Journal of Machine Learning Research 25 (2024) 1-49 Submitted 9/23; Revised 8/24; Published 10/24

Causal Discovery with Generalized Linear Models through Peeling Algorithms

Minjie Wang MWANG46@BINGHAMTON.EDU
Department of Mathematics and Statistics
Binghamton University, State University of New York
Binghamton, NY 13902, USA

Xiaotong Shen XSHEN@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Wei Pan PANXX014@UMN.EDU
Division of Biostatistics
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Jin Tian

Abstract

This article presents a novel method for causal discovery with generalized structural equation models suited for analyzing diverse types of outcomes, including discrete, continuous, and mixed data. Causal discovery often faces challenges due to unmeasured confounders that hinder the identification of causal relationships. The proposed approach addresses this issue by developing two peeling algorithms (bottom-up and top-down) to ascertain causal relationships and valid instruments. This approach first reconstructs a super-graph to represent ancestral relationships between variables, using a peeling algorithm based on nodewise GLM regressions that exploit relationships between primary and instrumental variables. Then, it estimates parent-child effects from the ancestral relationships using another peeling algorithm while deconfounding a child's model with information borrowed from its parents' models. The article offers a theoretical analysis of the proposed approach, establishing conditions for model identifiability and providing statistical guarantees for accurately discovering parent-child relationships via the peeling algorithms. Furthermore, the article presents numerical experiments showcasing the effectiveness of our approach in comparison to state-of-the-art structure learning methods without confounders. Lastly, it demonstrates an application to Alzheimer's disease (AD), highlighting the method's utility in constructing gene-to-gene and gene-to-disease regulatory networks involving Single Nucleotide Polymorphisms (SNPs) for healthy and AD subjects.

Keywords: Generalized linear models, large directed acyclic graphs, hierarchy, nonconvex minimization, mixed graphical models

1. Introduction

Discovering causal relationships among variables is crucial for scientific inquiries in various fields, including genetics, artificial intelligence, and social science. For instance, in genetics, biologists aim to uncover gene-gene regulatory relationships, while neuroscientists focus on causal influences between different regions of interest in a patient's brain. However, unmeasured confounders can arise when randomized experiments are unethical or infeasible, which distort the discovery process and obscure the relationship between exposures and the outcome variable, leading to false discoveries. Consider our motivating case study on inferring regulatory networks from the Alzheimer's disease

gene expression data. We study a subset of genes while other genes are not included and removed by the prescreening procedure, which introduces unmeasured confounders. Meanwhile, in neuroscience, existing technologies can only record from a small subset of neurons in the brain at once, also leading to confounders. This article proposes a novel approach to causal discovery using instrumental variables to correct confounding effects, yielding accurate causal discovery, particularly for discrete outcomes such as binary, count-valued, and multinomial.

Causal discovery necessitates estimating parent-child relationships, or equivalently, the graph structure of a directed acyclic graph (DAG). DAGs are an effective tool for describing directional effects in causal discovery, but reconstructing a DAG structure poses computational challenges due to the acyclicity constraint. Two popular approaches for reconstructing a Gaussian DAG structure without confounders are the sequential conditional independence tests, such as the PC algorithm (Spirtes et al. 2000), and the likelihood-based methods subject to the acyclicity constraint (Zheng et al. 2018; Yuan et al. 2019). Recently, Li et al. (2023) proposed a linear causal discovery method without confounders through interventions. Going beyond, for non-Gaussian outcomes, Zheng et al. (2020) extended the algebraic characterization of DAGs by Zheng et al. (2018) to nonparametric and semiparametric models including GLMs; Shi et al. (2023) proposed a new hypothesis testing method for nonlinear DAG models. However, despite recent work, causal discovery for discrete outcome data, particularly in the presence of confounders, has received limited attention, and unique challenges arise when handling such data. One challenge is the non-identifiability of the logistic DAG model, even without confounders (Park and Raskutti 2018). Moreover, in the presence of confounders, unmeasured confounders can distort causal effect estimation, making structural equation models non-identifiable. Another challenge is the typically intractable form of the marginal likelihood, despite an interpretable conditional likelihood and data-specific noise or variance. It also remains unclear how to separate confounders from causal effects in the discovery process. Some recent proposals focus on simple situations, such as the two-stage least squares (Theil 1992), an instrumental variable (IV) regression of continuous outcomes given a known causal order, the two-stage predictor substitution (2SPS, Cai et al. (2011)) and two-stage residual inclusion (2SRI, Hausman (1978); Terza et al. (2008)) for general nonlinear outcomes, including discrete outcome data. However, none of these approaches apply to causal discovery with an unknown causal order and multiple primary variables.

This article proposes a new approach called GAMPI (Generalized Linear Models with Peeling and Instruments) for causal discovery of multiple primary variables from various data types. GAMPI involves a two-step process. First, we propose a fidelity model as a simple surrogate for the original intractable marginal model, which retains intervention characteristics. Then, we design a bottom-up peeling algorithm to reconstruct the super-graph consisting of ancestral relationships while identifying valid instrumental variables (IVs) for each primary variable by exploiting the connections between the primary and instrumental variables to determine the causal order. For each primary variable, a constrained generalized linear model (GLM, Nelder and Wedderburn (1972)) subject to the truncated ℓ_1-penalty constraint (TLP, Shen et al. (2012)) is fit on the instrumental variables to identify nonzero-coefficient IVs, followed by a difference-of-convex (DC) algorithm to solve the corresponding nonconvex minimization. In the second step, given the identified super-graph, we develop a top-down peeling algorithm to estimate the direct causal effects of each primary variable while identifying its parents from ancestors. In this peeling process, we propose a novel deconfounding approach using the estimated confounders from the parents' equation models to correct the confounding effects of a child's equation model. This approach fits a TLP-constrained GLM to each primary variable on its ancestors and residuals from its ancestors' models to identify parents and estimate the direct causal effect of each parent-child relationship.

This article contributes to causal discovery. It introduces a comprehensive approach capable of handling diverse data types with unobserved confounders, ensuring the identification of parent-child relationships through valid instruments for each primary variable. This involves generalized linear models, addressing both discrete and mixed (continuous and discrete) outcomes while considering confounders beyond Gaussian data without confounders by Li et al. (2023). In particular,

(1) It establishes the identifiability of generalized structural equation models with confounders and instruments, valid and invalid. This result does not require additional assumptions for each primary variable with a nonlinear link, unlike the Gaussian case which requires valid instrumental variables to be the majority of the instrumental variables (Kang et al. 2016; Windmeijer et al. 2019).

(2) It introduces a fidelity model to handle intractable likelihoods and eliminate the confounding effects for identifying ancestral relationships.

(3) It designs a projection-based difference-convex (DC) algorithm to solve nonconvex minimization for a constrained generalized linear model regression. This algorithm delivers a global minimizer with high probability and a computational complexity of $q^2 \max(q, n) \log K^0$, where q, n, and K^0 are the numbers of regressors, the sample size, and the nonzero regression coefficients.

(4) It develops bottom-up and top-down peeling algorithms to estimate the causal order and the causal effects for primary variables. These algorithms require solving at most p generalized linear model regressions subject to the truncated ℓ_1-penalty constraint, where p is the number of primary variables.

(5) It shows that GAMPI yields the correct discovery of all parent-child relationships, providing statistical guarantees for GAMPI.

(6) It demonstrates the superior performance of GAMPI for logistic and Poisson models over state-of-the-art methods, NOTEARS (Zheng et al. 2018, 2020) and a faster version of NOTEARS, called DAGMA (Bello et al. 2022), especially in the presence of confounders. It suggests that GAMPI corrects the confounding effects without imposing additional noise variance structures to reconstruct a causal graph.

The rest of the article is structured as follows. Section 2 introduces generalized structural equation models with confounders and instruments. Section 3 introduces the fidelity model and three algorithms, one DC and two peeling algorithms, for identifying the ancestral and then parent-child relationships. Section 4 investigates the statistical properties of the proposed approach. Section 5 performs simulation studies, followed by Section 6 with an application to Alzheimer's disease to reconstruct a gene-to-gene and gene-to-disease regulatory network. Section 7 concludes the article. The Appendix contains illustrative examples, technical proofs, and additional simulations.

2. Generalized Structural Mean Models

2.1 Directed Acyclic Graphs, Confounders, and Interventions

Given a vector of primary variables of interest $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$, the joint probability of a generalized structural equation model (SEM, Pearl (2000)) with confounders $\boldsymbol{h} = (h_1, \ldots, h_p)^\top$ and instrumental variables $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$ can be factorized as:

$$\mathbb{P}\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{h}\right) = \prod_{j=1}^{p} \mathbb{P}\left(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j\right), \tag{1}$$

where $\mathbb{P}\left(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j\right)$ denotes the conditional probability of Y_j given $\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j$, which follows an exponential family distribution. Here, unmeasured confounders refer to variables that are not included in the model, but nonetheless affect the primary variables of interest. Confounders h_j and $h_{j'}$ can be correlated among equations for $j \neq j'$. An instrumental variable (IV) is a variable that affects the primary variables of interest, but not vice versa, i.e., the primary variable should not have an impact on the IVs. In practice, one may choose candidate IV sets based on scientific knowledge, as in Section 6. Note that (1) characterizes a DAG under the acyclicity constraint. Moreover, the conditional distribution of Y_j is characterized by a generalized linear model:

$$\psi_j(\mathbb{E}\left[Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j\right]) = \boldsymbol{U}_{\mathrm{pa}(j),j}^\top \boldsymbol{Y}_{\mathrm{pa}(j)} + \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + h_j, \quad j = 1, \ldots, p, \tag{2}$$

where $\psi_j(\cdot)$ is a monotone link function for a GLM chosen to be appropriate for the data type of Y_j (cf. Table 1), $\mathrm{pa}(j) \equiv \{k : u_{kj} \neq 0\} = \{k : Y_k \to Y_j\}$ denotes a set of parent variables of Y_j, defined by the parent-child relationship $Y_k \to Y_j$, $\mathrm{in}(j) \equiv \{l : w_{lj} \neq 0\} = \{l : X_l \to Y_j\}$ denotes a set of the associated instrumental variables of Y_j, defined by an intervention from X_l to Y_j: $X_l \to Y_j$, and $\boldsymbol{Y}_A = (\boldsymbol{Y}_{k_1}, \cdots, \boldsymbol{Y}_{k_M})^\top$, $k_m \in A$, is a sub-vector of $\boldsymbol{Y}$ indexed by A. Here, $\boldsymbol{U} = (u_{kj})$ and $\boldsymbol{W} = (w_{lj})$ are the $p \times p$ adjacency and $q \times p$ intervention matrices, and $\boldsymbol{U}_{\mathrm{pa}(j),j} = (u_{kj})_{k \in \mathrm{pa}(j)}$ and $\boldsymbol{W}_{\mathrm{in}(j),j} = (w_{lj})_{l \in \mathrm{in}(j)}$ are sub-vectors of the jth column vector of $\boldsymbol{U}$, $\boldsymbol{U}_{\bullet j} = (u_{kj})$ and the jth column vector of $\boldsymbol{W}$, $\boldsymbol{W}_{\bullet j} = (w_{lj})$. $^\top$ denotes the transpose. Note that the p structural equations can possess different ψ_js, depending on the data type of Y_j, reminiscent of the mixed graphical models framework (Yang et al. 2014). We refer the reader to Section 6 for an illustrative example.

The adjacency matrix $\boldsymbol{U}$ specifies a directed acyclic graph (DAG) with each primary variable as a node, and its non-zero elements represent directed edges between nodes. To prevent directed cycles, $\boldsymbol{U}$ is subject to the acyclicity constraint (Zheng et al. 2018; Yuan et al. 2019).

Table 1: Examples of distributions in generalized linear models

Distribution	Support	Link	Density
Bernoulli, $Bern(\mu)$	Integer: $\{0, 1\}$	$\psi_j(\mu) = \ln(\frac{\mu}{1-\mu})$	$\mu^y(1-\mu)^{1-y}$
Binomial, $Bin(N, \mu)$	Integer: $0, \ldots, N$	$\psi_j(\mu) = \ln(\frac{\mu}{1-\mu})$	$\binom{N}{y}\mu^y(1-\mu)^{N-y}$
Gaussian, $N(\mu, \sigma^2)$	Real: $(-\infty, \infty)$	$\psi_j(\mu) = \mu$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-\mu)^2}{2\sigma^2})$
Poisson, $Poisson(\mu)$	Integer: $0, 1, \ldots$	$\psi_j(\mu) = \ln \mu$	$\frac{\mu^y \exp(-\mu)}{y!}$
Multinomial, $Multi(\mu_1, \ldots, \mu_K)$	K-vector of integer: $[0, \ldots, N]$	$\psi_j(\mu) = \ln(\frac{\mu}{1-\mu})$	$\frac{n!}{y_1! \ldots y_K!} \prod_{k=1}^K \mu_k^{y_k}$

2.2 Identifiability

Model (2) encodes a DAG model describing multiple parent-child relationships, which, however, is generally not identifiable in the presence of unmeasured confounders $\boldsymbol{h}$. Note that (2) may not be identifiable even in the absence of confounders $\boldsymbol{h}$, for instance, a logistic model without instrumental variables and confounders (Park and Raskutti 2018). However, as suggested by Proposition 1, with suitable instruments, (2) is identifiable.

To proceed, we first categorize instrumental variables (IVs) into valid IVs and non-valid IVs (covariates). A valid instrument X_l for primary variable Y_j satisfies:

(i) Relevance: it intervenes on Y_j;

(ii) Exclusion: it does not intervene on other primary variables.

Otherwise, it is a non-valid IV that intervenes on none or multiple primary variables. Let $\mathrm{in}_*(j)$ denote a set of valid IV of Y_j. Next, we make some assumptions on instruments for model (2).

Assumption 1 *Assume that for $j = 1, \ldots, p$, model (2) satisfies:*

(A) (Local faithfulness) $\mathrm{Cov}(Y_j, X_l|\boldsymbol{X}_{\{1,\cdots,q\}\setminus\{l\}}) \neq 0$ *when X_l intervenes on an immediate parent of Y_j, where* Cov *denotes the covariance.*

(B) (Instrumental sufficiency) Each primary variable is intervened by at least one valid IV. If ψ_j is linear, then the number of valid IVs for Y_j is required to exceed 50% of its total number of IVs, known as the majority rule. Otherwise, the majority rule is not required for a specific nonlinear ψ_j.

(C) (Validity) Confounders $\boldsymbol{h} = (h_1, \ldots, h_p)^\top$ and valid instruments $\boldsymbol{X}_{in_*} = (X_{in_*(1)}, \ldots, X_{in_*(p)})^\top$ are independent. That is, for each pair of $\{(l, j), l \in \cup_{j=1}^p in_*(j)\}$, X_l and h_j are independent.

Assumption 1(A) guarantees that other interventions don't offset an intervention from X_l to Y_j, while Assumption 1(B) ensures that each primary variable has at least one valid IV. Both are necessary for the identifiability of a Gaussian structural model (Li et al. 2023). The second condition in Assumption 1(B) requires the majority rule for a linear link, which amounts to the so-called majority requirement for Gaussian data (Kang et al. 2016; Windmeijer et al. 2019). However, such a majority condition is not required for a nonlinear link function. We provide an illustrative example of the majority rule in Appendix A.2. Note Assumption 1(B) considers a GLM with the canonical link as well as the non-canonical link, defined by model (2). Assumption 1(C) is also required by the two-stage least squares and residual inclusion methods for the IVs (Lousdal 2018; Terza et al. 2008), known as the instrumental validity assumption. Further, the instrumental variables $\boldsymbol{X}$ and confounders are independent by parameterization, i.e., projecting h_j onto the space spanned by the non-valid IVs. Given Assumption 1, Proposition 1 suggests the identifiability of model (2).

Proposition 1 (Identifiability) *Under Assumption 1, model* (2) *is identifiable for model parameters* $(\boldsymbol{U}, \boldsymbol{W})$.

Proposition 1 suggests that a nonlinear link function permits the identification of the parents of a primary variable, which is unlike the linear link for Gaussian data. This new result highlights the importance of a link function concerning the model identifiability of causal effects.

3. Method

This section estimates $(\boldsymbol{U}, \boldsymbol{W})$ to identify parent-child relationships and the corresponding interventions in (2). Due to the model identifiability issue of (2), direct estimation of $\boldsymbol{U}$ is impossible without the help of instrumental variables $\boldsymbol{X}$. To estimate parent sets $\mathrm{pa}(j)$, $j = 1, \ldots, p$, and thus $\boldsymbol{U}$, we first need to determine the causal order, which amounts to determining ancestral relationships, including all parent-child relationships. Here, Y_k is an ancestor of Y_j, or Y_j is an offspring of Y_k, denoted by $Y_k \rightsquigarrow Y_j$, if there exists a directed pathway $Y_k \to Y_{k_1} \to \ldots \to Y_{k_m} \to Y_j$, where $Y_k \to Y_{k_1}$ is a parent-child relationship defined by $\boldsymbol{U}$. Subsequently, $\mathrm{an}(j)$ denotes a set of ancestors of Y_j. Once $\mathrm{an}(j)$ is identified, we then pinpoint $\mathrm{pa}(j)$ via a deconfounding approach in Section 3.3.

3.1 Fidelity Models

This subsection introduces a working model termed as the "fidelity model", to identify all ancestral relationships. The term "fidelity model" is named as it yields the same support as the marginal distribution of the original model. Towards this end, we exploit the connections between a primary variable and the associated instrumental variables, described by the conditional distribution of Y_j given $\boldsymbol{X}$ from (2), $\mathbb{P}(Y_j|\boldsymbol{X})$, to identify the causal orders among primary variables. However, $\mathbb{P}(Y_j|\boldsymbol{X})$ is generally intractable even given an analytic expression of $\mathbb{P}(Y_j|\boldsymbol{Y}_{\mathrm{pa(j)}}, \boldsymbol{X}, h_j)$ in (2). To overcome this difficulty, we introduce the fidelity model that is also a GLM:

$$\psi_j(\mathbb{E}(Y_j|\boldsymbol{X})) = \boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}, \quad j = 1, \cdots, p, \tag{3}$$

where ψ_j is set to be the same as in (2). Here, $\boldsymbol{V}_{\bullet j} = (V_{1j}, \ldots, V_{qj})$ is the jth column vector of a $q \times p$ matrix $\boldsymbol{V} = (\boldsymbol{V}_{\bullet 1}, \ldots, \boldsymbol{V}_{\bullet p})$. This model (3) is motivated by the observation that the conditional

distribution of Y_j given $\boldsymbol{X}$, denoted by $\mathbb{P}^*(Y_j|\boldsymbol{X})$ and defined by (3), satisfies $\frac{\partial \mathbb{P}^*(Y_j|\boldsymbol{X})}{\partial X_m} \neq 0$ if and only if $\frac{\partial \mathbb{P}(Y_j|\boldsymbol{X})}{\partial X_m} \neq 0$ based on (2) due to the properties of GLMs, as shown in Proposition 2, where $\frac{\partial}{\partial X_m}$ denotes the partial derivative with respect to X_m.

The conditional distribution $\mathbb{P}^*(Y_j \mid \boldsymbol{X})$ defined by the fidelity model (3) not only provides a simple form to work with, but also has the same support as the intractable marginal distribution $\mathbb{P}(Y_j|\boldsymbol{X})$ under (2), although with different intervention magnitudes. In particular, a nonzero l-th element of $\boldsymbol{V}_{\bullet j}$ indicates that Y_k is an ancestor of Y_j if X_l is a valid IV of Y_k. This property permits the identification of the super-graph characterizing all the ancestral relationships, as shown in Proposition 3.

We define the index set of $X_1, \cdots, X_q$ with nonzero coefficients in the fidelity model (3) and in the true model $\mathbb{P}(Y_j|\boldsymbol{X})$ marginalized from (2) as $S_j = \{m : V_{mj} \neq 0\}$ and $\widetilde{S}_j = \{m : \frac{\partial \mathbb{P}(Y_j|\boldsymbol{X})}{\partial X_m} \neq 0\}$, respectively, for $j = 1, \ldots, p$. The following Proposition 2 establishes the connections between the fidelity model and the marginal distribution of the true model $\mathbb{P}(Y_j|\boldsymbol{X})$.

Proposition 2 (Support preservation) *Assume that Assumption 1 is satisfied and the link function ψ_js in* (2) *are differentiable. Then, $\mathbb{P}^*(Y_j|\boldsymbol{X})$ defined by the fidelity model* (3) *has the same support as $\mathbb{P}(Y_j|\boldsymbol{X})$ under the full model* (2)*, that is, $S_j = \widetilde{S}_j$, $j = 1, \ldots, p$.*

Proposition 2 suggests that the fidelity model (3) retains the intervention structure of $\mathbb{P}(Y_j \mid \boldsymbol{X})$ in the original model concerning the presence or absence of a specific intervention. It is worth mentioning that the fidelity model (3) eliminates the confounding effects when identifying the support of $\mathbb{P}(Y_j \mid \boldsymbol{X})$ and hence the ancestral relationships or the causal order among $Y_1, \ldots, Y_p$. This property is due to Assumption 1(C) that $\boldsymbol{X}$ are independent of confounders $\boldsymbol{h}$. Consequently, the confounders are marginalized for $\boldsymbol{X}$ and thus have no impact on the support of $\mathbb{P}(Y_j \mid \boldsymbol{X})$.

To illustrate the fidelity model and Proposition 2, we here include a motivating example. Consider a generalized structural equation model for binary outcomes with $p = q = 5$:

$$\begin{aligned}
&\psi(\mathbb{E}[Y_1|X_1, h_1]) = 2X_1 + h_1, && \psi(\mathbb{E}[Y_2|Y_1, X_2, h_2]) = 1.5Y_1 + 2X_2 + h_2,\\
&\psi(\mathbb{E}[Y_3|Y_2, X_3, h_3]) = 1.5Y_2 + 2X_3 + h_3, && \psi(\mathbb{E}[Y_4|Y_3, Y_1, X_4, h_4]) = -1.5Y_1 + 1.5Y_3 + 2X_4 + h_4,\\
&\psi(\mathbb{E}[Y_5|X_5, h_5]) = 2X_5 + h_5, && \qquad (4)
\end{aligned}$$

where $\psi_1 = \cdots = \psi_5 = \psi$ is the logit link function. Here, (4) defines a DAG shown in Figure 1. Note that marginalizing each equation in (4) over $\boldsymbol{Y}_{-j}$ does not lead to closed-form expressions for $Y_j|\boldsymbol{X}$.

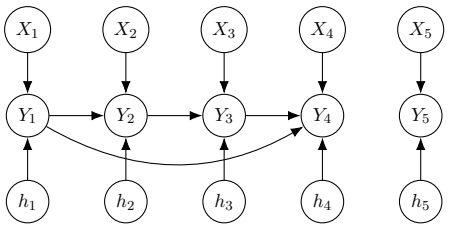

Figure 1: Example DAG defined by model (4).

The proposed fidelity model that has the same support as the marginal model of (4) is:

$$\begin{aligned}
&\psi(\mathbb{E}[Y_1|X_1]) = V_{11}X_1, && \psi(\mathbb{E}[Y_3|X_1, X_2, X_3]) = V_{13}X_1 + V_{23}X_2 + V_{33}X_3,\\
&\psi(\mathbb{E}[Y_2|X_1, X_2]) = V_{12}X_1 + V_{22}X_2, && \psi(\mathbb{E}[Y_4|X_1, X_2, X_3, X_4]) = V_{14}X_1 + V_{24}X_2 + V_{34}X_3 + V_{44}X_4,\\
&\psi(\mathbb{E}[Y_5|X_5]) = V_{55}X_5. &&
\end{aligned}$$

Note that the fidelity model has the same support as the true marginal model and in the next section, we show that ancestral relationships can be identified via $\boldsymbol{V}$.

3.2 Identifying Ancestral Relationships

This subsection proposes a novel structure learning method to identify the ancestral relationships. To start with, we introduce a proposition demonstrating the connections between the primary variables and instrumental variables via $\boldsymbol{V}$ in the fidelity model.

Proposition 3 (Identification of ancestral relationships via $\boldsymbol{V}$) *Assume that Assumption 1 is met. Then,*

(a) For a valid instrument X_l, if $V_{lj} \neq 0$, then X_l intervenes on Y_j or an ancestor of Y_j.

(b) Y_j is a leaf variable with no children if and only if there exists a valid instrument X_l such that $V_{lj} \neq 0$ and $\|V_{l\bullet}\|_0 = 1$.

(c) If $V_{lj} \neq 0$ and X_l is a valid instrument for Y_k, then Y_k is an ancestor of Y_j, that is, $Y_k \rightsquigarrow Y_j$.

Proposition 3 suggests that the topological order of a DAG can be reconstructed by recursively identifying and removing ("peeling off") leaf variables in the graph, as long as the non-zero elements of $\boldsymbol{V}$ are obtained. We define Y_j as a leaf variable if it has no children. Next, we first introduce a nodewise constrained GLM-based approach to estimate the non-zero elements of $\boldsymbol{V}$ and then propose a peeling algorithm to identify the ancestral relationships from $\boldsymbol{V}$ based on Proposition 3.

3.2.1 Nodewise constrained GLM regressions

This subsection proposes nodewise constrained GLM regressions subject to the ℓ_0-constraint based on the fidelity model to estimate nonzero elements of $\boldsymbol{V}$ in (3).

Consider the data matrix $(\boldsymbol{X}_{n\times q}, \boldsymbol{Y}_{n\times p})$ where $\boldsymbol{X}_{i\bullet}$ and $\boldsymbol{Y}_{i\bullet}$ refer to the ith row of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Given independent observations $(\boldsymbol{X}_{i\bullet}, \boldsymbol{Y}_{i\bullet})_{i=1}^n$, let $\mathcal{L}(\boldsymbol{V}_{\bullet j}) = n^{-1}\sum_{i=1}^n \ell(Y_{ij}, \boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet})$ denote the negative log-likelihood for a GLM, where $\ell(Y_{ij}, \boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet})$ is the negative log-likelihood for Y_{ij} given $\boldsymbol{X}_{i\bullet}$; refer to Table 1 and (12) for details. For example, $\ell(Y_{ij}, \boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}) = \left(-Y_{ij}\left(\boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}\right) + \log(1+\exp(\boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}))\right)$ for a logistic model.

For $j = 1, \ldots, p$, the nodewise constrained GLM regression solves the following minimization with a nonconvex constraint:

$$\widehat{\boldsymbol{V}}_{\bullet j} = \arg\min_{\boldsymbol{V}_{\bullet j}} \ \mathcal{L}(\boldsymbol{V}_{\bullet j}) \quad \text{subject to} \quad \sum_{l=1}^{q} I\left(V_{lj} \neq 0\right) \leq K_j, \tag{5}$$

where $1 \leq K_j \leq q$ is an integer-valued tuning parameter. Note that $K_j \geq 1$ ensures that each variable Y_j receives at least one valid IV, as required by Assumption 1(B). Here, we impose the ℓ_0-constraint to obtain the exact number of non-zeros as opposed to the ℓ_1 version. Note many other penalty functions in the literature induce sparsity such as the ℓ_1-penalty (Tibshirani 1996) and the minimax concave penalty (MCP, Zhang (2010)). However, these penalty functions do not yield the exact number of non-zero coefficients to ensure that each variable Y_j receives at least one valid IV, i.e., $K_j \geq 1$, required by Assumption 1(B).

To solve the nonconvex minimization (5), we propose a projection-based difference-convex (DC) algorithm for efficient computation. The constrained problem is equivalent to solving a penalized version of (5) by adding a penalty term to the objective function. Specifically, we minimize $\mathcal{L}(\boldsymbol{V}_{\bullet j}) + \lambda_j \sum_{l=1}^q I\left(V_{lj} \neq 0\right)$, where $\lambda_j > 0$ is a computational parameter corresponding to the constrained parameter K_j in (5). Next, we replace the ℓ_0-indicator function with its computational surrogate, the truncated ℓ_1-function (TLP) denoted by $J_\tau(\cdot)$, where $J_\tau(z) = \min(|z|/\tau, 1)$, as

suggested by Shen et al. (2012). We decompose J_τ into a difference of two convex functions: $J_\tau(z) = S_1(z) - S_2(z) \equiv |z|/\tau - \max(|z|/\tau - 1, 0)$, to construct an upper approximation of the cost function iteratively. At the t-th iteration, we approximate J_τ by $S_1(z) - S_2(z^{[t-1]}) - \nabla S_2(z^{[t-1]})^\top (z - z^{[t-1]}) = \frac{|z|}{\tau} \cdot I\left(|z^{[t-1]}| \le \tau\right) + 1 - I\left(|z^{[t-1]}| \le \tau\right)$ based on the DC decomposition. Then, we solve the unconstrained minimization problem:

$$\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} = \arg\min_{V_{lj}} \ \mathcal{L}(\boldsymbol{V}_{\bullet j}) + \gamma_j \tau_j \sum_{l=1}^{q} I\left(\left|\widetilde{V}_{lj}^{[t-1]}\right| \le \tau_j\right) |V_{lj}|\,, \tag{6}$$

where $\gamma_j = \lambda_j / \tau_j^2$. The DC algorithm iterates until a stopping criterion is met. In particular, let $\bar{f}(\cdot)$ denote the objective function in (6). The DC algorithm terminates at iteration T when $|\bar{f}(\widetilde{\boldsymbol{V}}_{\bullet j}^{[T]}) - \bar{f}(\widetilde{\boldsymbol{V}}_{\bullet j}^{[T-1]})| \le \epsilon_{\text{tol}}$, where ϵ_{tol} is the tolerance level. Finally, the estimated solution $\widehat{\boldsymbol{V}}_{\bullet j}$ is computed by projecting the penalized solution onto the constraint set $\left\{\|\boldsymbol{V}_{\bullet j}\|_0 \le K_j\right\}$. In this paper, $\|\cdot\|_q$ denotes the ℓ_q-norm of a vector and $\|\boldsymbol{x}\|_0 = \sum_j I(x_j \neq 0)$. In practice, we use either 5-fold cross-validation or the extended Bayesian information criterion (EBIC, Chen and Chen (2008)) to choose (τ_j, K_j). We recommend EBIC due to its computational efficiency and strong empirical performance. Algorithm 1 summarizes the DC algorithm for solving nonconvex minimization (5).

Algorithm 1: DC algorithm for nonconvex minimization (5)

1. **(Initialization)** Specify tuning parameters (τ_j, K_j). Initialize $\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[0]}\|_0 \le K_j$, and choose a sequence of γ_j so that $|C_j| \ge K_j$ in Step 4.
2. **(Relaxation)** Compute the penalized solution $\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]}$ of (6).
3. **(Termination)** Repeat Step 2 until a termination criterion is met. Compute $\widetilde{\boldsymbol{V}}$: $\widetilde{\boldsymbol{V}}_{\bullet j} = \text{argmin}_{V_{\bullet j}} \mathcal{L}(\boldsymbol{V}_{\bullet j})$ with $\boldsymbol{V}_{\bullet j} \in \left(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]}\right)_{t=1}^{T}$, where T is the iteration index at termination.
4. **(Projection)** Let $C_j = \{l : |\widetilde{V}_{lj}| > |\widetilde{V}_{\bullet j}|_{(K_j+1)}\}$, where $|\widetilde{V}_{\bullet j}|_{(K_j+1)}$ is the (K_j+1)th largest absolute value of the coefficients. Set $\widehat{\boldsymbol{V}}_{\bullet j} = \text{argmin}_{\boldsymbol{V}_{\bullet j}} \mathcal{L}(\boldsymbol{V}_{\bullet j})$ subject to $V_{lj} = 0$ for $l \notin C_j$.

