Ordinal-Poisson Causal Discovery

Joni Shaska, Urbashi Mitra
University of Southern California
{shaska,ubli}@usc.edu

Yingying Wang, Christine K. Johnson
University California Davis
{yxgwang,ckjohnson}@ucdavis.edu

***Abstract*—This paper deals with causal discovery in the setting where random variables in the causal graph can be either ordinal or Poisson. In particular, we show that it is possible to recover the causal network (modeled by a directed acyclic graph) from purely observational data. Such results are important in the field of causal discovery since, in general, it is impossible to recover causal relationships from purely observational data, and the identifiability of causal network structures relies on assumptions regarding the data-generating process. Causal discovery in the setting where all variables in the graph are ordinal and the setting where all variables are Poisson has been separately considered. However, the setting in this paper where the two are allowed to co-exist in the graph has, to the best of our knowledge, not been considered. In addition to proving the identifiability of ordinal-Poisson causal discovery, we show, through numerical simulations, that the Bayesian information criterion (BIC) tends to reliably recover the true causal graph as the number of observations increases.**

***Index Terms*—causal inference, causal discovery, identifiability**

I. Introduction

Understanding the underlying causes of phenomena affected by multiple variables can often be done via the representation of causal graphs [1]. These graphs are often assumed to be directed acyclic graphs. Applying causal graph discovery can have utility in disciplines as diverse as topology inference in wireless networks [2], gene networks in biology (e.g. [3]), the impact of medications, and optimizing the effect of advertising [4]. Unfortunately, it is well known that, in general, one cannot distinguish between cause and effect from observational data alone [5]. Hence, the causal discovery literature has emphasized studying which assumptions and system models allow one to identify causal relationships, with some notable examples including linear systems with additive homoscedastic Gaussian noise [6], linear models with additive non-Gaussian noise [7], and non-linear models with additive Gaussian noise [8]. The downstream application of the found graph can also influence the graph identification scheme. In [9], causal graph identification is optimized for a causal bandit application, and in [10], false negatives (or false positives) in link discovery of the graph are minimized.

In this work, we study causal discovery in the setting where the dataset contains both ordinal and Poisson count data. First, we show that it is possible to identify causal structures from observational data alone. Second, we derive an estimator based on the Bayesian Information Criterion (BIC) to identify causal relationships from observational data.

This work has been funded in part by one or more of the following grants: ARO W911NF1910269, ARO W911NF2410094, NSF RINGS-2148313, NSF CCF-2200221, NSF CCF-2311653, ONR N00014-22-1-2363, and the NSF Center for Pandemic Insights DBI-2412522

The problem of identifiability has received much attention in the causal discovery literature [6]–[8], [11]–[13]. Many classical results often deal with continuous variables, or assume an additive noise model such as in [6]–[8]. However, some recent results have proved identifiability for datasets consisting of discrete random variables. For instance, [11] shows that datasets consisting of ordinal random variables are identifiable. The work in [12] shows identifiability for datasets consisting of Poisson (count) data, and [13] generalizes this result to distributions of the generalized hyper-geometric family, allowing for some heterogeneity among the nodes in the underlying causal graph.

This paper seeks to generalize the results of [11] and [12] by showing that datasets consisting of *both* ordinal and Poisson data are identifiable, which, to the best of our knowledge, have not been considered in the causal discovery literature. These datasets are found in epidemiology, such as [14], which contains the list of endangered species (red list) described by ordinal values, population trends (is the population increasing or decreasing, *etc.*, also ordinal), and the number of zoonotic viruses carried by different animal host species, a count that is well-modeled by a Poisson random variable. While causal discovery has been applied to epidemiological scenarios [15], [16], typically human-expert designed graphs are constructed and regression methods employed to determine the weights of the graph. In contrast, our approach is completely data driven, but exploits real data to determine meaningful underlying distributions for the observations. This approach is needed since all causal relationships cannot be determined *a priori*.

We adapt the model and techniques used in [11] to show that the set of model parameters that results in a non-identifiable system has zero Lebesgue measure for our mixed ordinal-Poisson observation model. Hence, if model parameters are chosen randomly according to any continuous probability distribution, a non-identifiable model occurs with zero probability. We also propose an estimator to recover the causal structure from observational data based on the Bayesian Information Criterion (BIC) for our model, similar to [6] and [11] that use the BIC to identify causal structures in linear systems with additive homoscedastic Gaussian noise and the ordinal model, respectively. We note that our models are not strictly linear in the classical sense. Moreover, it is worth noting that while our model generalizes the work of [11] and [12], even minor generalizations of identifiable systems do not necessarily yield identifiable systems. A notable example is

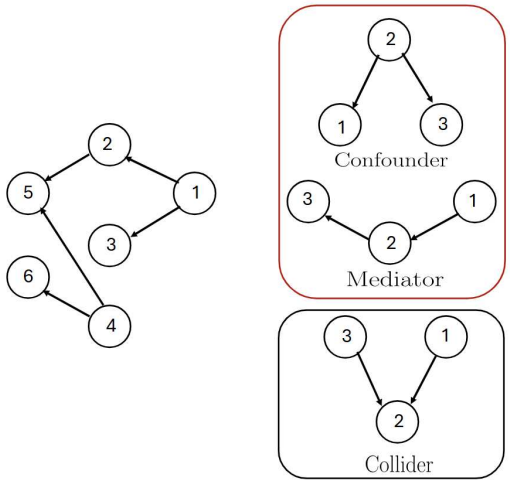

Fig. 1: An example graph (left). The three-node structures, confounder, mediator, and collider, are grouped by their Markov equivalence classes (right).

that linear systems with additive Gaussian noise, where error terms have equal variance, are identifiable, but linear systems with general Gaussian noise are not.

