that $\mathrm{E}\left[f^*_{jk}(X_k)\right] = \mathrm{E}\left[g^*_{jk}(X_k)\right] = 0$, we have $pa^*(j) \setminus pa(j) = \emptyset$. Similarly, we obtain $pa^*(j) \setminus pa(j) = \emptyset$, which implies $pa(j) = pa^*(j)$.

Lastly, we show that if p_j, q_j, and h_j are fixed (i.e., $p_j = p^*_j$, $q_j = q^*_j$, and $h_j = h^*_j$), we can deduce from $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ that $(f_{jk}, g_{jk}) = (f^*_{jk}, g^*_{jk})$ for $k \in pa^\cap(j)$, $\mu_j = \mu^*_j$, $\nu_j = \nu^*_j$, and $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ are equivalent to $\boldsymbol{a}^*_j$ and $\boldsymbol{b}^*_j$ up to permutations. Here, we denote $pa^\cap(j) = pa(j) = pa^*(j)$ as in the proof of Theorem 6. Consider the ratio of $p(x_j|\boldsymbol{x}_{pa(j)}) = p^*(x_j|\boldsymbol{x}_{pa^*(j)})$ at $x_j + 1$ and x_j. We observe that if $x_j \neq 0$,

$$\begin{aligned} &\frac{(a_{j1} + x_j)\cdots(a_{jp_j} + x_j)H_j(\sum_{k\in pa^\cap(j)} g_{jk}(x_k) + \nu_j)}{(b_{j1} + x_j)\cdots(b_{jq_j} + x_j)(x_j + 1)} \\ &= \frac{(a^*_{j1} + x_j)\cdots(a^*_{jp_j} + x_j)H_j(\sum_{k\in pa^\cap(j)} g^*_{jk}(x_k) + \nu^*_j)}{(b^*_{j1} + x_j)\cdots(b^*_{jq_j} + x_j)(x_j + 1)}, \end{aligned} \tag{16}$$

and otherwise,

$$\begin{aligned} &\frac{a_{j1}\cdots a_{jp_j} H_j(\sum_{k\in pa^\cap(j)} g_{jk}(x_k) + \nu_j)}{b_{j1}\cdots b_{jq_j}\{1 + \exp(\sum_{k\in pa^\cap(j)} f_{jk}(x_k) + \mu_j)_{p_j} F_{q_j}(\boldsymbol{a}_j; \boldsymbol{b}_j; H_j(\sum_{k\in pa^\cap(j)} g_{jk}(x_k) + \nu_j))\}} \\ &= \frac{a^*_{j1}\cdots a^*_{jp_j} H_j(\sum_{k\in pa^\cap(j)} g^*_{jk}(x_k) + \nu_j)}{b^*_{j1}\cdots b^*_{jq_j}\{1 + \exp(\sum_{k\in pa^\cap(j)} f^*_{jk}(x_k) + \mu^*_j)_{p_j} F_{q_j}(\boldsymbol{a}^*_j; \boldsymbol{b}^*_j; H_j(\sum_{k\in pa^\cap(j)} g^*_{jk}(x_k) + \nu^*_j))\}}. \end{aligned} \tag{17}$$

Note that (16) and (17) hold for all possible values of x_k for $k \in pa^\cap(j)$. If it is combined with $\mathrm{E}\left[f_{jk}(X_k)\right] = \mathrm{E}\left[g_{jk}(X_k)\right] = 0$, our identifiability condition for the nonlinear functions f_{jk} and g_{jk}, we get $\mu_j = \mu^*_j$, $\nu_j = \nu^*_j$, and $f_{jk}(x_k) = f^*_{jk}(x_k)$ and $g_{jk}(x_k) = g^*_{jk}(x_k)$ for all possible values of x_k for $k \in pa^\cap(j)$. Then, it is clear that $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$, respectively, are equivalent to a permutation of $\boldsymbol{a}^*_j$ and $\boldsymbol{b}^*_j$, which completes the proof. ■

References

Holly Andersen. When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5):672–683, 2013.

Simon C Barry and Alan H Welsh. Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157(2-3):179–188, 2002.

Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

Federico Castelletti, Guido Consonni, Marco L Della Vedova, and Stefano Peluso. Learning markov equivalence classes of directed acyclic graphs: an objective bayes approach. *Bayesian Analysis*, 13(4):1235–1260, 2018.

Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Junsouk Choi, Robert Chapkin, and Yang Ni. Bayesian causal structural learning with zero-inflated poisson bayesian networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Gerda Claeskens, Nils Lid Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.

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

Higinio Dopeso, Paulo Rodrigues, Josipa Bilic, Sarah Bazzocco, Fernando Cartón-García, Irati Macaya, Priscila Guimarães De Marcondes, Estefanía Anguita, Marc Masanas, Lizbeth M Jiménez-Flores, et al. Mechanisms of inactivation of the tumour suppressor gene rhoa in colorectal cancer. *British journal of cancer*, 118(1):106–116, 2018.

Jean-Paul Fox. Multivariate zero-inflated modeling with latent predictors: Modeling feedback behavior. *Computational Statistics & Data Analysis*, 68:361–374, 2013.

Heonjong Han et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386, 2018.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

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.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf, et al. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, pages 689–696. Citeseer, 2008.

HE Hua, TANG Wan, WANG Wenjuan, and CRITS-CHRISTOPH Paul. Structural zeroes and zero-inflated models. *Shanghai Archives of Psychiatry*, 26(4):236, 2014.

Yasumichi Inoue and Takeshi Imamura. Regulation of tgf-β family signaling by e3 ubiquitin ligases. *Cancer science*, 99(11):2107–2112, 2008.

Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Yun Kang, Michael H Norris, Jan Zarzycki-Siek, William C Nierman, Stuart P Donachie, and Tung T Hoang. Transcript amplification from single bacterium for transcriptome analysis. *Genome Research*, 21(6):925–935, 2011.

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

Adrienne Winifred Kemp. *Studies in Univariate Discrete Distribution Theory Based on the Generalized Hypergeometric Function and Associated Differential Equations*. PhD thesis, Queens' University of Belfast, 1968b.

Huipeng Li, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, 49(5):708–718, 2017.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. CRC Press, 2018.

Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1(1):1–10, 2007.

Gunwoong Park and Hyewon Park. Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 158–166. PMLR, 2019.

Gunwoong Park and Garvesh Raskutti. Learning large-scale poisson dag models based on overdispersion scoring. *Advances in Neural Information Processing Systems*, 28:631–639, 2015.

Judea Pearl. *Causality: Models, reasoning and inference*. Cambridge university press, 2009.

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

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 589–598. AUAI Press, 2011.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

Paulo Rodrigues, Irati Macaya, Sarah Bazzocco, Rocco Mazzolini, Elena Andretta, Higinio Dopeso, Silvia Mateo-Lozano, Josipa Bilić, Fernando Cartón-García, Rocio Nieto, et al. Rhoa inactivation enhances wnt signalling and promotes colorectal cancer. *Nature communications*, 5(1):1–15, 2014.

Constance E Runyan, Tomoko Hayashida, Susan Hubchak, Jessica F Curley, and H William Schnaper. Role of sara (smad anchor for receptor activation) in maintenance of epithelial cell phenotype. *Journal of Biological Chemistry*, 284(37):25181–25189, 2009.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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

Kevin E Staub and Rainer Winkelmann. Consistent estimation of zero-inflated count models. *Health Economics*, 22(6):673–686, 2013.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

Y Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59, 2020.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Directed graphical models and causal discovery for zero-inflated data. *arXiv preprint arXiv:2004.04150*, 2020.