Remark: Computing $\widehat{\boldsymbol{V}} = (\widehat{\boldsymbol{V}}_{\bullet 1}, \ldots, \widehat{\boldsymbol{V}}_{\bullet p})$ amounts to applying Algorithm 1 p times. The computational complexity of Algorithm 1 to solve one ℓ_0-constrained regression in (5) is the number of DC iterations multiplied by that of solving a weighted Lasso regression for a GLM, which is $q^2 \max(q, n) \log K_j^0$ (Efron et al. 2004).

3.2.2 Identifying Ancestral Relationships via Peeling

Given the nonzero elements of $\widehat{\boldsymbol{V}}$ obtained by Algorithm 1, we now introduce a bottom-up peeling algorithm to estimate ancestral relationships through the nonzero elements of $\widehat{\boldsymbol{V}}$ using Proposition 3. This algorithm constructs a hierarchy of different layers of primary variables, defined by the causal ordering of the variables. The algorithm begins with leaf variables at the bottom, and proceeds by recursively identifying and peeling off one leaf layer of primary variables along with the associated instrumental variables in the graph. Specifically, at iteration h, based on Proposition 3 (b), the algorithm first identifies all leaf nodes Y_k in the subgraph with $\widehat{V}_{lk}^{[h]} \neq 0$ and instrumental variables X_l such that $\|\widehat{V}_{l\bullet}^{[h]}\|_0 = 1$. In practice, the condition $\|\widehat{V}_{l\bullet}^{[h]}\|_0 = 1$ may not hold due to estimation error. To address this issue, we identify the rows of $\widehat{\boldsymbol{V}}^{[h]}$ with the smallest positive ℓ_0-norm, that is, $\left\{l^* : l^* = \arg\min_{l=1}^{q} \|\widehat{V}_{l\bullet}^{[h]}\|_0, \text{ s.t } \|\widehat{V}_{l\bullet}^{[h]}\|_0 \ge 1\right\}$, followed by identifying the largest absolute value element index $k^* = \arg\max_{k=1}^{p} \left|\widehat{V}_{l^*k}^{[h]}\right|$ of the l^*th row for each l^*. By Proposition 3 (b), $X_{l^*} \to Y_{k^*}$.

Moreover, the algorithm identifies the ancestral relationship $Y_{k^*} \rightsquigarrow Y_j$ if an instrument X_{l^*} for the primary variable Y_{k^*} also satisfies $\widehat{V}_{l^*j} \neq 0$ for a previously peeled off Y_j, according to Proposition 3 (c). The algorithm continues by peeling off all the current leaf-instrument $X_l \to Y_k$ pairs (i.e., removing the lth row and kth column from the current $\widehat{\boldsymbol{V}}^{[h]}$) to focus on the subgraph. This peeling process repeats until all primary variables are removed. The super-graph $\hat{\mathcal{S}}$ contains all the ancestral relationships identified during this process. Lastly, the algorithm computes the causal ordering from the super-graph $\hat{\mathcal{S}}$, which is defined as a linear ordering of the nodes where each node appears before all nodes to which it has edges.

Algorithm 2: Peeling algorithm for identifying all ancestral relationships

1. **(Initialization)** $\widehat{\boldsymbol{V}}^{[1]} = \widehat{\boldsymbol{V}}$ and $\hat{\mathcal{S}} = \emptyset$.
Begin iteration $h = 1, \cdots$: at iteration h,
2. **(Leaf-IV pairs)**

   (a) Identify rows of $\widehat{\boldsymbol{V}}^{[h]}$ with the smallest positive ℓ_0-norm. Store indices of all IVs associated with leaf variables in $A^{[h]} = \left\{ l^* : l^* = \arg\min \|\widehat{V}^{[h]}_{l\bullet}\|_0 \right\}$.

   (b) Identify the largest absolute value element index of the l^*th row for each $l^* \in A^{[h]}$: $B^{[h]}_{l^*} = \left\{ k^* : k^* = \arg\max \left| \widehat{V}^{[h]}_{l^*k} \right| \right\}$. Identify all leaf-IV pairs: $X_{l^*} \to Y_{k^*}$. Let $B^{[h]} = \bigcup_{l^*} B^{[h]}_{l^*}$.

3. **(Ancestral relationships)** Identify ancestral relationships $Y_{k^*} \rightsquigarrow Y_j$ if i) $X_{l^*} \to Y_{k^*}$ for $l^* \in A^{[h]}$ and ii) $\widehat{V}_{l^*j} \neq 0$ where Y_j has been previously removed. Update $\hat{\mathcal{S}} = \hat{\mathcal{S}} \cup \{(k^*, j)\}$.
4. **(Peeling)** Remove leaf variables and associated IVs. Let $\widehat{\boldsymbol{V}}^{[h+1]} = \widehat{\boldsymbol{V}}^{[h]}_{\setminus (A^{[h]}, B^{[h]})}$ where $\widehat{\boldsymbol{V}}^{[h]}_{\setminus (A^{[h]}, B^{[h]})}$ is a submatrix by removing the rows and columns indexed by $A^{[h]}$ and $B^{[h]}$ from $\widehat{\boldsymbol{V}}^{[h]}$.
5. **(Termination)** Let $h \to h + 1$ and repeat steps 2-4 until all Y_j's are removed. Update $\hat{\mathcal{S}} = \hat{\mathcal{S}} \cup \{(k, j) : Y_k \to \cdots \to Y_j \text{ in } \hat{\mathcal{S}}\}$. Compute the causal ordering $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_p)$ from $\hat{\mathcal{S}}$. Return the ancestors and IVs identified for each Y_j, $(\overline{\text{an}}(j), \overline{\text{in}}(j))$.

Algorithm 2 summarizes the peeling process for identifying all ancestral relationships or the causal order among primary variables. We include an illustrative example of the peeling algorithm in Appendix A.1. In Step 3, the peeling algorithm identifies all ancestral relationships via Proposition 3, reconstructing a superset that includes all parent-child relationships. Given the superset, we propose a deconfounding approach to identify parent-child relationships.

3.3 Identifying Parent-Child Relationships via Deconfounding

This subsection identifies parent-child relationships given the estimated ancestral relationships from Algorithm 2.

3.3.1 Deconfounding

Given estimated ancestral relationships from the first stage, we develop a novel deconfounding approach based on residual inclusion, called DRI, to estimate parent-child relationships in the

presence of confounders. From (2),

$$\psi_j\left(\mathbb{E}(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)},\boldsymbol{X},h_j)\right)=\boldsymbol{U}_{\mathrm{pa}(j),j}^{\top}\boldsymbol{Y}_{\mathrm{pa}(j)}+\boldsymbol{W}_{\mathrm{in}(j),j}^{\top}\boldsymbol{X}_{\mathrm{in}(j)}+h_j,\quad j=1,\cdots,p, \tag{7}$$

where $h_1,\ldots,h_p$ may be correlated. When there is no confounder, we could identify parents by fitting a constrained GLM regression of Y_j on its ancestors $\boldsymbol{Y}_{\mathrm{an}(j)}$ and instruments $\boldsymbol{X}_{\mathrm{in}(j)}$. However, in the presence of confounders, unobserved confounders h_j and $\boldsymbol{h}_{\mathrm{pa}(j)}$ can be correlated. Thus Y_j's parent variables $\boldsymbol{Y}_{\mathrm{pa}(j)}$ depend on h_j through $\boldsymbol{h}_{\mathrm{pa}(j)}$, which biases the estimation of $\boldsymbol{U}_{\mathrm{pa}(j),j}$ as h_j is one resource of the model error for the regression of Y_j.

To address the confounding issue, we propose a novel deconfounding approach, DRI, to correct the confounding effects in the child structural equations by treating the residuals from its parent GLM regression as predictors. In this way, this approach utilizes the connections between confounders in a parent and its child equations. To facilitate DRI, we make the practically sensible assumption that the confounders $h_1,\cdots,h_p$ are jointly normal. Assumption 2 simplifies the implementation of DRI and makes it computationally efficient.

Assumption 2 *The confounders $h_1,\cdots,h_p$ are jointly normal with an unknown mean and an unknown covariance.*

Remark: Assumption 2 can be relaxed to the assumption that each confounder can be represented as a linear function of other confounders along with an independent error, $h_j=\sum_k\beta_{kj}h_k+\epsilon_j$. In the literature, most assume one common underlying confounding (i.e., one h across all equations) while we here consider a more general case of $h_1,\cdots,h_p$. For complex problems, Assumption 2 is sensible as the confounder is in fact an ensemble of many confounding effects. Under the dense confounding setting in Figure 3, the confounder for each variable h_j is added up by many independent confounding effects from unobserved variables. Therefore, asymptotics holds and the confounders are jointly normal by the central limit theorem. In practice, many variables are unobserved and each is associated with many primary variables of interest, satisfying the dense confounding setting.

To implement DRI, we estimate the confounding effect h_j using the parent equations for each Y_j based on Assumption 2, that is, $h_j|\{h_k,k\in\mathrm{an}(j)\}\sim N(\sum_{k\in\mathrm{an}(j)}\alpha_{kj}h_k,\sigma^2)$, or $h_j=\sum_{k\in\mathrm{an}(j)}\alpha_{kj}h_k+\boldsymbol{e}_j$, where $\boldsymbol{e}_j\sim N(0,\sigma^2)$ is the unobserved error orthogonal to the projection space spanned by $\{h_k:k\in\mathrm{an}(j)\}$, and uncorrelated with and thus independent of $\{h_k:k\in\mathrm{an}(j)\}$ and $\boldsymbol{Y}_{\mathrm{pa}(j)}$. By Assumption 1(C) and reparameterization (projecting $\boldsymbol{e}_j$ onto the space spanned by the non-valid IVs), $\boldsymbol{e}_j$ is also independent of $\boldsymbol{X}$. Then,

$$\psi_j(\mathbb{E}\left[Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)},\boldsymbol{X},h_j\right])=\boldsymbol{U}_{\mathrm{pa}(j),j}^{\top}\boldsymbol{Y}_{\mathrm{pa}(j)}+\boldsymbol{W}_{\mathrm{in}(j),j}^{\top}\boldsymbol{X}_{\mathrm{in}(j)}+\sum_{k\in\mathrm{an}(j)}\alpha_{kj}h_k+\boldsymbol{e}_j, \tag{8}$$

where DRI replaces h_k with the residuals $\widehat{h}_k$ estimated from the parent equations of Y_j. As a result, $\boldsymbol{e}_j$ is independent of $\boldsymbol{Y}_{\mathrm{pa}(j)},\boldsymbol{X}_{\mathrm{in}(j)},\sum_{k\in\mathrm{an}(j)}\alpha_{kj}h_k$ in (8), resolving the dependence issue of $\boldsymbol{Y}_{\mathrm{pa}(j)}$ on h_j in (7) due to confounding.

We propose a top-down algorithm to estimate parent-child relationships through deconfounding, given the causal ordering of the primary variables $\hat{\pi}$, and $(\overline{\mathrm{an}}(j),\overline{\mathrm{in}}(j))$, $j=1,\ldots,p$, identified by Algorithm 2. Note that the causal ordering represents the direction of edges in a DAG in that for every directed edge (k,j), i.e., $Y_k\to Y_j$, k appears before j in the ordering. The algorithm proceeds from the top to the bottom of a hierarchy defined by the causal order while identifying the parents for each primary variable and iterates this process until the last element of the ordering.

The algorithm starts from a root variable Y_k that has no parents. First, a GLM regression of Y_k is fit on its valid IVs $\boldsymbol{X}_{\overline{\mathrm{in}}(k)}$ via the model: $\psi_k(\mathbb{E}\left[Y_k|\boldsymbol{X}\right])=\boldsymbol{W}_{\overline{\mathrm{in}}(k),k}^{\top}\boldsymbol{X}_{\overline{\mathrm{in}}(k)}$. Then, we compute the

residuals $Y_{ik} - \varphi_k(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k}^\top \boldsymbol{X}_{i,\overline{\text{in}}(k)})$ to estimate the confounding effect h_{ik}, where $\varphi_k(\cdot)$ is the inverse link function for the k-th GLM model. It is important to note that the confounders do not bias the estimation of residuals in root equations by the independence assumption of the IVs and confounders. Our simulations and theory suggest that this approach works well, as in the IV regression (Johnston et al. 2008). Alternatively, we can also fit a generalized linear mixed-effects model for root equations when the data has repeated measurements. Details are given in Algorithm 5 of the Appendix.

The algorithm then moves to a non-root variable Y_j and considers the GLM regression on its ancestors $\boldsymbol{Y}_{\overline{\text{an}}(j)}$, its IVs $\boldsymbol{X}_{\overline{\text{in}}(j)}$, and the estimated confounder $\widehat{h}_k$ from the ancestor equations via the model: $\psi_j(\mathbb{E}\left[Y_j|\boldsymbol{Y}_{\text{pa}(j)}, \boldsymbol{X}, h_j\right]) = \boldsymbol{U}_{\overline{\text{an}}(j),j}^\top \boldsymbol{Y}_{\overline{\text{an}}(j)} + \boldsymbol{W}_{\overline{\text{in}}(j),j}^\top \boldsymbol{X}_{\overline{\text{in}}(j)} + \sum_{k\in\overline{\text{an}}(j)} \alpha_{kj}\widehat{h}_k + \boldsymbol{e}_j$, where $(\text{pa}(j), h_k)$ in (8) is replaced by $(\overline{\text{an}}(j), \widehat{h}_k)$. Specifically, it fits TLP-constrained GLM regressions:

$$
\begin{aligned}
&(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}, \widehat{\boldsymbol{\alpha}}_{\overline{\text{an}}(j),j}) \\
&= \underset{\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j}, \boldsymbol{\alpha}_{\overline{\text{an}}(j),j}}{\text{argmin}} \mathcal{L}(\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j}, \boldsymbol{\alpha}_{\overline{\text{an}}(j),j} | \boldsymbol{X}_{\overline{\text{in}}(j)}, \boldsymbol{Y}_{\overline{\text{an}}(j)}, \widehat{\boldsymbol{h}}_{\overline{\text{an}}(j)}) \\
&\text{subject to} \quad \sum_{k\in\overline{\text{an}}(j)} I(U_{kj} \neq 0) \leq \overline{K_j}, \quad \sum_{k\in\overline{\text{an}}(j)} I(\alpha_{kj} \neq 0) \leq \overline{K_j'}, \quad j = 1, \ldots, p,
\end{aligned} \tag{9}
$$

where $0 \leq \overline{K_j} \leq |\overline{\text{an}}(j)|$ and $0 \leq \overline{K_j'} \leq |\overline{\text{an}}(j)|$ can be tuned as in (5), with $|\cdot|$ denoting the size of a set; $\boldsymbol{W}_{\overline{\text{in}}(j),j}$ is unconstrained so that Assumption (1)(B) continues to satisfy; $\mathcal{L}(\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j}, \boldsymbol{\alpha}_{\overline{\text{an}}(j),j} | \boldsymbol{X}_{\overline{\text{in}}(j)}, \boldsymbol{Y}_{\overline{\text{an}}(j)}, \widehat{\boldsymbol{h}}_{\overline{\text{an}}(j)}) = n^{-1}\sum_{i=1}^n \ell(Y_{ij}, \boldsymbol{W}_{\overline{\text{in}}(j),j}^\top \boldsymbol{X}_{i,\overline{\text{in}}(j)} + \boldsymbol{U}_{\overline{\text{an}}(j),j}^\top \boldsymbol{Y}_{i,\overline{\text{an}}(j)} + \boldsymbol{\alpha}_{\overline{\text{an}}(j),j}^\top \widehat{\boldsymbol{h}}_{i,\overline{\text{an}}(j)})$; $\boldsymbol{h}_{i,\overline{\text{an}}(j)}$ denotes a column vector consisting of $\{h_{ik} : k \in \overline{\text{an}}(j)\}$ and $\boldsymbol{\alpha}_{\overline{\text{an}}(j),j}^\top \widehat{\boldsymbol{h}}_{i,\overline{\text{an}}(j)} = \sum_{k\in\widehat{\text{an}}(j)} \widehat{\alpha}_{kj}\widehat{h}_{ik}$. From (9), we obtain the estimated set $\widehat{\text{pa}}(j) = \{k \in \overline{\text{an}}(j) : \widehat{U}_{kj} \neq 0\} \subset \overline{\text{an}}(j)$, and $\widehat{\text{in}}(j) = \overline{\text{in}}(j)$. Finally, we compute the residuals

$$
\widehat{h}_{ij} = Y_{ij} - \varphi_j(\widehat{\boldsymbol{U}}_{\widehat{\text{pa}}(j),j}^\top \boldsymbol{Y}_{i,\widehat{\text{pa}}(j)} + \widehat{\boldsymbol{W}}_{\widehat{\text{in}}(j),j}^\top \boldsymbol{X}_{i,\widehat{\text{in}}(j)} + \sum_{k\in\widehat{\text{an}}(j)} \widehat{\alpha}_{kj}\widehat{h}_{ik}). \tag{10}
$$

Algorithm 3 summarizes the peeling process for identifying parent-child relationships using the proposed deconfounders.

Algorithm 3: Peeling algorithm for estimating parent-child relationships via DRI

---

1. Input $(\overline{\text{an}}(j), \overline{\text{in}}(j))_{j=1}^p$ and $\hat{\pi}$ from Algorithm 2. Input data matrix $(Y_{ij}, X_{ij})_{n\times(p+q)} = (\boldsymbol{Y}_{i\bullet}, \boldsymbol{X}_{i\bullet})_{i=1}^n$ of primary variables $\boldsymbol{Y}_{n\times p}$ and instruments $\boldsymbol{X}_{n\times q}$.
Begin Iteration: for $d = 1 \cdots, p$,
2. **(Estimating the confounding effects via IV regression)** If $\hat{\pi}_d$ is a root variable indexed by Y_k, compute $\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k}$ by fitting a GLM regression of Y_k on $\boldsymbol{X}$: $\mathbb{E}[Y_k|\boldsymbol{X}] = \varphi_k(\boldsymbol{W}_{\overline{\text{in}}(k),k}^\top \boldsymbol{X}_{\overline{\text{in}}(k)})$. Compute the residuals: $\widehat{h}_{ik} = Y_{ik} - \varphi_k(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k}^\top \boldsymbol{X}_{i,\overline{\text{in}}(k)})$.
3. **(Deconfounding)** If $\hat{\pi}_d$ is a non-root variable indexed by Y_j, compute $(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}, \widehat{\boldsymbol{\alpha}}_{\overline{\text{an}}(j),j})$ by fitting a TLP-constrained GLM regression of Y_j in (9). Compute the residuals $\widehat{h}_{ij}$ in (10).

Remark: The computational complexity of Algorithm 3 amounts to solving at most p TLP-constrained regressions in (5) of size $|\overline{\text{in}}(j)| + 2|\overline{\text{an}}(j)|$ via Algorithm 1, which is of order $p(p+q)^2 \max(n, (p+q)) \log K_j^0$.

3.3.2 Connections with 2SRI and 2SPS

DRI is reminiscent of, but fundamentally different from the two-stage predictor substitution (2SPS, (Cai et al. 2011)) and two-stage residual inclusion (2SRI, (Hausman 1978; Terza et al. 2008)), both of which require a known causal order between two primary variables. Similar to 2SRI, our DRI uses estimated residuals as additional predictors in subsequent GLM regressions to deconfound. In 2SRI, the residuals obtained at the first stage serve as an additional predictor as opposed to replacing the endogenous variables with their predicted values in 2SPS, which is also known as two-stage least squares for Gaussian data. However, neither applies to our situation of multiple primary variables with an unknown causal order and different confounders among equations.

For our problem, we also include a version of predictor substitution, referred to as DPS, to compare with DRI in the Appendix. In practice, we recommend DRI for causal discovery due to its superior performance and theoretical guarantees, and therefore integrate it with our top-down peeling algorithm for implementation. DRI explores the connection between parent and child equations to eliminate the confounding effect in a child equation through the residuals, whereas DPS cannot capture this aspect. This recommendation is consistent with the observation of Terza et al. (2008) and Ying et al. (2019) that 2SRI suits more than 2SPS for general nonlinear outcomes, including binary or discrete outcomes in our case. Moreover, in Algorithm 3, we use the residuals from a GLM to estimate the unmeasured confounders. Yet, one may employ different models to estimate the confounders based on their distribution. In Appendix B, we present a general framework of the deconfounding algorithm and then propose a generalized linear mixed model (GLMM) to estimate the confounders when the data has repeated measurements.

4. Theory

This section presents a novel theoretical analysis of the proposed approach, offering theoretical guarantees even in the presence of confounders. First, we demonstrate in Theorem 4 that the proposed DC algorithm, Algorithm 1, successfully recovers the true support of $\boldsymbol{V}^0$, terminates within a finite number of steps, and achieves a global minimizer for the nonconvex minimization (5), with probability approaching one. Based on this, our bottom-up peeling algorithm, Algorithm 2, retrieves the true super-graph $\mathcal{S}$. Secondly, we prove in Theorem 5 that our top-down peeling algorithm, Algorithm 3, accurately reconstructs the true causal graph, thereby identifying all parent-child relationships.

Consider a generalized linear model with the canonical link, where the negative log-likelihood of Y_{ij} given $\boldsymbol{X}_{i\bullet}$ based on independent observations $(Y_{ij}, \boldsymbol{X}_{i\bullet})_{i=1}^n$ can be expressed as:

$$\ell(Y_{ij}, \theta(\boldsymbol{X}_{i\bullet})) = -Y_{ij}\theta(\boldsymbol{X}_{i\bullet}) + A_j(\theta(\boldsymbol{X}_{i\bullet})), \quad i = 1, \ldots, n. \tag{11}$$

Here, $A_j(\theta)$ represents the cumulant function of an exponential family distribution, with θ denoting the regression function. For instance, in the case of the logistic regression, $A_j(\theta) = \log(1 + \exp(\theta))$. Given the canonical link, $A_j^{'}(\theta) = E(Y_j|\cdot) = \psi_j^{-1}(\theta) = \varphi_j(\theta)|_{\theta=\boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}}$. Hence, the log-likelihood of Y_{ij} given $\boldsymbol{X}_{i\bullet}$ for the fidelity model (3) can be written as:

$$\ell(Y_{ij}, \boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}) = -Y_{ij}\big(\boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}\big) + A_j(\boldsymbol{V}_{\bullet j}^\top \boldsymbol{X}_{i\bullet}), \quad i = 1, \ldots, n. \tag{12}$$

Subsequently, we denote ^{0} as the true parameter; for example, $\boldsymbol{V}^0$ means the true parameter values of $\boldsymbol{V}$. Denote $S_j^0 = \{l : V_{lj}^0 \neq 0\}$. Let $K_j^0 = \|\boldsymbol{V}_{\bullet j}^0\|_0 = |S_j^0|$ and $K_{\max}^0 = \max_{1 \leq j \leq p} K_j^0$. The following technical conditions are assumed for the fidelity model (3).

Assumption 3 (GLM residuals) *Assume there exists an interval $[K_1, K_2]$ such that ${\boldsymbol{V}^0_{\bullet j}}^\top \boldsymbol{X}_{i\bullet} \in [K_1, K_2]$. Further, assume that for any $\theta \in [K_1 - \epsilon, K_2 + \epsilon]$ with some constant $\epsilon > 0$, there exist some positive constants L_1 and L_2, such that $|A_j''(\theta)| \le L_1$, $|A_j'''(\theta)| \le L_2$, $j = 1, \ldots, p$, where $''$ and $'''$ denote the second and third derivatives. Moreover, $\{\xi_{ij}\}_{i=1}^n$ with $\xi_{ij} = Y_{ij} - \mathbb{E}[Y_{ij}| \cdot\ ]$ is sub-exponential with mean zero, so that for any real $t > 0$,*

$$\mathbb{P}\big(|n^{-1}\sum_{i=1}^n \xi_{ij}| \ge t\big) \le 2\exp\big(-\min\big(\frac{t^2}{2M^2}, \frac{t}{2M}\big)n\big), \quad j = 1, \ldots, p.$$

Note for the fidelity model, $\mathbb{E}[Y_{ij}|\boldsymbol{X}] = \varphi_j({\boldsymbol{V}^0_{\bullet j}}^\top \boldsymbol{X}_{i\bullet})$. Similar conditions have been suggested in Assumption E.1 of Ning and Liu (2017). Assumption 3 includes a wide range of exponential family distributions such as Poisson, and holds for a large class of GLMs including the Poisson regression. In particular, Ning and Liu (2017) and Yang et al. (2015) computed the exact value and thus showed the existence of L_1 and L_2 for specific GLMs including the logistic, exponential, and Poisson regressions. For linear and logistic models, Assumption 3 can be relaxed to sub-Gaussian residuals as all sub-Gaussian and bounded variables are sub-exponential (Maurer and Pontil 2021).

Assumption 4 (Restricted strong convexity) *For a constant $m > 0$,*

$$\Lambda_{\min} = \min_{A:|A|\le 2K^0_{\max}} \min_{\{(\Delta, \boldsymbol{V}_{\bullet j}):\|\Delta_{A^c}\|_1 \le 3\|\Delta_A\|_1, \boldsymbol{V}_{\bullet j} \in (\boldsymbol{V}^0_{\bullet j} - \Delta, \boldsymbol{V}^0_{\bullet j} + \Delta)\}} \frac{\Delta^\top \nabla^2 \mathcal{L}(\boldsymbol{V}_{\bullet j})\Delta}{\|\Delta\|_2^2} \ge m. \tag{13}$$

Note that (13) is the restricted strong convexity (eigenvalue) condition and requires the log-likelihood $\mathcal{L}(\boldsymbol{V}_{\bullet j})$ to be strongly convex in a neighborhood of $\boldsymbol{V}^0_{\bullet j}$, where $\nabla^2\mathcal{L}(\boldsymbol{V}^0_{\bullet j}) = \boldsymbol{X}^\top \boldsymbol{M}^j \boldsymbol{X}$ and $\boldsymbol{M}^j$ is a diagonal matrix with $\boldsymbol{M}^j_{ii} = A_j''({\boldsymbol{V}^0_{\bullet j}}^\top \boldsymbol{X}_{i\bullet})$ depending on $\boldsymbol{X}$ and $\boldsymbol{V}^0$ only. This condition has been commonly used for the analysis of the error bound of parameter estimation and the convergence analysis of optimization algorithms (Lee et al. 2015; Negahban et al. 2012; Hastie et al. 2015; Zhang 2017). Note that Assumption 4 permits correlated designs $\boldsymbol{X}$ and is a weaker condition than the irrepresentable condition required by the Lasso (van de Geer and Bühlmann 2009).

Assumption 5 (Bounded domain for interventions) *For some constants c_0-c_2 and $C_1 > 0$,*

$$\|\boldsymbol{X}\|_\infty \le c_1, \quad \|\boldsymbol{V}^0_{\bullet j}\|_2 \le C_1, \quad \|(\boldsymbol{X}^\top_{S^0_j}\boldsymbol{M}^j\boldsymbol{X}_{S^0_j}/n)^{-1}\boldsymbol{X}^\top_{S^0_j}\|_\infty \le c_2, \quad \Omega_{\max}(\boldsymbol{X}^\top_{S^0_j}\boldsymbol{X}_{S^0_j}/n) \le c_0,$$

where $\Omega_{\max}(\cdot)$ refers to the maximum eigenvalue of a matrix.

Assumption 6 (Minimum signal strength)

$$\min_{V^0_{lj} \ne 0} |V^0_{lj}| \ge 100 M c_2 \sqrt{\frac{\log q + \log n}{n}}.$$

Assumption 6 specifies the minimal signal strength over candidate interventions. Such an assumption has been widely used for establishing selection consistency in high-dimensional variable selection (Zhao et al. 2018; Fan and Lv 2011).

Theorem 4 (Reconstruction of super-graph via Algorithm 1) *Under Assumptions 3-6, for $j = 1, \ldots, p$, if the tuning parameters (τ_j, K_j) of Algorithm 1 satisfy:*

(1) (Computation) $\gamma_j \in [8\tau_j^{-1} \cdot Mc_1\sqrt{(\log q + \log n)/n}, m/6]$,

(2) (Tuning parameters) $8Mc_2\sqrt{\frac{\log q+\log n}{n}} \le \tau_j \le 0.4\min_{V^0_{lj}\neq 0}|V^0_{lj}|$, $K_j = K^0_j$,

then Algorithm 1 terminates in at most $1+\left\lceil \log(2K^0_j)/\log 4 \right\rceil$ *iterations for* (5)*, where* $\lceil\cdot\rceil$ *is the ceiling function. Moreover, for* $1 \le j \le p$,

$$\mathbb{P}\left(\widetilde{\boldsymbol{V}}_{\bullet j} \text{ is not a global minimizer of } (5)\right) \le 8q\exp(-2(\log(q)+\log(n))) = 8q^{-1}n^{-2}.$$

As a result, Algorithm 1 yields a global minimizer of (5)*,* $\widetilde{\boldsymbol{V}}_{\bullet j}$*, with probability tending to 1 as* $n \to \infty$*. Importantly, Algorithm 1, together with Algorithm 2, recovers the true super-graph* $\mathcal{S}^0$ *containing ancestral relations with probability*

$$\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}^0) \le 8pq^{-1}n^{-2},$$

where $\widehat{\mathcal{S}}$ *is obtained from Algorithm 2 and* $\mathcal{S}^0 \equiv \{(k,j) : k \in an(j)\}$*. Under Assumption 1(C) (i.e.,* $p \le q$*), with probability tending to one,* $\widehat{\mathcal{S}}$ *correctly reconstructs the true super-graph* $\mathcal{S}^0$ *and thus the causal order of* $Y_1,\ldots,Y_p$ *as* $n \to \infty$*.*

Theorem 4 ensures the consistent reconstruction of the super-graph $\mathcal{S}^0$ by Algorithm 1 and Algorithm 2, which characterizes ancestral relationships and determines the causal order of primary variables. Also, it says that Algorithm 1 (DC algorithm) attains a global minimizer almost surely as $n \to \infty$ under the data generating distribution. This result is in contrast to the strong hardness result of Chen et al. (2019) that there does not exist a polynomial-time algorithm achieving the globality of the ℓ_0-constrained optimization (5) in the worst-case scenario. We here show that with probability tending to one, this problem can be solved. In other words, the probability of the worst-case scenario tends to zero. Note that Algorithm 1 is indeed a polynomial-time algorithm with time complexity $O(q^2\max(q,n)\log K^0_j)$ for solving one ℓ_0-constrained regression in (5). In addition, since K_j is discrete, the assumption $K_j = K^0_j$ corresponds to the requirement that the optimal parameter λ for the Lasso has to be within a range of values for consistency. In practice, K^0_j is unknown and K_j is tuned via parameter selection methods.

