Causal Effect Identification in Heterogeneous Environments from Higher-Order Moments

Yaroslav Kivva[1] **Sina Akbari**[1] **Saber Salehkaleybar**[2] **Negar Kiyavash**[1]

[1]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Leiden Institute of Advanced Computer Science, Leiden University, Netherlands

Abstract

We investigate the estimation of the causal effect of a treatment variable on an outcome in the presence of a latent confounder. We first show that the causal effect is identifiable under certain conditions when data is available from multiple environments, provided that the target causal effect remains invariant across these environments. Secondly, we propose a moment-based algorithm for estimating the causal effect as long as only a single parameter of the data-generating mechanism varies across environments – whether it be the exogenous noise distribution or the causal relationship between two variables. Conversely, we prove that identifiability is lost if both exogenous noise distributions of both the latent and treatment variables vary across environments. Finally, we propose a procedure to identify which parameter of the data-generating mechanism has varied across the environments and evaluate the performance of our proposed methods through experiments on synthetic data.

1 INTRODUCTION

Identifying the causal effect of a treatment on an outcome is a fundamental objective in various fields, including economics [Card, 1993, Angrist and Krueger, 1991], social sciences [Rosenbaum and Rubin, 1983, Imbens, 2024], epidemiology [Robins et al., 2000], and artificial intelligence [Pearl, 2009, 2014]. One of the primary challenges in causal effect identification is the presence of latent confounders – unobserved variables that influence both the treatment and the outcome. Ignoring latent confounders and simply regressing the outcome on the treatment can lead to biased estimates of the causal effect. To address the challenge of latent confounding, one might conduct a randomized controlled trial (RCT). However, RCTs are often too expensive, time-consuming, or even infeasible due to ethical or legal constraints.

In many real-life applications, data collected from different domains often exhibit heterogeneity due to variations in the causal mechanisms that generate each variable from its direct causes. There is extensive research in the literature (see Section 4 for the related work) on causality that leverages data from multiple environments to recover causal relationships. In particular, several studies [Ghassami et al., 2017, Huang et al., 2020, Jaber et al., 2020] have shown that data collected from multiple environments can narrow the set of possible causal graphs compatible with the observed data, compared to using data from a single environment.

Research in multi-environment settings follows two main approaches (see Section 4 for more details). The first aims to identify an equivalence class of causal structures by leveraging distributional shifts across environments, assuming these arise from unknown interventions. The second approach focuses on identifying the direct causes of a target variable rather than the entire causal graph, often assuming linear causal mechanisms and estimating causal effects corresponding to the direct causes.

In this paper, we study the problem of identifying the causal effect of a treatment T on an outcome Y within linear structural causal models (SCMs) in a multi-environment setting (see Figure 1), where a latent confounder U between T and Y is present. This problem closely relates to the second approach discussed earlier. However, prior work either does not account for latent confounding between treatment and outcome or imposes restrictive assumptions on which causal mechanisms can vary across environments (see Section 4 for more details).

The main contributions of our work are as follows:

- For the setup considered in Figure 1 with two environments, we show (Theorem 3.1) that if there is only a single *unknown* change across two domains, then we can classify whether this change occurs in the causal

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:2223–2254.

effect of the latent confounder U on treatment T or outcome Y (i.e. only α or γ varies between the two environments), or in the exogenous noises of variables T or U. Furthermore, in the case that the varying parameter is either the causal effect of U on T or that of U on Y, then the causal effect of the treatment T on Y is identifiable uniquely, and otherwise it can be recovered up to two possible candidates. We provide an estimation procedure tailored to each of these cases.

- We provide a non-identifiability result (Theorem 3.6) showing that the causal effect is not identifiable from two environments if both the exogenous noises of T and U vary across the environments.
- We provide extensive experimental results validating our algorithms. These results show that our proposed estimators consistently converge to the true value of the treatment effect, whereas linear regression baselines exhibit systematic bias. In addition, we analyze the typical range of key parameters in our algorithms. Our code is provided online for reproducibility purposes.

2 PRELIMINARIES

2.1 NOTATION

Let $G = \langle \mathbf{V}, \mathbf{E} \rangle$ be an *directed acyclic graph* (DAG), such that each vertex represents a random variable and each edge corresponds to the direct causal relationship between the random variables it connects. A DAG $\mathcal{G}$ with causal relations defines a structural causal model $\mathcal{M}$ (SCM) such that any random variable X in the graph satisfies the structural equation

$$X = f_x(Pa(X), \epsilon_x),$$

where $Pa(X)$ denotes parents of X in DAG $\mathcal{G}$, ϵ_x is the exogenous noise corresponding to X, and $f_x(\cdot)$ is a function capturing how variable X causally depends on its parents in the causal graph. The subscript in each exogenous noise denotes the variable to which it corresponds, e.g., ϵ_x is the noise pertaining to X. All exogenous noises in a structural equation model are assumed to be jointly independent. To indicate that the structural causal model $\mathcal{M}$ corresponds to an environment i, (i) is added as a superscript, i.e. $\mathcal{M}^{(i)}$. Similarly, $X^{(i)}$, $\epsilon^{(i)}$, $f_x^{(i)}$, denote a random variable, an exogenous noise, and a causal relationship in the environment i. For ease of presentation, we may omit the superscript (i) if the index of the environment is not important or is clear from the context.

For any variable $X \in \mathbf{V}$, the intervention $do(X = x)$ is an operation that converts SCM $\mathcal{M}$ to a new one where the equation of X in $\mathcal{M}$ is replaced by the constant x. Intuitively, this operation can be seen as performing an experiment where one forces a variable X to take a specific value x.

2.2 IDENTIFIABILITY FROM MULTI-DOMAIN OBSERVATIONS

Let $\mathbf{V}$ be the set of random variables in SCM $\mathcal{M}$ and let observed random variables $T, Y \in \mathbf{V}$ denote the treatment and the outcome variable, respectively. In this paper, we focus on linear SCMs, a widely adopted assumption in the literature.

Assumption 1 (Linear SCM)**.** *For any random variable $X \in \mathbf{V}$, function $f_x(\cdot)$ is a linear function:*

$$X = \sum_{S \in Pa(X)} \alpha_{S,X} S + \epsilon_x.$$

Note that coefficient $\alpha_{S,X}$ represents the direct causal effect of the variable S on X. The causal effect of the treatment T on the outcome Y is defined as $\mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)]$. Under Assumption 1 (Linear SCM), finding this causal effect is equivalent to learning coefficient $\beta := \alpha_{Y,T}$.

In this paper, we assume the observational data comes from a collection of linear SCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)} \dots, \mathcal{M}^{(n)}$ that satisfy the following assumption.

Assumption 2. *The causal effect of the treatment T on the outcome Y is invariant across domains, that is $\beta^{(1)} = \beta^{(2)} = \cdots = \beta^{(n)}$.*

Assumption 2 states that the treatment effect remains the same across the domains. For instance, in example proposed by Shi et al. [2021], one may be interested in the effect of sleeping pills on lung disease using electronic health records collected from multiple hospitals. The causal effect of sleeping pills on lung disease is assumed to remain consistent across different hospitals.

We denote by $\mathcal{F}\left(\mathcal{M}^{(i)}, \mathcal{M}^{(j)}\right)$, the set of the **non-invariant** coefficients and exogenous noises between two SCMs $\mathcal{M}^{(i)}$ and $\mathcal{M}^{(j)}$. For example, consider two equations $X^{(i)} = \alpha^{(i)}\epsilon_u^{(i)} + \gamma^{(i)}\epsilon_d^{(i)} + \epsilon_x^{(i)}$ and $X^{(j)} = \alpha^{(j)}\epsilon_u^{(j)} + \gamma^{(j)}\epsilon_d^{(j)} + \epsilon_x^{(j)}$, where $\alpha^{(i)} = \alpha^{(j)}$, $\gamma^{(i)} \neq \gamma^{(j)}$, $\epsilon_u^{(i)} \not\sim \epsilon_u^{(j)}$, $\epsilon_x^{(i)} \not\sim \epsilon_x^{(j)}$ and $\epsilon_d^{(i)} \sim \epsilon_d^{(j)}$ (notation $\sim$ means the random variables are drawn from the same distribution). Then $\alpha, \epsilon_d \notin \mathcal{F}\left(\mathcal{M}^{(i)}, \mathcal{M}^{(j)}\right)$ while $\gamma, \epsilon_u, \epsilon_x \in \mathcal{F}\left(\mathcal{M}^{(i)}, \mathcal{M}^{(j)}\right)$.

Additionally, we require all exogenous noises to have finite moments and be "well-defined" given the moments.

Assumption 3 (Finite moments)**.** *Given an SCM $\mathcal{M}$ on the set of variables $\mathbf{V}$, for any $X \in \mathbf{V}$ and for any $n \in \mathbb{N}$, $\mathbb{E}\left[\epsilon_x^n\right] < \infty$.*

Assumption 4. *Given an SCM $\mathcal{M}$ on the set of variables $\mathbf{V}$, for any $X \in \mathbf{V}$, there exists some $s > 0$ such that the power series $\sum_k \mathbb{E}\left[\epsilon_x^k\right] r^k/k!$ converges for any $0 < r < s$.*

The last assumption implies that the distribution of the random variable X is *uniquely* determined given its moments.

Next, we formalize the definition of identifiability of the treatment effect.

Definition 1. ***(Identifiability)*** *Suppose Assumptions 1 through 4 hold. Moreover, assume that there are n environments, each with a true underlying SCM denoted by $\mathcal{M}^{(i)}$ for environment i. The treatment effect β is said to be* identifiable *from merely observational distributions of n environments if for any collection of SCMs $\{\tilde{\mathcal{M}}^{(i)}\}_{i=1}^{n}$ such that $\tilde{\mathcal{M}}^{(i)}$ entails the same observational distribution as $\mathcal{M}^{(i)}$ for every $i \in \{1, \cdots, n\}$, then the treatment effect in the collection $\{\tilde{\mathcal{M}}^{(i)}\}_{i=1}^{n}$ is equal to the one in $\{\mathcal{M}^{(i)}\}_{i=1}^{n}$; that is, $\tilde{\beta} = \beta$.*

In this work, we address the problem for the canonical case of two environments (i.e., $n = 2$). This characterization suffices as the identifiability results extend to larger values of n simply by considering the environments in pairs.

3 MAIN RESULT

We consider the problem of estimating the causal effect of treatment T on the outcome Y in DAG $\mathcal{G}$ given in Figure 1. It is well-known that given observational data from a single environment, this causal effect is not identifiable [Salehkaleybar et al., 2020]. We will show that the causal effect can be identified given observational data from two environments under certain mild assumptions.

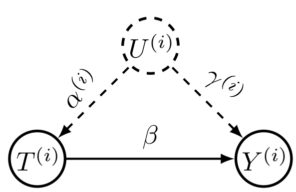


Figure 1: Causal graph of a linear SCM in the i-th domain.

More specifically, in each domain $i \in \{1, 2\}$, we consider the following linear SCM (with the corresponding causal graph in Figure 1) with a treatment variable $T^{(i)}$, an outcome variable $Y^{(i)}$, and a latent confounder $U^{(i)}$:

$$\begin{cases} U^{(i)} := \epsilon_u^{(i)}, \\ T^{(i)} := \alpha^{(i)} U^{(i)} + \epsilon_t^{(i)}, \\ Y^{(i)} := \beta T^{(i)} + \gamma^{(i)} U^{(i)} + \epsilon_y^{(i)}, \end{cases} \tag{1}$$

where $\epsilon_u^{(i)}$, $\epsilon_t^{(i)}$, $\epsilon_y^{(i)}$ are exogenous noises corresponding to $U^{(i)}$, $T^{(i)}$, and $Y^{(i)}$, respectively. In the sequel, we use β and 'the treatment effect' interchangeably.

In this section, we will show that if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)$ is known and $|\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)| = 1$, and additionally $\epsilon_y \notin \mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)$, then the treatment effect can be uniquely identified under mild non-Gaussianity assumptions. We will propose a procedure to learn β for any given $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)$ satisfying the aforementioned conditions. Note that the case $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_y\}$ is not of interest, since intuitively, the change in the distribution of ϵ_y does not provide any new information on the treatment mechanism. Moreover, it can be shown that the treatment effect β is not uniquely identifiable for such a scenario. For completeness, we provide proof of this statement in Proposition 1 in Appendix A.

In practice, we might only know that $|\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)})| = 1$, without knowing which parameter has changed across environments. The following result indicates that even in such a scenario, β can be uniquely identified in some cases, and identified up to a finite set in the others.

Theorem 3.1. *Consider two linear SCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ compatible with the graph of Figure 1, such that $|\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)| = 1$. The treatment effect β can be uniquely identified if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \subset \{\alpha, \gamma\}$ under some additional case-specific mild assumptions; otherwise, if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \subset \{\epsilon_t, \epsilon_u\}$, β can be identified only up to two possible candidates.*

The proofs of all our results are given in Appendix A. Below, we first show that as long as we *know* which single parameter or exogenous noise variable has changed across domains, we can identify β uniquely. Specifically, the procedures for identifying β when $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$, $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$, $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$ and $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$ are given in Sections 3.1, 3.2, 3.3 and 3.4, respectively. Next, in Section 3.5, we outline the procedure for identifying the varying parameter across the two environments given that $|\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)| = 1$. In particular, we can always distinguish whether the source of change was γ or α across the domains. However, if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \subset \{\epsilon_u, \epsilon_t\}$, we cannot pinpoint whether the change was due to variation in the distribution of ϵ_u, or ϵ_t. Therefore, we need to apply both procedures in Sections 3.1 and 3.2 to recover two candidates for the treatment effect β.

In Section 3.6, we prove a non-identifiability result, namely, if the distributions of both exogenous noises ϵ_t and ϵ_u vary across the two environments, then β is not identifiable.

Note that all the results presented in this work can be easily generalized to settings with observed confounders $\mathbf{X} = \{X_1, \ldots, X_m\}$, e.g., Figure 2. More specifically, by regressing the treatment and outcome on the observed covariates and working with the residuals, the problem reduces to the case without observed covariates – a similar procedure was done in [Kivva et al., 2024].

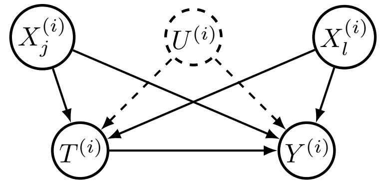


Figure 2: Causal graph with observed covariates $\mathbf{X}$.

3.1 THE CASE $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$

Here we consider the case where the distribution of ϵ_t is changing across the two environments. The corresponding SCMs for environments 1 and 2 can be simplified to[1]

$$\begin{cases} U^{(i)} := \epsilon_u, \\ T^{(i)} := \alpha U^{(i)} + \epsilon_t^{(i)}, \\ Y^{(i)} := \beta T^{(i)} + \gamma U^{(i)} + \epsilon_y. \end{cases} \tag{2}$$

Theorem 3.2. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$. Then under Assumptions 1-4, the treatment effect β can be recovered uniquely.*

The proof of the above result can be found in Appendix A. Algorithm 1 follows from the proof and outlines the procedure for estimating the treatment effect β in this case.

Algorithm 1 $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$

Input: $\{T^{(i)}, Y^{(i)}\}$ and $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$

1: $k \leftarrow 1$
2: **while** $\mathbb{E}\big[\big(T^{(1)}\big)^k\big] = \mathbb{E}\big[\big(T^{(2)}\big)^k\big]$ **do**
3: $\quad k \leftarrow k + 1$
4: $\beta \leftarrow \frac{\mathbb{E}\left[Y^{(1)}(T^{(1)})^{k-1} - Y^{(2)}(T^{(2)})^{k-1}\right]}{\mathbb{E}\left[(T^{(1)})^k - (T^{(2)})^k\right]}$
5: **RETURN:** β

Here, at the end of **while** loop, we find the smallest k such that

$$\mathbb{E}\big[\big(\epsilon_t^{(1)}\big)^k\big] \neq \mathbb{E}\big[\big(\epsilon_t^{(2)}\big)^k\big],$$

which is required to estimate β with the formula given in line 4 of Algorithm 1.

Remark 1. *Note that under Assumption 4, both distributions $\epsilon_t^{(1)}$ and $\epsilon_t^{(2)}$ are uniquely defined given all the moments. Since these distributions are different, they differ at least in one of their moments, which guarantees that such a k exists.*

[1]Please note that while the distributions of ϵ_u and ϵ_y remain unchanged across the domains, the realizations of ϵ_u and ϵ_y differ between the two domains.

3.2 THE CASE $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$

In this section, we assume that only the distribution of ϵ_u is changing across the two environments. The SCM equations can be reduced similarly to Equation 2, as in the previous case. However, the assumptions for the identifiability of the treatment effect β are slightly stronger.

Theorem 3.3. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$. Suppose that $\exists n \in \mathbb{N}$ such that $\mathbb{E}[\epsilon_t^n] \neq (n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$. Then under Assumptions 1-4, the treatment effect β can be recovered uniquely.*

In the statement of the theorem above, we have an additional restriction on the exogenous noise of the treatment, which under Assumption 4, is equivalent to ϵ_t not being Gaussian. Based on the proof of the theorem, we provide Algorithm 2 for estimating the treatment effect β in this case.