Our contributions are as follows:

1) We show that for datasets consisting of a mix of ordinal and Poisson random variables, the underlying causal graph is identifiable. This is a generalization of the results in [11] and [12], which show identifiability in the purely ordinal case and purely Poisson case, respectively.
2) We derive an estimator based on the Bayesian information criterion to recover the underlying graph in ordinal-Poisson systems.
3) We present numerical results showing that our algorithm correctly identifies three fundamental causal structures: the mediator, the confounder, and the collider. This shows that our algorithm can indeed distinguish not only the correct Markov equivalence class, but also the correct causal structure within that class. We also present results for larger graphs, showing the utility of our proposed algorithm.

This paper is organized as follows: Section II introduces the problem setup and formulation. Section III presents the main results of the paper. Section IV presents the estimator we use to learn causal structures in ordinal-Poisson systems, and Section V conducts numerical experiments for both small and moderate-sized graphs to illustrate the effectiveness of our proposed estimator.

Notation: Boldface symbols are used to denote vectors. Calligraphic letters denote sets and graphs. $\mathbb{P}$ is used to denote both probabilities and mass functions since all random variables considered are discrete.

II. Problem Setup

We briefly describe the system model. We consider a graph $\mathcal{G}$ with node set $\mathcal{V} = \{X_i\}_{i=1}^d$ and directed edge set $\mathcal{E}$. We assume $\mathcal{G}$ is a *weighted directed acyclic graph* (DAG) with adjacency matrix A. For a given node X_i, we denote its parent set as $\mathcal{P}(X_i)$, and let $\boldsymbol{P}_i$ denote the vector whose elements are the nodes in $\mathcal{P}(X_i)$. In the left diagram of Figure 1, we see that $\mathcal{P}(X_5) = \{2, 4\}, \mathcal{P}(X_6) = \{4\}$, and $\mathcal{P}(X_2) = \emptyset$.

We assume all nodes in $\mathcal{V}$ take on discrete values, and have a probability mass function conditioned on their parents, which we denote as $\mathbb{P}(X_i|\boldsymbol{P}_i)$. For each node, the probability mass function may take one of two forms:

- *Ordinal*:
  1) If $\mathcal{P}(X_i) \neq \emptyset$,
  $$\mathbb{P}(X_i = l|\boldsymbol{P}_i = \boldsymbol{p}) = F(\theta_{i,l} - \boldsymbol{\beta}_i^\top \boldsymbol{p}) - F(\theta_{i,l-1}^i - \boldsymbol{\beta}_i^\top \boldsymbol{p}), \tag{1}$$
  where $l \in \{1, 2, ..., L\}$, $-\infty = \theta_{i,0} < 0 = \theta_{i,1} \leq \theta_{i,2} \leq \cdots \leq \theta_{i,L} = \infty$, and $\mathbb{P}(X_i \leq l|\boldsymbol{P}_i = \boldsymbol{p}) = F(\theta_{i,l} - \boldsymbol{\beta}_i^\top \boldsymbol{p})$ where $F(x)$ is an *analytic* function in x, and whose derivative is nowhere zero. In this work, we assume the ordered logit model,
  $$F(\theta_{i,l} - \boldsymbol{\beta}_i^\top \boldsymbol{p}) = \frac{1}{1 + e^{-(\theta_{i,l} - \boldsymbol{\beta}_i^\top \boldsymbol{p})}}. \tag{2}$$
  However, our theoretical results hold for any analytic F, such as the *probit* function [17], given as
  $$F(\theta_{i,l} - \boldsymbol{\beta}_i^\top \boldsymbol{p}) = \Phi(1 + e^{-(\theta_{i,l} - \boldsymbol{\beta}_i^\top \boldsymbol{p})}), \tag{3}$$
  where Φ is the cumulative distribution function of the standard Gaussian distribution.
  2) If $\mathcal{P}(X_i) = \emptyset$,
  $$\mathbb{P}(X_i = l) = \pi_{i,l}. \tag{4}$$
- *Poisson*:
  1) If $\mathcal{P}(X_i) \neq \emptyset$,
  $$\mathbb{P}(X_i = j|\boldsymbol{P}_i = \boldsymbol{p}) = \frac{\mu_{i,\boldsymbol{p}}^j e^{-\mu_{i,\boldsymbol{p}}}}{j!}, \tag{5}$$
  where $j \in \{1, 2, ...\}$ and $\mu_{i,\boldsymbol{p}} = e^{\boldsymbol{\alpha}_i^\top \boldsymbol{p}}$.
  2) If $\mathcal{P}(X_i) = \emptyset$,
  $$\mathbb{P}(X_i = j) = \frac{\lambda_i^j e^{-\lambda_i}}{j!}, \tag{6}$$
  for some $\lambda > 0$.

It is important to note that this system model differs from the widely used *structural equation models* (SEMs). The *parameters* of our disributions are *functions* of the probability mass functions of the parents. In classical SEMs, the nodes are functions of their parents, that is, for $i \in \{1, 2, ..., d\}$, if $\boldsymbol{P_i} = p$,
$$X_i = f_i(\boldsymbol{p}, \zeta_i), \tag{7}$$
where ζ_i is a random variable. For example, a common SEM is an additive noise mode, i.e.,
$$X_i = f_i(\boldsymbol{p}, \zeta_i) = h_i(\boldsymbol{p}) + \zeta_i, \tag{8}$$
for some deterministic function h_i. Notice that for a fixed noise term ζ_i, and $\boldsymbol{p}$, X_i becomes a constant. This is not true in our

case, since even if the parent outcomes are held fixed, X_i is still a random variable due to its underlying distribution.