Next, we establish causal graph selection consistency of the estimated causal graph based on the estimates $\widehat{\boldsymbol{U}}_{\bullet j}$ by Algorithm 3. On this ground, we ensure that all parent-child relationships are correctly identified. Let $\overline{K_j}^0 = \|\boldsymbol{U}^0_{\bullet j}\|_0$, $\overline{K'_j}^0 = |\mathrm{an}(j)|$, $s = \max_{1\le j\le p}|\mathrm{an}(j)|$, $\tilde{s} = \max_{1\le j\le p}\|\boldsymbol{W}^0_{\bullet j}\|_0$, and $\widetilde{\boldsymbol{Z}} = [\boldsymbol{X}_{\mathrm{in}(j)}, \boldsymbol{Y}_{\mathrm{pa}(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}(j)}]$. Under Assumption 5 with $\widetilde{\boldsymbol{Z}}$, $\|\widetilde{\boldsymbol{Z}}^\top\|_\infty \le b_1$, $\|(\widetilde{\boldsymbol{Z}}^\top\boldsymbol{M}\widetilde{\boldsymbol{Z}}/n)^{-1}\widetilde{\boldsymbol{Z}}^\top\|_\infty \le b_2$, and $\Omega_{\max}(\widetilde{\boldsymbol{Z}}^\top\widetilde{\boldsymbol{Z}}/n) \le b_0$.

Theorem 5 (Reconstruction of causal graph via Algorithm 3) *Under Assumptions 3-5 with* $\widetilde{\boldsymbol{Z}} = [\boldsymbol{X}_{in(j)}, \boldsymbol{Y}_{pa(j)}, \widehat{\boldsymbol{h}}_{an(j)}]$ *in the GLM regression* (9)*, if tuning parameters of Algorithm 3 satisfy:*

(1) (Computation) $\gamma_j \in [\tau_j^{-1}\cdot 8Mb_1\sqrt{(\log(2s+\tilde{s})+\log p)/n}, m/6]$,

(2) (Tuning parameters) $C\sqrt{\frac{\log(2s+\tilde{s})+\log p}{n}} \le \tau_j \le 0.4\min_{U^0_{kj}\neq 0}|U^0_{kj}|$, $\overline{K_j} = \overline{K_j}^0$, $\overline{K'_j} = \overline{K'_j}^0$,

where C *is a constant depending on* b_1*,* b_2 *and* b_0*, then, Algorithm 3 reconstructs the causal graph consistently with probability tending to one, or*

$$P(\widehat{E} = E^0) \to 1, \qquad \text{as } n \to \infty,$$

where $\widehat{E} = \{(k,j) : \widehat{U}_{kj} \neq 0\}$ *and* $E^0 = \{(k,j) : U^0_{kj} \neq 0\}$.

Theorem 5 suggests that Algorithm 3 recovers the true causal graph and thus causal relationships with probability tending to one as the sample size is sufficiently large. In Appendix D.5, we prove this by establishing error bounds of the estimates $\widehat{\boldsymbol{U}}_{\bullet j}$, $\widehat{\widetilde{\boldsymbol{W}}}_{\bullet j}$ for estimating $\boldsymbol{U}$ and $\boldsymbol{W}$ by Algorithm 3.

Remark: By Theorem 4 and 5, our proposed GAMPI using Algorithm 1-3 reconstructs the causal graph consistently with probability at least $1 - 8pq^{-1}n^{-2} - 8(2s + \widetilde{s})^{-1}p^{-1}$. For fixed p case, the $\log p$ term in γ_j and τ_j of Theorem 5 can be modified to $\log(np)$ respectively and the probability is then $1 - 8pq^{-1}n^{-2} - 8(2s + \widetilde{s})^{-1}n^{-2}p^{-1}$, similar to the remarks of Ravikumar et al. (2010).

5. Simulations

This section investigates the empirical performance of the proposed method. We assess the performance of GAMPI and compare it against the structure learning method NOTEARS (Zheng et al. 2018, 2020), under various graph structures (hub, chain, and random graphs) and types of outcome variables. Further, we compare GAMPI with a recently proposed structure learning method DAGMA (Bello et al. 2022) based on a log-det constraint. For NOTEARS and DAGMA, we use the loss type that is appropriate for the data type of the outcome variables. Note that DAGMA is designed exclusively for Gaussian and logistic outcomes.

5.1 Simulation Setting

The data simulation process is as follows. Firstly, we generate an adjacency matrix $\boldsymbol{U}$ based on the graph structure and construct an intervention matrix $\boldsymbol{W}$ with $W_{jj} = 1$, for $j = 1, \cdots, p$ and $W_{lj} = 0$, for $1 \leq l \neq j \leq q$. For the hub graph, $U_{1j} = 1$, $j = 2, \cdots, p$, and 0 otherwise. The random graph is simulated similarly as Li et al. (2023). Secondly, we generate Gaussian instrumental variables $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q) \sim N(0, 1)$. We also investigate the case when the instrumental variables $\boldsymbol{X}$ are correlated in Appendix C.7. For the confounders, we simulate $\boldsymbol{h} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\rho_{ij} = 0.95$. In Appendix C.1, we explore the simulation setup where the data is generated without confounders, i.e., $\boldsymbol{h} = \boldsymbol{0}$. Given $\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{W}$, and $\boldsymbol{h}$, we generate random samples $\boldsymbol{Y}$ according to (2). In this section, we consider two data types for the outcome variable Y: binary and count outcomes. In the binary case, Y_j is generated from the Bernoulli distribution with $P(Y_j = 1)$ equal to $\frac{\exp(\alpha_0 \boldsymbol{w}_j^\top X_{\text{in}(j)} + h_j)}{1+\exp(\alpha_0 \boldsymbol{w}_j^\top X_{\text{in}(j)} + h_j)}$ if Y_j is a root variable, and $\frac{\exp(\beta_1 \boldsymbol{u}_j^\top Y_{\text{pa}(j)} + \alpha_1 \boldsymbol{w}_j^\top X_{\text{in}(j)} + h_j)}{1+\exp(\beta_1 \boldsymbol{u}_j^\top Y_{\text{pa}(j)} + \alpha_1 \boldsymbol{w}_j^\top X_{\text{in}(j)} + h_j)}$ otherwise. For the hub graph, we set $\alpha_0 = 5$, $\beta_1 = 2.5$, and $\alpha_1 = 2$. For the chain graph, we set $\alpha_0 = 5$, $\beta_1 = 2.5$, and $\alpha_1 = 3$. For the random graph, we set $\alpha_0 = 5$, $\beta_1 = 3$, and $\alpha_1 = 3$.

For the count outcome, to avoid extreme values, we employ standard copula transforms to simulate Y, as described by Yang et al. (2015) and Nelsen (2007). Specifically, we first generate data using $\tilde{Y}_j = \beta_1 \boldsymbol{u}_j^\top \boldsymbol{Y}_{\text{pa}(j)} + \alpha_1 \boldsymbol{w}_j^\top \boldsymbol{X}_{\text{in}(j)} + h_j + \epsilon_j$, where ϵ_j are i.i.d. Gaussian errors. We then use a standard copula transform to ensure that the marginals of the generated data Y_j are approximately Poisson. For the hub graph, we set $\alpha_0 = 5$, $\beta_1 = 0.5$, and $\alpha_1 = 2$. For the chain graph, we set $\alpha_0 = 5$, $\beta_1 = 0.5$, and $\alpha_1 = 3$. For the random graph, we set $\alpha_0 = 4$, $\beta_1 = 1$, and $\alpha_1 = 2$. We consider three different graph structures: the hub, chain (of length 4), and random graphs. In addition, we fix the sample size $n = 500$ while varying the number of variables from 100 to 300.

To evaluate the accuracy of estimating the directed edges of a graph, we consider five evaluation metrics: the false positive rate (FPR), the false discovery rate (FDR), the F-score, the Matthews correlation coefficient (MCC), and the structural Hamming distance (SHD). The Matthews correlation coefficient is a binary classification metric defined as

$$\frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})\}^{1/2}},$$

where TP, FP, TN and FN denote the true positive, false positive, true negative, and false negative rates for edge selection. A large MCC value close to 1 indicates that the estimated edge set is close to the true edge set. In addition, the structural Hamming distance measures edge directionality between two directed graphs, which is the number of edge insertions, deletions, or flips needed to transform one graph to another graph (Tsamardinos et al. 2006). A small structural Hamming distance between two graphs of the same size indicates their closeness.

5.2 Results

This subsection reports the simulation results in a situation where we simulate the data in the presence of confounders. Table 2 suggests that GAMPI outperforms NOTEARS across all setups in terms of causal graph recovery, as measured by five metrics: FPR, FDR, F-score, MCC, and SHD. Table 2 shows that NOTEARS can yield an empty graph with no edges selected when "NA" occurs. Table 6 in Appendix C.4 suggests that GAMPI outperforms DAGMA significantly in most scenarios, except for the simple case of the hub graph, where both methods perform equally well. Further, note that unlike GAMPI, NOTEARS and DAGMA do not guarantee acyclicity or estimate the parameters of causal effects. To conclude, our simulation results demonstrate the advantage of the proposed method for causal graph recovery in the presence of confounders.

Binary

Graph	(p,q,n)	FPR		FDR		F-score		MCC		SHD	
		No-tears	GAMPI	No-tears	GAMPI	No-tears	GAMPI	No-tears	GAMPI	No-tears	GAMPI
Hub	(100,100,500)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.05 (0.01)	0.16 (0.01)	0.96 (0.01)	0.29 (0.01)	0.96 (0.01)	90.30 (0.78)	8.10 (1.46)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.04 (0.01)	0.13 (0.02)	0.95 (0.01)	0.25 (0.03)	0.95 (0.01)	184.90 (2.37)	20.40 (3.95)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.01)	0.21 (0.02)	0.95 (0.01)	0.34 (0.02)	0.95 (0.01)	263.50 (3.87)	28.20 (7.61)
Chain	(100,100,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.16 (0.02)	NA (NA)	0.87 (0.01)	0.00 (0.00)	0.87 (0.01)	75.00 (0.00)	21.00 (2.72)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.21 (0.01)	NA (NA)	0.84 (0.01)	0.02 (0.01)	0.84 (0.01)	149.80 (0.13)	52.30 (2.31)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.22 (0.01)	NA (NA)	0.83 (0.01)	0.01 (0.01)	0.83 (0.01)	224.80 (0.13)	84.30 (5.17)
Random	(100,100,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.14 (0.01)	NA (NA)	0.74 (0.02)	0.08 (0.02)	0.74 (0.02)	73.00 (1.56)	33.90 (1.98)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	0.52 (0.08)	0.17 (0.01)	0.03 (0.01)	0.69 (0.01)	0.09 (0.01)	0.70 (0.01)	147.90 (5.32)	78.40 (3.25)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	0.61 (0.04)	0.26 (0.01)	0.03 (0.00)	0.64 (0.00)	0.07 (0.01)	0.65 (0.00)	224.30 (6.86)	144.00 (3.69)

Count

Graph	(p,q,n)	FPR		FDR		F-score		MCC		SHD	
		No-tears	GAMPI	No-tears	GAMPI	No-tears	GAMPI	No-tears	GAMPI	No-tears	GAMPI
Hub	(100,100,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	99.00 (0.00)	0.30 (0.30)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	199.00 (0.00)	1.40 (1.19)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	299.00 (0.00)	2.50 (0.79)
Chain	(100,100,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	0.95 (0.00)	0.00 (0.00)	0.95 (0.00)	75.00 (0.00)	7.30 (0.58)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	0.94 (0.01)	0.00 (0.00)	0.94 (0.01)	150.00 (0.00)	17.80 (1.58)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	0.92 (0.00)	0.00 (0.00)	0.92 (0.00)	225.00 (0.00)	32.60 (1.45)
Random	(100,100,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	0.92 (0.01)	0.00 (0.00)	0.92 (0.01)	73.00 (2.93)	11.10 (1.28)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.01 (0.00)	NA (NA)	0.89 (0.01)	0.00 (0.00)	0.89 (0.01)	154.60 (3.88)	31.40 (2.79)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	NA (NA)	0.00 (0.00)	NA (NA)	0.88 (0.01)	0.00 (0.00)	0.89 (0.01)	228.60 (2.93)	49.20 (2.45)

Table 2: Comparison of causal graph reconstruction accuracy of GAMPI and NOTEARS in the presence of confounders, with GAMPI employing EBIC for tuning parameter selection and NOTEARS applying the default value of 0.1. Metrics include FPR, FDR, F-score, MCC, and SHD. NA indicates that the method returns an empty graph with no edges selected.

In Appendix C.2, we further compare our deconfounding approach via DRI with employing the standard GLM in Algorithm 3 for binary outcomes. For the chain graph, the standard logistic regression without adjusting for confounders does not perform well in terms of causal discovery. This is because the unobserved confounders induce false positive edges between the node and its ancestors. By contrast, the deconfounding approach corrects the bias of the confounders and recovers the true graph structure. For the hub graph, though both two approaches recover the

true causal graph, the confounding approach still outperforms the standard logistic regression in terms of parameter estimation. Note that our peeling algorithm in the first stage identifies the correct ancestral relationships (super-graph) as the confounders are independent of the instrumental variables by assumption.

In addition, in Appendix C.1, we consider the special case when the data is simulated without confounders. The result suggests that our deconfounding approach performs well even when the data is simulated without confounders. Last, we consider the simulation setup where the data has repeated measurements. Still, our deconfounding approach using a mixed-effects model outperforms the standard GLM approach. To summarize, our deconfounding approach demonstrates strong empirical performance and outperforms the existing methods in most cases.

6. Mixed DAG Networks: Direct Effect to AD

This section applies GAMPI to a publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Our goal is to estimate a regulatory gene expression network of a subset of genes related to Alzheimer's disease (AD) and identify which of the genes have a direct causal effect on AD through gene-to-gene and gene-to-AD regulatory networks.

First, we download the raw data from the ANDI website (`https://adni.loni.usc.edu`), containing gene expression, DNA sequencing, and phenotypic data. Then, for preprocessing, we clean and merge these data to obtain 712 subjects with complete records. In addition, from the KEGG database (Kanehisa et al. 2002), we extract the AD reference pathway (hsa05010, `https://www.genome.jp/pathway/hsa05010`) and therefore obtain 146 genes from the ANDI data. Meanwhile, the subjects are categorized into four groups: Cognitive Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Alzheimer's Disease (AD). We treat the 247 CN individuals as the control group and the remaining 465 AD and MCI individuals as the case group. We then include the disease status, a binary outcome with 0/1 indicating normal/AD, as an additional variable (node) to identify which genes are directly related to AD. Moreover, we use SNPs as instrumental variables in this case study as it is known that biologically SNPs may have an impact on the genes, but not the other way around, therefore satisfying the IV requirement.

To perform data analysis, we first regress the gene expressions on the additional covariates, including age, gender, education, handedness, and intracranial volume. Next, for each SNP from a gene, we perform significance tests with the gene and disease status marginally and select the genes which have at least one SNP whose i) significance level with the gene is less than 0.05 and ii) significance level with the disease status is less than 0.02, rendering $p = 39$ primary variables. For these genes, we extract their two most correlated SNPs with the disease status based on the p-values given the significance level with the gene less than 0.05, yielding $q = 39 \times 2 = 78$ instrumental variables. Removing duplicate SNPs and the gene that has the same SNPs as other genes results in $p = 38$ and $q = 76$. To summarize, we use the gene expressions along with the disease status as primary variables and SNPs as instrumental variables to reconstruct a causal network for gene-to-gene and gene-to-disease regulatory relationships.

As shown in Figure 2, GAMPI identifies a direct causal effect of gene ATF6 on the AD status. In the literature, ATF6 is a transcription factor that acts during endoplasmic reticulum (ER) stress by activating UPR target genes, and ER stress is known to be closely associated with AD. Furthermore, Du et al. (2020) suggested that ATF6 could be a potential hub for targeting the treatment of AD, which protects the retention of spatial memory in AD model mice. Zhang et al. (2022) found that the expression of both ATF6 and CTH are decreased in AD patients and ATF6 positively regulates the expression of CTH so that the addition of CTH reduces the loss of spatial learning and memory

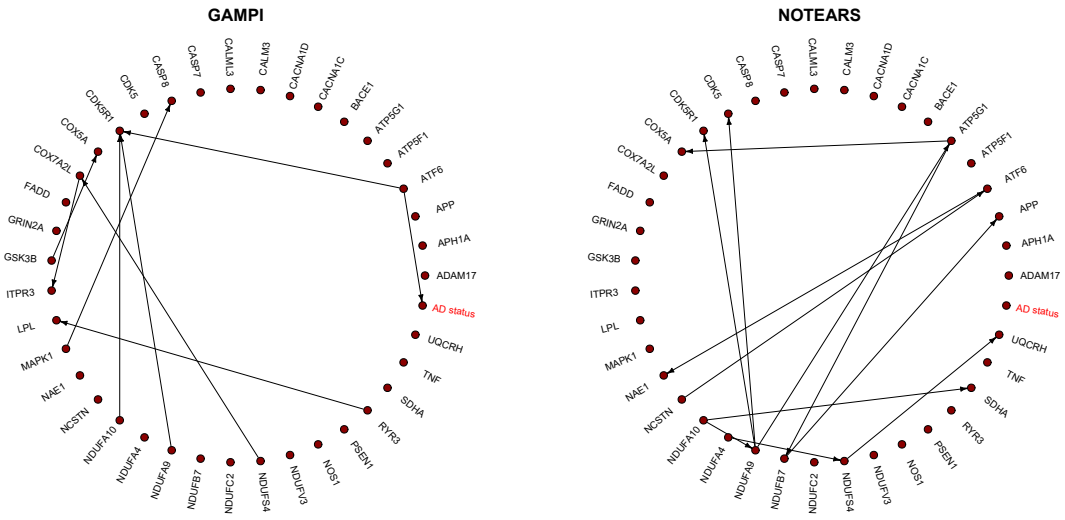

Figure 2: Reconstructed gene-to-gene and gene-to-AD regulatory network. "AD status" is a binary outcome with 0/1 indicating normal/AD. Directed edges indicate causal relationships identified by the proposed GAMPI (left) and existing method NOTEARS (right).

ability in mice caused by ATF6 reduction. In addition, GAMPI uncovers some known regulatory relationships related to AD in the literature for both the AD and control groups. For example, for the directed connection MAPK1 → CASP8, it has been shown that phosphorylation of p38 MAPK induced by oxidative stress is associated with the activation of caspase-8-mediated apoptotic pathways in dopaminergic neurons (Choi et al. 2004). The connection ATF6 → CDK5R1 is in the AD KEGG pathway `https://www.genome.jp/pathway/hsa05010`. Furthermore, the approach also identifies some potential gene regulatory relationships for future biological investigations. For example, the two genes in the connection RYR3 → LPL are among the 13 genes directly associated with AD in the DEX DFC geneset analysis (Sharma et al. 2021), while the two genes in the connection GSK3B → COX5A are in the same AD-related protein association network in AD-iPS5 neurons (Hossini et al. 2015). Further, we compare our proposed method GAMPI with the existing method NOTEARS. Both find common gene-to-gene causal relationship NDUFA9 → CDK5R1. Moreover, our proposed method identifies the meaningful gene-to-disease regulatory relationship validated biologically in the literature.

7. Discussion

The article introduces a new causal discovery approach, GAMPI, which reconstructs a directed acyclic graph using instruments in the presence of unmeasured confounders. GAMPI involves generalized structural equation models that are identifiable with the help of instruments under certain conditions. GAMPI involves two steps. First, we proposed a fidelity model that is also a generalized linear model, having the same support as the marginal model regarding instrumental interventions. On this ground, we designed a bottom-up peeling algorithm to identify ancestral relationships and valid instruments by exploiting the connection between primary and instrumental variables. In the second step, we proposed a deconfounding approach to further select parent-child relationships from the identified ancestral relationships. This approach estimates the confounding effects from the parent's equations and uses them in subsequent child equations to correct the confounding effects. The theoretical properties of GAMPI are also analyzed, including the globality

of the DC solution for nonconvex minimization, estimation accuracy, and causal graph selection consistency. A series of simulation results demonstrate causal graph selection consistency and the practical advantages of GAMPI for handling unmeasured confounders and non-Gaussian outcomes.

Overall, GAMPI provides a promising approach to causal discovery, with potential applications in various fields beyond Alzheimer's disease. For instance, the method can be used to explore causal relationships in complex systems with unmeasured confounders, such as in economics or public health. Furthermore, GAMPI's flexibility to adapt to different distributions of confounders and link functions makes it suitable for a wide range of scenarios. For instance, it can handle directed graphical models with mixed variables (Chowdhury et al. 2022). In conclusion, GAMPI offers a valuable contribution to causal inference by providing a practical method for identifying causal relationships under challenging situations.

The R implementation is available at `https://github.com/minjie-wang/GAMPI`.

Acknowledgments

The authors would like to thank the action editor and three anonymous reviewers for constructive comments and suggestions on this work. The research is supported in part by NSF grant DMS-1952539, NIH grants R01GM113250, R01GM126002, R01AG065636, R01AG074858, R01AG069895, U01AG073079.

Appendix A. Illustrative Examples

In this section, we delve into detailed examples that elucidate the peeling algorithm, the majority rule for a linear link as outlined in Assumption 1(B), and the dense confounding setting justifying Assumption 2.

A.1 Peeling Algorithm

We illustrate the peeling algorithm (Algorithm 2) with the motivating example in (4). From (4), we generate the data of sample size $n = 500$ and compute $\widehat{\boldsymbol{V}}$ using Algorithm 1. The estimated $\widehat{\boldsymbol{V}}$ is:

$$\widehat{\boldsymbol{V}}_{q\times p} = \begin{pmatrix} 2.06 & 0.35 & 0.00 & -0.35 & 0.00 \\ 0.00 & 1.84 & 0.46 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.77 & 0.39 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.76 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.96 \end{pmatrix}.$$

Algorithm 2 proceeds as follows.

- *Iteration 1:* X_4 is identified as an instrument of leaf node Y_4 ($X_4 \to Y_4$) as row 4 has the smallest row-wise ℓ_0-norm and $\widehat{V}_{44}$ is the only nonzero item in row 4.

  X_5 is identified as an instrument of leaf node Y_5 ($X_5 \to Y_5$) as row 5 has the smallest row-wise ℓ_0-norm and $\widehat{V}_{55}$ is the only nonzero item in row 5.

  Y_4, Y_5, X_4, and X_5 are removed.

- *Iteration 2:* X_3 is identified as an instrument of leaf node Y_3 ($X_3 \to Y_3$) in the subgraph for Y_1, Y_2 and Y_3 as row 3 has the smallest row-wise ℓ_0-norm of the submatrix for Y_1, Y_2 and Y_3, with $\widehat{V}_{33}$ the only nonzero item in row 3. Moreover, since $\widehat{V}_{34} \neq 0$ and Y_4 is removed in the previous iteration, $Y_3 \rightsquigarrow Y_4$.

Y_3 and X_3 are removed.

- *Iteration 3:* Similarly, X_2 is identified as an instrument of leaf node Y_2 ($X_2 \to Y_2$) in the subgraph for Y_1, Y_2. Moreover, since $\widehat{V}_{23} \neq 0$ and Y_3 is removed in the previous iteration, $Y_2 \rightsquigarrow Y_3$.

  Y_2 and X_2 are removed.

- *Iteration 4:* Similarly, X_1 is identified as an instrument of leaf node Y_1 ($X_1 \to Y_1$). Moreover, since $\widehat{V}_{12}$, $\widehat{V}_{14} \neq 0$ and Y_2, Y_4 are removed in the previous iterations, $Y_1 \rightsquigarrow Y_2$ and $Y_1 \rightsquigarrow Y_4$.

  Y_1 and X_1 are removed and the peeling process is terminated.

Finally, step 5 adds ancestral relations: $Y_1 \rightsquigarrow Y_3$ and $Y_2 \rightsquigarrow Y_4$. To conclude, Algorithm 2 identifies ancestral relationships: $Y_1 \rightsquigarrow Y_2$, $Y_1 \rightsquigarrow Y_3$, $Y_1 \rightsquigarrow Y_4$, $Y_2 \rightsquigarrow Y_3$, $Y_2 \rightsquigarrow Y_4$ and $Y_3 \rightsquigarrow Y_4$.

A.2 Majority Rule

Consider the following example of a generalized structural equation model:

$$\begin{aligned}
&\psi_1(\mathbb{E}[Y_1|X_1, X_2, X_3, h_1]) = W_{11}X_1 + W_{21}X_2 + W_{31}X_3 + h_1, \\
&\psi_2(\mathbb{E}[Y_2|Y_1, X_2, X_3, X_4, h_2]) = U_{12}Y_1 + W_{22}X_2 + W_{32}X_3 + W_{42}X_4 + h_2,
\end{aligned} \tag{14}$$

where X_1 is an valid IV of Y_1, X_4 is an valid IV of Y_2, and X_2, X_3 are invalid IVs. Note if ψ_1 is linear, then (14) is not identifiable as the majority rule is not satisfied (Kang et al. 2016; Windmeijer et al. 2019). If ψ_1 is non-linear, then the linear effect of the instruments in the first equation cannot be represented by the one in the second equation. Hence, identifiability is achieved through non-linearity and the majority rule is not required (details are given in the proof of Proposition 1).

A.3 Dense Confounding Setting

This section illustrates the dense confounding setting justifying Assumption 2, where the confounder for each variable h_j is added up by many independent confounding effects from unobserved variables. Therefore, asymptotics holds and the confounders are jointly normal by the central limit theorem.

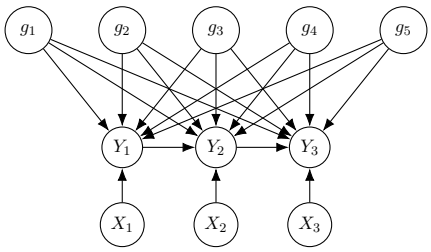

Figure 3: Dense confounding setting, where the confounder for each variable h_j is added up by many independent confounding effects from unobserved variables. For example, in this case, the confounders can be re-parameterized as: $h_1 = a_{11}g_1 + a_{21}g_2 + a_{31}g_3 + a_{41}g_4 + a_{51}g_5$, etc.

Appendix B. General Form of Deconfounding Algorithm

Algorithm 4 serves as a general version of Algorithm 3 in the main paper for estimating parent-child relationships in the presence of confounders. In Algorithm 3 of the main paper, we utilize residuals from a GLM to impute confounders. The underlying intuition of this deconfounding approach is to achieve accurate parameter estimates and construct a consistent estimate of unmeasured confounders

via residuals through the root equations. In this regard, different models can be employed to estimate confounding effects in the root equations, as described in Algorithm 4. For instance, the marginal likelihood that integrates the confounding effect h_k from the complete likelihood can be applied, in addition to the Markov Chain Monte Carlo approach (Knudson et al. 2021).

Algorithm 4: General peeling algorithm for estimating parent-child relationships via DRI

1. Input $(\overline{\text{an}}(j), \overline{\text{in}}(j))_{j=1}^{p}$ and $\hat{\pi}$ from Algorithm 2. Input data matrix $(Y_{ij}, X_{ij})_{n\times(p+q)} = (\boldsymbol{Y}_{i\bullet}, \boldsymbol{X}_{i\bullet})_{i=1}^{n}$ of primary variables $\boldsymbol{Y}_{n\times p}$ and instruments $\boldsymbol{X}_{n\times q}$.
Begin Iteration: for $d = 1 \cdots, p$,
2. **(Estimation of confounding effects)** If $\hat{\pi}_d$ is a root variable indexed by Y_k, obtain an estimate of the confounding effect $\widehat{h}_{ik}$.
3. **(Deconfounding)** If $\hat{\pi}_d$ is a non-root variable indexed by Y_j, compute $(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}, \widehat{\boldsymbol{\alpha}}_{\overline{\text{an}}(j),j})$ by fitting a TLP-constrained GLM regression of Y_j in (9). Compute the residuals $\widehat{h}_{ij}$ in (10).

Specifically, when the data has repeated measurements, we propose to use a generalized linear mixed model (GLMM) to estimate the unmeasured confounders in the root equations in Algorithm 5 as an alternative to Algorithm 3 in the main paper. Consider the structural equation model:

$$\psi_j(\mathbb{E}\left(\boldsymbol{Y}_{ij}|\boldsymbol{Y}_{i,\text{pa}(j)}, \boldsymbol{X}_{i\bullet}, h_{ij}\right)) = \boldsymbol{Y}_{i,\text{pa}(j)}\boldsymbol{U}_{\text{pa}(j),j} + \boldsymbol{X}_{i,\text{in}(j)}\boldsymbol{W}_{\text{in}(j),j} + h_{ij}\mathbf{1}_{n_i}, \quad j = 1, \cdots, p. \quad (15)$$

Here, $i = 1, \cdots, N$ represents a group index with n_i observations within a group. For each group, we observe an $n_i \times 1$ vector of responses, $\boldsymbol{Y}_{ij}$ and hence an $n_i \times p$ matrix, $\boldsymbol{Y}_{i\bullet}$. Let $\boldsymbol{X}_{i\bullet}$ be an $n_i \times q$ fixed-effects design matrix, $\boldsymbol{W}_{\text{in}(j),j}$ an $|\text{in}(j)| \times 1$, and $\boldsymbol{U}_{\text{pa}(j),j}$ a $|\text{pa}(j)| \times 1$ column vector of fixed regression coefficients. Further, h_{ij} denotes a group-specific vector of random intercepts.

Similarly, we adopt a two-stage deconfounding procedure to estimate parent-child relationships, with a GLMM in the root equation for improved confounder estimation. Specifically, we fit a GLMM on variable Y_k using instrumental variables $\boldsymbol{X}_{\text{in}(k)}$ and estimate confounding effects $\{\widehat{h}_{ik}\}_{i=1}^{N}$ via the estimated random effects. In the child equation, we impute unmeasured confounders using estimated values from the parent equation and fit a TLP-constrained GLM, similar to the previous approach.