Algorithm 2 $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$

Input: $\{T^{(i)}, Y^{(i)}\}$ and $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$

1: $k \leftarrow 1$
2: **while** $\mathbb{E}\big[\big(T^{(1)}\big)^k\big] = \mathbb{E}\big[\big(T^{(2)}\big)^k\big]$ **do**
3: $\quad k \leftarrow k + 1$
4: $r_1 \leftarrow \frac{\mathbb{E}\left[Y^{(1)}(T^{(1)})^{k-1} - Y^{(2)}(T^{(2)})^{k-1}\right]}{\mathbb{E}\left[(T^{(1)})^k - (T^{(2)})^k\right]}$
5: $r_2 \leftarrow$ *GetRatio* $\left(r_1 T^{(1)} - Y^{(1)}, T^{(1)}\right)$
6: $\beta \leftarrow r_1 - r_2$
7: **RETURN:** β

Similarly to the previous case, Algorithm 2 identifies through the **while** loop the smallest k for which

$$\mathbb{E}\big[\big(\epsilon_u^{(1)}\big)^k\big] \neq \mathbb{E}\big[\big(\epsilon_u^{(2)}\big)^k\big].$$

By knowing k then in line 4 we compute $\beta + \frac{\gamma}{\alpha}$ and denote it by r_1. The algorithm then obtains the value of $\frac{\gamma}{\alpha}$, denoted as r_2, using the function *GetRatio*$(\cdot)$, which was proposed by Kivva et al. [2024]. We recover β by subtracting r_2 from r_1. Below we explain the workings of function *GetRatio*$(\cdot)$. Suppose we observe two random variables X_1 and X_2 that can be represented as

$$X_1 = a\epsilon + \epsilon_1,$$
$$X_2 = b\epsilon + \epsilon_2,$$

where ϵ_1, ϵ_2 and ϵ are mutually independent. Then *GetRatio*(X_1, X_2) computes the ratio a/b under the same assumption on the distribution of ϵ as in Theorem 3.3 imposes on the distribution of ϵ_t. Moreover, it is important to emphasize that algorithm *GetRatio*$(\cdot)$ does not require the knowledge of the constant n; it only requires the existence of such n. For further details, see Kivva et al. [2024].

3.3 THE CASE $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$

Here, we assume that the causal effect of the latent confounder on the treatment varies across environments. For simplicity of notation, specifically for this setting, we rescale the exogenous noise ϵ_u together with α and $\gamma^{(i)}$ in the SCMs $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ so that $\epsilon_u \leftarrow \alpha\epsilon_u$, $\alpha \leftarrow 1$, $\gamma^{(i)} \leftarrow \gamma^{(i)}/\alpha$. Here we used that α and ϵ_u are invariant across environments, and therefore the corresponding SCM after rescaling will take the following form:

$$\begin{cases} U^{(i)} := \epsilon_u, \\ T^{(i)} := U^{(i)} + \epsilon_t, \\ Y^{(i)} := \beta T^{(i)} + \gamma^{(i)} U^{(i)} + \epsilon_y, \end{cases} \tag{3}$$

where ϵ_u and $\gamma^{(i)}$ are rescaled and the distributions of all exogenous noises are the same across environments, but not there realizations.

Theorem 3.4. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$. Suppose $\exists n \in \mathbb{N}$ such that $\mathbb{E}\left[\epsilon_t^n\right] \neq (n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$. Then under Assumptions 1-4, the treatment effect β can be recovered uniquely.*

The proof of this theorem can be found in Appendix A. Below, we provide a procedure for estimating the treatment effect β, which consists of multiple steps.

Step 1. First, we compute $2\beta + \gamma^{(1)} + \gamma^{(2)}$ as

$$2\beta + \gamma^{(1)} + \gamma^{(2)} = \frac{\mathbb{E}\left[\left(Y^{(2)}\right)^2 - \left(Y^{(1)}\right)^2\right]}{\mathbb{E}\left[Y^{(2)}T^{(2)} - Y^{(1)}T^{(1)}\right]}. \tag{4}$$

We then define new variables $X^{(1)}$, $X^{(2)}$ as follows

$$X^{(i)} := \left(2\beta + \gamma^{(i)} + \gamma^{(i)}\right) T^{(i)} - 2Y^{(i)}. \tag{5}$$

We also define $a := \gamma^{(2)} - \gamma^{(1)}$, and $b := \gamma^{(2)} + \gamma^{(1)}$.

Step 2. We use the following equations to compute the values $\tilde{a} := a\mathbb{E}[\epsilon_u^2]$ and $\tilde{b} := b\mathbb{E}[\epsilon_t^2]$:

$$a\mathbb{E}\left[\epsilon_u^2\right] = \frac{1}{2}\left(\mathbb{E}\left[T^{(1)}X^{(1)}\right] - \mathbb{E}\left[T^{(2)}X^{(2)}\right]\right),$$
$$b\mathbb{E}\left[\epsilon_t^2\right] = \frac{1}{2}\left(\mathbb{E}\left[T^{(1)}X^{(1)}\right] + \mathbb{E}\left[T^{(2)}X^{(2)}\right]\right).$$

Since $\mathbb{E}\left[\epsilon_u^2\right]$ and $\mathbb{E}\left[\epsilon_t^2\right]$ are always positive, $\tilde{a}$ and $\tilde{b}$ reveal the signs of a and b.

Step 3. For every n, we define $\phi_n^{(i)}$ as follows. If n is odd,

$$\phi_n^{(i)} = \mathbb{E}\big[\big(T^{(i)}\big)^{n-1} X^{(i)}\big],$$

and if n is even,

$$\phi_n^{(1)} = \mathbb{E}\big[X^{(1)}\big(T^{(1)}\big)^{n-1}\big] - (n-1)(\tilde{a}+\tilde{b})\mathbb{E}\big[\big(T^{(1)}\big)^{n-2}\big],$$
$$\phi_n^{(2)} = \mathbb{E}\big[X^{(2)}\big(T^{(2)}\big)^{n-1}\big] - (n-1)(\tilde{b}-\tilde{a})\mathbb{E}\big[\big(T^{(2)}\big)^{n-2}\big].$$

In this step, we find the smallest $n \in \mathbb{N}$ such that at least one of the values $\{\phi_n^{(i)}\}_{i=1}^2$ is non-zero. We denote this value by n^*. Note that such n^* exists under the assumption imposed on ϵ_t in Theorem 3.4 (see the proof for the details).

Step 4. Define $\psi_j^{(1)}$ and $\psi_j^{(2)}$ for every positive integer j as follows. If n^* is odd,

$$\psi_j^{(i)} = \mathbb{E}\big[\big(T^{(i)}\big)^{n^*-j}\big(X^{(i)}\big)^j\big],$$

and if n^* is even,

$$\psi_j^{(1)} = \mathbb{E}\big[\big(X^{(1)}\big)^{n^*-j}\big(T^{(1)}\big)^j\big] - (n^*-1)(\tilde{a}+\tilde{b})\mathbb{E}\big[\big(X^{(1)}\big)^{n^*-2}\big],$$
$$\psi_j^{(2)} = \mathbb{E}\big[\big(X^{(2)}\big)^{n^*-j}\big(T^{(2)}\big)^j\big] - (n^*-1)(\tilde{b}-\tilde{a})\mathbb{E}\big[\big(X^{(2)}\big)^{n^*-2}\big].$$

Finally, two possibilities may occur based on the values of $\phi_{n^*}^{(i)}$, which are dealt with in step 5.

Step 5. *Case 1*: $\phi_{n^*}^{(1)} - \phi_{n^*}^{(2)} \neq 0$. Choose $j = 3, l = 2$ if n^* is odd, and $j = 1, l = (n^*-1)$ otherwise. The absolute value of $a = \gamma^{(2)} - \gamma^{(1)}$ can be computed via

$$|a| = \left|\frac{\psi_j^{(1)} - \psi_j^{(2)}}{\phi_{n^*}^{(1)} - \phi_{n^*}^{(2)}}\right|^{1/l},$$

and as mentioned earlier, a has the same sign as $\tilde{a}$. As a result, $\beta + \gamma^{(1)}$ can be computed as $\beta + \gamma^{(1)} = \frac{1}{2}(2\beta + \gamma^{(1)} + \gamma^{(2)} - a)$. Finally, we recover β via

$$\beta = \beta + \gamma^{(1)} - \textit{GetRatio}\left(\left(\beta + \gamma^{(1)}\right)T^{(1)} - Y^{(1)}, T^{(1)}\right).$$

Step 5. *Case 2*: $\phi_{n^*}^{(1)} - \phi_{n^*}^{(1)} = 0$. In this case, by definition of n^*, $\phi_{n^*}^{(1)} + \phi_{n^*}^{(1)} \neq 0$. We choose $j = 2, l = 1$ if n^* is odd, and $j = 1, l = (n^*-1)$ otherwise. In this case we can recover b via

$$|b| = \left|\frac{\psi_j^{(1)} + \psi_j^{(2)}}{\phi_{n^*}^{(1)} + \phi_{n^*}^{(2)}}\right|^{1/l},$$

and the fact the b and $\tilde{b}$ have the same sign. Finally, $\beta = \frac{1}{2}(2\beta + \gamma^{(1)} + \gamma^{(2)} - b)$.

3.4 THE CASE $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$

Here, we assume that only the causal effect of the latent variable on the treatment varies across environments. The SCM $\mathcal{M}^{(i)}$ corresponding to the environment (i) is then

$$\begin{cases} U^{(i)} := \epsilon_u, \\ T^{(i)} := \alpha^{(i)} U^{(i)} + \epsilon_t, \\ Y^{(i)} := \beta T^{(i)} + \gamma U^{(i)} + \epsilon_y. \end{cases} \tag{6}$$

Algorithm 3 $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$

Input: $\{T^{(i)}, Y^{(i)}\}$ and $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$

1: $h(\beta) := \mathbb{E}\big[\big(Y^{(1)} - \beta T^{(1)}\big)^2\big] - \mathbb{E}\big[\big(Y^{(2)} - \beta T^{(2)}\big)^2\big]$
2: $\beta_1, \beta_2 \leftarrow roots\,(h(\cdot))$
3: $X_i^{(j)} := Y^{(j)} - \beta_i T^{(j)}$
4: $\phi_{i,m}^{(j)} := \mathbb{E}\big[\big(X_i^{(j)}\big)^m T^{(j)}\big]$
5: $\psi_m^{(j)} := \mathbb{E}\big[\big(X_i^{(j)}\big)^m\big]$
6: $n_1 \leftarrow 2, n_2 \leftarrow 2$
7: **while** $\phi_{i,n_i-1}^{(1)} - (n_i - 1)\phi_{i,1}^{(1)}\psi_{n_i-2}^{(1)} = 0$ **do**
8: $\quad n_i \leftarrow n_i + 1$
9: **if** $n_1 = n_2$ **then**
10: $\quad i \leftarrow \arg\min_i \left| \frac{\phi_{i,n-1}^{(1)}}{\phi_{i,n-1}^{(2)}} - \frac{\phi_{i,n-1}^{(1)} - (n-1)\phi_{i,1}^{(1)}\psi_{n-2}^{(1)}}{\phi_{i,n-1}^{(2)} - (n-1)\phi_{i,1}^{(2)}\psi_{n-2}^{(2)}} \right|$
11: **else**
12: $\quad i \leftarrow \arg\max_i\, [n_i]$
13: **RETURN:** β_i

Theorem 3.5. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$. Suppose $\exists n \in \mathbb{N}$ such that $\mathbb{E}\left[\epsilon_u^n\right] \neq (n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]$. Then under Assumptions 1-4 the treatment effect β can be recovered uniquely almost surely*[2].

See Appendix A for a formal proof. The procedure for estimating the treatment effect β is presented in Algorithm 3. This algorithm takes advantage of the fact that $Y^{(i)} - \hat{\beta}T^{(i)}$ does not depend on $\alpha^{(i)}$ if and only if $\hat{\beta} = \beta$. Therefore, solving the quadratic equation $h(\beta)$ (line 1 of Algorithm 3) for β (with coefficients that can be computed from observational data) provides us with two possible candidates.We can then identify the true value of β among these two candidates using the criterion in line 10 of Algorithm 3, which is equal to zero only for the correct value of β. See the proof of Theorem 3.5 for further details.

3.5 DETECTING THE SOURCE OF CHANGE

We begin by verifying whether the varying parameter is γ. This can be done through comparing the distributions of T and Y between the environments: if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$, then the distributions of $T^{(1)}$ and $T^{(2)}$ are identical, whereas the distributions of $Y^{(1)}$ and $Y^{(2)}$ are different. If we conclude that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \neq \{\gamma\}$, then we check whether $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$. To do so we need to consider the following quantities.

$$\mathbb{E}\big[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\big] / \mathbb{E}\big[\big(T^{(1)}\big)^2 - \big(T^{(2)}\big)^2\big],$$
$$\mathbb{E}\big[\big(Y^{(1)}\big)^2 - \big(Y^{(2)}\big)^2\big] / \mathbb{E}\big[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\big].$$

If these two quantities are equal or if both their denominators are equal to zero, we conclude that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \neq \{\alpha\}$; otherwise $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$. Finally, if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \not\subset \{\gamma, \alpha\}$, then we conclude that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \subset \{\epsilon_t, \epsilon_u\}$.

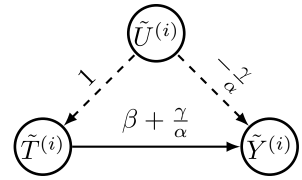


Figure 3: The causal structure corresponding to SCM $\tilde{\mathcal{M}}^{(i)}$.

3.6 NON-IDENTIFIABILITY

We shall now show that β is not identifiable if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u, \epsilon_t\}$. The SCM $\mathcal{M}^{(i)}$ corresponding to the environment i takes the following form:

$$\mathcal{M}^{(i)} = \begin{cases} U^{(i)} := \epsilon_u^{(i)}, \\ T^{(i)} := \alpha U^{(i)} + \epsilon_t^{(i)}, \\ Y^{(i)} := \beta T^{(i)} + \gamma U^{(i)} + \epsilon_y. \end{cases} \tag{7}$$

Theorem 3.6. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u, \epsilon_t\}$. The treatment effect β is not identifiable from the observational data from both domains and $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big)$.*

Proof. To prove that β is not identifiable, we will construct two new SCMs $\tilde{\mathcal{M}}^{(1)}$ and $\tilde{\mathcal{M}}^{(2)}$,

$$\tilde{\mathcal{M}}^{(i)} = \begin{cases} \tilde{U}^{(i)} := \tilde{\epsilon}_u^{(i)}, \\ \tilde{T}^{(i)} := \tilde{\alpha}\tilde{U}^{(i)} + \tilde{\epsilon}_t^{(i)}, \\ \tilde{Y}^{(i)} := \tilde{\beta}\tilde{T}^{(i)} + \tilde{\gamma}\tilde{U}^{(i)} + \tilde{\epsilon}_y. \end{cases} \tag{8}$$

such that $\mathcal{F}\big(\tilde{\mathcal{M}}^{(1)}, \tilde{\mathcal{M}}^{(2)}\big) = \{\tilde{\epsilon}_u, \tilde{\epsilon}_t\}$ and they induce the same observational distributions as $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, respectively, but the treatment effects differs for them from (7), i.e $\beta \neq \tilde{\beta}$. To do so, we utilize the counter- example presented in [Salehkaleybar et al., 2020]. Specifically, the causal structure of the new models $\tilde{\mathcal{M}}^{(i)}$ corresponding to environments $i \in \{0, 1\}$ can be seen in Figure 3 with the parameters defined as follows:

$$\tilde{\epsilon}_u^{(i)} = \epsilon_t^{(i)},\ \tilde{\epsilon}_t^{(i)} = \alpha\epsilon_u^{(i)},\ \tilde{\epsilon}_y = \epsilon_y,$$
$$\tilde{\alpha} = 1,\ \tilde{\gamma} = -\frac{\gamma}{\alpha},\ \tilde{\beta} = \beta + \frac{\gamma}{\alpha}.$$

[2] Here we consider the Lebesgue measure on the set of coefficients of linear SCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$. Then the causal effect is not identifiable only for a set of coefficients with measure zero.

Substituting these values into the set of equations 8, we obtain

$$\tilde{\mathcal{M}}^{(i)} = \begin{cases} \tilde{U}^{(i)} = \epsilon_t^{(i)}, \\ \tilde{T}^{(i)} = \epsilon_t^{(i)} + \alpha\epsilon_u^{(i)}, \\ \tilde{Y}^{(i)} = (\beta + \frac{\gamma}{\alpha})(\epsilon_t^{(i)} + \alpha\epsilon_u^{(i)}) + -\frac{\gamma}{\alpha}\epsilon_t^{(i)} + \epsilon_y, \end{cases}$$

and after regrouping and simplifications, it is easy to verify that

$$\tilde{T}^{(i)} = \alpha\epsilon_u^{(i)} + \epsilon_t^{(i)} = T^{(i)},$$
$$\tilde{Y}^{(i)} = (\alpha\beta + \gamma)\epsilon_u^{(i)} + \beta\epsilon_t^{(i)} + \epsilon_y = Y^{(i)},$$

and that $\mathcal{F}\big(\tilde{\mathcal{M}}^{(1)}, \tilde{\mathcal{M}}^{(2)}\big) = \{\tilde{\epsilon}_u, \tilde{\epsilon}_t\}$, while $\beta \neq \tilde{\beta}$. This concludes the proof. $\square$

4 RELATED WORK

In the multi-environment setting, there are two main lines of research in causality. The first aims to learn an equivalent class of possible causal structures from samples collected across multiple environments. In this context, it is typically assumed that the distributional changes across environments arise due to interventions on exogenous noises or causal mechanisms, which are unknown to the observer. Various methods have been proposed to leverage these shifts for causal discovery, including constraint-based approaches [Ghassami et al., 2017, Mooij et al., 2020, Jaber et al., 2020, Squires et al., 2020, Perry et al., 2022, Zhou et al., 2022] or score-based methods [Brouillard et al., 2020, Hägele et al., 2023, Mameche et al., 2024].

