Post-Nonlinear Gaussian Causal Models

Grigor Keropyan

A thesis presented for the degree of
Master of Science

Supervisor: Prof. Dr. Mathias Drton
Principal Advisor: M.Sc. David Strieder

Department of Mathematics
Technical University of Munich
Munich, Germany

Submission Date: July 27, 2022

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, July 27, 2022

Acknowledgements

I am extremely grateful to professor Mathias Drton for giving me the opportunity to work with him and write my thesis under his supervision. From our discussions I have learned how to dive into a problem and understand its true nature. Professor Drton has been a great mentor to me during these two year at TUM. In particular, his lecture ”Fundamental of Mathematical Statistics” helped me to carry out the proofs of the theoretical parts of this thesis. I would like to extend my sincere thanks to my advisor, M.Sc. David Strieder for his great help, patience and time. The discussions with him and his comments guided me throughout my work and helped to understand the concepts more thoroughly. Last but not least, I would like to thank Ph.D. Hongjian Shi for insightful discussions.

Abstract

Learning causal structures plays an important role in various fields, ranging from biology and clinical medicine to economics and many others. Since using controlled experiments is often not possible due to cost or ethical reasons, causal discovery based on only observational data is an interesting topic of research. In order to study causal structures researchers often employ Structural Equation Models (SEM). In general the true underlying causal structure can not be uniquely identified. To avoid this problem, constrained version of SEM's can be considered. However, if we would like to have a flexible model which can describe data generation process in real life, the constraints should not be too strict. Post-Nonlinear (PNL) causal models are quite general form of SEM's, which include many other models discussed in the literature, such as Linear SEM's and Additive Noise Models.

This thesis studies both bivariate and multivariate PNL models under the assumption of Gaussian noise. We employ Linear Transformation Models and estimate the involved parameters with Pairwise-Rank likelihood methods. Furthermore we prove asymptotic normality and consistency of our proposed estimates. Using those results, we developed computationally fast algorithms to estimate the causal ordering within PNL models and prove consistency with high probability. The performance of our method is evaluated on simulated data. At the end, general PNL models (not necessarily with Gaussian noise) have been discussed by showing how the same ideas can be transferred to general models.

Contents

1 Introduction

Discovering causal structure of a system is an important question in various disciplines, i.e. biology, economics, clinical medicine, neuroscience and many other domains [Opgen-Rhein and Strimmer, 2007, Glymour et al., 2019, Moneta et al., 2013]. Since using controlled experiments is often not possible due to cost or ethical reasons, causal discovery based on only observational data is an interesting topic of research [Spirtes and Zhang, 2016].

There are various methods for causal structure discovering such as constraint-based or scored-based but these methods identify the causal structure only up to Markov equivalence class [Spirtes et al., 2000], which can leave some causal directions undetermined. In order to deal with this problem Structural Equation Models (SEM) have been proposed, which in their general form have identifiable issues and so they should be constrained (discussed later in this section in more detail).

In the case when data generation process is simple (i.e. effect is a linear combination of its causes added by some noise) it would be sensible to use linear SEM's; however, in reality data generating process might be very complex and linear structural equation models will not be able to approximate the process properly. So, more flexible models are necessary in these situations and the following models have been proposed. Additive Noise Models (ANM) [Hoyer et al., 2008, Peters et al., 2014], where the effect is some nonlinear function of its causes added by a noise independent from the causes. Post-Nonlinear (PNL) models [Zhang and Hyvärinen, 2009], where on top of the ANM there is a nonlinear distortion to obtain the effect and more generally Functional Causal Models (FCM) [Peters et al., 2011], where effect is some nonlinear function from its causes and a noise independent from the causes.

In this work we are mainly focused on the PNL models where the noise variables are from a Gaussian distribution and call them Post-Nonlinear Gaussian (PNLG) causal models. Then, obtain algorithms which will discover underlying causal relationships in the case of PNLG assumptions are satisfied. The rest of the section introduce some notations, give a brief introduction about Directed Acyclic Graphs (which are used in the SEMs) and then give the definition of SEMs and then define PNL and PNLG models.

1.1 Notations

Throughout the work random variables and random vectors are denoted by upper case letters, i.e. X, Y, Z. Their corresponding observed values are denoted by lowercase letters, i.e. x, y, z. Vectors are assumed to be column vectors and for a random vector X its kth element is usually denoted by $X^{(k)}$. For a random vector X its joint distribution function is denoted by $\mathbb{P}^X$. If the joint distribution $\mathbb{P}^X$ of a random vector X has density with respect to Lebesgue measure, usually we denote the density function by $p_X(x)$ if otherwise is not mentioned. Similarly, we denote the distribution function of X by F_X. Matrices are denoted by bold upper case letters, i.e. $\mathbf{X}, \mathbf{A}, \mathbf{B}$. The indicator function is denoted by $I\{\cdot\}$, i.e. $I\{3 < 10\} = 1$ and $I\{2 + 3 = 9\} = 0$. The set $\{1, 2, \dots, m\}$ is denoted by $[1, m]$.

In the text we use different types of convergence of random vectors and here are their notations. For some random vectors $X, X_1, \dots,$ we write $X_n \overset{a.s.}{\to} X$ if we mean X_n converges to X almost surely. We write $X_n \overset{p}{\to} X$ if we mean X_n converges to X in probability and $X_n \overset{d}{\to} X$ if we mean X_n converges to X in distribution. Their corresponding definitions are stated in the Appendix.

1.2 Directed Acyclic Graphs

A graph $\mathcal{G} = (V, \mathcal{E})$ consists of a finite set of nodes V and edges $\mathcal{E} \subseteq V \times V$ of ordered pairs of distinct nodes. Given a set of random variables $X = (X^{(1)}, \dots, X^{(m)})$, $V := \{1, \dots, m\}$ and a graph $\mathcal{G} = (V, \mathcal{E})$ we associate every random variable $X^{(j)}$ with node $j \in V$. The joint distribution of X is denoted by $\mathbb{P}^X$ and marginal distribution of X_j by $\mathbb{P}^{X_j}$. A graph $\mathcal{G}_1 = (V_1, \mathcal{E}_1)$ is called a **subgraph** of $\mathcal{G}$ if $V_1 \subseteq V$ and $\mathcal{E}_1 \subseteq \mathcal{E}$ and $\mathcal{G}$ is called a **super graph** of $\mathcal{G}_1$. If $\mathcal{G}_1$ is a subgraph of $\mathcal{G}$ we write $\mathcal{G}_1 \leq \mathcal{G}$ and if $\mathcal{E}_1 \neq \mathcal{E}$ we say $\mathcal{G}_1$ is **proper subgraph** of $\mathcal{G}$ and and $\mathcal{G}$ is called a **proper super graph** of $\mathcal{G}_1$.

A node i is called a child of j if $(j, i) \in \mathcal{E}$ and is called a parent if $(i, j) \in \mathcal{E}$. If $(i, j) \in \mathcal{E}$ we also write $i \to j$. Children of j is denoted by $\mathbf{CH}_j^{\mathcal{G}} := \{i \in V : (j, i) \in \mathcal{E}\}$ and parents of j by $\mathbf{PA}_j^{\mathcal{G}} := \{i \in V : (i, j) \in \mathcal{E}\}$. Two nodes i and j are called **adjacent** if $(j, i) \in \mathcal{E}$ or $(i, j) \in \mathcal{E}$ and if both holds we say the edge between i and j is **undirected**, otherwise **directed**. A graph is called **complete** if every two nodes are adjacent. **Cliques** of a graph $\mathcal{G}$ are the maximal complete subgraphs of $\mathcal{G}$ (here maximal in a sense of set inclusion). A **path** in $\mathcal{G}$ is a sequence of distinct nodes $j_1, \dots, j_n$ such that j_k and j_{k+1} are adjacent $\forall k = 1, \dots, n-1$ and $n \geq 2$. If $j_k \to j_{k+1}$ $\forall k = 1, \dots, n-1$ path is called **directed** from j_1 to j_n. We say j is a **descendant** of i if there is a directed path from i to j and denote all the descendants of j by $\mathbf{DE}_j^{\mathcal{G}}$ and all non-descendants by $\mathbf{ND}_j^{\mathcal{G}}$. Note that descendants and non-descendants do not contain the node. j_k is called a **collider** in the path if $j_{k-1} \to j_k$ and $j_{k+1} \to j_k$. $\mathcal{G}$ is called a **partially directed acyclic graph (PDAG)** if there is no directed cycle, i.e., if there is no pair (i, j) such that there are directed paths from i to j and from j to i. $\mathcal{G}$ is called **directed acyclic graph (DAG)** if all edges are directed and there is no cycle in $\mathcal{G}$. Permutation π of $1, 2, \ldots, m$ is called a **order** of a DAG $\mathcal{G}$ if for every $i < k$, $\pi(i)$ is not a descendant of $\pi(k)$. Three nodes i, j, k are called **immorality** or **v-structure** if one of them, say j is a child of the others and these parents are not adjacent: $i \to j$, $k \to j$ and $(k, i) \notin \mathcal{E}, (i, k) \notin \mathcal{E}$. The **skeleton** of graph $\mathcal{G}$ is the set of all edges without taking the direction into account, that is all (i, j) such that $i \to j$ or $j \to i$.

In a DAG $\mathcal{G} = (V, \mathcal{E})$, a path between i and j is **blocked** by $\mathbf{S} \subsetneq V$ $(i, j \notin \mathbf{S})$ whenever there is a node k in the path and one of the following holds:

1. $k \in \mathbf{S}$ and k is not a collider in the path, or

2. $k \notin \mathbf{S}$ and k is a collider in the path and $\forall l \in \mathbf{DE}_k^{\mathcal{G}} \implies l \notin \mathbf{S}$.

Given disjoint subsets A, B, C, we say A and B are **d-separated** by C if every path between nodes in A and B is blocked by C. Independence (conditional) is denoted by $\perp\!\!\!\perp$. The joint distribution $\mathbb{P}^X$ of X is said to be **Markov with respect to the DAG** $\mathcal{G}$ if

$$A, B \text{ d-sep. by } C \implies A \perp\!\!\!\perp B | C.$$

for all disjoint sets $A, B, C \subseteq V$. We say $\mathbb{P}^X$ is **faithful to the DAG** $\mathcal{G}$ if

$$A, B \text{ d-sep. by } C \impliedby A \perp\!\!\!\perp B | C.$$

for all disjoint sets $A, B, C \subseteq V$. A distribution satisfies **causal minimality** with respect to graph $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$. Let's denote $\mathcal{M}(\mathcal{G}) := \{\mathbb{P}^X : \mathbb{P}^X \text{ is Markov w.r.t. } \mathcal{G}\}$ all the distributions which are Markov with respect to $\mathcal{G}$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are called **Markov equivalent** if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This holds if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy same set of d-separations. The set of all DAGs that are Markov equivalent to some DAG is called Markov equivalence class. A Markov equivalence class of DAGs can be uniquely represented by a **completed partially directed acyclic graph (CPDAG)**, which is a PDAG that satisfies the following: 1) $i \to j$ in the CPDAG if $i \to j$ in every DAG in the Markov equivalence class, 2) $i - j$ in CPDAG if Markov equivalence class contains a DAG where $i \to j$ and contains another DAG where $i \leftarrow j$.

1.3 Structural Equation Models

A structural equation model (SEM) is defined as a tuple $(\mathcal{S}, \mathbb{P}^\varepsilon)$, where $\mathcal{S} = (S_1, \ldots, S_m)$ is a collection of m equations

$$S_j : \quad X^{(j)} = f_j(\mathbf{PA}_j, \varepsilon_j), \quad j = 1, \ldots, m \tag{1}$$

and $\mathbb{P}^{\varepsilon} = \mathbb{P}^{\varepsilon_1,\ldots,\varepsilon_m}$ is the joint distribution of noise variables and we assume that noise variables are jointly independent. Given a SEM $(\mathcal{S}, \mathbb{P}^{\varepsilon})$ corresponding graph of the structural equation is a DAG where the nodes are $(1, \ldots, p)$ corresponding to the random variables $(X^{(1)}, \ldots, X^{(m)})$ and the edges are determined from the equations by drawing a edge from each node in $\mathbf{PA}_j$ (which appears on the right hand side of equation S_j) to node j.

Although SEM defined in (1) is in a general form and very flexible, it turns out that for each joint distribution of $\mathbb{P}^X$ which is Markov with respect to a graph $\mathcal{G}$ there is SEM with respect to the graph $\mathcal{G}$ that generates distribution $\mathbb{P}^X$. So, only the CPDAG is possible to recover from the joint distribution. The following proposition is the formal version of it.

Proposition 1.1. *(Proposition 9 of Peters et al. [2014])*
Let the joint distribution $\mathbb{P}^X$ of $X = (X^{(1)}, \ldots, X^{(m)})$ is Markov with respect to $\mathcal{G}$ and it has positive density with respect to Lebesgue measure. Then there exists an SEM with graph $\mathcal{G}$ that generates the distribution $\mathbb{P}^X$.

The above Proposition 1.1 shows that in general SEMs it is only possible to obtain the Markov equivalence class of the underlying true graph. So, in order to obtain a unique graph from a joint distribution it is natural to put some restriction on the functions f_j in SEMs. The following subsection defines the restricted SEMs included ANM and PNL models.

1.4 ANM, PNL and PNLG Definitions

In this subsection we introduce the ANM, PNL and PNLG causal models and discuss the identifiability of them. Bivariate ANM models studied in the paper [Hoyer et al., 2008], their identifiability results can be found in [Peters et al., 2014] and the following is the definition.

Definition 1.1. *(ANM Causal Models)*
Additive Noise Model (ANM) is a SEM $(\mathcal{S}, \mathbb{P}^{\varepsilon})$ with equations

$$S_j : \quad X^{(j)} = f_{j,1}(\boldsymbol{PA}_j) + \varepsilon_j, \quad j = 1, \ldots, m \tag{2}$$

where $\boldsymbol{PA}_j$ are the parents of X_j such that the corresponding graph is acyclic. Noise variables ε_j for $j = 1, 2, \ldots, m$ are independent of corresponding parents random variables $\boldsymbol{PA}_j$ i.e., $\varepsilon_j \perp\!\!\!\perp \boldsymbol{PA}_j$, and are jointly independent. The functions $f_{j,1}$ for $j = 1, 2, \ldots, m$ are potentially nonlinear effects of causes $\boldsymbol{PA}_j$.

ANM models allow only that noise variable can be part of the effect in additive way, which is quite restrictive and in real world data generating processes might be more complex. The following PNL model overcomes this problem and it is strictly larger class of ANM's, i.e. taking $f_{j,2}$ as identity gives ANM models and taking $f_{j,2}$ as an exponent, noise will be part of an effect multiplicative way.

Definition 1.2. *(PNL Causal Models)*
Post-Nonlinear (PNL) causal model is a SEM $(\mathcal{S}, \mathbb{P}^{\varepsilon})$ *with equations*

$$S_j: \quad X^{(j)} = f_{j,2}(f_{j,1}(\boldsymbol{PA}_j) + \varepsilon_j), \quad j = 1, \ldots, m \tag{3}$$

where $\boldsymbol{PA}_j$ *are the parents of* X_j *such that the corresponding graph is acyclic. Noise variables* ε_j *for* $j = 1, 2, \ldots, m$ *are independent of corresponding parents random variables* $\boldsymbol{PA}_j$ *i.e.,* $\varepsilon_j \perp\!\!\!\perp \boldsymbol{PA}_j$*, and are jointly independent. The functions* $f_{j,1}$ *for* $j = 1, 2, \ldots, m$ *are potentially nonlinear effects of causes* $\boldsymbol{PA}_j$ *and the functions* $f_{j,2}$ *are invertible post-nonlinear distortions in variable* $X^{(j)}$ *for* $j = 1, 2, \ldots, m$*.*

Bühlmann et al. [2014] studied ANM models in the case when $f_{j,1}$ functions are additive and developed an algorithm to discover underlying graph and proved that discovered causal order is consistent when number of observations goes to infinity. However, the PNL models are not so well studied. Zhang and Hyvärinen [2009] proved identifiability results of bivariate PNL case and suggested an algorithm for it. Peters et al. [2014] after studying identifiability of ANM models, mentions that similar identifiability results hold for the PNL multivariate case, but the estimation of the causal relations is not discussed. Lately Uemura and Shimizu [2020], Uemura et al. [2022] suggested neural network approach for estimating the underlying causal structure for bivariate and multivariate cases respectively. To my knowledge [Uemura et al., 2022] is the only practical algorithm exist for the multivariate PNL models and we discuss it in Section 4 more in detail.

Since, throughout the text we use PNL Gaussian models many time let us define them in the following, which adds a restriction to PNL model that noise variables are standard normal.

Definition 1.3. *(PNLG Causal Models)*
Post-Nonlinear Gaussian (PNLG) causal model is a SEM $(\mathcal{S}, \mathbb{P}^{\varepsilon})$ *with equations*

$$S_j: \quad X^{(j)} = f_{j,2}(f_{j,1}(\boldsymbol{PA}_j) + \varepsilon_j), \quad j = 1, \ldots, m \tag{4}$$

where $\boldsymbol{PA}_j$ *are the parents of* X_j *such that the corresponding graph is acyclic. Noise variables* ε_j *for* $j = 1, 2, \ldots, p$ *are independent of corresponding parents random variables* $\boldsymbol{PA}_j$ *i.e.,* $\varepsilon_j \perp\!\!\!\perp \boldsymbol{PA}_j$*, are jointly independent and distribution is standard normal, i.e.* $\varepsilon_i \sim \mathcal{N}(0, 1)$*. The functions* $f_{j,1}$ *for* $j = 1, 2, \ldots, m$ *are potentially* *nonlinear effects of causes $\boldsymbol{PA}_j$ and the functions $f_{j,2}$ are invertible post-nonlinear distortions in variable $X^{(j)}$ for $j = 1, 2, \ldots, m$.*

Note that we need the assumption of mean zero and unit variance of noise variables to have identifiability of the parameters in the model, if noise variables do not have mean zero and unit variance they can be consumed in functions $f_{j,2}$ and $f_{j,1}$, which is discussed later in the text in more detail.

Our main goal is to deal with PNLG models but we will also consider the PNL models their advantages and disadvantages.

Now, let us consider equations in PNL model (3) to understand the structure of the remaining text. Firstly, we will look at one equation in SEM for fixed $j \in [1, m]$ in order to understand its properties. For this purpose we will simplify more and assume $f_{j,1}$ is linear function and it will result $Y = f_{j,2}(X^T\beta + \varepsilon)$. In our first step we look at the case when the functions $f_{j,1}$ are linear which is called Linear Transformation Models and study it in Section 2. Then we will move in general case and study the model $Y = f_{j,2}(f_{j,1}(X) + \varepsilon)$ in Section 3. Note that in the above mentioned two Sections we will have slightly different notations for convenience and assume that the function (and its inverse) $f_{j,2}$ is strictly increasing. In Section 4 we study estimation of the causal structure of PNLG and PNL models in bivariate and multivariate cases separately using the results from Sections 2 and 3. Then, Section 5 reviews the experimental results and comparison with other existing methods and Section 6 concludes the text. Some of the proofs which are technical postponed in Appendix A.

2 Linear Transformation Models

This section is a review of Linear Transformation Models in the literature, how the parameters are estimated in that models and development of new estimation methods. The experimental results are presented in the Section 5.

Let's assume that our data (X, Y) is generated by the following Linear Transformation Model

$$h(Y) = X^T\beta_0 + \varepsilon, \tag{5}$$

where $X \in \mathbb{R}^m, Y \in \mathbb{R}$ for some $m \in \mathbb{N}$, $\beta_0 = [\beta_{0,1}, \ldots, \beta_{0,m}]^T$, ε is a noise variable independent of X, i.e. $\varepsilon \perp\!\!\!\perp X$ and $h : \mathbb{R} \to \mathbb{R}$ is invertible and strictly increasing function, i.e. $a < b$ implies $h(a) < h(b)$. In general X is not necessarily random; however, since in the Post-Nonlinear causal models it is generated by some structural equations, here we are assuming it is a random. In case it is assumed to be not random will be mentioned explicitly.

Before analyzing the above Linear Transformation Model let us show that it is a quite general model in a sense that it includes many other models that have been a interest in the literature.

1. Linear Regression Model:

   In the case of function h is specified model (5) is a simple linear regression model and we can estimate the β by maximum likelihood method or least squares method.

2. Box and Cox transformation:

   In the case of function h is specified in the following parametric way

   $$h(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

   gives the Box and Cox transformation model [Box and Cox, 1964]. The transformation is specified by a parameter λ, which produces from linear ($\lambda = 1$) to log ($\lambda = 0$) transformed regression models. Box and Cox [1964] considered Gaussian distributed error and estimated the parameter λ by maximizing the likelihood.

3. Proportional Hazard (Cox PH) Model

   Let us assume that the function h is specified in the following way

   $$h(y) = \log(-\log(1 - g(y))),$$

   where $g(y) = 1 - exp(-\int_0^y r(t)dt)$, $r(t) = \frac{f(y)}{1-F(y)}$ for for probability density and cumulative density functions of Y correspondingly and noise is from extreme value distribution. Then this models is called Proportional Hazard model and β is estimated using partial likelihood method [Cox, 1972, 1975].

Now let us look at the Linear Transformation model without any further assumptions. The following papers [Cuzick, 1988, Pettitt, 1982, 1984, 1987, Doksum, 1987, Clayton and Cuzick, 1985, 1986] are some existing results of Linear Transformation Models in the literature where the distribution of the noise variable is usually assumed to be known. In the [Han, 1987, Abrevaya, 1999a, Sherman, 1993] papers Maximum Rank Correlation (MRC) is introduced for estimation of β in Linear Transformation Model with unknown noise distribution and asymptotic properties proved. The Maximum Rank (MR) estimator is introduced by Cavanagh and Sherman [1998] and PDR (there are several versions of these estimators) estimators are introduced by Abrevaya [1999a,b, 2003]. Lately, Yu et al. [2021] introduced a Pairwise Rank Likelihood (PRL) method to estimate the β and distribution function of the noise variable and their asymptotic properties derives. We will consider this method in more detail later to obtain estimator in the case when the distribution of the noise variable is Gaussian.

Before looking at the methods to estimate β let us analyze the Linear Transformation Model and understand in which scenarios it is not possible to identify the parameters. Let $(X_i, Y_i)_{i=1}^n$ be n i.i.d. copies from the model (5), $R = (R_1, \dots, R_n)$ be the ranks of $Y = [Y_1, \dots, Y_n]^T$. Denote $Z_i := h(Y_i)$ and note, that by the assumption of strictly increasing h the ranks of Y and $Z := [Z_1, \dots, Z_n]^T$ are exactly the same. Assume observed data is $(x_i, y_i)_{i=1}^n$ and $r = (r_1, \dots, r_n)$ be the observed ranks, i.e. ranks of $[y_1, \dots, y_n]$ and denote $\mathbf{X} := [x_1^T, \dots, x_n^T]^T \in \mathbb{R}^{n \times m}$ the design matrix.

Note that, replacing $X^T\beta$ by $\alpha + X^T\beta$ and ε by $\varepsilon - \alpha$ in the model (5) all the model assumptions still be satisfied. Moreover, the same is true if for any $\sigma > 0$, we replace $h(Y)$ by $\sigma h(Y)$, ε by $\sigma\varepsilon$ and β_0 by $\sigma\beta_0$. So, without loss of generality we can assume that the noise variable satisfies the following requirements

$$\mathbb{E}[\varepsilon] = 0 \text{ and } Var(\varepsilon) = 1, \tag{6}$$

otherwise, it will not be possible to identify the parameters uniquely from the model (5), i.e. assume true parameters in the model are h, β_0, ε is a random variable from some arbitrary distribution and (X, Y) is generated from the model, then we also have $h'(Y) = \alpha + X^T\beta_0' + \varepsilon'$, where $h' := \sigma h, \beta_0' := \sigma\beta_0$ and $\varepsilon' = \sigma\varepsilon - \alpha$, which means that the parameters of the model will not be uniquely identified if we do not put any more restrictions like (6).

In the following most methods are based on some kind of rank likelihood. First, let us investigate some properties of the marginal rank likelihood and discuss some of the limits of information contained in ranks. We denote the marginal rank likelihood by

$$\begin{aligned} L(\beta; r, \mathbf{X}) &:= \mathbb{P}(Y_j^\beta \text{ has rank } r_j \text{ for } j \in [1, n]) = \mathbb{P}(R^\beta = r) \\ &= \mathbb{P}(Z_j^\beta \text{ has rank } r_j \text{ for } j \in [1, n]) \\ &= \int_{\{z \in r\}} \prod_{i=1}^n f_0(z_i - x_i^T\beta)\, dz_i \ , \end{aligned} \tag{7}$$

where the notation $\{z \in r\}$ is the set $\{z_1, \dots, z_n\}$ which have ranks r, $\mathbf{X} := [x_1^T, \dots, x_n^T]^T \in \mathbb{R}^{n \times m}$, f_0 is the probability density function of noise variable ε_i for $i \in [1, n]$ and $Y_j^\beta := h^{-1}(x_j^T\beta + \varepsilon_i)$, $Z_j^\beta := h(Y_j^\beta)$.