Throughout this work, we denote the set of nodes that have a Poisson distribution as $\mathcal{Q}$. We assume that the set of Poisson nodes is *known a priori* and remains *fixed*. That is, we do not need to infer whether a node is Ordinal or Poisson. One can easily extend our results to the case where $\mathcal{Q}$ is unknown if the support of the ordinal and Poisson random variables are distinct; in the case where the supports overlap, one can create a hypothesis test to distinguish between the the node types. However, in this latter case, the question of identifiability becomes a probabilistic one and merits further analysis beyond the scope of the current work.

Given n observations $\{\boldsymbol{X}^{(k)}\}_{k=1}^{n}$ where $\boldsymbol{X}^{(k)} = [X_1^{(k)}, ..., X_d^{(k)}]$ with $X_i^{(k)}$ being the kth observation of node i, the joint distribution induced by the graph $\mathcal{G}$ factors according to

$$\begin{aligned}\mathbb{P}_{\mathcal{G}}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(n)}) &= \prod_{k=1}^{n} \mathbb{P}_{\mathcal{G}}(\mathbf{X}^{(k)}) \\ &= \prod_{k=1}^{n}\prod_{i=1}^{d} \mathbb{P}(X_i^{(k)} | \boldsymbol{P}_i^{(k)}),\end{aligned} \tag{9}$$

where $\boldsymbol{P}_i^{(k)}$ is the vector whose entries are the kth observations of the parents of node i. Recall that the exact form of $\mathbb{P}(X_i^{(k)} | \boldsymbol{P}_i^{(k)})$ is dependent on whether node i is ordinal or Poisson, with the appropriate cases handled in equations (1) - (6).

There is an important concept related to equation (9) that appears repeatedly in our work called *Markov equivalence classes*, which we define below.

Definition 1. *Given a graph $\mathcal{G}$, the Markov equivalence class (MEC), denoted as $\mathcal{M}(\mathcal{G})$, is the set of all DAGs such that the joint distribution admits the same set of conditional independence as $\mathcal{G}$.*

An example of the Markov equivalence classes for three three-node structures is given in Figure 1. Since the confounder and mediator induce the same conditional independencies (namely, X_1 and X_3 are independent conditioned on X_2), they belong to the same MEC. A challenge with distinct DAGs belonging to the same MEC is that using only independence tests does not allow one to distinguish between the two DAGs. We will see that exploiting the additional structure in our problem allows for distinguishing DAGs in a common MEC.

Our goal is to recover the underlying DAG $\mathcal{G}$, together with the parameters given in Equations (1) - (6) that characterize the distributions of the nodes under the graph $\mathcal{G}$, given the observations $\{\boldsymbol{X}^{(k)}\}_{k=1}^{n}$. As stated before, it is, in general, not possible to recover the underlying causal graph given purely observational data [5]. Hence, to make theoretical claims regarding recovery of the graph $\mathcal{G}$ from observational data, one must make certain assumptions on the system model, such as [6]–[8], [11]–[13]. In particular, when $\mathcal{Q} = \emptyset$, our model reduces to that of [11], which shows that if all nodes are ordinal, then one can recover the underlying graph $\mathcal{G}$. Moreover, it is further shown in [12] that if all nodes are Poisson (which in our setting corresponds to $\mathcal{Q} = \{1, 2, ..., d\}$), one can also recover the underlying graph. Then, it is natural to ask whether, if the graph contains a mixture of ordinal and Poisson random variables, it is still possible to recover the underlying graph. The rest of this paper deals with this question, and it is shown in Section III that under our system model, one can still recover causal relationships in the bivariate case. It is worth noting that even slight generalizations of identifiable systems need not yield identifiable systems. The classical example is linear systems with additive Gaussian noise. If the noise terms have equal variance, the underlying graph is identifiable [6]. However, if this assumption (and only this assumption) is removed, identifiability is lost [5], [6]. Hence, there is no reason to believe that, despite ordinal and Poisson systems being separately identifiable, mixtures of these systems should also be identifiable.

III. Main Results

In this section, we show that in the bivariate case, one can identify causal relationships between ordinal and Poisson random variables. We take an approach similar to that in [11], in which we show that the set of parameters such that causal relationships are not identifiable has zero Lebesgue measure. Then, if the parameters of the system are chosen randomly according to any continuous probability distribution, the probability that one cannot identify the correct causal direction is exactly zero. Before we continue with the proof, we must introduce some important concepts and definitions.