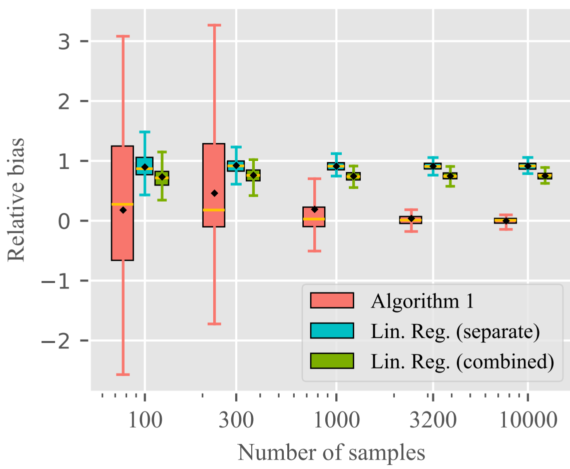
The second line of research focuses on identifying the direct causes of a target variable rather than inferring the entire causal structure. This approach is particularly relevant in settings where determining the parents of a specific variable is more critical than learning the whole causal graph. The primary objective here can be framed as a causal discovery task, which differs from causal effect identification, the main focus of our paper. However, this research direction is closely related to our work as it often assumes that causal mechanisms are linear (akin to us). Moreover, as part of the process to identify the parents of the target variable, the causal coefficients in the linear model are also often estimated.

As an example of the second line of research, Peters et al. [2016] assumed that interventions could be applied to any variable except the target variable. They proposed the Invariant Causal Prediction (ICP) method, which leverages the invariance of the conditional distribution of the target variable across environments to identify a subset of its direct causes. Additionally, they assumed no latent confounding exists between the covariates and the target variable. Subsequent research has extended this idea to more general settings, including linear models with additive interventions, nonlinear models [Heinze-Deml et al., 2018], and sequential data [Pfister et al., 2019]. For a comprehensive review of these developments, see [Bühlmann, 2020].

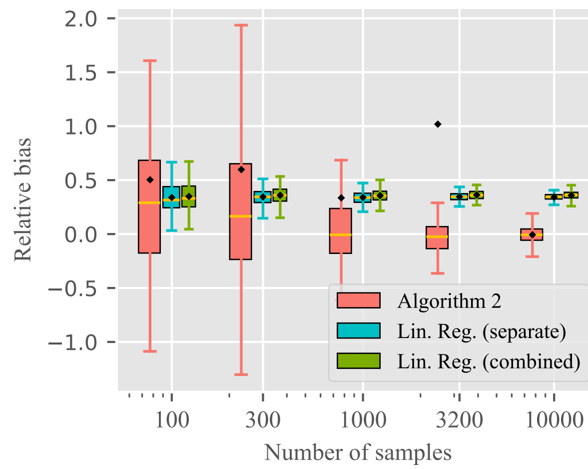
One of the drawbacks of the ICP and its extensions is high computational cost, as these approaches must search over all possible subsets of covariates to verify the invariance of the conditional distribution of the target variable. To help address this issue, there has been growing interest in leveraging optimization techniques to recover direct causes in the multi-environment setting [Rothenhäusler et al., 2019, Gimenez and Zou, 2020, Yin et al., 2024, Wang et al., 2024]. For instance, for linear models, Rothenhäusler et al. [2019] proposed Causal Dantzig exploiting "invariant inner-products" instead of the conditional invaraince in ICP. This method also allows latent confounding but not between covariates and the target variable. Gimenez and Zou [2020] proposed KL regression to identify the direct causes of a target variable in the presence of latent confounding for linear models. In this approach, the model parameters are optimized by minimizing the Kullback-Leibler (KL) divergence between two multivariate Gaussian distributions, one corresponding to the observed covariance matrix and the other to the parameterized model—across different environments. The method assumes that all causal coefficients in the linear model remain unchanged, while only the covariance matrix of the covariates or the target variable may vary across environments. Therefore, the distribution of the latent confounder is required to remain invariant.

Another related area of research is causal transportability [Bareinboim and Pearl, 2014, Lee et al., 2020], which is to identify the distribution of a causal effect in a target domain using experimental data from a source domain and observational data from the target, typically under the assumption that certain mechanisms remain invariant across domains. In contrast, our work identifies the average causal effect of a treatment on an outcome using data collected from multiple environments, without access to any experimental interventions. Moreover, our results are derived under a specific causal graph structure, for which, to the best of our knowledge, no existing transportability method offers identifiability guarantees [Bareinboim and Pearl, 2014, Lee et al., 2020].

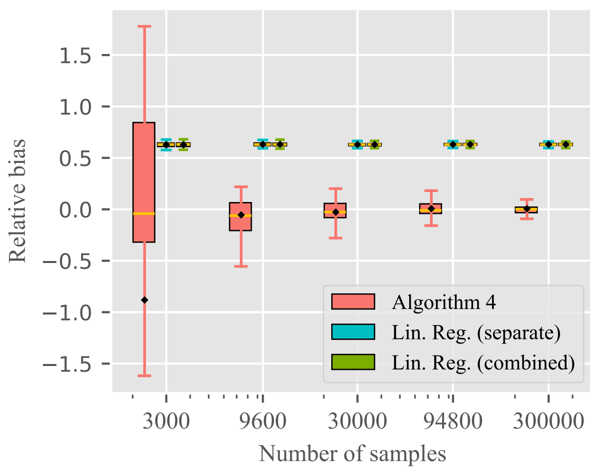
In the context of robust prediction rather than for the problem of causal effect identification, Arjovsky et al. [2019] introduced Invariant Risk Minimization (IRM), which incorporates an additional penalty term into the empirical risk function to encourage invariance of the predictor across environments. Since its introduction, IRM has been extended to various domains, including meta-learning [Bae et al., 2021], reinforcement learning [Zhang et al., 2020], and causal inference [Shi et al., 2021, Lu et al., 2021]. For linear models, Rothenhäusler et al. [2021] proposed anchor regression as a robust predictor which is an interpolation between the solutions of ordinary least squares and two-stage least square. It



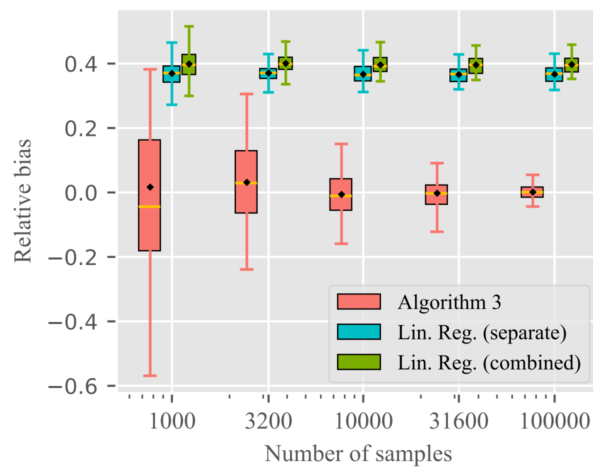
(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$



(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$



(c) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\gamma\}$



(d) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\alpha\}$

Figure 4: Relative estimation bias given data from two domains, when only ϵ_t (4a), only ϵ_t (4b), only γ (4c), and only α (4d) varies across domains.

is noteworthy that the main goal in methods such as IRM or anchor regression is to have robust prediction against distribution shifts and they do not provide any guarantee for recovering the direct causes.

5 EXPERIMENTAL RESULTS

We evaluated the performance of Algorithms 1 through 4, in terms of *relative estimation bias*, over a variety of settings. We define the relative bias as $(\frac{\hat{\beta}}{\beta} - 1)$, where β and $\hat{\beta}$ denote the true and estimated values of the parameter, respectively. Our Python code is accessible online[3].

Figure 4 illustrates the relative bias of our algorithms, and compares them to that of linear regression (ordinary least squares). For linear regression, we included two versions: (i) *separate*, which regresses the outcomes on the treatments separately in each domain, and takes the average of both estimates; and (ii) *combined*, which concatenates the data from both domains and performs a single linear regression to estimate β. Note that in our setup, most methods such as ICP reduce to linear regression. We sampled the noise variables from an exponential distribution with parameter λ chosen uniformly at random in the range $(0.9, 1.1)$. In the cases where ϵ_t or ϵ_u were changing between domains, we picked $\lambda \in (0.45, 0.55)$ as the alternate parameter. Parameters α, β, γ were uniformly sampled from $(0.4, 0.6)$, $(0.6, 0.7)$, and $(0.8, 0.9)$, respectively. In the case α or γ were changing, the alternative values were sampled uniformly from $(0.8, 0.9)$ and $(2, 2.1)$, respectively. Figures 4a, 4b, 4c and 4d represent the relative biases in estimation of β, when only one of ϵ_t, ϵ_u, γ, or α varied across the

[3]https://github.com/SinaAkbarii/IdentificationMultipleDomain

domains. The box plots show the median and 25% quantiles of relative estimation bias. As evident from all plots, our algorithms converge to the true parameter as the number of samples grows, whereas both linear regression baselines have a systematic bias regardless of the number of samples. However, for smaller number of samples, our algorithms show a higher variability (in terms of sample variance) compared to linear regression. This is expected due to more complex estimation procedures in our algorithms.

In Appendix C, we present complementary simulation results for when noises are sampled according to various probability distributions. Furthermore, Figures 12 through 16 in Appendix C depict some values of the parameters k and n which we observed running our algorithms. Interestingly, in Figure 16, for logistic distribution, the value of n is not recovered correctly when the sample size is not large enough. Therefore, Algorithm 3 will use the incorrect estimation formula to compute β. This issue is reflected in the results of the Figure 7c. However, when the sample size increases, the correct value of n is recovered, and in the next error boxes, it can be seen that our estimate is more accurate, unlike the one obtained by comparative methods. It is noteworthy that n is often known in the literature of causal effect estimation via high-order moments (e.g. [Schkoda et al., 2024]), and to the best of our knowledge, our work is the first one that does not assume it in the experiments. Additionally, the results show that for commonly encountered data distributions, the parameters k (or n) often are a small number.

6 CONCLUSION

We studied the problem of causal effect identification from observational data collected from two environments. We showed that when there is a single unknown change across two domains, we can detect whether the causal effect of the latent confounder on the treatment or the outcome had changed or the distribution of the exogenous noises of the treatment or latent confounder varied between the domains. We established that if the change occurs in the causal effect of latent confounder on other variables, the treatment effect is uniquely identifiable. Otherwise, it can be recovered up to two possible candidates. Additionally, we provided an estimation procedure tailored to each scenario and proved a non-identifiability result for identifying the treatment effect if the distribution of both exogenous noises corresponding to the treatment and latent varied across domains.

For future work, we shall explore generalizing our findings to more than two environments. While using our current results for each pair of environments, we can establish identifiability in multiple environments in some cases, there may be situations where the treatment effect is not identifiable through pairwise comparisons but becomes identifiable when more environments are considered.

References

Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jun-Hyun Bae, Inchul Choi, and Minho Lee. Meta-learned invariant risk minimization. *arXiv preprint arXiv:2103.12947*, 2021.

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27, 2014.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33: 21865–21877, 2020.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

David Card. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. *American Economic Review*, 1993.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. *Advances in Neural Information Processing Systems*, 30, 2017.

Jaime Roquero Gimenez and James Zou. Identifying invariant factors across multiple environments with kl regression. *arXiv preprint arXiv:2002.08341*, 2020.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 1411–1436. PMLR, 2023.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Guido W Imbens. Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11, 2024.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33: 9551–9561, 2020.

Yaroslav Kivva, Saber Salehkaleybar, and Negar Kiyavash. A cross-moment approach for causal effect estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

Sanghack Lee, Juan Correa, and Elias Bareinboim. General transportability–synthesizing observations and experiments from heterogeneous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10210–10217, 2020.

Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. *Advances in Neural Information Processing Systems*, 36, 2024.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

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

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):947–1012, 2016.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal dantzig. *The Annals of Statistics*, 47(3): 1688–1722, 2019.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83 (2):215–246, 2021.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.

Daniela Schkoda, Elina Robeva, and Mathias Drton. Causal discovery of linear non-gaussian causal models with unobserved confounding. *arXiv preprint arXiv:2408.04907*, 2024.

Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In *Uncertainty in artificial intelligence*, pages 1546–1555. PMLR, 2021.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

Zhenyu Wang, Yifan Hu, Peter Bühlmann, and Zijian Guo. Causal invariance learning via efficient optimization of a nonconvex objective. *arXiv preprint arXiv:2412.11850*, 2024.

Mingzhang Yin, Yixin Wang, and David M Blei. Optimization-based causal estimation from heterogeneous environments. *J. Mach. Learn. Res*, 25:1–44, 2024.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pages 11214–11224. PMLR, 2020.

Fangting Zhou, Kejun He, and Yang Ni. Causal discovery with heterogeneous observational data. In *Uncertainty in Artificial Intelligence*, pages 2383–2393. PMLR, 2022.

Causal Effect Identification in Heterogeneous Environments from Higher-Order Moments
(Supplementary Material)

Yaroslav Kivva[1] **Sina Akbari**[1] **Saber Salehkaleybar**[2] **Negar Kiyavash**[1]

[1]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Leiden Institute of Advanced Computer Science, Leiden University, Netherlands

A PROOFS

In this section we provide the proofs for the theorems introduced in main part.

Theorem 3.2. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$. Then under Assumptions 1-4, the treatment effect β can be recovered uniquely.*

Proof. Since α is not changing across the domains, the causal effect of the latent confounder $U^{(i)}$ on $T^{(i)}$ can always be set to one by appropriately rescaling both the latent confounder and γ. Hence, we can rewrite the structural equations in the two domains as follows:

$$\mathcal{M}^{(i)} = \begin{cases} U^{(i)} := \epsilon_u^{(i)}, \\ T^{(i)} := U^{(i)} + \epsilon_t^{(i)}, \\ Y^{(i)} := \beta T^{(i)} + \gamma U^{(i)} + \epsilon_y^{(i)}. \end{cases}$$

Remark: *Here γ and $\epsilon_u^{(1)}$ are already rescaled, and are actually different from the one in SCM 2.*

Let k be the smallest positive integer such that $\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^k\right] \neq \mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^k\right]$. Note, that such k will always exists, since otherwise $\epsilon_t^{(1)}$ and $\epsilon_t^{(2)}$ would be equal as distributions due to Assumption 4. Then,

$$\begin{aligned}
\mathbb{E}\left[(T^{(1)})^k - (T^{(2)})^k\right] &= \mathbb{E}\left[\left(\epsilon_u^{(1)} + \epsilon_t^{(1)}\right)^k - \left(\epsilon_u^{(2)} + \epsilon_t^{(2)}\right)^k\right] \\
&= \sum_{j=0}^{k} \binom{k}{j} \left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^j\right] \mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k-j}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^j\right] \mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k-j}\right]\right) \qquad (9) \\
&= \mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^k\right] - \mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^k\right],
\end{aligned}$$

where the third equality is due to the fact the difference term in the sum is equal to zero for $j \neq 0$ as the distribution of ϵ_u is not changing across the two domains and $\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^j\right] = \mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^j\right]$ for $j < k$ due to the definition of k. Now,

$$
\begin{aligned}
&\mathbb{E}\left[Y^{(1)}(T^{(1)})^{k-1} - Y^{(2)}(T^{(2)})^{k-1}\right] = \\
&= \mathbb{E}\left[\left((\beta+\gamma)\epsilon_u^{(1)} + \beta\epsilon_t^{(1)} + \epsilon_y^{(1)}\right)\left(\epsilon_u^{(1)} + \epsilon_t^{(1)}\right)^{k-1}\right] - \mathbb{E}\left[\left((\beta+\gamma)\epsilon_u^{(2)} + \beta\epsilon_t^{(2)} + \epsilon_y^{(2)}\right)\left(\epsilon_u^{(2)} + \epsilon_t^{(2)}\right)^{k-1}\right] \\
&\overset{(a)}{=} (\beta+\gamma)\sum_{j=0}^{k-1}\binom{k-1}{j}\left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{j+1}\right]\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k-1-j}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{j+1}\right]\mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k-1-j}\right]\right) \\
&\qquad + \beta\sum_{j=0}^{k-1}\binom{k-1}{j}\left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{j}\right]\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k-j}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{j}\right]\mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k-j}\right]\right) \\
&\overset{(b)}{=} \beta\left(\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k}\right] - \mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k}\right]\right),
\end{aligned}
\tag{10}
$$

where (a) is based on the fact that $\mathbb{E}\left[\epsilon_y^{(i)}\left(\epsilon_u^{(i)} + \epsilon_t^{(i)}\right)^{k-1}\right] = \mathbb{E}\left[\epsilon_y^{(i)}\right] E\left[\left(\epsilon_u^{(i)} + \epsilon_t^{(i)}\right)^{k-1}\right] = 0$ for $i \in \{1, 2\}$ as the exogenous noises are independent and mean zero. Moreover, (b) is due to the fact that all difference terms in the first sum are zero and in the second sum, only the difference for $j = 0$ is nonzero because of the definition of k and unchanging distribution of ϵ_u across the domains. Now, by dividing (10) by (9), we have:

$$
\beta = \frac{\mathbb{E}\left[Y^{(1)}(T^{(1)})^{k-1} - Y^{(2)}(T^{(2)})^{k-1}\right]}{\mathbb{E}\left[(T^{(1)})^{k} - (T^{(2)})^{k}\right]}, \tag{11}
$$

which shows the causal effect of T on Y is identifiable from the two domains in case only the distribution of ϵ_t is changing across the domains. $\square$

Theorem 3.3. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$. Suppose that $\exists n \in \mathbb{N}$ such that $\mathbb{E}[\epsilon_t^n] \neq (n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$. Then under Assumptions 1-4, the treatment effect β can be recovered uniquely.*

