Research Article

Raluca Cobzaru*, Roy Welsch, Stan Finkelstein, Kenney Ng, and Zach Shahn

Bias formulas for violations of proximal identification assumptions in a linear structural equation model

https://doi.org/10.1515/jci-2023-0039
received June 03, 2023; accepted March 20, 2024

Abstract: Causal inference from observational data often rests on the unverifiable assumption of no unmeasured confounding. Recently, Tchetgen Tchetgen and colleagues have introduced proximal inference to leverage negative control outcomes and exposures as proxies to adjust for bias from unmeasured confounding. However, some of the key assumptions that proximal inference relies on are themselves empirically untestable. In addition, the impact of violations of proximal inference assumptions on the bias of effect estimates is not well understood. In this article, we derive bias formulas for proximal inference estimators under a linear structural equation model. These results are a first step toward sensitivity analysis and quantitative bias analysis of proximal inference estimators. While limited to a particular family of data generating processes, our results may offer some more general insight into the behavior of proximal inference estimators.

Keywords: proximal causal inference, sensitivity analysis, bias analysis, negative control

MSC 2020: 62D20

1 Introduction

Causal inference using observational data often rests on the assumption of no unmeasured confounding. This assumption is not empirically verifiable, but sensitivity analysis methods [1–4] are available to assess robustness to possible violations. Alternatively, investigators might use methods such as instrumental variable analysis or difference-in-differences, which depend on different assumptions. Sensitivity analyses for violations of the assumptions required by these alternative methods are also available [5,6].

There has been recent interest in the use of negative control methods to detect and resolve confounder bias. A negative control outcome (NCO) is a variable known not to be causally affected by the treatment of interest, while a negative control exposure (NCE) is a variable known not to causally affect the outcome

* **Corresponding author: Raluca Cobzaru,** Operations Research Center, Massachusetts Institute of Technology, Cambridge, 02142 MA, United States of America; MIT-IBM Watson AI Lab, Cambridge, 02142 MA, United States of America, e-mail: rcobzaru@mit.edu
Roy Welsch: Operations Research Center, Massachusetts Institute of Technology, Cambridge, 02142 MA, United States of America; MIT-IBM Watson AI Lab, Cambridge, 02142 MA, United States of America
Stan Finkelstein: Operations Research Center, Massachusetts Institute of Technology, Cambridge, 02142 MA, United States of America; MIT-IBM Watson AI Lab, Cambridge, 02142 MA, United States of America; Institute for Data Systems and Society, Massachusetts Institute of Technology, Cambridge, 02142 MA, United States of America; Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, 02215 MA, United States of America
Kenney Ng: MIT-IBM Watson AI Lab, Cambridge, 02142 MA, United States of America; Center for Computational Health, IBM Research Cambridge, Cambridge, 02142 MA, United States of America
Zach Shahn: MIT-IBM Watson AI Lab, Cambridge, 02142 MA, United States of America; Department of Epidemiology and Biostatistics, City University of New York, New York, 10027 NY, United States of America

of interest [7]. Tchetgen Tchetgen et al. have developed a proximal inference framework [8–10], which uses NCE-NCO pairs sharing the same unmeasured confounders as the treatment–outcome relationship of interest as proxies to adjust for unmeasured confounding. Key assumptions of proximal inference include that the unmeasured confounders are associated with both the NCEs and NCOs ("U-relevance") and, roughly, that the NCEs and NCOs are sufficiently "rich" relative to the unmeasured confounders to serve as adequate proxies ("completeness"). Both U-relevance and completeness are themselves empirically untestable [8], and bias resulting from violations is not fully understood.

In this article, we characterize bias from violations of proximal inference assumptions in a linear structural equation model (LSEM). Our results build understanding of the sensitivity of proximal inference to assumption violations and enable "assumption-heavy" sensitivity analysis and quantitative bias analysis [11] tools for proximal inference. We hope that our results will also serve as a first step toward assumption-light sensitivity analysis.

As a motivating example, we consider the observational SUPPORT study estimating the effect of right heart catheterization (RHC) on 30 day survival in intensive care unit (ICU) patients. Initial analyses of these data [8] depended on the no unobserved confounding assumption and adjusted for a set of 71 baseline covariates. Despite extensive covariate adjustment, concern about unobserved confounding remains. Cui et al. [9] applied proximal inference to the SUPPORT data, using physiological measurements taken early in patients' ICU stay as both NCEs and NCOs. These variables were noisy early measurements of indicators of evolving underlying health conditions and did not themselves influence treatment decisions or health outcomes, making them valid NCEs. They also preceded treatment, making them valid NCOs. However, violations of U-relevance and completeness are still possible. For example, suppose that physician training is an unobserved confounder, with physicians who favor the heart catheterization procedure also tending to favor other posttreatment interventions which impact survival. As physician training would be independent of patient characteristics such as the NCEs and NCOs, U-relevance would be violated. Completeness could also be violated if there were many unobserved confounders, e.g., many dimensions of underlying health status that influence physicians' treatment decisions via aspects of the patient's physical appearance or behavior not captured by the covariates. The bias formulas we derive enable sensitivity analysis under a range of assumptions about the magnitude of completeness and U-relevance violations and under the strong simplifying assumption that the data generating process was an LSEM.

The organization of the article is as follows. In Section 2, we review proximal inference and motivate the need for bias analysis given the assumptions of this framework. In Sections 3 and 4, we derive and numerically explore bias formulas in a setting with two-dimensional unobserved confounder U and a setting of general-dimensional unobserved confounder U with no treatment–confounder interaction, respectively. In Section 5, we present an illustrative sensitivity analysis (based on the bias formula from Section 4) of a proximal inference analysis estimating the effect of RHC on survival. In Section 6, we conclude this study.

2 Proximal identification of the average treatment effect

2.1 Review of definitions and assumptions

We use the potential outcome framework [12] to define causal effects. Let A denote the binary treatment of interest, Y the observed posttreatment outcome, and $Y(a)$, $a = 0, 1$ the potential (counterfactual) outcome that would have been observed had treatment A been set to a. We implicitly make the no-interference assumption that the potential outcome of each individual does not depend on the treatments received by other individuals [13]. We aim to estimate the average causal effect (ACE) of A on Y, defined as $\psi = \mathbb{E}[Y(1) - Y(0)]$.

Let L denote the set of measured covariates. We make the standard consistency and positivity assumptions, defined below.

Assumption 1. (Consistency) $Y = Y(A)$ almost surely.

In other words, the observed value of Y under treatment A coincides with the counterfactual outcome that would have been observed under the same treatment value. Thus, we only observe the counterfactual outcome corresponding to the treatment value that was actually administered in our data.

Assumption 2. (Positivity) $0 < \mathbb{P}(A = a|L) < 1$ almost surely, for $a = 0, 1$.

Assumption 2 states that both exposure levels are observed at all levels of the observed covariates L.

Many analyses further make the assumption that there is no unobserved confounding, i.e., that observed covariates block all "backdoor" causal paths between treatment and outcome.

Assumption 3. (Exchangeability) $Y(a) \perp\!\!\!\perp A \mid L$, for $a = 0, 1$.

Under Assumptions 1–3, counterfactual mean $\mathbb{E}[Y(a)]$ is identified by the g-formula (introduced by Robins [14]):

$$\mathbb{E}[Y(a)] = \sum_l \mathbb{E}[Y|A = a, L = l]\mathbb{P}(L = l), \tag{1}$$

where L is assumed to be discrete. When L is continuous, the sum $\sum_l$ can be interpreted as integral $\int_l \mathbb{E}[Y|A = a, L = l]\mathrm{d}\mathbb{P}(L = l)$.

Exchangeability is a strong assumption that is empirically untestable, and much effort in causal inference research has been devoted to relaxing this assumption. Miao et al. [15] propose an alternative to Assumption 3 that allows identification of the counterfactual mean $\mathbb{E}[Y(a)]$ despite unobserved confounding. We review the alternative conditions developed by Miao et al. [15], leading to the *proximal g-formula*, a counterpart to (1) allowing for some unobserved confounding.

As shown by Cui et al. [9], we consider a (potentially multidimensional) variable L that can be partitioned into three types of variables (X, Z, W), such that

(1) X includes observed variables that may be common causes of A and Y (observed confounders),
(2) Z includes treatment-inducing confounding proxies, i.e., Z includes causes of A that share an unmeasured common cause U_Z with Y,
(3) W includes outcome-inducing confounding proxies, i.e., W includes causes of Y that share an unmeasured common cause U_W with A.

Figure 1 contains directed acyclic graphs (DAGs) representing each of the proxy types included in L. The covariates U_Z represent common causes of Y and treatment-inducing proxies Z, while covariates U_W represent common causes of A and outcome-inducing proxies W. In general, we will utilize U to denote unobserved common causes of A and Y.

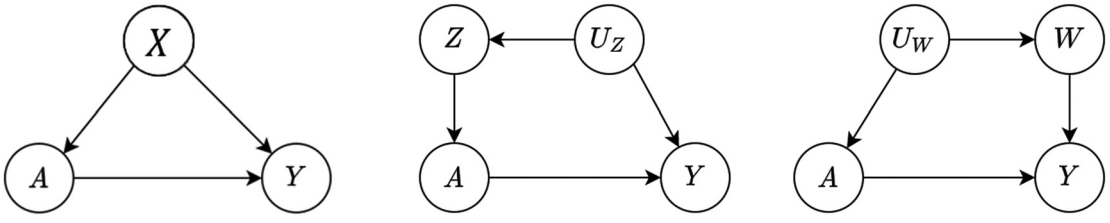

Figure 1: DAGs representing the three types of variables (X, Z, W) partitioning L.

In the study by Miao et al. [15], exchangeability is replaced with the following assumptions:

Assumption 4. (Treatment-inducing confounding proxy)

$$Y(a, z) = Y(a), \quad \text{for all } a, z \text{ almost surely.} \tag{2}$$

Assumption 5. (Outcome-inducing confounding proxy)

$$W(a, z) = W, \quad \text{for all } a, z \text{ almost surely.} \tag{3}$$

Assumption 6. (Latent unconfoundedness) If U denotes the set of unobserved confounders, then:

$$Z \perp\!\!\!\perp (Y(a), W) \mid (U, X), \tag{4}$$

$$W \perp\!\!\!\perp A \mid (U, X). \tag{5}$$

Assumption 4 states that Z does not have a direct effect on Y upon intervening on A, while Assumption 5 states that neither A nor Z have a causal effect on W. Past works [7] refer to variables Z satisfying (2) and (4) as NCE variables, and to variables W satisfying (3) and (5) as NCO variables. This terminology is based on negative control methods employing variables that share a confounding mechanism with the treatment–outcome relationship in view to detect bias in epidemiological research. In this article, we will use *treatment-inducing* (*outcome-inducing*) *confounding proxies* and *NCE* (*NCO*) variables interchangeably.

Moreover, to be valid proxies, variables (Z, W) must be U-relevant:

Assumption 7. (U-relevance)

$$Z \not\perp\!\!\!\perp U \mid (A, X), \tag{6}$$

$$W \not\perp\!\!\!\perp U \mid X. \tag{7}$$

The U-relevance assumption (also known as U-comparability [7]) requires the unmeasured confounders U of the A–Y relationship to be the same as the unmeasured confounders of the A–W and Z–Y secondary treatment–outcome associations. This is such that, by the negative control framework, any nonnull A–W or Z–Y association can be attributed to U confounding the A–Y relationship (while null associations imply no empirical evidence of unmeasured confounding).

Finally, in addition to Assumptions 1–7, Miao et al. [10] introduce the following *completeness* conditions for the identification of $\mathbb{E}[Y(a)]$:

Assumption 8. (Completeness) For any a, x and for any square-integrable function g:
(a) If $\mathbb{E}[g(U)|Z, A = a, X = x] = 0$ almost surely, then $g(U) = 0$ almost surely.
(b) If $\mathbb{E}[g(Z)|W, A = a, X = x] = 0$ almost surely, then $g(Z) = 0$ almost surely.

Assumption 8(a) can be interpreted as a requirement that the NCE Z has enough variability relative to the variability of U; similarly, Assumption 8(b) requires the variability of W to be large enough relative to the variability of Z. Under conditions 8(a) and (b), we can essentially account for U in our ACE estimate without either measuring or modeling the distribution of U. The role of completeness will be further explored in Section 2.2, where we outline the analytical framework by which the ACE is estimated using the proximal g-formula.

Completeness Assumption 8(a) has a simple interpretation in the case where confounders U and the negative control pair (Z, W) are all categorical. As mentioned by Cui et al. [9], if (U, Z, W) are categorical with respective number of categories (d_u, d_z, d_w), then completeness 8(a) requires that:

$$\min(d_z, d_w) \geq d_u. \tag{8}$$

In other words, proximal inference can account for unmeasured confounding if the number of categories of U is less than that of either Z or W. This leads to the practical recommendation to measure a rich set of baseline

characteristics (which can be used as negative controls), such that the proximal identification approach has a higher chance of mitigating unmeasured confounder bias [9]. There is not such a straightforward method for expressing the completeness condition in the case of continuous U and negative controls (Z, W), though some theory about completeness has been developed in some commonly used models (e.g., exponential families [16]). In Section 3, we investigate the behavior of proximal inference in LSEM setups in which the completeness Assumption 8(a) is violated.

2.2 Estimating the proximal g-formula via moment restriction

Miao et al. [15] introduce the notion of an *outcome confounding bridge function*, which transforms the NCO W to match the confounding effect of U on Y. More precisely, an outcome confounding bridge function $h(W, A, X)$ is a function satisfying:

$$\mathbb{E}[Y|U, A = a, X = x] = \mathbb{E}[h(W, A, X)|U, A = a, X = x], \tag{9}$$

for all values of a, x. In other words, if function $h(W, A, X)$ exists, then the confounding effect of U on the transformed variable $h(W, a, X)$ equals the confounding effect of U on Y at exposure level $A = a$. Given Assumptions 1, 5, 6, and 7, [15] infer that:

$$\mathbb{E}[Y(a)] = \mathbb{E}[h(W, a, X)] \quad \text{for all } a = 0, 1, \tag{10}$$

which means $\mathbb{E}[Y(a)]$ can be estimated following the identification of an outcome bridge function $h(W, A, X)$, if such a function is assumed to exist.

Cui et al. [9] and Miao et al. [10] established the following proximal identification result for the outcome confounding bridge function that leverages the distribution of a NCE Z.

Theorem 1. *Suppose there exists an outcome confounding bridge function $h(w, a, x)$ solving the Fredholm integral equation of the first kind,*

$$\mathbb{E}[Y|Z, A, X] = \int h(w, A, X)\mathrm{d}F(w|Z, A, X), \tag{11}$$

almost surely. Then, under Assumptions 1, 2, 4–6, *and* 8(a),

$$\mathbb{E}[Y|U, A, X] = \int h(w, A, X)\mathrm{d}F(w|U, X), \tag{12}$$

almost surely.

Under Assumption 6, we have $\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y|U, A = a, X]]$ for all a. The counterfactual mean $\mathbb{E}[Y(a)]$ can then be computed as follows:

Corollary 1.1. (Proximal g-formula) *If* (12) *holds almost surely, then the counterfactual mean $\mathbb{E}[Y(a)]$, $a = 0, 1$ is nonparametrically identified by*

$$\mathbb{E}[Y(a)] = \int\int_X h(w, a, x)\mathrm{d}F(w|x)\mathrm{d}F(x), \tag{13}$$

and the ACE is identified by

$$\psi = \int\int_X \{h(w, 1, x) - h(w, 0, x)\}\mathrm{d}F(w|x)\mathrm{d}F(x). \tag{14}$$

Assuming the outcome confounding bridge function $h(W, A, X)$ exists and is identifiable as a solution to (12), [8,15] provide a practical approach for estimating the proximal g-formula using the generalized method of moments (GMM). Suppose one has access to n i.i.d. samples $D_i = (A_i, Y_i, L_i)$, $L_i = (X_i, Z_i, W_i)$ (where Z, W are assumed to be correctly classified as treatment- and outcome-inducing confounding proxies, respectively).

Moreover, suppose one has specified a parametric model for the confounding bridge, $h(W, A, X) = h(W, A, X;\ b)$ (e.g., $h(W, A, X;\ b)$ is linear in W, A, X with unknown parameter b). The true model for $h(W, A, X)$ is unknown, but one can specify a fairly flexible model.

We define the target parameter $\theta = (b, \psi)$ to encode the parameters b of $h(W, A, X;\ b)$ and the ACE ψ, along with moment restrictions

$$h(D_i;\ \theta) = \begin{Bmatrix} \{Y_i - h(W_i, A_i, X_i;\ b)\} \times Q(Z_i, A_i, X_i) \\ \psi - \{h(W_i, 1, X_i;\ b) - h(W_i, 0, X_i;\ b)\} \end{Bmatrix}, \tag{15}$$

for some vector function Q (as in the study by Miao et al. [15]). For instance, for linear bridge function,

$$h(W_i, A_i, X_i;\ b) = (1 \quad A_i \quad W_i \quad X_i \quad A_iX_i \quad A_iW_i)^T b,$$

we may choose a function

$$Q(Z_i, A_i, X_i) = (1 \quad A_i \quad Z_i \quad X_i \quad A_iX_i \quad A_iZ_i)^T,$$

such that the dimension of Q is at least equal to that of h.

Then, for $m_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} h(D_i;\ \theta)$, the GMM estimator solves

$$\hat{\theta} = \arg\min_{\theta} m_n^T(\theta) m_n(\theta). \tag{16}$$

As established by Miao et al. [15], the estimates $(\hat{b}, \hat{\psi})$ obtained from (16) are consistent.

2.3 The need for bias analysis

We have so far collected a series of untestable Assumptions 4–8 that replace exchangeability and account for the effect of unmeasured confounders U without directly modeling or estimating U. The impact on the direction and/or magnitude of bias resulting from violations of these assumptions has not been explored. We trust the analyst to identify "true" NCEs and NCOs in this work (Assumptions 4 and 5), on the basis of subject–matter knowledge. That is, we assume that arrow $A \to Z$ is correctly identified, and that there are no additional arrows $Z \to Y$ or $A \to W$. Latent unconfoundedness (Assumption 6) presumably holds for some sufficiently rich U, but the richer (or higher-dimensional) the U required to satisfy Assumption 6, the less plausible it is that U-relevance (Assumption 7) or completeness (Assumption 8) hold. If many components of U are common causes of the NCEs and NCOs, then Assumption 8 is difficult to satisfy. In addition, if many components of U are required to block all backdoor paths between A and Y, then they are less likely to all be associated with both Z and W, violating Assumption 7.

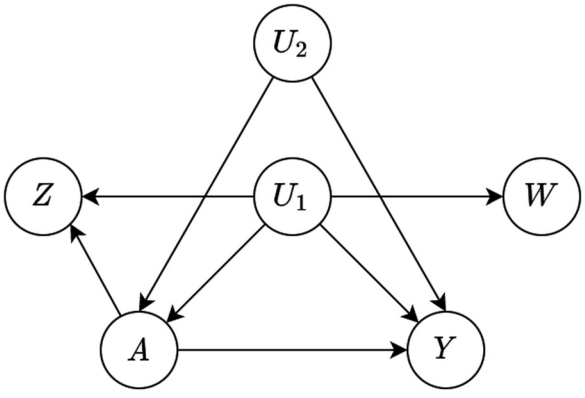

Figure 2: DAG encoding causal relationships among variables in (19) in which U-relevance Assumption 7 is violated.