Now, having defined the marginal rank likelihood, let's understand in what cases it is not possible to obtain any information from it. For simplicity let's consider the case when there is only one covariate i.e., $m = 1$ and analyze the range of β's which will not change the rank likelihood. The following proposition shows that for large β's the value of rank likelihood is the same with high probability in the Gaussian noise case, i.e. $\varepsilon_i \sim \mathcal{N}(0,1)$.

Proposition 2.1. *Assume there are n observed samples $(x_i, y_i)_{i=1}^n$ from model (5) for some fixed β_0, corresponding $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $x_i \neq x_j$ for all $i \neq j$. Let's fix arbitrary small $\delta > 0$ and define $K_{\delta,n} := \left\lfloor \sqrt{2\log\frac{2n}{\delta}} \right\rfloor + 1$ and $m_X := \min_{i\neq j}\{|x_i - x_j|\}$. Then with probability at least $1-\delta$*

$$L(\beta; r, \mathbf{X}) \text{ is constant for all } \beta > \frac{2K_{\delta,n}}{m_X} \text{ or } \beta < -\frac{2K_{\delta,n}}{m_X},$$

where r is the ranks of y_i's.

Proof. The idea of the proof is to obtain some bound for the noise variables with high probability and using it make the gaps between $x_i^T\beta$'s large enough that adding a noise will preserve the ranks with the same probability.

For exact proof let's estimate the probability that all noise variables are bounded by $K_{\delta,n}$ that is

$$\begin{aligned}\mathbb{P}(|\varepsilon_i| \leq K_{\delta,n} \text{ for all } i \in [1,n]) &= 1 - \mathbb{P}(\exists i \in [1,n] \text{ such that } |\varepsilon_i| > K_{\delta,n}) \\ &\geq 1 - \sum_{i=1}^n \mathbb{P}(|\varepsilon_i| > K_{\delta,n}) = 1 - n\mathbb{P}(|\varepsilon_1| > K_{\delta,n}) \\ &\geq 1 - 2ne^{-K_{\delta,n}^2/2} \geq 1 - \delta,\end{aligned}$$

where the first inequality follows from the union bound and the second one is a direct application of Chernoff(i. e., exponential Markov inequality) bound for standard normal random variables.

Now having a uniform bound on the noise variables, we can show that for arbitrary $\beta > \frac{2K_{\delta,n}}{m_X}$ the ranks of $Z_i^\beta := x_i^T\beta + \varepsilon_i$ are the same as ranks of x_i's. For that purpose assume that all noise variables are bounded by $K_{\delta,n}$ which we know from the uniform bound that holds with at least $1-\delta$ probability. Let's fix two different indices i and j. Without loss of generality we can assume that $x_i < x_j$ so,

$$\begin{aligned}Z_j^\beta - Z_i^\beta &= x_j\beta + \varepsilon_j - x_i\beta - \varepsilon_i = (x_j - x_i)\beta + \varepsilon_j - \varepsilon_i \\ &\geq m_X\beta + \varepsilon_j - \varepsilon_i > 2K_{\delta,n} + \varepsilon_j - \varepsilon_i \geq 0.\end{aligned}$$

This shows that the ranks of Z_i^β's are preserved for the mentioned range of β's. So, from 7 we have $L(\beta; r, \mathbf{X}) = \mathbb{P}(Z_j^\beta \text{ has rank } r_j \text{ for } j \in [1,n])$ and so, it is not changing with probability at least $1-\delta$ as the Z_j^β's will be fixed. Note that the case where $\beta < -\frac{K_{\delta,n}}{m_X}$ is the same besides we will have same ranks for $-Z_i^\beta$'s and x_i's, which completes the proof.

□

In order to have an idea how large can be the value $\frac{2K_{\delta,n}}{m_X}$ in the Proposition 2.1, let's assume $m_X = 1$, which means that minimum distance between x_i's is 1 and take $n = 1000$ samples with $\delta = 0.01$. Inserting this quantities in the formula stated in the Proposition we obtain $\frac{2K_{\delta,n}}{m_X} = 10$. This shows that in the case when x_i's are far far each other, i.e. difference is at least 1, then with probability 0.99 marginal rank likelihood cannot differentiate between betas larger than 10.

Note that in the Proposition 2.1 we needed the Gaussian distributed noise variables only to have exponential decaying tails of this distribution. So, the similar statement is true for all noise variables which have a exponentially decaying tails, i.e. distribution of subgaussian random variables. In case of heavy tail distributions like subexponentials, we will have larger bound on β.

The remaining part of this Section reviews some existing results on the Linear Transformation Models, describes the Algorithms to estimate the parameters, reviews their properties (asymptotic) and develops new method and studies its properties (asymptotic).

2.1 Monte Carlo Method

In order to be be able to use Monte-Carlo methods to estimate the rank likelihood, firstly, it should be represented as an expectation. Assuming noise variable is standard normal, from equation 7 we have

$$\begin{aligned}
L(\beta; r, \mathbf{X}) &= \int_{\{z\in r\}} \prod_{i=1}^n \phi(z_i - x_i^T\beta)\, dz_i \;= \int_{\{z\in r\}} \prod_{i=1}^n \frac{\phi(z_i - x_i^T\beta)}{\phi(z_i)} \phi(z_i)\, dz_i \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,I_n)} \left[\prod_{i=1}^n \frac{\phi(Z_i - x_i^T\beta)}{\phi(Z_i)} \middle| \{Z \in r\} \right] \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,I_n)} \left[\prod_{i=1}^n \frac{\phi(Z_{r_i} - x_i^T\beta)}{\phi(Z_{r_i})} \middle| \{Z \in r\} \right] = \frac{1}{n!}\mathbb{E}\left[\prod_{i=1}^n \frac{\phi(Z_{r_i} - x_i^T\beta)}{\phi(Z_{r_i})} \right],
\end{aligned}$$

where ϕ is a pdf function of standard normal random variable and $Z_1 < Z_2 < \cdots < Z_n$ are order statistics from standard normal distribution of size n. The last equality in the above follows from the fact that $(Z_i)_{i=1}^n$ are i.i.d. and so it is symmetric and $\{Z \in r\}$ is equiprobable for each r and since there are $n!$ possible rankings we have $\mathbb{P}(\{Z \in r\}) = \frac{1}{n!}$. Using the above equality, Doksum [1987] suggested to estimate it by Monte Carlo method. The Monte Carlo estimator of the rank likelihood can be written as

$$n!\hat{L}(\beta; r, \mathbf{X}) := \frac{\sum_{i=1}^{M} \prod_{j=1}^{n} \frac{\phi(Z^i_{r_j} - x_j^T \beta)}{\phi(Z^i_{r_j})}}{M}, \tag{8}$$

where $Z^i := \{Z^i_1, \ldots, Z^i_n\}^T$ is the order statistics of ith sample from the standard normal distribution and M is the number of samples. Using the strong law of large numbers, it is clear that

$$\hat{L}(\beta; r, \mathbf{X}) \to L(\beta; r, \mathbf{X}) \text{ almost surely as } M \to \infty. \tag{9}$$

In practice, we generate $z^i := \{z^i_1, \ldots, z^i_n\}^T$ order statistics of ith generation sample from the standard normal distribution of size n and substitute them in equation 8.

Now let's understand how the quantity in equation 8 can be maximized. Denoting

$$l_i := \prod_{j=1}^{n} \frac{\phi(z^i_{r_j} - x_j^T \beta)}{\phi(z^i_{r_j})}$$

and recalling that ϕ is the pdf of standard normal distribution we obtain

$$\begin{aligned}
-2 \log l_i &= -2 \sum_{j=1}^{n} \log \left(\frac{\phi(z^i_{r_j} - x_j^T \beta)}{\phi(z^i_{r_j})} \right) = \sum_{j=1}^{n} ((z^i_{r_j} - x_j^T \beta)^2 - (z^i_{r_j})^2) \\
&= -2 \sum_{j=1}^{n} z^i_{r_j} x_j^T \beta + \sum_{j=1}^{n} (x_j^T \beta)^2 = -2 z^i_r \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta \\
&= b^T \mathbf{X}^T \mathbf{X} b - 2 z^i_r \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta - b^T \mathbf{X}^T \mathbf{X} b \\
&= (b_i - \beta)^T \mathbf{X}^T \mathbf{X} (b_i - \beta) - SSF_i,
\end{aligned}$$

where $z^i_r := (z_{r_1}, \ldots, z_{r_n})$ is the ith sample of normal distribution of size n which has rank r, $b_i := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T z^i_r$ is the usual least square estimate of β for the response vector z^i_r and SSF_i is fitted sum of squares, i.e. $SSF_i := (\hat{z}^i_r)^T \hat{z}^i_r = b_i^T \mathbf{X}^T \mathbf{X} b_i$. Note that b_i and SSF_i does not depend on β. Inserting this in (8) gives the monte carlo estimate of the rank likelihood, i.e.

$$n!\hat{L}(\beta; r, \mathbf{X}) = \frac{\sum_{i=1}^{M} w_i exp(-\frac{1}{2}(b_i - \beta)^T \mathbf{X}^T \mathbf{X} (b_i - \beta))}{M}, \tag{10}$$

where $w_i := exp(\frac{1}{2} SSF_i)$. This estimate is proportional to mixture of multivariate normal densities. To maximize the likelihood we calculate the gradient

$$\nabla_\beta \hat{L}(\beta; r, \mathbf{X}) = \frac{1}{Mn!} \sum_{i=1}^{M} w_i u_i(\beta) \mathbf{X}^T \mathbf{X} (b_i - \beta) = \frac{\mathbf{X}^T \mathbf{X}}{Mn!} \sum_{i=1}^{M} w_i u_i(\beta)(b_i - \beta) = 0,$$

where

$$u_i(\beta) := exp\left(-\frac{1}{2} (b_i - \beta)^T \mathbf{X}^T \mathbf{X} (b_i - \beta) \right). \tag{11}$$

Assuming $\mathbf{X}$ has full column rank, it implies that $\mathbf{X}^T\mathbf{X}$ is invertible and so, the above is equivalent to

$$\begin{aligned}\sum_{i=1}^{M} w_i u_i(\beta)(b_i - \beta) = 0 &\iff \beta \sum_{i=1}^{M} w_i u_i(\beta) = \sum_{i=1}^{M} w_i u_i(\beta) b_i \\ &\iff \beta = \sum_{i=1}^{M} \frac{w_i u_i(\beta)}{\sum_{j=1}^{M} w_j u_j(\beta)} b_i.\end{aligned}$$

Using the reweighted method we obtain iteratively reweighted least square algorithm

$$\beta^{k+1} = \sum_{i=1}^{M} \frac{w_i u_i(\beta^k)}{\sum_{j=1}^{n} w_j u_j(\beta^k)} b_i, \tag{12}$$

where $k = 0, 1, \dots$ and β^0 is some initial value for β. Thus Monte Carlo algorithm can be described as in the following Algorithm 1.

Algorithm 1: Monte Carlo Algorithm

```
Require: $\mathbf{X}$, $Y$, M, $maxit$ (maximum iterations).
$r \leftarrow rank(Y)$
for $i = [1, M]$ do
    $z^i := (z_1^i, \dots, z_n^i) \leftarrow n$ i.i.d. sample from $\mathcal{N}(0,1)$
    $z_r^i \leftarrow (z_{r_1}^i, \dots, z_{r_n}^i)$
    $b_i \leftarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T z_r^i$
    $\log w_i \leftarrow b_i^T \mathbf{X}^T \mathbf{X} b_i / 2$
end for
$\beta^0 \leftarrow$ random vector with the size of columns of $\mathbf{X}$
for $k = [0, maxit - 1]$ do
    for $i = [1, M]$ do
        $\log u_i(\beta^k) \leftarrow -\frac{1}{2}(b_i - \beta^k)^T \mathbf{X}^T \mathbf{X}(b_i - \beta^k)$
        $logweights_i \leftarrow \log u_i(\beta^k) + \log w_i$
    end for
    $rw \leftarrow softmax(logweights)$
    $\beta^{k+1} = \sum_{i=1}^{M} rw_i b_i$
end for
$\hat{\beta}_{MC} \leftarrow \beta^{maxit}$
Return $\hat{\beta}_{MC}$.
```

In the Algorithm 1 logarithms of the weights are introduced in order to circumvent the problem of overflow/underflow of the weights and *softmax* function should be numerically stable. One such example is describe in the Algorithm 2.

Algorithm 2: Numerically Stable Softmax Algorithm

```
Require: $x := (x_1, \dots, x_n)^T$.
$c \leftarrow max_{i \in [1,n]}\{x_i\}$
$y \leftarrow (x_1 - c, \dots, x_n - c)^T$
$s \leftarrow \sum_{i=1}^{n} exp(y_i)$
$z \leftarrow (exp(y_1)/s, \dots, exp(y_n)/s)^T$
Return $z$.
```

Note that in Algorithm 2 the fact that $softmax(x) = softmax(x - c)$ is used and after subtracting the max element from all elements in the input x we will end up at least one element which is zero and others are at most zero, so, overflow is not an option already. For the case of underflow since there at least one element equal to zero implies that underflow cannot happen to all elements, so, it will produce reasonable result.

In the paper [Doksum, 1987] no general asymptotic property of $\hat{\beta}_{MC}$ has been showed, but it is mentioned (Remark 3.2 in [Doksum, 1987]) that for proportional hazard two-sample model case $\hat{\beta}_{MC}$ is asymptotically normal.

2.2 Substitution of Conditional Expectation Method

This subsection is a review and development of corresponding algorithm of [Cuzick, 1988] paper. In the paper it is assumed that distribution of the noise variable is known, but is not necessarily standard normal. From 7 we have that rank likelihood is the following

$$L(\beta; r, \mathbf{X}) = \int_{\{z \in r\}} \prod_{i=1}^{n} f_0(z_i - x_i^T \beta)\, dz_i \ , \tag{13}$$

where f_0 is the probability density function of the noise variable. In order to maximize the likelihood we calculate the gradient of the log likelihood, which gives

$$\nabla_\beta \log L(\beta; r, \mathbf{X}) = \sum_{i=1}^{n} x_i \mathbb{E}_R \left[-\frac{f_0'(z_i - x_i^T \beta)}{f_0(z_i - x_i^T \beta)} \right] = \sum_{i=1}^{n} x_i \mathbb{E}_R [h_0(z_i - x_i^T \beta)] = 0, \tag{14}$$

where

$$\mathbb{E}_R \left[g(z_i) \right] = \frac{\int_{\{z \in r\}} g(z_i) \prod_{j=1}^{n} f_0(z_j - x_j^T \beta)\, dz_j}{\int_{\{z \in r\}} \prod_{j=1}^{n} f_0(z_j - x_j^T \beta)\, dz_j}$$

is the conditional expectation of $g(z_i)$ given the ranks R and β for some function g and

$$h_0(t) := -\frac{f_0'(t)}{f_0(t)}.$$

Since, $\mathbb{E}_R[h_0(z_i - x_i^T \beta)]$ depends on β not in a simple way, some substitutions of it have been proposed. For instance Clayton and Cuzick [1985, 1986] suggested to replace it by $h_0(\mathbb{E}_R[z_i] - x_i^T\beta)$ which makes no replacement for the case of Gaussian noise variable. Cuzick [1988] suggested the following substitution and studied asymptotic normality. Let

$$\hat{F}(z) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{z_i \leq z\} \tag{15}$$

be the adjusted empirical distribution function of z_i and denoting the CDF of randomly chosen z_i by $F_\beta(z)$, we will have

$$F_\beta(z) = \frac{1}{n}\sum_{i=1}^{n} F_0(z - x_i^T\beta), \tag{16}$$

where F_0 is the CDF of the noise variable. So, the substitution is to replace $\mathbb{E}_R[h_0(z_i - x_i^T\beta)]$ by $h_0(\bar{z}_i^\beta - x_i^T\beta)$, where

$$\bar{z}_i^\beta := F_\beta^{-1}(\hat{F}(z_i)) \text{ for } z_i := h(y_i) \tag{17}$$

and denote $\bar{z}^\beta := [\bar{z}_1^\beta, \ldots, \bar{z}_n^\beta]$. Note that for the computation of $\bar{z}^\beta$ for fixed β we only need to know the ranks of z_i's which we know as it is the same as the ranks of y_i's. Thus, equation 14 will be replaced by

$$\sum_{i=1}^{n} x_i h_0(\bar{z}_i^\beta - x_i^T\beta) := 0 \tag{18}$$

After the estimation of β as $\hat{\beta}$, the value of $h(y_i)$ can be estimated by $\bar{z}_i^{\hat{\beta}}$, i.e.

$$\widehat{h}(y_i) := \bar{z}_i^{\hat{\beta}} \tag{19}$$

Now moving to the case for standard normal noise variable, we have $h_0(t) = t$, so equation 18 becomes

$$\begin{aligned} 0 =: \sum_{i=1}^{n} x_i(\bar{z}_i^\beta - x_i^T\beta) &= \mathbf{X}^T\bar{z}^\beta - \mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T F_\beta^{-1}(\hat{F}(z)) - \mathbf{X}^T\mathbf{X}\beta \\ &= \mathbf{X}^T F_\beta^{-1}\left(\frac{r}{n+1}\right) - \mathbf{X}^T\mathbf{X}\beta := G(\beta), \end{aligned}$$

where the CDF functions and its inverses applied element wise to vector arguments. Cuzick [1988] in the Extensions section part 5 claims that $G(\beta) = 0$ has a unique solution, but it turns out to be wrong as stated in the correction part of the paper that there is an error in the proof of the claim in Lemma 5 that there is a unique solution for $G(\beta)$ for $m = 1$ as well as $m > 1$ (note, $G(\beta)$ corresponds to the equation 7 in the paper when noise variable is standard normal). However, assuming that $\hat{\beta}$

which makes $G(\beta)$ zero we can find by the fixed point iteration algorithm, we will have

$$\beta^{k+1} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T F_{\beta^k}^{-1}\left(\frac{r}{n+1}\right) \tag{20}$$

for kth iteration, where β^0 is some initial value for β. So, it will produce an estimator which could satisfies the properties that are established in Cuzick [1988] and described at the end of this subsection. The description of the procedures is described in the following Algorithm 3.

Algorithm 3: Fixed Point Algorithm

```
Require: $\mathbf{X}$, $Y$, $maxit$ (maximum iterations), root finding algorithm.
$r \leftarrow rank(Y)$
$\beta^0 \leftarrow$ random vector with the size of columns of $\mathbf{X}$
for $k = [0, maxit - 1]$ do
    for $i = [1, n]$ do
        $t_i \leftarrow F_{\beta^k}^{-1}\left(\frac{r_i}{n+1}\right)$, using root finding algorithm for $F_{\beta^k}(t_i) - \frac{r_i}{n+1} = 0$
    end for
    $t \leftarrow (t_1, \ldots, t_n)^T$
    $\beta^{k+1} \leftarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T t$
end for
$\hat{\beta}_{FP} \leftarrow \beta^{maxit}$
Return $\hat{\beta}_{FP}$.
```

In the above Algorithm 3 the most computational expensive part is the computations of t_i's, which has to be done at each iteration of fixed point for all $i \in [1, n]$. It is not possible to circumvent this computation since it is not possible to obtain the closed form of the inverse of F_β function.

In order to reduce the computational complexity, we can propose stochastic version of this algorithm which will reduce the complexity significantly. Since, only the ranks of y_i's are used in the algorithm, reasonable size of random subset of y_i's could approximate the ranks and so we can take corresponding x_i's and do the fixed point updated at each step. Thus, it will be the following Algorithm 4.

```
Algorithm 4: Stochastic Fixed Point Algorithm
  Require: $\mathbf{X}$, $Y$, $bs$ (batch size), $maxit$ (maximum iterations), root finding
           algorithm.
  $\beta^0 \leftarrow$ random vector with the size of columns of $\mathbf{X}$
  for $k = [0, maxit - 1]$ do
      $I \leftarrow$ random subset of $[1, 2, \ldots, n]$ with size $bs$
      $Y_{bs} \leftarrow Y[I]$, take the elements from $Y$ corresponding to indices $I$
      $\mathbf{X}_{bs} \leftarrow \mathbf{X}[I]$ , take the rows from $\mathbf{X}$ corresponding to indices $I$
      $r_{bs} \leftarrow rank(Y)$
      for $i = [1, bs]$ do
          $t_i \leftarrow F_{\beta^k}^{-1}\left(\frac{r_{bs,i}}{n+1}\right)$ , using root finding algorithm for $F_{\beta^k}(t_i) - \frac{r_{bs,i}}{n+1} = 0$
      end for
      $t \leftarrow (t_1, \ldots, t_{bs})^T$
      $\beta^{k+1} \leftarrow (\mathbf{X}_{bs}^T \mathbf{X}_{bs})^{-1} \mathbf{X}_{bs}^T t$
  end for
  $\hat{\beta}_{SFP} \leftarrow \beta^{maxit}$
  Return $\hat{\beta}_{SFP}$.
```

In the Algorithm 4 batch size bs can be chosen as big as possible to run the algorithm, but since in the experiments different batch sizes produced similar results, we have fixed it to be 64.

The main theoretical result of Cuzick [1988] are the followings.

Theorem 2.1. *(Theorem 1 in [Cuzick, 1988])*
If the assumptions (A1)-(A6) in [Cuzick, 1988] are satisfied, then with probability tending to one as $n \to \infty$ there exists a solution $\hat{\beta}$ to the equation

$$\sum_{i=1}^{n} x_i h_0(\bar{z}_i^{\beta} - x_i^T \beta) = 0,$$

such that as $n \to \infty$

$$\sqrt{n}(\hat{\beta} - \beta_0) \to \mathcal{N}(0, \sigma^2),$$

where σ^2 as is specified in Theorem 1 at [Cuzick, 1988] and β_0 is the true value of β given in model (5).

Note that the proofs in [Cuzick, 1988] are presented only for the case of $m = 1$ but it is mentioned that they are straightforward for all m.

2.3 Expecting Rank Method

Since, Expected Rank method is using the expectation of the ranks of Y_i's, let's prove the following lemma about it.

Lemma 2.1. *Let R_j be the rank of Y_i among $\{Y_i\}_{i=1}^n$ in the model 5 as well as the rank of $Z_i : -f_2^{-1}(Y_i)$, for all $j \in [1, n]$ and the noise variable is standard normal. Then*

$$\mathbb{E}[R_j] = 1 + \sum_{i \neq j} \Phi\left(\frac{(x_j - x_i)^T \beta}{\sqrt{2}}\right),$$

where Φ is the CDF of standard normal distribution.

Proof. Let's fix any $j \in [1, n]$. Since the rank R_j is the number of Z_i's ($i \neq j$) less than Z_j plus one(note that since the distribution of Z_i's are continuous, i.e. Gaussian, equality holds with probability zero), we obtain

$$\begin{aligned}
\mathbb{E}[R_j] &= \mathbb{E}\left[1 + \sum_{i \neq j} \mathbb{1}_{\{Z_i < Z_j\}}\right] = 1 + \sum_{i \neq j} \mathbb{E}\left[\mathbb{1}_{\{Z_i < Z_j\}}\right] \\
&= 1 + \sum_{i \neq j} \mathbb{P}(Z_i < Z_j) = 1 + \sum_{i \neq j} \mathbb{P}(Z_i - Z_j < 0) \\
&= 1 + \sum_{i \neq j} \mathbb{P}\left(\frac{(Z_i - Z_j) - (x_i - x_j)^T \beta}{\sqrt{2}} < \frac{(x_j - x_i)^T \beta}{\sqrt{2}}\right) \\
&= 1 + \sum_{i \neq j} \Phi\left(\frac{(x_j - x_i)^T \beta}{\sqrt{2}}\right),
\end{aligned}$$

where the last equation follows from the fact that $Z_i - Z_j$ is Gaussian with mean $(x_i - x_j)^T$ and variance 2 (since Z_k's are independent Gaussians with mean $x_i^T \beta$ and variance one for all $k \in [1, n]$). □

The idea of expected rank method is to make observed ranks of Z_i's (note that Y_i's are observed and they have the same rank as Z_i's) close to their corresponding expected ranks. The meaning of the word close is not exact here and it can be realized in different ways, i.e. minimizes the L_p norm of observed and actual expected ranks for $p \geq 1$. In the paper [Pettitt, 1987] the author suggested to minimize the square of L_2 norm, that is

$$S(\beta) := \sum_{j=1}^{n} (r_j - \mathbb{E}[R_j])^2.$$

Using the Lemma 2.1 we obtain

$$S(\beta) = \sum_{j=1}^{n} \left(r_j - \left(1 + \sum_{i \neq j} \Phi\left(\frac{(x_j - x_i)^T \beta}{\sqrt{2}}\right)\right)\right)^2. \tag{21}$$

As Pettitt [1987] noted, if there is a value of β, i.e. β_0 such that $x_j^T \beta_0$'s have the same rank as Z_i's then letting $\beta_\lambda := \lambda \beta_0$ and $\lambda \to +\infty$ will minimize $S(\beta)$, since

$$\Phi\left(\frac{(x_j - x_i)^T \beta_\lambda}{\sqrt{2}}\right) \to 1 \text{ if } x_j^T \beta_\lambda > x_i^T \beta_\lambda \text{ and}$$

$$\Phi\left(\frac{(x_j - x_i)^T\beta_\lambda}{\sqrt{2}}\right) \to 0 \text{ if } x_j^T\beta_\lambda > x_i^T\beta_\lambda \text{ as } \lambda \to +\infty,$$

which gives that

$$1 + \sum_{i \neq j} \Phi\left(\frac{(x_j - x_i)^T\beta_\lambda}{\sqrt{2}}\right) \to R_j \text{ as } \lambda \to +\infty \implies S(\beta) \to 0 \text{ as } \lambda \to +\infty.$$

The above statement shows that in the case of perfect matching of the ranks of Z_i's and $x_i^T\beta$'s the minimization problem of $S(\beta)$ is ill-posed. In general (i.e. if the noise changes the ranks of z_i's from x_i's in one dimension) numerical experiments show that the function $S(\beta)$ can be be both increasing and decreasing when $\|\beta\| \to \pm\infty$, which suggests that in some cases minimum (local) of the function $S(\beta)$ is achieved when $\|\beta\| \to \pm\infty$ and in some cases minimum (local) will not be achieved in infinity. Moreover, numerical experiments show that it is not necessary that actual value of β is a local minimum of $S(\beta)$, i.e. the function in decreasing at that point, but for the case of $n \to \infty$ it might be improved.