A. Definitions

Let X_1 be a Poisson random variable and let X_2 be an ordinal random variable. In accordance with equation 9, if X_1 and X_2 have no causal relationship, then the joint distribution factors into the product of the marginals, and so this case can be distinguished by simple independence tests (namely testing wether X_1 and X_2 are mutually independent).Hence, we only need to distinguish between the two cases $X_1 \to X_2$ and $X_2 \to X_1$. To this end, we define the following two sets:

1)

$$\begin{aligned}\mathcal{D}_1 = \{\boldsymbol{\theta}_2, \beta_2, \lambda_1 : \exists \boldsymbol{\mu}_1, \boldsymbol{\pi}_2 \ni \\ \mathbb{P}_{X_1 \to X_2}(X_2 = l, X_1 = j; \boldsymbol{\theta}_2, \beta_2, \lambda_1) \\ = \mathbb{P}_{X_2 \to X_1}(X_2 = l, X_1 = j; \boldsymbol{\mu}_1, \boldsymbol{\pi}_2)\}.\end{aligned} \tag{10}$$

2)

$$\begin{aligned}\mathcal{D}_2 = \{\boldsymbol{\mu}_1, \boldsymbol{\pi}_2 : \exists \boldsymbol{\theta}_2, \beta_2, \lambda_1 \ni \\ \mathbb{P}_{X_1 \to X_2}(X_2 = l, X_1 = j; \boldsymbol{\theta}_2, \beta_2, \lambda_1) \\ = \mathbb{P}_{X_2 \to X_1}(X_2 = l, X_1 = j; \boldsymbol{\mu}_1, \boldsymbol{\pi}_2)\}.\end{aligned} \tag{11}$$

The two sets have simple, intuitive explanations. First, assume $X_1 \to X_2$ is the true causal model. Then, in this case, the joint distribution is characterized by the vector of parameters $\boldsymbol{\theta}_2 = [\theta_{2,1}, \theta_{2,2}, ...\theta_{2,L}]$, β_2, and λ_1, as described in equations (1) - (6). However, it could be the case that there exist parameters $\boldsymbol{\mu}_1 = [\mu_{1,1}, \mu_{1,2}, ..., \mu_{1,L}]$ and $\boldsymbol{\pi}_2 = [\pi_{2,1}, \pi_{2,2}, ..., \pi_{2,L}]$

that characterize the reverse model $X_2 \to X_1$ such that the likelihood of the two models are the same. Then, one cannot identify the two models from purely observational data. The set $\mathcal{D}_2$ is the analogous set for when $X_2 \to X_1$ is the true model. We show that both $\mathcal{D}_1$ and $\mathcal{D}_2$ have zero Lebesgue measure.

As mentioned in Section II, we assume the function F in equation (1) is an *analytic function*. This is a key assumption in our proof, and so we define analytic functions below.

Definition 2. *A real function F is real analytic if it is infinitely differentiable and matches its Taylor series in a neighborhood of every point.*

With the above definitions, we are ready to give the main results.

Theorem 1. *Let X_1 be a Poisson random variable, and let X_2 be an ordinal random variable with $L \geq 3$. Assume $X_1 \to X_2$ is the true causal model, and define $\mathcal{D}_1$ as in* (10)*. Then, $\mathcal{D}_1$ has zero Lebesgue measure.*

Proof. The proof is given in Appendix VII-A. □

Unfortunately, the problem is not symmetric. That is, Theorem 1 alone is not sufficient for identifiability. Hence, we must also deal with the case when $X_2 \to X_1$ is the true structure.

Theorem 2. *Let X_1 be a Poisson random variable, and let X_2 be an ordinal random variable with $L \geq 3$. Assume $X_2 \to X_1$ is the true causal model, and define $\mathcal{D}_2$ as in* (10)*. Then, $\mathcal{D}_2$ has zero Lebesgue measure.*

Proof. The proof is given in Appendix VII-B. □

The proofs of Theorems 1 and 2 focus on constructing two analytic functions g_1 (which is a function of $\boldsymbol{\theta}_2, \beta_2, \lambda_1$) and g_2 (which is a function of $\boldsymbol{\mu}_1, \boldsymbol{\pi}_2$). It is then shown that $\mathcal{D}_1$ and $\mathcal{D}_2$ are subsets of the zero sets of g_1 and g_2. Since the zero sets of analytic functions have zero Lebesgue measure, we conclude that so do $\mathcal{D}_1$ and $\mathcal{D}_2$.

As will be seen in our numerical results, we can devise an algorithm to identify these heterogeneous graphs with more than two nodes. We conjecture that our methods can be extended to the case of $n > 2$ as follows via an induction argument, since, if one conditions on the parents of two given nodes, our proof techniques for Theorems 1 and 2 hold. Hence, the main challenge in extending our results to the multi-node case is proving that source nodes can always be identified in general d-node DAGs. Then, our general technique should be easily applied to prove identifiability in the general d-node case.

IV. Algorithm for Ordinal-Poisson Discovery

We derive an algorithm for our setup based on the Bayesian information criterion. We discuss implementation details and provide numerical results in Section V.

We begin by defining our objective function. We consider the Bayesian Information Criterion (BIC) as our metric. To be precise, given a DAG $\mathcal{G}$, let $\boldsymbol{\eta}(\mathcal{G})$ be the parameters that characterize the distributions $\mathbb{P}(X_i|\boldsymbol{P}_i)$ according to $\mathcal{G}$ and equations (1)-(6) (we hope that it is clear that the specific nodes in $\boldsymbol{P}_i$ depend on the graph $\mathcal{G}$). Then, given the observations $\{\boldsymbol{X}^{(k)}\}_{k=1}^n$ and the graph $\mathcal{G}$ (which may not be the true DAG), let $\hat{\mathbb{P}}(X_i^{(k)}|\boldsymbol{P}_i^{(k)})$ be the likelihood of the kth observation of node i conditioned on the kth observations of the parents of node i according to graph $\mathcal{G}$, evaluated using the maximum likelihood estimates of the parameters $\eta(\mathcal{G})$. Then, our estimate of the underlying DAG $\hat{\mathcal{G}}$ is given as

$$\hat{\mathcal{G}} = \arg\min_{\mathcal{G}} -2\sum_{k=1}^{n}\sum_{i=1}^{d} \log \hat{\mathbb{P}}(X_i^{(k)}|\boldsymbol{P}_i^{(k)}) + |\mathcal{E}| \log n, \quad (12)$$

where $|\mathcal{E}|$ is the number of edges in the graph $\mathcal{G}$. For a given graph $\mathcal{G}$, we briefly discuss how we obtain the maximum likelihood parameters of $\eta(\mathcal{G})$, by highlighting all cases below.