Algorithm 5: Peeling algorithm for estimating parent-child relationships in the presence of confounders using GLMM and DRI

1. Input $(\overline{\text{an}}(j), \overline{\text{in}}(j))_{j=1}^{p}$ and $\hat{\pi}$ from Algorithm 2. Input $(\boldsymbol{Y}_{ij}, \boldsymbol{X}_{ij})_{n\times(p+q)} = (\boldsymbol{Y}_{i\bullet}, \boldsymbol{X}_{i\bullet})_{i=1}^{N}$ of primary variables $\boldsymbol{Y}_{n\times p}$ and instruments $\boldsymbol{X}_{n\times q}$. Here, $\boldsymbol{Y}_{i\bullet}$ is an $n_i \times p$ matrix and $\boldsymbol{X}_{i\bullet}$ is an $n_i \times q$ matrix. Input the grouping of subjects with $n = \sum_{i=1}^{N} n_i$.
Begin Iteration: for $d = 1 \cdots, p$,
2. **(Estimation of confounding effects using GLMM for root equations)** If $\hat{\pi}_d$ is a root variable indexed by Y_k, estimate the confounding effects $\{h_{ik}\}_{i=1}^{N}$ by fitting a GLMM on $\boldsymbol{Y}_{\bullet k}$:

$$\mathbb{E}[\boldsymbol{Y}_{ik}|\boldsymbol{X}_{i,\text{in}(k)}, h_{ik}] = \varphi_k(\boldsymbol{X}_{i,\overline{\text{in}}(k)}\boldsymbol{W}_{\overline{\text{in}}(k),k} + h_{ik}\mathbf{1}_{n_i}),$$

where h_{ik} denotes the random effect for ith group. Obtain estimated confounding effects $\{\widehat{h}_{ik}\}$.
3. **(Deconfounding)** If $\hat{\pi}_d$ is a non-root variable indexed by Y_j, compute $(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}, \widehat{\boldsymbol{\alpha}}_{\overline{\text{an}}(j),j})$ by fitting a TLP-constrained GLM regression of Y_j in (9). Compute the residuals $\{\widehat{h}_{i'j}\}_{i'=1}^{n}$ in (10).

Besides the residual inclusion approach proposed in the main paper for deconfounding, we also include a version of incorporating predictor substitution approach in GAMPI, referred to as DPS, in Algorithm 6. Here, $\mathcal{L}(\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j} | \boldsymbol{X}_{\overline{\text{in}}(j)}, \widehat{\boldsymbol{Y}}_{\overline{\text{an}}(j)}) = n^{-1} \sum_{i=1}^{n} \ell(Y_{ij}, \boldsymbol{W}_{\overline{\text{in}}(j),j}^{\top} \boldsymbol{X}_{i,\overline{\text{in}}(j)} + \boldsymbol{U}_{\overline{\text{an}}(j),j}^{\top} \widehat{\boldsymbol{Y}}_{i,\overline{\text{an}}(j)})$, which indicates the endogenous variables are replaced by their predicted values.

Algorithm 6: Peeling algorithm for estimating parent-child relationships via DPS

1. Input $(\overline{\text{an}}(j), \overline{\text{in}}(j))_{j=1}^{p}$ and $\hat{\pi}$ from Algorithm 2. Input data matrix $(Y_{ij}, X_{ij})_{n\times(p+q)} = (\boldsymbol{Y}_{i\bullet}, \boldsymbol{X}_{i\bullet})_{i=1}^{n}$ of primary variables $\boldsymbol{Y}_{n\times p}$ and instruments $\boldsymbol{X}_{n\times q}$.
Begin Iteration: for $d = 1 \cdots, p$,
2. **(Predictor substitution for root equation)** If $\hat{\pi}_d$ is a root variable indexed by Y_k, compute $\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k}$ by fitting a GLM regression of Y_k on $\boldsymbol{X}$: $\mathbb{E}[Y_k|\boldsymbol{X}] = \varphi_k(\boldsymbol{X}_{\overline{\text{in}}(k)} \boldsymbol{W}_{\overline{\text{in}}(k),k})$. Impute the predictor: $\widehat{Y}_k = \varphi_k(\boldsymbol{X}_{\overline{\text{in}}(k)} \widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k})$.
3. **(Predictor substitution for child equation)** If $\hat{\pi}_d$ is a non-root variable indexed by Y_j, compute $(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j})$ by fitting a TLP-constrained GLM regression of Y_j:

$$(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}) = \underset{\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j}}{\text{argmin}} \mathcal{L}(\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j} | \boldsymbol{X}_{\overline{\text{in}}(j)}, \widehat{\boldsymbol{Y}}_{\overline{\text{an}}(j)})$$
$$\text{subject to} \quad \sum_{k \in \overline{\text{an}}(j)} I(U_{kj} \neq 0) \leq \overline{K_j}, \quad j = 1, \ldots, p.$$

Impute the predictor: $\widehat{Y}_j = \varphi_j(\boldsymbol{Y}_{\widehat{\text{pa}}(j)} \widehat{\boldsymbol{U}}_{\widehat{\text{pa}}(j),j} + \boldsymbol{X}_{\widehat{\text{in}}(j)} \widehat{\boldsymbol{W}}_{\widehat{\text{in}}(j),j})$.

Appendix C. Additional Simulations

This section provides additional simulations in the paper to demonstrate the necessity of deconfounding in GAMPI.

Ideally, one might suggest estimating the causal relationships directly using the nodewise GLM regression subject to the ℓ_0-constraint in Algorithm 3 of the main paper, without employing the deconfounding approach or adjusting for confounders. That is,

$$(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}) = \underset{\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j}}{\text{argmin}} \mathcal{L}(\boldsymbol{W}_{\overline{\text{in}}(j),j}, \boldsymbol{U}_{\overline{\text{an}}(j),j} | \boldsymbol{X}_{\overline{\text{in}}(j)}, \boldsymbol{Y}_{\overline{\text{an}}(j)})$$
$$\text{subject to} \quad \sum_{k \in \overline{\text{an}}(j)} I(U_{kj} \neq 0) \leq \overline{K_j}, \quad j = 1, \ldots, p. \tag{16}$$

Similarly, to select parent-child relationships from the ancestral relationships identified in the first stage, we penalize the number of nonzero elements of $\boldsymbol{U}$. That is, if $\widehat{U}_{kj} \neq 0$, then Y_k is a parent of Y_j, or $Y_k \to Y_j$.

We show in Section C.1 and C.2 that GAMPI adjusting for confounders, performs as well as the above approach (16) when the data is simulated without confounders and outperforms it in the presence of confounders.

C.1 Absence of Confounders

This subsection considers the special case when the data is simulated without confounders for binary outcomes. Recall that the binary data is simulated from the Bernoulli distribution in Section 5.1 of

the main paper with $\boldsymbol{h} = 0$. Table 3 suggests that our deconfounding approach, GAMPI, performs well even when the data is simulated without confounders.

Graph	(p,q,n)	FPR		FDR		F-score		MCC		SHD	
		GAMPI-no deconf	GAMPI	GAMPI-no deconf	GAMPI	GAMPI-no deconf	GAMPI	GAMPI-no deconf	GAMPI	GAMPI-no deconf	GAMPI
Hub	(100,100,300)	0.00 (0.00)	0.00 (0.00)	0.03 (0.00)	0.04 (0.01)	0.98 (0.00)	0.98 (0.00)	0.98 (0.00)	0.98 (0.00)	3.70 (0.52)	4.80 (0.71)
	(100,100,400)	0.00 (0.00)	0.00 (0.00)	0.02 (0.01)	0.02 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	2.20 (0.55)	2.30 (0.52)
	(100,100,500)	0.00 (0.00)	0.00 (0.00)	0.02 (0.00)	0.02 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	2.10 (0.38)	1.80 (0.29)
Chain	(100,100,300)	0.00 (0.00)	0.00 (0.00)	0.08 (0.01)	0.07 (0.01)	0.82 (0.01)	0.82 (0.01)	0.83 (0.01)	0.83 (0.01)	23.90 (1.45)	23.80 (1.23)
	(100,100,400)	0.00 (0.00)	0.00 (0.00)	0.06 (0.00)	0.06 (0.00)	0.91 (0.01)	0.91 (0.01)	0.91 (0.01)	0.91 (0.01)	13.50 (0.95)	13.20 (0.95)
	(100,100,500)	0.00 (0.00)	0.00 (0.00)	0.05 (0.01)	0.04 (0.01)	0.94 (0.00)	0.95 (0.00)	0.94 (0.00)	0.95 (0.00)	8.30 (0.40)	7.70 (0.45)
Random	(100,100,300)	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	0.09 (0.01)	0.85 (0.01)	0.85 (0.01)	0.85 (0.01)	0.85 (0.01)	20.10 (1.75)	20.10 (1.67)
	(100,100,400)	0.00 (0.00)	0.00 (0.00)	0.07 (0.01)	0.07 (0.01)	0.93 (0.01)	0.93 (0.01)	0.93 (0.01)	0.93 (0.01)	10.80 (1.75)	10.60 (1.61)
	(100,100,500)	0.00 (0.00)	0.00 (0.00)	0.04 (0.01)	0.04 (0.00)	0.97 (0.00)	0.97 (0.01)	0.97 (0.00)	0.97 (0.01)	4.80 (0.59)	4.50 (0.64)

Table 3: Evaluating GAMPI's reconstruction accuracy for binary outcomes without confounders, utilizing the extended BIC (EBIC) for tuning parameter selection. Evaluation metrics include false positive rate (FPR), false discovery rate (FDR), F-score, Matthews correlation coefficient (MCC), and structural Hamming distance (SHD). "GAMPI-no deconf" method refers to employing the nodewise GLM regression approach without adjusting for confounders based on (16).

C.2 Presence of Confounders

This subsection compares the DRI approach with that without adjusting for confounders under the simulation setting in the presence of confounders. Furthermore, we compare our deconfounding approach, DRI in Algorithm 3 of the main paper, with that via predictor substitution (DPS) in Algorithm 6. In addition to the five metrics in the paper, we compute the estimation error $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}^0\|_F^2$ to evaluate the accuracy of parameter estimation, where $\|\cdot\|_F$ denotes the Frobenius norm.

Graph	(p,q,n)	F-score			$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}^*\|_F^2$		
		GAMPI-no deconf	GAMPI-DRI	GAMPI-DPS	GAMPI-no deconf	GAMPI-DRI	GAMPI-DPS
Hub	(100,100,500)	0.96 (0.01)	0.96 (0.01)	0.91 (0.01)	77.45 (5.28)	58.11 (6.66)	108.04 (12.11)
	(200,200,500)	0.95 (0.01)	0.95 (0.01)	0.89 (0.01)	185.12 (25.01)	140.73 (25.02)	255.65 (24.69)
	(300,300,500)	0.95 (0.01)	0.95 (0.01)	0.90 (0.01)	279.17 (35.82)	200.55 (45.36)	373.97 (47.27)
Chain	(100,100,500)	0.74 (0.01)	0.87 (0.01)	0.87 (0.01)	118.06 (6.37)	74.67 (4.74)	93.06 (5.2)
	(200,200,500)	0.71 (0.01)	0.84 (0.01)	0.84 (0.01)	281.96 (12.01)	189.05 (9.82)	217.29 (10.96)
	(300,300,500)	0.71 (0.01)	0.83 (0.01)	0.84 (0.01)	415.11 (20.26)	294.81 (12.57)	332.47 (13.82)
Random	(100,100,500)	0.71 (0.02)	0.74 (0.02)	0.74 (0.02)	282.58 (15.98)	278.67 (15.95)	321.49 (15.86)
	(200,200,500)	0.64 (0.01)	0.69 (0.01)	0.69 (0.01)	656.68 (30.57)	633.23 (31.49)	711.17 (31.42)
	(300,300,500)	0.59 (0.00)	0.64 (0.00)	0.62 (0.01)	1111.06 (37.05)	1053.11 (32.15)	1174.76 (37.16)

Table 4: Assessing GAMPI's reconstruction accuracy for binary outcomes with confounders, employing the extended BIC (EBIC) for tuning parameter selection. Evaluation metrics include F-score and parameter estimation error in the Frobenius norm. "GAMPI-DPS" employs the predictor substitution (DPS) approach for deconfounding as proposed in Algorithm 6. "GAMPI-DRI" uses the residual inclusion approach proposed in Algorithm 3 of the main paper. In other tables, "GAMPI" refers to the recommended "GAMPI-DRI" approach.

Table 4 suggests that our deconfounding approach outperforms the standard GLM approach (16) without adjusting for confounders in the presence of confounders. Moreover, deconfounding via DRI proposed in Algorithm 3 outperforms that using predictor substitution (DPS) in Algorithm 6 in terms of parameter estimation. Our simulation result indicates that DRI is more suited than DPS for binary or count outcomes, which is concordant with the observation of Terza et al. (2008).

C.3 Presence of Confounders with Replicates

In this subsection, we evaluate the performance of GAMPI using the generalized linear mixed models (GLMMs) for root equations proposed in Algorithm 5 under the simulation setting with repeated measurements. Table 5 suggests that our deconfounding approach using a GLMM outperforms the standard GLM approach, as it better estimates the confounders.

Graph	(p, q, n)	F-score			$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}^*\|_F^2$		
		GAMPI-no deconf	GAMPI	GAMPI-GLMM	GAMPI-no deconf	GAMPI	GAMPI-GLMM
Hub	(100,100,500)	0.94 (0.02)	0.94 (0.02)	0.95 (0.02)	102.96 (24.02)	79.71 (23.14)	61.6 (21.58)
	(200,200,500)	0.92 (0.01)	0.92 (0.01)	0.93 (0.01)	259.2 (24.82)	209.5 (25.42)	160.8 (21.94)
	(300,300,500)	0.91 (0.02)	0.91 (0.02)	0.93 (0.02)	357.33 (74.04)	321.79 (70.8)	251.28 (68.07)
Chain	(100,100,500)	0.75 (0.01)	0.87 (0.01)	0.93 (0.01)	125.64 (6.81)	81.8 (5.5)	65.99 (6.38)
	(200,200,500)	0.71 (0.01)	0.83 (0.01)	0.91 (0.00)	285.35 (12.08)	201.56 (4.04)	155.85 (3.66)
	(300,300,500)	0.69 (0.01)	0.80 (0.01)	0.89 (0.01)	500.19 (20.73)	366.98 (16.24)	287.8 (11.69)

Table 5: Evaluating GAMPI's reconstruction accuracy for binary outcomes in repeated measurements, using the extended BIC (EBIC) for tuning. Evaluation metrics include F-score and parameter estimation error in the Frobenius norm. "GAMPI-GLMM" refers to GAMPI using GLMM for root equations proposed in Algorithm 5.

C.4 Comparison with DAGMA

This subsection compares GAMPI with a recently proposed structure learning method, called DAGMA. DAGMA is designed only for Gaussian or logistic outcomes. Thus, we compare GAMPI with DAGMA for the binary outcomes case. Table 6 suggests that GAMPI continues to outperform DAGMA in most scenarios. Specifically, DAGMA performs equally well in the easy case, namely, the hub graph. However, in challenging situations like the chain and random graphs, GAMPI outperforms DAGMA significantly.

Binary Graph	(p, q, n)	FPR		FDR		F-score		MCC		SHD	
		DAGMA	GAMPI	DAGMA	GAMPI	DAGMA	GAMPI	DAGMA	GAMPI	DAGMA	GAMPI
Hub	(100,100,500)	0.00 (0.00)	0.00 (0.00)	0.04 (0.00)	0.05 (0.01)	0.98 (0.00)	0.96 (0.01)	0.98 (0.00)	0.96 (0.01)	4.20 (0.57)	8.10 (1.46)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	0.05 (0.01)	0.04 (0.01)	0.97 (0.00)	0.95 (0.01)	0.97 (0.00)	0.95 (0.01)	11.60 (1.60)	20.40 (3.95)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	0.07 (0.01)	0.04 (0.01)	0.96 (0.00)	0.95 (0.01)	0.96 (0.00)	0.95 (0.01)	24.40 (1.84)	28.20 (7.61)
Chain	(100,100,500)	0.01 (0.00)	0.00 (0.00)	0.43 (0.01)	0.16 (0.02)	0.68 (0.01)	0.87 (0.01)	0.70 (0.01)	0.87 (0.01)	51.70 (2.63)	21.00 (2.72)
	(200,200,500)	0.00 (0.00)	0.00 (0.00)	0.53 (0.01)	0.21 (0.01)	0.60 (0.01)	0.84 (0.01)	0.62 (0.01)	0.84 (0.01)	145.60 (2.84)	52.30 (2.31)
	(300,300,500)	0.00 (0.00)	0.00 (0.00)	0.57 (0.01)	0.22 (0.01)	0.56 (0.01)	0.83 (0.01)	0.59 (0.01)	0.83 (0.01)	247.60 (4.02)	84.30 (5.17)
Random	(100,100,500)	0.01 (0.00)	0.00 (0.00)	0.80 (0.01)	0.14 (0.01)	0.29 (0.01)	0.74 (0.02)	0.30 (0.02)	0.74 (0.02)	141.60 (3.54)	33.90 (1.98)
	(200,200,500)	0.01 (0.00)	0.00 (0.00)	0.82 (0.01)	0.17 (0.01)	0.26 (0.01)	0.69 (0.01)	0.29 (0.01)	0.70 (0.01)	342.90 (5.44)	78.40 (3.25)
	(300,300,500)	0.01 (0.00)	0.00 (0.00)	0.84 (0.01)	0.26 (0.01)	0.24 (0.01)	0.64 (0.00)	0.27 (0.01)	0.65 (0.00)	573.20 (6.88)	144.00 (3.69)

Table 6: Comparing reconstruction accuracy of GAMPI and DAGMA for binary outcomes with confounders, where GAMPI employs the extended BIC (EBIC) for tuning and DAGMA uses the default setting with a tuning parameter value of 0.02.

C.5 Tuning Parameter Selection

This subsection examines the performance of two tuning parameter selection approaches for GAMPI. We use either 5-fold cross-validation or the extended Bayesian information criterion (EBIC) to choose (τ_j, K_j) by minimizing the predictive likelihood or the EBIC criterion. For cross-validation, we adopt the one-standard error rule which is commonly used for the high-dimensional data. We consider the base simulation in the presence of confounders. Table 7 suggests that the EBIC approach outperforms cross-validation in all settings.

Graph	(p, q, n)	F-score		SHD	
		CV	EBIC	CV	EBIC
Hub	(100,100,500)	0.70 (0.03)	0.96 (0.01)	44.30 (3.97)	8.10 (1.46)
	(200,200,500)	0.70 (0.04)	0.95 (0.01)	89.20 (10.25)	20.40 (3.95)
	(300,300,500)	0.74 (0.05)	0.95 (0.01)	120.30 (15.74)	28.20 (7.61)
Chain	(100,100,500)	0.56 (0.03)	0.87 (0.01)	46.20 (1.96)	21.00 (2.72)
	(200,200,500)	0.59 (0.01)	0.84 (0.01)	88.20 (2.10)	52.30 (2.31)
	(300,300,500)	0.56 (0.01)	0.83 (0.01)	137.50 (2.93)	84.30 (5.17)
Random	(100,100,500)	0.55 (0.02)	0.74 (0.02)	46.20 (2.03)	33.90 (1.98)
	(200,200,500)	0.48 (0.02)	0.69 (0.01)	102.50 (4.47)	78.40 (3.25)
	(300,300,500)	0.47 (0.01)	0.64 (0.00)	158.90 (6.32)	144.00 (3.69)

Table 7: Reconstruction accuracy of causal graph of GAMPI for binary outcomes in the presence of confounders, where GAMPI uses cross-validation (CV) or the extended BIC (EBIC) for tuning parameter selection. Evaluation metrics include F-score and SHD.

C.6 Causal Graph Selection Consistency

In this subsection, we verify our theoretical statements and demonstrate the consistency of the proposed method, as proved in Theorem 5. Following Ravikumar et al. (2011), we evaluate the performance of the method in terms of the probability of correct causal graph selection. Figure 4 plots the probability of correct causal graph recovery against the sample size n, with varying number of nodes p. The probability of correct causal graph selection is calculated as the proportion of the $B = 50$ trials in which the proposed GAMPI recovers the directed edge sets exactly. We consider the hub and chain graphs for Poisson primary variables with the same setup as the base simulation in Table 2 except varying the sample size n and number of nodes p. For each curve, the probability of success starts at zero, and converges to one as the sample size increases, suggesting the causal graph selection consistency of our proposed method. Figure 4 shows that the proposed method performs well when the sample size n is large or the number of nodes p (graph size) is small, aligning with the theoretical results stated in Theorem 5. Figure 4 also suggests that a larger graph size requires a larger sample size for exact graph recovery, so that the curve for $p = 300$ is shifted to the right compared with the curve for $p = 100$.

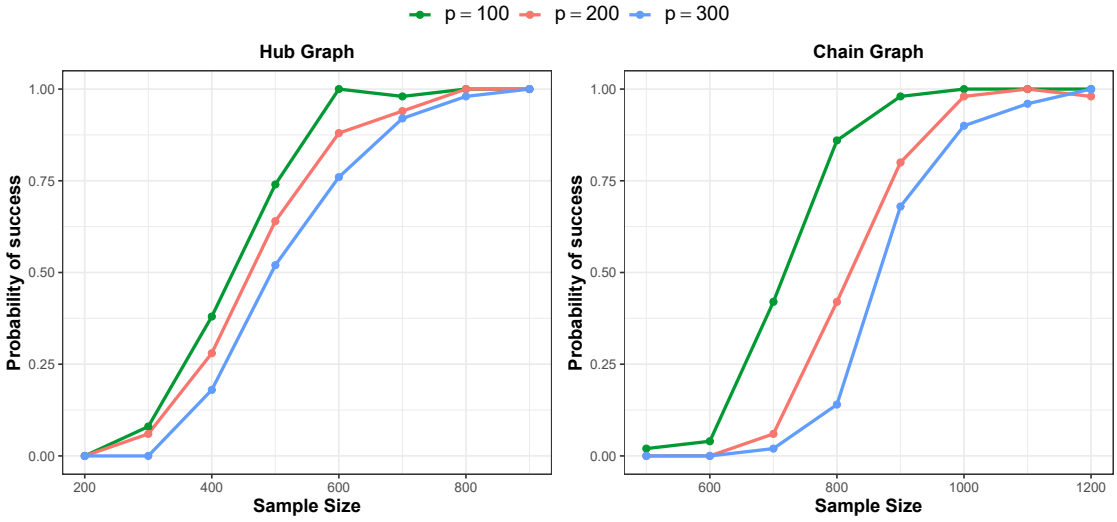

Figure 4: Simulation results of graph selection consistency for Poisson hub and chain graphs with varying number of nodes p; plots of probability of correct directed edge-set recovery versus the sample size n. Each point corresponds to the average over 50 trials.

C.7 Correlated Instrumental Variables

In this subsection, we evaluate the performance of the proposed method when the instrumental variables $\boldsymbol{X}$ are correlated. We consider the same setup for binary primary variables as the base simulation in Table 2 except that $\boldsymbol{X}$ is Gaussian with autoregressive covariance $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ where $(\boldsymbol{\Sigma}_X)_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. Figure 5 shows that our method still performs well in the case where the instrumental variables are correlated. We compare our proposed method with DAGMA as it is shown to outperform NOTEARS in Appendix C.4.

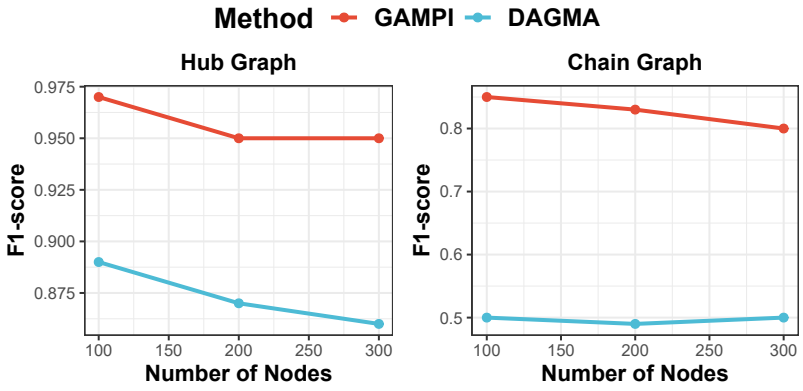

Figure 5: Simulation results when the instrumental variables $\boldsymbol{X}$ are correlated.

C.8 Comparison with Linear Deconfounding Structure Learning Algorithm

This subsection compares GAMPI with a recently proposed linear deconfounding algorithm, called GrIVET (Chen et al. 2023). Compared with Chen et al. (2023), we use the generalized linear models to account for different distributions of the outcome variables, which enhance model interpretation. Moreover, in contrast to the imputation-based approach by Chen et al. (2023), we propose a residual-inclusion-based deconfounding algorithm to address confounders, which has been shown to be more suitable for nonlinear outcomes. Going beyond, we present novel theoretical analysis including the fidelity model and consistency of residual inclusion for the GLMs. We compare the two methods for the Poisson outcomes case as in Table 2. Figure 6 suggests that our proposed method outperforms the existing linear deconfounding algorithm, demonstrating the advantage of the proposed method for handling non-Gaussian outcomes.

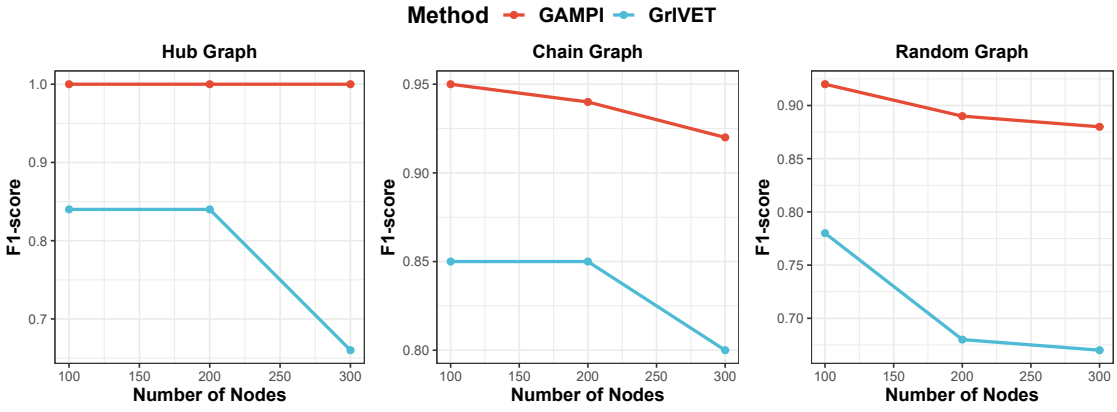

Figure 6: Simulation results of comparison with linear deconfounding algorithm, GrIVET.

Appendix D. Technical Proofs

D.1 Proof of Proposition 1

We prove the proposition in two steps. First, we show that the topological order and the corresponding DAG are identifiable. Next, we show that the model parameters are identifiable given the graph. Assume that two structural equation models as in equation (2), defined by $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{W})$ and $\tilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{U}}, \widetilde{\boldsymbol{W}})$, induce the same distribution of $(\boldsymbol{Y}, \boldsymbol{X})$. We will show that $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Identifying G. Let $G(\boldsymbol{\theta})$ and $G(\tilde{\boldsymbol{\theta}})$ be the DAGs corresponding to $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$. First, we show that the topological order of $Y_1, \cdots, Y_p$ is identifiable. For $G(\boldsymbol{\theta})$, assume, without loss of generality, that Y_1 is a leaf node in $G(\boldsymbol{\theta})$. By Assumption 1(B), there exists a valid instrument, say X_1, that intervenes on Y_1. By Assumptions 1(A) and (ii),

$$\operatorname{Cov}\left(Y_1, X_1 \mid \boldsymbol{Y}_S, \boldsymbol{X}_{\{2,\ldots,q\}}\right) \neq 0, \qquad \text{for any } S \subseteq \{2, \ldots, p\}, \tag{17}$$

$$\operatorname{Cov}\left(Y_j, X_1 \mid \boldsymbol{X}_{\{2,\ldots,q\}}\right) = 0, \qquad j = 2, \ldots, p. \tag{18}$$

Hence, (17) implies that $X_1 \to Y_1$ in $G(\tilde{\boldsymbol{\theta}})$. Now suppose Y_1 is not a leaf node in $G(\tilde{\boldsymbol{\theta}})$ and there exists Y_2 such that $Y_1 \to Y_2$. Then, $\operatorname{Cov}\left(Y_2, X_1 \mid \boldsymbol{X}_{\{2,\ldots,q\}}\right) = 0$ by (18) but $X_1 \to Y_1$ and $Y_1 \to Y_2$, which contradicts Assumption 1(A). Therefore, if Y_1 is a leaf node in $G(\boldsymbol{\theta})$, then Y_1 must also be a leaf node in $G(\tilde{\boldsymbol{\theta}})$. Therefore, we can identify the leaf nodes $\boldsymbol{Y}_{\mathcal{L}_1}$ in the graph. Further, following Li et al. (2023), the parents and instruments in $G(\boldsymbol{\theta})$ and $G(\tilde{\boldsymbol{\theta}})$ of Y_1 can be identified by:

$$\mathbb{E}(Y_1 \mid \boldsymbol{Y}_{-1}, \boldsymbol{X}, \boldsymbol{h}) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\mathrm{pa}_{G(\boldsymbol{\theta})}(1)}, \boldsymbol{X}, h_1) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\mathrm{pa}_{G(\boldsymbol{\theta})}(1)}, \boldsymbol{X}_{\mathrm{in}_{G(\boldsymbol{\theta})}(1)}, h_1), \tag{19}$$

$$\mathbb{E}(Y_1 \mid \boldsymbol{Y}_{-1}, \boldsymbol{X}, \boldsymbol{h}) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\mathrm{pa}_{G(\tilde{\boldsymbol{\theta}})}(1)}, \boldsymbol{X}, h_1) = \mathbb{E}(Y_1 \mid \boldsymbol{Y}_{\mathrm{pa}_{G(\tilde{\boldsymbol{\theta}})}(1)}, \boldsymbol{X}_{\mathrm{in}_{G(\tilde{\boldsymbol{\theta}})}(1)}, h_1), \tag{20}$$

where $\mathrm{pa}_{G(\boldsymbol{\theta})}(1)$ refers to the parent variables of Y_1 in $G(\boldsymbol{\theta})$ and $\mathrm{in}_{G(\boldsymbol{\theta})}(1)$ refers to the instrumental variables of Y_1 in $G(\boldsymbol{\theta})$; $\mathrm{pa}_{G(\tilde{\boldsymbol{\theta}})}(1)$ and $\mathrm{in}_{G(\tilde{\boldsymbol{\theta}})}(1)$ are similarly defined for $G(\tilde{\boldsymbol{\theta}})$. We have $\mathrm{pa}_{G(\boldsymbol{\theta})}(1) = \mathrm{pa}_{G(\tilde{\boldsymbol{\theta}})}(1)$ and $\mathrm{in}_{G(\boldsymbol{\theta})}(1) = \mathrm{in}_{G(\tilde{\boldsymbol{\theta}})}(1)$. To see this, if there exists Y_k such that $k \in \mathrm{pa}_{G(\boldsymbol{\theta})}(1)$ and $k \notin \mathrm{pa}_{G(\tilde{\boldsymbol{\theta}})}(1)$, we have $\operatorname{Cov}\left(Y_1, Y_k \mid \boldsymbol{Y}_{\{2,\ldots,p\}\setminus k}, \boldsymbol{X}, h_1\right) \neq 0$ by (19) and $\operatorname{Cov}\left(Y_1, Y_k \mid \boldsymbol{Y}_{\{2,\ldots,p\}\setminus k}, \boldsymbol{X}, h_1\right) = 0$ by (20), leading to a contradiction, and we conclude that $\mathrm{pa}_{G(\boldsymbol{\theta})}(1) = \mathrm{pa}_{G(\tilde{\boldsymbol{\theta}})}(1)$. Similarly, we have $\mathrm{in}_{G(\boldsymbol{\theta})}(1) = \mathrm{in}_{G(\tilde{\boldsymbol{\theta}})}(1)$. Therefore, the leaf node Y_1 has the same parents and instruments in $G(\boldsymbol{\theta})$ and $G(\tilde{\boldsymbol{\theta}})$.