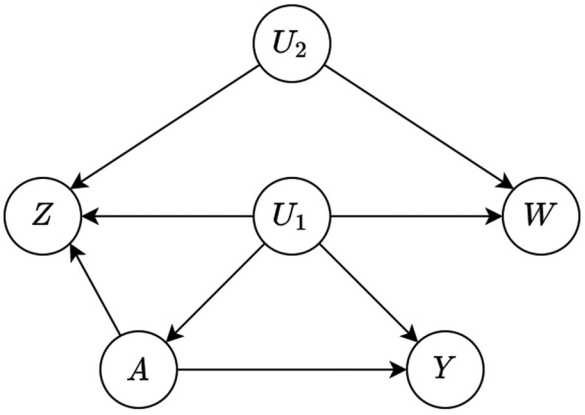

Figure 3: DAG encoding the causal relationships among variables in (21) in which completeness 7(a) is violated.

3 Bias formulas for two-dimensional U

In this section, we characterize the proximal inference estimator bias in an LSEM under scenarios in which each of Z and W are one-dimensional, but U (comprising common causes of any of A, Y, Z, and W) has two independent components. We first consider the case where one component of U is a common cause of A and Y but is not associated with either Z or W (which violates U-relevance Assumption 7 and is illustrated in Figure 2). Then, we consider the case where one component of U is an "extra" common cause of Z and W not associated with A or Y (which violates completeness Assumption 8 and is illustrated in Figure 3). We would argue that it is difficult to guard against violations of Assumptions 7 and 8 arising in this way using subject-matter knowledge, making sensitivity analysis for violations of these types particularly necessary.

In addition, for the settings of Figures 2 and 3, we compare the bias of the proximal estimator due to violations of Assumptions 7 and 8 to the bias of alternative estimators of the ACE which the analyst might implement under an incorrect unconfoundedness assumption. We consider

(1) an unadjusted estimator (referred to as "unadj"), which assumes no unobserved confounding and estimates $E[Y(a)]$ as $\hat{E}[Y|A = a]$ via sample means, and
(2) an outcome regression estimator (referred to as "OR"), which adjusts for (Z, W) via the g-formula (1) taking $L = \{Z, W\}$ and specifying outcome regression model $E[Y|A, L] = \beta^T(A, Z, W, AZ, AW)$.

3.1 Bias settings for two-dimensional U

As outlined at the beginning of this section, we derive formulas for the proximal inference estimator bias under scenarios depicted in Figures 2 and 3, under an LSEM based on [10]. The notation $U = (U_1, U_2)$ indicates that U is two-dimensional with components U_1 and U_2. Specifically, we consider i.i.d. data:

$$
\begin{aligned}
\begin{pmatrix} U \\ X \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \nu & \rho_1 \\ \nu & 1 & \rho_2 \\ \rho_1 & \rho_2 & 1 \end{bmatrix}\right), \quad \rho_1, \rho_2 \in (-1, 1), \\
\operatorname{logit}(\mathbb{P}(A = 1|X, U)) &= \alpha_0 + \alpha_x X + \boldsymbol{\alpha}_u^T U, \\
Z &= \theta_0 + \theta_a A + \theta_x X + \boldsymbol{\theta}_u^T U + \varepsilon_1, \\
W &= \mu_0 + \mu_x X + \boldsymbol{\mu}_u^T U + \varepsilon_2, \\
Y(a) &= \gamma_0 + \gamma_a a + \gamma_x X + \boldsymbol{\gamma}_u^T U + \gamma_{au_1} a U_1 + \varepsilon_3, \\
\varepsilon_1, \varepsilon_2, \varepsilon_3 &\sim \mathcal{N}(0, 1).
\end{aligned}
\tag{17}
$$

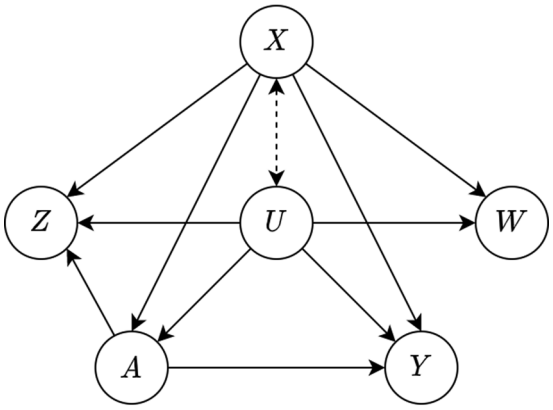

Figure 4: DAG encoding the causal relationships among variables in (17).

Figure 4 depicts the causal DAG corresponding to this LSEM. The dashed bidirectional arrow between X and U indicates an unrestricted association arising from an unspecified causal relationship (e.g., a shared common cause) between these variables. The parameter $\nu = \text{Corr}(U_1, U_2)$ encodes the correlation between two components of U, following standardization of the components. In addition, parameter $\boldsymbol{\alpha}_u = (\alpha_{u_1} \;\; \alpha_{u_2})^T$ encodes the magnitude of confounding, while $\boldsymbol{\theta}_u = (\theta_{u_1} \;\; \theta_{u_2})^T$ and $\boldsymbol{\mu}_u = (\mu_{u_1} \;\; \mu_{u_2})^T$ encode the association between confounder U and the NCE/NCO, respectively. We will explore the sensitivity of proximal inference bias to particular values of $(\boldsymbol{\alpha}_u, \boldsymbol{\theta}_u, \boldsymbol{\mu}_u)$.

The NCE Z is a posttreatment variable in this LSEM. We note that DAGs other than Figure 4 might also be compatible with proximal inference assumptions [7] (e.g., having an arrow $Z \to A$).

If U were one-dimensional and satisfied all the proximal inference assumptions, then the bridge function solving the outcome bridge function equation would take a linear form:

$$h(W, A, X;\, b) = b_0 + b_a A + b_w W + b_x X + b_{ax} AX + b_{aw} AW. \tag{18}$$

The proof of this claim is in Appendix A.1. Therefore, like an analyst unaware of the additional assumption-violating component of U, we consider bias of proximal estimators that specify a linear bridge function.

By Theorem 1, violating Assumption 8(a) leads to a potentially biased ACE estimate as the outcome confounding bridge function $\hat{h}(W, A, X)$ resulting from the GMM procedure no longer satisfies the bridge equation (12). The following theorem shows that the completeness Assumption 8(a) is indeed violated in DGP (17) when both components of U are associated with negative controls Z and W.

Theorem 2. *If $\boldsymbol{\theta}_u$ is nonzero* (*i.e., Z is U-relevant for at least one component of U*), *then the LSEM* (17) *with Gaussian* (X, U) *violates completeness Assumption* 8(a).

The proof of Theorem 2 is included in Appendix B.

3.2 Partial U-relevance for two-dimensional unobserved confounder U (as in Figure 2)

In this subsection, we consider the case where U-relevance Assumption 7 is violated because one component of the two-dimensional confounder U is not associated with the negative controls. We exclude X for simplicity, which results in the DAG from Figure 2. We consider i.i.d. data generated by:

$$
\begin{aligned}
U &\sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \nu\\ \nu & 1\end{pmatrix}\right), \quad \nu \in (-1,1),\\
\operatorname{logit}(\mathbb{P}(A=1|U)) &= \alpha_0 + \boldsymbol{\alpha}_u^T U,\\
Z &= \theta_0 + \theta_a A + \theta_{u_1} U_1 + \varepsilon_1,\\
W &= \mu_0 + \mu_{u_1} U_1 + \varepsilon_2,\\
Y(a) &= \gamma_0 + \gamma_a a + \boldsymbol{\gamma}_u^T U + \gamma_{au_1} a U_1 + \varepsilon_3,\\
\varepsilon_1, \varepsilon_2, \varepsilon_3 &\sim \mathcal{N}(0,1),
\end{aligned}
\tag{19}
$$

where $\boldsymbol{\alpha}_u$, $\boldsymbol{\gamma}_u$ have all nonzero entries.

From Theorem 2, we know that setup (19) violates Assumption 8(a). In addition, we do not have a derivation of the true outcome confounding bridge function, so a linear model might be misspecified. The following theorem provides a formula for this bias under a linear bridge function specification, which is still the specification an analyst unaware of U_2 would choose and therefore relevant to sensitivity analysis. Similar to the previous case, we further assume $\operatorname{Corr}(U_1, U_2) = 0$ to improve the interpretability of the resulting bias formula.

Theorem 3. *If* $(Z, W) \perp\!\!\!\perp U_2 \mid (A, U_1)$ *and* $\operatorname{Corr}(U_1, U_2) = 0$, *then fitting a linear outcome bridge function* $h(W, A, X) = b_0 + b_a A + b_w W + b_{aw} AW$ *under LSEM* (17) *yields a proximal outcome estimator bias equal to*:

$$
\begin{aligned}
\delta_{\text{POR}} = \Bigg[&\frac{(1 - \mathbb{E}[A] - \mathbb{E}[AU_1^2])\mathbb{E}[AU_1]\mathbb{E}[AU_1U_2]}{(\mathbb{E}[A]\mathbb{E}[AU_1^2] - \mathbb{E}[AU_1]^2)((1 - \mathbb{E}[A])(1 - \mathbb{E}[AU_1^2]) - \mathbb{E}[AU_1]^2)}\\
&+ \frac{(\mathbb{E}[AU_1^2](1 - \mathbb{E}[AU_1^2]) - \mathbb{E}[AU_1]^2)\mathbb{E}[AU_2]}{(\mathbb{E}[A]\mathbb{E}[AU_1^2] - \mathbb{E}[AU_1]^2)((1 - \mathbb{E}[A])(1 - \mathbb{E}[AU_1^2]) - \mathbb{E}[AU_1]^2)} \Bigg] \gamma_{u_2}.
\end{aligned}
\tag{20}
$$

The proof for Theorem 3 (as well as a more general formula for $\nu \in (-1, 1)$) is in Appendix C.1. It clearly follows that the proximal outcome estimator bias is proportional to the strength of association γ_{u_2} between outcome Y and U_2, as well as to the function $\mathbb{E}[AU_2]$ encoding the strength of association between treatment A and U_2.

3.3 Completeness violation: Association between negative controls through $U = (U_1, U_2)$ (as in Figure 3)

As shown in Figure 3, for simplicity, we consider a scenario with no covariates X, i.e.,:

$$
\begin{aligned}
U &\sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \nu\\ \nu & 1\end{pmatrix}\right), \quad \nu \in (-1,1),\\
\operatorname{logit}(\mathbb{P}(A=1|U)) &= \alpha_0 + \alpha_{u_1} U_1,\\
Z &= \theta_0 + \theta_a A + \boldsymbol{\theta}_u^T U + \varepsilon_1,\\
W &= \mu_0 + \boldsymbol{\mu}_u^T U + \varepsilon_2,\\
Y(a) &= \gamma_0 + \gamma_a a + \gamma_{u_1} U_1 + \gamma_{au_1} a U_1 + 2\varepsilon_3,\\
\varepsilon_1, \varepsilon_2, \varepsilon_3 &\sim \mathcal{N}(0,1),
\end{aligned}
\tag{21}
$$

where $\boldsymbol{\theta}_u$, $\boldsymbol{\mu}_u$ have all nonzero entries.

The aforementioned setup satisfies all assumptions except 8(a) (which is violated according to Theorem 2). Thus, solving for the parameters b of a linear outcome bridge function (which is the functional form an investigator unaware of U_2 would select) will lead to a biased estimate of the ACE, even if the linear bridge function is correctly specified. The following theorem (proved in Appendix C.2) provides a formula for this bias under a linear outcome bridge function in the case when $\nu = 0$:

Theorem 4. *If* $(A, Y) \perp\!\!\!\perp U_2 \mid U_1$ *and* $\mathrm{Cor}(U_1, U_2) = 0$, *then fitting a linear outcome bridge function* $h(W, A, X) = b_0 + b_a A + b_w W + b_{aw} AW$ *under LSEM* (21) *yields a proximal outcome estimator bias equal to*

$$\delta_{\mathrm{POR}} = \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A](1-\mathbb{E}[A])}\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}\cdot\left[\frac{(1-\mathbb{E}[A])S_2}{\mu_{u_1}+S_2\cdot\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{au_1}+\left(\frac{\mathbb{E}[A]S_1}{\mu_{u_1}+S_1\cdot\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}+\frac{(1-\mathbb{E}[A])S_2}{\mu_{u_1}+S_2\cdot\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\right)\gamma_{u_1}\right], \tag{22}$$

where

$$S_1 = \frac{(1-\mathbb{E}[A])^2}{(1-\mathbb{E}[A])(1-\mathbb{E}[AU_1^2])-\mathbb{E}[AU_1]^2},$$
$$S_2 = \frac{\mathbb{E}[A]^2}{\mathbb{E}[A]\mathbb{E}[AU_1^2]-\mathbb{E}[AU_1]^2}.$$

For $\gamma_{au} = 0$, *the bias simplifies to*

$$\delta_{\mathrm{POR}} = \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A](1-\mathbb{E}[A])}\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}\cdot\left(\frac{\mathbb{E}[A]S_1}{\mu_{u_1}+S_1\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}+\frac{(1-\mathbb{E}[A])S_2}{\mu_{u_1}+S_2\cdot\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\right)\gamma_{u_1}. \tag{23}$$

The more general formulas for arbitrary $\nu \in (-1, 1)$ *are included in Appendix* C.2. *For ease of interpretation, we restrict our attention to the case when* $\nu = 0$ *in this section's discussion.*

One implication of Theorem 4 is that the proximal outcome regression bias δ_{POR} can obtain arbitrarily large when one of the denominators $\mu_{u_1} + S_1\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}$ or $\mu_{u_1} + S_2\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}$ approaches zero. This will be illustrated in Figure 5 (the solid "PI" curve) of Numerical Experiments section 3.4.2, in which there exists values of θ_{u_2} and μ_{u_2} for which the proximal outcome bias becomes infinite.

Under the simplifying assumption that the unobserved confounder is not an effect modifier, i.e., $\gamma_{au} = 0$, in Theorem 5, we characterize when the proximal estimator will reduce bias relative to an unadjusted estimator, even when the proximal inference assumptions are violated. It turns out that if the components of U induce associations between Z and W in the same direction, then the proximal estimator is guaranteed to have lower bias. This will be illustrated via the plots in Section 3.4.2, through numerical comparisons between the proximal outcome and unadjusted bias curves under different setups of U-component associations.

Theorem 5. *Assuming* $\gamma_{au_1} = 0$, *the proximal g-computation bias* δ_{POR} *and the unadjusted estimator bias* δ_{unadj} *can be compared as follows*:

(*i*) *If* $\theta_{u_1}\mu_{u_1}$ *and* $\theta_{u_2}\mu_{u_2}$ *have the same sign (both positive or both negative), then* $|\delta_{\mathrm{POR}}| < |\delta_{\mathrm{unadj}}|$.

(*ii*) *If* $\theta_{u_1}\mu_{u_1}$ *and* $\theta_{u_2}\mu_{u_2}$ *have different signs, then*

$$\begin{cases} |\delta_{\mathrm{POR}}| > |\delta_{\mathrm{unadj}}| & \text{if } \dfrac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} > -S_1(1-\mathbb{E}[A]) - S_2\mathbb{E}[A], \\ |\delta_{\mathrm{POR}}| < |\delta_{\mathrm{unadj}}| & \text{if } \dfrac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} < -S_1(1-\mathbb{E}[A]) - S_2\mathbb{E}[A]. \end{cases}$$

The proof for Theorem 5 is in Appendix C.5.

3.4 Numerical experiments

We provide numerical examples based on the bias formulas derived earlier in this section to illustrate how the bias of different estimators (proximal and nonproximal) varies with different values of $(\alpha_{u_2}, \theta_{u_2}, \mu_{u_2}, \gamma_{u_2})$, which encode how strongly the proximal identification assumptions are violated in the presence of U_2.

3.4.1 Partial U-relevance for two-dimensional unobserved confounder U

Figure 6 illustrates the change in absolute bias for the proximal and unadjusted estimators relative to the value of α_{u_2}, for the same and opposite directions of associations γ_{u_1} and γ_{u_2}, respectively. In both cases, the distributions of bias appear almost shifted by translation. For γ_{u_1} and γ_{u_2} of opposite directions of association, we observe a reversal in which estimator has less bias compared to the case of $\gamma_{u_1}\gamma_{u_2} > 0$.

3.4.2 Completeness violation: Association between negative controls through $U = (U_1, U_2)$

Figures 5 and 7 illustrate the change in absolute bias for each of the three estimators relative to the value of θ_{u_1}, where it is assumed that $\mu_{u_2} = \theta_{u_2}$ in all cases, for the same sign and opposite signs of $\theta_{u_1}\mu_{u_1}$ and $\theta_{u_2}\mu_{u_2}$, respectively. We observe that the absolute unadjusted bias is always greater than the proximal estimator bias when $\theta_{u_1}\mu_{u_1}$, $\theta_{u_2}\mu_{u_2}$ have the same sign, as predicted by Theorem 5. Conversely, for different signs of $\theta_{u_1}\mu_{u_1}$, $\theta_{u_2}\mu_{u_2}$, the proximal estimation bias exceeds that of the unadjusted estimator beyond a certain threshold in the value of $|\theta_{u_2}|$ (and can even be infinite). Both setups are consistent with Theorem 5. Any ordering of the biases of the three estimators is possible depending on the parameter values, making the choice of estimator not straightforward.

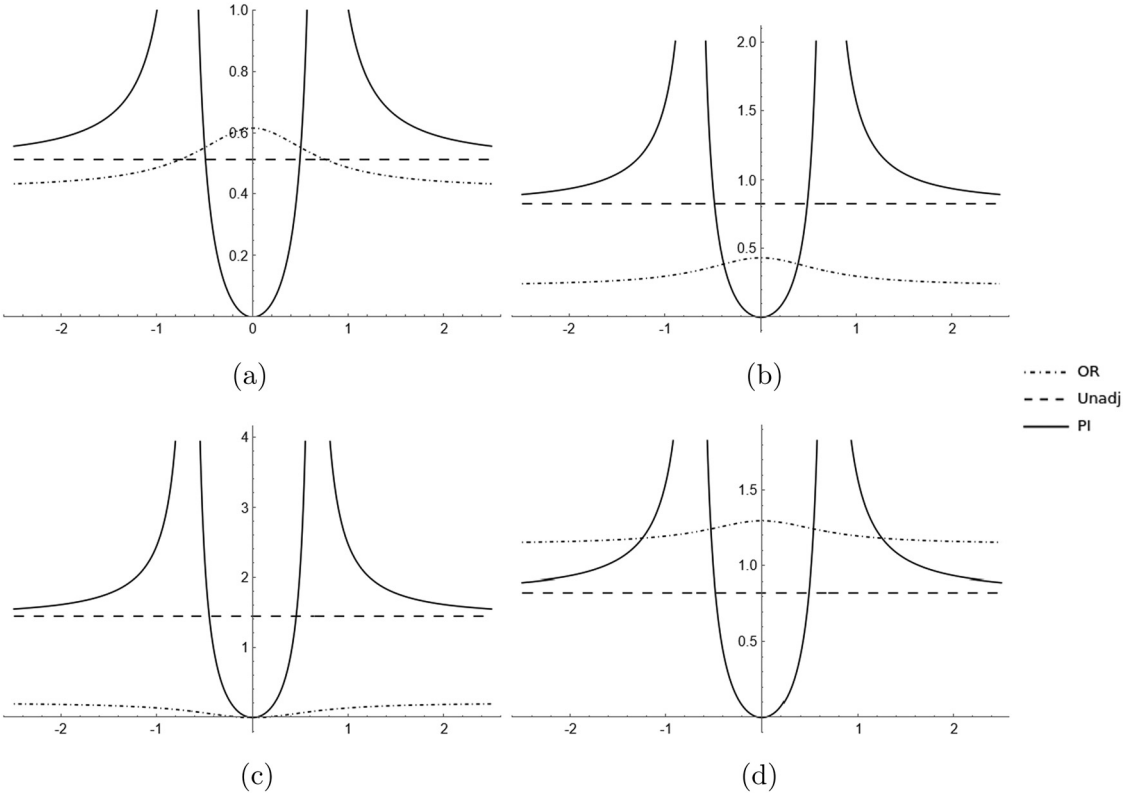

Figure 5: Plots of the ACE estimate bias under DGP (21) and under violations of completeness Assumption 8(a), where $\theta_{u_1}\mu_{u_1}$ and $\theta_{u_2}\mu_{u_2}$ have the same sign as in Theorem 5(b). The completeness violation is imposed by setting both components of U to be associated with both negative controls and only including one NCE and one NCO, as per Theorem 2. Along the x-axis, we vary the strength of association between U_2 and Z and W, by varying θ_{u_2} and μ_{u_2} (set to be equal to each other, $\theta_{u_2} = \mu_{u_2}$, so they always have positive product). The different panels correspond to different values of α_{u_1} governing the strength of confounding by U_1. The figure shows that, as predicted by Theorem 5(b), the bias of the proximal outcome estimator ("PI", **solid** line) can be greater or smaller than the bias of the unadjusted estimator ("Unadj", dashed line —) and can be arbitrarily large under certain parameter settings. Moreover, the relationship of PI with the outcome adjusted estimator ("OR", dotted line ⋯) varies with the strength of confounding α_{u_1} across panels. All other parameters are fixed, with values: $\alpha_0 = \gamma_0 = \theta_0 = \mu_0 = 0$, $\theta_a = \theta_{u_1} = 1$, $\gamma_{u_1} = 1$, $\gamma_{u_a1} = 1.5$, $\mu_{u_1} = 0.5$, $\gamma_a = 0.5$, similar to the simulation in [15]. (a) $\alpha_{u_1} = 0.3$, (b) $\alpha_{u_1} = 0.5$, (c) $\alpha_{u_1} = 1$, (d) $\alpha_{u_1} = 0.5$.