To circumvent the problem of $\|\beta\| \to \infty$, it can be useful to add a penalty term to $S(\beta)$ and then minimize it. Since penalty terms prevent the function to have local minimum in infinity, then every result of optimization algorithm will definitely be in bounded interval.

In the following we discussed only L_2 and L_1 penalties, but any other penalty could also work. Using a L_2 penalty term the loss function that we want to minimize will be

$$S_2(\beta) = \sum_{j=1}^{n}\left(r_j - \left(1 + \sum_{i \neq j} \Phi\left(\frac{(x_j - x_i)^T\beta}{\sqrt{2}}\right)\right)\right)^2 + \lambda\,\|\beta\|_2^2 \tag{22}$$

and for a L_1 penalty term loss function will be

$$S_1(\beta) = \sum_{j=1}^{n}\left(r_j - \left(1 + \sum_{i \neq j} \Phi\left(\frac{(x_j - x_i)^T\beta}{\sqrt{2}}\right)\right)\right)^2 + \lambda\,\|\beta\|_1\,, \tag{23}$$

where $\lambda > 0$ is a penalty strength. Since, L_2 penalty includes squares of entries of β and L_1 only the modules, values of β's in $S_2(\beta)$ are expected to be smaller than in $S_1(\beta)$ after minimization. However, L_1 penalty tends to do a variable selection, i.e. trying to make some entries of β zero.

For the minimization of one of the 21, 22 or 23 can be done using any minimization algorithm. Let's name this algorithm **Expected Rank Algorithm**, which is described in Algorithm 5.

```
Algorithm 5: Expected Rank Algorithm
Require: $\mathbf{X}$, $Y$, penalty, $\lambda$, minimization algorithm.
$r \leftarrow rank(Y)$
$S(\beta) \leftarrow \sum_{j=1}^{n} \left( r_j - \left( 1 + \sum_{i \neq j} \Phi \left( \frac{(x_j - x_i)^T \beta}{\sqrt{2}} \right) \right) \right)^2$
if $penalty$ == $no\ penalty$ then
    $\hat{\beta}_{ER} \leftarrow \underset{\beta}{argmin}\ S(\beta)$
else if $penalty$ == $L_2$ then
    $\hat{\beta}_{ER} \leftarrow \underset{\beta}{argmin}\ S(\beta) + \lambda \|\beta\|_2^2$
else if $penalty$ == $L_1$ then
    $\hat{\beta}_{ER} \leftarrow \underset{\beta}{argmin}\ S(\beta) + \lambda \|\beta\|_1$
end if
Return $\hat{\beta}_{ER}$.
```

Note that in the Algorithm 5 each computation of $S(\beta)$ (the same for $S_1(\beta)$ and $S_2(\beta)$) requires n^2 operations. The computation of the ranks is $n \log n$, since it is essentially a sorting of the elements in vector Y. So, if the minimization algorithm requires κ evaluations of the objective function, then the overall complexity of the algorithm will be $O(\kappa n^2)$. Experimental results and comparison with other methods are described in section 5.

Bennett [1968] discusses asymptotic efficiency of $\hat{\beta}_{ER}$, but it is only for the case when f_2^{-1} is a shift function, i.e. $f_2^{-1}(Y) = \mu + Y$ for some constant μ. Pettitt [1987] mentions that it is difficult to obtain any asymptotic property for $\hat{\beta}_{ER}$ in general setup.

2.4 Pairwise Rank Likelihood Method

In the previous subsections we have seen that computing the rank likelihood is computationally hard. We presented results to approximate the rank likelihood using Monte Carlo methods, but the estimation works in practice only for small β_0. In order to reduce the complexity of the computation, another idea is to only look at a reduced version of the marginal rank likelihood instead of considering whole marginal rank likelihood. Following Yu et al. [2021] we consider the pairwise rank likelihood and maximize it for the linear transformation models and we obtain consistency and asymptotic normality results.

Overview of their method is the following. Assume that distribution of the noise variable in the Linear Transformation model (5) is unknown. For i.i.d. sample $\{X_i, Y_i\}_{i=1}^n$ and corresponding i.i.d. noise $\{\varepsilon_i\}_{i=1}^n$ let us denote the Cumulative Distribution function of $\varepsilon_i - \varepsilon_j$ by $F(\cdot)$ for $i \neq j$, which does not depend on neither i nor j since noise variables are i.i.d. In (6) we have introduced initial conditions for the model for the sake of identifiability. As we know, initial conditions are not unique and they can be chosen in a different way (whichever is convenient). Yu et al. [2021] assume that true parameter β_0 has norm 1, i.e. $\|\beta_0\|_2 = 1$, but allowing any variance for the noise variables (i.e. $\sigma^2 := Var(\varepsilon_i)$ is not necessarily 1). Note that replacing β_0 by $\beta_0/\|\beta_0\|$, h by $h/\|\beta_0\|$ and ε by $\varepsilon/\|\beta_0\|$ in the model (5) we will have the initial conditions of Yu et al. [2021] with $\sigma = 1/\|\beta_0\|^2$. Given the monotonicity of function h we will have

$$\begin{aligned}\mathbb{P}(Y_j > Y_i|X_i, X_j) &= \mathbb{P}(h(Y_j) > h(Y_i)|X_i, X_j) = \mathbb{P}(X_j^T\beta + \varepsilon_j > X_i^T\beta + \varepsilon_i|X_i, X_j)\\ &= \mathbb{P}(\varepsilon_i - \varepsilon_j < X_j^T\beta - X_i^T\beta|X_i, X_j) = F\left((X_j - X_i)^T\beta\right),\end{aligned}$$

which gives that the following pairwise rank log-likelihood is

$$\begin{aligned}\ell(\beta, F) = &\binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i) \log \mathbb{P}(Y_j > Y_i|X_i, X_j)\\ &+ I(Y_j \leq Y_i) \log\left(1 - \mathbb{P}(Y_j \leq Y_i|X_i, X_j)\right)\\ = &\binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i) \log F\left((X_j - X_i)^T\beta\right)\\ &+ I(Y_j \leq Y_i) \log\left(1 - F\left((X_j - X_i)^T\beta\right)\right).\end{aligned} \tag{24}$$

Using the pairwise rank likelihood the estimators of β_0 and F will be

$$(\hat{\beta}, \hat{F}) = \underset{\beta\in\mathcal{B}, F\in\mathcal{F}}{argmax}\, \ell(\beta, F),$$

where $\mathcal{B}$ is a compact subspace of $\mathbb{R}^m$ and

$$\mathcal{F} = \{F(\cdot) : F(x) \in [0, 1] \text{ and is monotonically increasing}\}.$$

For the Algorithm of finding the estimators please refer to the Sections 3.1 (page 5-6) and 8.1 (page 94) of the paper. The idea is to use Pool Adjacent Violation Algorithm (PAVA, Ayer et al. [1955]) and active set methods (de Leeuw et al. [2009]) for the estimation of function F, then use standard optimization (i.e. Nelder–Mead method, since resulting objective function will not be differentiable) for the parameter β.

In our experiments we saw that the nonparametric estimation of the distribution function F is computationally expensive. Since, in our setup we assume that the distribution of the noise variable in the Linear Transformation Model (5) is Gaussian, which implies that function F is known up to a scaling factor (variance of the noise). However, in our initial conditions (6) we have incorporated it in the parameter β_0 as have been discussed at the beginning of this subsection.

Now let us define the Pairwise Rank Likelihood in our setup and derive its properties. From the initial conditions (6) we have $\varepsilon_i - \varepsilon_j \sim \mathcal{N}(0, 2)$ since they are i.i.d. standard normal and so

$$F(x) = \Phi\left(\frac{x}{\sqrt{2}}\right),$$

which gives that pairwise rank log-likelihood is the following

$$\begin{aligned}\ell_{prl}(\beta) = \binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i) \log \Phi\left(\frac{(X_j - X_i)^T\beta}{\sqrt{2}}\right) \\ + I(Y_j \le Y_i) \log\left(1 - \Phi\left(\frac{(X_j - X_i)^T\beta}{\sqrt{2}}\right)\right).\end{aligned} \tag{25}$$

So, using the above definition of pairwise rank log-likelihood we can define the PRL estimator which will maximize it, i.e.

$$\hat{\beta}_{PRL} := \underset{\beta}{argmax}\ \{\ell_{prl}(\beta)\}. \tag{26}$$

The procedure to obtain a PRL estimator is described in the following Algorithm 6.

```
Algorithm 6: Pairwise Rank Likelihood Algorithm
  Require: $\mathbf{X}$, $Y$, maximization algorithm.
  $\ell_{prl}(\beta) \leftarrow \binom{n}{2}^{-1} \sum_{i<j} I(y_j > y_i) \log \Phi\left(\frac{(x_j - x_i)^T\beta}{\sqrt{2}}\right) + I(y_j \le y_i) \log\left(1 - \Phi\left(\frac{(x_j - x_i)^T\beta}{\sqrt{2}}\right)\right)$
  $\hat{\beta}_{PRL} \leftarrow \underset{\beta}{argmax}\ \ell_{prl}(\beta)$
  Return $\hat{\beta}_{PRL}$.
```

From the optimization perspective in order to obtain the maximum of the function the optimization function is desired to be concave, so, that its stationary point will be maximum. It turns out that the pairwise rank log-likelihood function $\hat{\beta}_{PRL}$ is indeed concave, which is stated in the following Proposition 2.2.

In order to establish the desired properties, some technical conditions are necessary to hold, which are introduced in the Appendix A.1.

Proposition 2.2. *Log pairwise rank likelihood function $\ell_{prl}(\beta)$ defined in 25 is concave. Moreover, if we assume that the Condition 0 in the Appendix A.1 holds, we have $\ell_{prl}(\beta)$ is strictly concave.*

Proof. Please see the proof in Appendix A.4.1. □

The next Theorem establishes asymptotic properties of the estimator $\hat{\beta}_{PRL}$, in particular asymptotic normality.

Theorem 2.2. *Assume that Conditions 0-2 in Appendix A.1 hold, then*

1. $\hat{\beta}_{PRL} - \beta_0 = o_p(1)$,
2. $\sqrt{n}(\hat{\beta}_{PRL} - \beta_0) \xrightarrow{d} \Sigma^{-1}\mathcal{N}(0, \Sigma_\psi)$,

where Σ_ψ *is defined in Condition 2 in the Appendix A.1 and* $\Sigma := -\nabla^2_\beta \ell_{prl}(\beta_0)$.

Proof. Please see the proof in Appendix A.4.2. □

The above Theorem 2.2 shows that the estimator $\hat{\beta}_{PRL}$ is consistent and asymptotically normal. This theorem is different from the results in [Yu et al., 2021] in various ways. Firstly, initial conditions are different as we assume no condition on β_0 and they assume it has norm 1. Instead we assume the variance of noise variable is 1. These conditions should have similar effect on the estimation of β_0. Secondly, they have neither asymptotic normality nor root n consistency of their PRL estimator. Moreover they have conjectured that their estimator is root n consistent. Theorem 2.2 is a validation of their conjecture in the case when the distribution of the noise variable in the linear transformation model is known and it is Gaussian. In order to produce root n consistent estimator they have introduced score base version of the pairwise rank likelihood and in this case asymptotic normality and root n consistency hold. Finally, since the assumption of Gaussian noise is quite strict one, we expect to smaller variance in the estimator.

On top of the theoretical results of $\hat{\beta}_{PRL}$, it shows quite good results in practice. Concavity of the objective function makes the optimization problem much easier and in practice its maximization converges much faster than the other objective functions. Moreover, experimental results in Section 5 show that this estimator outperforms the others, especially when we have β_0 is not close to 0.

2.5 Estimation of the Transformation Function

Having estimated β_0 in the model (5), we will also need to estimate the transformation function h in order to obtain the estimate of the noise. We need to estimate the noise in order to carry out an independence test between the noise and the parents of the effect in the Post Nonlinear models to obtain the correct causal directions.

As mentioned briefly in the substitution of conditional expectation method (see the equation (17)), Cuzick [1988] also suggested an estimate for the function h.

Moreover, in the paper it has been showed that the estimator is asymptotically Gaussian process. Let us review it and restate the result. Assume that we have already estimated β_0 using some method and name it $\hat{\beta}$. For the values in the sample the estimator of the function h is

$$\hat{h}_{FP}(y_i) := F_{\hat{\beta}}^{-1}(\hat{F}(h(y_i))) \text{ for } i \in [1, n], \tag{27}$$

where the function $\hat{F}$ and $F_{\hat{\beta}}$ are defined in (15) and (16), respectively. Note that we can calculate the values of $\{\hat{F}(h(y_i))\}_{i=1}^n$ since $\hat{F}$ needs only the ranks and the ranks of $\{h(y_i)\}_{i=1}^n$ is the same as the ranks of $\{y_i\}_{i=1}^n$ by strict monotonicity of h.

Theorem 2.3. *(Theorem 2 in [Cuzick, 1988])*
If the assumptions (A1)-(A6) in [Cuzick, 1988] are satisfied and we define $\hat{h}_{FP}(y) := \hat{h}_{FP}(y_{(i)})$ for $y \in [y_{(i)}, y_{(i+1)})$ (recall that $\{y_{(i)}\}_{i=1}^n$ is the order statistics of $\{y_i\}_{i=1}^n$), then

$$\hat{h}_{FP}(y) \to h(y) \text{ as } n \to \infty$$

for all continuity points y of h. Moreover, if we further assume that h is continuously differentiable, then

$$\sqrt{n}(\hat{h}_{FP}(y) - h(y)) \to \gamma(y)$$

weakly in Skorohod space D on every bounded set, where $\gamma(\cdot)$ is a mean zero Gaussian process. For the covariance function of $\gamma(\cdot)$ please refer to [Cuzick, 1988].

Since, calculating the function $\hat{h}_{FP}$ only requires to do one step in the Fixed Point Algorithm 3, this is computationally feasible even for large sample size.

The estimators from [Horowitz, 1996, Chen, 2002, Zhang, 2013] can also be used, since they also have similar asymptotic properties as $\hat{h}_{FP}$.

Remark 2.1. *Note that the above Theorem 2.3 is stated and proved, where the estimation $\hat{\beta}$ of β_0 is carried out using the fixed point method (Algorithm 3). However, in the proof of the theorem no special representation of the $\hat{\beta}$ is used and only the asymptotic properties is enough. So, using the estimator $\hat{\beta}$ which is consistent and asymptotically normal is enough for the Theorem 2.3. We use Pairwise Rank Likelihood method to estimate β_0.*

2.6 Estimation of the Noise

In the previous subsections we discussed how it is possible to estimate the transformation function h and the parameter β_0 in the Linear Transformation Model (5). Their asymptotic properties have been derived. Here we discuss how to estimate the noise in this model and obtain its properties.

Using these methods we can estimate the noise in the following way. Assume we have n i.i.d. sample $\{X_j, Y_j\}_{j=1}^n$ and estimators $\hat{h}$ and $\hat{\beta}$ for h and β_0 respectively. Define

$$\hat{\varepsilon}_j = \hat{h}(Y_j) - X_j^T \hat{\beta}. \tag{28}$$

Since it is reasonable to assume that $\hat{h}$ is depends on $\hat{\beta}$ (i.e. see the equation (27) and $\hat{h}_{FP}$ clearly depends on $\hat{\beta}$ through the function $F_{\hat{\beta}}$), having only the asymptotic distributions of the estimators we cannot obtain the asymptotic distribution of $\hat{\varepsilon}_j$ as $Z_n \xrightarrow{d} Z$ and $T_n \xrightarrow{d} T$ does not always imply $Z_n + T_n \xrightarrow{d} Z + T$, for random variables Z_n, T_n, Z, T and the counterexample occurs when Z_n and T_n are dependent. I have not been able to obtain the joint distribution of $\hat{h}$ and $\hat{\beta}$. However, as the next results shows provided a consistency of the estimators we will have consistency of the errors also.

Lemma 2.2. *Let x, y be fixed observed sample from model* (5)*, i.e. $h(y) = x^T \beta_0 + \epsilon$. Assume $\hat{h}$ and $\hat{\beta}$ are consistent estimators of h and β_0 respectively based on i.i.d sample $\{X_j, Y_j\}_{j=1}^n$ from the model* (5)*, that is*

$$\hat{\beta} \xrightarrow{p} \beta \text{ and } \hat{h}(y) \xrightarrow{p} h(y) \text{ as } n \to \infty.$$

Then, for $\hat{\varepsilon} := \hat{h}(y) - x^T \hat{\beta}$ we have

$$\hat{\varepsilon} - \epsilon = o_p(1).$$

Proof. Using the Continuous Mapping Theorem (Theorem 2.3 in [Vaart, 1998]) and stochastic o notation, we obtain

$$\hat{\varepsilon} - \epsilon = \hat{h}(y) - h(y) - (x^T \hat{\beta} - x^T \beta_0) = o_p(1) - o_p(1) = o_p(1).$$

□

3 General Transformation Models

In this section we look at the General Transformation Models, which contain as special case Linear Transformation Models (5). This models can be used to estimate the causal order of PNL models without Gaussian noise assumption and so the most general case of PNL models.

Let's assume

$$h(Y) = g(X) + \varepsilon, \tag{29}$$

where $X \in \mathbb{R}^m, Y \in \mathbb{R}$ for some $m \in \mathbb{N}$, ε is a noise variable independent of X, i.e. $\varepsilon \perp\!\!\!\perp X$, $g : \mathbb{R}^m \to \mathbb{R}$ is some function (can be nonlinear as well) and $h : \mathbb{R} \to \mathbb{R}$ is invertible and strictly increasing function. Note that in case g is a linear function this models becomes the Linear transformation Model.

In this section we do not put any assumption on the noise variable ε, i.e. they can be Gaussian, exponential, Weibull and etc. The goal is to estimate the functions h and g based on only observational data. In the Linear Transformation Models we have estimated the function g (which is linear in that case) then estimated the function h. However, for the General Transformation Models it is exactly vice versa, i.e. firstly the function h is estimated then using that information the function g will be estimated.

For the Linear transformation Models (5) Horowitz [1996] suggested a way to estimate the transformation function h, based on a root n consistent estimator of β_0. Chiappori et al. [2015] exploited this idea for General Transformation Models (29) and able to estimate the function h at the first stage (non estimates is known for g at this point) using kernel smoothing methods. However, if the response Y is skewed with very long tails the method of Chiappori et al. [2015] does not work well. Then, Colling and Keilegom [2019] suggested a way to address this problem by working with the proper transformation of Y instead of Y directly, which we describe in the following.

In order to identify the model (29) the following normalization condition is imposed (to draw parallels with Linear Transformation Models, we put conditions on the noise variable, such as centered and variance 1):

$$\text{(N1) } h(\alpha_1) = a_1 \text{ and } h(\alpha_2) = a_2 \text{ for some } \alpha_1 < \alpha_2 \text{ and } a_1 < a_2.$$

The idea of (N1) normalization is to fix the location and scale of the model. Without loss of generality it is assumed that $\alpha_1 = a_1 = 0$ and $\alpha_2 = a_2 = 1$.

Since, in general Y can be skewed then kernel smoothing based on Y can work poorly. So, the idea of Colling and Keilegom [2019] is to work on the following transformation of Y. Let

$$T(Y) := \frac{F_Y(Y) - F_Y(0)}{F_Y(1) - F_Y(0)},$$

where F_Y is the cumulative distribution function of Y. Since, T is invertible (i.e. it is strictly increasing and right continuous) we can define $\Gamma(Y) = h(T^{-1}(Y))$, so under (N1) we have $\Gamma(0) = h(T^{-1}(0)) = 0$, $\Gamma(1) = h(T^{-1}(1)) = 1$ and General Transformation Model (29) becomes

$$\Gamma(U) = g(X) + \varepsilon, \tag{30}$$

where $U = T(Y)$ and $\Gamma(U) = h(Y)$. Since, U is defined by shifting and rescaling

$F_Y(Y)$ then it is uniformly distributed and kernel smoothing based on U should work better than kernel smoothing based on Y.

Now, let us state the identification results and present the estimators. For that purpose define $\varphi(u,x) = F_{U|X}(u,x)$ conditional distribution function of U given X. Then

$$\varphi(u,x) = \mathbb{P}(U \leq u|X=x) = \mathbb{P}(g(X)+\varepsilon \leq \Gamma(u)|X=x) = F_\varepsilon(\Gamma(u)-g(x)),$$

where we used the independence between X and ε, and F_ε is the cumulative distribution function of ε. Under some assumptions the following derivatives exists. The derivative of $\varphi(u,x)$ with respect to u is denoted by $\varphi_u(u,x)$ and can be calculated by

$$\varphi_u(u,x) = \frac{\partial}{\partial u}\varphi(u,x) = \Gamma'(u) f_\varepsilon(\Gamma(u)-g(x)),$$

where f_ε is the probability density function of ε. In the same way the derivative of $\varphi(u,x)$ with respect to x_ρ (for all $\rho \in [1,m]$) will be

$$\varphi_\rho(u,x) = \frac{\partial}{\partial u}\varphi(u,x) = -\frac{\partial}{\partial x_\rho} g(x) \cdot f_\varepsilon(\Gamma(u)-g(x)).$$

Assuming $\varphi_\rho(u,x)$ is not zero, the division of the above two equations will give

$$\Gamma'(u) = -\frac{\partial}{\partial x_\rho} g(x) \cdot \frac{\varphi_u(u,x)}{\varphi_\rho(u,x)}, \tag{31}$$

which is the basis for proving the following identification result.

Theorem 3.1. *(Theorem 3.1 in [Colling and Keilegom, 2019])*
Assume (A1)-(A4) in [Colling and Keilegom, 2019] hold. Then for any $\rho \in [1,m]$ such that $\mathcal{A}_\rho$ (defined in [Colling and Keilegom, 2019]) is not empty, Γ can be identified under (N1) in the following way

$$\Gamma(u) = \lambda_\rho(u,x) = \frac{S_\rho(u,x)}{S_\rho(1,x)},$$

where $s_\rho(u,x) = \frac{\varphi_u(u,x)}{\varphi_\rho(u,x)}$ and $S_\rho(u,x) = \int_0^u s_\rho(w,x)\,dw$. Moreover, $\lambda_\rho(u,x)$ does not depend neither ρ nor x.

The above shows the identifiability of the function Γ and now let us define its estimators. Let $\{(X_j,Y_j)\}_{j=1}^n$ be i.i.d. sample from the General Transformation Model (29) and define $U_i = T(Y_i)$.