Toward this end, we have identified the leaf nodes $\boldsymbol{Y}_{\mathcal{L}_1}$ along with their parents and instruments respectively. Next, after removing the leaf nodes $\boldsymbol{Y}_{\mathcal{L}_1}$, we apply the same argument and identify the leaf variables $\boldsymbol{Y}_{\mathcal{L}_2}$ in the sub-graph. We proceed until all the variables are removed, leading to $G(\boldsymbol{\theta}) = G(\tilde{\boldsymbol{\theta}})$ and $\mathrm{in}(j) = \widetilde{\mathrm{in}}(j)$, $\forall j$. In this way, the graph G and the topological order can be identified.

Identifying $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{W})$ **given** G. Second, we show that $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. Recall that for the jth equation, $\psi_j(\mathbb{E}\left(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j\right)) = \boldsymbol{u}_j^\top \boldsymbol{Y}_{\mathrm{pa}(j)} + \boldsymbol{w}_j^\top \boldsymbol{X}_{\mathrm{in}(j)} + h_j$, $j = 1, \cdots, p$. Let Y_k be a parent of Y_j. We can rewrite the above equation as $\psi_j(\mathbb{E}\left(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j\right)) = U_{kj}\boldsymbol{Y}_k + \boldsymbol{U}_{\mathrm{pa}(j)\setminus k,j}\boldsymbol{Y}_{\mathrm{pa}(j)\setminus k} + \boldsymbol{w}_j^\top \boldsymbol{X}_{\mathrm{in}(j)} + h_j$. Similarly, for the kth equation, $\mathbb{E}\left(\boldsymbol{Y}_k|\boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) = \psi_k^{-1}(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k)$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left(\psi_j(\mathbb{E}\left(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, h_j\right)) - \boldsymbol{U}_{\mathrm{pa}(j)\setminus k,j}\boldsymbol{Y}_{\mathrm{pa}(j)\setminus k}|\boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) \\
&= U_{kj}\psi_k^{-1}(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k) + \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + \mathbb{E}\left(h_j|\boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right).
\end{aligned} \tag{21}$$

Note that the left-hand side is not equal to $\psi_j(\mathbb{E}\left(Y_j|\boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_j\right))$ but still characterizes a proper conditional distribution. We next prove the identifiability of $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{W})$ by induction. Suppose $\boldsymbol{u}_k = \boldsymbol{U}_{\bullet k}$ and $\boldsymbol{w}_k = \boldsymbol{W}_{\bullet k}$ are identified. We will show that U_{kj} and $\boldsymbol{W}_{\mathrm{in}(j),j}$ are identifiable and

therefore $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{W})$ is also identifiable by induction on the topological depth. If there exist $\tilde{U}_{kj} \neq U_{kj}$ and $\widetilde{\boldsymbol{W}}_{\mathrm{in}(j),j} \neq \boldsymbol{W}_{\mathrm{in}(j),j}$ which render the same conditional distribution (21) in that $U_{kj}\psi_k^{-1}(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k) + \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + \mathbb{E}_{\boldsymbol{\theta}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) = \widetilde{U}_{kj}\psi_k^{-1}(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k) + \widetilde{\boldsymbol{W}}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + \mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right)$. Rearranging terms yields that

$$\begin{aligned} &U_{kj}\psi_k^{-1}(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k) - \widetilde{U}_{kj}\psi_k^{-1}(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k) \\ &= \widetilde{\boldsymbol{W}}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} - \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + \mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) - \mathbb{E}_{\boldsymbol{\theta}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right). \end{aligned} \tag{22}$$

If $\psi_k^{-1}(\cdot)$ is a non-linear function, then the left-hand side cannot be linearly represented by the right-hand side linear function of $\boldsymbol{X}_{\mathrm{in}(j)}$. To see this, note that there exists an instrumental variable X_l for Y_k, $l \in \mathrm{in}(k)$. In addition, by Assumption 2, $\mathbb{E}_{\boldsymbol{\theta}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) = \mathbb{E}_{\boldsymbol{\theta}}\left(h_j | \boldsymbol{h}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) = \sum_{m \in \{k \cup \mathrm{pa}(k)\}} \alpha_m h_m$, leading to $\mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) - \mathbb{E}_{\boldsymbol{\theta}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) = \sum_m (\tilde{\alpha}_m - \alpha_m) h_m$. Taking the second derivative of (22) with respect to X_l and then h_k yields that

$$(U_{kj} - \widetilde{U}_{kj}) W_{lk}^2 \cdot (\psi_k^{-1})''''(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + \boldsymbol{w}_k^\top \boldsymbol{X}_{\mathrm{in}(k)} + h_k) = 0,$$

where we use the property that the second derivative of a linear function is zero. This implies $U_{kj} = \widetilde{U}_{kj}$ and therefore U_{kj} is identifiable. Further, plugging $U_{kj} = \widetilde{U}_{kj}$ into (22) yields $\widetilde{\boldsymbol{W}}_{\mathrm{in}(j),j} = \boldsymbol{W}_{\mathrm{in}(j),j}$. Note the statement still holds in the absence of the confounders $\boldsymbol{h}$ by taking the second derivative of (22) with respect to X_l.

If $\psi_k^{-1}(\cdot)$ is a linear function, then the same conclusion holds under the majority rule that the number of valid IVs for Y_k exceeds 50% of its total number of IVs. To see this, let $\mathrm{in}_*(k)$ and $\widetilde{\mathrm{in}}_*(k)$ denote the valid IVs of Y_k in $G(\boldsymbol{\theta})$ and $G(\tilde{\boldsymbol{\theta}})$ respectively. Given a linear ψ_k^{-1}, (22) can be written as

$$\begin{aligned} U_{kj}\boldsymbol{W}_{\mathrm{in}(k),k}^\top \boldsymbol{X}_{\mathrm{in}(k)} - \widetilde{U}_{kj}\boldsymbol{W}_{\mathrm{in}(k),k}^\top \boldsymbol{X}_{\mathrm{in}(k)} = (\widetilde{U}_{kj} - U_{kj})(\boldsymbol{u}_k^\top \boldsymbol{Y}_{\mathrm{pa}(k)} + h_k) + \widetilde{\boldsymbol{W}}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} - \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} \\ + \mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right) - \mathbb{E}_{\boldsymbol{\theta}}\left(h_j | \boldsymbol{Y}_{\mathrm{pa}(k)}, \boldsymbol{X}, h_k\right). \end{aligned}$$

Let I denote the right-hand side term, which is not a function of $\boldsymbol{X}_{\mathrm{in}_*(k)}$ or $\boldsymbol{X}_{\widetilde{\mathrm{in}}_*(k)}$. Rearranging terms yields that

$$U_{kj}\boldsymbol{W}_{\mathrm{in}_*(k),k}^\top \boldsymbol{X}_{\mathrm{in}_*(k)} - \widetilde{U}_{kj}\boldsymbol{W}_{\widetilde{\mathrm{in}}_*(k),k}^\top \boldsymbol{X}_{\widetilde{\mathrm{in}}_*(k)} = \widetilde{U}_{kj}\boldsymbol{W}_{\mathrm{in}(k)\setminus\widetilde{\mathrm{in}}_*(k),k}^\top \boldsymbol{X}_{\mathrm{in}(k)\setminus\widetilde{\mathrm{in}}_*(k)} - U_{kj}\boldsymbol{W}_{\mathrm{in}(k)\setminus\mathrm{in}_*(k),k}^\top \boldsymbol{X}_{\mathrm{in}(k)\setminus\mathrm{in}_*(k)} + I.$$

By the majority rule, $|\mathrm{in}_*(k)| > |\mathrm{in}(k)|/2$ and $|\widetilde{\mathrm{in}}_*(k)| > |\mathrm{in}(k)|/2$. Hence, there must exist some valid IV, $l \in \mathrm{in}_*(k) \cap \widetilde{\mathrm{in}}_*(k)$, such that $(U_{kj} - \widetilde{U}_{kj})W_{lk}X_l$ cannot be linearly represented by $\boldsymbol{X}_{\mathrm{in}(k)\setminus(\mathrm{in}_*(k)\cap\widetilde{\mathrm{in}}_*(k))}$ or $\boldsymbol{X}_{\mathrm{in}(k)^c}$. Again, we have $\widetilde{U}_{kj} = U_{kj}$, $\widetilde{\boldsymbol{W}}_{\mathrm{in}(j),j} = \boldsymbol{W}_{\mathrm{in}(j),j}$. This completes the proof.

Before proving Proposition 2 and Proposition 3, we first introduce Lemma 6 which investigates the marginal distribution defined by the true model $\mathbb{P}(Y_j|\boldsymbol{X})$ and instruments.

Lemma 6 *For a valid instrument X_l, if the marginal distribution under the true model $\mathbb{P}(Y_j|\boldsymbol{X})$ satisfies: $\frac{\partial}{\partial X_l}\mathbb{P}(Y_j|\boldsymbol{X}) \neq 0$, then X_l intervenes on Y_j or an ancestor of Y_j.*

Proof of Lemma 6. Note that the marginal distribution can be written as

$$f(Y_j|\boldsymbol{X}) = \iint f(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}_{\mathrm{in}(j)}, h_j) f(\boldsymbol{Y}_{\mathrm{pa}(j)}|\boldsymbol{X}) f(h_j)\, d\boldsymbol{Y}_{\mathrm{pa}(j)}\, dh_j. \tag{23}$$

If $\frac{\partial}{\partial X_l}\mathbb{P}(Y_j|\boldsymbol{X}) \neq 0$, or equivalently, $\frac{\partial}{\partial X_l} f(Y_j|\boldsymbol{X}) \neq 0$, then, by the product rule and Assumption 1(C), i) $\frac{\partial}{\partial X_l} f(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}_{\mathrm{in}(j)}, h_j) \neq 0$, or ii) $\frac{\partial}{\partial X_l} f(\boldsymbol{Y}_{\mathrm{pa}(j)}|\boldsymbol{X}) \neq 0$. Note $f(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}_{\mathrm{in}(j)}, h_j) =$

$\exp(Y_j\big(\boldsymbol{U}_{\mathrm{pa}(j),j}^\top \boldsymbol{Y}_{\mathrm{pa}(j)} + \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + h_j\big) - A_j(\boldsymbol{U}_{\mathrm{pa}(j),j}^\top \boldsymbol{Y}_{\mathrm{pa}(j)} + \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + h_j))$ under (2). By the chain rule, $\frac{\partial f(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)},\boldsymbol{X}_{\mathrm{in}(j)},h_j)}{\partial X_l} = f(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)},\boldsymbol{X}_{\mathrm{in}(j)},h_j)(Y_j - \varphi_j(\boldsymbol{U}_{\mathrm{pa}(j),j}^\top \boldsymbol{Y}_{\mathrm{pa}(j)} + \boldsymbol{W}_{\mathrm{in}(j),j}^\top \boldsymbol{X}_{\mathrm{in}(j)} + h_j))W_{lj}$. Therefore, condition i) implies that $W_{lj} \neq 0$ and $l \in \mathrm{in}(j)$. Condition ii) implies that there exists an $m \in \mathrm{pa}(j)$ such that $\frac{\partial}{\partial X_l} f(Y_m|\boldsymbol{X}) \neq 0$. Similarly, this implies $l \in \mathrm{in}(m), m \in \mathrm{pa}(j)$, or there exists an $r \in \mathrm{pa}(m)$ such that $\frac{\partial}{\partial X_l} f(Y_r|\boldsymbol{X}) \neq 0$. By induction, we conclude that if $\frac{\partial}{\partial X_l}\mathbb{P}(Y_j|\boldsymbol{X}) \neq 0$, then (i) there exists an $l \in \mathrm{in}(j)$ such that $W_{lj} \neq 0$, or (ii) there exists an $k \in \mathrm{an}(j)$ and $l \in \mathrm{in}(k)$ such that $W_{lk} \neq 0$. Hence, X_l intervenes on Y_j or an ancestor of Y_j.

Remark: By the proof of Lemma 6 and (23), for a general IV (valid or non-valid), if $\frac{\partial}{\partial X_l}\mathbb{P}(Y_j|\boldsymbol{X}) \neq 0$, then i) $\frac{\partial}{\partial X_l} f(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)},\boldsymbol{X}_{\mathrm{in}(j)},h_j) \neq 0$, or ii) $\frac{\partial}{\partial X_l} f(\boldsymbol{Y}_{\mathrm{pa}(j)}|\boldsymbol{X}) \neq 0$, or iii) $\frac{\partial}{\partial X_l} f(h_j) \neq 0$. This suggests: i) X_l intervenes on Y_j or an ancestor of Y_j, or ii) X_l is correlated with h_j or $h_{\mathrm{an}(j)}$.

D.2 Proof of Proposition 2

Recall that $S_j = \{l : V_{lj} \neq 0\}$ for the fidelity model (3) and $\widetilde{S}_j = \{l : \frac{\partial \mathbb{P}(Y_j|\boldsymbol{X})}{\partial X_l} \neq 0\}$ for the true model $\mathbb{P}(Y_j|\boldsymbol{X})$. Next, we will show that $\{l : V_{lj} \neq 0\} = \widetilde{S}_j$, implying that the fidelity model $\mathbb{P}^*(Y_j|\boldsymbol{X})$ and the marginal model $\mathbb{P}(Y_j|\boldsymbol{X})$ have the same support.

For any $l \in \widetilde{S}_j$, $\frac{\partial \mathbb{P}(Y_j|\boldsymbol{X})}{\partial X_l} \neq 0$. By Lemma 6 and the remark, i) X_l intervenes on Y_j or an ancestor of Y_j, or ii) X_l is correlated with h_j or $h_{\mathrm{an}(j)}$. Case i) suggests that $l \in \{\mathrm{in}(j)\} \cup \mathrm{in}(\mathrm{an}(j))$. Hence, there exists a path in the graph from X_l to Y_j: $X_l \to Y_k \to \cdots \to Y_j$. By the local faithfulness in Assumption 1(A), $\mathrm{Cov}(X_l, Y_j) \neq 0$. This implies that $V_{lj} \neq 0$ in the fidelity model. Otherwise, suppose $V_{lj} = 0$. By (3), $\mathbb{E}(Y_j|X_l, \boldsymbol{X}_{-l}) = \mathbb{E}(Y_j|\boldsymbol{X}_{-l})$, implying that $\mathbb{P}(Y_j|X_l, \boldsymbol{X}_{-l}) = \mathbb{P}(Y_j|\boldsymbol{X}_{-l})$ and thus $\mathrm{Cov}(X_l, Y_j) = 0$ by the definition of conditional independence, which contradicts $\mathrm{Cov}(X_l, Y_j) \neq 0$. Hence, $V_{lj} \neq 0$ in the fidelity model. For case ii), $\mathrm{Cov}(X_l, h_j) \neq 0$ or $\mathrm{Cov}(X_l, h_{\mathrm{an}(j)}) \neq 0$ implies $\mathrm{Cov}(X_l, Y_j) \neq 0$. Following the same argument as in case i), we obtain $V_{lj} \neq 0$. Combining the two cases, $V_{lj} \neq 0$ in the fidelity model or $l \in S_j$, implying $\widetilde{S}_j \subset S_j$.

On the other hand, for any $l \in S_j$, $V_{lj} \neq 0$. Then, $\mathbb{E}[Y_j|X_l, \boldsymbol{X}_{-l}] \neq \mathbb{E}[Y_j|\boldsymbol{X}_{-l}]$. Now, suppose $\frac{\partial \mathbb{P}(Y_j|\boldsymbol{X})}{\partial X_l} = 0$. Then, as in (23), there does not exist a path in the graph from X_l to Y_j. Moreover, we have $\frac{\partial}{\partial X_l} f(h_j) = 0$ and $\frac{\partial}{\partial X_l} f(h_{\mathrm{an}(j)}) = 0$, implying $\mathrm{Cov}(X_l, h_j) = 0$ and $\mathrm{Cov}(X_l, h_{\mathrm{an}(j)}) = 0$. Thus, $\mathrm{Cov}(X_l, Y_j) = 0$, which contradicts $\mathbb{E}[Y_j|X_l, \boldsymbol{X}_{-l}] \neq \mathbb{E}[Y_j|\boldsymbol{X}_{-l}]$. Hence, $l \in \widetilde{S}_j$ and thus $S_j \subset \widetilde{S}_j$. This establishes that $S_j = \widetilde{S}_j$.

D.3 Proof of Proposition 3

If $V_{lj} \neq 0$, then $\frac{\partial}{\partial X_l}\mathbb{P}(Y_j|\boldsymbol{X}) \neq 0$ by Proposition 2. For a valid instrument X_l, by Lemma 6, X_l intervenes on Y_j or an ancestor of Y_j.

Moreover, following Li et al. (2023), for a leaf node Y_j, there exists a valid instrument $X_l \to Y_j$ by Assumption 1(B). If there exists $j' \neq j$ such that $V_{lj'} \neq 0$, then Y_j must be an ancestor of $Y_{j'}$, which contradicts the fact that Y_j is a leaf node. On the other hand, suppose $V_{lj} \neq 0$ and $V_{lj'} = 0$, $\forall j' \neq j$. If Y_j is not a leaf node, then there exists a $Y_{j'}$ such that Y_j is a parent of $Y_{j'}$. This implies $\frac{\partial}{\partial X_l} f(Y_{j'}|\boldsymbol{X}) \neq 0$ and thus $V_{lj'} \neq 0$, which contradicts $\|V_{l\bullet}\|_0 = 1$.

D.4 Proof of Theorem 4

Let $S_j^0 = \{l : V_{lj}^0 \neq 0\}$ and $S_j^{[t]} = \{l : |\widetilde{V}_{lj}^{[t]}| \geq \tau_j\}$ be the indices of the true and estimated non-zero elements of the jth columns $\widehat{\boldsymbol{V}}_{\bullet j}^0$ and $\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]}$ at the t-th iteration of Algorithm 1, respectively. Let the

corresponding false negative and positive sets be $\mathrm{FN}_j^{[t]} = S_j^0 \backslash S_j^{[t]}$ and $\mathrm{FP}_j^{[t]} = S_j^{[t]} \backslash S_j^0$ at iteration t. Let an event $\mathcal{E}_j = \{\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty \le 0.5\gamma_j\tau_j\} \cap \{\|\widehat{\boldsymbol{V}}_{\bullet j}^0 - \boldsymbol{V}_{\bullet j}^0\|_\infty \le 0.5\tau_j\}$, where $\widehat{\boldsymbol{\xi}}_j = \boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}_{\bullet j}^0)$ is the residual of the oracle MLE $\widehat{\boldsymbol{V}}_{\bullet j}^0$ for the GLM, with the support $\{l : \widehat{V}_{lj}^0 \neq 0\} = S_j^0$. Consider the data matrix $(\boldsymbol{X}_{n\times q}, \boldsymbol{Y}_{n\times p})$ and $\boldsymbol{Y}_j$ refers to the j-th column of $\boldsymbol{Y}$, that is, an $n \times 1$ vector.

Our proof consists of three steps. In **Step 1**, we show by induction that if $\big|S_j^0 \cup S_j^{[t-1]}\big| \le 2K_j^0$ on $\mathcal{E}_j$, then $|S_j^0 \cup S_j^{[t]}| \le 2K_j^0$, $t = 1, \ldots$, so that Assumption 4 applies. Recall that $K_j^0 = \|\boldsymbol{V}_{\bullet j}^0\|_0 = |S_j^0|$. In **Step 2**, we estimate the number of iterations to termination T. Particularly, we prove that $|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}| < 1$ or $|\mathrm{FP}_j^{[t]}| = |\mathrm{FN}_j^{[t]}| = 0$ and thus $S_j^{[t]} = S_j^0$, for $t \ge T$. In **Step 3**, we bound $\mathbb{P}(\mathcal{E}_j)$ and show that $1 - \mathbb{P}\big(\cup_{j=1}^p \mathcal{E}_j^c\big)$ has a high probability tending to one as $n \to \infty$.

Step 1: Suppose $|S_j^0 \cup S_j^{[t-1]}| \le 2K_j^0$ on $\mathcal{E}_j$. By the Taylor's expansion of the gradient $\nabla\mathcal{L}(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]})$ at $\widehat{\boldsymbol{V}}_{\bullet j}^0$,

$$\nabla\mathcal{L}(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]}) = \nabla\mathcal{L}(\widehat{\boldsymbol{V}}_{\bullet j}^0) + \nabla^2\mathcal{L}(\overline{\boldsymbol{V}}_{\bullet j})(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0), \tag{24}$$

where $\overline{\boldsymbol{V}}_{\bullet j}$ is a vector of intermediate values on the line between $\widehat{\boldsymbol{V}}_{\bullet j}^0$ and $\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]}$, $\nabla\mathcal{L}(\widehat{\boldsymbol{V}}_{\bullet j}^0) = n^{-1}\sum_{i=1}^n \boldsymbol{X}_{i\bullet}(-Y_{ij} + \varphi_j(\boldsymbol{X}_{i\bullet}^\top \widehat{\boldsymbol{V}}_{\bullet j}^0)) = -n^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j$ with $\widehat{\boldsymbol{\xi}}_j = \boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}_{\bullet j}^0)$. By the optimality condition of (6) at iteration t,

$$0 \le (\widehat{\boldsymbol{V}}_{\bullet j}^0 - \widetilde{\boldsymbol{V}}_{\bullet j}^{[t]})^\top \big(\nabla\mathcal{L}(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]}) + \gamma_j\tau_j\nabla\big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]})_{\left(S_j^{[t-1]}\right)^c}\big\|_1\big), \tag{25}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm. On the other hand, by the optimality condition of the oracle estimator $\widehat{\boldsymbol{V}}_{\bullet j}^0$: $\boldsymbol{X}_{S_j^0}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}_{\bullet j}^0)) = \boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j = \boldsymbol{0}$ on S_j^0, implying that $(\boldsymbol{R}_j)_{S_j^{[t-1]} \cap S_j^0} = \boldsymbol{0}$, where $\boldsymbol{R}_j = \boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n - \gamma_j\tau_j\nabla\big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]})_{\left(S_j^{[t-1]}\right)^c}\big\|_1$. Let $S_j^0 \Delta S_j^{[t-1]} = (S_j^0 \backslash S_j^{[t-1]}) \cup (S_j^{[t-1]} \backslash S_j^0)$, where Δ denotes the symmetric difference.

Hence, combination of (24) and (25) yields that

$$\begin{aligned}
&(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)^\top \nabla^2\mathcal{L}(\overline{\boldsymbol{V}}_{\bullet j})(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0) \le (\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)^\top \big(\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n - \gamma_j\tau_j\nabla\big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]})_{\left(S_j^{[t-1]}\right)^c}\big\|_1\big) \\
&\le (\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}^\top (\boldsymbol{R}_j)_{S_j^0 \Delta S_j^{[t-1]}} + (\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}^\top (\boldsymbol{R}_j)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c} \\
&\le \big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}\big\|_1 (\|\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j) \\
&+ \big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}\big\|_1 (\|\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n\|_\infty - \gamma_j\tau_j),
\end{aligned} \tag{26}$$

where the last inequality holds since $(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}^\top (\nabla\big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]})_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}\big\|_1) = \big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}\big\|_1$. Note that $(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)^\top \nabla^2\mathcal{L}(\overline{\boldsymbol{V}}_{\bullet j})(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0) \ge 0$ since $\nabla^2\mathcal{L}(\overline{\boldsymbol{V}}_{\bullet j})$ is positive-definite. By (26),

$$\big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}\big\|_1 (\gamma_j\tau_j - \big\|\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n\big\|_\infty) \le \big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}\big\|_1 (\big\|\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n\big\|_\infty + \gamma_j\tau_j).$$

Note, on event $\mathcal{E}_j$, $\big\|\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n\big\|_\infty \le \gamma_j\tau_j/2$, and thus

$$\Big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{(t-1)}\right)^c}\Big\|_1 \le 3\Big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}\Big\|_1 \le 3\Big\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \cup S_j^{[t-1]}}\Big\|_1.$$

Note that $\left|S_j^0 \cup S_j^{[t-1]}\right| \le 2K_j^0$. By Assumption 4 and (26),

$$\begin{aligned}
&m\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2^2 \le (\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)^\top \nabla^2 \mathcal{L}(\overline{\boldsymbol{V}}_{\bullet j})(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0) \\
&\le \big(\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j\big)\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}\|_1 + \big(\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty - \gamma_j\tau_j\big) \\
&\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{\left(S_j^0 \cup S_j^{[t-1]}\right)^c}\|_1 \\
&\le \big(\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty + \gamma_j\tau_j\big)\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}\|_1, \\
&\le 1.5\gamma_j\tau_j\sqrt{|S_j^0 \Delta S_j^{[t-1]}|} \cdot \|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2, \qquad (27)
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and $\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty \le 0.5\gamma_j\tau_j$ on $\mathcal{E}_j$. Hence,

$$\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2/\tau_j \le (1.5\gamma_j/m)\sqrt{2K_j^0} \le \sqrt{K_j^0}, \qquad (28)$$

since $|S_j^0 \Delta S_j^{[t-1]}| \le |S_j^0 \cup S_j^{[t-1]}| \le 2K_j^0$ and $\gamma_j \le m/6$ by Condition (1) of Theorem 4. Moreover, $\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2^2 \ge |\mathrm{FP}_j^{[t]}| \cdot \tau_j^2$ since $|\widetilde{\boldsymbol{V}}_{lj}^{[t]} - \widehat{\boldsymbol{V}}_{lj}^0| = |\widetilde{\boldsymbol{V}}_{lj}^{[t]}| > \tau_j$ for any $l \in \mathrm{FP}_j^{[t]} = S_j^{[t]} \backslash S_j^0$. By (28), $|\mathrm{FP}_j^{[t]}| \le \|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2^2/\tau_j^2 \le K_j^0$. Therefore, $|S_j^0 \cup S_j^{[t]}| = |S_j^0| + |\mathrm{FP}_j^{[t]}| \le 2K_j^0$.

Step 2: Suppose $|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}| \ge 1$. Similarly,

$$\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2^2 \ge (|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}|)(0.5\tau_j)^2,$$

since $|\widetilde{V}_{lj}^{[t]} - \widehat{V}_{lj}^0| \ge |\widetilde{V}_{lj}^{[t]} - V_{lj}^0| - |\widehat{V}_{lj}^0 - V_{lj}^0| \ge \tau_j - 0.5\tau_j$ for any $l \in \mathrm{FP}_j^{[t]} \cup \mathrm{FN}_j^{[t]}$, by Assumption 6. Therefore, $\sqrt{|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}|} \le \|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2/0.5\tau_j$. Moreover, by (27) and the Cauchy-Schwarz inequality, $m\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2^2 \le 1.5\gamma_j\tau_j\|(\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0)_{S_j^0 \Delta S_j^{[t-1]}}\|_1 \le 1.5\gamma_j\tau_j\sqrt{|S_j^0 \Delta S_j^{[t-1]}|} \cdot \|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2$. Hence, $\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2/\tau_j \le (1.5\gamma_j/m)\sqrt{|\mathrm{FP}_j^{[t-1]}| + |\mathrm{FN}_j^{[t-1]}|}$. By Conditions (1) and (2) of Theorem 4:

$$\sqrt{|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}|} \le \frac{\|\widetilde{\boldsymbol{V}}_{\bullet j}^{[t]} - \widehat{\boldsymbol{V}}_{\bullet j}^0\|_2}{0.5\tau_j} \le \frac{3\gamma_j}{m}\sqrt{|\mathrm{FP}_j^{[t-1]}| + |\mathrm{FN}_j^{[t-1]}|} \le 0.5\sqrt{|\mathrm{FP}_j^{[t-1]}| + |\mathrm{FN}_j^{[t-1]}|}.$$

Iterating this process implies that $\sqrt{|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}|} \le (\frac{1}{2})^t\sqrt{|S_j^0| + |S_j^{[0]}|}$, $t = 0, 1, \ldots$. If $t \ge T = 1 + \lceil \log(2K_j^0)/\log 4 \rceil$, then $|\mathrm{FP}_j^{[t]}| + |\mathrm{FN}_j^{[t]}| < 1$ or $\mathrm{FP}_j^{[t]} = \mathrm{FN}_j^{[t]} = \emptyset$ on event $\mathcal{E}_j$. Consequently, $\{l : \widetilde{V}_{lj}^{[T]} \ne 0\} = \{l : V_{lj}^0 \ne 0\} = S_j^0$.

Step 3: To bound $P\big(\bigcup_{j=1}^p \mathcal{E}_j^c\big)$, recall that $\mathcal{E}_j = \big\{\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty \le 0.5\gamma_j\tau_j\big\} \cap \big\{\|\widehat{\boldsymbol{V}}_{\bullet j}^0 - \boldsymbol{V}_{\bullet j}^0\|_\infty \le 0.5\tau_j\big\}$. Next, we bound the two events in $\mathcal{E}_j^c$ separately. For the first event, by the triangular inequality,

$$\begin{aligned}
&\mathbb{P}\big(\|\boldsymbol{X}^\top \widehat{\boldsymbol{\xi}}_j/n\|_\infty > 0.5\gamma_j\tau_j\big) = \mathbb{P}\big(\|\boldsymbol{X}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}_{\bullet j}^0))/n\|_\infty > 0.5\gamma_j\tau_j\big) \\
&\le \mathbb{P}\big(\|\boldsymbol{X}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\boldsymbol{V}_{\bullet j}^0))/n\|_\infty > 0.25\gamma_j\tau_j\big) \\
&+ \mathbb{P}\big(\|\boldsymbol{X}^\top(\varphi_j(\boldsymbol{X}\boldsymbol{V}_{\bullet j}^0) - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}_{\bullet j}^0))/n\|_\infty > 0.25\gamma_j\tau_j\big). \qquad (29)
\end{aligned}$$

By Assumption 5, $|X_{ik}| \le c_1$. By Assumption 3, $Y_{ij} - \varphi_j({\boldsymbol{V}^0_{\bullet j}}^\top \boldsymbol{X}_{i\bullet})$ is sub-exponential with the bound M. Hence, by Bernstein's inequality (Theorem 2.8.2 of Vershynin (2018)), for any given $k = 1, \cdots, q$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_{ik}(Y_{ij} - \mathbb{E}[Y_{ij}|\boldsymbol{X}])/n\right| \ge 0.25\gamma_j\tau_j\right) \le 2\exp\left(-\min\left(\frac{\gamma_j^2\tau_j^2 n}{32M^2c_1^2}, \frac{\gamma_j\tau_j n}{8Mc_1}\right)\right).$$

Note that $\left\|\boldsymbol{X}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\boldsymbol{V}^0_{\bullet j}))/n\right\|_\infty = \max_{k=1}^q |\sum_{i=1}^n X_{ik}(Y_{ij} - \mathbb{E}[Y_{ij}|\boldsymbol{X}])/n|$. The union bound yields, for the first quantity in (29), that

$$\begin{aligned}
&\mathbb{P}\big(\left\|\boldsymbol{X}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}\boldsymbol{V}^0_{\bullet j}))/n\right\|_\infty > 0.25\gamma_j\tau_j\big) \le 2q\exp\big(-\min\big(\frac{\gamma_j^2\tau_j^2 n}{32M^2c_1^2}, \frac{\gamma_j\tau_j n}{8Mc_1}\big)\big),\\
&\le 2\exp(-2\log n - \log q) = 2n^{-2}q^{-1},
\end{aligned} \tag{30}$$

by the choice of γ_j and τ_j, that is, $\gamma_j\tau_j \ge \sqrt{64M^2c_1^2(\log q + \log n)/n}$.

