Causal Discovery from Poisson Branching Structural Causal Model Using High-Order Cumulant with Path Analysis

Jie Qiao[1*], Yu Xiang[1*], Zhengming Chen[1], Ruichu Cai[1,2†], Zhifeng Hao[3]

[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]College of Science, Shantou University, Shantou, China
{qiaojie.chn, thexiang2000, chenzhengming1103, cairuichu}@gmail.com, haozhifeng@stu.edu.cn

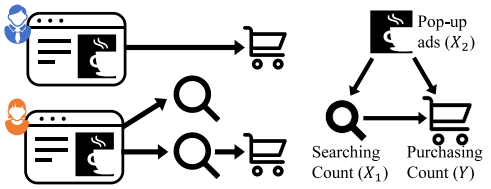
Abstract

Count data naturally arise in many fields, such as finance, neuroscience, and epidemiology, and discovering causal structure among count data is a crucial task in various scientific and industrial scenarios. One of the most common characteristics of count data is the inherent branching structure described by a binomial thinning operator and an independent Poisson distribution that captures both branching and noise. For instance, in a population count scenario, mortality and immigration contribute to the count, where survival follows a Bernoulli distribution, and immigration follows a Poisson distribution. However, causal discovery from such data is challenging due to the non-identifiability issue: a single causal pair is Markov equivalent, i.e., $X \to Y$ and $Y \to X$ are distributed equivalent. Fortunately, in this work, we found that the causal order from X to its child Y is identifiable if X is a root vertex and has at least two directed paths to Y, or the ancestor of X with the most directed path to X has a directed path to Y without passing X. Specifically, we propose a Poisson Branching Structure Causal Model (PB-SCM) and perform a path analysis on PB-SCM using high-order cumulants. Theoretical results establish the connection between the path and cumulant and demonstrate that the path information can be obtained from the cumulant. With the path information, causal order is identifiable under some graphical conditions. A practical algorithm for learning causal structure under PB-SCM is proposed and the experiments demonstrate and verify the effectiveness of the proposed method.

Introduction

Causal discovery from observational data especially for count data is a crucial task that arises in numerous applications in biology (Wiuf and Stumpf 2006), economic (Weiß and Kim 2014), network operation maintenance (Qiao et al. 2023; Cai et al. 2022), etc. In online services, for instance, the reason for the number of product purchases is of particular interest, while finding the underlying causal structure among user behavior from purely observational data is appealing and pivotal for online operation.

Much effort has been made to address the identification of causal structure from observational data (Spirtes, Glymour, and Scheines 2000; Zhang et al. 2018; Glymour, Zhang, and Spirtes 2019; Cai et al. 2018). In particular, constraint-based methods (Pearl 2009; Spirtes, Meek, and Richardson 1995), score-based methods (Chickering 2002; Tsamardinos, Brown, and Aliferis 2006) identify the causal structure by exploring the conditional independence relation among variables, but these methods only focus on the category domain and can only identify up to the Markov equivalent class (Pearl 2009). Thus, proper count data modeling is required to further identify the causal structure beyond the equivalence class. Recent work by (Park and Raskutti 2015) introduces a Poisson Bayesian network to model the count data and shows that it is identifiable using the overdispersion properties of Poisson BNs. Subsequently, it has been extended by accommodating a broader spectrum of distributions (Park and Raskutti 2017). In addition, the modeling of the zero-inflated Poisson data (Choi, Chapkin, and Ni 2020) and the ordinal relation data (Ni and Mallick 2022) and its identifiability of causal structure are investigated. However, the majority of these methods model the count data using Bayesian network ignoring the inherent branching structure among the counting relationship which is frequently encountered (Weiß 2018).



(a) Branching Structure (b) Causal Graph

Figure 1: Illustration of branching structure causal modeling.

Take Figure 1 as an example, the cause of the purchasing event can be inherited from some of the searching events, the pop-up ads event, or exogenously occurs. As a result, the causal relationship among counts constitutes a branching structure that can be modeled by a binomial thinning operator '$\circ$' (Steutel and van Harn 1979) with an additive independent Poisson distribution for innovation. That is, the purchasing count (Y) is affected by the pop-up ads count (X_2) and the searching count (X_1) which can be modeled by $Y = a_1 \circ X_1 + a_2 \circ X_2 + \epsilon$ where $a \circ X \coloneqq \sum_{n=1}^{X} \xi_n^{(a)}$, and

[*]These authors contributed equally.
[†]Corresponding author.

$\xi_n^{(a)} \sim \text{Bern}(a), \epsilon \sim \text{Pois}$. Generally speaking, the thinning operator models the branching structure that not every click will lead to purchasing while the additional noise models the general count of exogenous events. That is, a count represents the random size of an imaginary population, and the thinning operation randomly deletes some of the members of this population while concurrently introducing new immigration. This modeling approach finds widespread utility across various domains, notably within the context of the integer-value autoregressive model (Weiß 2018), which is first proposed by Al-Osh and Alzaid (1987); McKenzie (1985). Despite its extensive used, how to identify the causal structure in such type of model from purely observational data is still unclear.

To explicitly account for the branching structure, we propose a Poisson Branching Structural Causal Model (PB-SCM). We establish the identifiability theory for the proposed PB-SCM using high-order cumulant with path analysis. Theoretical results suggest that for any adjacent vertex X and Y, the causal order is identifiable if X is a root vertex and has at least two directed paths to Y, or the ancestor of X with the most directed path to X has a directed path to Y without passing X. Based on the results of the causal order we further propose an efficient causal skeleton learning approach featured with FFT acceleration. We demonstrate the effectiveness of the proposed causal discovery method using synthetic data and real data.

Poisson Branching Structural Causal Model

In this section, we first formalize the Poisson branching structural causal model, and then we introduce the preliminary of cumulant and some necessary properties in this model.

Problem Formulation

Our framework is in the causal graphical models. We use $Pa(i) = \{j | j \to i\}$, $An(i) = \{j | j \rightsquigarrow i\}$ denote the set of parents, ancestors of vertex i in a directed acyclic graph (DAG), respectively, and $An(i, j) = An(i) \cap An(j)$ denote the set of common ancestors of vertex i and vertex j. Moreover, we define a *directed path* $P = (i_0, i_1, ..., i_n)$ in G is a sequence of vertices of G where there is a directed edge from i_j to i_{j+1} for any $0 \leqslant j \leqslant n-1$ with the coefficient $\alpha_{i_j, i_{j+1}}$ of each edge. The set of vertices can be arranged in *causal order*, such that no later variable causes any earlier variable.