The function $\varphi(u,x)$ can also be written as

$$\varphi(u,x) = \frac{p(u,x)}{f_X(x)},$$

where

$$p(u,x) = \int_{-\infty}^{u} f_{U,X}(w,x)\,dw\,,\; f_X(x) = \int_{-\infty}^{+\infty} f_{U,X}(w,x)\,dw$$

and $f_{U,X}(u,x)$ is the joint probability density function of (U,X). For a univarite kernel K (in the simulations Epanechinkov kernel has been chosen, which also satisfies desired conditions) define $\mathcal{K}(u) = \int_{-\infty}^{u} K(w)\,dw$ and $\mathbf{K}(x) = \prod_{j=1}^{m} K(x^{(j)})$ be product kernel. For bandwidths h_x and h_u we can define $\mathbf{K}_{h_x}(x) := \mathbf{K}(x/h_x)/h_x^m$ and $\mathcal{K}_{h_u}(u) = \mathcal{K}(u/h_u)/h_u$. Since, we observe Y_j and not U_j we have to estimate U_j and natural way to do is to replace cumulative distribution function of Y_j by its empirical distribution function, that is

$$\hat{U}_j = \hat{T}(Y_j) = \frac{\hat{F}_Y(Y_j) - \hat{F}_Y(0)}{\hat{F}_Y(1) - \hat{F}_Y(0)} \text{ for } j \in [1,n],$$

where $\hat{F}_Y(y) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \leq y)$. Then kernel estimator of $\varphi(u,x)$ can be defined as

$$\hat{\varphi}(u,x) = \frac{\hat{p}(u,x)}{\hat{f}_X x},$$

where

$$\hat{p}(u,x) := \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}_{h_u}(u - \hat{U}_j)\cdot \mathbf{K}_{h_x}(X_i - x) \text{ and } \hat{f}_X(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{K}_{h_x}(X_i - x).$$

Now based on the above we can estimate the ingredients of λ_ρ defined in the Theorem 3.1. Since the Theorem hold for every value of ρ we can assume that without loss of generality $\rho = 1$. We have the following

$$\varphi_u(u,x) = \frac{\partial}{\partial u}\frac{p(u,x)}{f_X(x)} = \frac{p_u(u,x)}{f_X(x)},$$

and

$$\varphi_1(u,x) = \frac{\partial}{\partial x_1}\frac{p(u,x)}{f_X(x)} = \frac{p_1(u,x)f_X(x) - p(u,x)f_{X,1}(x)}{f_X^2(x)},$$

where $p_u(u,x) = \frac{\partial}{\partial u}p(u,x)$, $p_1(u,x) = \frac{\partial}{\partial x_1}p(u,x)$ and $f_{X,1}(x) = \frac{\partial}{\partial x_1}f_X(x)$. Now their corresponding kernel estimators will be

$$\hat{p}_u(u,x) = \frac{\partial}{\partial u}\hat{p}(u,x), \quad \hat{p}_1(u,x) = \frac{\partial}{\partial x_1}\hat{p}(u,x) \text{ and } \hat{f}_{X,1}(x) = \frac{\partial}{\partial x_1}\hat{f}_X(x)$$

and

$$\hat{\varphi}_u(u,x) == \frac{\hat{p}_u(u,x)}{\hat{f}_X(x)}, \quad \hat{\varphi}_1(u,x) = \frac{\hat{p}_1(u,x)\hat{f}_X(x) - \hat{p}(u,x)\hat{f}_{X,1}(x)}{\hat{f}_X^2(x)}.$$

Putting everything together, we obtain

$$\hat{\lambda}_1(u,x) = \frac{\hat{S}_1(u,x)}{\hat{S}_1(1,x)}, \tag{32}$$

where

$$\hat{S}_1(u,x) = \int_0^u \frac{\hat{\varphi}_u(w,x)}{\hat{\varphi}_1(w,x)}\, dw\,.$$

Since in the equation (32) λ_1 depends on x, while the actual λ_1 does not depend on x, we can integrate it over$x$ to have more reliable estimator. For a weighting function $v(x)$ Theorem 3.1 gives $\Gamma(u) = \lambda_1(u,x)$ and we have

$$\Gamma(u) = \underset{q}{argmin} \int v(x)\ell(\lambda_1(u,x) - q)\, dx$$

where ℓ is a loss function (i.e. $\ell(t) = t^2$ or $\ell(t) = |t|$). This is true since integrand is always non-negative and is zero when $\lambda_1(u,x) - q = 0$, which gives $\Gamma(u) = \lambda_1(u,x)$. Then, for $\ell(t) = t^2$ we have

$$\hat{\Gamma}_{LS}(u) = \int v(x)\hat{\lambda}_1(u,x)\, dx \tag{33}$$

and for $\ell(t) = |t|$ we have

$$\hat{\Gamma}_{LAD}(u) = \underset{q}{argmin} \int v(x)|\lambda_1(u,x) - q|\, dx\,. \tag{34}$$

In order to facilitate the theoretical analysis of $\hat{\Gamma}_{LAD}$ the authors suggested smoothed version of it, namely $\hat{\Gamma}_{LAD,b}$. For more detailed definitions please refer to Colling and Keilegom [2019].

Both of the above estimators are asymptotically Gaussian as the following Theorem states.

Theorem 3.2. *(Corollary 5.1 in [Colling and Keilegom, 2019])*
Assume (A1)-(A10) in [Colling and Keilegom, 2019] hold. Then the processes $\sqrt{n}(\hat{\Gamma}_{LS}(\hat{T}(y)) - h(y))$ and $\sqrt{n}(\hat{\Gamma}_{LAD,b}(\hat{T}(y)) - h(y))$ converge to centered Gaussian process $\tilde{\mathcal{N}}(y)$.

For the detailed definitions and proofs please refer to Colling and Keilegom [2019].

Note that here we obtained estimate for h without estimating the function g, which is the exact opposite of what we have done in the Linear Transformation Models. Now we can replace function h by its estimate $\hat{\Gamma} \circ \hat{T}$ (where $\hat{\Gamma}$ is either $\hat{\Gamma}_{LS}$ or $\hat{\Gamma}_{LAD,b}$) in the General Transformation Model (29) and obtain

$$\hat{\Gamma}_{LS}(\hat{T}(Y)) = g(X) + \varepsilon.$$

Since, in the above response $\hat{\Gamma}_{LS}(\hat{T}(Y))$ is already known we can estimate g using any method for the kernel regression. For instance, Nadaraya–Watson method gives

$$\hat{g}(x) = \frac{\sum_{j=1}^n \mathbf{K}_{h_x}(x - X_j)\cdot \hat{\Gamma}_{LS}(\hat{T}(Y_j))}{\sum_{j=1}^n \mathbf{K}_{h_x}(x - X_j)}. \tag{35}$$

Thus, the equations (33), (34) and (35) give a complete non-parametric estimation of the model (29).

4 Post-Nonlinear Gaussian Causal Models

Using the results from the previous section we know how to estimate the parameters in the model $Y = h^{-1}(g(X)+\varepsilon)$. Since the post nonlinear model is exactly the same form in the case when variables in X are the causes of Y, we can use these results to obtain a method which will discover the underlying true causal directions. For causal order estimation, the idea is to identify a sink node at each step and remove it from the causal graph and repeat the same procedure until there will be no node in the graph.

For the case of m nodes $X^{(1)}, \ldots, X^{(m)}$ in the causal graph, we have to understand which one is a sink node. For that purpose we can fit the models for each node as a sink, then report the one which fitted the best. In order to understand what means best precisely we can just check the conditions of the model, i.e. noise is independent of the parents of the effect. For instance, if we fit the model for $X^{(m)}$ as a sink node, then we have $X^{(m)} = h_m^{-1}(g_m(X^{(1)}, \ldots, X^{(m-1)})+\varepsilon_m)$. After estimating the functions h_m and g_m we obtain the estimate of the noise and we can test the independence between the noise and $X^{(1)}, \ldots, X^{(m-1)}$. Note that in the case of $X^{(m)}$ is actually a sink node then there is a model $X^{(m)} = h_m^{-1}(g_m(X^{(1)}, \ldots, X^{(m-1)}) + \varepsilon_m)$, where the noise is independent of $X^{(1)}, \ldots, X^{(m-1)}$. So, for the sink node we expect that the estimated noise is independent from the parents of the sink.

There are various methods to test independence based on observational data. For instance, different rank correlation tests, including ρ of Spearman [1904], τ of Kendall [1938] and Chatterjee's rank correlation Shi et al. [2020]. Berrett et al. [2021] use permutation U-statistics to test independence. Gretton et al. [2005] suggested Hilbert-Schmidt Independence Criterion (HSIC) which has found many applications.

We mainly use the HSIC (we will review it later in this section) criterion for independence since it characterizes independence completely, i.e. HSIC is zero if and only if the random vectors are independent.

The remaining of the section is organized as follows. Firstly, we briefly overview the Hilbert-Schmidt Independence Criterion and existing PNL methods. Then, consider the bivariate PNLG models and develop a method to estimate causal direction. Afterwards, we extend the bivariate method for multivariate setting and derive their properties. At the end discuss the PNL (without Gaussian noise assumption) models and show how our method can be extended to this case.

4.1 HSIC

Here we review the Hilbert-Schmidt Independence Criterion (HSIC) from [Gretton et al., 2005] paper. Assume that X and Y are random vectors and they have a joint distribution $\mathbb{P}^{X,Y}$. We would like to test the independence of X and Y based on the i.i.d. sample $S := \{(X_1, Y_1), \dots, (X_n, Y_n))\}$ from the distribution $\mathbb{P}^{X,Y}$. Let $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$ be kernels defined on the corresponding spaces of X and Y. Then, empirical HSIC is defined as follows

$$HSIC(S) := (n-1)^{-2} tr(KHLH), \tag{36}$$

where tr is a trace of a matrix, $H, K, L \in \mathbb{R}^{n \times n}$, $K_{ij} := k(X_i, X_j)$, $L_{ij} := l(Y_i, Y_j)$ and $H_{ij} := I(i = j) - n^{-1}$ for $i, j \in [1, n]$. Moreover, population version of the HSIC (not empirical one) is defined as $HSIC(\mathbb{P}^{X,Y})$ for the same kernels $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$ and $HSIC(\mathbb{P}^{X,Y}) = 0$ if and only if X and Y are independent. For more exact definition please refer to the paper.

Now let us state the properties of empirical HSIC from [Gretton et al., 2005]. The following theorem gives the bias term of the HSIC.

Theorem 4.1. *(Theorem 1 in [Gretton et al., 2005])*
The following equality holds

$$HSIC(\mathbb{P}^{X,Y}) = \mathbb{E}_S[HSIC(S)] + O(n^{-1}),$$

where $\mathbb{E}_S$ denotes the expectation taken over n independent copies of X, Y from the distribution $\mathbb{P}^{X,Y}$.

The above theorem shows that bias of empirical HSIC converges to the actual HSIC with the rate $O(n^{-1})$. The next result gives a quantitative bound on empirical HSIC.

Theorem 4.2. *(Theorem 3 in [Gretton et al., 2005])*
Assume that kernels $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$ are bounded almost surely by 1 and are non-negative. Then for $n > 1$ and arbitrary $\delta > 0$, with probability at least $1 - \delta$, for all $\mathbb{P}^{X,Y}$, the following holds

$$|HSIC(\mathbb{P}^{X,Y}) - HSIC(S)| \le \sqrt{\frac{\log(6/\delta)}{\alpha^2 n}} + \frac{C}{n},$$

where $\alpha^2 > 0.24$ and C are constants.

We will not review how exactly carry out a independence test using HSIC, because only the value of the Independence criterion HSIC is sufficient for us to understand which is the best fitted model.

4.2 Existing PNL methods

Firstly, let us review existing bivariate PNL estimation methods then move to the multivariate case.

As described in Zhang and Hyvärinen [2009], the authors use the multi-layer perceptrons (MLP's) to represent the functions f_1 and f_2^{-1} in the bivariate PNL model

$$X^{(2)} = f_2(f_1(X^{(1)}) + \varepsilon_2).$$

If $\hat{f}_1$ and $\hat{f}_2^{-1}$ are these representations, then noise can be expressed as

$$\hat{\varepsilon}_2 = \hat{f}_2^{-1}(X^{(2)}) - \hat{f}_1(X^{(1)}).$$

Since, ε_2 is independent of $X^{(1)}$ they learn $\hat{f}_1$ and $\hat{f}_2^{-1}$ by minimizing the mutual information between $\hat{\varepsilon}_2$ and $X^{(1)}$, i.e. $I(X^{(1)}, \hat{\varepsilon}_2)$. Then the direction $X^{(1)} \to X^{(2)}$ is supported if the resulting noise is independent of the cause according to some independence test.

Note that if $\hat{f}_1$ and $\hat{f}_2^{-1}$ are constants then their difference is again constant and it will be independent of $X^{(1)}$, which gives their mutual information is zero. So, for every model MLP's can learn constant functions by minimizing the mutual information.

To deal with this problem Uemura and Shimizu [2020] suggested to force the function $\hat{f}_2$ be invertible and instead of the mutual information use HSIC. In order to make $\hat{f}_2$ invertible the authors use auto-encoder structure in the loss in the following way. The functions f_1, f_2, f_2^{-1} are represented by Neural Networks $\hat{f}_1, \hat{f}_2, \hat{f}_2^{-1}$, respectively. Then again noise is estimated by $\hat{\varepsilon}_2 = \hat{f}_2^{-1}(X^{(2)}) - \hat{f}_1(X^{(1)})$ and in order to force $\hat{f}_2^{-1}$ be the inverse of $\hat{f}_2$, the reconstraction error of $X^{(2)}$ is provided by the euclidean norm of $X^{(2)} - \hat{f}_2(\hat{f}_2^{-1}(X^{(2)}))$. So, the final loss will be convex combination of the two losses, i.e.

$$\mathcal{L} = \lambda HSIC(X^{(1)}, \hat{\varepsilon}_2) + (1 - \lambda) \left\| X^{(2)} - \hat{f}_2(\hat{f}_2^{-1}(X^{(2)})) \right\|_2,$$

where $\lambda \in (0, 1)$. Then the both directions fitted, i.e. $X^{(1)} \to X^{(2)}$ and $X^{(2)} \to X^{(1)}$, and the direction which produced smaller loss have been chosen.

For the multivariate PNL models Uemura et al. [2022] extend the method of Uemura and Shimizu [2020] to estimate the true underlying causal graph of PNL models. The idea of the method is to identify the sink node at each step by using the same method as above for the bivariate case. After estimating the causal order the authors identify the nodes in the set of possible parents that are important to make the estimated noise independent of them and thus producing the estimation of the causal graph.

Note that in all the above mentioned methods the functions f_1 and f_2 are learned by minimizing the independence between noise and potential causes. Then testing the independence between noise and the causes. This can be problematic if we do not have infinite data. For instance, for a fixed sample size n models are powerful enough that can make the estimated noise independent of the causes then we end up supporting all the possible causal directions. However, if we are bale to estimate these function using only the structure of the model it can produce better results.

4.3 Bivariate Post-Nonlinear Gaussian Causal Models

In this section we are looking at the bivariate Gaussian causal model

$$X^{(2)} = f_2(f_1(X^{(1)}) + \varepsilon_2), \tag{37}$$

where $\varepsilon_2 \sim \mathcal{N}(0,1)$ is a noise variable independent of $X^{(1)}$ and f_2 is invertible.

Firstly, let us understand which are the identifiable cases in bivariate post-nonlinear causal models. Assuming the model is not identifiable, i.e. there is a backward direction

$$X^{(1)} = g_2(g_1(X^{(2)}) + \varepsilon_1),\ \varepsilon_1 \perp\!\!\!\perp X^{(2)}, \tag{38}$$

For some functions g_2 and g_1, where g_2 is invertible. Zhang and Hyvärinen [2009] show that under some conditions(which are satisfied in Gaussian PNL setting) all non-identifiable cases are listed in Table 1. For the distributions listed in the table please look at the paper.

Table 1 gives that in the Gaussian noise setting, only non-identifiable model is the first line of the the table, which shows that if bivariate pnl-Gaussian model is identifiable, then both h and h_1 functions must be linear. Since, these functions can't be linear if, for instance, function f_1 is not injective, of course there are some subtle points which should be taking into account. For strict examples, Corollaries 4.1 and 4.2 give an examples of non-identifiable and identifiable bivariate models, respectively, which are also used in the experiments.

Corollary 4.1. *The following post-nonlinear causal model is not identifiable*

$$X^{(2)} = (\beta X^{(1)} + \varepsilon_2)^{1/3}, \tag{39}$$

where $\beta \in \mathbb{R}, \varepsilon_2 \sim \mathcal{N}(0,1), X^{(1)} \sim \mathcal{N}(0,1)$ *and* $\varepsilon_2 \perp\!\!\!\perp X^{(1)}$.

Proof. Please see the proof in Appendix A.5.1. □

ε_2	$T_1 := g_2^{-1}(X^{(1)})$	$h := f_1 \circ g_2$	Remark ($h_1 := g_1 \circ f_2$)
Gaussian	Gaussian	linear	h_1 is linear
log-mix-lin-exp	log-mix-lin-exp	linear	$h_1^{'}$ strictly monotonic, $h_1^{'} \to 0$, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
log-mix-lin-exp	one-sided asymptotically exponential (but not log-mix-lin-exp)	h is strictly monotonic and $h^{'} \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$	-
log-mix-lin-exp	generalized mixture of two exponentials	same as above	-
generalized mixture of two exponentials	two-sided asymptotically exponential	same as above	-

Table 1: Non identifiable bivariate PNL causal models

Corollary 4.2. *The following post-nonlinear causal model is identifiable*

$$X^{(2)} = f_2(\beta(X^{(1)})^2 + \varepsilon_2), \tag{40}$$

where $\beta \in \mathbb{R}, \varepsilon_2 \sim \mathcal{N}(0,1), X^{(1)} \sim \mathcal{N}(0,1), \varepsilon_2 \perp\!\!\!\perp X^{(1)}$ and function f_2 is invertible. In particular the model $X^{(2)} = (\beta X^{(1)} + \varepsilon_2)^{1/3}$ is identifiable.

Proof. Please see the proof in Appendix A.5.1. □

Now having an impression in what cases the model is non-identifiable and having seen examples from both cases, let us develop a method to estimate the underlying true causal direction assuming the model is identifiable. This means, the model (37) and backward direction is not possible, i.e. (38) does not hold. Thus, if we fit the transformation model for both directions, i.e. $f_2^{-1}(X^{(2)}) = f_1(X^{(1)}) + \varepsilon_2$ and $g_2^{-1}(X^{(1)}) = g_1(X^{(2)}) + \varepsilon_1$ and test the independence of noise and the cause, only in correct direction it will be satisfied. For the Independence we can just take the HSIC value and compare for both fitted cases. Intuitively, the direction which produces smaller HSIC value should be a true causal direction. The method is described in the following Algorithm 7.

Algorithm 7: Bivariate PNL Estimation Algorithm

```
Require: ${(X_1^{(1)}, X_1^{(2)}), \dots (X_n^{(1)}, X_n^{(2)})}$, noise estimation algorithm.
$(u_1, \dots, u_n) \leftarrow$ estimated noise for the direction $X^{(1)} \to X^{(2)}$
$hsic_1 \leftarrow HSIC(\{X_j^{(1)}, u_j\}_{j=1}^n)$
$(w_1, \dots, w_n) \leftarrow$ estimated noise for the direction $X^{(2)} \to X^{(1)}$
$hsic_2 \leftarrow HSIC(\{X_j^{(2)}, w_j\}_{j=1}^n)$
if $hsic_1 < hsic_2$ then
    $order \leftarrow (X^{(1)}, X^{(2)})$
else
    $order \leftarrow (X^{(2)}, X^{(1)})$
end
Return $order$.
```

Since, above described method is a specific case of multivariate PNL method, discussed in the next subsection, we present the theoretical results only for multivariate case.

4.4 Multivariate Post-Nonlinear Gaussian Causal Models

Here we look at the multivariate PNLG (Definition 1.3) causal models, give a method to estimate the causal order of them and study the asymptotic properties of this method.

For PNLG models we have

$$X^{(j)} = f_{j,2}(f_{j,1}(\mathbf{PA}_j) + \varepsilon_j) \text{ for } \forall j \in [1, m],$$

where $\mathbf{PA}_j$ are the parents of $X^{(j)}$ in the causal graph $\mathcal{G}^0$ and $\varepsilon_j \sim \mathcal{N}(0, 1)$ with $\varepsilon_j \perp\!\!\!\perp \mathbf{PA}_j$.

The idea of estimating the causal order of the PNLG model is to identify a sink node in the causal graph in the same way as in the bivariate case and then remove it from the graph. Note that iteration of identifying a sink node and removing it from the graph gives causal order of the model. Moreover, we can obtain a sink node exactly in the same way as we did in the bivariate case using the noise estimation algorithm and then testing if the noise is independent of the causes or not.

Assume we have i.i.d. sample $\{X_j := (X_j^{(1)}, \dots, X_j^{(m)})\}_{j=1}^n$ from the PNLG model with sample size n and number of nodes in the graph $\mathcal{G}^0$ is m. Moreover, assume that we also have T (T is fixed) i.i.d. sample $\{V_j := (V_j^{(1)}, \dots, V_j^{(m)})\}_{j=1}^T$ from the same model with observed values $\{v_j := (v_j^{(1)}, \dots, v_j^{(m)})\}_{j=1}^T$, which are independent of $\{X_j\}_{j=1}^n$.

Remark 4.1. *The values $\{V_j\}_{j=1}^T$ can be seen as the test set and the values $\{X_j\}_{j=1}^n$ as training set. This is somewhat technical separation in order to obtain the consistency of the causal order estimation with high probability with precise arguments. However, I think consistency holds without this separation, when errors are evaluated directly on the same dataset and then tested Independence with potential causes. This is a interesting topic of future work. In the experiments the values in the test and train set have been chosen the same.*

Let π be a permutation of $\{1, 2, \dots, m\}$ and for a random vector

$$X_j = (X_j^{(1)}, \dots, X_j^{(m)})$$

define

$$X_{\pi,j} := (X_j^{\pi(1)}, \dots, X_j^{\pi(m)}), \text{ i.e. } X_{\pi,j}^{(k)} = X_j^{\pi(k)}.$$

We can define a fully connected DAG $\mathcal{G}^\pi$ based on a permutation π in the following way. There is an edge between $\pi(i)$ to $\pi(k)$ in DAG $\mathcal{G}^\pi$, i.e. $\pi(i) \to \pi(k)$ if and only if $k > i$. For instance, $\pi(m)$ will be a sink node in $\mathcal{G}^\pi$. Note that different permutations correspond different fully connected DAGs and each fully connected DAG gives a permutation, which means correspondence is one to one. Similar to Bühlmann et al. [2014] let us define the set of true permutations as

$$\Pi^0 := \{\pi^0 : \mathcal{G}^{\pi^0} \text{ is a super-graph of true causal graph } \mathcal{G}^0\}.$$

All the permutations in Π^0 correspond to the order of $\mathcal{G}^0$. If the true DAG $\mathcal{G}^0$ is not fully connected there can be more than one true permutations, i.e. Π^0 has more than one element. For example, for the DAG depicted in Figure 9 true permutations are $\{(1, 2, 3, 4), (1, 3, 2, 4)\}$, which contains two elements.

Having a method to estimate a sink node we can estimate a causal order $\hat{\pi}$ in the following way. Let

$$\hat{\pi}(m) = \text{ sink node that produces a sink node estimation method.}$$

Then we remove $\hat{\pi}(m)$ from the set $[1, m]$ and estimate $\hat{\pi}(m-1)$, i.e. using the sink estimation method we estimate $\hat{\pi}(m-1)$ in the set $[1, m] \setminus \hat{\pi}(m)$. Continuing this process we obtain

$$\hat{\pi} = (\hat{\pi}(1), \dots, \hat{\pi}(m)). \tag{41}$$

The following Proposition states that if we can estimate a sink node consistently then we can also identify a causal order consistently.

Proposition 4.1. *Assume sink node estimation method is consistent with high probability, i.e. for arbitrary small $\delta > 0$ we have*

$$\mathbb{P}(\hat{\pi}(m) \text{ is a sink node }) \geq 1 - \delta/m \text{ as } n \to \infty.$$

Then $\hat{\pi}$ defined in (41) *is a consistent estimator with high probability of true causal order, i.e.*

$$\mathbb{P}(\hat{\pi} \in \Pi^0) \geq 1 - \delta \text{ as } n \to \infty.$$

Proof. For a set $A \subset [1, m]$ define a subgraph $\mathcal{G}_A$ of a DAG $\mathcal{G}$, by removing all the nodes that are not in A, i.e. not in $[1, m] \setminus A$ and their adjacent edges. Let us estimate the probability that estimated order is in the true set of permutations in the following way

$$\begin{aligned}
\mathbb{P}(\hat{\pi} \in \Pi^0) &= \mathbb{P}(\hat{\pi}(j) \text{ is a sink node in } \mathcal{G}^0_{[1,m]\setminus\{\hat{\pi}(j+1),\ldots,\hat{\pi}(m)\}} \text{ for } \forall j \in [1, m]) \\
&= 1 - \mathbb{P}(\exists j \in [1, m] : \hat{\pi}(j) \text{ is not a sink node in } \mathcal{G}^0_{[1,m]\setminus\{\hat{\pi}(j+1),\ldots,\hat{\pi}(m)\}}) \\
&\geq 1 - \sum_{j=1}^{m} \mathbb{P}(\hat{\pi}(j) \text{ is not a sink node in } \mathcal{G}^0_{[1,m]\setminus\{\hat{\pi}(j+1),\ldots,\hat{\pi}(m)\}}) \\
&\geq 1 - \sum_{j=1}^{m} \delta/m = 1 - \delta, \text{ as } n \to \infty,
\end{aligned}$$

where the first equality follows from the fact that if $\hat{\pi}(j)$ is not a sink node in a subgraph $\mathcal{G}^0_{[1,m]\setminus\{\hat{\pi}(j+1),\ldots,\hat{\pi}(m)\}}$ for some j implies there is a directed path from $\hat{\pi}(i)$ to $\hat{\pi}(j)$ for some $i < k$, which is a contradiction of the definition of Π^0. The second equality is just the complement rule of probability. The first inequality is the union bound of probability and the last one is just an application of the assumption of Proposition. This completes the proof of the Proposition. □

The above Proposition shows that provided consistent sink estimation method we can estimate the causal order consistently as well. Now let us understand how can we obtain such a method.