Proof. Similar to the theorem 3.2, we write $\mathbb{E}\left[(T^{(1)})^k - (T^{(2)})^k\right]$ and $\mathbb{E}\left[Y^{(1)}(T^{(1)})^{k-1} - Y^{(2)}(T^{(2)})^{k-1}\right]$ based on the moments of exogenous noises. Herein, let k be the smallest positive integer such that $\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^k\right] \neq \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^k\right]$. Then,

$$
\begin{aligned}
\mathbb{E}\left[(T^{(1)})^k - (T^{(2)})^k\right] &= \mathbb{E}\left[\left(\epsilon_u^{(1)} + \epsilon_t^{(1)}\right)^k - \left(\epsilon_u^{(2)} + \epsilon_t^{(2)}\right)^k\right] \\
&= \sum_{j=0}^{k}\binom{k}{j}\left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{j}\right]\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k-j}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{j}\right]\mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k-j}\right]\right) \\
&= \mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{k}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{k}\right],
\end{aligned}
\tag{12}
$$

where the third equality is due to the fact the difference term in the sum is equal to zero for $j \neq k$ as the distribution of ϵ_u is not changing across the two domains and $\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^j\right] = \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^j\right]$ for $j < k$ due to the definition of k. Moreover,

$$
\begin{aligned}
&\mathbb{E}\left[Y^{(1)}(T^{(1)})^{k-1} - Y^{(2)}(T^{(2)})^{k-1}\right] = \\
&= \mathbb{E}\left[\left((\beta+\gamma)\epsilon_u^{(1)} + \beta\epsilon_t^{(1)} + \epsilon_y^{(1)}\right)\left(\epsilon_u^{(1)} + \epsilon_t^{(1)}\right)^{k-1}\right] - \mathbb{E}\left[\left((\beta+\gamma)\epsilon_u^{(2)} + \beta\epsilon_t^{(2)} + \epsilon_y^{(2)}\right)\left(\epsilon_u^{(2)} + \epsilon_t^{(2)}\right)^{k-1}\right] \\
&\overset{(a)}{=} (\beta+\gamma)\sum_{j=0}^{k-1}\binom{k-1}{j}\left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{j+1}\right]\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k-1-j}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{j+1}\right]\mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k-1-j}\right]\right) \\
&\qquad + \beta\sum_{j=0}^{k-1}\binom{k-1}{j}\left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{j}\right]\mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^{k-j}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{j}\right]\mathbb{E}\left[\left(\epsilon_t^{(2)}\right)^{k-j}\right]\right) \\
&\overset{(b)}{=} (\beta+\gamma)\left(\mathbb{E}\left[\left(\epsilon_u^{(1)}\right)^{k}\right] - \mathbb{E}\left[\left(\epsilon_u^{(2)}\right)^{k}\right]\right),
\end{aligned}
\tag{13}
$$

where (a) is based on the fact that $\mathbb{E}\left[\epsilon_y^{(i)}\left(\epsilon_u^{(i)}+\epsilon_t^{(i)}\right)^{k-1}\right]=\mathbb{E}\left[\epsilon_y^{(i)}\right]E\left[\left(\epsilon_u^{(i)}+\epsilon_t^{(i)}\right)^{k-1}\right]=0$ for $i\in\{1,2\}$ as the exogenous noises are independent and mean zero. Moreover, (b) is due to the fact that all difference terms in the second sum are zero, and in the first sum, only the difference for $j=k$ is nonzero because of the definition of k and unchanging distribution of ϵ_t across the domains. Therefore, based on (12) and (13), we can obtain the value of $\beta+\gamma$ as follows:

$$\beta+\gamma=\frac{\mathbb{E}\left[Y^{(1)}(D^{(1)})^{k-1}-Y^{(2)}(D^{(2)})^{k-1}\right]}{\mathbb{E}\left[(D^{(1)})^{k}-(D^{(2)})^{k}\right]}. \tag{14}$$

Now, in any domain $i\in\{1,2\}$, consider the following two equations:

$$\begin{cases} T^{(i)}:=U^{(i)}+\epsilon_t^{(i)}=\epsilon_u^{(i)}+\epsilon_t^{(i)},\\ Y^{(i)}-(\beta+\gamma)T^{(i)}=-\gamma\epsilon_t^{(i)}+\epsilon_y^{(i)}.\end{cases}$$

Utilizing the cross-moment approach [Kivva et al., 2024] and more specifically Kivva et al. [2024][Theorem 1], we can identify the value of γ from $T^{(i)}$ and $Y^{(i)}-(\beta+\gamma)T^{(i)}$ in any domain $i\in\{1,2\}$ given the assumption on ϵ_t in theorem. Therefore, this finishes the proof that β is uniquely identifiable. □

Theorem 3.4. *Suppose $\mathcal{M}^{(1)},\mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)},\mathcal{M}^{(2)}\big)=\{\gamma\}$. Suppose $\exists n\in\mathbb{N}$ such that $\mathbb{E}\left[\epsilon_t^n\right]\neq(n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$. Then under Assumptions 1-4, the treatment effect β can be recovered uniquely.*

Proof.

$$\mathcal{M}^{(i)}\begin{cases} U^{(i)}:=\epsilon_u^{(i)},\\ T^{(i)}:=U^{(i)}+\epsilon_t^{(i)},\\ Y^{(i)}:=\beta T^{(i)}+\gamma^{(i)}U^{(i)}+\epsilon_y^{(i)},\end{cases}$$

For ease of notation, we omit superscripts corresponding to the exogenous noises where this is not important, since their distributions remain the same across the domains. Then,

$$\begin{aligned}&\mathbb{E}\left[Y^{(2)}T^{(2)}-Y^{(1)}T^{(1)}\right]=\left(\gamma^{(2)}-\gamma^{(1)}\right)\mathbb{E}\left[\epsilon_u^2\right],\\ &\mathbb{E}\left[\left(Y^{(2)}\right)^2-\left(Y^{(1)}\right)^2\right]=\left(\left(\beta+\gamma^{(2)}\right)^2-\left(\beta+\gamma^{(1)}\right)^2\right)\mathbb{E}\left[\epsilon_u^2\right]=\left(\gamma^{(2)}-\gamma^{(1)}\right)\left(2\beta+\gamma^{(1)}+\gamma^{(2)}\right)\mathbb{E}\left[\epsilon_u^2\right].\end{aligned}$$

Note that in the above equations all the terms corresponding to the exogenous noises of observed variables are canceled out. From the two equations, we can compute $2\beta+\gamma^{(1)}+\gamma^{(2)}$ from the observational distribtuion.

Let us define the random variables $X^{(1)}$ and $X^{(2)}$ as follows:

$$\begin{aligned}X^{(1)}&=\left(2\beta+\gamma^{(1)}+\gamma^{(2)}\right)T^{(1)}-2Y^{(1)}=\left(2\beta+\gamma^{(1)}+\gamma^{(2)}\right)\left(\epsilon_u^{(1)}+\epsilon_t^{(1)}\right)-2\left(\left(\beta+\gamma^{(1)}\right)\epsilon_u^{(1)}+\beta\epsilon_t^{(1)}+\epsilon_y^{(1)}\right),\\ X^{(2)}&=\left(2\beta+\gamma^{(1)}+\gamma^{(2)}\right)T^{(2)}-2Y^{(2)}=\left(2\beta+\gamma^{(1)}+\gamma^{(2)}\right)\left(\epsilon_u^{(2)}+\epsilon_t^{(2)}\right)-2\left(\left(\beta+\gamma^{(2)}\right)\epsilon_u^{(2)}+\beta\epsilon_t^{(2)}+\epsilon_y^{(2)}\right).\end{aligned}$$

Let us define $a=\gamma^{(2)}-\gamma^{(1)}$, $b=\gamma^{(2)}+\gamma^{(1)}$, $\epsilon^{(i)}=-2\epsilon_y^{(i)}$. Then,

$$\begin{aligned}X^{(1)}&=\left(\gamma^{(2)}-\gamma^{(1)}\right)\epsilon_u^{(1)}+\left(\gamma^{(1)}+\gamma^{(2)}\right)\epsilon_t^{(1)}-2\epsilon_y^{(1)}=a\epsilon_u^{(1)}+b\epsilon_t^{(1)}+\epsilon^{(1)},\\ X^{(2)}&=\left(\gamma^{(1)}-\gamma^{(2)}\right)\epsilon_u^{(2)}+\left(\gamma^{(1)}+\gamma^{(2)}\right)\epsilon_t^{(2)}-2\epsilon_y^{(2)}=-a\epsilon_u^{(2)}+b\epsilon_t^{(2)}+\epsilon^{(2)}.\end{aligned}$$

Let us consider the following expectations:

$$\begin{aligned}&\mathbb{E}\left[\left(T^{(1)}\right)^2\right]=\mathbb{E}\left[\left(T^{(2)}\right)^2\right]=\mathbb{E}\left[\epsilon_u^2\right]+\mathbb{E}\left[\epsilon_t^2\right],\\ &\mathbb{E}\left[T^{(1)}X^{(1)}\right]=\left(\gamma^{(2)}-\gamma^{(1)}\right)\mathbb{E}\left[\epsilon_u^2\right]+\left(\gamma^{(1)}+\gamma^{(2)}\right)\mathbb{E}\left[\epsilon_t^2\right]=a\mathbb{E}\left[\epsilon_u^2\right]+b\mathbb{E}\left[\epsilon_t^2\right],\\ &\mathbb{E}\left[T^{(2)}X^{(2)}\right]=\left(\gamma^{(1)}-\gamma^{(2)}\right)\mathbb{E}\left[\epsilon_u^2\right]+\left(\gamma^{(1)}+\gamma^{(2)}\right)\mathbb{E}\left[\epsilon_t^2\right]=-a\mathbb{E}\left[\epsilon_u^2\right]+b\mathbb{E}\left[\epsilon_t^2\right].\end{aligned}$$

The difference and sum of the expectations above give us:

$$\hat{a} := \frac{1}{2}\left(\mathbb{E}\left[T^{(1)}X^{(1)} - T^{(2)}X^{(2)}\right]\right) = \left(\gamma^{(2)} - \gamma^{(1)}\right)\mathbb{E}\left[\epsilon_u^2\right] = a\mathbb{E}\left[\epsilon_u^2\right], \tag{15}$$

$$\hat{b} := \frac{1}{2}\left(\mathbb{E}\left[T^{(1)}X^{(1)} - T^{(2)}X^{(2)}\right]\right) = \left(\gamma^{(1)} + \gamma^{(2)}\right)\mathbb{E}\left[\epsilon_t^2\right] = b\mathbb{E}\left[\epsilon_t^2\right]. \tag{16}$$

Note that from (15)-(16), we can deduce the sign of a and b. Now, we show how to obtain either a or b. If we recover $b = \gamma^{(1)} + \gamma^{(2)}$, then from knowing $\beta + \gamma^{(1)} + \gamma^{(2)}$ we compute β. If we recover a, then from knowing $\beta + \gamma^{(1)} + \gamma^{(2)}$ we can compute $\beta + \gamma^{(i)}$. From this point, we can use the same method proposed in proof of theorem 3.3 to compute β.

Consider,

$$\mathbb{E}\left[\left(X^{(1)}\right)^3\right] = a^3\mathbb{E}\left[\epsilon_u^3\right] + b^3\mathbb{E}\left[\epsilon_t^3\right] + \mathbb{E}\left[\epsilon^3\right], \tag{17}$$

$$\mathbb{E}\left[\left(X^{(2)}\right)^3\right] = -a^3\mathbb{E}\left[\epsilon_u^3\right] + b^3\mathbb{E}\left[\epsilon_t^3\right] + \mathbb{E}\left[\epsilon^3\right], \tag{18}$$

$$\implies \mathbb{E}\left[\left(X^{(1)}\right)^3 - \left(X^{(2)}\right)^3\right] = a^3\mathbb{E}\left[\epsilon_u^3\right]. \tag{19}$$

On the other hand,

$$\mathbb{E}\left[X^{(1)}\left(T^{(1)}\right)^2\right] = a\mathbb{E}\left[\epsilon_u^3\right] + b\mathbb{E}\left[\epsilon_t^3\right] \tag{20}$$

$$\mathbb{E}\left[X^{(2)}\left(T^{(2)}\right)^2\right] = -a\mathbb{E}\left[\epsilon_u^3\right] + b\mathbb{E}\left[\epsilon_t^3\right] \tag{21}$$

$$\mathbb{E}\left[\left(X^{(1)}\right)\left(T^{(1)}\right)^2 - \left(X^{(2)}\right)\left(T^{(2)}\right)^2\right] = a\mathbb{E}\left[\epsilon_u^3\right]. \tag{22}$$

Case 1: $\mathbb{E}\left[\epsilon_u^3\right] \neq 0$. Then the ratio between (19) and (22) gives a^2. We can identify the value of a as we know the sign of a from (15).
Case 2: $\mathbb{E}\left[\epsilon_u^3\right] = 0$ and $\mathbb{E}\left[\epsilon_t^3\right] \neq 0$. Then:

$$\mathbb{E}\left[X^{(1)}\left(T^{(1)}\right)^2\right] = a\mathbb{E}\left[\epsilon_u^3\right] + b\mathbb{E}\left[\epsilon_t^3\right] = b\mathbb{E}\left[\epsilon_t^3\right],$$

$$\mathbb{E}\left[\left(X^{(1)}\right)^2 T^{(1)}\right] = a^2\mathbb{E}\left[\epsilon_u^3\right] + b^2\mathbb{E}\left[\epsilon_t^3\right] = b^2\mathbb{E}\left[\epsilon_t^3\right].$$

From the two above equations, we can compute the value of b.
Case 3: $\mathbb{E}\left[\epsilon_u^3\right] = 0$, $\mathbb{E}\left[\epsilon_t^3\right] = 0$ and $n \in \mathbb{N}$ - the smallest number such that one of the following equations hold:

- $\mathbb{E}\left[\epsilon_u^n\right] \neq (n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]$.
- $\mathbb{E}\left[\epsilon_t^n\right] \neq (n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$.

Then

$$\begin{aligned}
&\mathbb{E}\left[\left(T^{(1)}\right)^{n-1}X^{(1)}\right] = \mathbb{E}\left[\left(a\epsilon_u + b\epsilon_t + \epsilon\right)\left(\epsilon_u + \epsilon_t\right)^{n-1}\right] \\
&\overset{*}{=} a\sum_{k=0}^{n-1}\binom{n-1}{k}\mathbb{E}\left[\epsilon_u^{k+1}\right]\mathbb{E}\left[\epsilon_t^{n-k-1}\right] + b\sum_{k=0}^{n-1}\binom{n-1}{k}\mathbb{E}\left[\epsilon_t^{k+1}\right]\mathbb{E}\left[\epsilon_u^{n-k-1}\right] \\
&\overset{**}{=} a\sum_{k=1}^{n-1}\binom{n-1}{k}\mathbb{E}\left[\epsilon_u^{k+1}\right]\mathbb{E}\left[\epsilon_t^{n-k-1}\right] + b\sum_{k=1}^{n-1}\binom{n-1}{k}\mathbb{E}\left[\epsilon_t^{k+1}\right]\mathbb{E}\left[\epsilon_u^{n-k-1}\right],
\end{aligned}$$

where $(*)$ and $(**)$ are based on the facts that exogenous noises are independent and have mean zero. On the other hand,

$$\begin{aligned}
&(n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(T^{(1)}\right)^{n-2}\right] = (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[(\epsilon_u+\epsilon_t)^{n-2}\right] \\
&= (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\sum_{k=0}^{n-2}\binom{n-2}{k}\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-k-2}\right] \\
&= a\sum_{k=1}^{n-1}k\frac{(n-1)!}{k!(n-1-k)!}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-1}\right]\mathbb{E}\left[\epsilon_t^{n-k-1}\right] \\
&= a\sum_{k=1}^{n-1}k\binom{n-1}{k}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-1}\right]\mathbb{E}\left[\epsilon_t^{n-k-1}\right].
\end{aligned} \tag{23}$$

Note that $T^{(1)}$ is symmetric with respect to the exogenous noises $\epsilon_u^{(1)}$ and $\epsilon_t^{(1)}$. Therefore we can obtain similar equation to (23), where $\epsilon_u^{(1)}$ and $\epsilon_t^{(1)}$ are swapped. Hence, combining it with the knowledge that $\mathbb{E}\left[\epsilon_u^k\right] = (k-1)\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_u^2\right]$ and $\mathbb{E}\left[\epsilon_t^k\right] = (k-1)\mathbb{E}\left[\epsilon_t^{k-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$ for all $k < n$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(T^{(1)}\right)^{n-1}X^{(1)}\right] - (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(T^{(1)}\right)^{n-2}\right] - (n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(T^{(1)}\right)^{n-2}\right] \\
&= a\left(\mathbb{E}\left[\epsilon_u^n\right] - (n-1)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{n-2}\right]\right) + b\left(\mathbb{E}\left[\epsilon_t^n\right] - (n-1)\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\epsilon_t^{n-2}\right]\right).
\end{aligned} \tag{24}$$

Similarly we obtain,

$$\begin{aligned}
&\mathbb{E}\left[X^{(2)}\left(T^{(2)}\right)^{n-1}\right] - (n-1)(-a)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(T^{(2)}\right)^{n-2}\right] - (n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(T^{(2)}\right)^{n-2}\right] \\
&= -a\left(\mathbb{E}\left[\epsilon_u^n\right] - (n-1)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{n-2}\right]\right) + b\left(\mathbb{E}\left[\epsilon_t^n\right] - (n-1)\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\epsilon_t^{n-2}\right]\right).
\end{aligned} \tag{25}$$

Now we will compute:

$$\mathbb{E}\left[\left(X^{(1)}\right)^{n-1}T^{(1)}\right] - (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(X^{(1)}\right)^{n-2}\right] - (n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(X^{(1)}\right)^{n-2}\right] \tag{26}$$

$$\mathbb{E}\left[\left(X^{(2)}\right)^{n-1}T^{(2)}\right] - (n-1)(-a)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(X^{(2)}\right)^{n-2}\right] - (n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(X^{(2)}\right)^{n-2}\right] \tag{27}$$

For the $\mathbb{E}\left[\left(X^{(1)}\right)^{n-1}T^{(1)}\right]$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left(X^{(1)}\right)^{n-1}T^{(1)}\right] = \mathbb{E}\left[\left(a\epsilon_u^{(1)} + b\epsilon_t^{(1)} + \epsilon^{(1)}\right)^{n-1}\left(\epsilon_u^{(1)} + \epsilon_t^{(1)}\right)\right] \\
&\overset{*}{=} \mathbb{E}\left[\epsilon_u(a\epsilon_u + b\epsilon_t + \epsilon)^{n-1} + \epsilon_t(a\epsilon_u + b\epsilon_t + \epsilon)^{n-1}\right].
\end{aligned}$$

Note that in $(*)$ we omit superscript (1) since the moments of the exogenous noises are equal across domains. Then,

$$\begin{aligned}
&\mathbb{E}\left[\epsilon_u(a\epsilon_u + b\epsilon_t + \epsilon)^{n-1}\right] = \sum_{m=0}^{n-1}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\mathbb{E}\left[\epsilon_u(a\epsilon_u + b\epsilon_t)^{n-1-m}\right] \\
&= \sum_{m=0}^{n-1}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\sum_{k=1}^{n-m}a^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-1-m}{k-1} \\
&\overset{*}{=} \sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\sum_{k=2}^{n-m}a^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-1-m}{k-1}.
\end{aligned}$$

In the last equality $(*)$, the term in the second summation corresponding to $k = 1$ is equal to zero due to the fact that exogenous noises have zero mean.