Now, we show the causal relationship in a causal graph can be formalized as the **P**oisson **B**ranching **S**tructural **C**ausal **M**odel (PB-SCM). Let $X = \{X_1, \dots, X_{|V|}\}$ denotes a set of random Poisson counts, of which the causal relationship consist of a causal DAG $G(V, E)$ with the vertex set $V = \{1, 2, ..., |V|\}$ and edge set E such that each causal relation follows the PB-SCM:

Definition 1 (Poisson Branching Structural Causal Model)**.** *For each random variable $X_i \in X$, let $\epsilon_i \sim Pois(\mu_i)$ be the noise component of X_i, then X_i is generated by:*

$$X_i = \sum_{j \in Pa(i)} \alpha_{j,i} \circ X_j + \epsilon_i, \tag{1}$$

where $\alpha_{j,i} \in (0, 1]$ is the coefficient from vertex j to i, $Pa(i)$ is the parent set of X_i in G, and $\alpha \circ X_i := \sum_{n=1}^{X_i} \xi_n^{(\alpha)}$ is a Binomial thinning operation with $\xi_n^{(\alpha)} \overset{\text{i.i.d.}}{\sim}$ Bern(α), Bern(α) is the Bernoulli distribution with parameter α.

We further define some graphical concepts. We use $\mathbf{P}^{i \rightsquigarrow j} = \left\{P_k^{i \rightsquigarrow j}\right\}_{k=1}^{|\mathbf{P}^{i \rightsquigarrow j}|}$ denotes the set of all directed paths from vertex i to j, where $P_k^{i \rightsquigarrow j} = (i, k_1, k_2, ..., k_p, j)$, $p = |P_k^{i \rightsquigarrow j}| - 2$, denote the k-th directed path from vertex i to j. For each directed path $P_k^{i \rightsquigarrow j}$, we use $A_k^{i \rightsquigarrow j} = (\alpha_{i,k_1}, \alpha_{k_1,k_2}, \dots, \alpha_{k_p,j})$ denote the corresponding *coefficients sequence* of path $P_k^{i \rightsquigarrow j}$. We let $\mathbf{P}^{i \rightsquigarrow i} = \{P^{i \rightsquigarrow i}\}$ also be a valid directed path for simplicity. Besides, we use $A_k^{i \rightsquigarrow j} \circ X_i \coloneqq \alpha_{k_p,j} \circ \cdots \circ \alpha_{k_1,k_2} \circ \alpha_{i,k_1} \circ X_i$ denote to perform a consecutive thinning operation on X_i based on the path sequence.

Goal: Given i.i.d. samples $\mathcal{D} = \{x_1^{(j)}, \dots, x_{|V|}^{(j)}\}_{j=1}^m$ from the joint distribution $P(X)$, our goal is to identify the unknown causal structure G from $\mathcal{D}$, assuming the data generative mechanism follows PB-SCM.

Preliminary

To address the identification of PB-SCM, cumulant are used in our work for building a connection to the path, providing a solution to the identifiability issue. Here, we recall the definition of cumulant and some basis properties.

Definition 2 (k-th order joint cumulant tensor)**.** *The k-th order joint cumulant tensor of a random vector $X = [X_1, ..., X_n]^T$ is the k-way tensor $\mathcal{T}_X^{(k)}$ in $R^{n \times \cdots \times n} \equiv \left(R^n\right)^k$ whose entry in $(i_1, ..., i_k)$ is the joint cumulant:*

$$\mathcal{T}^{(k)}_{X\ i_1,\dots,i_k} = \kappa(X_{i_1}, \dots, X_{i_k}) := \sum_{(B_1,\dots,B_L)} (-1)^{L-1}(L-1)!\mathbb{E}\Big[\prod_{j \in B_1} X_j\Big] \cdots \mathbb{E}\Big[\prod_{j \in B_L} X_j\Big], \tag{2}$$

where the sum is taken over all partitions $(B_1, \dots, B_L)$ of the multiset $\{i_1, ..., i_k\}$.

In this work, we use the following specific cumulant form:

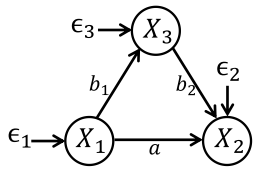
Definition 3 (2D slice of joint cumulant tensor)**.** *For a random vector X with k-th order joint cumulant tensor $\mathcal{T}_X^{(k)}$ where $k \geq 2$, denote its 2D matrix slice of k-th order joint cumulant tensor as $\mathcal{C}^{(k)}$, where*

$$\mathcal{C}_{i,j}^{(k)} := \kappa(X_i, \underbrace{X_j, \cdots, X_j}_{k-1 \text{ times}}). \tag{3}$$

Cumulant has the property of *multilinearity* such that $\kappa(X + Y, Z_1, \dots) = \kappa(X, Z_1, \dots) + \kappa(Y, Z_1, \dots)$. Furthermore, any cumulant involving two (or more) independent random variables equals zero, i.e., $\kappa(\epsilon_i, \epsilon_j, \dots) = 0$ if ϵ_i and ϵ_j are independent. More importantly, any two variables in cumulant are exchangeable, e.g., $\kappa(X, Y, \dots) = \kappa(Y, X, \dots)$.

Identifiability

In this section, we deal with the identification problem of causal structure under PB-SCM. Due to our identifiability result benefit from the 'reducibility' of cumulant in Poisson



$$X_1 = \epsilon_1$$
$$X_3 = b_1 \circ X_1 + \epsilon_3$$
$$X_2 = a \circ X_1 + b_2 \circ X_3 + \epsilon_2$$
$$\epsilon_i \sim Pois(\mu_i),\ i = 1,2,3.$$

Figure 2: Triangular structure. For simplicity, we denote directed path $P_1 : X_1 \xrightarrow{a} X_2$ and $P_2 : X_1 \xrightarrow{b_1} X_3 \xrightarrow{b_2} X_2$ with sequence of path coefficients $A_1 = (a)$ and $A_2 = (b_1, b_2)$.

distribution, we first characterize such property in Theorem 1. After which, an example is provided to reveal the intrinsic relation between the cumulant and the path in a causal graph under PB-SCM. Based on such connection, we complete the identifiability results that are divided into the case when the cause variable is root (Theorem 3) and the case when the cause variable is not root (Theorem 6).