For each $k \in [1, m]$ let us assume that

$$\{(X_j^{(1)}, \ldots, X_j^{(k-1)}, X_j^{(k+1)}, \ldots, X_j^{(m)}), X_j^{(k)}\}_{j=1}^n$$

are i.i.d. samples from the Linear Transformation Model (5) with standard normal noise. Define

$$X_j^{(-k)} := (X_j^{(1)}, \ldots, X_j^{(k-1)}, X_j^{(k+1)}, \ldots, X_j^{(m)}).$$

Using the Pairwise Rank Likelihood (26) and Fixed Point Transformation function estimation (27) we obtain consistent estimators $\hat{h}$ and $\hat{\beta}$ of h and β_0 respectively.

Now using $\hat{h}$ and $\hat{\beta}$ we can estimate the noise for the other T observations and test whether or not noise is independent of potential causes in the following way. From the noise estimation method from (28) we obtain

$$\hat{\varepsilon}_j^{(k)} = \hat{h}(v_j^{(k)}) - (v_j^{(-k)})^T \hat{\beta} \text{ for } j \in [1, T]. \tag{42}$$

Let us define

$$t_k := HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T). \tag{43}$$

Now, let us define the estimator of a sink node in the following way

$$\hat{\pi}(m) := \underset{k}{argmin} \left\{ t_k \right\}. \tag{44}$$

Note that if V_j happens to be a sink node in the PNLG model where the functions $f_{j,1}$ are linear and if we estimated the parameters perfectly, the HSIC value should be close to zero as the noise is independent of nodes $V_j^{(-k)}$. The next result shows that under some conditions $\hat{\pi}(m)$ is a consistent sink node estimator.

In the following we state some assumptions and discuss them. We need these assumptions to make the arguments of the consistency results precise.

(A1) *Assume $X := (X^{(1)}, \dots, X^{(m)})$ be a random vector distributed according to the PNLG model. For each $k \in [1, m]$ and $A \subset [1, m] \setminus \{k\}$ such that $X^{(A)}$ contains at least one child of $X^{(k)}$ define*

$$N := h(X^{(k)}) - (X^{(A)})^T \beta,$$

for arbitrary strictly increasing function h and arbitrary vector β. Then, we assume

$$HSIC(\mathbb{P}^{N, X^{(A)}}) > \xi,$$

for some constant $\xi > 0$, which does not depend k, A, h or β.

The above assumption **(A1)** is necessary to make the PNLG model identifiable (i.e. there exists unique causal graph corresponding to the distribution). Each time sink node can be identified from the distribution since otherwise noise would not be independent of the causes. So, there will be unique causal graph corresponding to the distribution. However, this condition might be quite strict as it can happen that sink node cannot be identified but if we continue further we see that at some point we come across to a contradiction. This situations cannot be handled with our method since if at some point we estimate a sink node wrongly we cannot change it and the causal order will be wrong. Thus, this condition is somehow necessary for our method.

(A2) *Assume conditions of Theorems 2.2 and 2.3 are satisfied.*

Assumption **(A2)** is necessary to obtain the asymptotic results of the Linear Transformation Models and is discussed in Appendix A.1.

(A3) *Assume that functions $f_{k,1}$ are linear in the PNLG model, i.e. $f_{k,1}(\mathbf{PA}_k) = \mathbf{PA}_k^T\beta_k$ for some vector β_k, functions $f_{k,2}$ are strictly increasing.*

Assumption **(A3)** is a particular case of PNLG models, in order to make analysis more easy and develop practical algorithms. Generalization ideas can be found in the next section, where we see that not only we can drop the assumption of the Gaussianity of the noise but also the assumption **(A3)**. However, it can payoff with the computation cost of the practical algorithms.

The following result states the consistency of a sink node estimation method under the above assumptions.

Proposition 4.2. *Assume **(A1)** – **(A3)** hold and fix arbitrary small $\delta > 0$. Then, we have $\hat{\pi}(m)$ defined in (44) estimates a sink node consistently with high probability, that is*

$$\mathbb{P}(\hat{\pi}(m) \text{ is a sink node}) \geq 1 - \delta/m \text{ as } n \to \infty,$$

provided

$$\sqrt{\frac{\log(6m/\delta)}{\alpha^2 T}} + \frac{C}{T} < \frac{\xi}{2},$$

*where α, C are defined in Theorem 4.2 and ξ is defined in **(A1)**.*

Proof. Please see the proof in Appendix A.5.2. □

Now, let us state and prove the main Theorem of this Section, which is a consistency of the causal order estimation. Note that previous two proposition should give the desired consistency, as we can see it in the proof.

Theorem 4.3. *Assume that the conditions of Proposition 4.2 are satisfied. Then, $\hat{\pi}$ defined in (41) with the sink node estimation from (44) is a consistent estimation of true causal order with high probability, that is, for arbitrary small $\delta > 0$ we have*

$$\mathbb{P}(\hat{\pi} \in \Pi^0) \geq 1 - \delta \text{ as } n \to \infty.$$

Proof. From Proposition 4.2 we have $\mathbb{P}(\hat{\pi}(m) \text{ is a sink node}) \geq 1 - \delta/m$ as $n \to \infty$. So, the assumption of Proposition 4.1 is satisfied, which gives that

$$\mathbb{P}(\hat{\pi} \in \Pi^0) \geq 1 - \delta \text{ as } n \to \infty.$$

□

The sink node identification procedure is described in the Algorithm 8.

Algorithm 8: Multivariate PNL Sink Estimation Algorithm

```
Require: $\{X_1 := (X_1^{(1)}, X_1^{(2)}, \dots, X_1^{(m)}), \dots X_n := (X_n^{(1)}, X_n^{(2)}, \dots, X_n^{(m)})\}$,
         noise estimation algorithm.
$hsic = []$
for $j = [1, m]$ do
    $(u_1, \dots, u_n) \leftarrow$ estimated noise for sink $X^{(j)}$, i.e. $Y := X^{(j)}$ and
     $X := (X^{(1)}, \dots, X^{(j-1)}, X^{(j+1)}, \dots, X^{(n)})$ in the LTM (2)
    $hsic_j \leftarrow HSIC(\{(X_i^{(1)}, \dots, X_i^{(j-1)}, X_i^{(j+1)}, \dots, X_i^{(n)}), u_i\}_{i=1}^n)$
    $hsic \leftarrow [hsic, hsic_j]$
end for
$sink \leftarrow \underset{j}{argmin}\ \{hsic\}$
Return $sink$.
```

Using the above sink estimation algorithm we can estimate the causal order in the following Algorithm 9.

Algorithm 9: Multivariate PNL Order Estimation Algorithm

```
Require: $\{X_1 := (X_1^{(1)}, X_1^{(2)}, \dots, X_1^{(m)}), \dots X_n := (X_n^{(1)}, X_n^{(2)}, \dots, X_n^{(m)})\}$,
         noise estimation algorithm.
$order = []$
$remained \leftarrow [1, m]$
for $j = [1, m]$ do
    $sink \leftarrow$ get sink node using Algorithm 8 for the nodes $X_i$ for
     $i \in remained$
    $order \leftarrow [sink, order]$
    remove $sink$ from $remained$
end for
Return $order$.
```

4.5 Post-Nonlinear Causal Models

In this subsection we consider generalizations of the previous subsection, namely how can we drop some of the assumptions that we made before. The most general form of Post-Nonlinear causal models (Definition 1.2) is the following structural equations

$$X^{(j)} = f_{j,2}(f_{j,1}(\mathbf{PA}_j) + \varepsilon_j) \text{ for } \forall j \in [1, m],$$

where $\mathbf{PA}_j$ are the parents of $X^{(j)}$ in the causal graph $\mathcal{G}^0$ and $\varepsilon_j \perp\!\!\!\perp \mathbf{PA}_j$.

Now consider the following cases, where each of them is a generalization of PNLG models that we discussed in the previous subsection and the last one is the PNL models without any restriction (only the technical assumptions for theoretical results)

1. *Assume that functions $f_{j,1}$ are linear in the PNLG model, i.e. $f_{j,1}(\mathbf{PA}_k) = \mathbf{PA}_j^T \beta_j$ for some vector β_k, functions $f_{j,2}$ are strictly increasing.*

   Note that this is the assumption **(A3)** in the previous subsection and we do not put any restriction on the noise variables. So, the structural equations for the PNL model will become

   $$X^{(j)} = f_{j,2}(\mathbf{PA}_j^T \beta_j + \varepsilon_j) \text{ for } \forall j \in [1, m].$$

   Note that defining $h_j := f_{j,2}^{-1}$ we obtain

   $$h_j(X^{(j)}) = \mathbf{PA}_j^T \beta_j + \varepsilon_j \text{ for } \forall j \in [1, m],$$

   which is exactly the Linear Transformation Models (5), that is

   $$h(Y) = X^T \beta + \varepsilon,$$

   where ε is independent of X. So, using this model we can identify a sink node each time and remove it from the graph until we obtain the causal order.

   In the previous sections we developed the case when we have Gaussian noise, but as mentioned similar results also hold without assuming Gaussianity. So, we can take any estimator from the [Han, 1987, Cavanagh and Sherman, 1998, Abrevaya, 1999a,b, 2003, Yu et al., 2021] list and do the exact same analysis as we did for the Gaussian noise case. For instance, we can take Han [1987] Maximum Rank Correlation (MRC) estimator which is consistent and asymptotically normal [Cavanagh and Sherman, 1998]. To obtain MRC estimator let us define the following objective function

   $$\begin{aligned}\ell_{mrc}(\beta) = \binom{n}{2}^{-1} \sum_{i<j} & I(Y_j > Y_i) I((X_j - X_i)^T \beta > 0) \\ & + I(Y_j < Y_i) I((X_j - X_i)^T \beta < 0),\end{aligned} \tag{45}$$

   and MRC estimator of β will be

   $$\hat{\beta} = \underset{\beta}{argmax}\{\ell_{mrc}(\beta)\}. \tag{46}$$

   Having a estimator of β like in the previous sections we can use it to obtain a estimator $\hat{h}$ of h, i.e. [Horowitz, 1996, Chen, 2002, Zhang, 2013]. Then the rest is the same as the PNLG models as we only need asymptotic results for the $\hat{h}$ and $\hat{\beta}$ which also hold in this case.

One disadvantage of the objective function in (45) is its discontinuity and its maximization is computationally expensive. To tackle this problem, the estimator based on the objective function suggested by Zhang [2013] can be used, which is continuous.

2. *Assume PNL models (Definition 1.2) without any further assumptions*

   Here the idea is to use the General Transformation Models 3 to obtain the estimate of the noise and proceed in the same way as the PNLG case. Since, structural equations have the following form

$$X^{(j)} = f_{j,2}(f_{j,1}(\mathbf{PA}_j) + \varepsilon_j) \text{ for } \forall j \in [1, m],$$

   as in the previous case we can look at the following transformation model

$$h(Y) = g(X) + \varepsilon,$$

   which is exactly the General Transformation Models (29). Using the method described in section 3 we can obtain estimators $\hat{h}$ and $\hat{g}$ of h and g correspondingly. Note that these estimators are again asymptotically normal and so the same analysis for the consistency of the causal order estimation holds in this case also. One of the drawbacks in this method is that even for the bivariate case and sample size 100 the method takes more than 7 hours to obtain the causal order.

The above two cases have been discussed very shortly, but they are interesting topic of future research. In particular, reducing the computational cost of the second case can produce reliable algorithm to estimate the casual directions in the most general PNL models.

5 Experimental Results

This section is an overview of the experimental results for the algorithms described in the previous sections and their comparison. Implantation of the Algorithms have been done in R and can be found in the following GitHub `https://github.com/grigor97/pnl_gaussian` page. For the implementations (Python/PyTorch) of the papers [Uemura and Shimizu, 2020, Uemura et al., 2022] please refer to `https://github.com/grigor97/ab_pnl`.

In the next first subsection reviews the results of the Linear Transformation Models and compares discussed algorithms with each other. Then the second subsection is the review of the PNL results with Gaussian noise (both bivariate and multivariate cases).

5.1 Results of Linear Transformation Models

Experiments for the Linear Transformation Models (5) have been carried out mainly for a sample size 1000 (for 500 and 3000 sample sizes similar results have been obtained) for one dimensional β_0 and each experiment (fixed β_0 and sample size) have been repeated 100 times. Since the model in (5) is symmetric for β_0 (i.e. having $-\beta_0$ instead of β_0 in the equation $h(Y) = X^T\beta_0 + \varepsilon$ we can just negate the values in X to have $h(Y) = (-X)^T(-\beta_0) + \varepsilon)$, we only considered non-negative β_0. The actual values of β_0 in the experiments are $\beta_0 \in [0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1, 10, 30, 50, 70, 100, 1000]$. Values of X have been simulated from standard normal distribution, i.e. $X \sim \mathcal{N}(0, 1)$ and function h cubic function, i.e. $h(Y) = (Y - c)^3$ for some constant c. The noise ε as described in the algorithms is standard normal distributed.

In the following are the results of the Algorithms described in the Section 2. The first we introduce the numerical comparison of described algorithms and their corresponding estimators then make visualization of the results.

Tables 2 3 and 4 are the Relative Bias (RB), Variance (Var) and Mean Squared Error (MSE) numerical results on the simulated datasets, respectively. Relative Bias of an estimator

$hat\beta$ is computed in the following way

$$RB_{\hat{\beta}} := \frac{\hat{\beta} - \beta_0}{\beta_0},$$

where β_0 is the true parameter and it is not zero. Since we have repeated each experiment 100 times and every time we obtain some estimate, we compute the mean of the all estimates and then use the above formula to obtain Relative Bias. Variance and the Mean Squared Error are the standard estimates (for the variance we have n in the denominator instead of $n - 1$). Rows in the Tables show the different values of true parameter β_0 and the columns are for different estimators. The smallest values for each value of β_0 is boldface in the Tables. For the Expected Rank Algorithm only the L_2 penalty version is reported in here as the others not even comparable, which we can see in plots discussed next.

Comparing the different estimators we see that Pairwise Rank Likelihood estimator $\hat{\beta}_{PRL}$ is the best among all. Its Relative Bias is almost all the cases smaller than the others. For the values it is not the smallest we can see that it shows comparable

Relative Bias (RB)				
β_0 \ Est.	$\hat{\beta}_{MC}$	$\hat{\beta}_{SPF}$	$\hat{\beta}_{ER}$ (L_2, $\lambda = 1$)	$\hat{\beta}_{PRL}$
0.01	**0.07463830**	-0.38476481	0.153509414	-0.1884277362
0.1	0.04116651	-0.11114560	-0.021546194	**0.0016160692**
0.3	-0.01656634	-0.08849482	0.006836647	**-0.0058695061**
0.5	-0.04880277	-0.1048635	0.000776208	**-0.0070462067**
0.7	-0.11633692	-0.11139786	-0.005621234	**-0.0014367806**
0.9	-0.20450478	-0.08167034	-0.005647065	**0.0004179769**
1	-0.24206052	-0.11180298	**-0.000550831**	-0.0008016458
10	-0.89321580	-0.70401173	0.026115395	**-0.0031277489**
30	-0.96424925	-0.89717817	-0.211420737	**-0.0012698640**
50	-0.97855179	-0.93869024	-0.411303826	**-0.0011620624**
70	-0.98468334	-0.95586403	-0.542852247	**0.0015114792**
100	-0.98927378	-0.96908991	-0.662049306	**0.0019802367**
1000	-0.99892702	-0.99692616	-0.963831477	**-0.0023829947**

Table 2: Relative Bias results for algorithms, $n = 1000$ and each experiment repeated 100 times.

results. For the case of variance we see that only for the first two smallest β_0 PRL estimator produces smallest variance and for large values it is significantly larger than the others. The reason for this is that the other estimators tend to estimate β_0 very small and so have smaller variance. However, if we compare the results of the MSE's PRL estimator outperforms all the others. Note that $\hat{\beta}_{MC}$ estimator have smaller variance and in some cases even smaller MSE than $\hat{\beta}_{PRL}$, but it cannot estimate β_0 properly for large values. Moreover, $\hat{\beta}_{MC}$ produces smaller MSE (which is again comparable to the MSE of $\hat{\beta}_{PRL}$) only for smaller values of β_0. For the case of Expected Rank Algorithm with L_2 penalty we see that it also has smaller variance and smaller MSE for one value of β_0. The reason for smaller MSE is expected for us since we penalize the estimator with L_2 norm and also it might perform well for smaller values of β_0. All in all, $\hat{\beta}_{PRL}$ estimator is the best of all especially for large values of β_0 and the others are not even comapreable in this case.

In order to have visual understanding how the different algorithms performed in different occasions let us visualize them in the following. The following plots represent the boxplots of the estimated parameter β_0 for a given algorithm and the red dots are the true values of β_0 which are also provided in the x axis.

Figure 1 shows the results of the Monte Carlo Algorithm 1 for $M = 1000$ and maximum iterations 100.

Variance (Var)				
$\beta_0 \setminus$ Est.	$\hat{\beta}_{MC}$	$\hat{\beta}_{SPF}$	$\hat{\beta}_{ER}$ (L_2, $\lambda = 1$)	$\hat{\beta}_{PRL}$
0.01	0.0011508935	0.012582517	0.0009262215	**8.309943e-04**
0.1	0.0011936741	0.012089770	0.0012562599	**9.562384e-04**
0.3	**0.0007916878**	0.013961263	0.0011547485	1.328395e-03
0.5	**0.0006302459**	0.010734662	0.0011522811	1.341047e-03
0.7	**0.0004535486**	0.009515386	0.0016379383	9.416307e-04
0.9	**0.0003130218**	0.007776599	0.0017335710	1.884255e-03
1	**0.0003260043**	0.007166803	0.0018085053	1.599025e-03
10	**0.0009515591**	0.018573563	3.9904497069	6.878423e-02
30	**0.0008952047**	0.019629901	1.9867132388	5.612309e-01
50	**0.0008756016**	0.017714052	3.1891878291	1.274070e+00
70	**0.0008986833**	0.016322445	3.3426258427	3.396206e+00
100	**0.0008614268**	0.019277317	4.2921893238	1.001695e+01
1000	**0.0009306356**	0.021337788	5.4034255887	2.168555e+03

Table 3: Variance results for algorithms, $n = 1000$ and each experiment repeated 100 times.

Mean Squared Error (MSE)				
$\beta_0 \setminus$ Est.	$\hat{\beta}_{MC}$	$\hat{\beta}_{SPF}$	$\hat{\beta}_{ER}$ (L_2, $\lambda = 1$)	$\hat{\beta}_{PRL}$
0.01	1.151451e-03	1.259732e-02	9.285781e-04	**8.345448e-04**
0.1	1.210621e-03	1.221330e-02	1.260902e-03	**9.562645e-04**
0.3	**8.163878e-04**	1.466608e-02	1.158955e-03	1.331496e-03
0.5	**1.225674e-03**	1.348375e-02	1.152432e-03	1.353459e-03
0.7	7.085345e-03	1.559603e-02	1.653421e-03	**9.426423e-04**
0.9	3.418901e-02	1.317933e-02	**1.759401e-03**	1.884397e-03
1	5.891930e-02	1.966671e-02	1.808809e-03	**1.599667e-03**
10	7.978440e+01	4.958183e+01	4.058651e+00	**6.976251e-02**
30	8.367999e+02	7.244554e+02	4.221557e+01	**5.626822e-01**
50	2.393910e+03	2.202866e+03	4.261163e+02	**1.277446e+00**
70	4.751047e+03	4.477029e+03	1.447317e+03	**3.407400e+00**
100	9.786627e+03	9.391372e+03	4.387385e+03	**1.005616e+01**
1000	9.978552e+05	9.938618e+05	9.289765e+05	**2.174233e+03**

Table 4: MSE results for algorithms, $n = 1000$ and each experiment repeated 100 times.

For the algorithms which estimate large β_0 as a small value a few plots have been showed in order to have clear visual understanding how they have estimated the values for the different scales of β_0. For instance, in the Figure 1 there are three plots the first one is only for β_0 not greater than 1 the second one is for not grater than 30 and the last one is for all values of β_0, which also can be seen in the Figure. Of course all the information available in the first plot is also available in the subsequent ones also but not in a proper visual scale. If we look for the value of $\beta_0 = 0.1$ only the first plot provides exact information how well it is estimated.

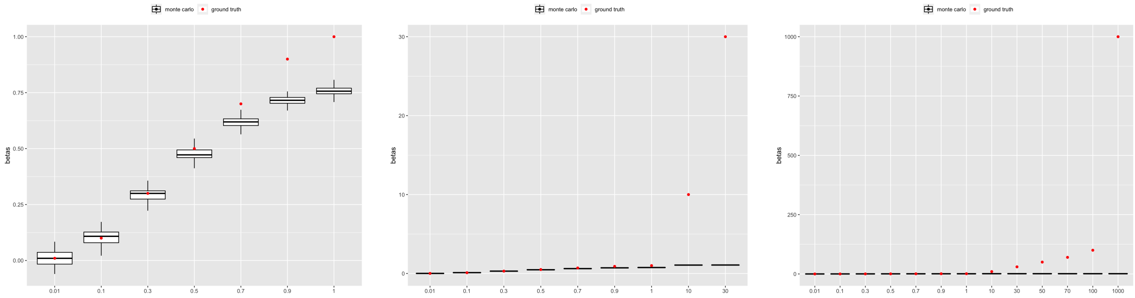

Figure 1: Monte Carlo Algorithm 1 results on sample size 1000.

In the following Figure 2 are the results of the Stochastic Fixed Point Algorithm 4 for batch size 64 and maximum iterations 100. Note that non stochastic version of this algorithm would have been computationally much more expensive and the batch size have been chosen some fixed number as larger ones also give similar results.

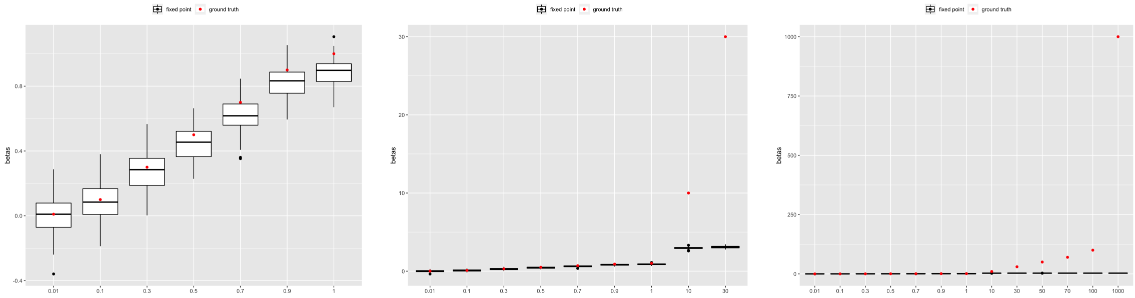

Figure 2: Fixed Point Algorithm 4 results on sample size 1000.

For the Expected Rank Algorithm 5 there are three different versions that we have discussed (no penalty and L_1, L_2 penalties). Figure 3 shows the results of the algorithm with no penalty (only on plot here since it estimates the values β_0 very large even for small β_0). Note that this algorithm estimates the values very large which is a expected result as discussed during the development of the Algorithm, i.e. objective function can be minimized (locally) by making the absolute value of β large.

For the case of L_1 penalty, three different values of penalty strength $\lambda = 1, 10, 100$

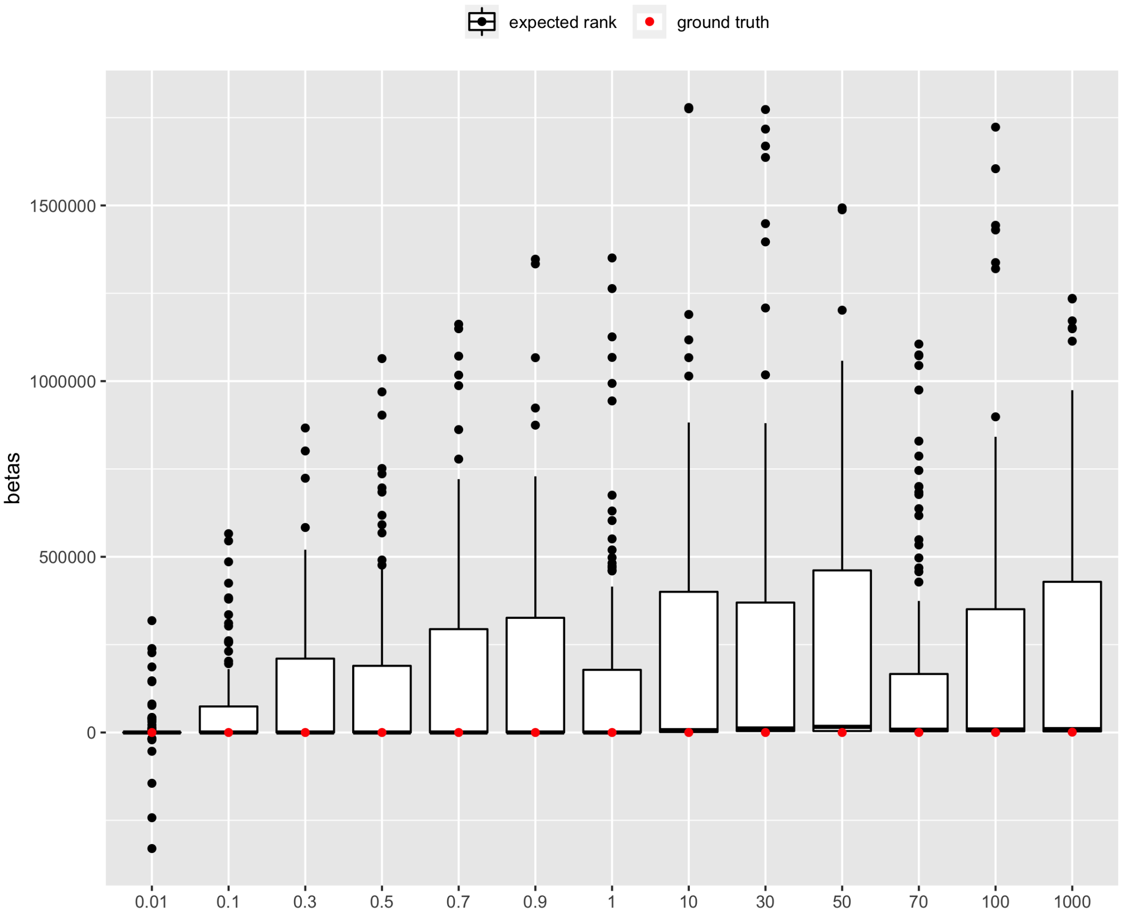

Figure 3: Expected Rank Algorithm 5 (with no penalty) results on sample size 1000.

have been tried. Figure 4 shows the results for $\lambda = 1, 10, 100$ from left to right, respectively (again one plot from each since the estimated values are large even for small β_0). Note that bigger the strength λ less noisy is the estimation, however even for the case $\lambda = 100$ still we have very large estimates.