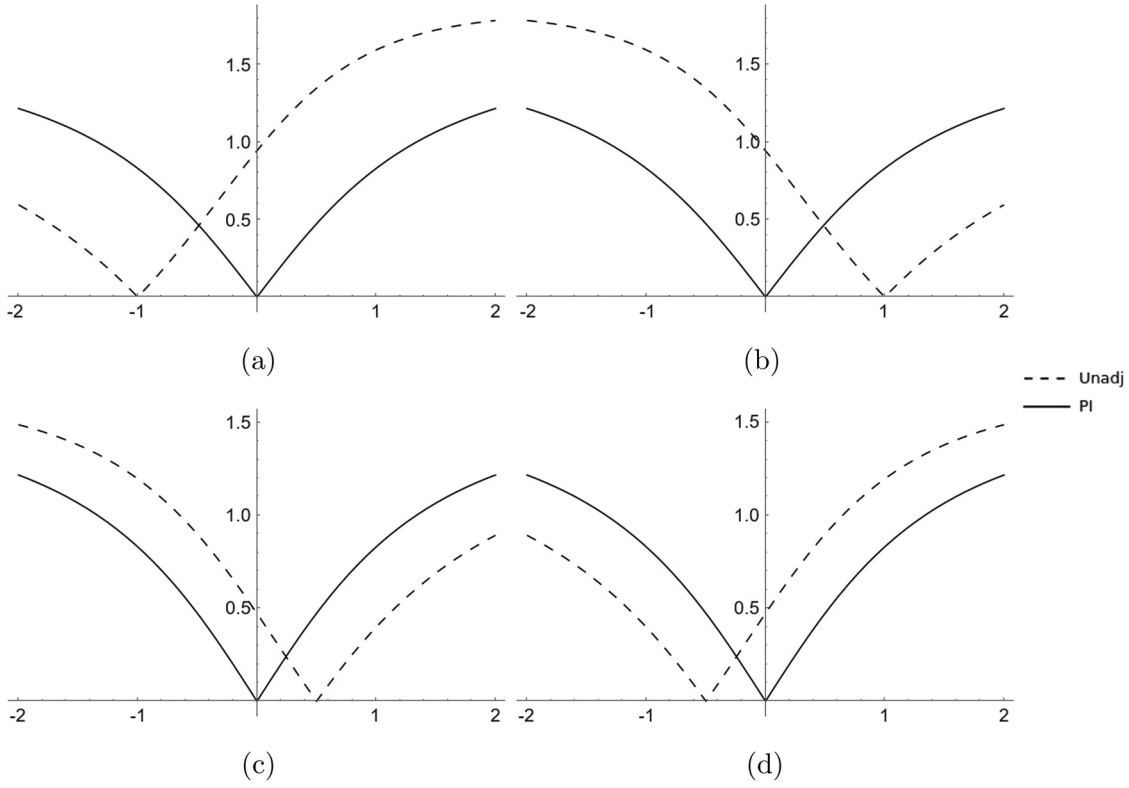

Figure 6: Plots of the ACE estimate bias under DGP (19) and under violations of U-relevance Assumption 7. We compare the proximal outcome estimator bias ("PI", **solid** line) to the unadjusted estimator bias ("Unadj", dashed line —) that assumes no unobserved confounding. The U-relevance violation is imposed by setting $\theta_{u_2} = \mu_{u_2} = 0$, making U_2 unassociated with the negative controls. Strength of confounding by U_2 is varied along the x-axis through the parameter α_{u_2} governing its association with treatment. The relative magnitude of bias of the proximal and unadjusted estimators flips depending on whether U_1 induces a positive (panels (a) and (d)) or negative (panels (b) and (c)) association between treatment and outcome. All other parameters are fixed, with values: $\alpha_0 = \gamma_0 = \theta_0 = \mu_0 = 0$, $\theta_a = \theta_{u_1} = 1$, $\mu_{u_1} = 1$, $\gamma_{au_1} = 1$, $\gamma_{u_2} = 1$, $\gamma_a = 0.5$. (a) $\alpha_{u_1} = 0.5$, $\gamma_{u_1} = 1.5$, (b) $\alpha_{u_1} = 0.5$, $\gamma_{u_1} = 1.5$, (c) $\alpha_{u_1} = 0.5$, $\gamma_{u_1} = 1.5$, and (d) $\alpha_{u_1} = 0.5$. $\gamma_{u_2} = 1.5$.

4 Bias formulas in arbitrary dimension with no confounder–treatment interaction

To tractably obtain bias formulas in the general case of multidimensional Z, W, U, X with $(\dim(Z), \dim(W), \dim(U), \dim(X)) = (m, n, p, q)$, we again make the simplifying assumption that $\gamma_{au} = 0$ – that is, the unobserved confounder is not an effect modifier. Moreover, we assume that the analyst is aware of the lack of interaction between A and U in the true outcome model, so we consider a simplified bridge function model $h(W, A, X) = b_0 + b_a A + b_w^T W + b_x^T X$. We further assume that the unobserved and observed confounders (U, X) jointly follow a multivariate normal distribution with mean $\mathbf{0}_{p+q}$, $\mathrm{Var}(U) = \Sigma_u$, $\mathrm{Var}(X) = \Sigma_x$ and some appropriate positive semidefinite covariance matrix such that $\mathrm{Cov}(U, X) = \rho \in (-1, 1)^{p \times q}$.

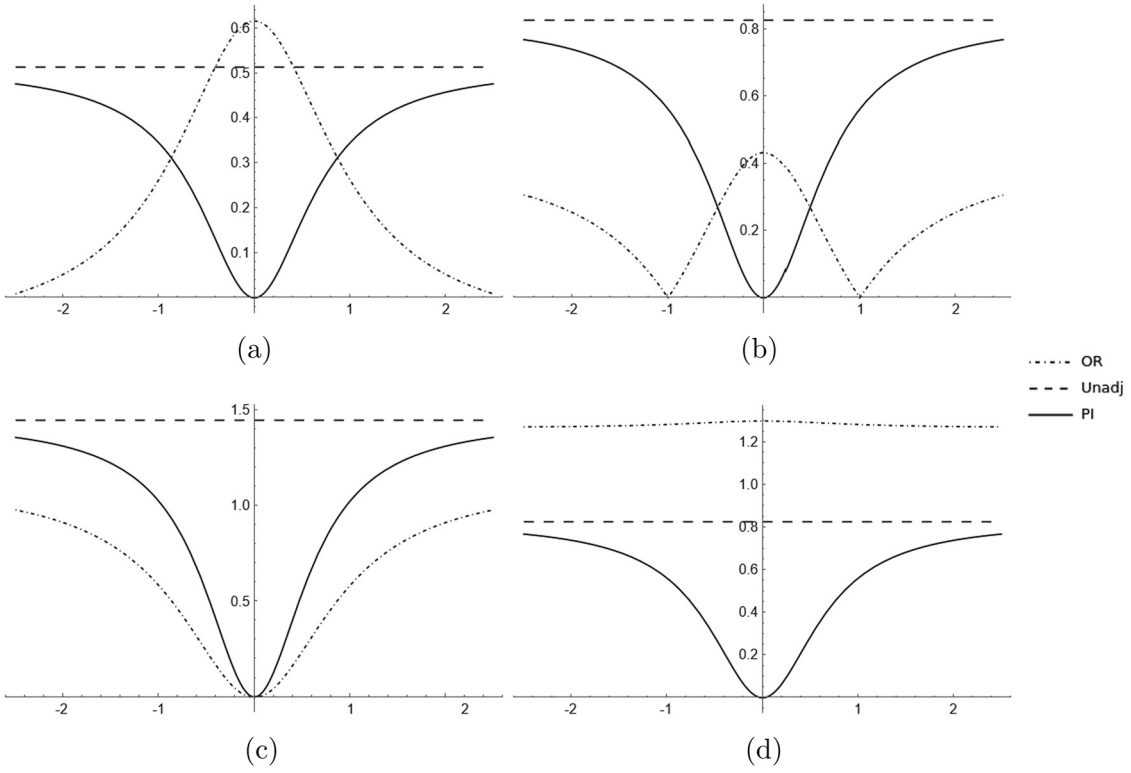

Figure 7: Plots of the ACE estimate bias under DGP (21) and under violations of completeness Assumption 8(a), where $\theta_{u_1}\mu_{u_1}$ and $\theta_{u_2}\mu_{u_2}$ have the same sign as in Theorem 5(a). The completeness violation is imposed by setting both components of U to be associated with both negative controls and only including one NCE and one NCO, as per Theorem 2. Along the x-axis, we vary the strength of association between U_2 and Z and W, by varying θ_{u_2} and μ_{u_2} (set to be equal to each other, $\theta_{u_2} = \mu_{u_2}$, so they always have positive product). The different panels correspond to different values of α_{u_1} governing the strength of confounding by U_1. The figure shows that, as predicted by Theorem 5(a), the bias of the proximal outcome estimator ("PI", **solid** line) is always less than the unadjusted bias ("Unadj", dashed line —). However, the relationship of PI with the outcome adjusted estimator ("OR", dotted line ⋯) varies with the strength of confounding α_{u_1} across panels. All other parameters are fixed, with values: $\alpha_0 = \gamma_0 = \theta_0 = \mu_0 = 0$, $\theta_a = \theta_{u_1} = 1$, $\gamma_{u_1} = 1$, $\gamma_{u_a1} = 1.5$, $\mu_{u_1} = 0.5$, $\gamma_a = 0.5$, similar to the simulation in [15]. (a) $\alpha_{u_1} = 0.3$, (b) $\alpha_{u_1} = 0.5$, (c) $\alpha_{u_1} = 1$, (d) $\alpha_{u_1} = 0.5$.

We consider i.i.d. data generated by:

$$\begin{aligned}
\begin{pmatrix} U \\ X \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0}_p \\ \mathbf{0}_q \end{bmatrix}, \begin{bmatrix} \Sigma_u & \rho \\ \rho^T & \mathbf{\Sigma}_x \end{bmatrix}\right), \quad \rho \in (-1,1)^{p\times q}, \\
\operatorname{logit}(\mathbb{P}(A=1|U,X)) &= \alpha_0 + \alpha_u^T U + \alpha_x^T X, \\
Z &= \theta_0 + \theta_a A + \theta_u^T U + \theta_x^T X + \varepsilon_1, \\
W &= \mu_0 + \mu_u^T U + \mu_x^T X + \varepsilon_2, \\
Y(a) &= \gamma_0 + \gamma_a a + \gamma_u^T U + \gamma_x X + \varepsilon_3, \\
\varepsilon_1, \varepsilon_2, \varepsilon_3 &\sim \mathcal{N}(0,1).
\end{aligned} \tag{24}$$

The following theorem provides a formula for the proximal outcome identification bias under a linear bridge function:

Theorem 6. *Let* $\mathbb{E}[AU] = (\mathbb{E}[AU_1], \ldots, \mathbb{E}[AU_p])$, $\mathbb{E}[AX] = (\mathbb{E}[AX_1], \ldots, \mathbb{E}[AX_p])$, *and*

$$B = \left(\Sigma_u - \rho\Sigma_x^{-1}\rho^T - \frac{(\mathbb{E}[AU] - \rho\Sigma_x^{-1}\mathbb{E}[AX])(\mathbb{E}[AU]^T - \mathbb{E}[AX]^T\Sigma_x^{-1}\rho^T)}{\mathbb{E}[A](1-\mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathbb{E}[AX]}\right)\theta_u.$$

If $(B^T\mu_u)^\dagger$ *denotes the Moore-Penrose inverse of* $B^T\mu_u$*, then fitting a linear outcome bridge function* $h(W, A, X) = b_0 + b_a A + b_w^T W + b_x^T X$ *under LSEM* (24) *yields a proximal outcome estimator bias equal to*

$$\delta_{\text{POR}} = \frac{\mathbb{E}[AU]^T - \mathbb{E}[AX]^T\Sigma_X^{-1}\rho^T}{\mathbb{E}[A](1-\mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_X^{-1}\mathbb{E}[AX]}(I_p - \mu_u(B^T\mu_u)^\dagger B^T)\gamma_u. \tag{25}$$

A proof of Theorem 6 (which also considers the case of general $\text{Var}(U) = \Sigma_u \in \mathbb{R}^p$) can be found in Appendix C.6.

Remark 1. If $m = n = p$ and $B^T\mu_u$ has full rank, then $\delta_{\text{POR}} = 0$. If $p < m$ or $p < n$, then we have a similar discussion as in [17] where we can either consider the Moore-Penrose inverse of $B^T\mu_u$, or reduce the dimensions of Z and W until they match the dimension of U.

Theorem 6 enables sensitivity analysis. Note that the terms $\mathbb{E}[A]$ and $\mathbb{E}[AX]$ in (25) can be estimated from data. Thus, to perform a sensitivity analysis using the bias formula (25), it remains for the analyst to specify parameters $\mathbb{E}[AU]$ (which is determined by α_u), μ_u, γ_u, and ρ. An analyst could specify a distribution over these parameters, which, via (25), would imply a distribution over δ as each realization of the parameters drawn from the distribution would correspond to a different bias δ. In the following section, we provide an example of this procedure applied to real data.

5 Illustration of sensitivity analysis on the SUPPORT data

In this section, we provide an illustrative sensitivity analysis of the proximal inference application in [8] using the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT) dataset. We do not aim to be prescriptive, but rather to provide an example of how one might apply the bias formulas we derived to explore the extent of likely bias in a proximal inference analysis. Importantly, if data were not generated from an LSEM, then clearly our bias formulas will be incorrect. Still, the bias ensuing from violations of proximal inference assumptions under the LSEM assumption may serve as an approximation to the actual bias of a proximal inference analysis even if the data were not generated by an LSEM. We hope bias formulas for more general data generating processes might be developed in future work.

The SUPPORT data comprise 5,735 individuals, of which 2,184 were treated by RHC and 3,551 belonged to the control group. Outcome variable Y encodes the number of days between admission to the ICU and death or censoring at 30 days. The goal of the analysis was to estimate the ACE of RHC on this 30-day survival outcome. As in the study by Cui et al. [9], we consider 71 baseline covariates, including demographics and physiological measures, to construct the bins (X, Z, W) for confounding adjustment and confounding proxies. Cui et al. [9] reason that the 10 variables measuring patients' physiological status during the initial 24 hours in the ICU, which provide a snapshot of underlying physiological state subject to measurement error, may be viewed as confounding proxies. They are valid NCOs because they precede treatment. They are valid NCEs because physicians did not base treatment decisions on their values, and as mere measurements, they could not directly impact health outcomes in any other way. (It is also important that they are noisy measurements, as the actual underlying values of what they seek to measure could influence both treatment decisions and outcomes.) Of these 10 measurements, four are allocated to the negative control bins Z = (pafi1, paco21) and W = (ph1, hema1) based on strength of association with the A and Y, respectively. The remaining 67 variables are collected under X. Like Cui et al. [9], we specify the outcome confounding bridge function $h(W, A, X; b) = b_0 + b_a A + b_x^T X + b_w^T W$ to compute the proximal outcome regression estimate $\widehat{\psi}_{\text{POR}}$ of the ACE.

We assess the potential impact of assumption violations by evaluating bias formula (25) under draws from an assumed distribution on the dimension of U and parameters $\alpha_u, \gamma_u, \theta_u, \mu_u, \rho$ from (24). We consider the following framework for drawing the sensitivity parameters:

- Draw $p = \dim(U)$ (i.e., number of independent components of unobserved confounders U) from a Poisson distribution with mean λ (which we set to $\lambda = 5$).

- Set a subset of components of U to violate U-relevance, i.e., be unassociated with Z and W, by setting the corresponding columns in θ_u and μ_u to zero. The set of U-relevance violating components is selected randomly as follows:
  - Set proportion of violating U components $\pi \in [0, 1]$,
  - For each $i = 1,\ldots, d$, introduce relevance violation on component U_i with probability π.
- We considered two approaches to drawing ρ, which determines the covariance between unobserved confounders U and observed covariates X:
  - **Empirical correlation:** Construct covariance matrix $\rho = \text{Cov}(U, X)$ such that covariances between elements of U and X are of similar magnitude to covariances between elements of X (as in Appendix D.2).
  - **Uncorrelated:** U and X are uncorrelated, i.e., $\rho = \text{Cov}(U, X) = \mathbf{0}_{p\times q}$.