We first introduce a fundamental property of cumulant in PB-SCM that the cumulant is reducible:

Theorem 1 (Reducibility)**.** *Given a Poisson random variable ϵ and n distinct sequences of coefficients $A_1, ..., A_n$, we have*

$$\begin{aligned}&\kappa(\underbrace{A_1 \circ \epsilon, ..., A_1 \circ \epsilon}_{k_1 \text{ times}}, ..., \underbrace{A_n \circ \epsilon, ..., A_n \circ \epsilon}_{k_n \text{ times}}) \\ &= \kappa(A_1 \circ \epsilon, \ldots, A_n \circ \epsilon)\end{aligned} \tag{4}$$

where each $A_i \circ \epsilon$ repeats $k_i \geq 1$ times in the original cumulant and only appears once in the reduced cumulant.

Such a result is a generalization of the property of the Poisson distribution since the cumulant of the Poisson distribution is identical in every order.

Motivating Example

Before describing our theoretical results, we use a motivating example to show the challenges of the non-identifiability issues and then introduce the basic intuition regarding in what case and how can we identify the PB-SCM.

To see the non-identifiability issue, we can show that a reversed model always exists in a two-variable system.

Remark 1. *For any two variables causal graph, the causal direction of PB-SCM is not identifiable and a distributed equivalent reversed model exists.*

For instance, consider $X_1 \to X_3$ in Fig. 2, the distributed equivalent reverse model satisfies $X_1 = \hat{b}_1 \circ X_3 + \hat{\epsilon}_1$, where $\hat{b}_1 = b_1\mu_1/(b_1\mu_1 + \mu_3)$ and $\hat{\epsilon}_1 \sim \text{Pois}(\mu_1 - b_1\mu_1)$ such that this direction is not identifiable.

Fortunately, we find that the causal direction is still possible to identify in a more general structure. Considering the causal relationship between X_1 and X_2 in Fig. 2, here we provide an intuitive example to show how to identify such causal direction by utilizing the relationship between cumulant and path. Considering the cumulant $\mathcal{C}_{1,2}$ with different orders, we can observe different behaviors of cumulant in the causal direction and the reverse direction. Thanks to the reducibility in Theorem 1, e.g., $\kappa(A_1 \circ \epsilon_1, \epsilon_1) = \kappa(A_1 \circ \epsilon_1, \epsilon_1, \epsilon_1)$, the cumulants with different orders for X_1 and X_2 is shown in Fig. 3(a) and Fig. 4(a). Interestingly, we have $\mathcal{C}_{2,1}^{(2)} = \mathcal{C}_{2,1}^{(3)}$ in the reverse direction (Fig. 4(a)) but $\mathcal{C}_{1,2}^{(2)} \neq \mathcal{C}_{1,2}^{(3)}$ in the causal direction (Fig. 3(a)), i.e., there exists an asymmetry in the inequality relations of cumulants. Such asymmetry intriguing possibility to identify the causal order between two variables using the cumulant.

To understand how this asymmetry occurs and hence use it to identify the causal relations. We first discuss the identification in the simple scenario that the cause variable is a root vertex in G, and then we generalize such results into the scenario that the cause variable is not root.

Identification When Cause Variable Is Root

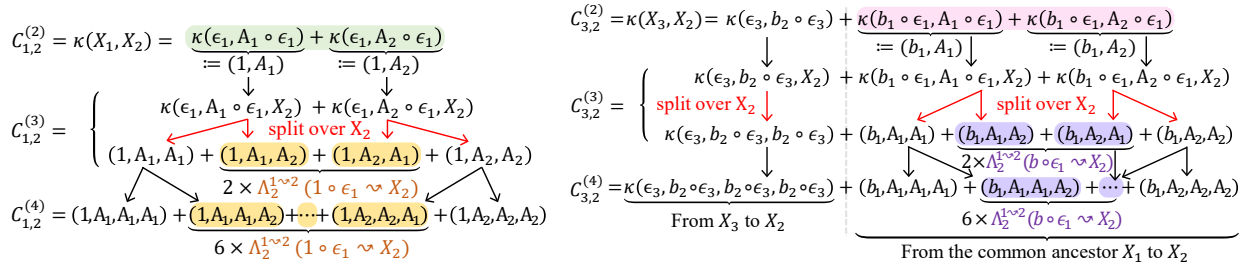
We start with the case that the cause variable is root vertex, in which our goal is to identify causal direction even though we do not know it is a root vertex. Recall the previous example, the key of identification is the inequality $\mathcal{C}_{1,2}^{(2)} \neq \mathcal{C}_{1,2}^{(3)}$ rendering an asymmetry for a causal pair. To understand how it occurs, we seek to character and leverage such inequality constraints of cumulants in a causal graph to infer the causal order (Theorem 4).

Here, we begin with two basic observations, which illustrate that inequality constraints of cumulants are driven by the number of paths between two variables. As shown in Fig. 3(a), one may see that (i) the decomposition of $\mathcal{C}_{1,2}$ is composed by a series of cumulants of the *common noise* (ϵ_1 in this example) between X_1 and X_2, which is due to the fact that any cumulant involving two (or more) independent random variables equals zero; (ii) moreover, such decomposition relates to the number of paths between X_1 and X_2 since $X_2 = A_1 \circ \epsilon_1 + A_2 \circ \epsilon_1 + b_2 \circ \epsilon_3 + \epsilon_2$ and by multilinearity, the cumulant will be split exponentially as the order of cumulant increase. With these observations, the reason why $\mathcal{C}_{1,2}^{(2)} \neq \mathcal{C}_{1,2}^{(3)}$ is that there exists more than one path in the causal direction while zero path in the reverse direction, i.e., $|\mathbf{P}^{1\rightsquigarrow 2}| = 2, |\mathbf{P}^{2\rightsquigarrow 1}| = 0$. As a result, $\mathcal{C}_{2,1}^{(2)} = \mathcal{C}_{2,1}^{(k)}$ for all $k \geq 2$ order cumulant in the reverse direction.