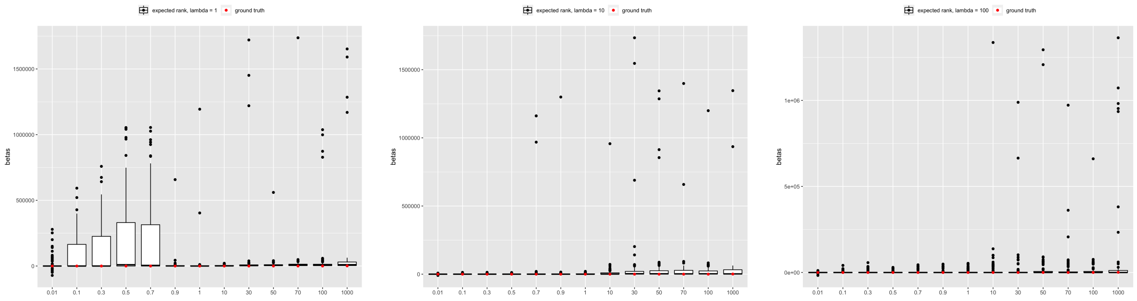

Figure 4: Expected Rank Algorithm 5 (with L_1 penalty) results on sample size 1000.

For the case of L_2 penalty, three different values of penalty strength $\lambda = 0.1, 1, 10$ have been tried for $\beta_0 \leq 1$. Figure 5 shows the results for $\lambda = 0.1, 1, 10$ from left to right, respectively. In this case we see that for the case of 0.1 strength there still could be large estimation, but for the case of larger strengths (i.e. $\lambda \geq 1$) we have reasonable estimation.

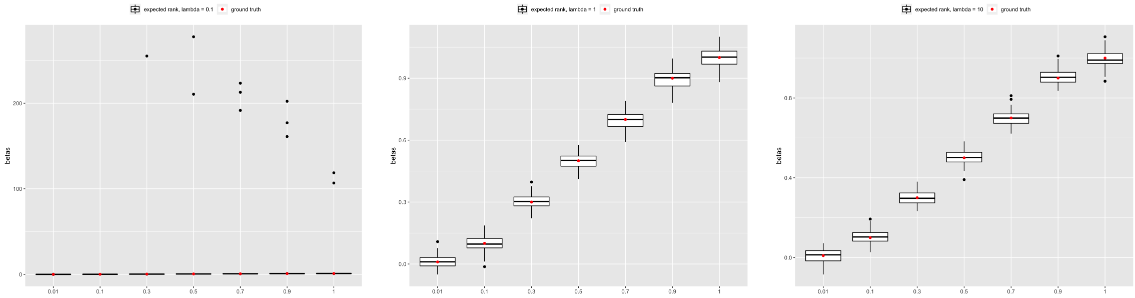

Figure 5: Expected Rank Algorithm 5 (with L_2 penalty) results on sample size 1000.

For larger β_0 (i.e. $\beta_0 = 100, 1000$) have been tried for penalty strength 1 and Figure 6 shows the corresponding results.

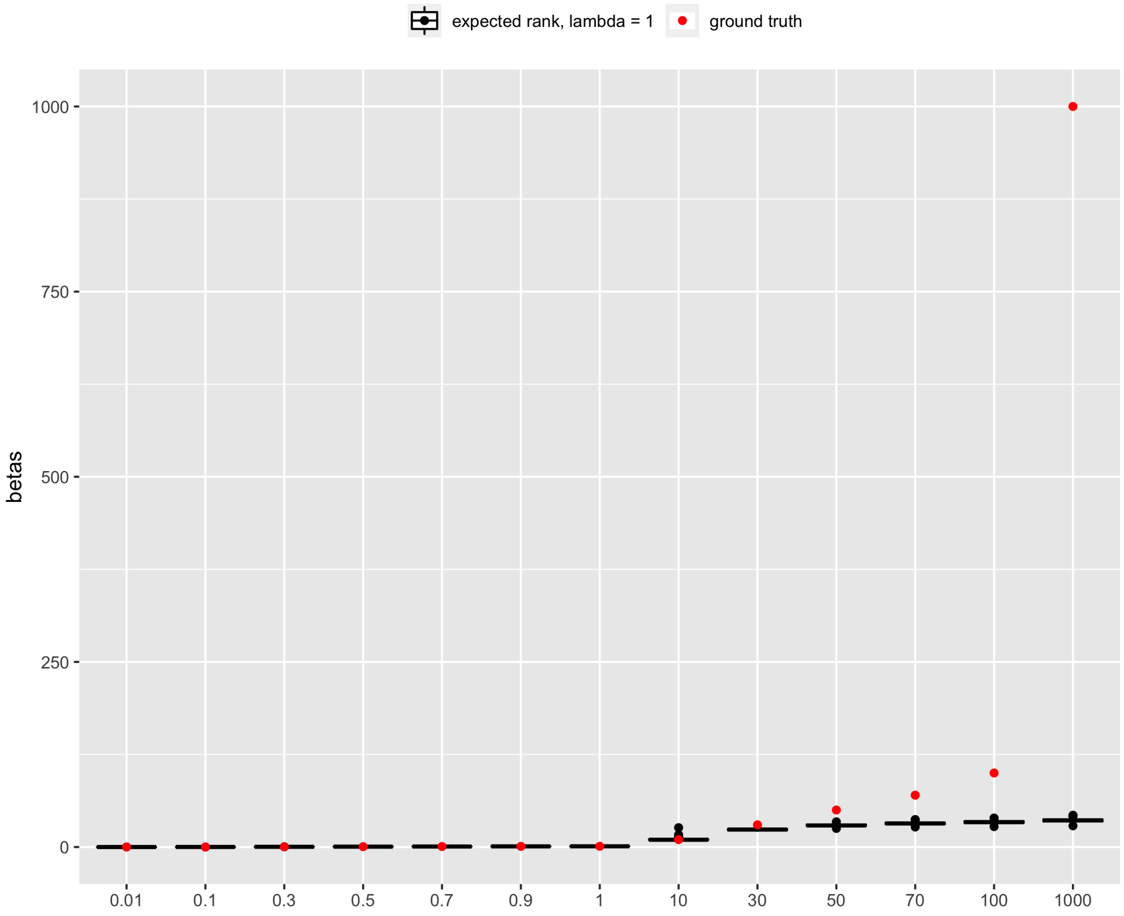

Figure 6: Expected Rank Algorithm 5 (with L_2 penalty) results on sample size 1000.

From the above results we understand that algorithms tend to shrink the size of estimated β_0 (except the non penalty and L_1 penalty Expected Rank Algorithm).

Moreover, the case of Expected Rank Algorithm with L_2 penalty term for some values of penalty strength the algorithm produces better estimates for $beta_0$ for larger than 1 case, but it also does not work for large $\beta_0 = 1000$.

Now let us look at the results of Pairwise Rank Likelihood Algorithm 6. Figure 7 shows these results for β_0 less than or equal 1, 100, 1000 from left to right, respectively. We see that even for the case $\beta_0 = 1000$ Pairwise Rank Likelihood Algorithm estimates it correctly (median of the estimators of 100 experiments correspond to the true value β_0 and there is almost no outliers).

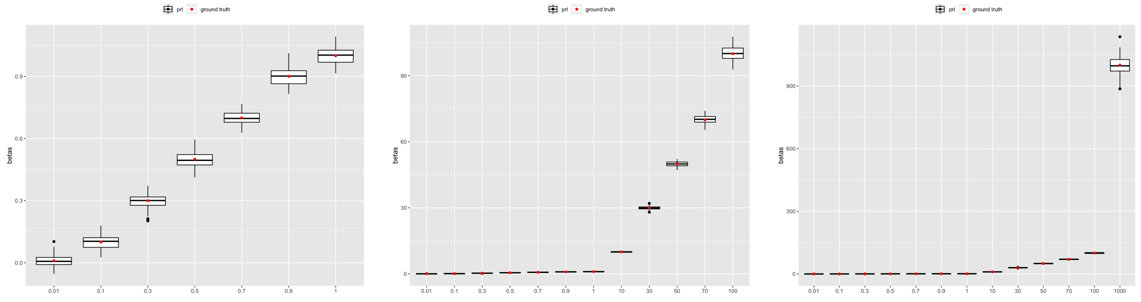

Figure 7: Pairwise Rank Likelihood Algorithm 6 on sample size 1000.

Comparing all the above results together we understand that Pairwise Rank Likelihood Algorithm performs significantly better than the others. For large β_0 the other algorithms always shrink the parameter value, but the last algorithm estimates it correctly.

The transformation function h estimation result is depicted in the Figure 8. The red line is the true function h and the black dots are the estimated results on the sample data points. We can see that estimation works quite good for all the points in the dataset.

5.2 Results of PNL Models

The experiments here are also carried out mainly for sample size 1000 and 100 times for each specific situation. The first part is for the bivariate PNL models and the second part is for multivariate PNL models.

5.2.1 Results of Bivariate PNL models

For the bivariate models we simulate a dataset in the following way. For the cause variable $X^{(1)}$ we simulate from standard normal distribution, i.e. $X^{(1)} \sim \mathcal{N}(0,1)$ and follow two models, namely identifiable (see Proposition 4.2) and non-identifiable (see Proposition 4.1). For the function f_2 we take $f_2(z) := z^{\frac{1}{3}} + 4.7$ for both cases

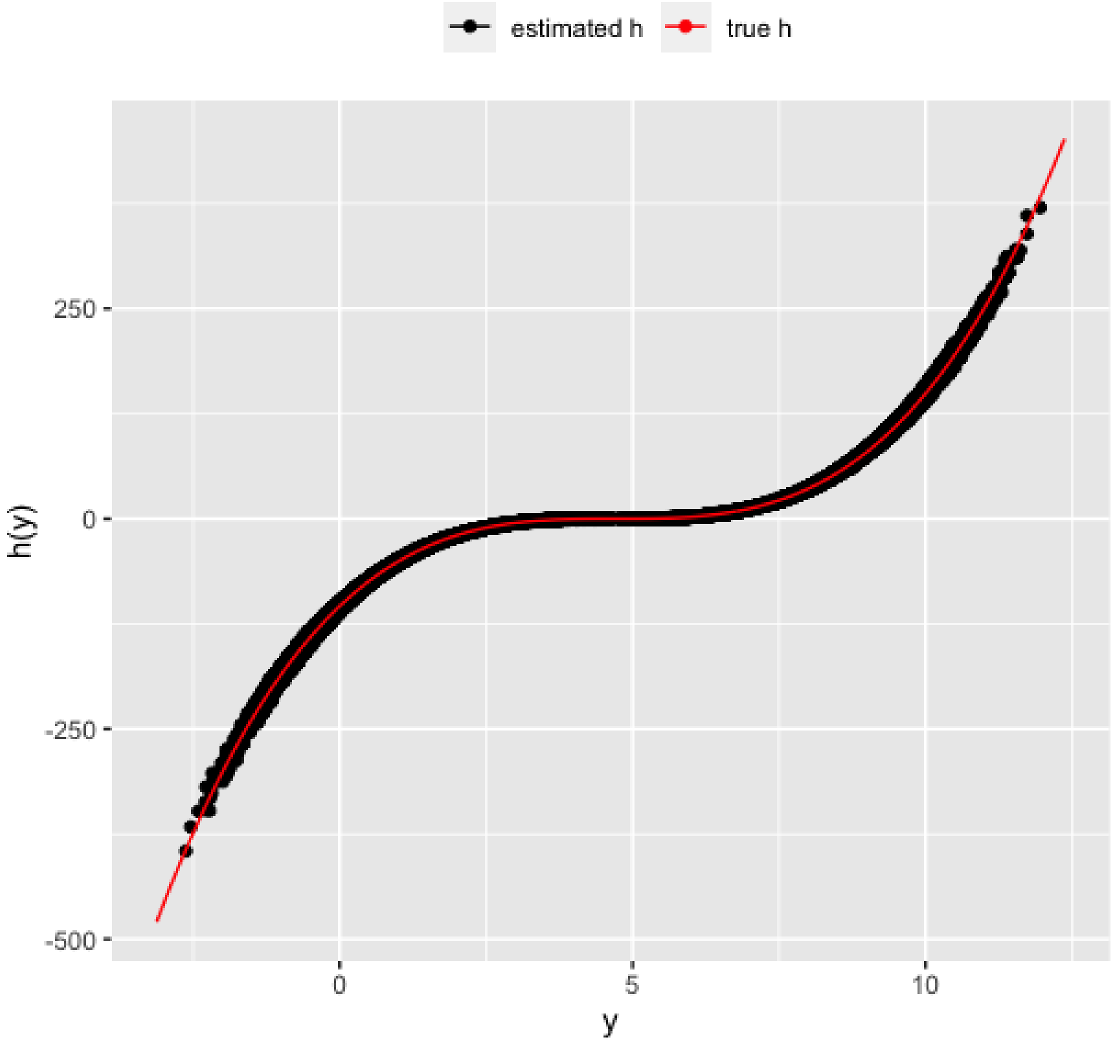

Figure 8: Transformation function estimation on sample size 1000.

and for the parameter β we simulate it from a uniform distribution on $(-100, 100)$, i.e. $\beta \sim U(-100, 100)$ in each experiment. So, for the identifiable case we have

$$X^{(2)} = (\beta(X^{(1)})^2 + \varepsilon)^{\frac{1}{3}} + 4.7,$$

where $X^{(1)}, \varepsilon_2 \sim \mathcal{N}(0, 1)$ and $\beta \sim U(-100, 100)$, $\varepsilon \perp\!\!\!\perp X^{(1)}$. For the non-identifiable case we have

$$X^{(2)} = (\beta X^{(1)} + \varepsilon)^{1/3} + 4.7,$$

where $X^{(1)}, \varepsilon_1 \sim \mathcal{N}(0, 1)$ and $\beta \sim U(-100, 100)$, $\varepsilon \perp\!\!\!\perp X^{(1)}$. For both cases we used the Algorithm 7, where the noise estimation algorithm is is the combination of Linear Transformation Algorithm(i.e. Algorithm 6) and (27) as described in Section 4. For both cases we fitted the models using the degree two polynomial for the cause. i.e. if we fit the model as $X^{(1)}$ is the cause and $X^{(2)}$ is the effect, we used the model $h(X^{(2)}) = \beta_1 X^{(1)} + \beta_2 (X^{(1)})^2 + \varepsilon$ in Linear Transformation Algorithms. From the Linear Transformation Models' results we saw that Pairwise Rank Likelihood Algorithm performed the best and as a results it produces the best results also for the bivariate PNL models as we can see in the Table 5.

LTM alg \ model	identifiable	non-identifiable
Fixed Point	72%	72%
Expected Rank L_1 ($\lambda = 10$)	72%	93%
Expected Rank L_2 ($\lambda = 10$)	96%	98%
PRL	**98%**	**99%**

Table 5: Bivariate PNL results, $n = 1000$ and each experiment repeated 100 times.

The above table shows the results of two different bivariate PNL models(identifiable and non-identifiable as described at the beginning of the subsection) using four different Linear Transformation Algorithms, namely (from top to bottom in the table) Fixed Point 4, Expected Rank 5 with L_1 penalty and $\lambda = 10$, Expected Rank 5 with L_2 penalty and $\lambda = 10$ and Pairwise Rank Likelihood 6.

Note that results in Table 5 are only for the cases if the underlying true model corresponds to our model assumptions, i.e. noise is standard normal, f_1 function in the PNL model is quadratic.

5.2.2 Results of Multivariate PNL models

Pairwise Rank Likelihood Algorithm 6 performed better than the others both in the Linear Transformation Models and in Bivariate PNL Models. Moreover, it requires less computational time than the others. On top of that, its estimator is asymptotically normal and consistent as showed in Theorem 2.2. Considering all that, we only use Pairwise Rank Likelihood for the multivariate PNL model estimation. For the estimation of the transformation function we use (27) as in the bivariate case. These two provide the estimation of the noise. First example for the multivariate PNL model we look at the diamond shape four node causal graph, i.e. Figure 9. Note that for this case in all experiments we have the same graph only the values of the nodes will change. For the second example we look at random Erdős–Rényi causal graph where each edge has a probability of $2/(m-1)$ where the m is the number of nodes in the graph. Example of Erdős–Rényi graph is depicted in Figure 10. Note that, in this case for each experiment we sample a random Erdős–Rényi graph and then simulate the PNL data according the the graph.

Having the graph, data simulation is carried out in the following way. If $X^{(j)}$ has no parents in the graph, it is simulated from a standard normal distribution,

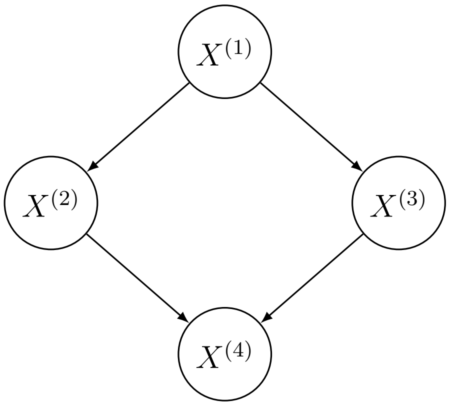

Figure 9: Diamond shape graph

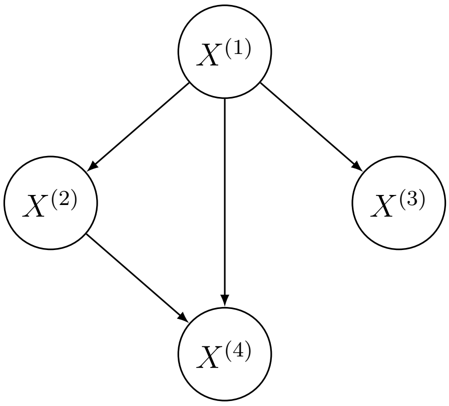

Figure 10: Example of Erdős–Rényi graph for 4 nodes

i.e. $X^{(j)} \sim \mathcal{N}(0,1)$. If $X^{(j)}$ has k parents, namely $X^{(j_1)}, \dots X^{(j_k)}$, then $X^{(j)}$ will be $X^{(j)} = (\sum_{i=1}^{k}(\beta_i X^{(j_i)} + \gamma_i (X^{(j_1)})^2) + \varepsilon_j)^{\frac{1}{3}} + 4.7$, where $\varepsilon_j \sim \mathcal{N}(0,1)$ and $\beta_i, \gamma_i \sim U(-100, 100)$ for $i \in [1, k]$. For the diamond case X^4 depends on its parent only in a linear way.

Table 6 shows the results of order estimation Algorithm 9 for sample size 1000 and 100 experiments. We can see that for the diamond case results are much better and reasonable explanation might be that the model is simpler than the other cases.

repetitions	Diamond (4 nodes)	Erdős–Rényi (4 nodes)	Erdős–Rényi (7 nodes)
100	67%	18%	2%
1000	62.8%	17.5%	1.5%

Table 6: Multivariate PNL results, $n = 1000$.

sample size	Diamond (4 nodes)	Erdős–Rényi (4 nodes)	Erdős–Rényi (7 nodes)
1000	67%	18%	2%
2000	71%	26%	6%
3000	77%	40%	2%

Table 7: Multivariate PNL results for different sample sizes. Number of datasets is 100.

For the case of Post-Nonlinear models (without Gaussian noise assumption and $f_{j,1}$ linear) using the General Transformation models computational time is very expensive even for the bivariate case. For the sample size 100 it takes more than 7 hours to estimate the causal direction. However, using the Pairwise Rank Likelihood for the PNLG models takes only a few minutes for the sample size 1000.

6 Discussion

In this thesis we studied Post-Nonlinear causal model with Gaussian noise. We employed Linear Transformation Models to obtain a sink node estimation method, which produced causal order of the underlying true data generation process. Estimated causal order is proved to be consistent with high probability. The performance of the method is tested on the simulated data.

For the Linear Transformation Models (LTM) with Gaussian noise we developed a Pairwise Rank Likelihood method to estimate the parameters. Estimated parameters proved to be consistent and asymptotically normal. We also studied existing other LTMs, but the Pairwise Rank Likelihood gives the best result both in theory and in practice.

Using LTMs we have estimated the noise based on the potential sink node (and remaining nodes considered potential causes). Having the estimated noise we used HSIC conditional independence criterion to test if the noise is independent of its potential causes. The node which produced the least independence have been chosen as sink node. Then, we removed estimated sink node and iterated same method until a causal order. Using the quantitative property of the HSIC criterion we proved the estimated causal order is consistent with high probability.

In order to generalize the above idea for general PNL models (without Gaussian noise) we reviewed the General Transformation Models. Using the kernel smoothing method estimators of the model have been obtained, which have similar asymptotic properties as LTM estimators. This fact later on used to show how the method developed for PNL models with Gaussian noise can be transferred to the general case.

A Appendix

Here we review some of the necessary results to obtain theoretical properties of our estimators and presented the proofs which are more involved.

In A.1 we state and review the conditions which are necessary to obtain a asymptotic results of Linear Transformation Models. In A.2 stochastic Landau symbols and consistency definition have been reviewed. A.3 discusses some of the relevant results for U-statistics which later used to prove the asymptotic results of Linear Transformation Models. Then in A.4 and A.5 the proofs of Linear Transformation Models and Post-Nonlinear Models are presented, respectively.

A.1 Conditions

Here are some technical conditions, which are used to prove results in the text. Most of them are standard assumptions used in statistical theory and some are easily satisfied, which are discussed after the conditions.

Condition 0: With probability one the design matrix $\mathbf{X} := [\mathbf{1}^T, X_1^T, \dots, X_n^T]^T$ has full column rank.

Condition 1: Probability of ties in Y_i's is zero, i.e. $\mathbb{P}(Y_i = Y_j) = 0$ for each $i \neq j$.

Condition 2: The following requirements are satisfied

$$\mathbb{E}_X \left[\frac{\phi^2\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)\Phi\left(U_{12}^T\beta_0\right)} \cdot (U_{21})^2 \right] < \infty$$

and

$$\Sigma_\psi \text{ has full rank}$$

where $U_{21} = X_2 - X_1$ and $U_{12} = X_1 - X_2$ and functions ϕ and Φ are probability density and cumulative density functions of standard normal random variable, respectively. For a kernel

$$\psi((X_1,Y_1),(X_2,Y_2)) := I(Y_2 > Y_1)\frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + I(Y_2 \leq Y_1)\frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12}$$

define $\psi_1((x_1,y_1)) := \mathbb{E}[\psi((x_1,y_1),(X_2,Y_2))]$ and $\Sigma_\psi := Var(\psi_1((X_1,Y_1)))$.

First of all let us see that Condition 0 is easily satisfied. For the condition be satisfied it is required that number of columns in $\mathbf{X}$ is at least the number of rows, which is just a requirement that sample size is at least one bigger than number of dimension of each data point. Now if we take arbitrary square sub-matrix in $\mathbf{X}$ and compute the determinant of it, we obtain some non-zero polynomial of $X_1, \dots, X_n$.

The Lemma (only one lemma in the paper) in [Okamoto, 1973] states that such a polynomial can be zero only on the Lebesgue measure zero. So, Condition 0 is a reasonable assumption.

Condition 1 is satisfied if random variable Y has a continuous distribution with respect to the Lebesgue measure, which covers many cases.

Condition 2 is a standard assumption in order to prove asymptotic results for U-statistics, i.e. see the proof the asymptotic results in [Yu et al., 2021].