For the $(n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[(X^{(1)})^{n-2}\right]$ we have:

$$
\begin{aligned}
&(n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(X^{(1)}\right)^{n-2}\right] = (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[(a\epsilon_u + b\epsilon_t + \epsilon)^{n-2}\right] \\
&= (n-1)\sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-2}{m}\sum_{k=0}^{n-m-2} a^{k+1}b^{n-m-k-2}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-m-k-2}\right]\binom{n-m-2}{k} \\
&= (n-1)\sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-2}{m}\sum_{k=2}^{n-m} a^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-m-2}{k-2} \\
&= (n-1)\sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\frac{n-1-m}{n-1}\sum_{k=2}^{n-m} a^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-m-1}{k-1}\frac{k-1}{n-1-m} \\
&= \sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\sum_{k=2}^{n-m} a^{k-1}b^{n-m-k}(k-1)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-m-1}{k-1}.
\end{aligned}
$$

Note that $\epsilon_t(a\epsilon_u + b\epsilon_t + \epsilon)^{n-1}$ can be obtained from $\epsilon_u(a\epsilon_u + b\epsilon_t + \epsilon)^{n-1}$ by substitutions $\epsilon_u \leftrightarrow \epsilon_t$ and $a \leftrightarrow b$. Hence we have:

$$
\mathbb{E}\left[\epsilon_t\left(a\epsilon_u + b\epsilon_t + \epsilon\right)^{n-1}\right] = \sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\sum_{k=2}^{n-m} b^{k-1}a^{n-m-k}\mathbb{E}\left[\epsilon_t^k\right]\mathbb{E}\left[\epsilon_u^{n-m-k}\right]\binom{n-1-m}{k-1}.
$$

With similar logic,

$$
\begin{aligned}
&(n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(X^{(1)}\right)^{n-2}\right] = \\
&\qquad \sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon^m\right]\binom{n-1}{m}\sum_{k=2}^{n-m} b^{k-1}a^{n-m-k}(k-1)\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\epsilon_t^{k-2}\right]\mathbb{E}\left[\epsilon_u^{n-m-k}\right]\binom{n-m-1}{k-1}.
\end{aligned}
$$

Note that $\mathbb{E}\left[\epsilon_u^k\right] = (k-1)\,\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_u^2\right]$ and $\mathbb{E}\left[\epsilon_t^k\right] = (k-1)\,\mathbb{E}\left[\epsilon_t^{k-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$ for all $k < n$. Therefore (26) can be simplified as,

$$
\begin{aligned}
&\mathbb{E}\left[\left(X^{(1)}\right)^{n-1}T^{(1)}\right] - (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(X^{(1)}\right)^{n-2}\right] - (n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(X^{(1)}\right)^{n-2}\right] \\
&= a^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right] - (n-1)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{n-2}\right]\right) + b^{n-1}\left(\mathbb{E}\left[\epsilon_t^n\right] - (n-1)\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\epsilon_t^{n-2}\right]\right).
\end{aligned}
\tag{28}
$$

Since the second domain can be obtained from the first by simple substitution of $a \leftrightarrow -a$, hence

$$
\begin{aligned}
&\mathbb{E}\left[\left(X^{(2)}\right)^{n-1}T^{(2)}\right] - (n-1)a\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(X^{(2)}\right)^{n-2}\right] - (n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(X^{(2)}\right)^{n-2}\right] \\
&= (-a)^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right] - (n-1)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{n-2}\right]\right) + b^{n-1}\left(\mathbb{E}\left[\epsilon_t^n\right] - (n-1)\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\epsilon_t^{n-2}\right]\right).
\end{aligned}
\tag{29}
$$

If n is even then we can recover a or b from ((28) - (29)) \((24) - (25)) or ((28) + (29)) \((24) + (25)),respectively.

If n is odd, then $\mathbb{E}[\epsilon_u^k] = 0$ and $\mathbb{E}[\epsilon_t^k] = 0$ for all k natural odd numbers smaller than n. Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\left(T^{(1)}\right)^{n-1}X^{(1)}\right] = a\mathbb{E}\left[\epsilon_u^n\right] + b\mathbb{E}\left[\epsilon_t^n\right], \\
&\mathbb{E}\left[\left(T^{(2)}\right)^{n-1}X^{(2)}\right] = -a\mathbb{E}\left[\epsilon_u^n\right] + b\mathbb{E}\left[\epsilon_t^n\right], \\
&\mathbb{E}\left[\left(T^{(1)}\right)^{n-3}\left(X^{(1)}\right)^3\right] - \mathbb{E}\left[\left(T^{(2)}\right)^{n-3}\left(X^{(2)}\right)^3\right] = 2a^3\mathbb{E}\left[\epsilon_u^n\right].
\end{aligned}
$$

The last equation is easy to verify, since all other terms of $\mathbb{E}\left[\left(T^{(1)}\right)^{n-3}\left(X^{(1)}\right)^{3}\right]$ except $a^3\mathbb{E}\left[\epsilon_u^n\right]$ are equal to zero or have identical one in $\mathbb{E}\left[\left(T^{(2)}\right)^{n-3}\left(X^{(2)}\right)^{3}\right]$. If $\mathbb{E}\left[\epsilon_u^n\right] \neq 0$ then we can compute a. In case when $\mathbb{E}\left[\epsilon_u^n\right] = 0$ we additionally compute the following expressions,

$$\mathbb{E}\left[\left(T^{(1)}\right)^{n-1} X^{(1)}\right] = a\mathbb{E}\left[\epsilon_u^n\right] + b\mathbb{E}\left[\epsilon_t^n\right],$$
$$\mathbb{E}\left[\left(T^{(1)}\right)^{n-2}\left(X^{(2)}\right)^{2}\right] = a^2\mathbb{E}\left[\epsilon_u^n\right] + b^2\mathbb{E}\left[\epsilon_t^n\right].$$

Since $\mathbb{E}\left[\epsilon_u^n\right] = 0$, then from the above equations we can recover b. □

Theorem 3.5. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$. Suppose $\exists n \in \mathbb{N}$ such that $\mathbb{E}\left[\epsilon_u^n\right] \neq (n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]$. Then under Assumptions 1-4 the treatment effect β can be recovered uniquely almost surely*[1].

Proof.

$$\mathcal{M}^{(i)} \begin{cases} U^{(i)} := \epsilon_u^{(i)}, \\ T^{(i)} := \alpha^{(i)} U^{(i)} + \epsilon_t^{(i)}, \\ Y^{(i)} := \beta T^{(i)} + \gamma U^{(i)} + \epsilon_y^{(i)}. \end{cases}$$

Let us we consider the following quadratic equation with respect to parameter $\hat{\beta}$

$$\mathbb{E}\left[\left(Y^{(1)} - \hat{\beta}T^{(1)}\right)^2\right] - \mathbb{E}\left[\left(Y^{(2)} - \hat{\beta}T^{(2)}\right)^2\right] = 0,$$

that simplifies as

$$\mathbb{E}\left[\left(Y^{(1)}\right)^2 - \left(Y^{(2)}\right)^2\right]\hat{\beta}^2 - 2\mathbb{E}\left[Y^{(1)}T^{(1)} - Y^{(2)}T^{(2)}\right]\hat{\beta} + \mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right] = 0 \tag{30}$$

It easy to see that the following equations holds

$$Y^{(1)} - \beta T^{(1)} = Y^{(2)} - \beta T^{(2)},$$

so β will be one of the roots of the Eq. (30).

Let us suppose β^* is one of the roots of Eq. (30) and $X^{(1)}, X^{(2)}$ are defined as follows

$$X^{(i)} = Y^{(i)} - \beta^* T^{(i)} = \Big(\left(\beta - \beta^*\right)\alpha^{(i)} + \gamma\Big)\epsilon_u^{(i)} + \left(\beta - \beta^*\right)\epsilon_t^{(i)} + \epsilon_y^{(i)}.$$

For simplicity of notation let us define $a^{(i)} := \left(\beta - \beta^*\right)\alpha^{(i)} + \gamma$, $b := \beta - \beta^*$, and so

$$X^{(i)} = a^{(i)}\epsilon_u^{(i)} + b\epsilon_t^{(i)} + \epsilon_y^{(i)}.$$

Since $\mathbb{E}\left[\left(X^{(1)}\right)^2\right] = \mathbb{E}\left[\left(X^{(2)}\right)^2\right]$ it implies that $\left(a^{(1)}\right)^2 = \left(a^{(2)}\right)^2$. In case, when $\beta^* \neq \beta$ it only possible that $a^{(1)} = -a^{(2)}$, so

$$\left(\beta - \beta^*\right)\alpha^{(1)} + \gamma = -\Big(\left(\beta - \beta^*\right)\alpha^{(2)} + \gamma\Big) \tag{31}$$
$$\implies \alpha^{(1)} + \alpha^{(2)} = -2\frac{\gamma}{\beta - \beta^*} \neq 0 \tag{32}$$

[1]Here we consider the Lebesgue measure on the set of coefficients of linear SCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$. Then the causal effect is not identifiable only for a set of coefficients with measure zero.

Additionally, we have:

$$\begin{aligned}\mathbb{E}\left[\left(X^{(i)}\right)^{n-1}T^{(i)}\right] &= \mathbb{E}\left[\left(\alpha^{(i)}\epsilon_u+\epsilon_t\right)\left(a^{(i)}\epsilon_u+b\epsilon_t+\epsilon_y\right)^{n-1}\right]\\ &= \alpha^{(i)}\mathbb{E}\left[\epsilon_u\left(a^{(i)}\epsilon_u+b\epsilon_t+\epsilon_y\right)^{n-1}\right]+\mathbb{E}\left[\epsilon_t\left(a^{(i)}\epsilon_u+b\epsilon_t+\epsilon_y\right)^{n-1}\right].\end{aligned} \tag{33}$$

As it was done in the proof of theorem 3.4 we can get

$$\begin{aligned}\mathbb{E}\left[\epsilon_u(a^{(i)}\epsilon_u+b\epsilon_t+\epsilon_y)^{n-1}\right] &= \sum_{m=0}^{n-1}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-1}{m}\mathbb{E}\left[\epsilon_u(a^{(i)}\epsilon_u+b\epsilon_t)^{n-1-m}\right]\\ &= \sum_{m=0}^{n-1}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-1}{m}\sum_{k=1}^{n-m}(a^{(i)})^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-1-m}{k-1}\\ &\overset{*}{=} \sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-1}{m}\sum_{k=2}^{n-m}(a^{(i)})^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-1-m}{k-1},\end{aligned} \tag{34}$$

and

$$\begin{aligned}&(n-1)a^{(i)}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(X^{(i)}\right)^{n-2}\right] = (n-1)a^{(i)}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\left(a^{(i)}\epsilon_u+b\epsilon_t+\epsilon_y\right)^{n-2}\right]\\ &= (n-1)\sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-2}{m}\sum_{k=0}^{n-m-2}(a^{(i)})^{k+1}b^{n-m-k-2}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^k\right]\mathbb{E}\left[\epsilon_t^{n-m-k-2}\right]\binom{n-m-2}{k}\\ &= (n-1)\sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-2}{m}\sum_{k=2}^{n-m}(a^{(i)})^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-m-2}{k-2}\\ &= (n-1)\sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-1}{m}\frac{n-1-m}{n-1}\sum_{k=2}^{n-m}(a^{(i)})^{k-1}b^{n-m-k}\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-m-1}{k-1}\frac{k-1}{n-1-m}\\ &= \sum_{m=0}^{n-2}\mathbb{E}\left[\epsilon_y^m\right]\binom{n-1}{m}\sum_{k=2}^{n-m}(a^{(i)})^{k-1}b^{n-m-k}(k-1)\mathbb{E}\left[\epsilon_u^2\right]\mathbb{E}\left[\epsilon_u^{k-2}\right]\mathbb{E}\left[\epsilon_t^{n-m-k}\right]\binom{n-m-1}{k-1}.\end{aligned} \tag{35}$$

Note that the similar equations can be obtained for $\mathbb{E}\left[\epsilon_t(a^{(i)}\epsilon_u+b\epsilon_t+\epsilon_y)^{n-1}\right]$ and $(n-1)b\mathbb{E}\left[\epsilon_t^2\right]\mathbb{E}\left[\left(X^{(i)}\right)^{n-2}\right]$ through the substitutions $\epsilon_u \leftrightarrow \epsilon_t$ and $a^{(i)} \leftrightarrow b$. Moreover,

$$\mathbb{E}\left[X^{(i)}T\right] = \alpha^{(i)}a^{(i)}\mathbb{E}\left[\epsilon_u^2\right]+b\mathbb{E}\left[\epsilon_t^2\right] \tag{36}$$

Then combining the Equations (33)-(36) we obtain

$$\begin{aligned}\Phi^{(i)}(\beta^*) &:= \mathbb{E}\left[\left(X^{(i)}\right)^{n-1}T^{(i)}\right]-(n-1)\mathbb{E}\left[X^{(i)}T\right]\mathbb{E}\left[\left(X^{(i)}\right)^{n-2}\right]\\ &= \alpha^{(i)}(a^{(i)})^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\right)+b^{n-1}\left(\mathbb{E}\left[\epsilon_t^n\right]-(n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]\right)\end{aligned} \tag{37}$$

Note that $b=0$ for $\beta^*=\beta$.

Suppose that n is the smallest natural number such that $\Phi^{(i)}(\beta^*)\neq 0$ for some i. This also implies that one of the following inequalities holds

- $\mathbb{E}\left[\epsilon_u^n\right]\neq(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]$,
- $\mathbb{E}\left[\epsilon_t^n\right]\neq(n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]$.

Then there are possible the following cases.

1. $\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]=0$, then $\Phi^{(1)}(\beta^*)=\Phi^{(2)}(\beta^*)\neq 0$. However the last equation for $\Phi^{(i)}$ can not happen if $\beta^*=\beta$. Indeed, if $\beta^*=\beta$ then

$$\Phi^{(i)}(\beta^*)=\alpha^{(i)}\gamma^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\right).$$

Moreover,

$$\Phi^{(1)}(\beta^*)-\Phi^{(2)}(\beta^*)=0=\left(\alpha^{(1)}-\alpha^2\right)\gamma^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\right)$$

however, the right-hand side of the equation can not be zero. This follows from inequalities $\gamma\neq 0$, $\alpha^{(1)}\neq\alpha^2$ and $\Phi^{(2)}(\beta^*)\neq 0$. Consequently, this means that if $\Phi^{(1)}(\beta^*)=\Phi^{(2)}(\beta^*)\neq 0$ then we pick wrong $\beta*$ and we should pick another root of the quadratic equation as β.

2. $\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\neq 0$. Let us assume for a moment that $\beta^*=\beta$. Then,

$$X^{(i)}=\gamma\epsilon_u^{(i)}+\epsilon_y^{(i)}\Longrightarrow\mathbb{E}\left[\left(X^{(i)}\right)^{n-1}T^{(i)}\right]\overset{(1)}{=}\mathbb{E}\left[\alpha^{(i)}\epsilon_u\left(\gamma\epsilon_u+\epsilon_y\right)^{n-1}\right]$$
$$\Longrightarrow\frac{\mathbb{E}\left[\left(X^{(1)}\right)^{n-1}T^{(1)}\right]}{\mathbb{E}\left[\left(X^{(2)}\right)^{n-1}T^{(2)}\right]}=\frac{\alpha^{(1)}}{\alpha^{(2)}}$$

and

$$\Phi^{(i)}(\beta^*)=\alpha^{(i)}\gamma^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\right)$$
$$\Longrightarrow\frac{\Phi^{(1)}(\beta^*)}{\Phi^{(2)}(\beta^*)}=\frac{\alpha^{(1)}}{\alpha^{(2)}}=\frac{\mathbb{E}\left[\left(X^{(1)}\right)^{n-1}T^{(1)}\right]}{\mathbb{E}\left[\left(X^{(2)}\right)^{n-1}T^{(2)}\right]}.$$

Moreover,

$$\frac{\mathbb{E}\left[X^{(1)}T^{(1)}\right]}{\mathbb{E}\left[X^{(2)}T^{(2)}\right]}=\frac{\alpha^{(1)}}{\alpha^{(2)}}.$$

Now let us consider the case when $\beta^*\neq\beta$. Then we have,

$$\frac{\mathbb{E}\left[X^{(1)}T^{(1)}\right]}{\mathbb{E}\left[X^{(2)}T^{(2)}\right]}=\frac{a^{(1)}\alpha^{(1)}\epsilon_u^2+b\epsilon_t^2}{a^{(2)}\alpha^{(2)}\epsilon_u^2+b\epsilon_t^2}$$
$$\frac{\Phi^{(1)}(\beta^*)}{\Phi^{(2)}(\beta^*)}=\frac{\alpha^{(1)}(a^{(1)})^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\right)+b^{n-1}\left(\mathbb{E}\left[\epsilon_t^n\right]-(n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]\right)}{\alpha^{(2)}(a^{(2)})^{n-1}\left(\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\right)+b^{n-1}\left(\mathbb{E}\left[\epsilon_t^n\right]-(n-1)\mathbb{E}\left[\epsilon_t^{n-2}\right]\mathbb{E}\left[\epsilon_t^2\right]\right)}.$$

Note that the equality

$$\frac{\mathbb{E}\left[X^{(1)}T^{(1)}\right]}{\mathbb{E}\left[X^{(2)}T^{(2)}\right]}=\frac{\Phi^{(1)}(\beta^*)}{\Phi^{(2)}(\beta^*)}\tag{38}$$

holds for the parameters $\alpha^{(1)},\alpha^{(2)},\gamma$ only for the set of Lebesgue measure zero. Indeed, an Eq 38 is equivalent to

$$\mathbb{E}\left[X^{(1)}T^{(1)}\right]\Phi^{(2)}(\beta^*)-\mathbb{E}\left[X^{(2)}T^{(2)}\right]\Phi^{(1)}(\beta^*)=0$$

that can be considered as polynomial with respect to the parameter $a^{(1)}$. It is easy to see that the coefficient near the highest degree of $a^{(1)}$ is non-zero because $\mathbb{E}\left[\epsilon_u^n\right]-(n-1)\mathbb{E}\left[\epsilon_u^{n-2}\right]\mathbb{E}\left[\epsilon_u^2\right]\neq 0$, $a^{(1)}=-a^{(2)}$, $\alpha^{(1)}\neq\alpha^{(2)}$ and $\alpha^{(1)}+\alpha^{(2)}\neq 0$ (Eq. (32)).