In the following, we articulate the underlying law of the cumulant in PB-SCM and propose a closed-form solution to it. The first important observation is that due to the reducibility and the exchangeability of cumulant, the $\mathcal{C}_{1,2}^{(k)}$ for $k \geq 3$ is only composed by three distinct cumulants: $\kappa(\epsilon_1, A_1 \circ \epsilon_1)$, $\kappa(\epsilon_1, A_2 \circ \epsilon_1)$, and $\kappa(\epsilon_1, A_1 \circ \epsilon_1, A_2 \circ \epsilon_1)$ with varying number of these cumulants. In particular, if we define the summation of cumulants that only contains one path as $\Lambda_1^{1\rightsquigarrow 2}(\epsilon_1 \rightsquigarrow X_2) \coloneqq \kappa(\epsilon_1, A_1 \circ \epsilon_1) + \kappa(\epsilon_1, A_2 \circ \epsilon_1)$ and the summation of cumulants that contains two paths as $\Lambda_2^{1\rightsquigarrow 2}(\epsilon_1 \rightsquigarrow X_2) \coloneqq \kappa(\epsilon_1, A_1 \circ \epsilon_1, A_2 \circ \epsilon_1)$, we will have the following closed-form solution:

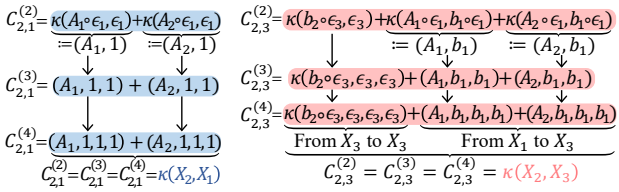
$$\mathcal{C}_{1,2}^{(4)} = \Lambda_1^{1\rightsquigarrow 2}(\epsilon_1 \rightsquigarrow X_2) + \sum_{\substack{m1+m2=3 \\ m_1,m_2>0}} \binom{3}{m_1\ m_2} \Lambda_2^{1\rightsquigarrow 2}(\epsilon_1 \rightsquigarrow X_2) \tag{5}$$

where $\binom{3}{m_1\ m_2}$ is the multinomial coefficient, indicating the number of ways of placing 3 distinct objects into 2 distinct bins with m_1 objects in the first bin, m_2 objects in the second



(a) Cumulant decomposition of the causal pair $X_1 \rightsquigarrow X_2$ where X_1 is root. (b) Cumulant decomposition of the causal pair $X_3 \rightsquigarrow X_2$ where X_3 is not root.

Figure 3: Illustration of decomposing the cumulant of causal direction, $\mathcal{C}_{1,2}$ and $\mathcal{C}_{3,2}$, in triangular structure (Fig. 2). For simplicity, we denote $\kappa(\epsilon_i, A_i \circ \epsilon_i, ..., A_j \circ \epsilon_i)$ by $(1, A_i, ..., A_j)$ and denote $\kappa(b_1 \circ \epsilon_i, A_i \circ \epsilon_i, ..., A_j \circ \epsilon_i)$ by $(b_1, A_i, ..., A_j)$.



(a) Cumulant decomposition of $X_2 \rightsquigarrow X_1$. (b) Cumulant decomposition of $X_2 \rightsquigarrow X_3$.

Figure 4: Illustration of decomposing the cumulant of reverse direction, $\mathcal{C}_{2,1}$ and $\mathcal{C}_{2,3}$, in triangular structure (Fig. 2).

bin. As a result, we will eventually have $6 \times \Lambda_2^{1\rightsquigarrow 2}(\epsilon_1 \rightsquigarrow X_2)$ as shown in Fig. 3(a). Generally, we define $\Lambda_k^{i\rightsquigarrow j}(A \circ \epsilon_i \rightsquigarrow X_j)$ as the summation of cumulants that contain k paths from root vertex i to j:

Definition 4 (k-path cumulants summation for root vertex)**.** *Given two vertices i and j, for $k \leqslant |\mathbf{P}^{i\rightsquigarrow j}|$, the k-path cumulants summation from vertex i to j is given by:*

$$\Lambda_k^{i\rightsquigarrow j}(A \circ \epsilon_i \rightsquigarrow X_j) = \sum_{1\leq l_1<l_2<\ldots<l_k\leq|\mathbf{P}^{i\rightsquigarrow j}|} \kappa(A \circ \epsilon_i, A_{l_1}^{i\rightsquigarrow j} \circ \epsilon_i, ..., A_{l_k}^{i\rightsquigarrow j} \circ \epsilon_i), \quad (6)$$

where $l_1, \ldots, l_k \in \mathbb{Z}^+$, A is an arbitrary sequence of coefficients. For $k > |\mathbf{P}^{i\rightsquigarrow j}|$, $\Lambda_k^{i\rightsquigarrow j} \equiv 0$ and for $k = 1$, $\Lambda_1^{i\rightsquigarrow i}(A \circ \epsilon_i \rightsquigarrow X_i) = \kappa(A \circ \epsilon_i, \epsilon_i)$, and $k > 1, \Lambda_k^{i\rightsquigarrow i} \equiv 0$.

Intuitively, Eq. (6) is a summation of all cumulants that contain k paths information from vertex i to j , and $\Lambda_1^{i\rightsquigarrow i}$ denotes the relation from the noise to itself. Based on the k-path cumulants summation, $\mathcal{C}_{i,j}^{(n)}$ can be decomposed as follows:

Theorem 2. *For any two vertices i and j where i is root vertex, i.e., vertex i has an empty parent set, the 2D slice of joint cumulant $\mathcal{C}_{i,j}^{(n)}$ satisfies:*

$$\mathcal{C}_{i,j}^{(n)} = \sum_{k=1}^{n-1} \sum_{\substack{m_1+\cdots+m_k=n-1 \\ m_l>0}} \binom{n-1}{m_1\ m_2 \cdots m_k} \Lambda_k^{i\rightsquigarrow j}(1 \circ \epsilon_i \rightsquigarrow X_j). \quad (7)$$

where $\binom{n-1}{m_1\ m_2\cdots m_k} = \frac{(n-1)!}{m_1!m_2!\cdots m_k!}$ is the multinomial coefficients.