A.2 Stochastic Landau Symbols and Consistency

Assume we have random vectors $X_1, X_2, \dots$. Then, the followings are the definitions of Stochastic Landau Symbols:

1. $X_n = o_p(1)$ if $X_n \overset{p}{\to} 0$.
2. $X_n = O_p(1)$ if $\forall \epsilon > 0 \quad \exists C > 0$ such that $sup_{n \in \mathbb{N}} \mathbb{P}(\|X_n\| > C) < \epsilon$.
3. $X_n = o_p(\alpha_n)$ if $X_n = \alpha_n Y_n$ with $Y_n = o_p(1)$
4. $X_n = O_p(\alpha_n)$ if $X_n = \alpha_n Y_n$ with $Y_n = O_p(1)$

We say X_n is consistent to X if $\|X_n - X\| = o_p(1)$, X_n is strongly consistent to X if $X_n \overset{a.s.}{\to} X$, where a.s. means almost sure convergence. X_n is α_n consistent to X if $\|X_n - X\| = O_p(\alpha_n)$.

Since, we use $\|X_n - X\| = o_p(1)$ and $X_n - X = o_p(1)$ interchangeably let us prove their equivalence.

Proposition A.1. *Assume $\{X_n\}_{n=1}^{\infty}$ and X are defined on the same probability space. Then, $\|X_n - X\| = o_p(1)$ if and only if $X_n - X = o_p(1)$.*

Proof. For the direction $\implies$ assume $\|X_n - X\| = o_p(1)$. By the norm property we have

$$|X_n^{(j)} - X_n^{(j)}| \leq \|X_n - X\| \text{ for } j \in [1, m],$$

where m is the dimension of X. So, we have

$$\mathbb{P}(|X_n^{(j)} - X_n^{(j)}| > \epsilon) \leq \mathbb{P}(\|X_n - X\| > \epsilon) \to 0 \text{ as } n \to \infty,$$

which is the same as $X_n^{(j)} - X_n^{(j)} = o_p(1)$ for each j and so $X_n - X = o_p(1)$.

For the direction $\impliedby$ assume $X_n - X = o_p(1)$. Now for every $\epsilon > 0$ we have

$$\mathbb{P}(\|X_n - X\| > \epsilon) = \mathbb{P}(\|X_n - X\|^2 > \epsilon^2) = \mathbb{P}(\sum_{j=1}^{m} |X_n^{(j)} - X_n^{(j)}|^2 > \epsilon^2)$$

$$
\begin{aligned}
&= \mathbb{P}(\exists j \in [1,m] \text{ s.t. } |X_n^{(j)} - X_n^{(j)}|^2 > \epsilon^2/m) \\
&\leq \sum_{j=1}^{m} \mathbb{P}(|X_n^{(j)} - X_n^{(j)}|^2 > \epsilon^2/m) \\
&= \sum_{j=1}^{m} \mathbb{P}(|X_n^{(j)} - X_n^{(j)}| > \epsilon/\sqrt{m}) \to 0 \text{ as } n \to \infty,
\end{aligned}
$$

where the first equality is just the definition of the norm. Second equality follows from the fact that in order to some of the m variables be greater then ϵ^2 we need at least one of them is greater than ϵ^2/m. The inequality is just a union bound. Last equality is application of square root on both sides of the inequality and the last convergence follows from the assumption $X_n - X = o_p(1)$ and that m is fixed finite number. □

A.3 U-statistics

Since we are going to use the properties of U-statistics in some of the proofs, we will introduce them in the following subsection A.3 and restate the results that we will use.

The following is mostly from the books [Serfling, 1980] and [Vaart, 1998]. For more detailed summary of the U-statistics please refer to the books.

Let $X_1, X_2, \ldots$ be independent samples from some distribution F. Assume for a parametric function $\theta := \theta(F)$, i.e. the mean of the distribution, there is an unbiased estimator. That is

$$\theta = \mathbb{E}[\psi(X_1, \ldots, X_k)],$$

for some symmetric function ψ, called a "kernel".

The following estimator for the estimation of θ based on the sample $X_1, \ldots, X_n$ of size $n \geq m$

$$U_n := U(X_1, \ldots, X_n) = \binom{n}{k}^{-1} \sum \psi(X_{i_1}, \ldots, X_{i_k})$$

is called U-statistic, where the summation is over all possible $\binom{n}{k}$ combinations of distinct elements $\{i_1, \ldots, i_k\}$ from $\{1, \ldots, n\}$. The linearity of the expectation gives that U_n is an unbiased estimator for θ.

The following Theorem A.1 is a generalization of the classical (i.e. for i.i.d. sample) Strong Law of Large Numbers (SLLN), which is the Theorem A in the [Serfling, 1980] of section 5.4.

Theorem A.1. *If $\mathbb{E}[|\psi(X_1, \ldots, X_m)|] < \infty$, then $U_n \overset{a.s.}{\to} \theta$, where a.s. means almost sure convergence.*

Asymptotic normality of the U-statistics can be established by projection method. The projection of $U_n - \theta$ onto the set of all statistics of the form $\sum_{j=1}^{n} g_j(X_j)$ for some function $g_1, \dots, g_n$, is defined as

$$\bar{U}_n = \sum_{j=1}^{n} \mathbb{E}[U_n - \theta | X_j] = \frac{k}{n} \sum_{j=1}^{n} \psi_1(X_j),$$

where

$$\psi_1(x) := \mathbb{E}[\psi(x, X_2, \dots, X_k)] - \theta$$

and define the first order variance of the U-statistics as

$$\Sigma_1 := Var(\psi_1(X_1)).$$

The following establishes the asymptotic normality of the U-statistics under the condition that first order variance exists, i.e. see [Vaart, 1998] Theorem 12.3.

Theorem A.2. *If $\mathbb{E}[(\psi(X_1, \dots, X_m))^2] < \infty$, then $\sqrt{n}(U_n - \theta - \bar{U}_n) = o_p(1)$. Consequently,*

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, k^2 \Sigma_1).$$

The idea of $\bar{U}_n$ is to project the U-statistics U_n that their difference will be asymptotically negligible and the projection is sum of i.i.d. samples and use the Central Limit Theorem for it.

A.4 Proofs of Section 2

A.4.1 Proof of Proposition 2.2

Before proving proposition 2.2 let's state and prove two lemma's which immediately imply 2.2 in order to keep the proofs readable.

Lemma A.1. *The following inequality holds for arbitrary $z \in \mathbb{R}$*

$$\phi'(z)\Phi(z) - (\phi(z))^2 < 0.$$

Proof. Denoting $h(z) := \phi'(z)\Phi(z) - (\phi(z))^2$ we need to show that $h(z) < 0$. Substituting the values of the derivative of the ϕ we will have

$$h(z) = -z\phi(z)\Phi(z) - (\phi(z))^2 = \phi(z)(-z\Phi(z) - \phi(z)).$$

Since $\phi(z) > 0$ for arbitrary z, $h(z) < 0$ is equivalent to $g(z) := -z\Phi(z) - \phi(z) < 0$ which is obvious for $z \geq 0$. In order to prove that $g(z) < 0$ let's look at the derivative of g which is

$$g'(z) = -\Phi(z) - z\phi(z) + z\phi(z) = -\Phi(z) < 0.$$

So, g is strictly decreasing function, which means that it will be maximum when $z \to -\infty$ which is

$$\begin{aligned}\lim_{z\to-\infty} g(z) &= \lim_{z\to-\infty}(-z\Phi(z) - \phi(z)) = -\lim_{z\to-\infty} z\Phi(z) = -\lim_{z\to-\infty}\frac{\Phi(z)}{\frac{1}{z}}\\ &= \lim_{z\to-\infty}\frac{\phi(z)}{\frac{1}{z^2}} = \lim_{z\to-\infty}\frac{1}{\sqrt{2\pi}}\frac{z^2}{e^{z^2/2}} = 0,\end{aligned}$$

where the first equality of the second line follows from the L'Hôpital's rule. Thus, $g(z)$ is strictly decreasing and the limit in $-\infty$ is zero whcih gives $g(z) < 0$ for all $z \in \mathbb{R}$ and this completes the proof of the lemma. □

Lemma A.2. *The function*

$$f(x) = \log\Phi(c^T x)$$

is concave, where Φ is the CDF function of standard normal distribution and $x, c \in \mathbb{R}^m$ for some nonzero constant c and $m \in \mathbb{N}$. Moreover, if $v^T\nabla^2 f(x)v = 0$ for some vector v and Hessian matrix $\nabla^2 f(x)$ if and only if $v^T c = 0$.

Proof. Since function f is twice differentiable, it is enough to show that Hessian of $f(x)$ is negative definite. The gradient of f is the following

$$\nabla f(x) = \frac{\phi(c^T x)}{\Phi(c^T x)}c,$$

where ϕ is the pdf function of standard normal distribution, which gives that the Hessian is

$$\nabla^2 f(x) = \frac{\phi'(c^T x)\Phi(c^T x) - (\phi(c^T x))^2}{(\Phi(c^T x))^2}\cdot cc^T.$$

So, for arbitrary $v \in \mathbb{R}^m$ we have

$$\begin{aligned}v^T\nabla^2 f(x)v &= v^T\frac{\phi'(c^T x)\Phi(c^T x) - (\phi(c^T x))^2}{(\Phi(c^T x))^2}\cdot cc^T v\\ &= \frac{\phi'(c^T x)\Phi(c^T x) - (\phi(c^T x))^2}{(\Phi(c^T x))^2}\cdot v^T cc^T v\\ &= \frac{\phi'(c^T x)\Phi(c^T x) - (\phi(c^T x))^2}{(\Phi(c^T x))^2}\cdot (v^T c)^2 \le 0,\end{aligned}$$

where the last step follows from the lemma A.1 and the fact that $(v^T c)^2 \ge 0$. Moreover, $v^T\nabla^2 f(x)v = 0$ if and only if $v^T c = 0$, which completes the proof the the lemma.

□

Now let's prove the proposition 2.2.

Proof. From 25 we have

$$\ell_{prl}(\beta) = \binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i) \log \Phi\left(\frac{(X_j - X_i)^T\beta}{\sqrt{2}}\right)$$
$$+ I(Y_j \leq Y_i) \log \Phi\left(\frac{(X_i - X_j)^T\beta}{\sqrt{2}}\right)$$

and since indicator function return either one or zero, lemma A.2 gives that $\ell_{prl}(\beta)$ is a sum of concave functions. As sum preserves the concavity, we have $\ell_{prl}(\beta)$ is concave.

Now assume that $\ell_{prl}(\beta)$ is not strictly concave. This imples that there is a vector v such that $v^T\nabla^2\ell_{prl}(\beta)v = 0$ for the Hessian matrix $\nabla^2\ell_{prl}(\beta)$ of $\ell_{prl}(\beta)$. Since the Hessian operator is linear, we have $\nabla^2\ell_{prl}(\beta)$ is a sum of the Hessians, that is

$$\nabla^2\ell_{prl}(\beta) = \binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i)\nabla^2 \log \Phi\left(\frac{(X_j - X_i)^T\beta}{\sqrt{2}}\right)$$
$$+ I(Y_j \leq Y_i)\nabla^2 \log \Phi\left(\frac{(X_i - X_j)^T\beta}{\sqrt{2}}\right).$$

Lemma A.2 gives that $v^T\nabla^2 \log \Phi\left(\frac{(X_j-X_i)^T\beta}{\sqrt{2}}\right) v = 0$ if and only if $v^T(X_j - X_i) = 0$, but the same things implies $v^T\nabla^2 \log \Phi\left(\frac{(X_i-X_j)^T\beta}{\sqrt{2}}\right) v = 0$ condition. Moreover, we have for each fixed i, j, exactly one of the $I(Y_j > Y_i)$ or $I(Y_j \leq Y_i)$ is 1. So, $v^T\nabla^2\ell_{prl}(\beta)v = 0$ implies that $v^T(X_j - X_i) = 0$ for all i and j. This means that $X_j^T v$ is constant for all j, i.e. $X_j^T v = c$ for some constant $c \in \mathbb{R}$, which gives the matrix $\mathbf{X} := [\mathbf{1}^T, X_1^T, \ldots, X_n^T]$ does not have full column rank, i.e. taking $u = (-c, v^T)^T$ implies $\mathbf{X}u = 0$. This contradicts to the assumption of the Proposition, which means $\ell_{prl}(\beta)$ is strictly concave function, which completes the proof.

□

A.4.2 Proof of Theorem 2.2

In order to keep the formulas compact and readable, define $U_{ij} = \frac{X_i - X_j}{\sqrt{2}}$, which gives

$$\ell_{prl}(\beta) = \binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i) \log \Phi\left(U_{ji}^T\beta\right) + I(Y_j \leq Y_i) \log \Phi\left(U_{ij}^T\beta\right).$$

Since we will use the Taylor expansion of $\ell_{prl}(\hat{\beta}_{PRL})$ around β_0, we need some properties of the gradient of $\ell_{prl}(\beta)$ at β_0, which are established in the following lemmas.

Lemma A.3. *Assume that the Condition 1 in the Appendix A.1 holds, then* $\nabla_\beta\ell_{prl}(\beta_0)$ *converges to zero almost surely, that is*

$$\nabla_\beta\ell_{prl}(\beta_0) \overset{a.s.}{\to} 0.$$

Proof. The gradient is

$$\nabla_\beta \ell_{prl}(\beta_0) = \binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i) \frac{\phi\left(U_{ji}^T\beta_0\right)}{\Phi\left(U_{ji}^T\beta_0\right)} \cdot U_{ji} + I(Y_j \le Y_i) \frac{\phi\left(U_{ij}^T\beta_0\right)}{\Phi\left(U_{ij}^T\beta_0\right)} \cdot U_{ij},$$

which we recognize, to be a U-statistic for the kernel

$$\psi((X_1,Y_1),(X_2,Y_2)) := I(Y_2 > Y_1) \frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + I(Y_2 \le Y_1) \frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12}.$$

Note that, $\psi((X_1,Y_1),(X_2,Y_2))$ is a symmetric kernel as Condition 1 in the Appendix A.1 assumes that there is no ties in Y_i's and so $I(Y_2 \le Y_1) = I(Y_2 < Y_1)$, which implies that $\psi((X_1,Y_1),(X_2,Y_2)) = \psi((X_2,Y_2),(X_1,Y_1))$.

Since we would like to use the Theorem A.1 let us compute the expectation of the kernel ψ and then show that its absolute value has finite first moment.

$$\begin{aligned}
&\mathbb{E}[\psi((X_1,Y_1),(X_2,Y_2))] \\
&\quad = \mathbb{E}\left[I(Y_2 > Y_2) \frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + I(Y_j \le Y_i) \frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12} \right] \\
&\quad = \mathbb{E}_X\left[\mathbb{E}_Y\left[I(Y_j > Y_i) | X \right] \frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + \mathbb{E}_Y\left[I(Y_j \le Y_i) | X \right] \frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12} \right] \\
&\quad = \mathbb{E}_X\left[\mathbb{P}(Y_j > Y_i | X) \frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + \mathbb{P}(Y_j \le Y_i | X) \frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12} \right] \\
&\quad = \mathbb{E}_X\left[\Phi\left(U_{21}^T\beta_0\right) \frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + \Phi\left(U_{12}^T\beta_0\right) \frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12} \right] \\
&\quad = \mathbb{E}_X\left[\phi\left(U_{21}^T\beta_0\right) \cdot U_{21} + \phi\left(U_{12}^T\beta_0\right) \cdot U_{12} \right] = \mathbb{E}_X\left[\phi\left(U_{21}^T\beta_0\right) \cdot \left[U_{21} + U_{12}\right] \right] \\
&\quad = \mathbb{E}_X\left[\phi\left(U_{21}^T\beta_0\right) \cdot \left[\frac{X_j - X_i}{\sqrt{2}} + \frac{X_i - X_j}{\sqrt{2}} \right] \right] = \mathbb{E}_X[0] = 0,
\end{aligned}$$

where the first two equalities follow from the linearity of the expectation and the tower rule of the expectation, i.e. $\mathbb{E}[Q(X,Y)] = \mathbb{E}[\mathbb{E}[Q(X,Y)|X]]$ for any function Q. The third equality is just an application of expectation on the indicator function is the probability. The fourth equality from the monotonicity of function h in the Linear Transformation model. The fifth step is just a cancellation of the equal members in the fractions. Note that here they will be canceled only if we are considering the true parameter β_0 in $\mathbb{E}[\nabla_\beta \ell_{prl}(\beta_0)]$. Finally, the sixth equality follows from the fact that $\phi(x) = \phi(-x)$ for the probability density function of standard normal random variable ϕ.

Now let us show that absolute value the kernel ψ has finite expectation, where by the absolute value of a vector $c = (c_1, \ldots, c_d)^T$ for some $d \in \mathbb{N}$ we mean by element wise absolute values, i.e. $|c| := (|c_1|, \dots, |c_d|)$. So,

$$\begin{aligned}
&\mathbb{E}[|\psi((X_1, Y_1), (X_2, Y_2))|] \\
&\quad \leq \mathbb{E}\left[I(Y_2 > Y_2)\frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot |U_{21}| + I(Y_j \leq Y_i)\frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot |U_{12}|\right] \\
&\quad = \mathbb{E}_X\left[\Phi\left(U_{21}^T\beta_0\right)\frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot |U_{21}| + \Phi\left(U_{12}^T\beta_0\right)\frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot |U_{12}|\right] \\
&\quad = \phi\left(U_{12}^T\beta_0\right) \cdot \frac{2 \cdot |X_i - X_j|}{\sqrt{2}} < \infty,
\end{aligned}$$

where in the first step we used the triangle inequality for the absolute value and the others are similar to the calculation for the expectation of the kernel ψ. The last quantity is finite since we have probability density function of the standard normal distribution is bounded on the whole real line.

Having the requirements satisfied in the Theorem A.1 and the expectation of the kernel is zero, we have

$$\nabla_\beta \ell_{prl}(\beta_0) \overset{a.s.}{\to} 0,$$

which completes the proof of Lemma. □

Lemma A.4. *Assume that the Condition 1-2 in the Appendix A.1 hold, then* $\sqrt{n}\nabla_\beta \ell_{prl}(\beta_0)$ *is asymptotically normal in the following way*

$$\sqrt{n}\nabla_\beta \ell_{prl}(\beta_0) \overset{d}{\to} \mathcal{N}(0, \Sigma_\psi),$$

where Σ_ψ *defined in Condition 2.*

Proof. Similar to the proof of the Lemma A.3 we have

$$\nabla_\beta \ell_{prl}(\beta_0) = \binom{n}{2}^{-1} \sum_{i<j} I(Y_j > Y_i)\frac{\phi\left(U_{ji}^T\beta_0\right)}{\Phi\left(U_{ji}^T\beta_0\right)} \cdot U_{ji} + I(Y_j \leq Y_i)\frac{\phi\left(U_{ij}^T\beta_0\right)}{\Phi\left(U_{ij}^T\beta_0\right)} \cdot U_{ij},$$

is a U-statistics with symmetric kernel

$$\psi((X_1, Y_1), (X_2, Y_2)) := I(Y_2 > Y_1)\frac{\phi\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot U_{21} + I(Y_2 \leq Y_1)\frac{\phi\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot U_{12}.$$

Since we would like to use the Theorem A.2 we need to show that $\mathbb{E}[(\psi((X_1, Y_1), (X_2, Y_2)))^2] < \infty$, that is

$$\begin{aligned}
&\mathbb{E}[(\psi((X_1, Y_1), (X_2, Y_2)))^2] \\
&\quad = \mathbb{E}\left[I(Y_2 > Y_1)\frac{\phi^2\left(U_{21}^T\beta_0\right)}{\Phi^2\left(U_{21}^T\beta_0\right)} \cdot (U_{21})^2 + I(Y_2 \leq Y_1)\frac{\phi^2\left(U_{12}^T\beta_0\right)}{\Phi^2\left(U_{12}^T\beta_0\right)} \cdot (U_{12})^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_X \left[\Phi\left(U_{21}^T\beta_0\right) \frac{\phi^2\left(U_{21}^T\beta_0\right)}{\Phi^2\left(U_{21}^T\beta_0\right)} \cdot (U_{21})^2 + \Phi\left(U_{12}^T\beta_0\right) \frac{\phi^2\left(U_{12}^T\beta_0\right)}{\Phi^2\left(U_{12}^T\beta_0\right)} \cdot (U_{12})^2 \right] \\
&= \mathbb{E}_X \left[\frac{\phi^2\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)} \cdot (U_{21})^2 + \frac{\phi^2\left(U_{12}^T\beta_0\right)}{\Phi\left(U_{12}^T\beta_0\right)} \cdot (U_{12})^2 \right] \\
&= \mathbb{E}_X \left[\frac{\phi^2\left(U_{21}^T\beta_0\right)}{\Phi\left(U_{21}^T\beta_0\right)\Phi\left(U_{12}^T\beta_0\right)} \cdot (U_{21})^2 \right] < \infty
\end{aligned}
$$

where the square of a vector is taken element wise. The first equality follows from the facts that square of the indicator function is the same and $I(Y_2 > Y_1) \cdot I(Y_2 \leq Y_1) = 0$. The second equality is the tower rule of the expectation and that function h is increasing. The third one is just cancellation of terms. In the fourth equality we used that $U_{21} = -U_{12}, U_{21}^2 = U_{12}^2, \Phi\left(U_{12}^T\beta_0\right) = 1 - \Phi\left(U_{21}^T\beta_0\right)$. The final inequality is just the Condition 2.

Now we have that the conditions of the Theorem A.2 are satisfied for U-statistics $\nabla_\beta \ell_{prl}(\beta_0)$ and note that the expectation of the kernel is zero, it gives

$$\sqrt{n}\nabla_\beta \ell_{prl}(\beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_\psi),$$

which complete the proof.

□

Lemma A.5. *Assume* $\|\beta - \beta_0\|_2 = a$ *for some* $a > 0$ *and Condition 0 in the in the Appendix A.1 is satisfied. Then*

$$\mathbb{P}(\ell_{prl}(\beta) < \ell_{prl}(\beta_0) \quad \forall \beta, \|\beta - \beta_0\| = a) \to 1.$$

Proof. Taylor expansion around β_0 gives

$$\ell_{prl}(\beta) - \ell_{prl}(\beta_0) = (\beta - \beta_0)^T \nabla_\beta \ell_{prl}(\beta_0) + \frac{1}{2}(\beta - \beta_0)^T \nabla_\beta^2 \ell_{prl}(\beta^*)(\beta - \beta_0),$$

where $\|\beta^* - \beta_0\|_2 \leq \|\beta - \beta_0\|_2 = a$, i.e. β^* is in the closed ball around β_0 with radius a. Clearly, $\nabla_\beta^2 \ell_{prl}(\beta^*)$ is continuous with respect to β^*). Moreover, Proposition 2.2 gives that maximum eigenvalue of $\nabla_\beta^2 \ell_{prl}(\beta')$ negative for all β' such that $\|\beta^* - \beta_0\|_2 \leq a$. So, continuity of the eigenvalues of the matrix gives that $\nabla_\beta^2 \ell_{prl}(\beta')$ eigenvalues of all the matrices such that $\|\beta' - \beta_0\|_2 \leq a$ has a maximum eigenvalue less than zero by the compactness of the ball around β_0 and radius a. In particular, $\nabla_\beta^2 \ell_{prl}(\beta^*)$ matrix has maximum eigenvalue $\lambda_{max} < 0$, which gives that

$$\frac{1}{2}(\beta - \beta_0)^T \nabla_\beta^2 \ell_{prl}(\beta^*)(\beta - \beta_0) \leq \frac{\lambda_{max}}{2} \|\beta - \beta_0\|_2 .$$

On the other hand, the term $(\beta - \beta_0)^T \nabla_\beta \ell_{prl}(\beta_0)$ can be made arbitrarily small by the previous Lemma, that is, Lemma A.3 gives with probability tends to 1

$$|(\beta - \beta_0)^T \nabla_\beta \ell_{prl}(\beta_0)| < -\frac{\lambda_{max}}{4} \|\beta - \beta_0\|_2 .$$

From the above inequalities, we obtain, with probability tends to 1

$$(\beta-\beta_0)^T\nabla_\beta\ell_{prl}(\beta_0)+\frac{1}{2}(\beta-\beta_0)^T\nabla^2_\beta\ell_{prl}(\beta^*)(\beta-\beta_0)<\frac{\lambda_{max}}{4}\left\|\beta-\beta_0\right\|_2<0.$$

Thus

$$\mathbb{P}(\ell_{prl}(\beta)<\ell_{prl}(\beta_0)\quad\forall\beta,\left\|\beta-\beta_0\right\|_2=a)\to 1.$$

□

Now let us prove the first part of the Theorem 2.2. i.e. $\hat{\beta}_{PRL}-\beta_0=o_p(1)$.