  In this setup, we assume that $\Sigma_u = I_p$ (i.e., the components of U are all uncorrelated) for illustrative purposes. In practice, one might consider a similar procedure for drawing Σ_u, informed by the empirical covariances between the components of X and other subject-matter input about the nature of unobserved confounding.
- Draw the parameters in θ_u and μ_u from uniform distributions over corresponding intervals $\theta_u \in [\boldsymbol{\theta}_{u,l}, \boldsymbol{\theta}_{u,r}]$, $\mu_u \in [\boldsymbol{\mu}_{u,l}, \boldsymbol{\mu}_{u,r}]$, where we set element-wise interval ends $\boldsymbol{\theta}_{u,l}, \boldsymbol{\theta}_{u,r}, \boldsymbol{\mu}_{u,l}, \boldsymbol{\mu}_{u,r}$. Details regarding how we selected these intervals are in Appendix D.2.
- The remaining parameters γ_u and $\mathbb{E}[AU]$ encoding the strength of association between U and (Y, A) are then constrained in terms of previously drawn sensitivity parameters and covariances that can be estimated from the data according to the following formulas (derived in Appendix D.1):

Constraining $\mathbb{E}[AU]$ (assuming fixed ρ):

$$\mathbb{E}[AU] = \rho\Sigma_x^{-1}\mathbb{E}[AX] + (\mu_u^T)^{\dagger}(\text{Cov}(W, A) - \text{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX]). \tag{26}$$

Constraining γ_u (assuming fixed $\mathbb{E}[AU]$ and ρ):

$$\begin{aligned}\gamma_u = [&\mu_u^T(I_p - \rho\Sigma_x^{-1}\rho^T) - (\text{Cov}(W, A) - \text{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX])\cdot \\ &\cdot\frac{(\mathbb{E}[AU]^T - \mathbb{E}[AX]^T\Sigma_x^{-1}\rho^T)}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathbb{E}[AX]}\Bigg]^{\dagger}\cdot[\text{Cov}(W, Y) - \text{Cov}(W, X)\Sigma_x^{-1}\text{Cov}(X, Y) \\ &- (\text{Cov}(W, A) - \text{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX])\cdot\frac{(\text{Cov}(A, Y) - \mathbb{E}[AX]^T\Sigma_x^{-1}\text{Cov}(X, Y))}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathbb{E}[AX]}\Bigg].\end{aligned} \tag{27}$$

To account for sampling variability of the covariance matrices plugged into the above formulas, we employ a bootstrapping strategy (Appendix D.2).

We would expect that settings with lower expected dimension λ of U, with lower probabilities π of U-relevance violations, and with ρ drawn according to the empirical correlation regime (ensuring that observed covariates X are good proxies for U) would lead to less bias. Table 1 contains sensitivity-adjusted

Table 1: Sensitivity-adjusted confidence intervals of the average treatment effect, where the intervals are computed using $[\delta_{j,.05} + \hat{\psi}_{\text{POR}} - 1.96 \times SE, \delta_{j,.95} + \hat{\psi}_{\text{POR}} + 1.96 \times SE]$

Row #	Setup	Sensitivity-adjusted CIs for $\widehat{\psi}_{\text{POR}}$
1	No bias-inducing U	(–2.65, –0.94)
2	No U-relevance violation ($\pi = 0$) + empirical correlation	(–2.67, –0.92)
3	$\pi = 0$ + uncorrelated (X, U)	(–3.36, –0.37)
4	$\pi = 1/3$ + empirical correlation	(–2.67, –0.92)
5	$\pi = 1/3$ + uncorrelated (X, U)	(–2.91, –0.31)
6	$\pi \sim \text{Unif}([0.2, 0.5])$ + empirical correlation	(–2.68, –0.91)
7	$\pi \sim \text{Unif}([0.2, 0.5])$ + uncorrelated (X, U)	(–2.94, 0.36)
8	$\pi \sim \text{Unif}([0.2, 0.8])$ + empirical correlation	(–2.68, –0.91)
9	$\pi \sim \text{Unif}([0.2, 0.8])$ + uncorrelated (X, U)	(–2.79, 0.55)

confidence intervals (CIs) for the ACE under various distributions of the sensitivity parameters within the framework outlined earlier. The first row replicates results from Cui et al. [9], assuming no bias. Let $\delta_{j,k}$ denote the k-quantile of the bias distribution under the setup for drawing sensitivity parameters in row j of Table 1. So $\delta_{1,k}$ is 0 for all k, since row 1 assumes no bias. (We note that $\delta_{j,k}$ usually takes extreme values for k near 0 or 1 when completeness is potentially violated, as can be gleaned from Figure 5.) We compute the sensitivity-adjusted CI in row j as $[\delta_{j,.05} + \hat{\psi}_{\text{POR}} - 1.96 \times SE, \delta_{j,.95} + \hat{\psi}_{\text{POR}} + 1.96 \times SE]$. Since $\dim(Z) = \dim(W) = 2$ and $\dim(U)$ is taken from a Poisson distribution with mean 5, rows 2 and 3 result from a mixture of nonviolating and completeness-violating structures, while rows 4–9 result from a mixture of unbiased, completeness-violating, and U-relevance violating structures.

In the presence of a rich adjustment set with significant correlation between X and U (i.e., the even-numbered rows of Table 1 using the empirical correlation setup for ρ), the impact of proximal inference assumption violations on $\widehat{\psi}_{\text{POR}}$ is quite small, presumably because X acts as a good proxy for U. However, in the uncorrelated (X, U) case, the sensitivity-adjusted CIs are significantly wider. If only completeness (not U-relevance) is violated, the sensitivity-adjusted intervals still exclude 0. Only in rows 7 and 9 (which allow a high proportion of U components to be independent of X and the negative controls) does the sensitivity-adjusted CI indicate that the data are compatible with a point estimate having the wrong sign.

Due to the interconnectedness of biological systems, we believe that most unmeasured confounders related to patients' pretreatment health status would be associated with both the covariates X and the NCEs and NCOs (which also reflect pretreatment health status). In Section 1, we posited that physician preference might be an unobserved confounder that violates U-relevance as it is unrelated to patient state. Physicians who prefer to perform RHCs may tend to have other preferences for posttreatment interventions that also impact the outcome. Perhaps time of admission could be another U-relevance violating confounder, if practice but not patient state varies with time of admission. However, it is difficult to conceive of large numbers of confounders independent of patient state, and the ones we identified are likely weak. Thus, we find settings with empirical correlation more plausible and interpret the sensitivity analysis to suggest that the results are probably robust to proximal inference assumption violations.

6 Discussion

By deriving bias formulas for proximal inference estimators under violations of completeness and U-relevance, we begin to gain insight into the sensitivity of proximal inference estimators to these bias sources. For example, under some LSEM settings, it is possible for completeness violations alone (i.e., too many common causes of the NCE and NCO) to lead to arbitrarily more bias in the proximal inference estimator than in an unadjusted estimator completely subject to unobserved confounding (Figure 5). However, under the conditions of Theorem 5, if the different components of the unobserved confounder induce associations between the NCE and NCO in the same direction, then the proximal inference estimator is guaranteed to perform better than an unadjusted one. Neither of these scenarios (infinite bias or guaranteed improvement over unadjusted) imposes any constraints on the observed data, highlighting the utility of bias analysis.

We have also shown how our bias formulas enable assumption-heavy sensitivity analysis of proximal inference estimates. While (25) was derived under the strong assumptions that data were generated by an LSEM and U is not an effect modifier, an analyst might reasonably conduct a sensitivity analysis using (25) as we described even if they did not believe the assumptions held for the data and did not construct their proximal inference estimators according to an LSEM. There is a long history of simplifying assumptions in sensitivity analysis. For example, VanderWeele and Arah [13] and Rosenbaum [18] assume a one-dimensional binary confounder for tractable sensitivity analysis of no unobserved confounding. Later, Ding and VanderWeele [2] developed an approach that made far fewer restrictions, allowing multidimensional and nonbinary unobserved confounders that may interact arbitrarily with the treatment. We are in the early stages of proximal inference, so we currently need to settle for preliminary insights into the behavior of proximal inference estimators under strong simplifying assumptions. However, because proximal inference is a

promising approach to causal inference that has rightfully garnered much attention from methodological researchers, it is important to begin probing its operating characteristics under violations of its assumptions.

Funding information: Funding was provided by the MIT-IBM Watson AI Lab.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Ethical approval: The research related to human use has been complied with all the relevant national regulations, institutional policies and in accordance the tenets of the Helsinki Declaration, and has been approved by the authors institutional review board or equivalent committee.

Informed consent: Informed consent was obtained from all individuals included in this study.

Data availability statement: These are secondary analyses that are de-identified and are publicly available. The datasets analysed during the current study are available from the corresponding author on reasonable request.

References

[1] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. Int J Epidemiol. 2009 Oct;38(5):1175–91. https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyp289.

[2] Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology. 2016 May;27(3):368–77.

[3] Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, editors. Statistical models in epidemiology, the environment, and clinical trials. The IMA Volumes in Mathematics and its Applications. New York, NY: Springer; 2000. p. 1–94.

[4] Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J R Stat Soc Ser B (Methodological). 1983;45(2):212–8. https://www.jstor.org/stable/2345524.

[5] Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostat. 2007;3(1):Article 14.

[6] Rambachan A, Roth J. A more credible approach to parallel trends. Rev Econ Stud. 2023 Oct;90(5):2555–91. doi: https://doi.org/10.1093/restud/rdad018.

[7] Shi X, Miao W, Tchetgen Tchetgen EJ. A selective review of negative control methods in epidemiology. arXiv:200905641 [stat]. 2020 Sep. ArXiv: 2009.05641. Available from: http://arxiv.org/abs/2009.05641.

[8] Tchetgen Tchetgen EJ, Ying A, Cui Y, Shi X, Miao W. An introduction to proximal causal learning. arXiv:200910982 [stat]. 2020 Sep. ArXiv: 2009.10982. Available from: http://arxiv.org/abs/2009.10982.

[9] Cui Y, Pu H, Shi X, Miao W, Tchetgen Tchetgen EJ. Semiparametric proximal causal inference. arXiv:201108411 [math, stat]. 2020 Nov. ArXiv: 2011.08411. Available from: http://arxiv.org/abs/2011.08411.

[10] Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika. 2018 Dec;105(4):987–93. https://academic.oup.com/biomet/article/105/4/987/5073056.

[11] Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014 Dec;43(6):1969–85.

[12] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol. 1974;66(5):688–701.

[13] VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology. 2011 Jan;22(1):42–52. https://journals.lww.com/00001648-201101000-00008.

[14] Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. Math Model. 1986 Jan;7(9):1393–512. https://www.sciencedirect.com/science/article/pii/0270025586900886.

[15] Miao W, Shi X, Tchetgen Tchetgen EJ. A confounding bridge approach for double negative control inference on causal effects. arXiv:180804945 [stat]. 2020 Sep. ArXiv: 1808.04945. Available from: http://arxiv.org/abs/1808.04945.

[16] Newey WK, Powell JL. Instrumental variable estimation of nonparametric models. Econometrica. 2003;71(5):1565–78. https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00459.
[17] Shi X, Miao W, Nelson JC, Tchetgen Tchetgen EJ. Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. arXiv; 2019. ArXiv:1808.04906 [stat]. Available from: http://arxiv.org/abs/1808.04906.
[18] Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika. 1987;74(1):13–26. https://www.jstor.org/stable/2336017.

Appendix

A Bridge function parameters for post-treatment NCE

A.1 Bridge functions derivation for one-dimensional unobserved U – case of no violations

We identify coefficients $(b_0, b_a, b_x, b_w, b_{ax}, b_{aw})$ and (t_0, t_a, t_x, t_z) such that

$$\mathbb{E}[Y|U, a, X] = \int h(w, a, X)\mathrm{d}F(w|U, X), a = 0, 1, \tag{A1}$$

$$\frac{1}{\mathbb{P}[A = a|U, X]} = \int q(z, a, X)\mathrm{d}F(z|U, a, X), a = 0, 1. \tag{A2}$$

Coefficients of h:

We have that $\mathbb{E}[Y|U, A, X] = \gamma_0 + \gamma_a A + \gamma_x X + \gamma_u U + \gamma_{au} AU$, so (A1) implies

$$\gamma_0 + \gamma_a A + \gamma_x X + \gamma_u U + \gamma_{au} AU = b_0 + b_a A + b_x X + b_{ax} AX + \int (b_w + b_{aw} A)w \cdot \mathrm{d}F(w|U, X) \Leftrightarrow$$
$$\gamma_0 + \gamma_a A + \gamma_x X + \gamma_u U + \gamma_{au} AU = b_0 + b_a A + b_x X + b_{ax} AX + (b_w + b_{aw} A)\mathbb{E}[W|U, X].$$

Since $W|U, X \sim \mathcal{N}(\mu_0 + \mu_x X + \mu_u U, 1)$, we obtain

$$\gamma_0 + \gamma_a A + \gamma_x X + \gamma_u U + \gamma_{au} AU = b_0 + b_a A + b_x X + b_{ax} AX + (b_w + b_{aw} A)(\mu_0 + \mu_x X + \mu_u U).$$

Assigning values $A = 0, 1$, we obtain the following system

$$0 = \gamma_0 - b_0 - b_w\mu_0 + (\gamma_x - b_x - \mu_x b_w)X + (\gamma_u - b_w\mu_u)U, \tag{A3}$$

$$0 = (\gamma_0 + \gamma_a) - (b_0 + b_a) - (b_w + b_{aw})\mu_0 + (\gamma_x - b_x - b_{ax} - \mu_x(b_w + b_{aw}))X + (\gamma_u + \gamma_{au} - (b_w + b_{aw})\mu_u)U. \tag{A4}$$

Multiplying (A3) by U and X and taking the expectation in each resulting equations yields

$$0 = \rho(\gamma_x - b_x - \mu_x b_w) + (\gamma_u - b_w\mu_u),$$
$$0 = (\gamma_x - b_x - \mu_x b_w) + \rho(\gamma_u - b_w\mu_u).$$

Since $\rho \in (-1, 1)$, we obtain $\gamma_x - b_x - \mu_x b_w = \gamma_u - b_w\mu_u = 0$. From (A3), this additionally implies $\gamma_0 - b_0 - b_w\mu_0 = 0$.

Similarly, from (A4), we obtain $\gamma_a - b_a - \mu_0 b_{aw} = -b_{ax} - \mu_x b_{aw} = \gamma_{au} - b_{aw}\mu_u = 0$. Solving for the coefficients of h, we obtain the unique solution:

$$(b_0, b_a, b_x, b_w, b_{ax}, b_{aw}) = \left(\gamma_0 - \frac{\mu_0\gamma_u}{\mu_u}, \gamma_a - \frac{\mu_0\gamma_{au}}{\mu_u}, \gamma_x - \frac{\mu_x\gamma_u}{\mu_u}, \frac{\gamma_u}{\mu_u}, -\frac{\mu_x\gamma_{au}}{\mu_u}, \frac{\gamma_{au}}{\mu_u}\right).$$

Coefficients of q:

We have that $\mathbb{P}[A|X, U] = \frac{1}{1 + \exp\{(-1)^A(\alpha_0 + \alpha_x X + \alpha_u U)\}}$, such that (A2) implies

$$1 + \exp\{(-1)^A(\alpha_0 + \alpha_x X + \alpha_u U)\} = 1 + \exp\{(-1)^{1-A}(t_0 + t_a A + t_x X)\}\int \exp\{(-1)^{1-A} t_z Z\}\mathrm{d}F(z|U, A, X).$$

Since $Z|U, A, X \sim \mathcal{N}(\theta_0 + \theta_a A + \theta_u U + \theta_x X, 1)$, we obtain

$$
\begin{aligned}
&1 + \exp\{(-1)^A(\alpha_0 + \alpha_x X + \alpha_u U)\} \\
&\quad = 1 + \exp\{(-1)^{1-A}(t_0 + t_a A + t_x X)\} \int \exp\{(-1)^{1-A} t_z Z\} \mathrm{d}F(z|U, A, X) \\
&\quad = 1 + \exp\{(-1)^{1-A}(t_0 + t_a A + t_x X)\} \int \frac{1}{\sqrt{2\pi}} \exp\{(-1)^{1-A} t_z Z + 0.5(Z - \theta_0 - \theta_a A - \theta_u U - \theta_x X)^2\} \\
&\quad = 1 + \exp\left\{(-1)^{1-A}(t_0 + t_a A + t_x X) + (-1)^{1-A} t_z(\theta_0 + \theta_a A + \theta_u U + \theta_x X) + \frac{t_z^2}{2}\right\},
\end{aligned}
$$

for each $A = 0, 1$. This is equivalent to

$$
(-1)^A(\alpha_0 + \alpha_x X + \alpha_u U) = (-1)^{1-A}(t_0 + t_a A + t_x X) + (-1)^{1-A} t_z(\theta_0 + \theta_a A + \theta_u U + \theta_x X) + 0.5 t_z^2.
$$

Assigning values $A = 0, 1$, we obtain the system

$$
0 = \alpha_0 + t_0 + \theta_0 t_z - 0.5 t_z^2 + (\alpha_x + t_x + \theta_x t_z) X + (\alpha_u + \theta_u t_z) U, \tag{A5}
$$

$$
0 = \alpha_0 + (t_0 + t_a) + (\theta_0 + \theta_a) t_z + 0.5 t_z^2 + (\alpha_x + t_x + \theta_x t_z) X + (\alpha_u + \theta_u t_z) U. \tag{A6}
$$

As in the outcome bridge function case, it follows that the coefficients of 1 (the constant term), X, and U must be identically 0. We then obtain $\alpha_0 + t_0 + \theta_0 t_z - 0.5 t_z^2 = t_a + \theta_a t_z + t_z^2 = \alpha_x + t_x + \theta_x t_z = \alpha_u + \theta_u t_z = 0$, which yields the unique solution

$$
(t_0, t_a, t_x, t_z) = \left(-\alpha_0 + \frac{\theta_0}{\theta_u}\alpha_u + \frac{0.5}{\theta_u^2}\alpha_u^2, -\frac{1}{\theta_u^2}\alpha_u^2 + \frac{\theta_a}{\theta_u}\alpha_u, \frac{\theta_x}{\theta_u}\alpha_u - \alpha_x, -\frac{\alpha_u}{\theta_u}\right).
$$

B Proving violations of completeness Assumption 8(a)

We will prove that completeness Assumption 8(a) is violated under the DGP (17) with $\boldsymbol{\theta}_u = (\theta_{u_1} \;\; \theta_{u_2})^T$, $\theta_{u_1} \neq 0$. We note that case $\theta_{u_2} \neq 0$ can be treated symmetrically, by appropriately exchanging $u1$ and $u2$ in the following computations.