For the second quantity in (29), we bound $\|\boldsymbol{X}^\top(\varphi_j(\boldsymbol{X}\boldsymbol{V}^0_{\bullet j}) - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}^0_{\bullet j}))/n\|_\infty$. Towards this end, note that $V^0_{kj} = \widehat{V}^0_{kj} = 0$ on $k \notin S^0_j$. Therefore, for $\boldsymbol{V}^0_{\bullet j}$ and $\widehat{\boldsymbol{V}}^0_{\bullet j}$ constrained on the set S^0_j, $\boldsymbol{V}^0_{\bullet j} = \boldsymbol{V}^0_{S^0_j, j}$ and $\widehat{\boldsymbol{V}}^0_{\bullet j} = \widehat{\boldsymbol{V}}^0_{S^0_j, j}$. Then, by Assumption 3,

$$\|\boldsymbol{X}^\top(\varphi_j(\boldsymbol{X}\boldsymbol{V}^0_{\bullet j}) - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}^0_{\bullet j}))\|_\infty \le L_1\|\boldsymbol{X}^\top\boldsymbol{X}_{S^0_j}(\boldsymbol{V}^0_{S^0_j, j} - \widehat{\boldsymbol{V}}^0_{S^0_j, j})\|_\infty, \tag{31}$$

for some Lipschitz constant $L_1 > 0$. Moreover, by Lemma 7, for the oracle estimator constrained on S^0_j, namely, $\hat{\boldsymbol{V}}^0_{S^0_j, j}$, $\widehat{\boldsymbol{V}}^0_{S^0_j, j} - \boldsymbol{V}^0_{S^0_j, j} = (\boldsymbol{X}^\top_{S^0_j}\boldsymbol{M}\boldsymbol{X}_{S^0_j})^{-1}\boldsymbol{X}^\top_{S^0_j}(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0 - \boldsymbol{r})$, where $\boldsymbol{M}$, $\boldsymbol{\zeta}^0$ and $\boldsymbol{r}$ will be defined in Lemma 7. Let $\boldsymbol{K} = \boldsymbol{X}^\top\boldsymbol{X}_{S^0_j}(\boldsymbol{X}^\top_{S^0_j}\boldsymbol{M}\boldsymbol{X}_{S^0_j})^{-1}\boldsymbol{X}^\top_{S^0_j}$. Plugging the above expression into (31) yields that

$$\begin{aligned}
&\mathbb{P}\big(\|\boldsymbol{X}^\top(\varphi_j(\boldsymbol{X}\boldsymbol{V}^0_{\bullet j}) - \varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}^0_{\bullet j}))/n\|_\infty > 0.25\gamma_j\tau_j\big) \le \mathbb{P}\big(\|\boldsymbol{K}(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0 - \boldsymbol{r})/n\|_\infty > \frac{\gamma_j\tau_j}{4L_1}\big)\\
&\le \mathbb{P}\big(\|\boldsymbol{K}(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0)/n\|_\infty > \frac{\gamma_j\tau_j}{4L_1} - \|\boldsymbol{K}\boldsymbol{r}/n\|_\infty\big).
\end{aligned} \tag{32}$$

By Assumption 5, there exists a constant $c_3 > 0$ such that $\|\boldsymbol{X}^\top\boldsymbol{X}_{S^0_j}(\boldsymbol{X}^\top_{S^0_j}\boldsymbol{M}\boldsymbol{X}_{S^0_j})^{-1}\boldsymbol{X}^\top_{S^0_j}\|_\infty \le c_3$ or $\max_{l=1}^q |K_{li}| \le c_3$. Then, by Lemmas 7 and 8 with the choice of τ_j and γ_j,

$$\begin{aligned}
\|\boldsymbol{K}\boldsymbol{r}/n\|_\infty &= \max_{l=1}^q |\sum_{i=1}^n K_{li}r_i|/n \le \max_l \sum_{i=1}^n |K_{li}||r_i|/n \le c_3\sum_{i=1}^n |r_i|/n\\
&\le c_3L_2(\hat{\boldsymbol{V}}^0_{S^0_j, j} - \boldsymbol{V}^0_{S^0_j, j})^\top(\boldsymbol{X}^\top_{S^0_j}\boldsymbol{X}_{S^0_j}/n)(\hat{\boldsymbol{V}}^0_{S^0_j, j} - \boldsymbol{V}^0_{S^0_j, j})\\
&\le c_3c_{\max}L_2\|\hat{\boldsymbol{V}}^0_{S^0_j, j} - \boldsymbol{V}^0_{S^0_j, j}\|_2^2 \le c_3c_{\max}L_2\frac{16M^2c_1^2}{m^2}\cdot\frac{K^0_j\log(nK^0_j)}{n} \le \frac{1}{2}\cdot\frac{\gamma_j\tau_j}{4L_1},
\end{aligned}$$

with probability at least $1 - 2\exp(-\log(K^0_j) - 2\log n) = 1 - 2(K^0_j)^{-1}n^{-2}$.

Next, in (32), we bound $\boldsymbol{K}(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0)/n$. Note that $\max_{k=1}^q |K_{ki}| \le c_3$. By Assumption 3, $Y_{ij} - \zeta^0_{ij} = (Y_{ij} - \mathbb{E}[Y_{ij}|\boldsymbol{X}])$ is sub-exponential with the bound M. By Theorem 2.8.2 of Vershynin

(2018) and a union bound as in (30),

$$\begin{aligned}
&\mathbb{P}\Big(\|\boldsymbol{X}^\top(\varphi_j(\boldsymbol{X}\boldsymbol{V}^0_{\bullet j})-\varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}^0_{\bullet j}))/n\|_\infty>0.25\gamma_j\tau_j\Big)\le\mathbb{P}\Big(\|\boldsymbol{K}(\boldsymbol{Y}_j-\boldsymbol{\zeta}^0)/n\|_\infty>0.5\frac{\gamma_j\tau_j}{4L_1}\Big)\\
&\le 2q\exp\Big(-\min\big(\frac{\gamma_j^2\tau_j^2 n}{8M^2c_3^2\cdot 16L_1^2},\frac{\gamma_j\tau_j n}{16Mc_3\cdot L_1}\big)\Big)+2(K_j^0)^{-1}n^{-2}\\
&\le 2q\exp\Big(-\frac{\gamma_j^2\tau_j^2 n}{8M^2c_3^2\cdot 16L_1^2}\Big)+2(K_j^0)^{-1}n^{-2}. \qquad (33)
\end{aligned}$$

Combining (29), (30), and (33) yields that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\boldsymbol{X}^\top\widehat{\boldsymbol{\xi}}_j/n\right\|_\infty>0.5\gamma_j\tau_j\right)=\mathbb{P}\left(\left\|\boldsymbol{X}^\top(\boldsymbol{Y}_j-\varphi_j(\boldsymbol{X}\widehat{\boldsymbol{V}}^0_{\bullet j}))/n\right\|_\infty>0.5\gamma_j\tau_j\right)\\
&\le 2q\exp\left(-\frac{\gamma_j^2\tau_j^2 n}{32M^2c_1^2}\right)+2q\exp\left(-\frac{\gamma_j^2\tau_j^2 n}{8M^2c_3^2\cdot 16L_1^2}\right)+2(K_j^0)^{-1}n^{-2}\\
&\le 2n^{-2}q^{-1}+2n^{-2}q^{-1}+2(K_j^0)^{-1}n^{-2}\le 6n^{-2}q^{-1}.
\end{aligned}$$

Next, we bound the second event $\{\|\widehat{\boldsymbol{V}}^0_{\bullet j}-\boldsymbol{V}^0_{\bullet j}\|_\infty\le 0.5\tau_j\}$ in $\mathcal{E}_j^c$. Since $\widehat{\boldsymbol{V}}^0_{\bullet j}=\boldsymbol{V}^0_{\bullet j}=0$ on $(S_j^0)^c$, it suffices to consider the entries of $\widehat{\boldsymbol{V}}^0_{\bullet j}$ constrained on S_j^0, or the oracle estimator $\widehat{\boldsymbol{V}}^0_{S_j^0,j}$. By Lemma 7, $\widehat{\boldsymbol{V}}^0_{S_j^0,j}-\boldsymbol{V}^0_{S_j^0,j}=\boldsymbol{H}(\boldsymbol{Y}_j-\boldsymbol{\zeta}^0-\boldsymbol{r})$, where $\boldsymbol{H}=\big(\boldsymbol{X}^\top_{S_j^0}\boldsymbol{M}\boldsymbol{X}_{S_j^0}\big)^{-1}\boldsymbol{X}^\top_{S_j^0}=(H_{ki})$.

To bound $\|\boldsymbol{H}\boldsymbol{r}\|_\infty$, by Assumption 5, $\|(\boldsymbol{X}^\top_{S_j^0}\boldsymbol{M}\boldsymbol{X}_{S_j^0}/n)^{-1}\boldsymbol{X}^\top_{S_j^0}\|_\infty\le c_2$ or $\|\boldsymbol{H}\|_\infty\le c_2n^{-1}$, for some constant $c_2>0$. By Lemmas 7 and 8 with the choice of τ_j,

$$\begin{aligned}
\|\boldsymbol{H}\boldsymbol{r}\|_\infty&=\max_k|\boldsymbol{H}_{k\bullet}\boldsymbol{r}|=\max_k|\sum_{i=1}^n H_{ki}r_i|\le\max_k\sum_{i=1}^n|H_{ki}||r_i|\le c_2\sum_{i=1}^n|r_i|/n\\
&\le c_2L_2(\widehat{\boldsymbol{V}}^0_{S_j^0,j}-\boldsymbol{V}^0_{S_j^0,j})^\top(\boldsymbol{X}^\top_{S_j^0}\boldsymbol{X}_{S_j^0}/n)(\widehat{\boldsymbol{V}}^0_{S_j^0,j}-\boldsymbol{V}^0_{S_j^0,j})\\
&\le c_2c_{\max}L_2\|\widehat{\boldsymbol{V}}^0_{S_j^0,j}-\boldsymbol{V}^0_{S_j^0,j}\|_2^2\le c_2c_{\max}L_2\frac{16M^2c_1^2}{m^2}\frac{K_j^0\log(nK_j^0)}{n}\le 0.25\tau_j. \qquad (34)
\end{aligned}$$

To bound $\|\boldsymbol{H}(\boldsymbol{Y}_j-\boldsymbol{\zeta}^0)\|_\infty$, note that $\max_{k=1}^q|H_{ki}|\le c_2/n$. By Assumption 3, $Y_{ij}-\zeta_{ij}^0=(Y_{ij}-\mathbb{E}[Y_{ij}|\boldsymbol{X}])$ is sub-exponential with the bound M. By the triangular inequality, Theorem 2.8.2 of Vershynin (2018) and the same argument as in (30), we obtain that

$$\begin{aligned}
&\mathbb{P}\big(\|\widehat{\boldsymbol{V}}^0_{\bullet j}-\boldsymbol{V}^0_{\bullet j}\|_\infty>0.5\tau_j\big)\le\mathbb{P}\big(\boldsymbol{H}(\boldsymbol{Y}_j-\boldsymbol{\zeta}^0)\|_\infty>0.5\tau_j-\|\boldsymbol{H}\boldsymbol{r}\|_\infty\big)\\
&\le\mathbb{P}\big(\|\boldsymbol{H}(\boldsymbol{Y}_j-\boldsymbol{\zeta}^0)\|_\infty>0.25\tau_j\big)\le\mathbb{P}\big(\|\boldsymbol{H}\boldsymbol{\xi}_j\|_\infty>0.25\tau_j\big)\\
&\le 2K_j^0\exp\big(-\min\big(\frac{\tau_j^2 n}{32M^2c_2^2},\frac{\tau_j n}{8Mc_2}\big)\big)+2(K_j^0)^{-1}n^{-2}.
\end{aligned}$$

To conclude, on $\mathcal{E}_j$, $\widehat{S}_j\equiv S_j^{[T]}=S_j^0$, which means $\widehat{\boldsymbol{V}}_{\bullet j}=\widetilde{\boldsymbol{V}}^{[T]}_{\bullet j}=\widehat{\boldsymbol{V}}^0_{\bullet j}$. Hence, for $j=1,\dots,p$,

$$\begin{aligned}
&\mathbb{P}(\widehat{\boldsymbol{V}}_{\bullet j}\neq\widehat{\boldsymbol{V}}^0_{\bullet j})\le\mathbb{P}(\mathcal{E}_j^c)\le 2q\exp\big(-\frac{\gamma_j^2\tau_j^2 n}{32M^2c_1^2}\big)+2q\exp\big(-\frac{\gamma_j^2\tau_j^2 n}{8M^2c_3^2\cdot 16L_1^2}\big)\\
&+2K_j^0\exp\big(-\frac{\tau_j^2 n}{32M^2c_2^2}\big)+2(K_j^0)^{-1}n^{-2}\le 8n^{-2}q^{-1}. \qquad (35)
\end{aligned}$$

It remains to show that $\widehat{\boldsymbol{V}}^0_{\bullet j}$ is a global minimizer of (5) with high probability. Towards this end, we will show that Assumptions 4 and 6 imply the degree of separation condition (3) of Shen et al. (2013). To see this, let $g(y_{ij}|\theta, \boldsymbol{x}_i) = e^{-\ell(y_{ij}, \theta^\top \boldsymbol{x}_i)}$ be a probability density for y_{ij} where we denote $\theta = \boldsymbol{V}_{\bullet j}$ for notation simplicity. In addition, denote $\theta^0 = (\boldsymbol{V}_{S^0_j, j}, \boldsymbol{0})$ and $\theta_{A_j} = (\boldsymbol{0}, \boldsymbol{V}_{A_j, j})$. By the mean value theorem, there exists $\overline{\theta}_0$ between θ_{A_j} and θ^0 such that

$$\begin{aligned}
&\left(g^{1/2}(y_{ij}|\theta_{A_j}, \boldsymbol{x}_i) - g^{1/2}(y_{ij}|\theta^0, \boldsymbol{x}_i)\right)^2 = \left(\left(\nabla g^{1/2}(y_{ij}|\overline{\theta}_0, \boldsymbol{x}_i)\right)^\top (\theta_{A_j} - \theta^0)\right)^2 \\
&= \left((\nabla e^{-\ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)/2})^\top (\theta_{A_j} - \theta^0)\right)^2 = \frac{1}{4} e^{-\ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)} \left(\nabla \ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)^\top (\theta_{A_j} - \theta^0)\right)^2.
\end{aligned}$$

Then, the Hellinger distance can be written as:

$$\begin{aligned}
&h^2\left(\theta_{A_j}, \theta^0\right) = \frac{1}{4}\left(\int \left(g^{1/2}(y_{ij}|\theta_{A_j}, \boldsymbol{x}_i) - g^{1/2}(y_{ij}|\theta^0, \boldsymbol{x}_i)\right)^2 d\mu(y_{ij})\right) \\
&= \frac{1}{16}\left(\int e^{-\ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)} (\theta_{A_j} - \theta^0)^\top \nabla \ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i) \nabla \ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)^\top (\theta_{A_j} - \theta^0) d\mu(y_{ij})\right) \\
&= \frac{1}{16}(\theta_{A_j} - \theta^0)^\top \left(\int e^{-\ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)} \cdot \nabla \ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i) \nabla \ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)^\top d\mu(y_{ij})\right) (\theta_{A_j} - \theta^0) \\
&= \frac{1}{16}(\theta_{A_j} - \theta^0)^\top \mathbb{E}_{\overline{\theta}_0}\left[\nabla^2 \ell(y'_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)\right] (\theta_{A_j} - \theta^0),
\end{aligned}$$

where $\mathbb{E}_{\overline{\theta}_0}$ is the expectation with respect to $Y'_{ij} \sim g(y'_{ij}|\overline{\theta}_0^\top, \boldsymbol{x}_i)$ while the last equality follows by the fact that $\mathbb{E}_\theta\left[\nabla \log g(y_{ij}|\theta, \boldsymbol{x}_i) \nabla \log g(y_{ij}|\theta, \boldsymbol{x}_i)^\top\right] = -\mathbb{E}_\theta\left[\nabla^2 \log g(y_{ij}|\theta, \boldsymbol{x}_i)\right]$ for any θ.

Let $\widetilde{\theta} = \theta_{A_j} - \theta^0$. Then, $\|\widetilde{\theta}\|_2^2 \geq |S^0_j \setminus A_j| \|\boldsymbol{V}_{S^0_j, j}\|_2^2$. By the definition of $C_{\min}$ of Shen et al. (2013) and Assumption 4,

$$\begin{aligned}
C_{\min} &= \min_{A_j \neq S^0_j, |A_j| \leq K^0_j} \frac{h^2\left(\theta_{A_j}, \theta^0\right)}{\max(|S^0_j \setminus A_j|, 1)} \\
&\geq \min_{A_j \neq S^0_j, |A_j| \leq K^0_j} |S^0_j \setminus A_j|^{-1} (\theta_{A_j} - \theta^0)^\top \mathbb{E}\left[\nabla^2 \ell(y_{ij}, \overline{\theta}_0^\top \boldsymbol{x}_i)\right] (\theta_{A_j} - \theta^0) \\
&\geq m \|\boldsymbol{V}_{S^0_j, j}\|_2^2 \geq m(100 M c_2)^2 \frac{\log q + \log n}{n} \geq m(100 M c_2)^2 \frac{\log q}{n},
\end{aligned}$$

where the last inequality uses Assumptions 4 and 6 and the fact $\overline{\theta}_0 = \theta_{A_j} + t(\theta_{A_j} - \theta^0)$, $t \in [0, 1]$ so that $\|\overline{\theta}_0\|_0 \leq 2K^0_j$. This implies the degree of separation condition (3) of Shen et al. (2013). By Theorem 2 there, $\mathbb{P}\left(\widehat{\boldsymbol{V}}^0_{\bullet j} \text{ is not a global minimizer of (5)}\right) \leq 3\exp(-2(\log(q) + \log(n))), 1 \leq j \leq p$, implying that

$$\mathbb{P}\left(\widehat{\boldsymbol{V}}^0_{\bullet j} \text{ is not a global minimizer of (5)}, 1 \leq j \leq p\right) \leq 3p\left(\exp(-2(\log(q) + \log(n)))\right).$$

Hence, $\widehat{\boldsymbol{V}}_{\bullet j}$ is a global minimizer of (5) with probability tending to 1 as $n \to \infty$; note that $q \geq p$ by Assumption 1(B). Finally, we have shown that $\widehat{S}_j \equiv \{l : \widehat{V}_{lj} \neq 0\} = S^0_j \equiv \{l : V^0_{lj} \neq 0\}$, implying $\{(l, j) : \widehat{V}_{lj} \neq 0\} = \{(l, j) : V^0_{lj} \neq 0\}$. By Proposition 3, the estimated $\widehat{\mathcal{S}}$ via $\widehat{\boldsymbol{V}}$ reconstructs the true super-graph $\mathcal{S}^0$ correctly. This completes the proof. We next present proofs of the lemmas.

Lemma 7 (Expression of the oracle MLE $\widehat{\boldsymbol{V}}^0_{S_j^0,j}$) *Let $\widehat{\boldsymbol{V}}^0_{S_j^0,j}$ be the oracle MLE, defined as the minimizer of $\mathcal{L}(\boldsymbol{V}_{S_j^0,j}|\boldsymbol{Y}_j,\boldsymbol{X}_{S_j^0}) = n^{-1}\sum_{i=1}^n \left(-Y_{ij}(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}_{S_j^0,j}) + A_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}_{S_j^0,j})\right)$ over $\boldsymbol{V}_{S_j^0,j}$. Then,*

$$\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j} = (\boldsymbol{X}_{S_j^0}^\top \boldsymbol{M}\boldsymbol{X}_{S_j^0})^{-1}\boldsymbol{X}_{S_j^0}^\top(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0 - \boldsymbol{r}), \tag{36}$$

where $\boldsymbol{\zeta}^0 = (\zeta_1^0,\ldots,\zeta_n^0)$ with $\zeta_i^0 = \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}^0_{S_j^0,j})$, $\boldsymbol{M}$ is a diagonal matrix with the ith diagonal $\boldsymbol{M}_{ii} = A_j''(\boldsymbol{X}_{i\bullet}^\top \boldsymbol{V}^0_{\bullet j})$ and $\boldsymbol{r} = (r_1,\cdots,r_p)$ is the integral form of the reminder for Taylor's expansion satisfying

$$\sum_{i=1}^n |r_i| \le L_2(\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j})^\top(\sum_{i=1}^n \boldsymbol{x}_{i,S_j^0}\boldsymbol{x}_{i,S_j^0}^\top)(\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}), \tag{37}$$

for L_2 defined in Assumption 3.

Proof of Lemma 7. By the optimality condition for the constrained oracle MLE, for $k \in S_j^0$,

$$\begin{aligned}
&\sum_{i=1}^n X_{ik}(y_{ij} - \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \widehat{\boldsymbol{V}}^0_{S_j^0,j})) = 0,\\
&\sum_{i=1}^n X_{ik}(y_{ij} - \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}^0_{S_j^0,j}) + \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}^0_{S_j^0,j}) - \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \widehat{\boldsymbol{V}}^0_{S_j^0,j})) = 0.
\end{aligned} \tag{38}$$

A Taylor series expansion of $\varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}_{S_j^0,j})$ at $\boldsymbol{V}^0_{S_j^0,j}$ yields that

$$\varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \widehat{\boldsymbol{V}}^0_{S_j^0,j}) = \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}^0_{S_j^0,j}) + w_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}^0_{S_j^0,j})\boldsymbol{x}_{i,S_j^0}^\top(\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}) + r_i, \tag{39}$$

where

$$r_i = \int_0^1 \varphi_j''\big(\boldsymbol{x}_{i,S_j^0}^\top(\widehat{\boldsymbol{V}}^0_{S_j^0,j} + t(\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}))\big)(1-t)\,dt\,\left((\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j})^\top \boldsymbol{x}_{i,S_j^0}\boldsymbol{x}_{i,S_j^0}^\top(\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j})\right),$$

is the integral form of the remainder for Taylor's expansion as Li and Lederer (2019) and $w_j(\boldsymbol{x}^\top\boldsymbol{u}) = \varphi_j'(\boldsymbol{x}^\top\boldsymbol{u}) = A_j''(\boldsymbol{x}^\top\boldsymbol{u})$. Let $\boldsymbol{M}$ be a diagonal matrix whose ith diagonal $\boldsymbol{M}_{ii} = A_j''({\boldsymbol{V}^0_{\bullet j}}^\top \boldsymbol{X}_{i\bullet})$. Write (38) in a matrix form using (39):

$$\begin{aligned}
\boldsymbol{X}_{S_j^0}^\top \boldsymbol{M}\boldsymbol{X}_{S_j^0}(\hat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}) &= \boldsymbol{X}_{S_j^0}^\top(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0 - \boldsymbol{r})\\
\hat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j} &= (\boldsymbol{X}_{S_j^0}^\top \boldsymbol{M}\boldsymbol{X}_{S_j^0})^{-1}\boldsymbol{X}_{S_j^0}^\top(\boldsymbol{Y}_j - \boldsymbol{\zeta}^0 - \boldsymbol{r}),
\end{aligned} \tag{40}$$

where $\zeta_i^0 = \varphi_j(\boldsymbol{x}_{i,S_j^0}^\top \boldsymbol{V}^0_{S_j^0,j})$. Further, note that $\varphi_j(\boldsymbol{x}^\top\boldsymbol{u}) = A_j'(\boldsymbol{x}^\top\boldsymbol{u})$. Then, in the expression for r_i, $\varphi_j''\big(\boldsymbol{x}_{i,S_j^0}^\top(\widehat{\boldsymbol{V}}^0_{S_j^0,j} + t(\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}))\big) = g_i(t)$, where $g_i(t) = \varphi_j''(\eta_i(t)) = A_j'''(\eta_i(t))$ with $\eta_i(t) = \boldsymbol{x}_{i,S_j^0}^\top(\hat{\boldsymbol{V}}^0_{S_j^0,j} + t(\hat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}))$, $t \in (0,1)$. By Assumption 3, $|g_i(t)| \le L_2$. Hence, (37) holds.

Lemma 8 (Rate of convergence under the ℓ_2-norm) *Under Assumption 4 (restricted strong convexity),*

$$\|\widehat{\boldsymbol{V}}^0_{S_j^0,j} - \boldsymbol{V}^0_{S_j^0,j}\|_2 \le \frac{2}{m}\sqrt{K_j^0}\big\|\boldsymbol{X}_{S_j^0}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{S_j^0}\boldsymbol{V}^0_{S_j^0,j}))/n\big\|_\infty \le \frac{4Mc_1}{m}\sqrt{\frac{K_j^0\log(nK_j^0)}{n}},$$

with probability at least $1 - 2\exp(-\log(K_j^0) - 2\log n) = 1 - 2(K_j^0)^{-1}n^{-2}$.

Proof of Lemma 8. We follow the proof of Lee et al. (2015) and consider the entries of $\boldsymbol{V}^0_{\bullet j}$ on S^0_j. The negative log-likelihood is

$$\mathcal{L}(\boldsymbol{V}_{S^0_j,j}|\boldsymbol{Y}_j,\boldsymbol{X}_{S^0_j}) = n^{-1}\sum_{i=1}^n \Big(-Y_{ij}(\boldsymbol{x}^\top_{i,S^0_j}\boldsymbol{V}_{S^0_j,j}) + A_j(\boldsymbol{x}^\top_{i,S^0_j}\boldsymbol{V}_{S^0_j,j})\Big),$$

where $\boldsymbol{x}_{i,S^0_j}$ is a subvector of $\boldsymbol{x}_i$ with elements constrained on the indices S^0_j. By the definition of the oracle MLE, $\mathcal{L}(\widehat{\boldsymbol{V}}^0_{S^0_j,j}) \le \mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})$. Taylor's expansion of $\mathcal{L}(\cdot)$ at $\boldsymbol{V}^0_{S^0_j,j}$ yields that

$$0 \ge \nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}) + \frac{1}{2}(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j})^\top\nabla^2\mathcal{L}(\overline{\boldsymbol{V}}^0_{S^0_j,j})(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}).$$

Let $\Delta = \widehat{\boldsymbol{V}}^0_{\bullet j} - \boldsymbol{V}^0_{\bullet j}$. Clearly, $0 = \|\Delta_{(S^0_j)^c}\|_1 \le 3\|\Delta_{S^0_j}\|_1$. Therefore, by the restricted strong convexity condition $\frac{1}{2}(\widehat{\boldsymbol{V}}^0_{\bullet j} - \boldsymbol{V}^0_{\bullet j})^\top\nabla^2\mathcal{L}(\overline{\boldsymbol{V}}^0_{\bullet j})(\widehat{\boldsymbol{V}}^0_{\bullet j} - \boldsymbol{V}^0_{\bullet j}) \ge \frac{m}{2}\|\widehat{\boldsymbol{V}}^0_{\bullet j} - \boldsymbol{V}^0_{\bullet j}\|_2^2$, which implies $\frac{1}{2}(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j})^\top\nabla^2\mathcal{L}(\overline{\boldsymbol{V}}^0_{S^0_j,j})(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}) \ge \frac{m}{2}\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2^2$ as $\widehat{\boldsymbol{V}}^0_{(S^0_j)^c,j} = \boldsymbol{V}^0_{(S^0_j)^c,j} = \boldsymbol{0}$.

By the restricted strong convexity condition,

$$\nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}) + \frac{m}{2}\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2^2 \le 0.$$

By the Hölder's inequality,

$$\frac{m}{2}\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2^2 \le -\nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})(\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}) \le \|\nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})\|_\infty\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_1.$$

By the Cauchy–Schwarz inequality,

$$\frac{m}{2}\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2^2 \le \sqrt{K^0_j}\|\nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})\|_\infty\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2,$$
$$\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2 \le \frac{2}{m}\sqrt{K^0_j}\|\nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j})\|_\infty,$$

where $\nabla\mathcal{L}(\boldsymbol{V}^0_{S^0_j,j}) = n^{-1}\boldsymbol{X}^\top_{S^0_j}(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{S^0_j}\boldsymbol{V}^0_{S^0_j,j}))$. Therefore,

$$\mathbb{P}\big(\big\|\boldsymbol{X}^\top_{S^0_j}(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{S^0_j}\boldsymbol{V}^0_{S^0_j,j}))/n\big\|_\infty > \epsilon\big) \le 2K^0_j\exp\big(-\min\big(\frac{n\epsilon^2}{2M^2c_1^2}, \frac{n\epsilon}{2Mc_1}\big)\big).$$

Hence, $\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2 \le \frac{2}{m}\sqrt{K^0_j}\epsilon = \frac{4Mc_1}{m}\sqrt{\frac{K^0_j\log(nK^0_j)}{n}}$, with probability $1 - 2K^0_j\exp\big(-\min\big(\frac{n\epsilon^2}{2M^2c_1^2}, \frac{n\epsilon}{2Mc_1}\big)\big) \ge 1 - 2\exp(-\log(K^0_j) - 2\log n) = 1 - 2(K^0_j)^{-1}n^{-2}$, $\|\widehat{\boldsymbol{V}}^0_{S^0_j,j} - \boldsymbol{V}^0_{S^0_j,j}\|_2 \le \frac{2}{m}\sqrt{K^0_j}\epsilon = \frac{4Mc_1}{m}\sqrt{\frac{K^0_j\log(nK^0_j)}{n}}$, where $\epsilon = 2Mc_1\sqrt{\frac{\log(nK^0_j)}{n}}$.