Consequently, by verifying whether the Eq. (38) holds we can conclude which one of the roots is the correct one. □

Theorem 3.1. *Consider two linear SCMs $\mathcal{M}^{(1)},\mathcal{M}^{(2)}$ compatible with the graph of Figure 1, such that $|\mathcal{F}(\mathcal{M}^{(1)},\mathcal{M}^{(2)})|=1$. The treatment effect β can be uniquely identified if $\mathcal{F}(\mathcal{M}^{(1)},\mathcal{M}^{(2)})\subset\{\alpha,\gamma\}$ under some additional case-specific mild assumptions; otherwise, if $\mathcal{F}(\mathcal{M}^{(1)},\mathcal{M}^{(2)})\subset\{\epsilon_t,\epsilon_u\}$, β can be identified only up to two possible candidates.*

Proof. To prove this theorem we specify a step-by-step procedure that determines which of the parameters $\alpha, \gamma, \epsilon_u, \epsilon_t$ varies across domains $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ under the assumption of infinite data.

Step 1. First we show that, we can verify that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$. Indeed, in such a case, we can statistically test whether treatment T and outcome Y are different as distributions in these two distributions. Since we assume that $\epsilon_y^{(1)}, \epsilon_y^{(2)}$ are equal as distributions ($\epsilon_y^{(1)} \overset{d}{=} \epsilon_y^{(2)}$), then under Assumption 4 the equalities $T^{(1)} \overset{d}{=} T^{(2)}$ and $Y^{(1)} \overset{d}{\neq} Y^{(2)}$ hold if and only if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$. In practice, to verify the equalities $T^{(1)} \overset{d}{=} T^{(2)}$ and $Y^{(1)} \overset{d}{\neq} Y^{(2)}$ we may use Kolmogorov-Smirnov test or any other statistical test appropriate for it.

Step 2. Knowing that $\gamma^{(1)} = \gamma^{(2)}$ we introduce a test that determines whether $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$. Let us consider the following quantities,

$$\mathbb{E}\left[\left(T^{(i)}\right)^2\right] = \mathbb{E}\left[\left(\alpha^{(i)}\right)^2\left(\epsilon_u^{(i)}\right)^2 + \left(\epsilon_t^{(i)}\right)^2\right], \tag{39}$$

$$\mathbb{E}\left[T^{(i)}Y^{(i)}\right] = \mathbb{E}\left[\left(\alpha^{(i)}\beta + \gamma\right)\left(\epsilon_u^{(i)}\right)^2 + \beta\left(\epsilon_t^{(i)}\right)^2\right], \tag{40}$$

$$\mathbb{E}\left[\left(Y^{(i)}\right)^2\right] = \mathbb{E}\left[\left(\alpha^{(i)}\beta + \gamma\right)^2\left(\epsilon_u^{(i)}\right)^2 + \beta^2\left(\epsilon_t^{(i)}\right)^2 + \epsilon_y^2\right]. \tag{41}$$

If $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$ then,

$$\begin{aligned}
\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right] &= \mathbb{E}\left[\left(\left(\alpha^{(1)}\right)^2 - \left(\alpha^{(1)}\right)^2\right)\epsilon_u^2\right], \\
\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right] &= \mathbb{E}\left[\beta\left(\left(\alpha^{(1)}\right)^2 - \left(\alpha^{(2)}\right)^2\right)\epsilon_u^2 + \gamma\left(\alpha^{(1)} - \alpha^{(2)}\right)\epsilon_u^2\right], \\
\mathbb{E}\left[\left(Y^{(1)}\right)^2 - \left(Y^{(2)}\right)^2\right] &= \mathbb{E}\left[\beta^2\left(\left(\alpha^{(1)}\right)^2 - \left(\alpha^{(2)}\right)^2\right)\epsilon_u^2 + 2\beta\gamma\left(\alpha^{(1)} - \alpha^{(2)}\right)\epsilon_u^2\right].
\end{aligned}$$

On the other hand, if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$

$$\begin{aligned}
\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right] &= \mathbb{E}\left[\alpha\left(\left(\epsilon_u^{(1)}\right)^2 - \left(\epsilon_u^{(2)}\right)^2\right)\right], \\
\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right] &= \mathbb{E}\left[\alpha\left(\alpha\beta + \gamma\right)\left(\left(\epsilon_u^{(1)}\right)^2 - \left(\epsilon_u^{(2)}\right)^2\right)\right], \\
\mathbb{E}\left[\left(Y^{(1)}\right)^2 - \left(Y^{(2)}\right)^2\right] &= \mathbb{E}\left[\left(\alpha\beta + \gamma\right)^2\left(\left(\epsilon_u^{(1)}\right)^2 - \left(\epsilon_u^{(2)}\right)^2\right)\right],
\end{aligned}$$

and if $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$

$$\begin{aligned}
\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right] &= \mathbb{E}\left[\left(\epsilon_t^{(1)}\right)^2 - \left(\epsilon_t^{(2)}\right)^2\right], \\
\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right] &= \mathbb{E}\left[\beta\left(\left(\epsilon_t^{(1)}\right)^2 - \left(\epsilon_t^{(2)}\right)^2\right)\right], \\
\mathbb{E}\left[\left(Y^{(1)}\right)^2 - \left(Y^{(2)}\right)^2\right] &= \mathbb{E}\left[\beta^2\left(\left(\epsilon_t^{(1)}\right)^2 - \left(\epsilon_t^{(2)}\right)^2\right)\right].
\end{aligned}$$

Note that for the $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$ it is easy to see that at least one of the quantities $\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right]$ or $\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right]$ is not equal to zero. Moreover,

$$\frac{\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right]}{\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right]} \neq \frac{\mathbb{E}\left[\left(Y^{(1)}\right)^2 - \left(Y^{(2)}\right)^2\right]}{\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right]}.$$

However for the case $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$ or $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$ either the following equation holds

$$\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right] = \mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right] = 0,$$

or

$$\frac{\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right]}{\mathbb{E}\left[\left(T^{(1)}\right)^2 - \left(T^{(2)}\right)^2\right]} = \frac{\mathbb{E}\left[\left(Y^{(1)}\right)^2 - \left(Y^{(2)}\right)^2\right]}{\mathbb{E}\left[T^{(1)}Y^{(1)} - T^{(2)}Y^{(2)}\right]}.$$

Since both of these equations are impossible for the case $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$, then we can use them for the verification procedure.

Now, knowing that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) \in \{\epsilon_u, \epsilon_t\}$ we will show that it is impossible to identify β uniquely. To prove it, it is enough to consider the similar construction of models $\mathcal{M}^{(i)}$ and $\hat{\mathcal{M}}^{(i)}$ presented in the proof of Theorem 3.6. Indeed, since we do not know which parameter of the parameters ϵ_u or ϵ_t may vary across the environments, then both of the models are possible SCMs that concludes the proof. □

Proposition 1. *Suppose $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ are linear SCMs compatible with the DAG of Figure 1, such that $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_y\}$. Then treatment causal effect β is not identifiable.*

Proof. To prove that β is not identifiable, we will construct two new SCMs $\tilde{\mathcal{M}}^{(1)}$ and $\tilde{\mathcal{M}}^{(2)}$,

$$\tilde{\mathcal{M}}^{(i)} = \begin{cases} \tilde{U}^{(i)} := \tilde{\epsilon}_u, \\ \tilde{T}^{(i)} := \tilde{\alpha}\tilde{U}^{(i)} + \tilde{\epsilon}_t, \\ \tilde{Y}^{(i)} := \tilde{\beta}\tilde{T}^{(i)} + \tilde{\gamma}\tilde{U}^{(i)} + \tilde{\epsilon}_y^{(i)}. \end{cases} \tag{42}$$

such that $\mathcal{F}\big(\tilde{\mathcal{M}}^{(1)}, \tilde{\mathcal{M}}^{(2)}\big) = \{\tilde{\epsilon}_y\}$ and they induce the same observational distributions as $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, respectively, but the treatment effects are different, i.e $\beta \neq \tilde{\beta}$. To do so, we again utilize the counter- example presented in [Salehkaleybar et al., 2020].

$$\tilde{\epsilon}_u = \epsilon_t,\ \tilde{\epsilon}_t = \alpha\epsilon_u,\ \tilde{\epsilon}_y^{(i)} = \epsilon_y^{(i)},$$
$$\tilde{\alpha} = 1,\ \tilde{\gamma} = -\frac{\gamma}{\alpha},\ \tilde{\beta} = \beta + \frac{\gamma}{\alpha}.$$

Substituting these values into the set of equations 42, we obtain

$$\tilde{\mathcal{M}}^{(i)} = \begin{cases} \tilde{U}^{(i)} = \epsilon_t, \\ \tilde{T}^{(i)} = \epsilon_t + \alpha\epsilon_u, \\ \tilde{Y}^{(i)} = (\beta + \frac{\gamma}{\alpha})(\epsilon_t + \alpha\epsilon_u) + -\frac{\gamma}{\alpha}\epsilon_t + \epsilon_y^{(i)}, \end{cases}$$

and after regrouping and simplifications, it is easy to verify that

$$\tilde{T}^{(i)} = \alpha\epsilon_u + \epsilon_t = T^{(i)},$$
$$\tilde{Y}^{(i)} = (\alpha\beta + \gamma)\epsilon_u + \beta\epsilon_t + \epsilon_y^{(i)} = Y^{(i)},$$

and $\mathcal{F}\big(\tilde{\mathcal{M}}^{(1)}, \tilde{\mathcal{M}}^{(2)}\big) = \{\tilde{\epsilon}_y\}$. This concludes the proof. □

B OMITTED PSEUDO-CODE

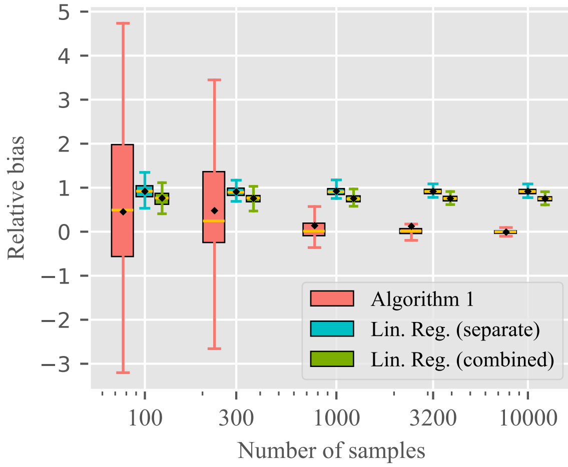
We present the pseudo-code pertaining to the estimation procedure of β when γ changes across domains, which was omitted from the main text due to space limitations.

Algorithm 4 $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$

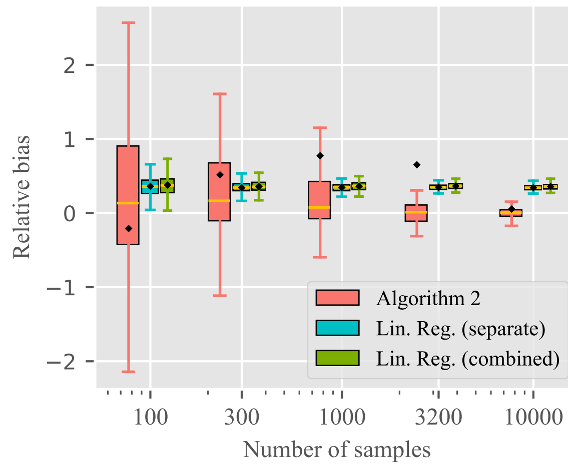
Input: $\{T^{(i)}, Y^{(i)}\}$ and $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$

$r \leftarrow \dfrac{\mathbb{E}\left[\left(Y^{(2)}\right)^2 - \left(Y^{(1)}\right)^2\right]}{\mathbb{E}\left[Y^{(2)}T^{(2)} - Y^{(1)}T^{(1)}\right]}$ $\quad \{r = 2\beta + \gamma^{(1)} + \gamma^{(2)}\}$
$X^{(i)} \leftarrow rT^{(i)} - 2Y^{(i)}$
$\tilde{a} \leftarrow \frac{1}{2}\big(\mathbb{E}\big[T^{(1)}X^{(1)}\big] - \mathbb{E}\big[T^{(2)}X^{(2)}\big]\big), \tilde{b} \leftarrow \frac{1}{2}\big(\mathbb{E}\big[T^{(1)}X^{(1)}\big] + \mathbb{E}\big[T^{(2)}X^{(2)}\big]\big)$
$n^* \leftarrow 2$
while $\phi_n^{(1)} = 0$ and $\phi_n^{(2)} = 0$ **do**
  $n^* \leftarrow n^* + 1$
if $\phi_n^{(1)} - \phi_n^{(2)} \neq 0$ **then**
  if n^* is odd **then**
    $j \leftarrow 3, l \leftarrow 2$
  else
    $j \leftarrow 1, l \leftarrow (n^* - 1)$
  $a \leftarrow sign(\tilde{b}) \left| \frac{\psi_j^{(1)} - \psi_j^{(2)}}{\phi_{n^*}^{(1)} - \phi_{n^*}^{(2)}} \right|^{1/l}$ $\quad \{a = \gamma^{(2)} - \gamma^{(1)}\}$
  $\tilde{r} \leftarrow \frac{1}{2}(r - a)$ $\quad \{\tilde{r} = \beta + \gamma^{(1)}\}$
  $\beta \leftarrow \tilde{r} - \textit{GetRatio}\big(\tilde{r}T^{(1)} - Y^{(1)}, T^{(1)}\big)$
else
  if n^* is odd **then**
    $j \leftarrow 2, l \leftarrow 1$
  else
    $j \leftarrow 1, l \leftarrow (n^* - 1)$
  $b \leftarrow sign(\tilde{b}) \left| \frac{\psi_j^{(1)} + \psi_j^{(2)}}{\phi_{n^*}^{(1)} + \phi_{n^*}^{(2)}} \right|^{1/l}$ $\quad \{b = \gamma^{(1)} + \gamma^{(2)}\}$
  $\beta \leftarrow \frac{1}{2}(r - b)$
return β

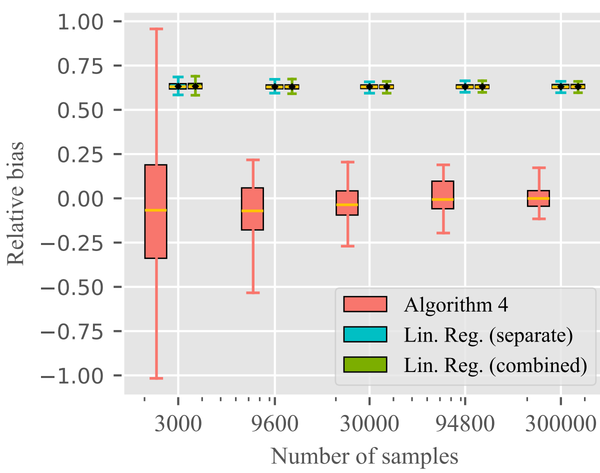
C COMPLEMENTARY EXPERIMENTAL RESULTS



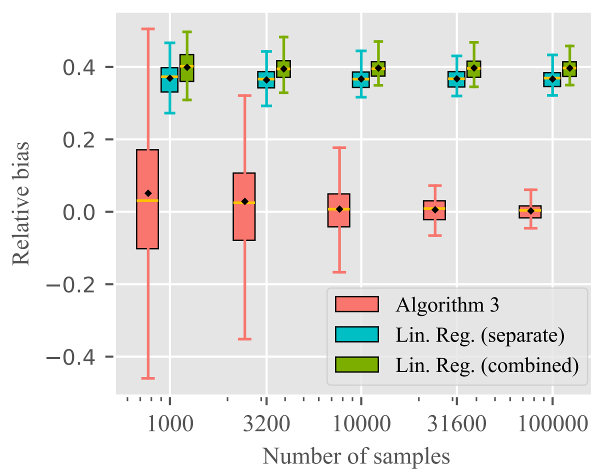
(a) $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_t\}$