For any values u, z, a, x, we have that

$$
\begin{aligned}
\mathbb{P}[U = u|Z = z, A = a, X = x] &= \frac{\mathbb{P}[U = u, Z = z|A = a, X = x]}{\mathbb{P}[Z = z|A = a, X = x]} \\
&= \frac{\mathbb{P}[Z = z|U = u, A = a, X = x]\mathbb{P}[U = u|A = a, X = x]}{\mathbb{P}[Z = z|A = a, X = x]} \\
&= \frac{\mathbb{P}[\varepsilon_1 = z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u]\frac{\mathbb{P}[A = a \mid U = u, X = x]\mathbb{P}[U = u \mid X = x]}{\mathbb{P}[A = a \mid X = x]}}{\mathbb{P}[Z = z|A = a, X = x]}.
\end{aligned}
$$

Using

$$
\mathbb{P}[U = u|X = x] = \frac{\exp\left\{\frac{(\rho_2 u_1 - \rho_1 u_2)^2 - (u_2 - \rho_2 x)^2 - (u_1 - \rho_1 x)^2}{2(1 - \nu^2 - \rho_1^2 - \rho_2^2 + 2\nu\rho_1\rho_2)}\right\}}{2\pi\sqrt{1 - \rho_1^2 - \rho_2^2}},
$$

we obtain

$$
\begin{aligned}
&\mathbb{P}[U = u|Z = z, A = a, X = x] \\
&\quad = \frac{\frac{1}{\sqrt{2\pi}} \exp\{-0.5(Z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u)^2\}}{\mathbb{P}[Z = z, A = a|X = x](1 + \exp\{(-1)^a(\alpha_0 + \alpha_x x + \alpha_u^T u)\})} \\
&\qquad \cdot \frac{1}{2\pi\sqrt{1 - \rho_1^2 - \rho_2^2}} \exp\left\{\frac{(\rho_2 u_1 - \rho_1 u_2)^2 - (u_2 - \rho_2 x)^2 - (u_1 - \rho_1 x)^2}{2(1 - \nu^2 - \rho_1^2 - \rho_2^2 + 2\nu\rho_1\rho_2)}\right\} \\
&\quad = \frac{\exp\left\{\frac{1}{2}\frac{(\rho_2 u_1 - \rho_1 u_2)^2 - (u_2 - \rho_2 x)^2 - (u_1 - \rho_1 x)^2}{(1 - \nu^2 - \rho_1^2 - \rho_2^2 + 2\nu\rho_1\rho_2)} - \frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u)^2\right\}}{(2\pi)^{3/2}\sqrt{1 - \rho_1^2 - \rho_2^2}\,\mathbb{P}[Z = z, A = a|X = x](1 + \exp\{(-1)^a(\alpha_0 + \alpha_x x + \alpha_u^T u)\})}.
\end{aligned}
$$

Let us consider

$$
\begin{aligned}
g(U) = u_2\Bigg\{u_2^2 - 3 - \alpha_{u_2}^2 - \frac{\alpha_{u_1}^2\theta_{u_2}^2}{\theta_{u_1}^2} + \frac{2\alpha_{u_1}\alpha_{u_2}\theta_{u_2}}{\theta_{u_1}}\Bigg\}\exp\left\{-\frac{u_2^2}{2}\right\} \\
\cdot\exp\left\{-\frac{(\rho_2 u_1 - \rho_1 u_2)^2 - (u_2 - \rho_2 x)^2 - (u_1 - \rho_1 x)^2}{2(1 - \nu^2 - \rho_1^2 - \rho_2^2 + 2\nu\rho_1\rho_2)}\right\}\cdot \\
\cdot(2 + \exp(-\alpha_0 - \alpha_x x - \alpha_u^T u) + \exp(\alpha_0 + \alpha_x x + \alpha_u^T u)).
\end{aligned} \tag{A7}
$$

We will prove that $\mathbb{E}[g(U)|Z = z, A = a, X = x] = 0$ for any values z, a, x. We have

$$
\begin{aligned}
&\mathbb{E}[g(U)|Z = z, A = a, X = x] \\
&= \int\limits_{(-\infty,\infty)^2} g(u)\mathbb{P}[U = u|Z = z, A = a, X = x]\mathrm{d}u_1\mathrm{d}u_2 \\
&= \frac{1}{(2\pi)^{3/2}\sqrt{1 - \rho_1^2 - \rho_2^2}\,\mathbb{P}[Z = z, A = a|X = x]}\int\limits_{(-\infty,\infty)^2} u_2\Bigg\{u_2^2 - 3 - \alpha_{u_2}^2 - \frac{\alpha_{u_1}^2\theta_{u_2}^2}{\theta_{u_1}^2} + \frac{2\alpha_{u_1}\alpha_{u_2}\theta_{u_2}}{\theta_{u_1}}\Bigg\} \\
&\quad\cdot\exp\left\{-\frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u)^2 - \frac{u_2^2}{2}\right\}(1 + \exp\{(-1)^{1-a}(\alpha_0 + \alpha_x x + \alpha_u^T u)\})\mathrm{d}u_1\mathrm{d}u_2.
\end{aligned}
$$

Let

$$
\begin{aligned}
T_1 &= \int\limits_{(-\infty,\infty)^2} u_2\exp\left\{-\frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u)^2 - \frac{1}{2}u_2^2\right\}\cdot \\
&\quad\cdot(1 + \exp\{(-1)^{1-a}(\alpha_0 + \alpha_x x + \alpha_u^T u)\})\mathrm{d}u_1\mathrm{d}u_2, \\
T_2 &= \int\limits_{(-\infty,\infty)^2} u_2^3\exp\left\{-\frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u)^2 - \frac{1}{2}u_2^2\right\}\cdot \\
&\quad\cdot(1 + \exp\{(-1)^{1-a}(\alpha_0 + \alpha_x x + \alpha_u^T u)\})\mathrm{d}u_1\mathrm{d}u_2.
\end{aligned}
$$

We have that

$$
\begin{aligned}
&\int\limits_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_u^T u)^2\right\}(1 + \exp\{(-1)^{1-a}(\alpha_0 + \alpha_x x + \alpha_u^T u)\})\mathrm{d}u_1 \\
&= \exp\left\{-\frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2)^2\right\}\left\{\int\limits_{-\infty}^{\infty}\exp\left\{\theta_{u_1}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2)u_1 - \frac{1}{2}\theta_{u_1}^2 u_1^2\right\}\mathrm{d}u_1\right. \\
&\quad\left. + \exp\{(-1)^{1-a}(\alpha_0 + \alpha_x x + \alpha_{u_2}u_2)\}\cdot\right. \\
&\quad\left.\cdot\int\limits_{-\infty}^{\infty}\exp\left\{\theta_{u_1}\left(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2 - \frac{(-1)^a\alpha_{u_1}}{\theta_{u_1}}\right)u_1 - \frac{1}{2}\theta_{u_1}^2 u_1^2\right\}\mathrm{d}u_1\right) \\
&= \exp\left\{-\frac{1}{2}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2)^2\right\}\left(\frac{\sqrt{2\pi}}{|\theta_{u_1}|}\exp\left\{\frac{\theta_{u_1}^2(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2)^2}{2\theta_{u_1}^2}\right\}\right. \\
&\quad\left. + \exp\{(-1)^{1-a}(\alpha_0 + \alpha_x x + \alpha_{u_2}u_2)\}\cdot\frac{\sqrt{2\pi}}{|\theta_{u_1}|}\exp\left\{\frac{\theta_{u_1}^2\left(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2 - \frac{(-1)^a\alpha_{u_1}}{\theta_{u_1}}\right)^2}{2\theta_{u_1}^2}\right\}\right) \\
&= \frac{\sqrt{2\pi}}{|\theta_{u_1}|}\left(1 + \exp\left\{(-1)^{1-a}\left(\alpha_0 + \alpha_x x + \alpha_{u_2}u_2 + \frac{\alpha_{u_1}}{\theta_{u_1}}(Z - \theta_0 - \theta_a a - \theta_x x - \theta_{u_2}u_2)\right) + \frac{\alpha_{u_1}^2}{2\theta_{u_1}^2}\right\}\right) \\
&= \frac{\sqrt{2\pi}}{|\theta_{u_1}|}\left(1 + \exp\left\{(-1)^{1-a}\left(\alpha_0 + \alpha_x x + \frac{\alpha_{u_1}}{\theta_{u_1}}(Z - \theta_0 - \theta_a a - \theta_x x)\right) + \frac{\alpha_{u_1}^2}{2\theta_{u_1}^2} + (-1)^a\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}} - \alpha_{u_2}\right)u_2\right\}\right),
\end{aligned}
$$

which implies

$$
\begin{aligned}
T_1 &= \frac{\sqrt{2\pi}}{|\theta_{u_1}|}\exp\left\{(-1)^{1-a}\left(\alpha_0+\alpha_x x+\frac{\alpha_{u_1}}{\theta_{u_1}}(Z-\theta_0-\theta_a a-\theta_x x)\right)+\frac{\alpha_{u_1}^2}{2\theta_{u_1}^2}\right\}\\
&\quad\cdot\int_{-\infty}^{\infty}u_2\exp\left\{(-1)^a\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)u_2-\frac{1}{2}u_2^2\right\}\mathrm{d}u_2\\
&= \frac{\sqrt{2\pi}}{|\theta_{u_1}|}\exp\left\{(-1)^{1-a}\left(\alpha_0+\alpha_x x+\frac{\alpha_{u_1}}{\theta_{u_1}}(Z-\theta_0-\theta_a a-\theta_x x)\right)+\frac{\alpha_{u_1}^2}{2\theta_{u_1}^2}\right\}\\
&\quad\cdot\sqrt{2\pi}(-1)^a\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)\exp\left\{\frac{1}{2}\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)^2\right\}\\
&= \frac{2\pi(-1)^a}{|\theta_{u_1}|}\exp\left\{(-1)^{1-a}\left(\alpha_0+\alpha_x x+\frac{\alpha_{u_1}}{\theta_{u_1}}(Z-\theta_0-\theta_a a-\theta_x x)\right)\right.\\
&\quad\left.+\frac{\alpha_{u_1}^2(1+\theta_{u_2}^2)}{2\theta_{u_1}^2}-\frac{\theta_{u_1}\alpha_{u_1}\alpha_{u_2}}{\theta_{u_1}}-\frac{1}{2}\alpha_{u_2}^2\right\}\cdot\left\{\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
T_2 &= \frac{\sqrt{2\pi}}{|\theta_{u_1}|}\exp\left\{(-1)^{1-a}\left(\alpha_0+\alpha_x x+\frac{\alpha_{u_1}}{\theta_{u_1}}(Z-\theta_0-\theta_a a-\theta_x x)\right)+\frac{\alpha_{u_1}^2}{2\theta_{u_1}^2}\right\}\\
&\quad\cdot\int_{-\infty}^{\infty}u_2^3\exp\left\{(-1)^a\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)u_2-\frac{1}{2}u_2^2\right\}\mathrm{d}u_2\\
&= \frac{\sqrt{2\pi}}{|\theta_{u_1}|}\exp\left\{(-1)^{1-a}\left(\alpha_0+\alpha_x x+\frac{\alpha_{u_1}}{\theta_{u_1}}(Z-\theta_0-\theta_a a-\theta_x x)\right)+\frac{\alpha_{u_1}^2}{2\theta_{u_1}^2}\right\}\\
&\quad\cdot\sqrt{2\pi}(-1)^a\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)\left[3+\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)^2\right]\exp\left\{\frac{1}{2}\left(\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right)^2\right\}\\
&= \frac{2\pi(-1)^a}{|\theta_{u_1}|}\exp\left\{(-1)^{1-a}\left(\alpha_0+\alpha_x x+\frac{\alpha_{u_1}}{\theta_{u_1}}(Z-\theta_0-\theta_a a-\theta_x x)\right)\right.\\
&\quad\left.+\frac{\alpha_{u_1}^2(1+\theta_{u_2}^2)}{2\theta_{u_1}^2}-\frac{\theta_{u_1}\alpha_{u_1}\alpha_{u_2}}{\theta_{u_1}}-\frac{1}{2}\alpha_{u_2}^2\right\}\cdot\left\{\frac{\alpha_{u_1}\theta_{u_2}}{\theta_{u_1}}-\alpha_{u_2}\right\}\left(3+\alpha_{u_2}^2+\frac{\alpha_{u_1}^2\theta_{u_2}^2}{\theta_{u_1}^2}-\frac{2\alpha_{u_1}\alpha_{u_2}\theta_{u_2}}{\theta_{u_1}}\right),
\end{aligned}
$$

using the fact that $\int_{-\infty}^{\infty}u_2\exp\left\{-\frac{1}{2}u_2^2\right\}\mathrm{d}u_2=0$ and $\int_{-\infty}^{\infty}u_2^3\exp\left\{-\frac{1}{2}u_2^2\right\}\mathrm{d}u_2=0$ (as integrals of odd functions). We then obtain

$$
\begin{aligned}
\mathbb{E}[g(U)|Z=z,A=a,X=x] &= \frac{1}{(2\pi)^{3/2}\sqrt{1-\rho_1^2-\rho_2^2}\,\mathbb{P}[Z=z,A=a|X=x]}\\
&\quad\cdot\left(T_2-\left(3+\alpha_{u_2}^2+\frac{\alpha_{u_1}^2\theta_{u_2}^2}{\theta_{u_1}^2}-\frac{2\alpha_{u_1}\alpha_{u_2}\theta_{u_2}}{\theta_{u_1}}\right)T_1\right)=0,
\end{aligned}
$$

for any z, a, x. However, we clearly do not have $g(U)\equiv 0$ a.s., so completeness Assumption 8(a) does not hold.

C Bias computations

C.1 Computing the (asymptotic) bias obtained through method of moments estimator under setup (19)

We will compute the asymptotic bias obtained from the method of moments solver using bridge function $h(W, A, 0; b) = b_0 + b_a A + b_w W + b_{aw} AW$ and vector function $Q(A, Z, 0) = (1, A, Z, AZ)^T$.

We define the moment restrictions $H(D_i; \theta) = \begin{Bmatrix} \{Y_i - h(W_i, A_i, 0; b)\} \times Q(A_i, Z_i, 0) \\ \Delta - (h(W_i, 1, 0; b) - h(W_i, 0, 0; b)) \end{Bmatrix}$, and let $m(\theta) = \mathbb{E}[H(D; \theta)] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(D_i; \theta)$. The estimate of $\theta = (b, \Delta)$ is given by

$$\hat{\theta} = \arg\min_{\theta} m^T(\theta) m(\theta).$$

C.1.1 Case of Corr(U_1, U_2) = 0 (used in main paper)

By using $\mathbb{E}[U_1] = \mathbb{E}[U_2] = 0$, $\mathbb{E}[U_1^2] = \mathbb{E}[U_2^2] = 1$, and $\mathbb{E}[U_1 U_2] = 0$, we express the coordinates of $\mathbb{E}[h(D; \theta)] = (m_1, m_2, m_3, m_4, m_5)$ as follows:

$$m_1 = -b_0 - \mathbb{E}[A] b_a - \mu_0 b_w - (\mathbb{E}[A]\mu_0 + \mathbb{E}[AU_1]\mu_{u_1}) b_{aw} + \gamma_0 + \mathbb{E}[A]\gamma_a + \mathbb{E}[AU_1]\gamma_{au_1}, \tag{A8}$$

$$\begin{aligned} m_2 = &-\mathbb{E}[A] b_0 - \mathbb{E}[A] b_a - (\mathbb{E}[A]\mu_0 + \mathbb{E}[AU_1]\mu_{u_1}) b_w - (\mathbb{E}[A]\mu_0 + \mathbb{E}[AU_1]\mu_{u_1}) b_{aw} \\ &+ (\mathbb{E}[A](\gamma_0 + \gamma_a) + \mathbb{E}[AU_1](\gamma_{u_1} + \gamma_{au_1})) + \mathbb{E}[AU_2]\gamma_{u_2}, \end{aligned} \tag{A9}$$

$$\begin{aligned} m_3 = &-(\theta_0 + \mathbb{E}[A]\theta_a) b_0 - (\mathbb{E}[A](\theta_0 + \theta_a) + \mathbb{E}[AU_1]\theta_{u_1}) b_a \\ &- (\mu_0\theta_0 + \mu_{u_1}\theta_{u_1} + \mathbb{E}[A]\mu_0\theta_a + \mathbb{E}[AU_1]\mu_{u_1}\theta_a) b_w - (\mathbb{E}[A]\mu_0(\theta_0 + \theta_a) \\ &+ \mathbb{E}[AU_1](\mu_0\theta_{u_1} + \mu_{u_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\mu_{u_1}\theta_{u_1}) b_{aw} + \gamma_0\theta_0 + \gamma_{u_1}\theta_{u_1} + \mathbb{E}[A](\gamma_0\theta_a + \gamma_a(\theta_0 + \theta_a)) \\ &+ \mathbb{E}[AU_1](\gamma_a\theta_{u_1} + \gamma_{u_1}\theta_a + \gamma_{au_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\gamma_{au_1}\theta_{u_1} + \mathbb{E}[AU_2]\gamma_{u_2}\theta_a, \end{aligned} \tag{A10}$$

$$\begin{aligned} m_4 = &-(\mathbb{E}[A](\theta_0 + \theta_a) + \mathbb{E}[AU_1]\theta_{u_1}) b_0 - (\mathbb{E}[A](\theta_0 + \theta_a) + \mathbb{E}[AU_1]\theta_{u_1}) b_a \\ &- (\mathbb{E}[A]\mu_0(\theta_0 + \theta_a) + \mathbb{E}[AU_1](\mu_0\theta_{u_1} + \mu_{u_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\mu_{u_1}\theta_{u_1}) b_w \\ &- (\mathbb{E}[A]\mu_0(\theta_0 + \theta_a) + \mathbb{E}[AU_1](\mu_0\theta_{u_1} + \mu_{u_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\mu_{u_1}\theta_{u_1}) b_{aw} \\ &+ \mathbb{E}[A](\gamma_0 + \gamma_a)(\theta_0 + \theta_a) + \mathbb{E}[AU_1]((\gamma_0 + \gamma_a)\theta_{u_1} + (\gamma_{u_1} + \gamma_{au_1})(\theta_0 + \theta_a)) \\ &+ \mathbb{E}[AU_1^2](\gamma_{u_1} + \gamma_{au_1})\theta_{u_1} + \mathbb{E}[AU_2]\gamma_{u_2}(\theta_0 + \theta_a) + \mathbb{E}[AU_1U_2]\gamma_{u_2}\theta_{u_1}. \end{aligned} \tag{A11}$$

Let

$$R_1 = \frac{1}{(1 - \mathbb{E}[A])(1 - \mathbb{E}[AU_1^2]) - \mathbb{E}[AU_1]^2}, \tag{A12}$$

$$R_2 = (1 - \mathbb{E}[A])\mathbb{E}[AU_1U_2] + \mathbb{E}[AU_1]\mathbb{E}[AU_2]. \tag{A13}$$

We obtain the estimated bridge function parameters

$$\begin{aligned} \hat{b}_0 &= \gamma_0 - \frac{\mu_0}{\mu_{u_1}}\gamma_{u_1} + \left(\frac{\mu_0}{\mu_{u_1}} R_2 - \mathbb{E}[AU_1]\mathbb{E}[AU_1U_2] - (1 - \mathbb{E}[AU_1^2]\mathbb{E}[AU_2])\right) R_1 \cdot \gamma_{u_2}, \\ \hat{b}_w &= \frac{1}{\mu_{u_1}}\gamma_{u_1} - \frac{1}{\mu_{u_1}} R_1 R_2 \gamma_{u_2}, \\ \hat{b}_{aw} &= \frac{1}{\mu_{u_1}}\gamma_{au_1} + \frac{1}{\mu_{u_1}}\left(R_1 R_2 + \frac{\mathbb{E}[A]\mathbb{E}[AU_1U_2] - \mathbb{E}[AU_1]\mathbb{E}[AU_2]}{\mathbb{E}[A]\mathbb{E}[AU_1U_2] - \mathbb{E}[AU_1^2]}\right)\gamma_{u_2}. \end{aligned} \tag{A14}$$

The estimated effect resulting from $\hat{h}(W, A, 0; b)$ is then

$$\begin{aligned}\hat{\Delta} &= \hat{b}_a + \hat{b}_{aw}\mathbb{E}[W] = \hat{b}_a + \hat{b}_{aw}\mu_0 \\ &= \gamma_a + \frac{R_2\mathbb{E}[AU_1] - \mathbb{E}[AU_1^2](\mathbb{E}[AU_1]\mathbb{E}[AU_1U_2] + (1 - \mathbb{E}[AU_1^2])\mathbb{E}[AU_2])}{\mathbb{E}[AU_1]^2 - \mathbb{E}[A]\mathbb{E}[AU_1^2]}R_1 \cdot \gamma_{u_2},\end{aligned}$$

which yields a bias equal to

$$\delta = \frac{R_2\mathbb{E}[AU_1] - \mathbb{E}[AU_1^2](\mathbb{E}[AU_1]\mathbb{E}[AU_1U_2] + (1 - \mathbb{E}[AU_1^2])\mathbb{E}[AU_2])}{\mathbb{E}[AU_1]^2 - \mathbb{E}[A]\mathbb{E}[AU_1^2]}R_1 \cdot \gamma_{u_2}. \tag{A15}$$