Theorem 2 plays an important role in the identification of the causal order as it introduces the connection between the joint cumulant and path information. Moreover, since every order of the 2D slice joint cumulant can be obtained by Eq. (3), and thus every order of Λ_k can also be obtained by solving the equation in Eq. (7). By using Λ_k we are able to understand the identifiability in the following theorem:

Theorem 3 (Identifiability for root vertex)**.** *For any vertex i and j, where i is the root vertex in graph G, if $\mathcal{C}_{i,j}^{(3)} - \mathcal{C}_{i,j}^{(2)} \neq 0$, then $\mathcal{C}_{j,i}^{(3)} - \mathcal{C}_{j,i}^{(2)} = 0$ and X_i is the ancestor of X_j.*

Intuitively, based on Theorem 2, we have $\mathcal{C}_{i,j}^{(3)} - \mathcal{C}_{i,j}^{(2)} = \Lambda_2^{i\rightsquigarrow j}(1 \circ \epsilon_i \rightsquigarrow X_j)$, and thus $\mathcal{C}_{i,j}^{(3)} - \mathcal{C}_{i,j}^{(2)} \neq 0$ indicates that there exists more than one path from i to j than the reverse direction. That is, the causal direction for root vertex is identifiable if there are at least two directed paths:

Theorem 4 (Graphical Implication of Identifiability for Root Vertex)**.** *For a pair of vertices i and j in graph G, if vertex i is a root vertex and exists at least two directed paths from i to j, i.e., $|\mathbf{P}^{i\rightsquigarrow j}| \geq 2$, then the causal order between i and j is identifiable.*

Identification When Cause Variable Is Not Root

In this section, we aim to generalize the identification result from the root vertex to the non-root vertex.

When vertex i is not root, the main difference is that there might exist more than one common noise between two variables due to the possible common ancestor. Therefore, one may extend the result from the root vertex by considering each noise term as the separated root vertex. We present a general version of k-path cumulants summation as follows, which can be expressed as the aggregation of the k-path cumulants summations for the root vertices.

Definition 5 (k-path cumulants summation)**.** *The k-path*

cumulants summation from vertex i to vertex j is given by:

$$\tilde{\Lambda}_k(X_i \rightsquigarrow X_j) = \Lambda_k^{i\rightsquigarrow j}(1 \circ \epsilon_i \rightsquigarrow X_j) + \sum_{m \in An(i,j)\cup\{j\}} \sum_{h=1}^{|\mathbf{P}^{m\rightsquigarrow i}|} \Lambda_k^{m\rightsquigarrow j}\left(A_h^{m\rightsquigarrow i} \circ \epsilon_m \rightsquigarrow X_j\right). \tag{8}$$

where Λ_k is the k-path cumulants summation for root vertex, $|\mathbf{P}^{m\rightsquigarrow i}|$ is the number of directed paths from m to i.

With the general k-path cumulants summation, the general joint cumulant can be decomposed as follows:

Theorem 5. *For any two vertices i and j, the 2D slice of joint cumulant $\mathcal{C}_{i,j}^{(n)}$ satisfies:*

$$\mathcal{C}_{i,j}^{(n)} = \sum_{k=1}^{n-1} \sum_{\substack{m_1+\cdots+m_k=n-1 \\ m_l>0}} \binom{n-1}{m_1\ m_2 \cdots m_k} \tilde{\Lambda}_k(X_i \rightsquigarrow X_j), \tag{9}$$

where $\binom{n-1}{m_1\ m_2\cdots m_k} = \frac{(n-1)!}{m_1!m_2!\cdots m_k!}$ is the multinomial coefficients.

To see the connection with the case of root vertex, we take $X_3 \to X_2$ in Fig. 2 as example. Since X_3 can be expressed as $X_3 = b_1 \circ \epsilon_1 + \epsilon_3$, as shown in Fig. 3(b), we can separate the cumulant into two parts $\kappa(\epsilon_3, X_2)$, $\kappa(b_1 \circ \epsilon_1, X_2)$, which can be considered as the cumulant starting from vertex X_3 to X_2 and X_1 to X_2, respectively. As a result, the general k-path cumulants summation can be expressed as the aggregate of all different Λ_k starting with the corresponding noise terms. For instance, for $X_3 \to X_2$ in Fig. 2, we have:

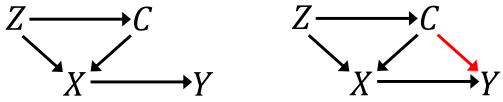
$$\tilde{\Lambda}_2(X_3 \rightsquigarrow X_2) = \underbrace{\Lambda_2^{3\rightsquigarrow 2}(1 \circ \epsilon_3 \rightsquigarrow X_2)}_{=0} + \underbrace{\Lambda_2^{1\rightsquigarrow 2}(b_1 \circ \epsilon_1 \rightsquigarrow X_2)}_{=\kappa(b_1\circ\epsilon_1, A_1\circ\epsilon_1, A_2\circ\epsilon_1)} \neq 0, \tag{10}$$

where Eq. (10) contains two different terms starting from ϵ_3 and ϵ_1, respectively. In particular, since there only exists one directed path from X_3 to X_2, $\Lambda_2^{3\rightsquigarrow 2}$ is zero while X_1 to X_2 has two paths and thus $\Lambda_2^{1\rightsquigarrow 2}$ is not zero. Similarly, for the reverse direction, we have

$$\tilde{\Lambda}_2(X_2 \rightsquigarrow X_3) = \underbrace{\Lambda_2^{2\rightsquigarrow 3}(1 \circ \epsilon_2 \rightsquigarrow X_3)}_{=0} + \underbrace{\Lambda_2^{3\rightsquigarrow 3}(b_2 \circ \epsilon_3 \rightsquigarrow X_3)}_{=0} + \underbrace{\Lambda_2^{1\rightsquigarrow 3}(A_1 \circ \epsilon_1 \rightsquigarrow X_3)}_{=0} + \underbrace{\Lambda_2^{1\rightsquigarrow 3}(A_2 \circ \epsilon_1 \rightsquigarrow X_3)}_{=0} = 0, \tag{11}$$

where $\tilde{\Lambda}_2$ is zero since there are 0 directed path from X_2 to X_3 and only 1 directed path from X_1 or ϵ_3 to X_3. Intuitively, the general k-path cumulants summation $\tilde{\Lambda}(X_i \rightsquigarrow X_j)$ captures the number of directed paths from the common ancestor to j. Moreover, for any two adjacency vertex $i \to j$ and their common ancestor m, the number of directed paths from m to j is greater or equal to that from m to i, and thus, the causal order can be identified using the following strategy:

Theorem 6 (Identification of PB-SCM)**.** *If there exist $k \in \mathbb{Z}^+$ such that $\tilde{\Lambda}_k(X_i \rightsquigarrow X_j) \neq 0$ and $\tilde{\Lambda}_k(X_j \rightsquigarrow X_i) = 0$ for any two adjacency vertex i and j, then X_i is the parent of X_j.*



(a) Not identifiable. (b) Identifiable.