D.5 Proof of Theorem 5

For the jth equation, the TLP estimator minimizes:

$$\begin{aligned}
&(\widehat{\boldsymbol{W}}_{\overline{\text{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\text{an}}(j),j}, \widehat{\boldsymbol{\alpha}}_{\overline{\text{an}}(j),j})\\
&= \underset{\boldsymbol{W}_{\overline{\text{in}}(j),j},\boldsymbol{U}_{\overline{\text{an}}(j),j},\boldsymbol{\alpha}_{\overline{\text{an}}(j),j}}{\operatorname{argmin}} \; n^{-1}\sum_{i=1}^n -Y_{ij}\Big(\boldsymbol{W}^\top_{\overline{\text{in}}(j),j}\boldsymbol{X}_{i,\overline{\text{in}}(j)} + \boldsymbol{U}^\top_{\overline{\text{an}}(j),j}\boldsymbol{Y}_{i,\overline{\text{an}}(j)} + \boldsymbol{\alpha}^\top_{\overline{\text{an}}(j),j}\widehat{\boldsymbol{h}}_{i,\overline{\text{an}}(j)}\Big)\\
&\qquad\qquad + A_j\Big(\boldsymbol{W}^\top_{\overline{\text{in}}(j),j}\boldsymbol{X}_{i,\overline{\text{in}}(j)} + \boldsymbol{U}^\top_{\overline{\text{an}}(j),j}\boldsymbol{Y}_{i,\overline{\text{an}}(j)} + \boldsymbol{\alpha}^\top_{\overline{\text{an}}(j),j}\widehat{\boldsymbol{h}}_{i,\overline{\text{an}}(j)}\Big)\\
&\text{subject to} \quad \sum_{k\in\overline{\text{an}}(j)} I(U_{kj}\neq 0) \le K_j, \quad \sum_{k\in\overline{\text{an}}(j)} I(\alpha_{kj}\neq 0) \le K'_j, \quad j = 1,\ldots,p.
\end{aligned}$$

In the absence of confounders ($\boldsymbol{h}_{i,\mathrm{an}(j)} = 0$), if we use standard constrained GLM regression without deconfounding, ($\widehat{\boldsymbol{h}}_{i,\overline{\mathrm{an}}(j)} = 0$), it is straightforward to show that $\widehat{\boldsymbol{U}}_{\overline{\mathrm{an}}(j),j} \to \boldsymbol{U}^0_{\mathrm{an}^0(j),j}$ and $\widehat{\boldsymbol{W}}_{\overline{\mathrm{in}}(j),j} \to \boldsymbol{W}^0_{\mathrm{in}^0(j),j}$ by standard high-dimensional statistics results.

We now show the causal graph selection consistency of the TLP estimator in the presence of the confounders. We follow the same proof procedure of Theorem 4. Denote the oracle M-estimator $\widehat{\boldsymbol{\theta}}^{ml} = (\widehat{\boldsymbol{W}}^{ml}_{\overline{\mathrm{in}}(j),j}, \widehat{\boldsymbol{U}}^{ml}_{\overline{\mathrm{an}}(j),j}, \widehat{\boldsymbol{\alpha}}^{ml}_{\overline{\mathrm{an}}(j),j})) = \operatorname{argmin} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{Y}_{\overline{\mathrm{an}}(j)}, \boldsymbol{X}_{\overline{\mathrm{in}}(j)}, \widehat{\boldsymbol{h}}_{\overline{\mathrm{an}}(j)})$ such that $\{k : \widehat{U}^{ml}_{kj} \neq 0\} = \{k : U^0_{kj} \neq 0\} = \mathrm{pa}^0(j)$, $\{l : \widehat{W}^{ml}_{lj} \neq 0\} = \{l : W^0_{lj} \neq 0\} = \mathrm{in}^0(j)$ and $\{k : \widehat{\alpha}^{ml}_{kj} \neq 0\} = \{k : \alpha^0_{kj} \neq 0\}$. Further, denote A^0_j as the set of non-zero indices of the concatenated vector $\boldsymbol{\theta}^0 = (\boldsymbol{W}^0_{\mathrm{in}^0(j),j}, \boldsymbol{U}^0_{\mathrm{an}^0(j),j}, \boldsymbol{\alpha}^0_{\mathrm{an}^0(j),j})$. Therefore, $\boldsymbol{\theta}^0_{A^0_j} = (\boldsymbol{W}^0_{\mathrm{in}^0(j),j}, \boldsymbol{U}^0_{\mathrm{pa}^0(j),j}, \boldsymbol{\alpha}^0_{\mathrm{an}^0(j),j})$ and $\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} = (\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}, \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}, \widehat{\boldsymbol{\alpha}}^{ml}_{\mathrm{an}^0(j),j})$; $\mathrm{supp}(\boldsymbol{\theta}^0) = \mathrm{supp}(\widehat{\boldsymbol{\theta}}^{ml}) = A^0_j$. By the proof of Theorem 4, it suffices to bound the event $\left\{\left\|\widehat{\boldsymbol{\theta}}^{ml} - \boldsymbol{\theta}^0\right\|_\infty \leq 0.5\tau_j\right\}$, or equivalently, $\left\{\left\|\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\right\|_\infty \leq 0.5\tau_j\right\}$, as $\widehat{U}^{ml}_{kj} = U^0_{kj} = 0$ on $k \in (A^0_j)^c$. Alternatively, by Proposition 1 of Shen et al. (2012),

$$P\left(\widehat{\boldsymbol{\theta}} \neq \widehat{\boldsymbol{\theta}}^{ml}\right) \leq \exp\left(-c_2 n C_{\min}(\boldsymbol{\theta}^0) + 2\log(p+1) + 3\right),$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{W}}_{\overline{\mathrm{in}}(j),j}, \widehat{\boldsymbol{U}}_{\overline{\mathrm{an}}(j),j}, \widehat{\boldsymbol{\alpha}}_{\overline{\mathrm{an}}(j),j})$ is the final TLP estimator at iteration T, i.e., $\widehat{\boldsymbol{\theta}}^{[T]}$. Therefore, $\left\|\widehat{\boldsymbol{\theta}}^{ml} - \boldsymbol{\theta}^0\right\|_\infty \leq 0.5\tau_j$ implies that $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\|_\infty \leq 0.5\tau_j$.

For the root equations, note that the confounder $\boldsymbol{h}_k$ is independent of the instrumental variable $\boldsymbol{X}_{\mathrm{in}^0(k)}$. Hence, the confounders do not interfere with the estimation of the coefficient $\boldsymbol{W}_{\mathrm{in}(k),k}$. By the standard GLM result, $\|\boldsymbol{W}^0_{\mathrm{in}^0(k),k} - \widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(k),k}\|_\infty \propto \sqrt{\frac{\log(p\widetilde{s})}{n}}$. We prove the error bound in detail in Lemma 11.

For the child equations, let $\boldsymbol{\theta}_{A^0_j} = (\boldsymbol{W}_{\mathrm{in}^0(j),j}, \boldsymbol{U}_{\mathrm{pa}^0(j),j}, \boldsymbol{\alpha}_{\mathrm{an}^0(j),j})$ and $\widetilde{\boldsymbol{Z}} = [\boldsymbol{X}_{\mathrm{in}^0(j)}, \boldsymbol{Y}_{\mathrm{pa}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)}]$. Let $s = \max_{1\leq j\leq p} \|\boldsymbol{U}^0_{\bullet j}\|_0$ and $\widetilde{s} = \max_{1\leq j\leq p} \|\boldsymbol{W}^0_{\bullet j}\|_0$. The log-likelihood that $\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} = (\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}, \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}, \widehat{\boldsymbol{\alpha}}^{ml}_{\mathrm{an}^0(j),j})$ minimizes is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_{A^0_j}|\widetilde{\boldsymbol{Z}}) &= \mathcal{L}(\boldsymbol{W}_{\mathrm{in}^0(j),j}, \boldsymbol{U}_{\mathrm{pa}^0(j),j}, \boldsymbol{\alpha}_{\mathrm{an}^0(j),j}|\boldsymbol{X}_{\mathrm{in}^0(j)}, \boldsymbol{Y}_{\mathrm{pa}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)}) \\
&= \frac{1}{n}\sum_{i=1}^n -Y_{ij}\left(\boldsymbol{W}^\top_{\mathrm{in}^0(j),j}\boldsymbol{X}_{i,\mathrm{in}^0(j)} + \boldsymbol{U}^\top_{\mathrm{pa}^0(j),j}\boldsymbol{Y}_{i,\mathrm{pa}^0(j)} + \boldsymbol{\alpha}^\top_{\mathrm{an}^0(j),j}\widehat{\boldsymbol{h}}_{i,\mathrm{an}^0(j)}\right) \\
&\quad + A_j\left(\boldsymbol{W}^\top_{\mathrm{in}^0(j),j}\boldsymbol{X}_{i,\mathrm{in}^0(j)} + \boldsymbol{U}^\top_{\mathrm{pa}^0(j),j}\boldsymbol{Y}_{i,\mathrm{pa}^0(j)} + \boldsymbol{\alpha}^\top_{\mathrm{an}^0(j),j}\widehat{\boldsymbol{h}}_{i,\mathrm{an}^0(j)}\right).
\end{aligned}$$

Since $\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} = (\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}, \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}, \widehat{\boldsymbol{\alpha}}^{ml}_{\mathrm{an}^0(j),j})$ minimizes $\mathcal{L}(\boldsymbol{\theta}_{A^0_j}|\boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{X}_{\mathrm{in}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)})$, by the KKT condition for the oracle MLE constrained on the true set:

$$\begin{aligned}
&\sum_{i=1}^n \widetilde{Z}_{ik}(y_{ij} - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j})) = 0, \\
&\sum_{i=1}^n \widetilde{Z}_{ik}(y_{ij} - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) + \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j})) = 0. \qquad (41)
\end{aligned}$$

As in Lemma 7, applying Taylor series expansion, (41) can be written in matrix form:

$$\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M}\widetilde{\boldsymbol{Z}}(\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}) = \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y}_j - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \boldsymbol{r}).$$

Therefore, $\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j} = (\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M} \widetilde{\boldsymbol{Z}})^{-1} \widetilde{\boldsymbol{Z}}^\top (\boldsymbol{Y}_j - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \boldsymbol{r})$. Then we calculate the ℓ_∞-norm of the estimation error:

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\|_\infty &= \|(\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M} \widetilde{\boldsymbol{Z}})^{-1} \widetilde{\boldsymbol{Z}}^\top (\boldsymbol{Y}_j - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \boldsymbol{r})\|_\infty \\
&= \|(\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M} \widetilde{\boldsymbol{Z}})^{-1} \widetilde{\boldsymbol{Z}}^\top (\boldsymbol{Y}_j - \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) + \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \boldsymbol{r})\|_\infty \\
&\le \|\boldsymbol{H}(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty + \|\boldsymbol{H}(\varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty + \|\boldsymbol{H}\boldsymbol{r}\|_\infty,
\end{aligned}$$

where $\boldsymbol{H} = (\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M} \widetilde{\boldsymbol{Z}})^{-1} \widetilde{\boldsymbol{Z}}^\top$. Denote $\boldsymbol{Z} = [\boldsymbol{X}_{\mathrm{in}^0(j)}, \boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{h}_{\mathrm{an}^0(j)}]$ as the true predictor variable. Again, by the bounded domain for interventions condition, there exists b_2 such that $\|n(\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M} \widetilde{\boldsymbol{Z}})^{-1} \widetilde{\boldsymbol{Z}}^\top\|_\infty \le b_2$. Note $\widetilde{\boldsymbol{Z}} = [\boldsymbol{X}_{\mathrm{in}^0(j)}, \boldsymbol{Y}_{\mathrm{pa}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)}] \in \mathbb{R}^{2s+\widetilde{s}}$; $\boldsymbol{h}_{\mathrm{an}^0(j)}$ refers to a submatrix consisting of $\boldsymbol{h}_k, k \in \mathrm{an}^0(j)$ and $\boldsymbol{h}_{\mathrm{an}^0(j)} \boldsymbol{\alpha}^0_{\mathrm{an}^0(j),j} = \sum_{k \in \mathrm{an}^0(j)} \alpha^0_{kj} \boldsymbol{h}_k$.

Since $\mathbb{E}[\boldsymbol{Y}_j | \boldsymbol{Z}] = \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)} \boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)} \boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_{k \in \mathrm{an}^0(j)} \alpha^0_{kj} \boldsymbol{h}_k) = \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j})$, the first term can be bounded by the Bernstein's inequality. That is,

$$\mathbb{P}(\|\boldsymbol{H}(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty > \epsilon) \le 2(2s + \widetilde{s}) \exp\left(- \min\left(\frac{n\epsilon^2}{2M^2 b_2^2}, \frac{n\epsilon}{2Mb_2} \right) \right).$$

Setting $\epsilon = 2Mb_2 \sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}}$ leads to

$$\|\boldsymbol{H}(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty \le 2Mb_2 \sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}},$$

with probability at least $1 - 2\exp(-2\log p - \log(2s+\widetilde{s})) = 1 - 2p^{-2}(2s+\widetilde{s})^{-1}$.

Note that for the second term,

$$\begin{aligned}
\varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) &= (\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j} - \widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) \odot \varphi_j'(\boldsymbol{\xi}) = \sum_k \alpha^0_{kj} (\boldsymbol{h}_k - \widehat{\boldsymbol{h}}_k) \odot \varphi_j'(\boldsymbol{\xi}) \\
&= -\sum_k \alpha^0_{kj} \Big(\Delta_k + (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k | \boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k]) \Big) \odot \varphi_j'(\boldsymbol{\xi}), \quad \text{by (43)},
\end{aligned}$$

where we use the fact that $\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j} = \boldsymbol{X}_{\mathrm{in}^0(j)} \boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)} \boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj} \boldsymbol{h}_k$ and $\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j} = \boldsymbol{X}_{\mathrm{in}^0(j)} \boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)} \boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj} \widehat{\boldsymbol{h}}_k$. Therefore,

$$\|\boldsymbol{H}\big(\varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j})\big)\|_\infty \le \|\boldsymbol{H} \sum_k \alpha^0_{kj} \Delta_k \odot \varphi_j'(\boldsymbol{\xi})\|_\infty + \|\boldsymbol{H} \sum_k \alpha^0_{kj} (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k | \boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k]) \odot \varphi_j'(\boldsymbol{\xi})\|_\infty.$$

Note that $|\varphi_j'(z)| \le L_1$ and $|H_{ij}| \le \frac{b_2}{n}$. The first quantity above is bounded by

$$\begin{aligned}
\|\sum_k \alpha^0_{kj} \Delta_k \odot \varphi_j'(\boldsymbol{\xi})\|_\infty &\le L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty, \\
\|\boldsymbol{H} \sum_k \alpha^0_{kj} \Delta_k \odot \varphi_j'(\boldsymbol{\xi})\|_\infty &\le n \cdot \frac{b_2}{n} \cdot L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty = b_2 L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty.
\end{aligned}$$

For the second quantity, note that $\|\boldsymbol{H} \sum_k \alpha^0_{kj} (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k | \boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k]) \odot \varphi_j'(\boldsymbol{\xi})\|_\infty = \|\sum_k \alpha^0_{kj} \boldsymbol{H}(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k | \boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k]) \odot \varphi_j'(\boldsymbol{\xi})\|_\infty \le L_1 \|\sum_k \alpha^0_{kj} \boldsymbol{H}(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k | \boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k])\|_\infty$. By Assumption 3,

$\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k]$ is sub-exponential and therefore $\sum_k \alpha^0_{kj}\boldsymbol{H}(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k])$ is also sub-exponential. We conclude that $\|\sum_k \alpha^0_{kj}\boldsymbol{H}(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k])\|_\infty = o(\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}})$ with probability tending to 1 by the Bernstein's inequality, and therefore converges to zero with increased sample size. Therefore, $\|\boldsymbol{H}(\varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty \le b_2 L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty$.

Last, for the remainder of Taylor series expansion $\boldsymbol{r}$, similar to Theorem 4 and Lemma 7, $|\boldsymbol{H}_{k\bullet}\boldsymbol{r}| = |\sum_{i=1}^n H_{ki} r_i| \le b_2 \sum_{i=1}^n |r_i|/n \le b_2 D(\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j})^\top (\boldsymbol{Z}^\top \boldsymbol{Z}/n)(\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}) \le b_2 c_0 D \|\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}\|_2^2$. Further, by Lemma 9, $\|\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\|_2 \le \frac{2}{m}\sqrt{2s+\widetilde{s}} \cdot \|\nabla \mathcal{L}(\boldsymbol{\theta}^0_{A^0_j}|\boldsymbol{Y}_{\text{pa}^0(j)}, \boldsymbol{X}_{\text{in}^0(j)}, \widehat{\boldsymbol{h}}_{\text{an}^0(j)})\|_\infty \le \frac{2}{m}\sqrt{2s+\widetilde{s}}\left[\eta_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + b_1 L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty\right]$ with $\eta_1 = 2Mb_1$. Therefore,

$$
\begin{aligned}
&|\boldsymbol{H}_{k\bullet}\boldsymbol{r}| \le b_2 c_0 D \|\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}\|_2^2 \\
&\le b_2 c_0 D \left(\frac{2}{m}\sqrt{2s+\widetilde{s}} \cdot \left(\eta_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + b_1 L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty\right)\right)^2 \\
&\le b_2 c_0 D \left(\left(\frac{2}{m}\sqrt{2s+\widetilde{s}}\right)^2 \cdot 2\left(\eta_1^2 \frac{\log(p(2s+\widetilde{s}))}{n} + \left(b_1 L_1 \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty\right)^2\right)\right) \\
&\le 2 b_2 c_0 D \left(\frac{2}{m}\right)^2 \left(\eta_1^2 (2s+\widetilde{s}) \frac{\log(p(2s+\widetilde{s}))}{n} + (2s+\widetilde{s}) \cdot (b_1 L_1)^2 \cdot s \cdot \sum_k |\alpha^0_{kj}|^2 \cdot \|\Delta_k\|_\infty^2\right) \\
&\le b_2 c_0 D \left(\frac{8}{m^2}\eta_1^2 \cdot \sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + \frac{8}{m^2}(b_1 L_1)^2 \sum_k |\alpha^0_{kj}|^2 \cdot \|\Delta_k\|_\infty\right),
\end{aligned}
$$

where the last inequality holds true as $n > (2s+\widetilde{s})^2 \log(p(2s+\widetilde{s}))$ and $(2s+\widetilde{s})s\|\Delta_k\|_\infty \le (2s+\widetilde{s})s\sqrt{\frac{\log(p\widetilde{s})}{n}} \le 1$. Also, we use the property $(\sum_{i=1}^s a_i)^2 \le s\sum_{i=1}^s a_i^2$. Combining the three terms leads to

$$
\begin{aligned}
\|\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\|_\infty &\le a_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + \sum_k (b_2 L_1 |\alpha^0_{kj}| + \frac{8 b_2 c_0 D b_1^2 L_1^2 |\alpha^0_{kj}|^2}{m^2})\|\Delta_k\|_\infty \\
&\le a_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + \max_k \left(b_2 L_1 |\alpha^0_{kj}| + \frac{8 b_2 c_0 D b_1^2 L_1^2 |\alpha^0_{kj}|^2}{m^2}\right) \cdot \sum_k \|\Delta_k\|_\infty,
\end{aligned}
$$

with probability greater than $1 - 4\exp(-2\log p - \log(2s+\widetilde{s})) = 1 - 4p^{-2}(2s+\widetilde{s})^{-1}$. Here, $a_1 = 2Mb_2 + b_2 c_0 D(\frac{32}{m^2}M^2 b_1^2)$. Denote $\widehat{A}_j$ as the set of non-zero indices of the concatenated vector $\widehat{\boldsymbol{\theta}}$. Similar to the proof of Theorem 4, if τ_j is chosen such that $\tau_j \ge 2\left\|\widehat{\boldsymbol{\theta}}^{ml} - \boldsymbol{\theta}^0\right\|_\infty$, then $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{ml}$ and $\widehat{A}_j = A^0_j$, that is, $\widehat{\text{pa}}(j) = \text{pa}^0(j)$, $\widehat{\text{in}}(j) = \text{in}^0(j)$ and $\widehat{\text{an}}(j) = \text{an}^0(j)$. Additionally, note that $\max(\|\widehat{\boldsymbol{U}}_{\bullet j} - \boldsymbol{U}^0_{\bullet j}\|_\infty, \|\widehat{\boldsymbol{W}}_{\bullet j} - \boldsymbol{W}^0_{\bullet j}\|_\infty) \le \|\widehat{\boldsymbol{\theta}}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\|_\infty$.

We have calculated the parameter estimation errors for the jth equation. Now we calculate the accumulated error for the confounder $\boldsymbol{h}_j$. Note that, by construction, $\widehat{\boldsymbol{h}}_j = \sum_k \widehat{\alpha}_{kj}\widehat{\boldsymbol{h}}_k + \widehat{\boldsymbol{\epsilon}}_j$ where $\widehat{\boldsymbol{\epsilon}}_j = \boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{\widehat{\text{in}}(j)}\widehat{\boldsymbol{W}}_{\widehat{\text{in}}(j),j} + \boldsymbol{Y}_{\widehat{\text{pa}}(j)}\widehat{\boldsymbol{U}}_{\widehat{\text{pa}}(j),j} + \sum_{k\in\widehat{\text{an}}(j)} \widehat{\alpha}_{kj}\widehat{h}_k)$ is the residual estimated from the jth equation. In practice, we replace $\widehat{\boldsymbol{h}}_j$ with $\widehat{\boldsymbol{\epsilon}}_j$ in the algorithm as $\widehat{\boldsymbol{h}}_j$ is a linear combination of $\widehat{\boldsymbol{h}}_k$ and $\widehat{\boldsymbol{\epsilon}}_j$; including $\widehat{\boldsymbol{h}}_k$ and $\widehat{\boldsymbol{h}}_j$ in the GLM regression model is equivalent to including $\widehat{\boldsymbol{h}}_k$ and $\widehat{\boldsymbol{\epsilon}}_j$.

Note $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{ml}$ implies $\widehat{\boldsymbol{U}}_{\widehat{\mathrm{pa}}(j),j} = \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}$, $\widehat{\boldsymbol{W}}_{\widehat{\mathrm{in}}(j),j} = \widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}$, and $\widehat{\alpha}_{kj} = \widehat{\alpha}^{ml}_{kj}$. We therefore have

$$\widehat{\boldsymbol{\epsilon}}_j = \boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j} + \sum_k \widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k).$$

On the other hand, by Assumption 2, $\boldsymbol{h}_j = \sum_k \alpha^0_{kj}\boldsymbol{h}_k + \boldsymbol{\epsilon}_j$, where $\boldsymbol{\epsilon}_j$ is orthogonal to the space spanned by $\{h_k : k \in \mathrm{an}(j)\}$. Therefore, (2) can be written as:

$$\mathbb{E}\left[\boldsymbol{Y}_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, \boldsymbol{h}\right] = \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\boldsymbol{h}_k + \boldsymbol{\epsilon}_j).$$

Similar to the root node case, we use $\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\boldsymbol{h}_k)$ to approximate $\boldsymbol{\epsilon}_j$ so that $\widehat{\boldsymbol{\epsilon}}_j$ and $\boldsymbol{\epsilon}_j$ can be compared at the same scale. Note by our previous calculation, the approximation error $\mathbb{E}\left[Y_{ij}|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}, \boldsymbol{h}\right] - Y_{ij}$ is sub-exponential with mean zero and the aggregate impact of this error term across samples is of the order of $o(\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}})$ by the Bernstein's inequality in subsequent equations. Toward this end,

$$\begin{aligned}
\Delta_j = \widehat{\boldsymbol{h}}_j - \boldsymbol{h}_j &= (\sum_k \widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k + \widehat{\boldsymbol{\epsilon}}_j) - (\sum_k \alpha^0_{kj}\boldsymbol{h}_k + \boldsymbol{\epsilon}_j) = \sum_k (\widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k - \alpha^0_{kj}\boldsymbol{h}_k) + (\widehat{\boldsymbol{\epsilon}}_j - \boldsymbol{\epsilon}_j) \\
&= I_1 - \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j} + \sum_k \widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k) + \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\boldsymbol{h}_k) \\
&= I_1 + \varphi_j'(\boldsymbol{\xi}) \odot \big(\boldsymbol{X}_{\mathrm{in}^0(j)}(\boldsymbol{W}^0_{\mathrm{in}^0(j),j} - \widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}) + \boldsymbol{Y}_{\mathrm{pa}^0(j)}(\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} - \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}) + \sum_k \alpha^0_{kj}\boldsymbol{h}_k - \sum_k \widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k\big),
\end{aligned}$$

where $I_1 = \sum_k(\widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k - \alpha^0_{kj}\boldsymbol{h}_k)$. Further, we have $I_1 = \sum_k(\widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k - \alpha^0_{kj}\boldsymbol{h}_k) = \sum_k(\widehat{\alpha}^{ml}_{kj}\widehat{\boldsymbol{h}}_k - \alpha^0_{kj}\widehat{\boldsymbol{h}}_k + \alpha^0_{kj}\widehat{\boldsymbol{h}}_k - \alpha^0_{kj}\boldsymbol{h}_k)$. On the other hand, note that $\widetilde{\boldsymbol{Z}}(\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}) = \boldsymbol{X}_{\mathrm{in}^0(j)}(\boldsymbol{W}^0_{\mathrm{in}^0(j),j} - \widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}) + \boldsymbol{Y}_{\mathrm{pa}^0(j)}(\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} - \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}) + \sum_k(\alpha^0_{kj} - \widehat{\alpha}^{ml}_{kj})\widehat{\boldsymbol{h}}_k$. Therefore, Δ_j can be written as

$$\begin{aligned}
\Delta_j &= \sum_k(\widehat{\alpha}^{ml}_{kj} - \alpha^0_{kj})\widehat{\boldsymbol{h}}_k + \sum_k \alpha^0_{kj}(\widehat{\boldsymbol{h}}_k - \boldsymbol{h}_k) + \varphi_j'(\boldsymbol{\xi}) \odot \widetilde{\boldsymbol{Z}}(\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}) + \varphi_j'(\boldsymbol{\xi}) \odot \sum_k \alpha^0_{kj}(\boldsymbol{h}_k - \widehat{\boldsymbol{h}}_k) \\
&= \sum_k(\widehat{\alpha}^{ml}_{kj} - \alpha^0_{kj})\widehat{\boldsymbol{h}}_k + \varphi_j'(\boldsymbol{\xi}) \odot \widetilde{\boldsymbol{Z}}(\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}) + (1 - \varphi_j'(\boldsymbol{\xi})) \odot \sum_k \alpha^0_{kj}(\widehat{\boldsymbol{h}}_k - \boldsymbol{h}_k).
\end{aligned}$$

Note $\|\sum_k(\widehat{\alpha}^{ml}_{kj} - \alpha^0_{kj})\widehat{\boldsymbol{h}}_k\|_\infty \le \sum_k \|(\widehat{\alpha}^{ml}_{kj} - \alpha^0_{kj})\widehat{\boldsymbol{h}}_k\|_\infty \le \max_k |\widehat{\alpha}^{ml}_{kj} - \alpha^0_{kj}| \cdot \sum_k \|\widehat{\boldsymbol{h}}_k\|_\infty \le \|\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}\|_\infty \cdot \sum_k \|\widehat{\boldsymbol{h}}_k\|_\infty \le b_1 s \cdot \|\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j}\|_\infty$, and $\|\sum_k \alpha^0_{kj}(\widehat{\boldsymbol{h}}_k - \boldsymbol{h}_k)\|_\infty \le \sum_k |\alpha^0_{kj}| \cdot \|\Delta_k\|_\infty$. Moreover,

$$\begin{aligned}
\|\widetilde{\boldsymbol{Z}}(\boldsymbol{\theta}^0_{A^0_j} - \widehat{\boldsymbol{\theta}}^{ml}_{A^0_j})\|_\infty &= \|\widetilde{\boldsymbol{Z}}(\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M}\widetilde{\boldsymbol{Z}})^{-1}\widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y}_j - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \boldsymbol{r})\|_\infty \\
&= \|\boldsymbol{H}_2(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) + \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \boldsymbol{r})\|_\infty \\
&\le \|\boldsymbol{H}_2(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty + \|\boldsymbol{H}_2(\varphi_j(\boldsymbol{z}_i^\top \boldsymbol{\theta}^0_{A^0_j}) - \varphi_j(\widetilde{\boldsymbol{z}}_i^\top \boldsymbol{\theta}^0_{A^0_j}))\|_\infty + \|\boldsymbol{H}_2\boldsymbol{r}\|_\infty,
\end{aligned}$$

where $\boldsymbol{H}_2 = \widetilde{\boldsymbol{Z}}(\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M}\widetilde{\boldsymbol{Z}})^{-1}\widetilde{\boldsymbol{Z}}^\top$. Again, by Assumption 5, there exists b_3 such that $\|n\widetilde{\boldsymbol{Z}}(\widetilde{\boldsymbol{Z}}^\top \boldsymbol{M}\widetilde{\boldsymbol{Z}})^{-1}\widetilde{\boldsymbol{Z}}^\top\|_\infty \leq b_3$. Similarly, following the calculation of $\|\widehat{\boldsymbol{\theta}}^{ml}_{A_j^0} - \boldsymbol{\theta}^0_{A_j^0}\|_\infty$, we obtain

$$
\begin{aligned}
||\Delta_j||_\infty &\leq b_1 s \cdot \|\boldsymbol{\theta}^0_{A_j^0} - \widehat{\boldsymbol{\theta}}^{ml}_{A_j^0}\|_\infty + L_1\Bigg[2Mb_3\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + b_3L_1\sum_k |\alpha^0_{kj}|\cdot\|\Delta_k\|_\infty \\
&\quad + b_3c_0D\left(\frac{32}{m^2}M^2b_1^2\cdot\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + \frac{8}{m^2}b_1^2L_1^2\sum_k|\alpha^0_{kj}|^2\cdot\|\Delta_k\|_\infty\right)\Bigg] + (1+L_1)\sum_k|\alpha^0_{kj}|\cdot\|\Delta_k\|_\infty \\
&\leq a_4\sum_k\|\Delta_k\|_\infty + a_3\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} \leq a_4 s\cdot\max_k\|\Delta_k\|_\infty + a_3\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}},
\end{aligned}
$$

where $a_4 = \max_k\left(b_3L_1^2|\alpha^0_{kj}| + \frac{8b_3c_0Db_1^2L_1^3|\alpha^0_{kj}|^2}{m^2} + b_1s\left(b_2L_1|\alpha^0_{kj}| + \frac{8b_2c_0Db_1^2L_1^2|\alpha^0_{kj}|^2}{m^2}\right) + (1+L_1)|\alpha^0_{kj}|\right)$ and $a_3 = \left(2Mb_3L_1 + b_3c_0D\frac{32}{m^2}M^2b_1^2L_1 + b_1s(2Mb_2 + b_2c_0D\frac{32}{m^2}M^2b_1^2)\right)$. The above inequality can be written as: $\|\Delta_j\|_\infty + c \leq a_4 s\,(\max_k\|\Delta_k\|_\infty + c)$, where $c = \frac{a_3}{a_4 s - 1}\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}}$. Therefore,

$$
\|\Delta_j\|_\infty + c \leq (a_4 s)^{d_j}(\|\Delta_0\|_\infty + c),
$$

where d_j denotes the topology depth of the primary variable Y_j defined as the maximal length of a directed path in the graph from a root variable with depth zero; therefore, $0 \leq d_j \leq d_{\max} \leq p-1$ with $d_{\max}$ the maximal length of a directed path. Rearranging terms yields

$$
\begin{aligned}
\|\Delta_j\|_\infty &\leq (a_4 s)^{d_j}\|\Delta_0\|_\infty + ((a_4 s)^{d_j} - 1)c \\
&= (a_4 s)^{d_j}\|\Delta_0\|_\infty + ((a_4 s)^{d_j} - 1)\frac{a_3}{a_4 s - 1}\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}}.
\end{aligned}
$$

In this way, we derive the general form of the accumulated error for $\|\Delta_j\|_\infty$ for the multi-layer case. To conclude, if τ_j satisfies: $\tau_j \geq 2\left\|\widehat{\boldsymbol{\theta}}^{ml} - \boldsymbol{\theta}^0\right\|_\infty = C\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}}$, then the deconfounding algorithm reconstructs the causal graph consistently, i.e., $\{(k,j) : \widehat{U}_{kj} \neq 0\} = \{(k,j) : U^0_{kj} \neq 0\}$, with probability $1 - 8p\cdot p^{-2}(2s+\widetilde{s})^{-1}$ by the union bound, tending to one as $p \to \infty$ and thus $n \to \infty$. This completes the proof. We next present proofs of the lemmas.