We note that the expectations

$$\begin{aligned}\mathbb{E}[A] &= \mathbb{E}[\mathbb{E}[A|U_1, U_2]] = \mathbb{E}[\mathbb{P}[A = 1|U_1, U_2]] \\ &= \mathbb{E}\left[\frac{1}{1 + \exp\{-\alpha_0 - \alpha_{u_1}U_1 - \alpha_{u_2}U_2\}}\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\frac{1}{2\pi}\exp\left\{-\frac{u^2+v^2}{2}\right\}dudv}{1 + \exp\{-\alpha_0 - \alpha_{u_1}u - \alpha_{u_2}v\}},\end{aligned} \tag{A16}$$

$$\begin{aligned}\mathbb{E}[AU_1] &= \mathbb{E}[\mathbb{E}[AU_1|U_1, U_2]] = \mathbb{E}[U_1\mathbb{E}[A|U_1, U_2]] \\ &= \mathbb{E}\left[\frac{U_1}{1 + \exp\{-\alpha_0 - \alpha_{u_1}U_1 - \alpha_{u_2}U_2\}}\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\frac{1}{2\pi}u\exp\left\{-\frac{u^2+v^2}{2}\right\}dudv}{1 + \exp\{-\alpha_0 - \alpha_{u_1}u - \alpha_{u_2}v\}},\end{aligned} \tag{A17}$$

$$\begin{aligned}\mathbb{E}[AU_2] &= \mathbb{E}[\mathbb{E}[AU_2|U_1, U_2]] = \mathbb{E}[U_2\mathbb{E}[A|U_1, U_2]] \\ &= \mathbb{E}\left[\frac{U_2}{1 + \exp\{-\alpha_0 - \alpha_{u_1}U_1 - \alpha_{u_2}U_2\}}\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\frac{1}{2\pi}v\exp\left\{-\frac{u^2+v^2}{2}\right\}dudv}{1 + \exp\{-\alpha_0 - \alpha_{u_1}u - \alpha_{u_2}v\}},\end{aligned} \tag{A18}$$

$$\begin{aligned}\mathbb{E}[AU_1^2] &= \mathbb{E}[\mathbb{E}[AU_1^2|U_1, U_2]] = \mathbb{E}[U_1^2\mathbb{E}[A|U_1, U_2]] \\ &= \mathbb{E}\left[\frac{U_1^2}{1 + \exp\{-\alpha_0 - \alpha_{u_1}U_1 - \alpha_{u_2}U_2\}}\right] \\ &= \int_{-\infty}^{\infty}\frac{\frac{1}{2\pi}u^2\exp\left\{-\frac{u^2+v^2}{2}\right\}dudv}{1 + \exp\{-\alpha_0 - \alpha_{u_1}u - \alpha_{u_2}v\}},\end{aligned} \tag{A19}$$

$$\begin{aligned}\mathbb{E}[AU_1U_2] &= \mathbb{E}[\mathbb{E}[AU_1U_2|U_1, U_2]] = \mathbb{E}[U_1U_2\mathbb{E}[A|U_1, U_2]] \\ &= \mathbb{E}\left[\frac{U_1U_2}{1 + \exp\{-\alpha_0 - \alpha_{u_1}U_1 - \alpha_{u_2}U_2\}}\right] \\ &= \int_{-\infty}^{\infty}\frac{\frac{1}{2\pi}uv\exp\left\{-\frac{u^2+v^2}{2}\right\}dudv}{1 + \exp\{-\alpha_0 - \alpha_{u_1}u - \alpha_{u_2}v\}},\end{aligned} \tag{A20}$$

cannot be computed in closed form but can be obtained numerically using software like Mathematica or Maple once we provide the values of α_0 and α_u.

C.1.2 General case of $\mathrm{Corr}(U_1, U_2) = \nu$

Using $\mathbb{E}[U_1] = \mathbb{E}[U_2] = 0$, $\mathbb{E}[U_1^2] = \mathbb{E}[U_2^2] = 1$, and $\mathbb{E}[U_1U_2] = \nu$, the new coordinates of $\mathbb{E}[h(D;\theta)] = (m_1, m_2, m_3, m_4, m_5)$ result from

$$(m_1, m_2, m_3, m_4, m_5) = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[AY] \\ \mathbb{E}[ZY] \\ \mathbb{E}[AZY] \end{pmatrix} - \begin{pmatrix} 1 & \mathbb{E}[A] & \mathbb{E}[W] & \mathbb{E}[AW] \\ \mathbb{E}[A] & \mathbb{E}[A] & \mathbb{E}[AW] & \mathbb{E}[AW] \\ \mathbb{E}[Z] & \mathbb{E}[AZ] & \mathbb{E}[ZW] & \mathbb{E}[AZW] \\ \mathbb{E}[AZ] & \mathbb{E}[AZ] & \mathbb{E}[AZW] & \mathbb{E}[AZW] \end{pmatrix} b.$$

Proceeding similarly as in the case $\nu = 0$, we obtain a bias equal to

$$\begin{aligned} \delta &= \frac{R_2\mathbb{E}[AU_1] - \mathbb{E}[AU_1^2](\mathbb{E}[AU_1](\mathbb{E}[AU_1U_2] - \nu\mathbb{E}[A]) + (1 - \mathbb{E}[AU_1^2]\mathbb{E}[AU_2])) - \mathbb{E}[AU_1]^3\nu}{\mathbb{E}[AU_1]^2 - \mathbb{E}[A]\mathbb{E}[AU_1^2]} R_1\gamma_{u_2} \\ &= \frac{\mathbb{E}[AU_1](1 - \mathbb{E}[A] - \mathbb{E}[AU_1^2])\mathbb{E}[AU_1U_2] + \nu\mathbb{E}[AU_1](\mathbb{E}[A]\mathbb{E}[AU_1^2] - \mathbb{E}[AU_1]^2)}{\mathbb{E}[AU_1]^2 - \mathbb{E}[A]\mathbb{E}[AU_1^2]} R_1\gamma_{u_2} \\ &\quad + \frac{(\mathbb{E}[AU_1]^2 - (1 - \mathbb{E}[AU_1^2])\mathbb{E}[AU_1^2])\mathbb{E}[AU_2]}{\mathbb{E}[AU_1]^2 - \mathbb{E}[A]\mathbb{E}[AU_1^2]} R_1\gamma_{u_2}. \end{aligned}$$

C.2 Computing the (asymptotic) bias obtained through method of moments estimator under setup (21)

We will compute the asymptotic bias obtained from the method of moments solver using bridge function $h(W, A, 0; b) = b_0 + b_aA + b_wW + b_{aw}AW$ and vector function $Q(A, Z, 0) = (1, A, Z, AZ)^T$.

We define the moment restrictions $H(D_i;\theta) = \begin{Bmatrix} \{Y_i - h(W_i, A_i, 0; b)\} \times Q(A_i, Z_i, 0) \\ \Delta - (h(W_i, 1, 0; b) - h(W_i, 0, 0; b)) \end{Bmatrix}$, and let $m(\theta) = \mathbb{E}[H(D;\theta)] = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n h(D_i;\theta)$. The estimate of $\theta = (b, \Delta)$ is given by

$$\hat{\theta} = \arg\min_{\theta} m^T(\theta)m(\theta).$$

C.2.1 Case of $\mathrm{Corr}(U_1, U_2) = 0$ (used in main paper):

Using $\mathbb{E}[U_1] = \mathbb{E}[U_2] = 0$, $\mathbb{E}[U_1^2] = \mathbb{E}[U_2^2] = 1$, and $\mathbb{E}[U_1U_2] = 0$, we express the coordinates of $\mathbb{E}[h(D;\theta)] = (m_1, m_2, m_3, m_4, m_5)$ as follows:

$$\begin{aligned} m_1 &= -b_0 - \mathbb{E}[A]b_a - \mu_0 b_w - (\mathbb{E}[A]\mu_0 + \mathbb{E}[AU_1]\mu_{u_1})b_{aw} + \gamma_0 + \mathbb{E}[A]\gamma_a + \mathbb{E}[AU_1]\gamma_{au_1}, \\ m_2 &= -\mathbb{E}[A]b_0 - \mathbb{E}[A]b_a - (\mathbb{E}[A]\mu_0 + \mathbb{E}[AU]\mu_{u_1})b_w - (\mathbb{E}[A]\mu_0 + \mathbb{E}[AU_1]\mu_{u_1})b_{aw} \\ &\quad + (\mathbb{E}[A](\gamma_0 + \gamma_a) + \mathbb{E}[AU_1](\gamma_{u_1} + \gamma_{au_1})), \\ m_3 &= -(\theta_0 + \mathbb{E}[A]\theta_a)b_0 - (\mathbb{E}[A](\theta_0 + \theta_a) + \mathbb{E}[AU_1]\theta_{u_1})b_a \\ &\quad - (\mu_0\theta_0 + \mu_{u_1}\theta_{u_1} + \mu_{u_2}\theta_{u_2} + \mathbb{E}[A]\mu_0\theta_a + \mathbb{E}[AU_1]\mu_{u_1}\theta_a)b_w \\ &\quad - (\mathbb{E}[A](\mu_0(\theta_0 + \theta_a) + \mu_{u_2}\theta_{u_2}) + \mathbb{E}[AU_1](\mu_0\theta_{u_1} + \mu_{u_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\mu_{u_1}\theta_{u_1})b_{aw} \\ &\quad + \gamma_0\theta_0 + \gamma_{u_1}\theta_{u_1} + \mathbb{E}[A](\gamma_0\theta_a + \gamma_a(\theta_0 + \theta_a)) + \mathbb{E}[AU_1](\gamma_a\theta_{u_1} + \gamma_{u_1}\theta_a + \gamma_{au_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\gamma_{au_1}\theta_{u_1}, \end{aligned}$$

$$
\begin{aligned}
m_4 = &- (\mathbb{E}[A](\theta_0 + \theta_a) + \mathbb{E}[AU]\theta_{u_1})b_0 - (\mathbb{E}[A](\theta_0 + \theta_a) + \mathbb{E}[AU_1]\theta_{u_1})b_a \\
&- (\mathbb{E}[A](\mu_0(\theta_0 + \theta_a) + \mu_{u_2}\theta_{u_2}) + \mathbb{E}[AU_1](\mu_0\theta_{u_1} + \mu_{u_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\mu_{u_1}\theta_{u_1})b_w \\
&- (\mathbb{E}[A](\mu_0(\theta_0 + \theta_a) + \mu_{u_2}\theta_{u_2}) + \mathbb{E}[AU_1](\mu_0\theta_{u_1} + \mu_{u_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\mu_{u_1}\theta_{u_1})b_{aw} \\
&+ \mathbb{E}[A](\gamma_0 + \gamma_a)(\theta_0 + \theta_a) + \mathbb{E}[AU_1](\gamma_{u_1}\theta_a + \gamma_{au_1}(\theta_0 + \theta_a)) + \mathbb{E}[AU_1^2]\gamma_{au_1}\theta_{u_1}.
\end{aligned}
$$

Let

$$
\begin{aligned}
S_1 &= \frac{(1 - \mathbb{E}[A])^2}{(1 - \mathbb{E}[A])(1 - \mathbb{E}[AU_1^2]) - \mathbb{E}[AU_1]^2}, \\
S_2 &= \frac{\mathbb{E}[A]^2}{\mathbb{E}[A]\mathbb{E}[AU_1^2] - \mathbb{E}[AU_1]^2}.
\end{aligned}
$$

We obtain the estimated bridge function parameters

$$
\begin{aligned}
\hat{b}_0 &= \gamma_0 - \left(\frac{\mu_0}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} + \frac{S_1}{1 - \mathbb{E}[A]} \cdot \frac{\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1}, \\
\hat{b}_a &= \gamma_a - \frac{\mu_0 - \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A]}S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{au_1} + \left(\frac{\mu_0 + \frac{\mathbb{E}[AU_1]}{1 - \mathbb{E}[A]}S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} - \frac{\mu_0 - \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A]}S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1}, \\
\hat{b}_w &= \frac{1}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{u_1}, \\
\hat{b}_{aw} &= \frac{1}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{au_1} - \left(\frac{1}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} - \frac{1}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1}.
\end{aligned}
\tag{A21}
$$

The estimated effect resulting from $\hat{h}(W, A, 0; b)$ is then

$$
\begin{aligned}
\hat{\Delta} &= \hat{b}_a + \hat{b}_{aw}\mathbb{E}[W] = \hat{b}_a + \hat{b}_{aw}\mu_0 \\
&= \gamma_a + \frac{\frac{\mathbb{E}[AU_1]}{\mathbb{E}[A]}S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{au_1} + \left(\frac{\frac{\mathbb{E}[AU_1]}{1 - \mathbb{E}[A]}S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} + \frac{\frac{\mathbb{E}[AU_1]}{\mathbb{E}[A]}S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1} \\
&= \gamma_a + \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A](1 - \mathbb{E}[A])}\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}\left[\frac{(1 - \mathbb{E}[A])S_2}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{au_1} + \left(\frac{\mathbb{E}[A]S_1}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} + \frac{(1 - \mathbb{E}[A])S_2}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1} \right],
\end{aligned}
$$

which yields a bias equal to

$$
\delta = \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A](1 - \mathbb{E}[A])}\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}\left[\frac{(1 - \mathbb{E}[A])S_2}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}}\gamma_{au_1} + \left(\frac{\mathbb{E}[A]S_1}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} + \frac{(1 - \mathbb{E}[A])S_2}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1} \right]. \tag{A22}
$$

In the particular case $\gamma_{au_1} = 0$, we obtain a bias equal to

$$
\delta = \frac{\mathbb{E}[AU_1]}{\mathbb{E}[A](1 - \mathbb{E}[A])}\frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}\left(\frac{\mathbb{E}[A]S_1}{\mu_{u_1} + S_1 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} + \frac{(1 - \mathbb{E}[A])S_2}{\mu_{u_1} + S_2 \cdot \frac{\theta_{u_2}}{\theta_{u_1}}\mu_{u_2}} \right)\gamma_{u_1}. \tag{A23}
$$

Similarly to Proof C.1, we note that the expectations

$$\begin{aligned}\mathbb{E}[A] &= \mathbb{E}[\mathbb{E}[A|U_1]] = \mathbb{E}[\mathbb{P}[A=1|U_1]] \\ &= \mathbb{E}\left[\frac{1}{1+\exp\{-\alpha_0-\alpha_{u_1}U_1\}}\right] = \int_{-\infty}^{\infty}\frac{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2}\right\}}{1+\exp\{-\alpha_0-\alpha_{u_1}u\}}\mathrm{d}u,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[AU_1] &= \mathbb{E}[\mathbb{E}[AU_1|U_1]] = \mathbb{E}[U_1\mathbb{E}[A|U_1]] \\ &= \mathbb{E}\left[\frac{U_1}{1+\exp\{-\alpha_0-\alpha_u U_1\}}\right] = \int_{-\infty}^{\infty}\frac{\frac{1}{\sqrt{2\pi}}u\exp\left\{-\frac{u^2}{2}\right\}}{1+\exp\{-\alpha_0-\alpha_{u_1}u\}}\mathrm{d}u,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[AU_1^2] &= \mathbb{E}[\mathbb{E}[AU_1^2|U_1]] = \mathbb{E}[U_1^2\mathbb{E}[A|U_1]] \\ &= \mathbb{E}\left[\frac{U_1^2}{1+\exp\{-\alpha_0-\alpha_u U_1\}}\right] = \int_{-\infty}^{\infty}\frac{\frac{1}{\sqrt{2\pi}}u^2\exp\left\{-\frac{u^2}{2}\right\}}{1+\exp\{-\alpha_0-\alpha_{u_1}u\}}\mathrm{d}u,\end{aligned}$$

cannot be computed in closed form but can be obtained numerically using software like Mathematica or Maple once we provide the values of α_0 and α_u.

C.2.2 General case of Corr(U_1, U_2) = ν:

Using $\mathbb{E}[U_1] = \mathbb{E}[U_2] = 0$, $\mathbb{E}[U_1^2] = \mathbb{E}[U_2^2] = 1$, and $\mathbb{E}[U_1U_2] = \nu$, the new coordinates of $\mathbb{E}[h(D;\theta)] = (m_1, m_2, m_3, m_4, m_5)$ result from

$$(m_1, m_2, m_3, m_4, m_5) = \begin{pmatrix}\mathbb{E}[Y]\\ \mathbb{E}[AY]\\ \mathbb{E}[ZY]\\ \mathbb{E}[AZY]\end{pmatrix} - \begin{pmatrix}1 & \mathbb{E}[A] & \mathbb{E}[W] & \mathbb{E}[AW]\\ \mathbb{E}[A] & \mathbb{E}[A] & \mathbb{E}[AW] & \mathbb{E}[AW]\\ \mathbb{E}[Z] & \mathbb{E}[AZ] & \mathbb{E}[ZW] & \mathbb{E}[AZW]\\ \mathbb{E}[AZ] & \mathbb{E}[AZ] & \mathbb{E}[AZW] & \mathbb{E}[AZW]\end{pmatrix} b.$$

Let

$$\begin{aligned}T_1 &= \mathbb{E}[AU_1^2]\mathbb{E}[AU_2] - \mathbb{E}[AU_1]\mathbb{E}[AU_1U_2],\\ T_2 &= \mathbb{E}[AU_2^2]\mathbb{E}[AU_1] - \mathbb{E}[AU_2]\mathbb{E}[AU_1U_2],\\ V_{11} &= \mathbb{E}[AU_1]^2 - \mathbb{E}[A]\mathbb{E}[AU_1^2],\\ V_{22} &= \mathbb{E}[AU_2]^2 - \mathbb{E}[A]\mathbb{E}[AU_2^2],\\ V_{12} &= \mathbb{E}[AU_1]\mathbb{E}[AU_2] - \mathbb{E}[A]\mathbb{E}[AU_1U_2].\end{aligned}$$

Proceeding similarly as in the case $\nu = 0$, we obtain an estimated effect resulting from $\hat{h}(W, A, 0; b)$ equal to

$$\begin{aligned}\hat{\Delta} &= \hat{b}_a + \hat{b}_{aw}\mathbb{E}[W] = \hat{b}_a + \hat{b}_{aw}\mu_0\\ &= \gamma_a + \frac{\mu_{u_2}}{\theta_{u_1}}\left[\frac{T_1 - \frac{\theta_{u_2}}{\theta_{u_1}}T_2}{\mu_{u_1}\left(V_{11} + \frac{\theta_{u_2}}{\theta_{u_1}}V_{12}\right) + \mu_{u_2}\left(V_{12} + \frac{\theta_{u_2}}{\theta_{u_1}}V_{22}\right)}(\gamma_{u_1} + \gamma_{au_1})\right.\\ &\quad \left. + \frac{-T_1 + (\mathbb{E}[AU_2] - \nu\mathbb{E}[AU_1]) + \frac{\theta_{u_2}}{\theta_{u_1}}(T_2 - (\mathbb{E}[AU_1] - \nu\mathbb{E}[AU_2]))}{F}\gamma_{u_1}\right],\end{aligned}$$

where

$$F = \mu_{u_1}\left(V_{11} + \mathbb{E}[AU_1^2] + \frac{\theta_{u_2}}{\theta_{u_1}}(V_{12} + \mathbb{E}[AU_1U_2]) - (1 - \mathbb{E}[A])\left(1 + \frac{\theta_{u_2}}{\theta_{u_1}}\nu\right)\right) + \mu_{u_2}\left(V_{12} + \mathbb{E}[AU_1U_2] + \frac{\theta_{u_2}}{\theta_{u_1}}(V_{22} + \mathbb{E}[AU_2^2]) - (1 - \mathbb{E}[A])\left(\nu + \frac{\theta_{u_2}}{\theta_{u_1}}\right)\right).$$