- *Ordinal*:
  1) If $\mathcal{P}(X_i) \neq \emptyset$, the maximum likelihood estimates $\hat{\theta}_{i,2}, ..., \hat{\theta}_{i,L}$ and $\hat{\boldsymbol{\beta}}_i$ are found by solving

$$\max_{\theta_{i,2},...,\theta_{i,L},\boldsymbol{\beta}_i} \sum_{k=1}^{n} \log F(\theta_{i,X_i^{(k)}} - \boldsymbol{\beta}_i^\top \boldsymbol{P_i^{(k)}}) - F(\theta_{i,X_i^{(k)}-1} - \boldsymbol{\beta}_i^\top \boldsymbol{P_i^{(k)}}). \quad (13)$$

     We use the OrderedModel module from the 8statsmodels.miscmodels.ordinal_model Python package to perform this ordinal regression.
  2) If $\mathcal{P}(X_i) = \emptyset$, the maximum likelihood estimates of the paratmers $\pi_{i,1}, ..., \pi_{i,L}$ are simply

$$\hat{\pi}_{i,l} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}\{X_i^{(k)} = l\}. \quad (14)$$

- *Poisson*:
  1) If $\mathcal{P}(X_i) \neq \emptyset$, the maximum likelihood estimates $\hat{\boldsymbol{\alpha}}_i$ are given by solving

$$\max_{\boldsymbol{\alpha}_i} \sum_{k=1}^{n} X_i^{(k)} \boldsymbol{\alpha}_i^\top \boldsymbol{P}_i^{(k)} - e^{\boldsymbol{\alpha}_i \top \boldsymbol{P}_i^{(k)}}. \quad (15)$$

     We use the Poisson module from the statsmodels.discrete.discrete_model Python package to perform this Poisson regression.
  2) If $\mathcal{P}(X_i) = \emptyset$, the maximum likelihood estimate of λ_i is

$$\hat{\lambda}_i = \frac{1}{n}\sum_{k=1}^{n} X_i^{(k)}. \quad (16)$$

Once all necessary regressions are performed, we can compute the BIC score for the given graph $\mathcal{G}$. Then, in order to find the graph that optimizes the BIC score, we can take one of two approaches:

1) *Exhaustive Search*: For small graph sizes, we can perform exhaustive search. That is, we can compute the BIC for every graph $\mathcal{G}$ and simply select the graph with the minimum BIC score.

2) *Greedy Search*: For larger graph sizes, exhaustive search becomes infeasible as the number of DAGs grows super-exponentially in the number of nodes. Hence, for larger graphs, we employ a greedy search algorithm. That is, for a given graph $\mathcal{G}$, we compute the BIC score of $\mathcal{G}$ and all of its neighbors. A neighbor is defined as a DAG that differs from $\mathcal{G}$ by at most an edge addition, edge deletion, or edge reversal. If the BIC score of $\mathcal{G}$ is smaller than all of its neighbors, we output $\mathcal{G}$. Otherwise, we go to the neighboring DAG with the smallest BIC score and repeat the process until termination.

V. Numerical Results

We present some numerical results to highlight the utility of our estimator defined in Section IV. We begin by showing that our estimator recovers three important three-node structures in causal discovery: the mediator, the confounder, and the collider. We also perform some experiments on larger graphs.

A. Three-Node Structures

We highlight three important three-node structures common in causal discovery:

- *Mediator*: $X_1 \to X_2 \to X_3$.
- *Confounder*: $X_1 \leftarrow X_2 \to X_3$.
- *Collider*: $X_1 \to X_2 \leftarrow X_3$.

The three structures are also displayed in Figure 1 along with their respective MECs. There are some important notes regarding these structures. First, as stated before, the mediator and confounder belong to the same MEC since they induce the same conditional independence on the data, namely that X_1 and X_3 are independent conditioned on X_2. The collider, however, belongs to a different Markov equivalence class, since under the collider structure, X_1 and X_3 are independent a priori, but become dependent when conditioned on X_2. Thanks to Equation (9), one can distinguish between Markov equivalence classes by testing conditional independence (namely, whether or not X_1 and X_3 are independent conditioned on X_2), hence distinguishing the collider from the mediator and confounder. However, since the mediator and confounder belong to the same Markov equivalence class, one cannot distinguish them via independence tests alone. Hence, we use our algorithm to distinguish between the three structures.