Lemma 9 bounds the quantity $\|\widehat{\boldsymbol{\theta}}^{ml}_{A_j^0} - \boldsymbol{\theta}^0_{A_j^0}\|_2$ in child equations.

Lemma 9 (Rate of convergence under the ℓ_2-norm for child equations)

$$
\begin{aligned}
\|\widehat{\boldsymbol{\theta}}^{ml}_{A_j^0} - \boldsymbol{\theta}^0_{A_j^0}\|_2 &\leq \frac{2}{m}\sqrt{2s+\widetilde{s}}\cdot\|\nabla\mathcal{L}(\boldsymbol{\theta}^0_{A_j^0}|\boldsymbol{Y}_{pa^0(j)}, \boldsymbol{X}_{in^0(j)}, \widehat{\boldsymbol{h}}_{an^0(j)})\|_\infty \\
&\leq \frac{2}{m}\sqrt{2s+\widetilde{s}}\left[2Mb_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + b_1L_1\sum_k|\alpha^0_{kj}|\cdot\|\Delta_k\|_\infty\right],
\end{aligned}
$$

with probability at least $1 - 2\exp(-2\log p - \log(2s+\widetilde{s})) = 1 - 2p^{-2}(2s+\widetilde{s})^{-1}$.

Proof of Lemma 9. Since $\widehat{\boldsymbol{\theta}}^{ml}_{A_j^0} = (\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(j),j}, \widehat{\boldsymbol{U}}^{ml}_{\mathrm{pa}^0(j),j}, \widehat{\boldsymbol{\alpha}}^{ml}_{\mathrm{an}^0(j),j})$ minimizes $\mathcal{L}(\boldsymbol{\theta}_{A_j^0}|\boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{X}_{\mathrm{in}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)})$, $\mathcal{L}(\widehat{\boldsymbol{\theta}}^{ml}_{A_j^0}|\boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{X}_{\mathrm{in}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)}) \leq \mathcal{L}(\boldsymbol{\theta}^0_{A_j^0}|\boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{X}_{\mathrm{in}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)})$.

As in Lemma 8, $\|\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\|_2 \le \frac{2}{m}\sqrt{2s+\widetilde{s}}\cdot\|\nabla\mathcal{L}(\boldsymbol{\theta}^0_{A^0_j}|\boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{X}_{\mathrm{in}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)})\|_\infty$. Meanwhile,

$$
\begin{aligned}
&\|\nabla\mathcal{L}(\boldsymbol{\theta}^0_{A^0_j}|\boldsymbol{Y}_{\mathrm{pa}^0(j)}, \boldsymbol{X}_{\mathrm{in}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)})\|_\infty \\
&= n^{-1}\|\widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\widehat{\boldsymbol{h}}_k))\|_\infty \\
&= n^{-1}\big\|T_1 + T_2\|_\infty \le n^{-1}\big\|T_1\|_\infty + n^{-1}\big\|T_2\|_\infty \\
&\le n^{-1}\|T_1\|_\infty + n^{-1}\big\|\widetilde{\boldsymbol{Z}}^\top\cdot\big(\sum_k \alpha^0_{kj}(\boldsymbol{h}_k - \widehat{\boldsymbol{h}}_k)\odot\varphi_j'(\boldsymbol{\xi})\big)\big\|_\infty \\
&\le n^{-1}\|T_1\|_\infty + n^{-1}\big\|\widetilde{\boldsymbol{Z}}^\top\cdot\big(\sum_k \alpha^0_{kj}\left(\Delta_k + (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k])\right)\odot\varphi_j'(\boldsymbol{\xi})\big)\big\|_\infty \qquad \text{by (43)} \\
&\le n^{-1}\|T_1\|_\infty + b_1L_1\sum_k|\alpha^0_{kj}|\cdot\|\Delta_k\|_\infty + L_1n^{-1}\big\|\widetilde{\boldsymbol{Z}}^\top\big(\sum_k \alpha^0_{kj}(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k])\big)\big\|_\infty,
\end{aligned}
$$

where $T_1 = \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y}_j - \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\boldsymbol{h}_k))$ and $T_2 = \widetilde{\boldsymbol{Z}}^\top(\varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\boldsymbol{h}_k) - \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\widehat{\boldsymbol{h}}_k))$. The last inequality holds as $|\varphi_j'(z)| = |A_j^{''}(z)| \le L_1$.

For the last term above, $n^{-1}\|\widetilde{\boldsymbol{Z}}^\top\big(\sum_k \alpha^0_{kj}(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k])\big)\|_\infty = n^{-1}\|\sum_k \alpha^0_{kj}\widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k])\|_\infty$. Similarly, by Assumption 3 and the Bernstein's inequality, we conclude that $n^{-1}\|\sum_k \alpha^0_{kj}\widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\mathrm{in}^0(k)}, \boldsymbol{h}_k])\|_\infty = o(\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}})$ with probability tending to 1.

Note that $\|\widetilde{\boldsymbol{Z}}\|_\infty \le b_1$ and $\widetilde{\boldsymbol{Z}} = [\boldsymbol{X}_{\mathrm{in}^0(j)}, \boldsymbol{Y}_{\mathrm{pa}^0(j)}, \widehat{\boldsymbol{h}}_{\mathrm{an}^0(j)}] \in \mathbb{R}^{2s+\widetilde{s}}$. The first term is also bounded by the Bernstein's inequality since $\mathbb{E}[\boldsymbol{Y}_j|\boldsymbol{Z}] = \varphi_j(\boldsymbol{X}_{\mathrm{in}^0(j)}\boldsymbol{W}^0_{\mathrm{in}^0(j),j} + \boldsymbol{Y}_{\mathrm{pa}^0(j)}\boldsymbol{U}^0_{\mathrm{pa}^0(j),j} + \sum_k \alpha^0_{kj}\boldsymbol{h}_k)$.

$$
\mathbb{P}(n^{-1}\|T_1\|_\infty > \epsilon) \le 2(2s+\widetilde{s})\exp\left(-\min\left(\frac{n\epsilon^2}{2M^2b_1^2}, \frac{n\epsilon}{2Mb_1}\right)\right).
$$

Setting $\epsilon = \eta_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}}$ yields

$$
\mathbb{P}\left(n^{-1}\|T_1\|_\infty \le \eta_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}}\right) \ge 1 - 2\exp\left(-\frac{1}{2}(\frac{\eta_1^2}{M^2b_1^2} - 2)\log(2s+\widetilde{s}) - \frac{\eta_1^2}{2M^2b_1^2}\log p\right).
$$

In particular, setting $\eta_1 = 2Mb_1$ yields

$$
\|\widehat{\boldsymbol{\theta}}^{ml}_{A^0_j} - \boldsymbol{\theta}^0_{A^0_j}\|_2 \le \frac{2}{m}\sqrt{2s+\widetilde{s}}\left[2Mb_1\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}} + b_1L_1\sum_k|\alpha^0_{kj}|\cdot\|\Delta_k\|_\infty\right],
$$

with probability at least $1 - 2\exp(-2\log p - \log(2s+\widetilde{s})) = 1 - 2p^{-2}(2s+\widetilde{s})^{-1}$.

Lemma 10 bounds the quantity $\|\widehat{\boldsymbol{W}}^{ml}_{\mathrm{in}^0(k),k} - \boldsymbol{W}^0_{\mathrm{in}^0(k),k}\|_2$ in root equations.

Lemma 10 (Rate of convergence under the ℓ_2-norm for root equations)

$$
\|\widehat{\boldsymbol{W}}^{ml}_{in^0(k),k} - \boldsymbol{W}^0_{in^0(k),k}\|_2 \le \frac{2}{m}\sqrt{\widetilde{s}}\left[2Mb_1\sqrt{\frac{\log(p\widetilde{s})}{n}}\right],
$$

with probability at least $1 - 2\exp(-2\log p - \log\widetilde{s}) = 1 - 2p^{-2}\widetilde{s}^{-1}$.

Proof of Lemma 10. Consider the log-likelihood for a root variable Y_k:

$$\mathcal{L}(\boldsymbol{W}_{\text{in}^0(k),k}|\boldsymbol{Y}_k, \boldsymbol{X}_{\text{in}^0(k)}) = n^{-1}\sum_{i=1}^{n} -Y_{ik}\left(\boldsymbol{W}_{\text{in}^0(k),k}^\top \boldsymbol{X}_{i,\text{in}^0(k)}\right) + A_k\left(\boldsymbol{W}_{\text{in}^0(k),k}^\top \boldsymbol{X}_{i,\text{in}^0(k)}\right),$$

where the oracle estimator $\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k}$ is its minimizer with respect to $\boldsymbol{W}_{\text{in}^0(k),k}$.

By the definition of $\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k}$, it follows from Lemma 8 that $\|\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_2 \leq \frac{2}{m}\sqrt{\widetilde{s}} \cdot \|\nabla\mathcal{L}(\boldsymbol{W}^0_{\text{in}^0(k),k}|\boldsymbol{Y}_k, \boldsymbol{X}_{\text{in}^0(k)})\|_\infty$, where $\nabla\mathcal{L}(\boldsymbol{W}^0_{\text{in}^0(k),k}|\boldsymbol{Y}_k, \boldsymbol{X}_{\text{in}^0(k)})$ is the gradient of $\mathcal{L}(\boldsymbol{W}^0_{\text{in}^0(k),k}|\boldsymbol{Y}_k, \boldsymbol{X}_{\text{in}^0(k)})$. By the triangular inequality,

$$\begin{aligned}
&\|\nabla\mathcal{L}(\boldsymbol{W}^0_{\text{in}^0(k),k}|\boldsymbol{Y}_k, \boldsymbol{X}_{\text{in}^0(k)})\|_\infty = n^{-1}\|\boldsymbol{X}^\top_{\text{in}^0(k)}(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\|_\infty \\
&\leq n^{-1}\|\boldsymbol{X}^\top_{\text{in}^0(k)}(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k))\|_\infty \\
&\quad + n^{-1}\|\boldsymbol{X}^\top_{\text{in}^0(k)}(\varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k) - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\|_\infty \\
&\equiv G_1 + G_2. \qquad (42)
\end{aligned}$$

Note that $\|\boldsymbol{X}_{\text{in}^0(k)}\|_\infty \leq b_1$ and $\mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k] = \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k)$. By the Bernstein's inequality, the first term in (42) is bounded by

$$\mathbb{P}(G_1 > \epsilon) \leq 2\widetilde{s}\exp\left(-\min\left(\frac{n\epsilon^2}{2M^2b_1^2}, \frac{n\epsilon}{2Mb_1}\right)\right).$$

Setting $\epsilon = 2Mb_1\sqrt{\frac{\log(p\widetilde{s})}{n}}$ leads to $G_1 \leq 2Mb_1\sqrt{\frac{\log(p\widetilde{s})}{n}}$ with probability at least $1 - 2\exp(-2\log p - \log\widetilde{s}) = 1 - 2p^{-2}\widetilde{s}^{-1}$.

For G_2 in (42), by the Taylor series expansion,

$$\begin{aligned}
G_2 &= n^{-1}\|\boldsymbol{X}^\top_{\text{in}^0(k)} \cdot \left(\varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) \odot \boldsymbol{h}_k\right)\|_\infty \\
&= n^{-1}\|(\boldsymbol{X}^\top_{\text{in}^0(k)}\text{diag}(\varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))) \cdot \boldsymbol{h}_k\|_\infty.
\end{aligned}$$

Note that $\boldsymbol{X}_{\text{in}^0(k)}$ and $\boldsymbol{h}_k$ are independent. Hence, $\mathbb{E}[\left(\boldsymbol{X}^\top_{\text{in}^0(k)}\text{diag}(\varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\right) \cdot \boldsymbol{h}_k] = 0$. By the bounded domain for interventions condition, there exists b_4 such that $\|\boldsymbol{X}^\top_{\text{in}^0(k)}\text{diag}(\varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\|_\infty \leq b_4$. For $j \in \text{in}^0(k)$, by the Hoeffding's inequality,

$$\mathbb{P}(n^{-1}\|\boldsymbol{X}_j^\top\text{diag}(\varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\boldsymbol{h}_k\|_\infty > \epsilon) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma_j^2b_4^2}\right).$$

Applying the union bound and setting $\epsilon = 2\sigma_j b_4\sqrt{\frac{\log(p\widetilde{s})}{n}}$ yield $G_2 \leq 2\sigma_j b_4\sqrt{\frac{\log(p\widetilde{s})}{n}}$ with probability at least $1 - 2\exp(-2\log p - \log\widetilde{s}) = 1 - 2p^{-2}\widetilde{s}^{-1}$. For simplicity, set $G_2 = o(\sqrt{\frac{\log(p\widetilde{s})}{n}})$.

Finally, combining the two terms in (42) yields:

$$\|\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_2 \leq \frac{2}{m}\sqrt{\widetilde{s}} \cdot \|\nabla\mathcal{L}(\boldsymbol{W}^0_{\text{in}^0(k),k}|\boldsymbol{Y}_k, \boldsymbol{X}_{\text{in}^0(k)})\|_\infty \leq \frac{2}{m}\sqrt{\widetilde{s}}\big(2Mb_1\sqrt{\frac{\log(p\widetilde{s})}{n}}\big).$$

Lemma 11 derives the estimation bound for $\|\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_\infty$ in root equations.

Lemma 11 (Rate of convergence under the ℓ_∞-norm for root equations)

$$\|\widehat{\boldsymbol{W}}_{\overline{in}(k),k} - \boldsymbol{W}^0_{in^0(k),k}\|_\infty \leq \left(2Mb_2 + b_2c_0D\frac{16}{m^2}M^2b_1^2\right)\sqrt{\frac{\log(p\widetilde{s})}{n}},$$

with probability at least $1 - 4\exp(-2\log p - \log\widetilde{s}) = 1 - 4p^{-2}\widetilde{s}^{-1}$. *Further, the estimation error of the confounder* $\boldsymbol{h}_k$ *satisfies:* $||\Delta_k||_\infty \leq \left(2Mb_3 + b_3c_0D\frac{16}{m^2}M^2b_1^2\right)\sqrt{\frac{\log(p\widetilde{s})}{n}}$.

Proof of Lemma 11. Note that by Theorem 4, $\{l : \widehat{V}_{lk} \neq 0\} = \{l : V^0_{lk} \neq 0\}$, implying that $\overline{\text{in}}(k) = \text{in}^0(k)$ in root equations. Therefore, by construction, $\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k} = \widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k}$, as both are GLM estimators constrained on the same set. It suffices to derive the error bound for the oracle estimator.

To establish the ℓ_∞-norm of the oracle estimator, as in Lemma 7, we apply the Taylor series expansion of $\varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}_{\text{in}^0(k),k})$ as:

$$\boldsymbol{X}^\top_{\text{in}^0(k)}\boldsymbol{M}\boldsymbol{X}_{\text{in}^0(k)}(\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}) = \boldsymbol{X}^\top_{\text{in}^0(k)}(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) - \boldsymbol{r}).$$

This implies that $\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k} = \boldsymbol{H}(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) - \boldsymbol{r})$, where $\boldsymbol{H} = (\boldsymbol{X}^\top_{\text{in}^0(k)}\boldsymbol{M}\boldsymbol{X}_{\text{in}^0(k)})^{-1}\boldsymbol{X}^\top_{\text{in}^0(k)}$. Then,

$$\begin{aligned}
&\|\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_\infty = \|\boldsymbol{H}(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k) \\
&\quad + \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k) - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) - \boldsymbol{r})\|_\infty \\
&\leq \|\boldsymbol{H}(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k))\|_\infty \\
&\quad + \|\boldsymbol{H}(\varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k) - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\|_\infty + \|\boldsymbol{H}\boldsymbol{r}\|_\infty, \quad \equiv I_1 + I_2 + I_3.
\end{aligned}$$

By Assumption 5, there exists b_2 such that $\|n(\boldsymbol{X}^\top_{\text{in}^0(k)}\boldsymbol{M}\boldsymbol{X}_{\text{in}^0(k)})^{-1}\boldsymbol{X}^\top_{\text{in}^0(k)}\|_\infty \leq b_2$. Note that $\mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k] = \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k)$. Then, by the Bernstein's inequality:

$$\mathbb{P}(I_1 > \epsilon) \leq 2\widetilde{s}\exp\left(-\min\left(\frac{n\epsilon^2}{2M^2b_2^2}, \frac{n\epsilon}{2Mb_2}\right)\right).$$

Setting $\epsilon = 2Mb_2\sqrt{\frac{\log(p\widetilde{s})}{n}}$ yields $I_1 \leq 2Mb_2\sqrt{\frac{\log(p\widetilde{s})}{n}}$ with probability at least $1 - 2\exp(-2\log p - \log\widetilde{s}) = 1 - 2p^{-2}\widetilde{s}^{-1}$. On the other hand, as in Lemma 10, we have

$$\begin{aligned}
I_2 &= \|\boldsymbol{H}\left(\boldsymbol{h}_k \odot \varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k})\right)\|_\infty \\
&= \|(\boldsymbol{X}^\top_{\text{in}^0(k)}\boldsymbol{M}\boldsymbol{X}_{\text{in}^0(k)})^{-1}\boldsymbol{X}^\top_{\text{in}^0(k)} \cdot \left(\boldsymbol{h}_k \odot \varphi_k'(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k})\right)\|_\infty = o(\sqrt{\frac{\log(p\widetilde{s})}{n}}).
\end{aligned}$$

Finally, as in Theorem 4 and Lemma 7,

$$\begin{aligned}
|\boldsymbol{H}_{k\bullet}\boldsymbol{r}| &= |\sum_{i=1}^n H_{ki}r_i| \leq \sum_{i=1}^n |H_{ki}||r_i| \leq b_2\sum_{i=1}^n |r_i|/n \\
&\leq b_2D(\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k})^\top(\boldsymbol{X}^\top_{\text{in}^0(k)}\boldsymbol{X}_{\text{in}^0(k)}/n)(\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}) \\
&\leq b_2c_0D\|\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_2^2.
\end{aligned}$$

By Lemma 10, $\|\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_2 \leq \frac{2}{m}\sqrt{\widetilde{s}}\left[2Mb_1\sqrt{\frac{\log(p\widetilde{s})}{n}}\right]$. Therefore,

$$I_3 = \max_k |\boldsymbol{H}_{k\bullet}\boldsymbol{r}| \leq b_2c_0D\left(\frac{2}{m}\sqrt{\widetilde{s}}\left[2Mb_1\sqrt{\frac{\log(p\widetilde{s})}{n}}\right]\right)^2 \leq b_2c_0D\frac{16}{m^2}M^2b_1^2\sqrt{\frac{\log(p\widetilde{s})}{n}},$$

where the last inequality holds as $n > \widetilde{s}^2\log(p\widetilde{s})$. Therefore, combining I_1, I_2 and I_3 yields

$$\|\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_\infty \leq a_1\sqrt{\frac{\log(p\widetilde{s})}{n}},$$

with probability greater than $1 - 4\exp(-2\log p - \log\widetilde{s}) = 1 - 4p^{-2}\widetilde{s}^{-1}$. Here, $a_1 = 2Mb_2 + b_2c_0D\frac{16}{m^2}M^2b_1^2$. Lastly, recall that $\overline{\text{in}}(k) = \text{in}^0(k)$ and $\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k} = \widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k}$. Therefore, $\|\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}\|_\infty \leq a_1\sqrt{\frac{\log(p\widetilde{s})}{n}}$.

To compute the estimation error of the confounder $\boldsymbol{h}_k$, note that, by construction,

$$\widehat{\boldsymbol{h}}_k = \boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\overline{\text{in}}(k)}\widehat{\boldsymbol{W}}_{\overline{\text{in}}(k),k}) = \boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k}).$$

On the other hand, by (2) and following (A5)-(A7) in Appendix A of Johnston et al. (2008),

$$\mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k] = \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k) = \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) + \varphi_k'(\boldsymbol{\xi}) \odot \boldsymbol{h}_k.$$

Rearranging terms yields $\boldsymbol{h}_k = \left[\mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k] - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k})\right] \odot (\varphi_k'(\boldsymbol{\xi}))^{-1}$. We now use $\mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k] - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k})$ to approximate $\boldsymbol{h}_k$ as we estimate the coefficient of the confounder in subsequent child equations. This reparametrization and approximations permits a comparison of $\widehat{\boldsymbol{h}}_k$ and $\boldsymbol{h}_k$ at the same scale; see Appendix A of Johnston et al. (2008) for some details about such approximations. Hence,

$$\begin{aligned}
\widehat{\boldsymbol{h}}_k - \boldsymbol{h}_k &= -\varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k}) + \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) + (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k]) \\
&= -\varphi_k'(\boldsymbol{\xi}) \odot \boldsymbol{X}_{\text{in}^0(k)}(\widehat{\boldsymbol{W}}^{ml}_{\text{in}^0(k),k} - \boldsymbol{W}^0_{\text{in}^0(k),k}) + (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k]) \\
&= -\varphi_k'(\boldsymbol{\xi}) \odot \boldsymbol{H}_2(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) - \boldsymbol{r}) + (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k]) \\
&= \Delta_k + (\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k]), \qquad (43)
\end{aligned}$$

where $\boldsymbol{H}_2 = \boldsymbol{X}_{\text{in}^0(k)}(\boldsymbol{X}^\top_{\text{in}^0(k)}\boldsymbol{M}\boldsymbol{X}_{\text{in}^0(k)})^{-1}\boldsymbol{X}^\top_{\text{in}^0(k)}$ and $\Delta_k = -\varphi_k'(\boldsymbol{\xi}) \odot \boldsymbol{H}_2(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) - \boldsymbol{r})$. Hence, the estimation error of the confounder consists of two terms: the prediction error Δ_k and an approximation error. For the approximation error, by Assumption 3, Y_{ik} is sub-exponential and $Y_{ik} - \mathbb{E}[Y_{ik}|\cdot]$ is sub-exponential with mean zero. Therefore, we have $\boldsymbol{\epsilon} = \boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{X}_{\text{in}^0(k)}, \boldsymbol{h}_k]$, where ϵ_i is sub-exponential with mean zero. We show in the proof of Theorem 5 and Lemma 9 that the aggregate impact of this term across all samples is of the order of $o(\sqrt{\frac{\log(p(2s+\widetilde{s}))}{n}})$ when calculating the estimation error of parameters in the subsequent child equations. For the prediction error Δ_k, by the triangular inequality,

$$\begin{aligned}
&\|\Delta_k\|_\infty \leq L_1 \cdot \|\boldsymbol{H}_2(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}) - \boldsymbol{r})\|_\infty \\
\leq\ & L_1 \cdot \big(\|\boldsymbol{H}_2(\boldsymbol{Y}_k - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k))\|_\infty \\
&+ \|\boldsymbol{H}_2(\varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k} + \boldsymbol{h}_k) - \varphi_k(\boldsymbol{X}_{\text{in}^0(k)}\boldsymbol{W}^0_{\text{in}^0(k),k}))\|_\infty + \|\boldsymbol{H}_2\boldsymbol{r}\|_\infty\big).
\end{aligned}$$

By the bounded domain for interventions condition, there exists b_3 such that $\|n\boldsymbol{X}_{\text{in}^0(k)}(\boldsymbol{X}_{\text{in}^0(k)}^\top \boldsymbol{M}\boldsymbol{X}_{\text{in}^0(k)})^{-1}\boldsymbol{X}_{\text{in}^0(k)}^\top\|_\infty \leq b_3$. Similarly, $||\Delta_k||_\infty \leq a_2 L_1\sqrt{\frac{\log(p\widetilde{s})}{n}}$, with probability greater than $1-4\exp(-2\log p-\log\widetilde{s}) = 1-4p^{-2}\widetilde{s}^{-1}$. Here, $a_2 = 2Mb_3+b_3c_0D\frac{16}{m^2}M^2b_1^2$. This completes the proof.

References

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, pages 8226–8239, 2022.

Bing Cai, Dylan S Small, and Thomas R Ten Have. Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. *Statistics in Medicine*, 30(15):1809–1824, 2011.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Li Chen, Chunlin Li, Xiaotong Shen, and Wei Pan. Discovery and inference of a causal network with hidden confounding. *Journal of the American Statistical Association*, pages 1–13, 2023.

Yichen Chen, Yinyu Ye, and Mengdi Wang. Approximation hardness for a class of sparse optimization problems. *Journal of Machine Learning Research*, 20(38):1–27, 2019.

Won-Seok Choi, Dae-Seok Eom, Baek S Han, Won K Kim, Byung H Han, Eui-Ju Choi, Tae H Oh, George J Markelonis, Jin W Cho, and Young J Oh. Phosphorylation of p38 MAPK induced by oxidative stress is linked to activation of both caspase-8-and-9-mediated apoptotic pathways in dopaminergic neurons. *Journal of Biological Chemistry*, 279(19):20451–20460, 2004.

Shrabanti Chowdhury, Ru Wang, Qing Yu, Catherine J Huntoon, Larry M Karnitz, Scott H Kaufmann, Steven P Gygi, Michael J Birrer, Amanda G Paulovich, Jie Peng, et al. DAGBagM: learning directed acyclic graphs of mixed variables with an application to identify protein biomarkers for treatment response in ovarian cancer. *BMC Bioinformatics*, 23(1):1–19, 2022.

Yayun Du, Xiaoli Liu, Xilin Zhu, Ying Liu, Xinru Wang, and Xiaopan Wu. Activating transcription factor 6 reduces Aβ1–42 and restores memory in Alzheimer's disease model mice. *International Journal of Neuroscience*, 130(10):1015–1023, 2020.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Jerry A Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

Amir M Hossini, Matthias Megges, Alessandro Prigione, Bjoern Lichtner, Mohammad R Toliat, Wasco Wruck, Friederike Schröter, Peter Nuernberg, Hartmut Kroll, Eugenia Makrantonaki, et al. Induced pluripotent stem cell-derived neuronal cells from a sporadic Alzheimer's disease donor as a model for investigating AD-associated gene regulatory networks. *BMC Genomics*, 16:1–22, 2015.

KM Johnston, P Gustafson, AR Levy, and P Grootendorst. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27(9):1539–1556, 2008.

Minoru Kanehisa et al. The KEGG database. In *Novartis Foundation Symposium*, pages 91–100. Wiley Online Library, 2002.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Christina Knudson, Sydney Benson, Charles Geyer, and Galin Jones. Likelihood-based inference for generalized linear mixed models: Inference with the R package glmm. *Stat*, 10(1):e339, 2021.

Jason D Lee, Yuekai Sun, and Jonathan E Taylor. On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics*, 9(1):608–642, 2015.

Chunlin Li, Xiaotong Shen, and Wei Pan. Inference for a large directed acyclic graph with unspecified interventions. *Journal of Machine Learning Research*, 24(73):1–48, 2023.

Wei Li and Johannes Lederer. Tuning parameter calibration for ℓ_1-regularized logistic regression. *Journal of Statistical Planning and Inference*, 202:80–98, 2019.

Mette Lise Lousdal. An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, 15(1):1, 2018.

Andreas Maurer and Massimiliano Pontil. Concentration inequalities under sub-Gaussian and sub-exponential conditions. In *Advances in Neural Information Processing Systems*, volume 34, pages 7588–7597, 2021.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Roger B. Nelsen. *An Introduction to Copulas*. Lecture Notes in Statistics. Springer, 2nd edition, 2007.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017. doi: 10.1214/16-AOS1448.

Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (QVF) DAG models via overdispersion scoring (ODS). *Journal of Machine Learning Research*, 18(224):1–44, 2018.

Judea Pearl. Models, reasoning and inference. *Cambridge University Press*, 19(2):3, 2000.

Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. doi: 10.1214/11-EJS631.

Ayati Sharma, Alisha Chunduri, Asha Gopu, Christine Shatrowsky, Wim E Crusio, and Anna Delprato. Common genetic signatures of Alzheimer's disease in Down Syndrome. *F1000Research*, 9:1299, 2021.

Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832, 2013.

Chengchun Shi, Yunzhe Zhou, and Lexin Li. Testing directed acyclic graph via structural, supervised and generative adversarial learning. *Journal of the American Statistical Association*, pages 1–14, 2023.

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

Joseph V Terza, Anirban Basu, and Paul J Rathouz. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3): 531–543, 2008.

Henri Theil. Estimation and simultaneous correlation in complete equation systems. *Henri Theil's Contributions to Economics and Econometrics: Econometric Theory and Methodology*, pages 65–107, 1992.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed Graphical Models via Exponential Families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 1042–1050. PMLR, 2014.

Eunho Yang, Pradeep Ravikumar, Genevera I. Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(115): 3813–3847, 2015. URL `http://jmlr.org/papers/v16/yang15a.html`.

Andrew Ying, Ronghui Xu, and James Murphy. Two-stage residual inclusion for survival data and competing risks — An instrumental variable approach with application to SEER-Medicare linked data. *Statistics in Medicine*, 38(10):1775–1801, 2019.

Yiping Yuan, Xiaotong Shen, Wei Pan, and Zizhuo Wang. Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika*, 106(1):109–125, 2019.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. doi: 10.1214/09-AOS729.

Hui Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. *Optimization Letters*, 11(4):817–833, 2017.

Jun-Yuan Zhang, Shuang Ma, Xiaoli Liu, Yayun Du, Xilin Zhu, Ying Liu, and Xiaopan Wu. Activating transcription factor 6 regulates cystathionine to increase autophagy and restore memory in Alzheimer's disease model mice. *Biochemical and Biophysical Research Communications*, 615: 109–115, 2022.

Tuo Zhao, Han Liu, and Tong Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics*, 46(1):180–218, 2018.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.