This yields a bias equal to

$$\delta = \frac{\mu_{u_2}}{\theta_{u_1}}\left[\frac{T_1 - \frac{\theta_{u_2}}{\theta_{u_1}}T_2}{\mu_{u_1}\left(V_{11} + \frac{\theta_{u_2}}{\theta_{u_1}}V_{12}\right) + \mu_{u_2}\left(V_{12} + \frac{\theta_{u_2}}{\theta_{u_1}}V_{22}\right)}(\gamma_{u_1} + \gamma_{au_1}) + \frac{-T_1 + (\mathbb{E}[AU_2] - \nu\mathbb{E}[AU_1]) + \frac{\theta_{u_2}}{\theta_{u_1}}(T_2 - (\mathbb{E}[AU_1] - \nu\mathbb{E}[AU_2]))}{F}\gamma_{u_1}\right].$$

As in the previous case, expectations $\mathbb{E}[A]$, $\mathbb{E}[AU_1]$, and $\mathbb{E}[AU_1^2]$, as well as

$$\begin{aligned}\mathbb{E}[AU_2] &= \mathbb{E}[\mathbb{E}[AU_2|U_1]] = \mathbb{E}[\mathbb{E}[AU_2|U_1, A = 1]\mathbb{P}[A = 1|U_1] + \mathbb{E}[AU_2|U_1, A = 0]\mathbb{P}[A = 0|U_1]]\\ &= \mathbb{E}[\mathbb{E}[U_2|U_1, A = 1]\mathbb{P}[A = 1|U_1]] = \mathbb{E}[\mathbb{E}[U_2|U_1]\mathbb{P}[A = 1|U_1]] = \mathbb{E}[\nu U_1\mathbb{P}[A = 1|U_1]]\\ &= \nu\mathbb{E}[U_1\mathbb{E}[A|U_1]] = \nu\mathbb{E}[AU_1],\end{aligned}$$

$$\begin{aligned}\mathbb{E}[AU_1U_2] &= \mathbb{E}[\mathbb{E}[AU_1U_2|U_1]] = \mathbb{E}[\mathbb{E}[AU_1U_2|U_1, A = 1]\mathbb{P}[A = 1|U_1] + \mathbb{E}[AU_1U_2|U_1, A = 0]\mathbb{P}[A = 0|U_1]]\\ &= \mathbb{E}[\mathbb{E}[U_1U_2|U_1, A = 1]\mathbb{P}[A = 1|U_1]] = \mathbb{E}[U_1\mathbb{E}[U_2|U_1]\mathbb{P}[A = 1|U_1]] = \mathbb{E}[\nu U_1^2\mathbb{P}[A = 1|U_1]]\\ &= \nu\mathbb{E}[U_1^2\mathbb{E}[A|U_1]] = \nu\mathbb{E}[AU_1^2],\end{aligned}$$

$$\begin{aligned}\mathbb{E}[AU_2^2] &= \mathbb{E}[\mathbb{E}[AU_2^2|U_1]] = \mathbb{E}[\mathbb{E}[AU_2^2|U_1, A = 1]\mathbb{P}[A = 1|U_1] + \mathbb{E}[AU_2^2|U_1, A = 0]\mathbb{P}[A = 0|U_1]]\\ &= \mathbb{E}[\mathbb{E}[U_2^2|U_1, A = 1]\mathbb{P}[A = 1|U_1]] = \mathbb{E}[(1 - \nu^2 + \nu^2U_1^2)\mathbb{P}[A = 1|U_1]]\\ &= (1 - \nu^2)\mathbb{E}[\mathbb{P}[A = 1|U_1]] + \nu^2\mathbb{E}[U_1^2\mathbb{P}[A = 1|U_1]] = (1 - \nu^2)\mathbb{E}[A] + \nu^2\mathbb{E}[\mathbb{E}[AU_1^2|U_1]]\\ &= (1 - \nu^2)\mathbb{E}[A] + \nu^2\mathbb{E}[AU_1^2],\end{aligned}$$

which help simplify the bias formula.

C.3 Computing the OR estimator bias under setup (19)

Let $M = (1 \quad Z \quad W \quad A \quad AZ \quad AW)$. By the typical formula $\hat{\boldsymbol{b}} = (M^TM)^{-1}M^TY$ for the OLS estimator and the following

$$
\begin{aligned}
\mathbb{E}[Z] &= \theta_0 + \theta_a\mathbb{E}[A],\\
\mathbb{E}[W] &= \mu_0,\\
\mathbb{E}[AZ] &= (\theta_0 + \theta_a)\mathbb{E}[A] + \theta_{u_1}\mathbb{E}[AU_1],\\
\mathbb{E}[AW] &= \mu_0\mathbb{E}[A] + \mu_{u_1}\mathbb{E}[AU_1],\\
\mathbb{E}[Z^2] &= \theta_0^2 + \theta_{u_1}^2 + 1 + (\theta_a^2 + 2\theta_0\theta_a)\mathbb{E}[A] + 2\theta_0\theta_{u_1}\mathbb{E}[AU_1],\\
\mathbb{E}[W^2] &= \mu_0^2 + \mu_{u_1}^2 + 1,\\
\mathbb{E}[AZ^2] &= (1 + (\theta_0 + \theta_a)^2)\mathbb{E}[A] + 2(\theta_0 + \theta_a)\theta_{u_1}\mathbb{E}[AU_1] + \theta_{u_1}^2\mathbb{E}[AU_1^2]\\
\mathbb{E}[AW^2] &= (1 + \mu_0^2)\mathbb{E}[A] + 2\mu_0\mu_{u_1}\mathbb{E}[AU_1] + \mu_{u_1}^2\mathbb{E}[AU_1^2]\\
\mathbb{E}[AZW] &= ((\theta_0 + \theta_a)\mu_0 + \theta_{u_2}\mu_{u_2})\mathbb{E}[A] + ((\theta_0 + \theta_a)\mu_{u_1} + \theta_{u_1}\mu_0)\mathbb{E}[AU_1] + \theta_{u_1}\mu_{u_1}\mathbb{E}[AU_1^2]\\
\mathbb{E}[Y] &= \gamma_0 + \gamma_a\mathbb{E}[A] + \gamma_{au_1}\mathbb{E}[AU_1]\\
\mathbb{E}[ZY] &= \theta_0\gamma_0 + \theta_{u_1}\gamma_{u_1} + ((\theta_0 + \theta_a)\gamma_a + \theta_a\gamma_0)\mathbb{E}[A] + (\theta_{u_1}\gamma_a + \theta_a\gamma_{u_1} + (\theta_0 + \theta_a)\gamma_{au_1})\mathbb{E}[AU_1]\\
&\quad + \theta_{u_1}\gamma_{au_1}\mathbb{E}[AU_1^2] + \theta_a\gamma_{u_2}\mathbb{E}[AU_2]\\
\mathbb{E}[WY] &= \mu_0\gamma_0 + \mu_{u_1}\gamma_{u_1} + \mu_0\gamma_a\mathbb{E}[A] + (\mu_0\gamma_{au_1} + \mu_{u_1}\gamma_a)\mathbb{E}[AU_1] + \mu_{u_1}\gamma_{au_1}\mathbb{E}[AU_1^2]\\
\mathbb{E}[AY] &= (\gamma_0 + \gamma_a)\mathbb{E}[A] + (\gamma_{u_1} + \gamma_{au_1})\mathbb{E}[AU_1] + \gamma_{u_2}\mathbb{E}[AU_2]\\
\mathbb{E}[AZY] &= (\theta_0 + \theta_a)(\gamma_0 + \gamma_a)\mathbb{E}[A] + ((\theta_0 + \theta_a)(\gamma_{u_1} + \gamma_{au_1}) + \theta_{u_1}(\gamma_0 + \gamma_a))\mathbb{E}[AU_1]\\
&\quad + \theta_{u_1}(\gamma_{u_1} + \gamma_{au_1})\mathbb{E}[AU_1^2] + (\theta_0 + \theta_a)\gamma_{u_2}\mathbb{E}[AU_2] + \theta_{u_1}\gamma_{u_2}\mathbb{E}[AU_1U_2]\\
\mathbb{E}[AWY] &= \mu_0(\gamma_0 + \gamma_a)\mathbb{E}[A] + (\mu_0(\gamma_{u_1} + \gamma_{au_1}) + \mu_{u_1}(\gamma_0 + \gamma_a))\mathbb{E}[AU_1] + \mu_{u_1}(\gamma_{u_1} + \gamma_{au_1})\mathbb{E}[AU_1^2]\\
&\quad + \mu_0\gamma_{u_2}\mathbb{E}[AU_2] + \mu_{u_1}\gamma_{u_2}\mathbb{E}[AU_1U_2],
\end{aligned}
$$

we obtain a linear regression estimator bias equal to

$$
\begin{aligned}
\delta_{\mathrm{OR}} = {} & \frac{\left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\theta_{u_1}\right)(1+\mu_{u_2}^2) + \left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\frac{\mu_{u_1}\mu_{u_2}}{\theta_{u_2}}\right)\theta_{u_2}^2}{\left(1+\frac{\theta_{u_1}^2}{S_2}\right)(1+\mu_{u_2}^2) + \left(1+\frac{\mu_{u_1}^2}{S_2}\right)(1+\theta_{u_2}^2) - \left(1+2\frac{\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}}{S_2}\right)}\gamma_{u_1}\\
& + \Bigg[\frac{(1+\theta_{u_2}^2+\mu_{u_2}^2)\mathbb{E}[AU]}{\mathbb{E}[A](1-\mathbb{E}[A])}((1+\theta_{u_2}^2+\mu_{u_2}^2)\\
& + \left(1 - \frac{\mathbb{E}[AU^2]}{\mathbb{E}[A](1-\mathbb{E}[A])}\right)(\theta_{u_1}^2(1+\mu_{u_2}^2) + \mu_{u_1}^2(1+\theta_{u_2}^2) - 2\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2})\Bigg)\\
& + \theta_a\theta_{u_1}\left(1 + \mu_{u_2} - \frac{\mu_{u_1}\theta_{u_2}\mu_{u_2}}{\theta_{u_1}}\right)\left(-\left(\frac{\mathbb{E}[AU^2](1-\mathbb{E}[AU^2])}{\mathbb{E}[A](1-\mathbb{E}[A])} + \mathbb{E}[AU]^2\left(\frac{1}{\mathbb{E}[A]^2S_1} + \frac{1}{(1-\mathbb{E}[A])^2S_2}\right)\right)\right.\\
& \cdot(\theta_{u_1}^2(1+\mu_{u_2}^2) + \mu_{u_1}^2(1+\theta_{u_2}^2) - 2\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}) + \frac{(1+\theta_{u_2}^2+\mu_{u_2}^2)}{\mathbb{E}[A]^2(1-\mathbb{E}[A])^2}\\
& \cdot(\mathbb{E}[A]^4 - \mathbb{E}[A]^3(1+2\mathbb{E}[AU^2]) + 3\mathbb{E}[A]^2(\mathbb{E}[AU^2] + \mathbb{E}[AU]^2) - \mathbb{E}[A](3\mathbb{E}[AU]^2 + \mathbb{E}[AU^2])\\
& \left. + \mathbb{E}[AU^2]))\right]\cdot \prod_{i=1,2}\frac{1}{\left(1+\frac{\theta_{u_1}^2}{S_i}\right)(1+\mu_{u_2}^2) + \left(1+\frac{\mu_{u_1}^2}{S_i}\right)(1+\theta_{u_2}^2) - \left(1+2\frac{\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}}{S_i}\right)}\gamma_{au_1}.
\end{aligned}
\tag{A24}
$$

In particular, for $\gamma_{au_1} = 0$, we obtain a bias equal to

$$
\delta_{\mathrm{OR}} = \frac{\left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\theta_{u_1}\right)(1+\mu_{u_2}^2) + \left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\frac{\mu_{u_1}\mu_{u_2}}{\theta_{u_2}}\right)\theta_{u_2}^2}{\left(1+\frac{\theta_{u_1}^2}{S_2}\right)(1+\mu_{u_2}^2) + \left(1+\frac{\mu_{u_1}^2}{S_2}\right)(1+\theta_{u_2}^2) - \left(1+2\frac{\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}}{S_2}\right)}\gamma_{u_1}
\tag{A25}
$$

C.4 Computing the OR estimator bias under setup (21)

Let $M = (1 \;\; Z \;\; W \;\; A \;\; AZ \;\; AW)$. By the typical formula $\hat{\boldsymbol{b}} = (M^T M)^{-1} M^T Y$ for the OLS estimator and the following

$$
\begin{aligned}
\mathbb{E}[Z] &= \theta_0 + \theta_a \mathbb{E}[A],\\
\mathbb{E}[W] &= \mu_0,\\
\mathbb{E}[AZ] &= (\theta_0 + \theta_a)\mathbb{E}[A] + \theta_{u_1}\mathbb{E}[AU_1],\\
\mathbb{E}[AW] &= \mu_0 \mathbb{E}[A] + \mu_{u_1}\mathbb{E}[AU_1],\\
\mathbb{E}[Z^2] &= \theta_0^2 + \theta_{u_1}^2 + \theta_{u_2}^2 + 1 + (\theta_a^2 + 2\theta_0\theta_a)\mathbb{E}[A] + 2\theta_0\theta_{u_1}\mathbb{E}[AU_1],\\
\mathbb{E}[W^2] &= \mu_0^2 + \mu_{u_1}^2 + \mu_{u_2}^2 + 1,\\
\mathbb{E}[AZ^2] &= (1 + (\theta_0 + \theta_a)^2 + \theta_{u_2}^2)\mathbb{E}[A] + 2(\theta_0 + \theta_a)\theta_{u_1}\mathbb{E}[AU_1] + \theta_{u_1}^2\mathbb{E}[AU_1^2],\\
\mathbb{E}[AW^2] &= (1 + \mu_0^2 + \mu_{u_2}^2)\mathbb{E}[A] + 2\mu_0\mu_{u_1}\mathbb{E}[AU_1] + \mu_{u_1}^2\mathbb{E}[AU_1^2],\\
\mathbb{E}[AZW] &= ((\theta_0 + \theta_a)\mu_0 + \theta_{u_2}\mu_{u_2})\mathbb{E}[A] + ((\theta_0 + \theta_a)\mu_{u_1} + \theta_{u_1}\mu_0)\mathbb{E}[AU_1] + \theta_{u_1}\mu_{u_1}\mathbb{E}[AU_1^2],\\
\mathbb{E}[Y] &= \gamma_0 + \gamma_a \mathbb{E}[A] + \gamma_{au_1}\mathbb{E}[AU_1],\\
\mathbb{E}[ZY] &= \theta_0\gamma_0 + \theta_{u_1}\gamma_{u_1} + ((\theta_0 + \theta_a)\gamma_a + \theta_a\gamma_0)\mathbb{E}[A] + (\theta_{u_1}\gamma_a + \theta_a\gamma_{u_1} + (\theta_0 + \theta_a)\gamma_{au_1})\mathbb{E}[AU_1] + \theta_{u_1}\gamma_{au_1}\mathbb{E}[AU_1^2],\\
\mathbb{E}[WY] &= \mu_0\gamma_0 + \mu_{u_1}\gamma_{u_1} + \mu_0\gamma_a\mathbb{E}[A] + (\mu_0\gamma_{au_1} + \mu_{u_1}\gamma_a)\mathbb{E}[AU_1] + \mu_{u_1}\gamma_{au_1}\mathbb{E}[AU_1^2],\\
\mathbb{E}[AY] &= (\gamma_0 + \gamma_a)\mathbb{E}[A] + (\gamma_{u_1} + \gamma_{au_1})\mathbb{E}[AU_1],\\
\mathbb{E}[AZY] &= (\theta_0 + \theta_a)(\gamma_0 + \gamma_a)\mathbb{E}[A] + ((\theta_0 + \theta_a)(\gamma_{u_1} + \gamma_{au_1}) + \theta_{u_1}(\gamma_0 + \gamma_a))\mathbb{E}[AU_1] + \theta_{u_1}(\gamma_{u_1} + \gamma_{au_1})\mathbb{E}[AU_1^2],\\
\mathbb{E}[AWY] &= \mu_0(\gamma_0 + \gamma_a)\mathbb{E}[A] + (\mu_0(\gamma_{u_1} + \gamma_{au_1}) + \mu_{u_1}(\gamma_0 + \gamma_a))\mathbb{E}[AU_1] + \mu_{u_1}(\gamma_{u_1} + \gamma_{au_1})\mathbb{E}[AU_1^2],
\end{aligned}
$$

we obtain a linear regression estimator bias equal to

$$
\begin{aligned}
\delta_{\mathrm{OR}} = {} & \frac{\left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\theta_{u_1}\right)(1+\mu_{u_2}^2) + \left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\frac{\mu_{u_1}\mu_{u_2}}{\theta_{u_2}}\right)\theta_{u_2}^2}{\left(1+\frac{\theta_{u_1}^2}{S_2}\right)(1+\mu_{u_2}^2) + \left(1+\frac{\mu_{u_1}^2}{S_2}\right)(1+\theta_{u_2}^2) - \left(1+2\frac{\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}}{S_2}\right)}\gamma_{u_1}\\
& + \Bigg[\frac{(1+\theta_{u_2}^2+\mu_{u_2}^2)\mathbb{E}[AU]}{\mathbb{E}[A](1-\mathbb{E}[A])}((1+\theta_{u_2}^2+\mu_{u_2}^2)\\
& + \left(1 - \frac{\mathbb{E}[AU^2]}{\mathbb{E}[A](1-\mathbb{E}[A])}\right)(\theta_{u_1}^2(1+\mu_{u_2}^2) + \mu_{u_1}^2(1+\theta_{u_2}^2) - 2\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2})\Bigg)\\
& + \theta_a\theta_{u_1}\left(1 + \mu_{u_2} - \frac{\mu_{u_1}\theta_{u_2}\mu_{u_2}}{\theta_{u_1}}\right)\Bigg(-\left(\frac{\mathbb{E}[AU^2](1-\mathbb{E}[AU^2])}{\mathbb{E}[A](1-\mathbb{E}[A])} + \mathbb{E}[AU]^2\left(\frac{1}{\mathbb{E}[A]^2 S_1} + \frac{1}{(1-\mathbb{E}[A])^2 S_2}\right)\right)\\
& \cdot(\theta_{u_1}^2(1+\mu_{u_2}^2) + \mu_{u_1}^2(1+\theta_{u_2}^2) - 2\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}) + \frac{(1+\theta_{u_2}^2+\mu_{u_2}^2)}{\mathbb{E}[A]^2(1-\mathbb{E}[A])^2}\\
& \cdot(\mathbb{E}[A]^4 - \mathbb{E}[A]^3(1+2\mathbb{E}[AU^2]) + 3\mathbb{E}[A]^2(\mathbb{E}[AU^2] + \mathbb{E}[AU]^2) - \mathbb{E}[A](3\mathbb{E}[AU]^2 + \mathbb{E}[AU^2])\\
& + \mathbb{E}[AU^2]))\Bigg] \cdot \prod_{i=1,2}\frac{1}{\left(1+\frac{\theta_{u_1}^2}{S_i}\right)(1+\mu_{u_2}^2) + \left(1+\frac{\mu_{u_1}^2}{S_i}\right)(1+\theta_{u_2}^2) - \left(1+2\frac{\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}}{S_i}\right)}\gamma_{au_1}.
\end{aligned}
\tag{A26}
$$

In particular, for $\gamma_{au_1} = 0$, we obtain a bias equal to

$$
\delta_{\mathrm{OR}} = \frac{\left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\theta_{u_1}\right)(1+\mu_{u_2}^2) + \left(\frac{\mathbb{E}[AU]}{\mathbb{E}[A]} - \frac{(1-\mathbb{E}[A])\theta_a}{S_2}\frac{\mu_{u_1}\mu_{u_2}}{\theta_{u_2}}\right)\theta_{u_2}^2}{\left(1+\frac{\theta_{u_1}^2}{S_2}\right)(1+\mu_{u_2}^2) + \left(1+\frac{\mu_{u_1}^2}{S_2}\right)(1+\theta_{u_2}^2) - \left(1+2\frac{\theta_{u_1}\mu_{u_1}\theta_{u_2}\mu_{u_2}}{S_2}\right)}\gamma_{u_1}. \tag{A27}
$$

C.5 Comparison of proximal and unadjusted estimator biases under setup (21)

We begin by proving that both $S_1, S_2 > 0$:

Proof that $S_1, S_2 > 0$:

We have that

$$|\mathrm{Cov}(A, AU)| = |\mathbb{E}[A^2U]| = |\mathbb{E}[AU]| \leq \sqrt{\mathrm{Var}(A)\mathrm{Var}(AU)} = \sqrt{\mathrm{Var}(A)}\sqrt{\mathbb{E}[AU^2] - \mathbb{E}[AU]^2}$$

which implies $\mathbb{E}[AU]^2 \leq \mathrm{Var}(A)(\mathbb{E}[AU^2] - \mathbb{E}[AU]^2)$. It follows that

$$\begin{aligned}\mathbb{E}[A]\mathbb{E}[AU^2] &\geq \mathbb{E}[A]\cdot \frac{\mathbb{E}[AU]^2(1 + \mathrm{Var}(A))}{\mathrm{Var}(A)} = \mathbb{E}[A]\cdot \frac{\mathbb{E}[AU]^2(1 + \mathrm{Var}(A))}{\mathbb{E}[A] - \mathbb{E}[A]^2} \\ &= \frac{\mathbb{E}[AU]^2(1 + \mathrm{Var}(A))}{1 - \mathbb{E}[A]} \geq \mathbb{E}[AU]^2,\end{aligned}$$

since $1 + \mathrm{Var}(A) \geq 1$ and $1 - \mathbb{E}[A] \in (0, 1)$. Thus, $S_2 > 0$.