For our setup, we have X_1 and X_2 as ordinal random variables with $L = 3$, and X_3 as a Poisson random variable. The parameters of the model depend on which structure is being considered:

- *Mediator*: $\pi_{1,1} = .5, \pi_{1,2} = .38, \pi_{1,3} = .12$, $\theta_{2,1} = 0, \theta_{2,2} = 2, \theta_{2,3} = \infty$, $\beta_2 = 2$, and $\alpha_3 = 1.5$.
- *Confounder*: $\theta_{1,1} = 0, \theta_{1,2} = 2, \theta_{1,3} = \infty$, $\beta_1 = 2$, $\pi_{2,1} = .5, \pi_{2,2} = .38, \pi_{2,3} = .12$, and $\alpha_3 = 1.5$.
- *Collider*: $\pi_{1,1} = .5, \pi_{1,2} = .38, \pi_{1,3} = .12$, $\theta_{2,1} = 0, \theta_{2,2} = 2, \theta_{2,3} = \infty$, $\boldsymbol{\beta}_2 = [2, 1.5]^\top$, and $\lambda_3 = 1.5$.

We pick these values so that in all cases, the different outcomes of the ordinal variables are all balanced (that is, one outcome does not occur too rarely or is too common) since this imbalance further challenges graph finding. For each structure, we generate n observations (we vary n) and run our algorithm on the n instances for each structure. We compute the structural Hamming distance (SHD) between the true structure and the structure estimated using our method, which is the number of missing edges, false edges, and reversed edges between the estimated graph and the true graph. The results are given in Figure 2, from which we see that as n increases, the SHD goes to zero for each of the three structures, and so our method correctly identifies each structure. Moreover, we see that our method can not only correctly identify the Markov equivalence class, but it also correctly identifies *the causal structure within the Markov equivalence class*, as seen by its ability to distinguish between the mediator and confounder structures correctly.

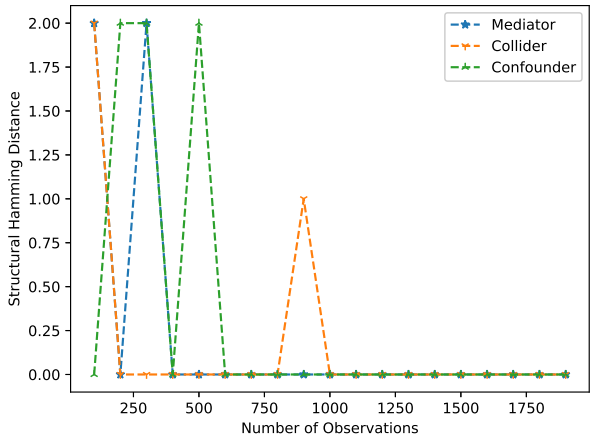

Fig. 2: Structural Hamming distance between the true structure and the estimated structure for each model.

B. Larger Graphs

We also explore our estimator's performance on slightly larger graphs ($d = 6$). We consider the graph given in Figure 1. Notice that this graph contains all three three-node structures. The set of Poisson nodes is given as $\mathcal{Q} = \{1, 6\}$, and for the ordinal nodes, we have $L = 3$ as before. We iterate over n, and for each n, generate data according to a DAG whose parameters are selected as follows:

- *Ordinal with* $\mathcal{P}(X_i) = \emptyset$: $\pi_{i,1} = .5, \pi_{i,2} = .38, \pi_{i,3} = .12$
- *Ordinal with* $\mathcal{P}(X_i) \neq \emptyset$: $\theta_{1,1} = 0, \theta_{1,2} = 2, \theta_{1,3} = \infty$, and the coefficients of the vector $\boldsymbol{\beta}_i$ are chosen i.i.d uniformly at random from the set $[-1, -.1] \cup [.1, 1]$.
- *Poisson with* $\mathcal{P}(X_i) = \emptyset$: $\lambda_i = 2$
- *Poisson with* $\mathcal{P}(X_i) \neq \emptyset$: The coefficients of the vector $\boldsymbol{\alpha}_i$ are chosen i.i.d uniformly at random from the set $[-1, -.1] \cup [.1, 1]$.

We plot the structural hamming distance in Figure 3, from which we see that the SHD decreases to zero, indicating that

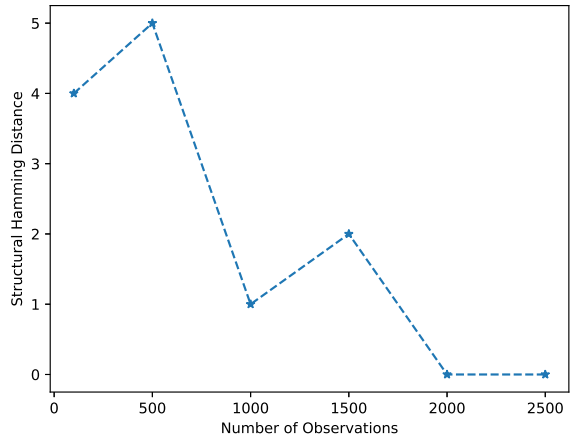

Fig. 3: Average Structural Hamming distance between the true structures and the estimated structures graph sized of $d = 6$.

our estimator can be used to recover larger graph sizes as well. This is an encouraging result, since this indicates that not only can our estimator recover the three-node structures separately, but it can also identify larger graphs with each of the three structures present as subgraphs. This facilitates further analysis to extend our results in Section III to the multi-node case.

VI. Conclusions

We have shown that ordinal-Poisson systems are identifiable from observational data. In particular, we have shown that the set of parameters such that one cannot identify the true causal structure occurs with probability zero when parameters are chosen according to any continuous probability distribution. Moreover, we derive an estimator based on the Bayesian Information Criterion (BIC) to identify causal structures from observational data. We show numerically that our estimator recovers correct causal structures as the number of observations increases, and that our estimator can be applied to larger graph sizes as well. Considering the cases where the node type is not known *a priori* is a topic for future work.