(b) $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\epsilon_u\}$

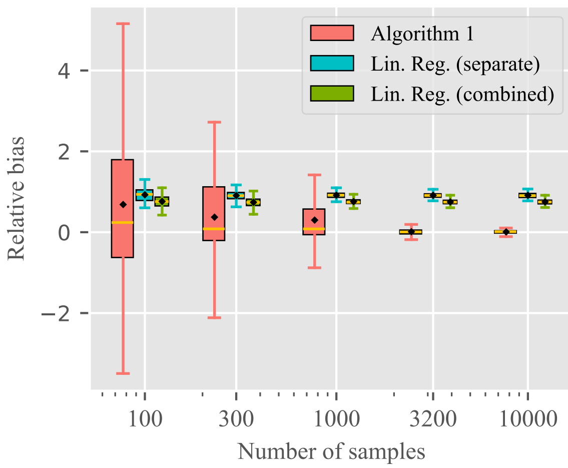


(c) $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\gamma\}$

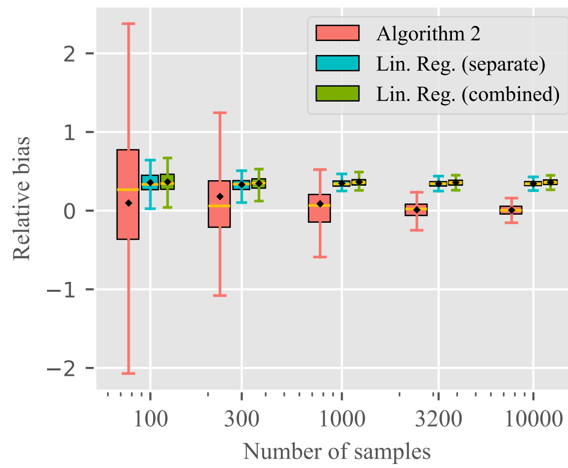


(d) $\mathcal{F}\big(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}\big) = \{\alpha\}$

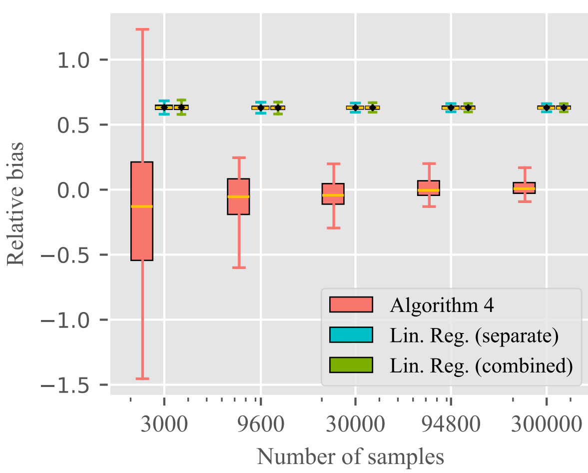
Figure 5: Relative estimation bias given data from two domains, when only ϵ_t (5a), only ϵ_t (5b), only γ (5c), and only α (5d) varies across domains. Noise variables are sampled from a Gamma distribution.



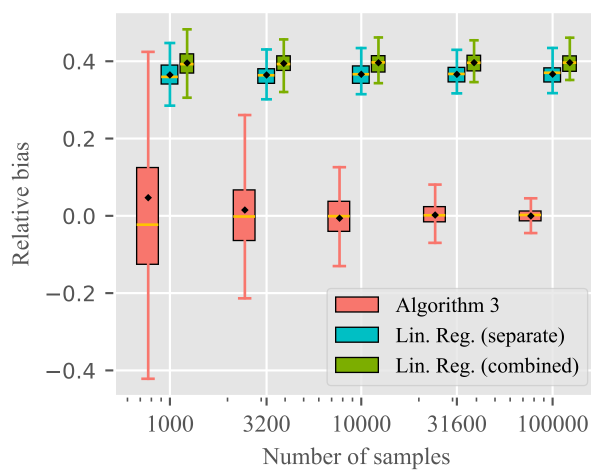
(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$



(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$

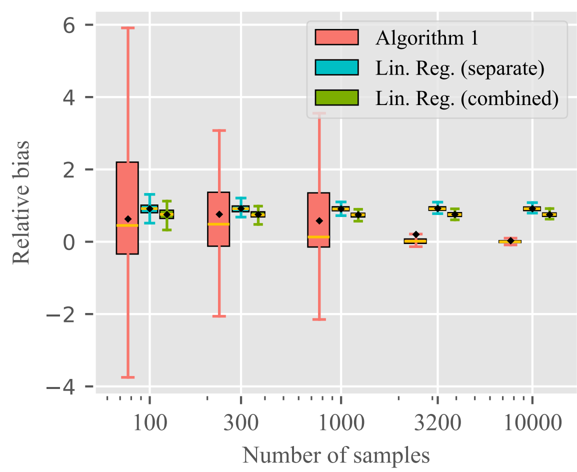


(c) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\gamma\}$

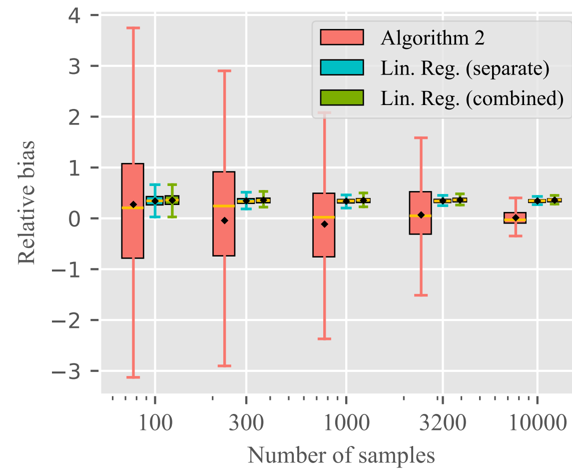


(d) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\alpha\}$

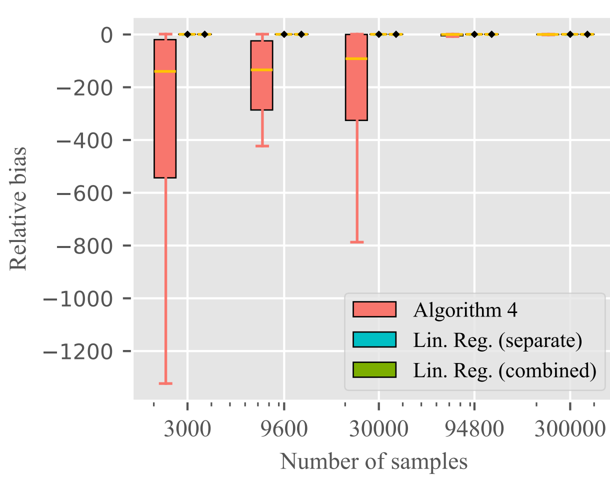
Figure 6: Relative estimation bias given data from two domains, when only ϵ_t (6a), only ϵ_t (6b), only γ (6c), and only α (6d) varies across domains. Noise variables are sampled from a Gumbel distribution.



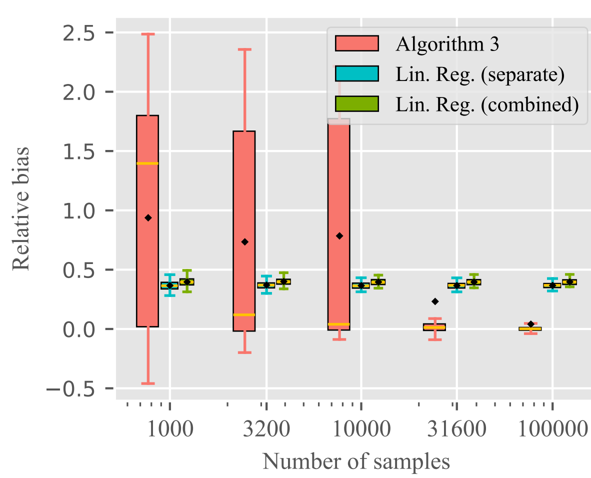
(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$



(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$

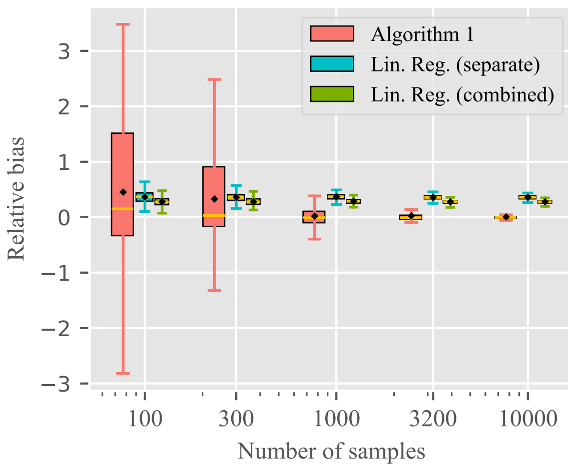


(c) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\gamma\}$

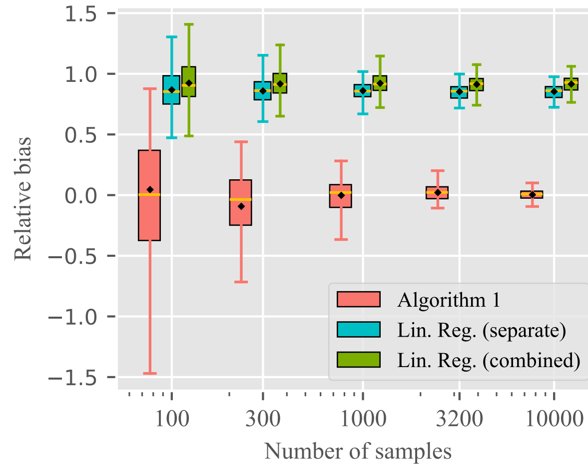


(d) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\alpha\}$

Figure 7: Relative estimation bias given data from two domains, when only ϵ_t (7a), only ϵ_t (7b), only γ (7c), and only α (7d) varies across domains. Noise variables are sampled from a Logistic distribution.

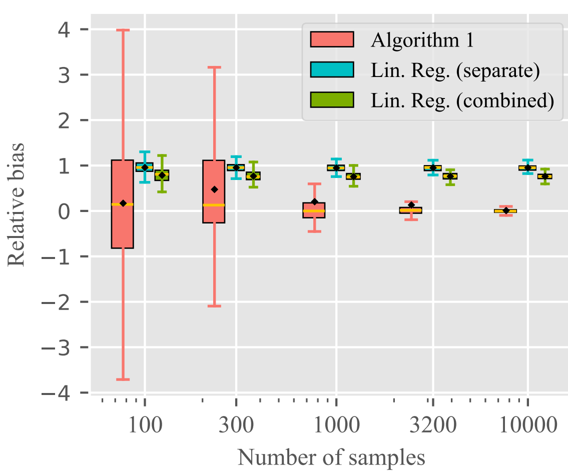


(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$

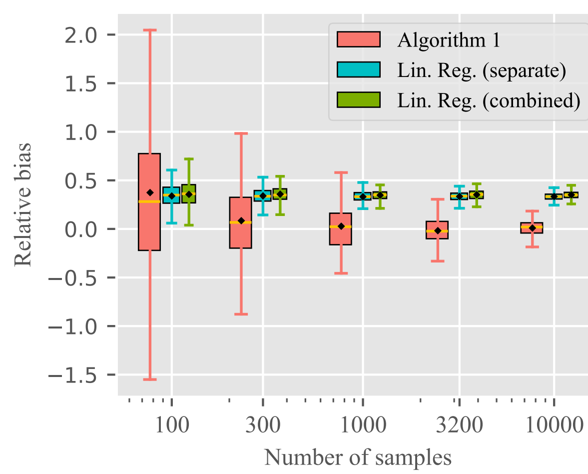


(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$

Figure 8: Relative estimation bias given data from two domains, when only ϵ_t (8a), and only ϵ_t (8b) varies across domains. All noise variables are sampled from an exponential distribution, except the alternating noise variable in the second domain which is sampled from a logistic distribution.

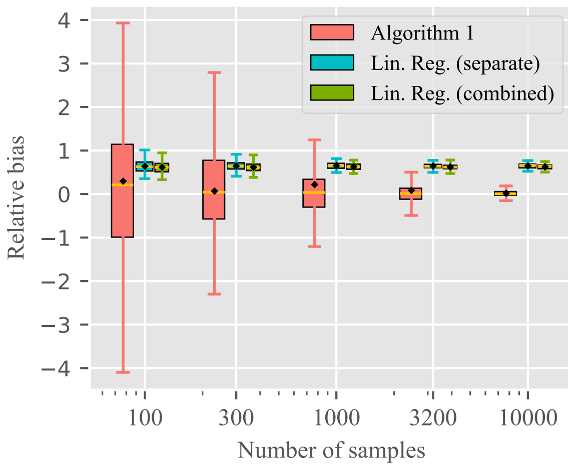


(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$

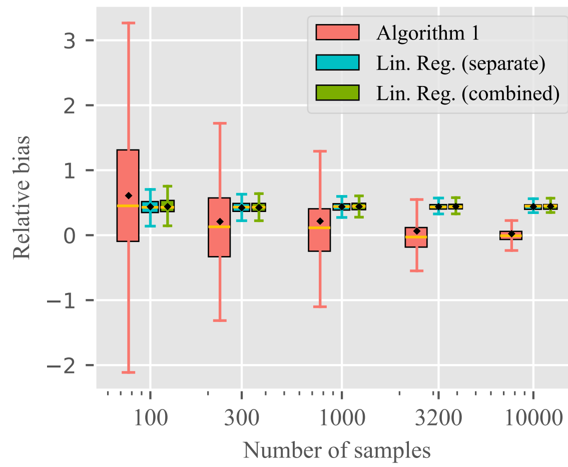


(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$

Figure 9: Relative estimation bias given data from two domains, when only ϵ_t (9a), and only ϵ_t (9b) varies across domains. All noise variables are sampled from a Gamma distribution, except the alternating noise variable in the second domain which is sampled from a uniform distribution.

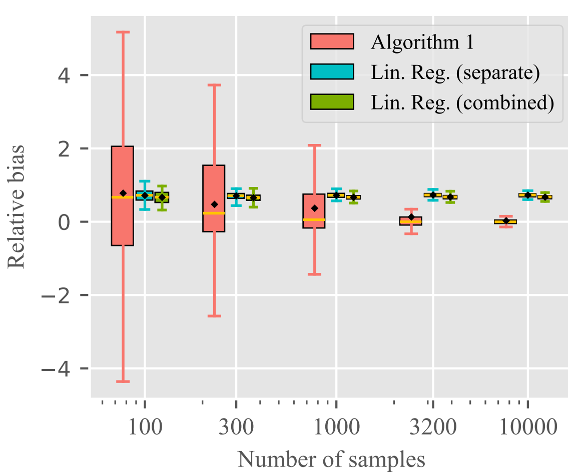


(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$

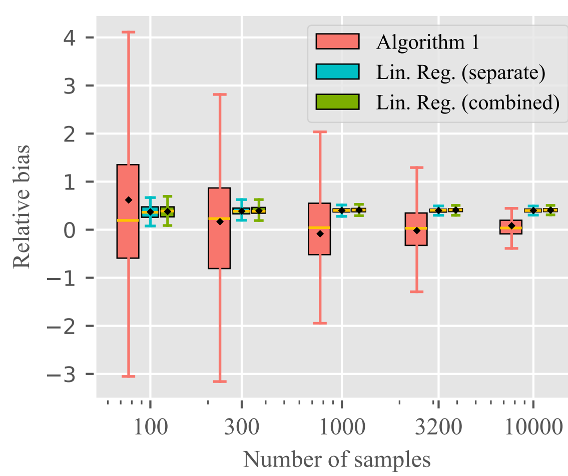


(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$

Figure 10: Relative estimation bias given data from two domains, when only ϵ_t (10a), and only ϵ_t (10b) varies across domains. All noise variables are sampled from a Gumbel distribution, except the alternating noise variable in the second domain which is sampled from an exponential distribution.