Figure 5: Illustration of the identifiability of $X \to Y$.

In addition, the k-path cumulants summation $\tilde{\Lambda}_k(X_i \rightsquigarrow X_j)$ will be 'dominated' by the variables (might be the common ancestor or i itself) that has the most paths to j since it is the aggregation of all the directed paths from both common ancestor and i. Therefore, for a non-root vertex, it is possible to be non-identifiable by Theorem 3 if the dominant variable is the common ancestor. Specifically, we provide the graphical implication of such identifiability given as follows:

Theorem 7 (Graphical Implication of Identifiability)**.** *For a pair of causal relationship $i \to j$. The causal order of i, j is identifiable by Theorem 6, if (i) vertex i is a root vertex and $|\mathbf{P}^{i\rightsquigarrow j}| \geq 2$; or (ii) there exists a common ancestor $k \in \arg\max_l \left\{|\mathbf{P}^{l\rightsquigarrow i}| | l \in An(i,j)\right\}$ has a directed path from k to j without passing i in G.*

One of the examples is given in Fig. 5, in which Fig. 5(a) is not identifiable but Fig. 5(b) is identifiable. The reason is that Z is the dominant common ancestor of X, Y, and all directed paths from Z to Y will pass X making it unidentifiable based on Theorem 7. In contrast, Fig. 5(b) includes an additional directed path $Z \to C \to Y$ without passing X making $X \to Y$ identifiable. This intriguingly implies that a denser structure would facilitate the effectiveness of our method.

Generally speaking, once the causal order is identified, one may identify the complete causal structure by orienting edges based on the causal order in the causal skeleton. Such implementation will be provided in the next section. By this, the identifiability of causal structure under PB-SCM is answered.

Learning Casual Structure For PB-SCM

In this section, we propose a causal structure learning algorithm for PB-SCM. Our method involves two steps: learning the skeleton of DAG G and inferring the causal direction using the results developed in Theorem 6.

Learning Causal Skeleton To learn the causal skeleton, instead of using the constraint-based method, we propose a likelihood-based method. This boosts sample efficiency as the likelihood of PB-SCM captures its branching structure but the constraint-based method does not.

Given a set of count data $\mathcal{D}$ and model parameters $\Theta = \left\{\mathbf{A} = [\alpha_{i,j}] \in [0,1]^{|V|\times|V|}, \boldsymbol{\mu} = [\mu_i] \in \mathbb{R}_{\geq 0}^{|V|}\right\}$, the log-likelihood is Markov respect to G, that is $\mathcal{L}(G, \Theta; \mathcal{D}) = \sum_{j=1}^{|\mathcal{D}|}\sum_{i=1}^{|V|} \log P_\Theta\left(X_i = x_i^{(j)} | X_{Pa(i)} = x_{Pa(i)}^{(j)}\right)$. However, calculating the likelihood directly using the probability mass function is costly. Therefore, we propose to calculate the probability mass function by using the probability-generating function (PGF). In detail, for each conditional distribution of X_i, the likelihood can be calculated as follows:

Theorem 8. *Let $G_{X_i|X_{Pa(i)}}(s)$ be the PGF of random variable X_i given its parents variable $X_{Pa(i)}$, we have:*

$$P(X_i = k|X_{Pa(i)} = x_{Pa(i)}) = \frac{1}{k!}\frac{\partial^k G_{X_i|X_{Pa(i)}}(s)}{(\partial s)^k}\Big|_{s=0}$$
$$= \sum_{t_i + \sum_{j\in Pa(i)} t_j = k} \frac{\mu_i^{t_i}\exp(-\mu_i)}{t_i!}\prod_{j\in Pa(i)}\frac{(x_j)_{t_j}\alpha_{j,i}^{t_j}(1-\alpha_{j,i})^{x_j-t_j}}{t_j!}, \tag{12}$$

where $t_j \le x_j$, $(x_j)_{t_j} := \frac{x_j!}{(x_j - t_j)!}$ is the falling factorial, $\mu_i = E[\epsilon_i]$, and ϵ_i is the noise component of X_i.

The result of Eq. (12) can be converted to a polynomial coefficient after taking polynomial multiplication, which can be accelerated via Fast Fourier Transform (FFT) (Cormen et al. 2022). A detailed discussion is given in the supplement.

Generally, the likelihood-based method will tend to produce excessive redundant causal edges. Such effect can be alleviated by introducing the Bayesian Information Criterion (BIC) penalty $d\log(m)/2$ into the $\mathcal{L}(G, \Theta; \mathcal{D})$, where d is the number of edge of G and m is the size of dataset $\mathcal{D}$. The penalized objective function is updated as follows:

$$\mathcal{L}_p(G, \Theta; \mathcal{D}) = \mathcal{L}(G, \Theta; \mathcal{D}) - d\log(m)/2 \tag{13}$$

We maximum the objective function $\mathcal{L}_p(G, \Theta; \mathcal{D})$ by using a Hill-Climbing-based algorithm as shown in Lines 2-6 of Algorithm 1. It mainly consists of two phases. First, we perform a structure searching scheme by taking one step adding, deleting, and reversing the graph G^* in the last iteration, i.e., in Line 4, $\mathcal{V}(G^*)$ represents a collection of the one-step modified graph of G^*. Second, by fixing the graph G', we estimate the parameter Θ' of the model via optimizer with initial values from approximated covariance estimates and then calculate the $\mathcal{L}_p'(G', \Theta'; \mathcal{D})$ in Lines 5. Iterating the two steps above until the likelihood no longer increases. In the end, we transform G^* into a skeleton (Line 6). The correctness of such a procedure can be guaranteed by the consistent property of BIC which is discussed in (Chickering 2002).

Learning Causal Direction Given the learned skeleton, we orient each undirected edge using the k-path cumulants summation, according to Theorem 6. In detail, for each undirected edge $(i, j) \in E$, we calculate $\tilde{\Lambda}_k(X_i \rightsquigarrow X_j)$ and $\tilde{\Lambda}_k(X_j \rightsquigarrow X_i)$ for $k = 1, \ldots, K$ until one of them being zero or k reaches the upper limit K. We then orient the direction based on Theorem 6 (Lines 11-14).