VII. Appendix

A. Proof of Theorem 1

First, for simplicity, we drop the subscript that denotes the node. We do this since it should be clear which node each parameter is referring to. We begin by writing for any $\boldsymbol{\theta}, \beta, \lambda$ in this set, we have for all l, j

$$\begin{aligned}&\mathbb{P}_{X_1 \to X_2}(X_2 = l, X_1 = j; \boldsymbol{\theta}, \beta, \lambda) \\ &\qquad = \mathbb{P}_{X_2 \to X_1}(X_2 = l, X_1 = j; \boldsymbol{\mu}, \boldsymbol{\pi})\end{aligned} \tag{17}$$

$$\implies \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big]\frac{\lambda^j e^{-\lambda}}{j!} = \frac{\mu_l^j e^{-\mu_l}}{j!}\pi_l \tag{18}$$

the left hand side of (18) is a telescoping sum in l, so summing over l yields

$$\big[F(\theta_L - \beta j) - F(\theta_0 - \beta j)\big]\frac{\lambda^j e^{-\lambda}}{j!} = \sum_{l=1}^{L}\frac{\mu_l^j e^{-\mu_l}}{j!}\pi_l. \tag{19}$$

Since $\theta_L = \infty$ and $\theta_0 = -\infty$, we have that $F(\theta_L - \beta j) = 1$ and $F(\theta_0 - \beta j) = 0$, which gives

$$\frac{\lambda^j e^{-\lambda}}{j!} = \sum_{l=1}^{L}\frac{\mu_l^j e^{-\mu_l}}{j!}\pi_l. \tag{20}$$

Then,

$$\big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] = \frac{\mu_l^j e^{-\mu_l}\pi_l}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l} \tag{21}$$

Letting $l = 1$ and assuming without loss of generality that $\theta_1 = 0$, we get

$$F(-\beta j) = \frac{\mu_1^j e^{-\mu_1}\pi_1}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l} \tag{22}$$

$$\implies \beta j = -F^{-1}\left(\frac{\mu_1^j e^{-\mu_1}\pi_1}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l}\right). \tag{23}$$

Substituting bacj into (21) for $l^* \neq l$ (we can do at least one substitution since $L \geq 3$), we get

$$\theta_{l^*} = F^{-1}\left(\frac{\sum_{l=1}^{l^*}\mu_l^j e^{-\mu_l}\pi_l}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l}\right) - F^{-1}\left(\frac{\mu_1^j e^{-\mu_1}\pi_1}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l}\right). \tag{24}$$

Since the left-hand side of (24) does not depend on j, we can write for any $j^* \neq j$,

$$\begin{aligned}0 = {} & F^{-1}\left(\frac{\sum_{l=1}^{l^*}\mu_l^j e^{-\mu_l}\pi_l}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l}\right) - F^{-1}\left(\frac{\mu_1^j e^{-\mu_1}\pi_1}{\sum_{l=1}^{L}\mu_l^j e^{-\mu_l}\pi_l}\right) \\ & - F^{-1}\left(\frac{\sum_{l=1}^{l^*}\mu_l^{j^*} e^{-\mu_l}\pi_l}{\sum_{l=1}^{L}\mu_l^{j^*} e^{-\mu_l}\pi_l}\right) + F^{-1}\left(\frac{\mu_1^{j^*} e^{-\mu_1}\pi_1}{\sum_{l=1}^{L}\mu_l^{j^*} e^{-\mu_l}\pi_l}\right).\end{aligned} \tag{25}$$

We make a series of important observations. First, the function xe^{-x} is analytic since it is the product of elementary functions. Since xe^{-x} is nowhere zero on $\mathbb{R}^+$ and the sum of analytic functions is analytic, the denominators in each term are analytic. Since the products of analytic functions are analytic, each argument in (25) is analytic. Since we have assumed F is analytic with derivative nowhere zero, F^{-1} is also analytic. Hence, the right side of (25) is an analytic function. Since the zero set of any analytic function (other than the zero function) has zero Lebesgue measure, and the above argument holds for any j, l $j^* \neq j$, and $l^* \neq l$, we are done.

B. Proof of Theorem 2

We begin by writing for any $\boldsymbol{\mu}, \boldsymbol{\pi} \in \mathcal{D}_{\in}$, for all l, j

$$\begin{aligned}&\mathbb{P}_{X_1 \to X_2}(X_2 = l, X_1 = j; \boldsymbol{\theta}, \beta, \lambda) \\ &\qquad = \mathbb{P}_{X_2 \to X_1}(X_2 = l, X_1 = j; \boldsymbol{\mu}, \boldsymbol{\pi})\end{aligned} \tag{26}$$

$$\implies \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!} = \frac{\mu_l^j e^{-\mu_l}}{j!} \pi_l. \tag{27}$$

Since $\sum_{j=0}^{\infty} \frac{\mu_l^j e^{-\mu_l}}{j!} = 1$, summing over j yields,

$$\sum_{j=0}^{\infty} \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!} = \pi_l. \tag{28}$$

Substituting back into (27) yields

$$\frac{\mu_l^j e^{-\mu_l}}{j!} = \frac{\big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!}}{\sum_{j=0}^{\infty} \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!}}. \tag{29}$$

Since this holds for every j, substituting $j = 0$ we get

$$e^{-\mu_l} = \frac{\big[F(\theta_l) - F(\theta_{l-1})\big] e^{-\lambda}}{\sum_{j=0}^{\infty} \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!}}. \tag{30}$$