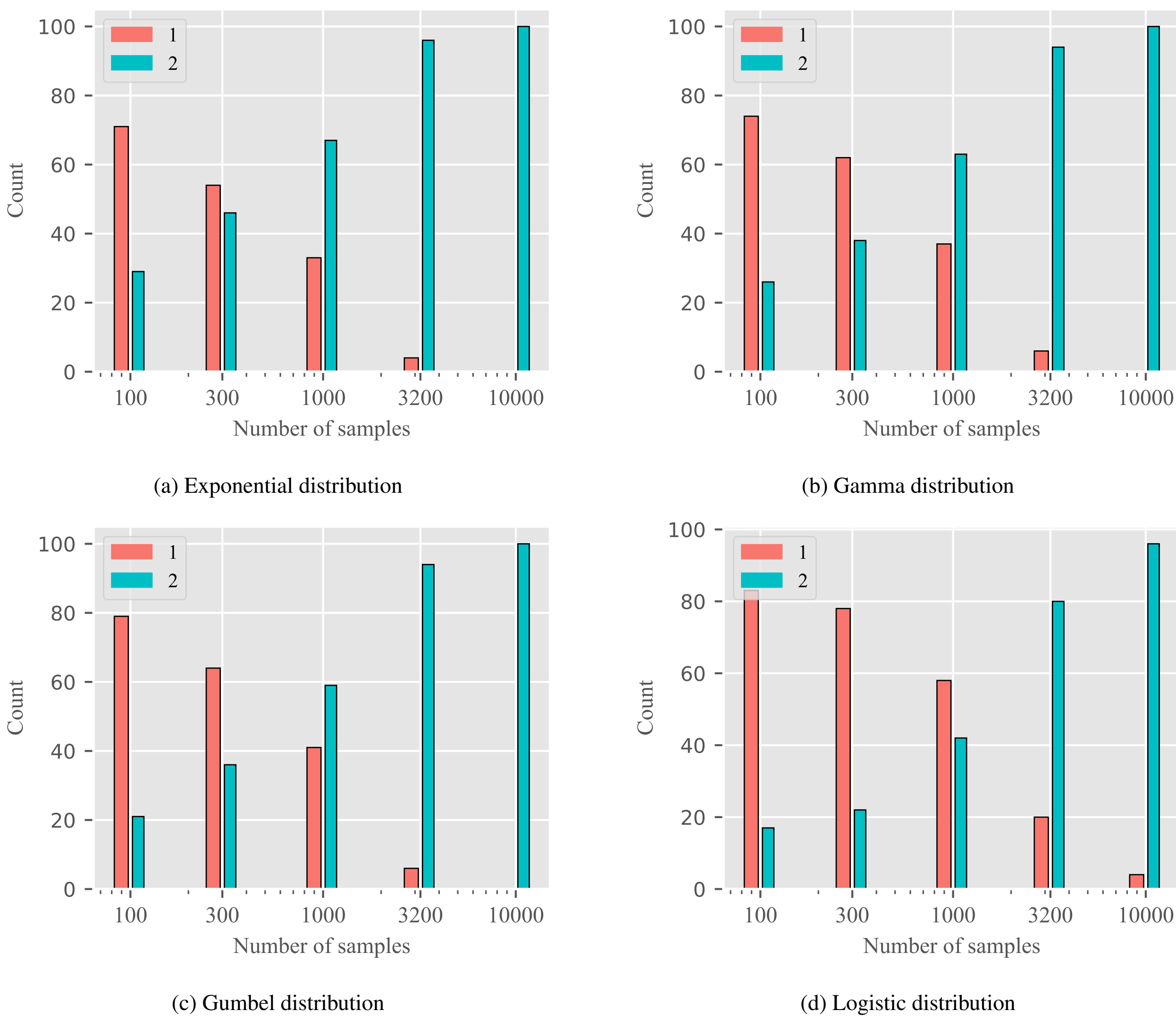


(a) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_t\}$



(b) $\mathcal{F}(\mathcal{M}^{(1)}, \mathcal{M}^{(2)}) = \{\epsilon_u\}$

Figure 11: Relative estimation bias given data from two domains, when only ϵ_t (11a), and only ϵ_t (11b) varies across domains. All noise variables are sampled from a logistic distribution, except the alternating noise variable in the second domain which is sampled from a Gamma distribution.



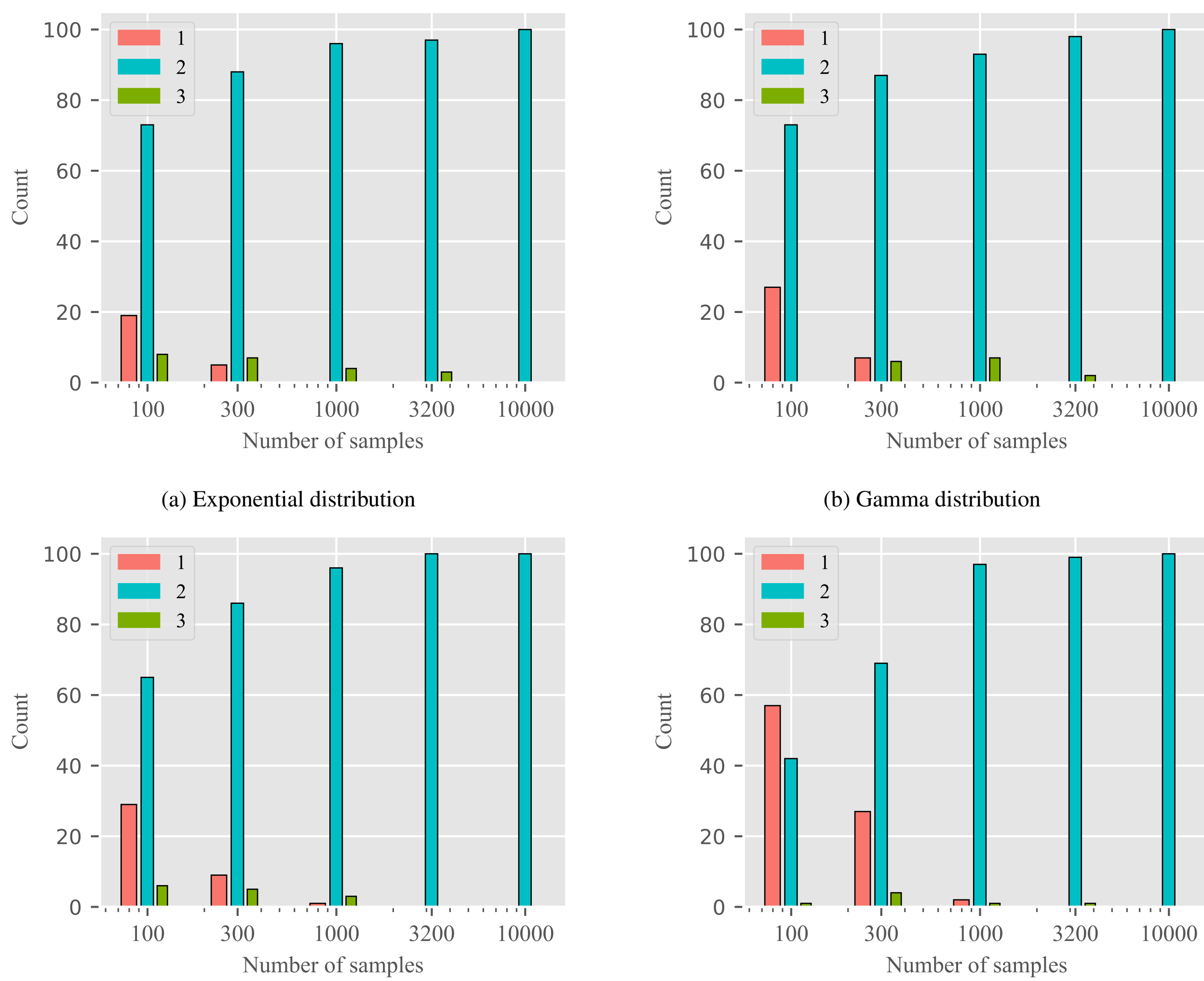
(a) Exponential distribution

(b) Gamma distribution

(c) Gumbel distribution

(d) Logistic distribution

Figure 12: Histogram of k in Algorithm 1.



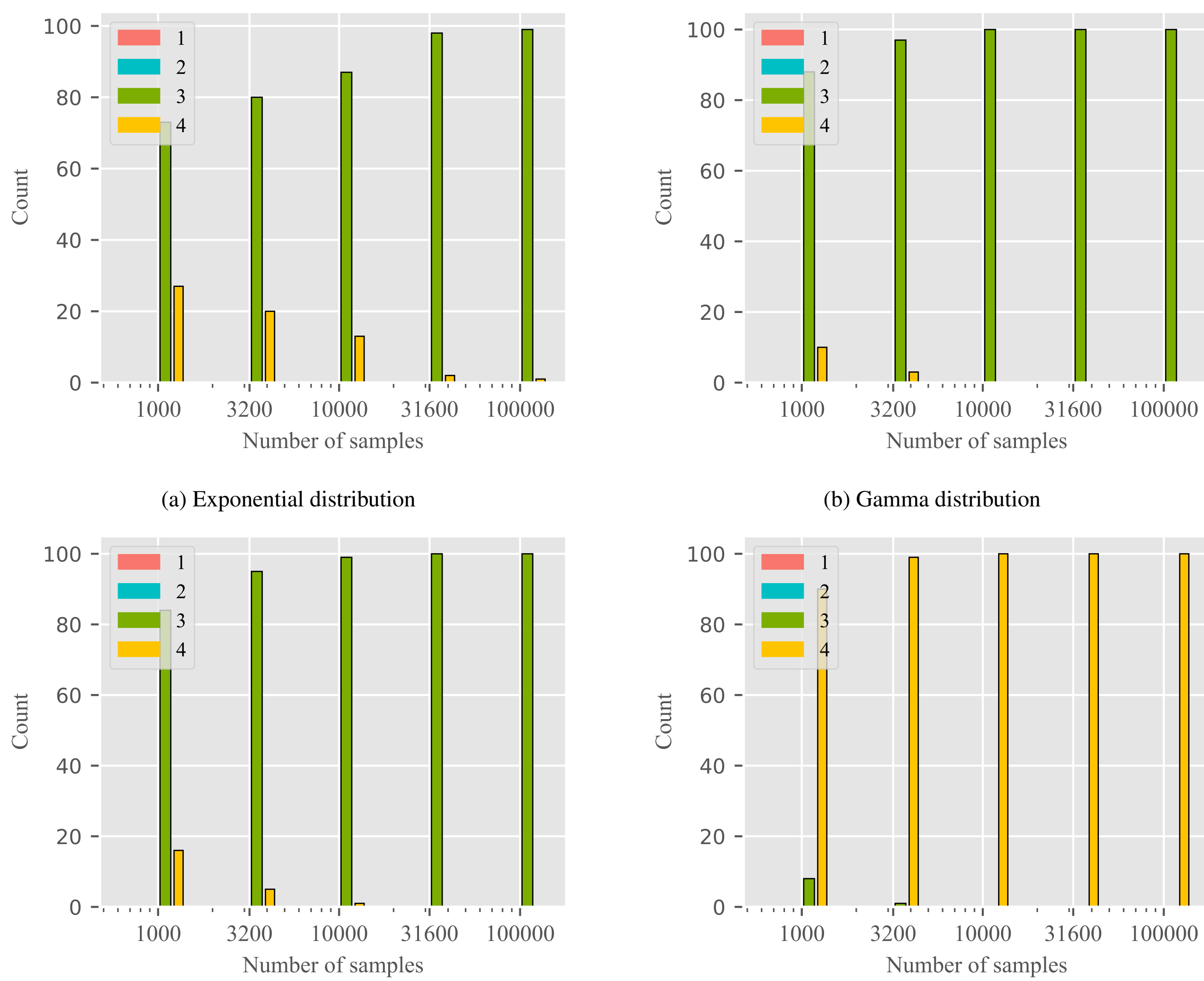
(a) Exponential distribution

(b) Gamma distribution

(c) Gumbel distribution

(d) Logistic distribution

Figure 13: Histogram of k in Algorithm 2.



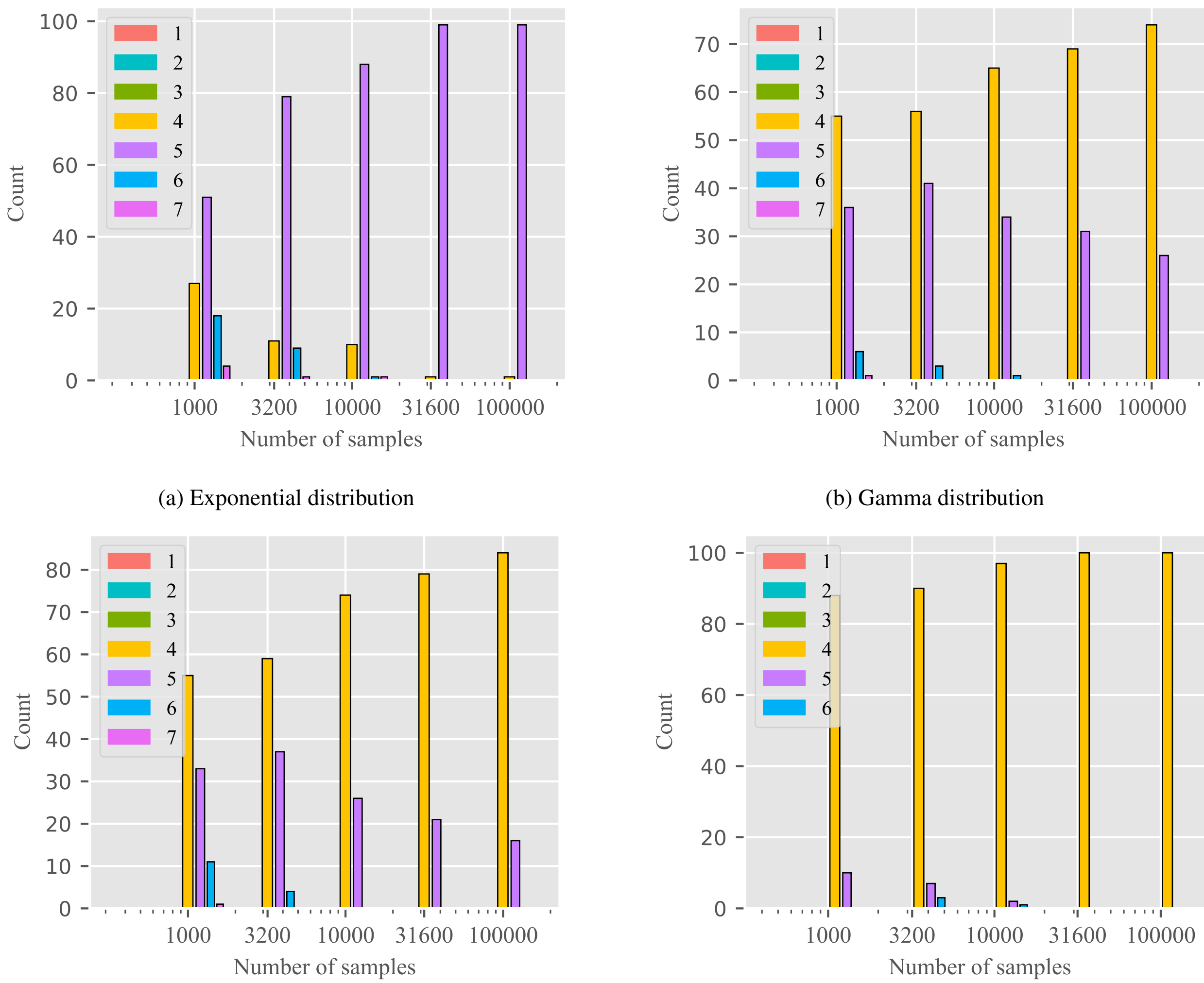
(a) Exponential distribution

(b) Gamma distribution

(c) Gumbel distribution

(d) Logistic distribution

Figure 14: Histogram of n_1 in Algorithm 3.



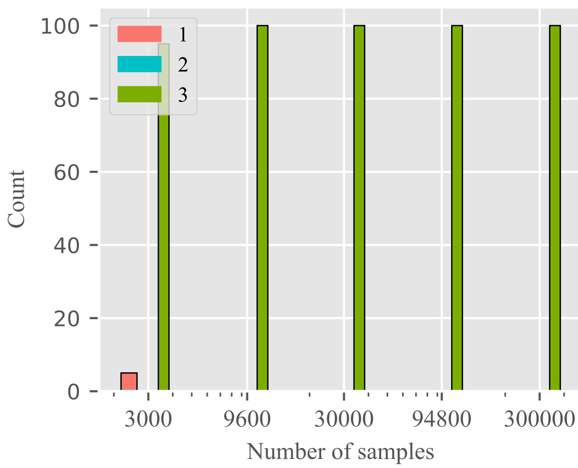
(a) Exponential distribution

(b) Gamma distribution

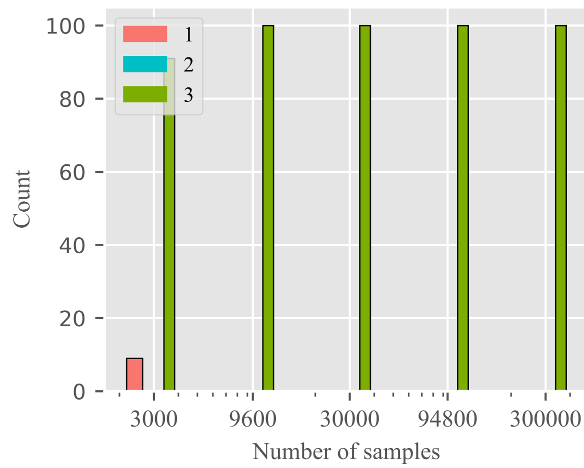
(c) Gumbel distribution

(d) Logistic distribution

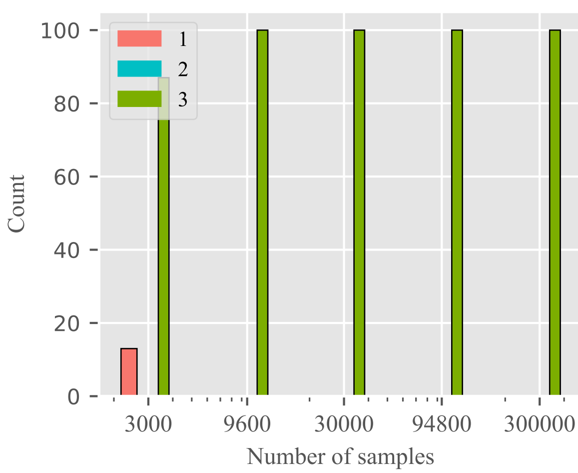
Figure 15: Histogram of n_2 in Algorithm 3.



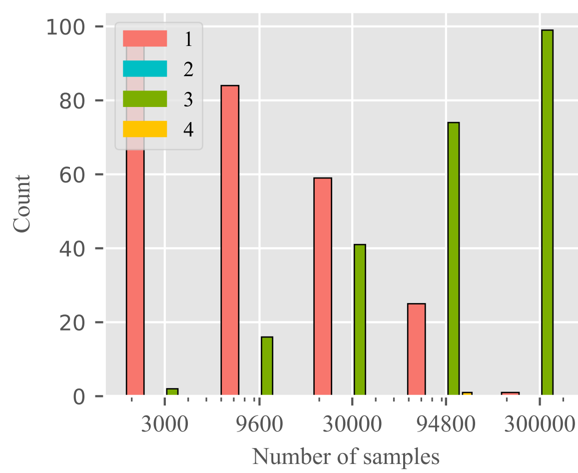
(a) Exponential distribution



(b) Gamma distribution



(c) Gumbel distribution



(d) Logistic distribution

Figure 16: Histogram of n in Algorithm 4.