Proof. From Lemma A.5, we have for arbitrary fixed $a>0$ there is a maximum of $\ell_{prl}(\beta)$ with probability tends to 1, such that $\left\|\beta-\beta_0\right\|_2<a$, but $\hat{\beta}_{PRL}$ is the maximum of the $\ell_{prl}(\beta)$, so,

$$\mathbb{P}\left(\left\|\hat{\beta}_{PRL}-\beta_0\right\|_2<a\right)\to 1 \text{ as } n\to\infty,$$

which is the same as $\hat{\beta}_{PRL}-\beta_0=o_p(1)$. □

Now let us prove the second part of the Theorem.

Proof. Since $\hat{\beta}_{PRL}$ maximizes the pairwise rank log-likelihood, it make sthe gradient of it zero. Then, first order Taylor expansion of the gradient of pairwise rank log-likelihood gives

$$0=\nabla_\beta\ell_{prl}(\hat{\beta}_{PRL})=\nabla_\beta\ell_{prl}(\beta_0)+\nabla^2_\beta\ell_{prl}(\beta^*)(\hat{\beta}_{PRL}-\beta_0),$$

for some β^* such that $\left\|\beta^*-\beta_0\right\|_2\leq\left\|\hat{\beta}_{PRL}-\beta_0\right\|_2=o_p(1)$ (from the first part of the theorem). The above equality gives

$$\sqrt{n}(\hat{\beta}_{PRL}-\beta_0)=(-\nabla^2_\beta\ell_{prl}(\beta^*))^{-1}(\sqrt{n}\nabla_\beta\ell_{prl}(\beta_0)).$$

The second part of the Theorem will be proved if we show that

$$\sqrt{n}\nabla_\beta\ell_{prl}(\beta_0)\xrightarrow{d}\mathcal{N}(0,\Sigma_\psi) \text{ and} \tag{47}$$

$$-\nabla^2_\beta\ell_{prl}(\beta^*)\xrightarrow{p}\Sigma, \tag{48}$$

Since by Slutsky we will have

$$(-\nabla^2_\beta\ell_{prl}(\beta^*))^{-1}(\sqrt{n}\nabla_\beta\ell_{prl}(\beta_0))\xrightarrow{d}\Sigma^{-1}\mathcal{N}(0,\Sigma_\psi)$$

From Lemma A.4 we have that (47) holds. In order to prove (48) let us calculate the hessian of the $\ell_{prl}(\beta^*)$). Using the fact that for a function $f(\beta)=\log\Phi(c^T\beta)$ the hessian is $\nabla_\beta f(\beta)=\frac{\phi'(c^T\beta)\Phi(c^T\beta)-(\phi(c^T\beta))^2}{(\Phi(c^T\beta))^2}\cdot cc^T$ for a constant vector c, we have

$$\begin{aligned}\nabla^2_\beta \ell_{prl}(\beta^*) &= \binom{n}{2}^{-1}\sum_{i<j} I(Y_j > Y_i)\frac{\phi'(U_{ji}^T\beta^*)\Phi(U_{ji}^T\beta^*) - (\phi(U_{ji}^T\beta^*))^2}{(\Phi(U_{ji}^T\beta^*))^2}\cdot U_{ji}U_{ji}^T \\ &+ \binom{n}{2}^{-1}\sum_{i<j} I(Y_j \leq Y_i)\frac{\phi'(U_{ij}^T\beta^*)\Phi(U_{ij}^T\beta^*) - (\phi(U_{ij}^T\beta^*))^2}{(\Phi(U_{ij}^T\beta^*))^2}\cdot U_{ij}U_{ij}^T \\ &\xrightarrow{p} \binom{n}{2}^{-1}\sum_{i<j} I(Y_j > Y_i)\frac{\phi'(U_{ji}^T\beta_0)\Phi(U_{ji}^T\beta_0) - (\phi(U_{ji}^T\beta_0))^2}{(\Phi(U_{ji}^T\beta_0))^2}\cdot U_{ji}U_{ji}^T \\ &+ \binom{n}{2}^{-1}\sum_{i<j} I(Y_j \leq Y_i)\frac{\phi'(U_{ij}^T\beta_0)\Phi(U_{ij}^T\beta_0) - (\phi(U_{ij}^T\beta_0))^2}{(\Phi(U_{ij}^T\beta_0))^2}\cdot U_{ij}U_{ij}^T \\ &= \nabla^2_\beta \ell_{prl}(\beta_0),\end{aligned}$$

where the convergence in probability follows from the fact that $\beta^* \xrightarrow{p} \beta_0$ and the Continuous Mapping Theorem (i.e. Theorem 2.3 in [Vaart, 1998]). Now, once again using the Continuous Mapping Theorem we obtain

$$-\nabla^2_\beta \ell_{prl}(\beta^*) \xrightarrow{p} -\nabla^2_\beta \ell_{prl}(\beta_0) = \Sigma,$$

where Σ is positive definite as $\ell_{prl}(\beta)$ is strictly concave provided Condition 0 holds. This completes the proof of (48) and so the proof of the Theorem.

□

A.5 Proofs of Section 4

Here are the proofs of the results in Section 4

A.5.1 Proofs of Bivariate PNL

The following is the proof of Corollary 4.1.

Proof. Choosing $\varepsilon_1 := \frac{1}{\sqrt{\beta^2+1}}X^{(1)} - \frac{\beta}{\sqrt{\beta^2+1}}\varepsilon_2$ we obtain

$$X^{(1)} = \frac{1}{\sqrt{\beta^2+1}}\left(\frac{\beta}{\sqrt{\beta^2+1}}(X^{(2)})^3 + \varepsilon_1\right). \tag{49}$$

Using equation 49, Corollary will be proved if we show that $\varepsilon_1 \sim \mathcal{N}(0,1)$ and $\varepsilon_1 \perp\!\!\!\perp X^{(2)}$. Since ε_1 is a linear combination of two independent standard normal random variables, it follows that ε_1 is a normal random variable. Moreover,

$$\mathbb{E}[\varepsilon_1] = \mathbb{E}\left[\frac{1}{\sqrt{\beta^2+1}}X^{(1)} - \frac{\beta}{\sqrt{\beta^2+1}}\varepsilon_2\right] = \frac{1}{\sqrt{\beta^2+1}}\mathbb{E}[X^{(1)}] - \frac{\beta}{\sqrt{\beta^2+1}}\mathbb{E}[\varepsilon_2] = 0,$$

and

$$\mathbb{V}ar(\varepsilon_1) = \left(\frac{1}{\sqrt{\beta^2+1}}\right)^2 + \left(\frac{\beta}{\sqrt{\beta^2+1}}\right)^2 = 1,$$

which implies that $\varepsilon_1 \sim \mathcal{N}(0,1)$. Now it remained to show that ε_1 and $X^{(2)}$ are independent. For that purpose let's look at the transformation from $(X^{(1)}, \varepsilon_2)$ to $(X^{(2)}, \varepsilon_1)$. The Jacobian of the transformation is

$$\mathbf{J} = \begin{bmatrix} \frac{\beta}{3}(\beta x^{(1)} + e_2)^{-2/3} & \frac{1}{3}(\beta x^{(1)} + e_2)^{-2/3} \\ \frac{1}{\sqrt{\beta^2+1}} & -\frac{\beta}{\sqrt{\beta^2+1}} \end{bmatrix}.$$

So, $|det(\mathbf{J})| = \frac{\sqrt{\beta^2+1}}{3}(\beta x^{(1)} + e_2)^{-2/3} = \frac{\sqrt{\beta^2+1}}{3}(x^{(2)})^{-2}$, which gives that

$$\begin{aligned}
p_{X^{(2)},\varepsilon_1}(x^{(2)}, e_1) = p_{X^{(1)},\varepsilon_2}(x^{(1)}, e_2)|det(\mathbf{J})|^{-1} &= \frac{1}{2\pi} exp\left[-\frac{(x^{(1)})^2}{2} - \frac{e_2^2}{2}\right] \frac{3(x^{(2)})^2}{\sqrt{\beta^2+1}} \\
&= \frac{1}{2\pi} exp\left[-\frac{1}{2}\left((x^{(1)})^2 + ((x^{(2)})^3 - \beta x_1)^2\right)\right] \frac{3x_2^2}{\sqrt{\beta^2+1}} \\
&= \frac{1}{2\pi} exp\left[-\frac{1}{2}\left(\frac{\beta}{\beta^2+1}(x^{(2)})^3 + \frac{1}{\sqrt{\beta^2+1}}e_1\right)^2\right] \\
&\quad exp\left[-\frac{1}{2}\left((x^{(2)})^3 - \frac{\beta^2}{\beta^2+1}(x^{(2)})^3 - \frac{\beta}{\sqrt{\beta^2+1}}e_1\right)^2\right] \frac{3(x^{(2)})^2}{\sqrt{\beta^2+1}} \\
&= \frac{3(x^{(2)})^2}{2\pi\sqrt{\beta^2+1}} exp\left[-\frac{1}{2}\left(\frac{\beta}{\beta^2+1}(x^{(2)})^3 + \frac{1}{\sqrt{\beta^2+1}}e_1\right)^2\right] \\
&\quad exp\left[-\frac{1}{2}\left(\frac{1}{\beta^2+1}(x^{(2)})^3 - \frac{\beta}{\sqrt{\beta^2+1}}e_1\right)^2\right] \\
&= \left[\frac{3(x^{(2)})^2}{\sqrt{2\pi}\sqrt{\beta^2+1}} exp\left[-\frac{1}{2}\frac{1}{\beta^2+1}(x^{(2)})^6\right]\right] \times \left[\frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}e_1^2\right]\right].
\end{aligned}$$

Above equality gives $X^{(2)}$ and ε_1 are independent, which completes the proof. □

The following is the proof of Corollary 4.2.

Proof. The intuition that the model is identifiable is that the values of $X^{(2)}$ depend only on the absolute value of $X^{(1)}$. So, the sign information is lost in $X^{(2)}$ and it should not be possible to obtain $X^{(1)}$ from $X^{(2)}$.

For the strict proof let's assume that backward direction is also possible, that is there exist invertible function g_2 and potentially nonlinear function g_1 such that

$$X^{(1)} = g_2(g_1(X^{(2)}) + \varepsilon_1),$$

where $\varepsilon_1 \perp\!\!\!\perp X^{(2)}$. Since noise variable ε_2 is standard normal, Table 1 gives that $T := g_2^{-1}(X^{(1)})$ is normal random variable and $\beta(g_2(T))^2$ function is linear in T. This is not possible since normality gives that values of T are the whole real line and $\beta(g_2(T))^2$ always has the sign of β. So, backward direction is not possible and this is the same as being identifiable. □

A.5.2 Proofs of Multivariate PNL

In order to facilitate the proofs let us state and prove the following lemma.

Lemma A.6. *Assume X is a p dimensional random vector and denote $Y = g(X)$ for some function $g : \mathbb{R}^p \to \mathbb{R}$. Let $\{X_j\}_{j=1}^n$ be n i.i.d. copies of X and $\hat{g}$ be a estimator of g based on $\{X_j\}_{j=1}^n$. Moreover, $\{x_j\}_{j=1}^n$ are observed sample and $y_j = g(x_j)$, $\hat{y}_j = \hat{g}(x_j)$ for $j \in [1, n]$. Denote $S := \{x_j, y_j\}_{j=1}^n$ and $\hat{S} := \{x_j, \hat{y}_j\}_{j=1}^n$. Then, provided*

$$\underset{j}{max}|\hat{y}_j - y_j)| = o_p(1),$$

we have

$$|HSIC(\hat{S}) - HSIC(S)| = o_p(1)$$

where HSIC is defined with the kernels $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ and $l(u, v)$ is function of $u - v$, which is Lipschitz with constant C_l and $k(\cdot,\cdot)$ is bounded by one.

Proof. Since x_i's are the same in both datasets $\hat{S}$ and S we have

$$HSIC(\hat{S}) = (n-1)^{-2} tr(KH\hat{L}H) \text{ and } HSIC(S) = (n-1)^{-2} tr(KHLH),$$

where H is the same as defined in the definition of HSIC, $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$ and $\hat{L}_{ij} = l(\hat{y}_i, \hat{y}_j)$. Denoting $A := HKH$, we obtain

$$\begin{aligned} A = HKH &= (I - n^{-1}\mathbf{1}\mathbf{1}^T)K(I - n^{-1}\mathbf{1}\mathbf{1}^T) = (I - n^{-1}\mathbf{1}\mathbf{1}^T)(K - n^{-1}K\mathbf{1}\mathbf{1}^T) \\ &= K - \frac{2}{n}K\mathbf{1}\mathbf{1}^T + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T K\mathbf{1}\mathbf{1}^T = K - \frac{2}{n}K\mathbf{1}\mathbf{1}^T + \frac{1}{n^2}\mathbf{1}^T K\mathbf{1}\mathbf{1}\mathbf{1}^T. \end{aligned}$$

So,

$$\begin{aligned} |A_{i,j}| = |K_{ij} - \frac{2}{n}\sum_{t=1}^n K_{it} + \frac{1}{n^2}\sum_{t,s=1}^n K_{ts}| &\leq |K_{ij}| + \frac{2}{n}\sum_{t=1}^n |K_{it}| + \frac{1}{n^2}\sum_{t,s=1}^n |K_{ts}| \\ &\leq 1 + 2 + 1 = 4, \end{aligned}$$

where the first inequality follows from the triangle inequality of the absolute value and the second one from the fact that kernel $k(\cdot,\cdot)$ is bounded by one. So, the definition of HSIC gives

$$|HSIC(\hat{S}) - HSIC(S)| \overset{(E.1)}{=} |(n-1)^{-2} tr(KH\hat{L}H) - (n-1)^{-2} tr(KHLH)|$$

$$
\begin{aligned}
&\overset{(E.2)}{=} \frac{1}{(n-1)^2}|tr(HKH\hat{L}) - tr(HKHL)| \\
&\overset{(E.3)}{=} \frac{1}{(n-1)^2}|tr(HKH(\hat{L}-L)| = \frac{1}{(n-1)^2}|tr(A(\hat{L}-L)| \\
&\overset{(E.4)}{=} \frac{1}{(n-1)^2}|\sum_{i,j=1}^{n} A_{ij}(\hat{L}_{ij} - L_{ij})| \\
&\overset{(I.1)}{\leq} \frac{1}{(n-1)^2}\sum_{i,j=1}^{n} |A_{ij}| \cdot |(\hat{L}_{ij} - L_{ij})| \\
&\overset{(I.2)}{\leq} \frac{4}{(n-1)^2}\sum_{i,j=1}^{n} |(\hat{L}_{ij} - L_{ij})| \\
&\overset{(E.5)}{=} \frac{4}{(n-1)^2}\sum_{i\neq j} |(\hat{L}_{ij} - L_{ij})| \\
&\overset{(E.6)}{=} \frac{4}{(n-1)^2}\sum_{i\neq j} |l(\hat{y}_i,\hat{y}_j) - l(y_i,y_j)| \\
&\overset{(I.3)}{\leq} \frac{4C_l}{(n-1)^2}\sum_{i\neq j} |\hat{y}_i - \hat{y}_j - (y_i - y_j)| \\
&\overset{(I.4)}{\leq} \frac{4C_l}{(n-1)^2}\sum_{i\neq j} [|\hat{y}_i - y_i)| + |\hat{y}_j - y_j)|] \\
&\overset{(E.7)}{=} \frac{8C_l(n-1)}{(n-1)^2}\sum_{j=1}^{n} |\hat{y}_j - y_j)| \overset{(E.8)}{=} \frac{8C_l}{n-1}\sum_{j=1}^{n} |\hat{y}_j - y_j)| \\
&\overset{(I.5)}{\leq} \frac{8C_l n}{n-1} \cdot \underset{j}{max}|\hat{y}_j - y_j)| \xrightarrow{p} 0,
\end{aligned}
$$

where $(E.1)$ follows from the definition of HSIC, $(E.2)$ is the positive homogeneity of absolute value and the fact that $tr(AB) = tr(BA)$ for any square matrices A and B. $(E.3)$ is the linearity of trace operator. Recalling the property of the trace: $tr(AB) = \sum_{i,j} A_{ij}B_{ij}$ (or just calculating it), gives $(E.4)$. Triangle inequality of absolute value gives $(I.1)$ and $(I.4)$. The above bound on A_{ij}, gives $(I.2)$ and $(E.5)$ follows form the fact that $\hat{L}_{i,j} = L_{ij} = 1$. $(E.6)$ is just the definition of matrices $\hat{L}$ and L. For $(I.3)$ we used that kernel $l(u,v)$ is a Lipschitz function of $u - v$ with the constant C_l. $(E.7)$ and $(E.8)$ are simple calculations. In $I.5$ we just replaced all the items in the summation by the maximum of them and final convergence follows from the assumptions. This concludes the proof. □

The following is the proof of Proposition 4.2.

Proof. Let us fix some $k \in [1.m]$ and consider two cases:

Case 1: $X^{(k)}$ is a sink node.

From the assumption **(A3)** we have that there is strictly increasing function $f_{k,2}$

and a vector β_k such that

$$f_{k,2}^{-1}(X^{(k)}) = \mathbf{PA}_k^T\beta_k + \varepsilon_k,$$

where $\mathbf{PA}_k$ are the parents of $X^{(k)}$ in $\mathcal{G}^0$ and ε_k is standard normal distributed. Note that this is exactly the Linear Transformation Model (5). Since, **(A2)** gives that the assumptions of Theorems 2.2 and 2.3 are satisfied we have estimators $\hat{h}$ and $\hat{\beta}$ satisfy the conditions of Lemma 2.2 satisfied, which gives for

$$\hat{\varepsilon}_j^{(k)} = \hat{h}(v_j^{(k)}) - (v_j^{(-k)})^T\hat{\beta} \text{ for } j \in [1, T]$$

we have

$$\hat{\varepsilon}_j^{(k)} - \epsilon_j^k = o_p(1) \text{ for } j \in [1, T],$$

where $\epsilon_j^k = f_{k,2}^{-1}(X_j^{(-k)}) - \mathbf{PA}_k^T\beta_j$. Since, T is fixed the above is the same as

$$\underset{j}{max}|\hat{\varepsilon}_j^{(k)} - \epsilon_j^k| = o_p(1). \tag{50}$$

Note that having T fixed is important here to be able to argue that the above holds.

The equation (50) gives that the assumptions of Lemma A.6 are satisfied and we have

$$HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T) = HSIC(\{v_j^{(-k)}, \epsilon_j^k\}_{j=1}^T) + o_p(1). \tag{51}$$

On the other hand since $X^{(k)}$ is a sink node the PNLG model gives that noise is independent form the causes, so $HSIC(\mathbb{P}^{\varepsilon^k, X^{(-k)}}) = 0$, where $(X^{(-k)}, X^{(k)})$ is distributed according to the PNLG model and ε^k is the noise variable corresponding to the node $X^{(k)}$). Thus, $HSIC(\mathbb{P}^{\varepsilon^k, X^{(-k)}}) = 0$ and Theorem 4.2 gives, with probability at least $1 - \delta/m$ we have

$$\begin{aligned}|HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T)| &= |HSIC(\mathbb{P}^{\varepsilon^k, X^{(-k)}}) - HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T)| \\ &< \sqrt{\frac{\log(6m/\delta)}{\alpha^2 T}} + \frac{C}{T} < \frac{\xi}{2},\end{aligned}$$

where the last inequality is the assumption of the Theorem. Combining the above with (51) gives with probability at least $1 - \delta/m$ we have

$$t_k := HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T) < \xi/2 \text{ as } n \to \infty.$$

Case 2: $X^{(k)}$ is not a sink node.
In **(A1)** putting $A := [1, m] \cap \{k\}$ we will have $X^{(A)}$ contains a child of $X^{(k)}$ as $X^{(k)}$ is not a sink node.

Let $\hat{h}$, $\hat{\beta}$, $X^{(k)}$, $X^{(-k)}$ and $\hat{\varepsilon}_j^{(k)}$ be defined similar to the previous case and define $N := \hat{h}(X^{(k)}) - (X^{(-k)})^T\hat{\beta}$. So, **(A1)** gives

$$HSIC(\mathbb{P}^{N,X^{(-k)}}) > \xi. \tag{52}$$

On the other hand Theorem 4.2 gives with probability at least $1 - \delta/m$ we have

$$|HSIC(\mathbb{P}^{N,X^{(-k)}}) - HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T)| \leq \sqrt{\frac{\log(6m/\delta)}{\alpha^2 T}} + \frac{C}{T} < \frac{\xi}{2},$$

where the last inequality follows from the assumption of Theorem. Combining the above with (52) and using the triangle inequality gives with probability at least $1 - \delta/m$ we have

$$\begin{aligned} t_k &:= HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T) \\ &\geq HSIC(\mathbb{P}^{N,X^{(-k)}}) - |HSIC(\mathbb{P}^{N,X^{(-k)}}) - HSIC(\{v_j^{(-k)}, \hat{\varepsilon}_j^{(k)}\}_{j=1}^T)| \\ &> \xi - \xi/2 = \xi/2 \text{ as } n \to \infty. \end{aligned}$$

Now let us combine the above two cases. They give us that with probability at least $1 - \delta/m$ and sample size $n \to \infty$, we have $t_k < \xi/2$ in case of $X^{(k)}$ is a sink node and $t_k > \xi/2$ in case of $X^{(k)}$ is not a sink node. Thus, we conclude for

$$\hat{\pi}(m) = \underset{k}{argmin}\{t_k\},$$

is a consistent sink node with probability at least $1 - \delta/m$, that is

$$\mathbb{P}(\hat{\pi}(m) \text{ is a sink node}) \geq 1 - \delta/m \text{ as } n \to \infty.$$

□

References

Jason Abrevaya. Computation of the maximum rank correlation estimator. *Economics Letters*, 62(3):279–285, 1999a.

Jason Abrevaya. Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *Journal of Econometrics*, 93(2):203–228, 1999b.

Jason Abrevaya. Pairwise-difference rank estimation of the transformation model. *Journal of Business & Economic Statistics*, 21(3):437–447, 2003.

Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An Empirical Distribution Function for Sampling with Incomplete Information. *The Annals of Mathematical Statistics*, 26(4):641 – 647, 1955.

B. M. Bennett. Rank-order tests of linear hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(3):483–489, 1968.

Thomas B. Berrett, Ioannis Kontoyiannis, and Richard J. Samworth. Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics*, 49(5):2457 – 2490, 2021.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014.

Christopher L. Cavanagh and Robert P. Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84:351–381, 1998.

Songnian Chen. Rank estimation of transformation models. *Econometrica*, 70(4): 1683–1697, 2002.

Pierre-André Chiappori, Ivana Komunjer, and Dennis Kristensen. Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22–39, 2015.

David Clayton and Jack Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)*, 148 (2):82–117, 1985.

David Clayton and Jack Cuzick. The semiparametric pareto model for regression analysis of survival times. *Papers on Semiparametric Models*, pages 19–31, 1986.

Benjamin Colling and Ingrid Van Keilegom. Estimation of fully nonparametric transformation models. *Bernoulli*, 25(4B):3762 – 3795, 2019.

D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 08 1975.

Jack Cuzick. Rank regression. *The Annals of Statistics*, 16(4):1369–1389, 1988.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

Kjell A. Doksum. An Extension of Partial Likelihood Methods for Proportional Hazard Models to General Transformation Models. *The Annals of Statistics*, 15 (1):325 – 345, 1987.

C Glymour, K Zhang, , and P Spirtes. Review of causal discovery methods based on graphical models. *Front. Genet.*, 2019.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, page 63–77, Berlin, Heidelberg, 2005. Springer-Verlag.

Aaron K. Han. Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35(2):303–316, 1987.

Joel L. Horowitz. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, 64(1):103–137, 1996.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 06 1938.

Alessio Moneta, Doris Entner, Patrik O. Hoyer, and Alex Coad. Causal Inference by Independent Component Analysis: Theory and Applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, October 2013.

Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763–765, 1973.

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*, 2007.

J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. pages 589–598, Corvallis, OR, USA, July 2011. AUAI Press.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

A. N. Pettitt. Proportional odds models for survival data and estimates using ranks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(2):169–175, 1984.

A. N. Pettitt. Estiamtes for a regression parameter using ranks. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):58–67, 1987.

Anthony N. Pettitt. Inference for the linear model using a likelihood based on ranks. *Journal of the royal statistical society series b-methodological*, 44:234–243, 1982.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, 1980.

Robert P. Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61(1):123–137, 1993.

Hongjian Shi, Mathias Drton, and Fang Han. On the power of Chatterjee rank correlation. *arXiv e-prints*, art. arXiv:2008.11619, August 2020.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 2016.

Kento Uemura and Shohei Shimizu. Estimation of post-nonlinear causal models using autoencoding structure. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3312–3316, 2020.

Kento Uemura, Takuya Takagi, Kambayashi Takayuki, Hiroyuki Yoshida, and Shohei Shimizu. A multivariate causal discovery based on post-nonlinear model. In *First Conference on Causal Learning and Reasoning*, 2022.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Tao Yu, Pengfei Li, Baojiang Chen, Ao Yuan, and Jing Qin. Maximum pairwise-rank-likelihood-based inference for the semiparametric transformation model. *arXiv e-prints*, art. arXiv:2103.13435, March 2021.

Junyi Zhang. Estimation and testing methods for monotone transformation models. PhD thesis, Columbia University, 2013.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. UAI'09, page 647–655, 2009.