Similarly, substituting $j = 1$ into (29) yields

$$\mu_l e^{-\mu_l} = \frac{\big[F(\theta_l - \beta) - F(\theta_{l-1} - \beta)\big] \lambda e^{-\lambda}}{\sum_{j=0}^{\infty} \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!}}. \tag{31}$$

Substituting (30) into (31) yields

$$\mu_l = \frac{\big[F(\theta_l - \beta) - F(\theta_{l-1} - \beta)\big] \lambda}{\big[F(\theta_l) - F(\theta_{l-1})\big]}. \tag{32}$$

Moreover, substituting $j = 2$ in (29) further yields,

$$\frac{\mu_l^2 e^{-\mu_l}}{2} = \frac{\big[F(\theta_l - 2\beta) - F(\theta_{l-1} - 2\beta)\big] \frac{\lambda^2 e^{-\lambda}}{2}}{\sum_{j=0}^{\infty} \big[F(\theta_l - \beta j) - F(\theta_{l-1} - \beta j)\big] \frac{\lambda^j e^{-\lambda}}{j!}}, \tag{33}$$

which by substituting (30) into (33) yields,

$$\mu_l = \sqrt{\frac{\big[F(\theta_l - 2\beta) - F(\theta_{l-1} - 2\beta)\big] \lambda^2}{\big[F(\theta_l) - F(\theta_{l-1})\big]}}. \tag{34}$$

Then, subtracting (32) and (34) we get

$$0 = \frac{\big[F(\theta_l - \beta) - F(\theta_{l-1} - \beta)\big] \lambda}{\big[F(\theta_l) - F(\theta_{l-1})\big]} - \sqrt{\frac{\big[F(\theta_l - 2\beta) - F(\theta_{l-1} - 2\beta)\big] \lambda^2}{\big[F(\theta_l) - F(\theta_{l-1})\big]}}. \tag{35}$$

By assumption, F is analytic, and since differences of analytic functions are analytic, both numerators and denominators are analytic. Moreover, $\sqrt{x}$ is analytic for $x > 0$, and so the second term of (35) is analytic. Hence, the right side of (35) is an analytic function. Since the zero set of any analytic function (other than the zero function) has zero Lebesgue measure, and the above argument holds for any l, we are done.

REFERENCES

[1] J. Pearl, *Causality*. Cambridge University Press, 2009.

[2] E. Testi and A. Giorgetti, "Blind wireless network topology inference," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1109–1120, 2020.

[3] G. L. Lozano, J. I. Bravo, M. F. G. Diago, H. B. Park, A. Hurley, S. B. Peterson, E. V. Stabb, J. M. Crawford, N. A. Broderick, and J. Handelsman, "Introducing THOR, a model microbiome for genetic dissection of community behavior," *mBio*, vol. 10, no. 2, pp. e02 846–18, 2019. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/mBio.02846-18

[4] E. Van den Broeck, K. Poels, and M. Walrave, "An experimental study on the effect of ad placement, product involvement and motives on facebook ad avoidance," *Telematics and Informatics*, vol. 35, no. 2, pp. 470–479, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585317307645

[5] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[6] J. Peters and P. Bühlmann, "Identifiability of Gaussian structural equation models with equal error variances," *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.

[7] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[8] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *Journal of Machine Learning Research*, vol. 15, no. 58, pp. 2009–2053, 2014. [Online]. Available: http://jmlr.org/papers/v15/peters14a.html

[9] C. Peng, D. Zhang, and U. Mitra, "Asymmetric graph error control with low complexity in causal bandits," *IEEE Transactions on Signal Processing*, 2025.

[10] J. Shaska and U. Mitra, "Neyman-Pearson causal inference," *2025 IEEE Trasnactions on Signal Processing (Early Access)*, pp. 1–14, 2025.

[11] Y. Ni and B. Mallick, "Ordinal causal discovery," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, J. Cussens and K. Zhang, Eds., vol. 180. PMLR, 01–05 Aug 2022, pp. 1530–1540. [Online]. Available: https://proceedings.mlr.press/v180/ni22a.html

[12] G. Park and G. Raskutti, "Learning large-scale poisson dag models based on overdispersion scoring," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/fccb60fb512d13df5083790d64c4d5dd-Paper.pdf

[13] G. Park and H. Park, "Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 158–166. [Online]. Available: https://proceedings.mlr.press/v89/park19a.html

[14] C. K. Johnson, P. L. Hitchens, P. S. Pandit, J. Rushmore, T. S. Evans, C. C. Young, and M. M. Doyle, "Global shifts in mammalian population trends reveal key predictors of virus spillover risk," *Proceedings of the Royal Society B*, vol. 287, no. 1924, p. 20192736, 2020.

[15] B. C. Saul, M. G. Hudgens, and M. E. Halloran, "Chapter 9 - causal inference in the study of infectious disease," in *Disease Modelling and Public Health, Part A*, ser. Handbook of Statistics, A. S. Srinivasa Rao, S. Pyne, and C. Rao, Eds. Elsevier, 2017, vol. 36, pp. 229–246. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169716117300093

[16] L. A. Barrero Guevara, S. C. Kramer, T. Kurth, and M. Domenech de Cellès, "Causal inference concepts can guide research into the effects of climate on infectious diseases," *Nature Ecology & Evolution*, vol. 9, no. 2, pp. 349–363, 2025.

[17] A. Agresti and M. Kateri, "Categorical data analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 206–208.