To assess whether $\tilde{\Lambda}_k$ is equal to 0, a bootstrap hypothesis test is conducted (Efron and Tibshirani 1994) while a threshold can be used for orientation once such testing fails. In detail, we calculate the statistic $\tilde{\Lambda}_k^+$ from N resampling dataset $\mathcal{D}^+ \in \{\mathcal{D}_i^+ | \mathcal{D}_{i=1,..,N}^+ \subset \mathcal{D}, \}$. Then, we estimate the distribution $P(\tilde{\Lambda}_k^+)$ by kernel density estimator and centralize it to mean zero. Finally, the p-value of $\tilde{\Lambda}_k$ from the original dataset can be obtained.

Complexity Analysis We provide the complexity of calculating likelihood in the worst cases—when graph is complete. Specifically, the complexity of Eq. (13) is $\mathcal{O}(\sum_{j=1}^{m}\sum_{i=1}^{|V|}\frac{(|V|+x_i^{(j)}-i)!}{(|V|-i)!x_i^{(j)}!})$, by using FFT acceleration, this complexity can be reduced to $\mathcal{O}(\sum_{j=1}^{m}\sum_{i=1}^{|V|}(|V|-i+1)^2 x_i^{(j)}\log(|V|-i+1)^2 x_i^{(j)})$, where m is the sample size.

Algorithm 1: Causal Discovery for PB-SCM

Input: Data set $\mathcal{D}$, Max order K
Output: Learning Causal Graph G

1 $G' \leftarrow$ *empty graph*, $\mathcal{L}_p^* \leftarrow -\infty$;
// Learning Causal Skeleton
2 **while** $\mathcal{L}_p^*(G^*, \Theta^*; \mathcal{D}) < \mathcal{L}_p'(G', \Theta'; \mathcal{D})$ **do**
3 $G^* \leftarrow G'$ with largest $\mathcal{L}_p'(G', \Theta'; \mathcal{D})$
4 **for** *every* $G' \in \mathcal{V}(G^*)$ **do**
5 Estimate Θ' and record score $\mathcal{L}_p'(G', \Theta'; \mathcal{D})$
6 $G \leftarrow$ Transfer G^* to a skeleton
// Learning Causal Direction
7 **for** *each pair* $X_i - X_j \in G$ **do**
8 **for** $k \leftarrow 1 : K$ **do**
9 Obtain $\tilde{\Lambda}_k$ at each side by solving Eq. (9)
10 Test whether $\tilde{\Lambda}_k$ equal to 0 for each side
11 **if** $\tilde{\Lambda}_k(X_i \rightsquigarrow X_j) \neq 0 \wedge \tilde{\Lambda}_k(X_j \rightsquigarrow X_i) = 0$ **then**
12 Orient "$X_i \rightarrow X_j$" in G
13 **if** $\tilde{\Lambda}_k(X_i \rightsquigarrow X_j) = 0 \wedge \tilde{\Lambda}_k(X_j \rightsquigarrow X_i) \neq 0$ **then**
14 Orient "$X_i \leftarrow X_j$" in G
15 **Return** G

Experiment

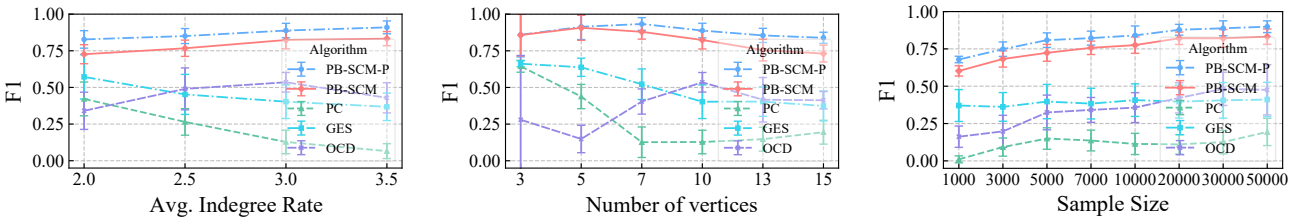
Synthetic Experiments

In this section, we test the proposed PB-SCM on synthetic data. We design control experiments using synthetic data to test the sensitivity of sample size, number of vertices, and different indegree rate. The baseline methods include OCD (Ni and Mallick 2022), PC (Spirtes, Glymour, and Scheines 2000), GES (Chickering 2002). We further provide the results using the true skeleton as prior knowledge (PB-SCM-P) to demonstrate the effectiveness of learning causal direction.

In the sensitivity experiment, we synthesize data with fixed parameters while traversing the target parameter as shown in Fig. 6. The default settings are as follows, sample size=30000, number of vertices=10, indegree rate=3.0, range of causal coefficient $\alpha_{i,j} \in [0.1, 0.5]$, range of the mean of Poisson noise $\mu_i \in [1.0, 3.0]$, the max order of cumulant $K = 4$. Each simulation is repeated 30 times.

As shown in Fig.6, we conduct three different control experiments for PB-SCM. Overall, our method outperforms all the baseline methods in all three control experiments.

In the control experiments of the indegree rate given in Fig. 6(a), as the indegree rate controls the sparse of causal structure, the higher the indegree rate, the less sparse in causal structure leading to a decrease of performance of the baseline methods. In contrast, PB-SCM keeps giving the best results



(a) Sensitivity to Avg. Indegree Rate (b) Sensitivity to Number of vertices (c) Sensitivity to Sample Size

Figure 6: F1 in the Sensitivity Experiments

in all indegree rates. The reason is that our method benefits from the sparsity of the graph and the denser structure would result in more causal order being identified which verified the theoretical result in our work.

In the control experiments of the number of vertices given in Fig.6(b). Our method outperforms all the baseline methods, showing a slight decrease as the number of nodes increases, yet still demonstrating reasonable performance. The reason might be that with an increasing number of vertices, the number of paths for both directions also increases, which requires a higher-order cumulant to obtain the asymmetry. However, estimating high-order cumulant is difficult and has a large variance which leads to a decrease in performance.

In the control experiments of sample size shown in Fig.6(c), as the sample size increases, our method's performance continues to improve and outperforms all the baseline methods. This suggests a sufficient sample size is beneficial for estimating accurate cumulant.

Real World Experiments

We also test the proposed PB-SCM on a real-world football events dataset[1], which contains 941,009 events from 9,074 football games across Europe. For this experiment, we focus on the causal relation in the following count of events: Foul, Yellow card, Second yellow card (abbreviated as 2nd Y. card), Red card, and Substitution. These events possess clear causal relationships according to the rules of the football game. Our goal is to identify the causal relationship from the observed count data while reasoning the possible number of paths between two events as a byproduct of our method.

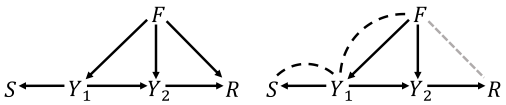
In detail, we employ the bootstrap hypothesis test with 0.05 significance level to test whether $\tilde{\Lambda}_k$ is equal to zero. The result is shown in Table 1. The column of $X \to Y$ shows the highest order of cumulants summation $\tilde{\Lambda}_k(X \rightsquigarrow Y)$ that is not equal to zero while the column of $Y \to X$ shows the lowest order of cumulants summation that equals zero.

The results are given in Fig. 7(b). Generally, PB-SCM successfully identifies five cause-effect pairs, except for Foul → Red card. The possible reason might be attributed to the weak causal influence since only a few serious fouls will result in a red card. Interestingly, We find $\tilde{\Lambda}_2$ (Foul → Yellow card) $\neq 0$, indicating two paths from F or its ancestor to Yellow card.

[1]https://www.kaggle.com/datasets/secareanualin/football-events

Cause (X)	Effect (Y)	$X \to Y$	$Y \to X$
Foul	Yellow card	$\tilde{\Lambda}_{k=2} \neq 0$	$\tilde{\Lambda}_{k=2} = 0$
	2nd Y. card	$\tilde{\Lambda}_{k=3} \neq 0$	$\tilde{\Lambda}_{k=2} = 0$
	Red card	$\tilde{\Lambda}_{k=1} = 0$	$\tilde{\Lambda}_{k=1} = 0$
Yellow card	2nd Y. card	$\tilde{\Lambda}_{k=3} \neq 0$	$\tilde{\Lambda}_{k=2} = 0$
	Substitution	$\tilde{\Lambda}_{k=2} \neq 0$	$\tilde{\Lambda}_{k=2} = 0$
2nd Y. card	Red card	$\tilde{\Lambda}_{k=2} \neq 0$	$\tilde{\Lambda}_{k=2} = 0$

Table 1: The result of real-world dataset experiment.



(a) Ground Truth (b) Result

Figure 7: Football Dataset Result (F:Foul, Y_1: Yellow card, Y_2: Second yellow card, R: Red card, S: Substitution).

This suggests a hidden confounder between Foul and Yellow card, possibly related to the football team's style which also coincides with other path findings. Moreover, the causal direction between Yellow card and Substitution is identified suggesting a hidden confounder or indirect relation exists. This result suggests the effectiveness of our method when dealing with complex real-world scenarios.

Conclusion

In this work, we study the identification of the Poisson branching structural causal model using high-order cumulant. We establish a link between cumulants and paths in the causal graph under PB-SCM, showing that cumulants encompass information about the number of paths between two vertices, which is retrievable. By leveraging this link, we propose the identifiability of the causal order of PB-SCM and its graphical implication. With the identifiability result, we propose a causal structure learning algorithm for PB-SCM consisting of learning causal skeleton and learning causal direction. Our theoretical results and the practical algorithm will hopefully further inspire a series of future methods to deal with count data and move the research of causal discovery further toward achieving real-world impacts in different respects.

Acknowledgments

This research was supported in part by National Key R&D Program of China (2021ZD0111501), National Science Fund for Excellent Young Scholars (62122022), Natural Science Foundation of China (61876043, 61976052), the major key project of PCL (PCL2021A12). ZM's research was supported by the China Scholarship Council (CSC).

References

Al-Osh, M. A.; and Alzaid, A. A. 1987. First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, 8(3): 261–275.

Cai, R.; Qiao, J.; Zhang, Z.; and Hao, Z. 2018. Self: structural equational likelihood framework for causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Cai, R.; Wu, S.; Qiao, J.; Hao, Z.; Zhang, K.; and Zhang, X. 2022. THPs: Topological Hawkes Processes for Learning Causal Structure on Event Sequences. *IEEE Transactions on Neural Networks and Learning Systems*.

Chickering, D. M. 2002. Optimal Structure Identification with Greedy Search. *Journal of machine learning research*, 3(Nov): 507–554.

Choi, J.; Chapkin, R.; and Ni, Y. 2020. Bayesian causal structural learning with zero-inflated poisson bayesian networks. *Advances in neural information processing systems*, 33: 5887–5897.

Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; and Stein, C. 2022. *Introduction to algorithms*. MIT press.

Efron, B.; and Tibshirani, R. J. 1994. *An introduction to the bootstrap*. CRC press.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

McKenzie, E. 1985. Some simple models for discrete variate time series 1. *JAWRA Journal of the American Water Resources Association*, 21(4): 645–650.

Ni, Y.; and Mallick, B. 2022. Ordinal causal discovery. In *Uncertainty in Artificial Intelligence*, 1530–1540. PMLR.

Park, G.; and Raskutti, G. 2015. Learning large-scale poisson dag models based on overdispersion scoring. *Advances in neural information processing systems*, 28.

Park, G.; and Raskutti, G. 2017. Learning Quadratic Variance Function (QVF) DAG Models via OverDispersion Scoring (ODS). *Journal of Machine Learning Research*, 18(224): 1–44.

Pearl, J. 2009. *Causality*. Cambridge university press.

Qiao, J.; Cai, R.; Wu, S.; Xiang, Y.; Zhang, K.; and Hao, Z. 2023. Structural Hawkes Processes for Learning Causal Structure from Discrete-Time Event Sequences. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 5702–5710.

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

Spirtes, P.; Meek, C.; and Richardson, T. 1995. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 499–506.

Steutel, F. W.; and van Harn, K. 1979. Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 893–899.

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

Weiß, C. H. 2018. *An introduction to discrete-valued time series*. John Wiley & Sons.

Weiß, C. H.; and Kim, H.-Y. 2014. Diagnosing and modeling extra-binomial variation for time-dependent counts. *Applied Stochastic Models in Business and Industry*, 30(5): 588–608.

Wiuf, C.; and Stumpf, M. P. 2006. Binomial subsampling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 462(2068): 1181–1195.

Zhang, K.; Schölkopf, B.; Spirtes, P.; and Glymour, C. 2018. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1): 26–29.