Similarly, if we consider $\bar{A} = 1 - A$ (such that $\bar{A}^2 = \bar{A}, \mathbb{E}[\bar{A}] = 1 - \mathbb{E}[A]$, $\mathrm{Var}(\bar{A}) = \mathrm{Var}(A), \mathbb{E}[\bar{A}U] = -\mathbb{E}[AU]$, and $\mathbb{E}[\bar{A}U^2] = 1 - \mathbb{E}[AU^2]$), we obtain

$$(1 - \mathbb{E}[A])(1 - \mathbb{E}[AU^2]) = \mathbb{E}[\bar{A}]\mathbb{E}[\bar{A}U^2] \geq \mathbb{E}[\bar{A}U]^2 = \mathbb{E}[AU]^2.$$

Thus, $S_1 > 0$ as well. □

Taking the ratio of magnitudes for the two biases, we have

$$\left|\frac{\delta_{\mathrm{POR}}}{\delta_{\mathrm{unadj}}}\right| = \left|\mathbb{E}[A]\cdot \frac{S_1}{\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} + S_1} + (1 - \mathbb{E}[A])\cdot \frac{S_2}{\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} + S_2}\right|.$$

Let $f(r) = \mathbb{E}[A]\cdot \frac{S_1}{r + S_1} + (1 - \mathbb{E}[A])\cdot \frac{S_2}{r + S_2}$ for $r \in (-\infty, -\min\{S_1, S_2\})$. We note that $f(r)$ is strictly increasing in r, that $\lim_{r\to-\infty} f(r) = 0$, and that $f(r) = 1$ has the unique solution $r^* = -S_1(1 - \mathbb{E}[A]) - S_2\mathbb{E}[A] < 0$. We consider the following four cases:

I. If $\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} \geq 0$, then $S_1, S_2 > 0$ imply that $\frac{S_i}{\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} + S_i} \in (0, 1)$ for $i = 1, 2$. Since $\mathbb{E}[A] \in (0, 1)$, it follows that $0 < \left|\frac{\hat{\delta}_{\mathrm{POR}}}{\hat{\delta}_{\mathrm{unadj}}}\right| < 1$.

II. If $-\min\{S_1, S_2\} \leq \frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} < 0$, then $S_1, S_2 > 0$ imply that $\frac{S_i}{\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} + S_i} > 1$ for $i = 1, 2$. Similarly, it follows that $\left|\frac{\hat{\delta}_{\mathrm{POR}}}{\hat{\delta}_{\mathrm{unadj}}}\right| > 1$. In particular, if $\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} = -\min\{S_1, S_2\}$, the proximal estimator bias can be arbitrarily large.

III. If $r^* \leq \frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} < -\min\{S_1, S_2\}$, then $\left|\frac{\hat{\delta}_{\mathrm{POR}}}{\hat{\delta}_{\mathrm{unadj}}}\right| \geq 1$.

IV. If $\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} < r^*$, then $0 \leq \left|\frac{\hat{\delta}_{\mathrm{POR}}}{\hat{\delta}_{\mathrm{unadj}}}\right| < 1$. In particular, $\frac{\theta_{u_1}\mu_{u_1}}{\theta_{u_2}\mu_{u_2}} = -\infty$ implies that the proximal estimator is unbiased (as either $\theta_{u_2} = 0$ or $\mu_{u_2} = 0$). □

C.6 Computing the proximal estimator bias under $\gamma_{au} = 0$ and $h(W, A, X) = b_0 + b_a A + b_x^T X + b_w^T W$

We will compute the asymptotic bias obtained from the method of moments solver using bridge function $h(W, A, X; b) = b_0 + b_a A + b_w^T W + b_x^T X$ and vector function $Q(A, Z, X) = (1, A, Z, X)$. We assume the general

case of multidimensional U, Z, W, X with $Z \in \mathbb{R}^m$, $W \in \mathbb{R}^n$, $U \in \mathbb{R}^p$, $X \in \mathbb{R}^q$. Throughout this section, we use the shorthand $\mathbb{E}[AU] = (\mathbb{E}[AU_1], \ldots, \mathbb{E}[AU_p])$ and $\mathbb{E}[AX] = (\mathbb{E}[AX_1], \ldots, \mathbb{E}[AX_q])$.

We define the moment restrictions $H(D_i; \theta) = \begin{Bmatrix} \{Y_i - h(W_i, A_i, X_i; b)\} \times Q(A_i, Z_i, X_i) \\ \Delta - (h(W_i, 1, X_i; b) - h(W_i, 0, X_i; b)) \end{Bmatrix}$, and let $m(\theta) = \mathbb{E}[H(D; \theta)] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(D_i; \theta)$. The estimate of $\theta = (b, \Delta)$ is given by

$$\hat{\theta} = \arg\min_{\theta} m^T(\theta) m(\theta).$$

Using $\mathbb{E}[U_i] = 0$, $\forall i = 1, \ldots, p$ and $\mathbb{E}[UU^T] = \Sigma_u$, as well as $\mathbb{E}[X_j] = 0, \forall j = 1, \ldots, q$, $\mathbb{E}[XX^T] = \Sigma_x$, and $\mathbb{E}[UX^T] = \rho$, we express the coordinates of $\mathbb{E}[h(D; \theta)] = (m_1, m_2, \boldsymbol{m}_3, \boldsymbol{m}_4)$ with $m_1, m_2 \in \mathbb{R}$, $\boldsymbol{m}_3 \in \mathbb{R}^m$, $\boldsymbol{m}_4 \in \mathbb{R}^q$ as follows:

$$\begin{aligned}
m_1 &= -b_0 - \mathbb{E}[A] b_a - \mu_0^T b_w + \gamma_0 + \mathbb{E}[A]\gamma_a, \\
m_2 &= -\mathbb{E}[A] b_0 - \mathbb{E}[A] b_a - (\mathbb{E}[A]\mu_0^T + \mathbb{E}[AU]^T \mu_u + \mathbb{E}[AX]^T \mu_x) b_w - \mathbb{E}[AX]^T b_x \\
&\quad + (\gamma_0 + \gamma_a)\mathbb{E}[A] + \mathbb{E}[AU]^T \gamma_u + \mathbb{E}[AX]^T \gamma_x, \\
m_3 &= -(\theta_0 + \theta_a \mathbb{E}[A]) b_0 - (\mathbb{E}[A](\theta_0 + \theta_a) + \theta_u^T \mathbb{E}[AU] + \theta_x^T \mathbb{E}[AX]) b_a \\
&\quad - ((\theta_0 + \mathbb{E}[A]\theta_a)\mu_0^T + (\theta_a \mathbb{E}[AU]^T + \theta_u^T \Sigma_u + \theta_x^T \rho^T)\mu_u + (\theta_a \mathbb{E}[AX]^T + \theta_x^T \Sigma_x + \theta_u^T \rho)\mu_x) b_w \\
&\quad - (\theta_a \mathbb{E}[AX]^T + \theta_u^T \rho + \theta_x^T \Sigma_x) b_x + (\theta_0 + \theta_a \mathbb{E}[A])\gamma_0 \\
&\quad + ((\theta_0 + \theta_a)\mathbb{E}[A] + \theta_u^T \mathbb{E}[AU] + \theta_x^T \mathbb{E}[AX])\gamma_a \\
&\quad + (\theta_a \mathbb{E}[AU]^T + \theta_u^T \Sigma_u + \theta_x^T \rho^T)\gamma_u + (\theta_a \mathbb{E}[AX]^T + \theta_u^T \rho + \theta_x^T \Sigma_x)\gamma_x, \\
m_4 &= -\mathbb{E}[AX] b_a - (\Sigma_x \mu_x + \rho^T \mu_u) b_w - \Sigma_x b_x + \mathbb{E}[AX]\gamma_a + \Sigma_x \gamma_x + \rho^T \gamma_u.
\end{aligned}$$

Under assumption $m = n$ and $p > m$: Let us define

$$\begin{aligned}
\beta &= \frac{\mathbb{E}[AU]^T - \mathbb{E}[AX]^T \Sigma_x^{-1} \rho^T}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T \Sigma_x^{-1} \mathbb{E}[AX]}, \\
B &= \left(\Sigma_u - \rho \Sigma_x^{-1} \rho^T - \frac{(\mathbb{E}[AU] - \rho \Sigma_x^{-1} \mathbb{E}[AX])(\mathbb{E}[AU]^T - \mathbb{E}[AX]^T \Sigma_x^{-1} \rho^T)}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T \Sigma_x^{-1} \mathbb{E}[AX]} \right) \theta_u.
\end{aligned}$$

Setting $m_1 = m_2 = m_{3i} = m_{4i} = 0$ for all $i = 1, \ldots, m$, $j = 1, \ldots, q$, we obtain solution

$$\begin{aligned}
b_0 &= \gamma_0 - \mathbb{E}[A]\beta\gamma_u - (\mu_0^T - \mathbb{E}[A]\beta\mu_u)(B^T\mu_u)^\dagger B^T \gamma_u, \\
b_a &= \gamma_a + \beta(I_p - \mu_u (B^T \mu_u)^\dagger B^T)\gamma_u, \\
b_w &= (B^T \mu_u)^\dagger B^T \gamma_u, \\
b_x &= \gamma_x + \Sigma_x^{-1}(\rho^T - \mathbb{E}[AX]\beta)\gamma_u - (\mu_x + \Sigma_x^{-1}(\rho^T - \mathbb{E}[AX]\beta)\mu_u)(B^T\mu_u)^\dagger B^T \gamma_u,
\end{aligned}$$

where $(B^T\mu_u)^\dagger$ denotes the Moore-Penrose inverse of $B^T\mu_u$. If $B^T\mu_u$ has full column rank, then $(B^T\mu_u)^\dagger$ corresponds to $(B^T\mu_u)^\dagger = (\mu_u^T B B^T \mu_u)^{-1} \mu_u^T B \gamma_u$.

The estimated effect resulting from $\hat{h}(W, A, X; b)$ is

$$\hat{\Delta} = \hat{b}_a = \gamma_a + \frac{\mathbb{E}[AU]^T - \mathbb{E}[AX]^T \Sigma_x^{-1} \rho^T}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T \Sigma_x^{-1} \mathbb{E}[AX]} (I_p - \mu_u (B^T \mu_u)^\dagger B^T)\gamma_u,$$

which yields a bias equal to

$$\delta = \frac{\mathbb{E}[AU]^T - \mathbb{E}[AX]^T \Sigma_x^{-1} \rho^T}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T \Sigma_x^{-1} \mathbb{E}[AX]} (I_p - \mu_u (B^T \mu_u)^\dagger B^T)\gamma_u.$$

D Details for illustrative sensitivity analysis on real data

D.1 Extracting relationships between U-parameters from the data

We have that

$$\begin{aligned}
\mathrm{Cov}(Z, A) &= \theta_a \mathbb{E}[A](1 - \mathbb{E}[A]) + \theta_u^T \mathbb{E}[AU] + \theta_x^T \mathbb{E}[AX], \\
\mathrm{Cov}(W, A) &= \mu_u^T \mathbb{E}[AU] + \mu_x^T \mathbb{E}[AX], \\
\mathrm{Cov}(X, Z) &= \mathbb{E}[AX]\theta_a^T + \rho^T \theta_u + \Sigma_x \theta_x, \\
\mathrm{Cov}(X, W) &= \rho^T \mu_u + \Sigma_x \mu_x, \\
\mathrm{Cov}(Z, W) &= \theta_a(\mathbb{E}[AU]^T \mu_u + \mathbb{E}[AX]^T \mu_x) + \theta_u^T \mu_u + \theta_u^T \rho \mu_x + \theta_x^T \rho^T \mu_u + \theta_x^T \Sigma_x \mu_x,
\end{aligned}$$

where $\mathbb{E}[A]$, $\mathbb{E}[AX]$, Σ_x, and the five covariance terms can be computed empirically from the data. Eliminating terms θ_x and μ_x, we obtain

$$\mathbb{E}[AU] - \rho\Sigma_x^{-1}\mathbb{E}[AX] = (\mu_u^T)^\dagger(\mathrm{Cov}(W, A) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX]), \tag{A28}$$

$$\theta_a = \frac{\mathrm{Cov}(Z, A) - \mathrm{Cov}(Z, X)\Sigma_x^{-1}\mathbb{E}[AX] - \theta_u^T(\mu_u^T)^\dagger(\mathrm{Cov}(W, A) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX])}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathbb{E}[AX]}, \tag{A29}$$

$$\mu_u^T(I_p - \rho\Sigma_x^{-1}\rho^T)\theta_u = \mathrm{Cov}(W, Z) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathrm{Cov}(X, Z) - (\mathrm{Cov}(W, A) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX])\theta_a^T, \tag{A30}$$

as well as

$$\begin{aligned}
\theta_x &= \Sigma_x^{-1}(\mathrm{Cov}(X, Z) - \mathbb{E}[AX]\theta_a^T - \rho^T\theta_u), \\
\mu_x &= \Sigma_x^{-1}(\mathrm{Cov}(X, W) - \rho^T\mu_u).
\end{aligned}$$

The aforementioned equations show that parameterizing θ_u and μ_u suffice towards identifying terms $\mathbb{E}[AU] - \rho\Sigma_x^{-1}\mathbb{E}[AX]$ and $(I_p - \rho\Sigma_x^{-1}\rho^T)$ in the bias formula (as long as the terms are identified via the pseudoinverses).

Moreover, we have that

$$\begin{aligned}
\mathrm{Cov}(Z, Y) &= \mathrm{Cov}(Z, X)\gamma_x + \mathrm{Cov}(Z, A)\gamma_a + (\theta_a\mathbb{E}[AU]^T + \theta_u^T + \theta_x^T\rho^T)\gamma_u, \\
\mathrm{Cov}(W, Y) &= \mathrm{Cov}(W, X)\gamma_x + \mathrm{Cov}(W, A)\gamma_a + (\mu_x^T\rho^T + \mu_u^T)\gamma_u, \\
\mathrm{Cov}(A, Y) &= \gamma_a\mathbb{E}[A](1 - \mathbb{E}[A]) + \gamma_u^T\mathbb{E}[AU] + \gamma_x^T\mathbb{E}[AX], \\
\mathrm{Cov}(X, Y) &= \gamma_a\mathbb{E}[AX] + \rho^T\gamma_u + \Sigma_x\gamma_x.
\end{aligned}$$

which imply

$$\begin{aligned}
&\mathrm{Cov}(W, Y) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathrm{Cov}(X, Y) \\
&- \frac{(\mathrm{Cov}(W, A) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX])(\mathrm{Cov}(A, Y) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathrm{Cov}(X, Y))}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathbb{E}[AX]} \\
&= \Bigg[\mu_u^T(I_p - \rho\Sigma_x^{-1}\rho^T) \\
&\quad - \frac{(\mathrm{Cov}(W, A) - \mathrm{Cov}(W, X)\Sigma_x^{-1}\mathbb{E}[AX])(\mathbb{E}[AU]^T - \mathbb{E}[AX]^T\Sigma_x^{-1}\rho^T)}{\mathbb{E}[A](1 - \mathbb{E}[A]) - \mathbb{E}[AX]^T\Sigma_x^{-1}\mathbb{E}[AX]}\Bigg]\gamma_u.
\end{aligned}$$

D.2 In-depth rationale for choice of sensitivity parameters

D.2.1 Choice of distribution for $p = \dim(U)$

The rate of the Poisson distribution can be adjusted depending on the expected number of independent unobserved confounders. In this case, we assume the large number of observed covariates in X (i.e.,

$\dim(X) = 67$) accounts for most confounding of the $A - Y$ association, and thus, the number of unobserved U is small compared to the dimension of X.

D.2.2 Drawing ρ – the base case

Construct covariance matrix $\rho = \mathrm{Cov}(U, X)$ such that covariances between elements of U and X are of similar magnitude to covariances between elements of X, as follows:

- Draw elements ρ_{ij} from the empirical distribution of pairwise covariances $\{(\Sigma_x)_{ij} : 1 \le i < j \le q\}$.
- Rescale each ρ_{ij} such that $\sum_j |\rho_{ij}| < 1$ for each $i = 1, \ldots, p$ and $\sum_j |\rho_{ji}| + \sum_k |\Sigma_x|_{ki} < 1$ for each $i = 1, \ldots, q$, to ensure positive semidefinite covariance matrix for $\begin{pmatrix} U \\ X \end{pmatrix}$.

We operate under the assumption that the pairwise covariances ρ_{ij} follow roughly the same distribution as the observed covariates' covariances in $\mathbf{\Sigma}_x$ (with a slight downwards shift in magnitude), given no additional information about the nature of unobserved U.

D.2.3 Choice of $\boldsymbol{\theta}_{u,l}$, $\boldsymbol{\theta}_{u,r}$, $\boldsymbol{\mu}_{u,l}$, $\boldsymbol{\mu}_{u,l}$

In the absence of additional priors on each of the unobserved confounders U, the take element-wise intervals $[(\boldsymbol{\theta}_{u,l})_{i,j}, (\boldsymbol{\theta}_{u,r})_{i,j}]$ to be equal to some interval $[\theta_l, \theta_r]$ for all $i = 1, \ldots, m$, $j = 1, \ldots, p$. Similarly, we take intervals $[(\boldsymbol{\mu}_{u,l})_{i,j}, (\boldsymbol{\mu}_{u,r})_{i,j}]$ to be equal to some $[\mu_l, \mu_r]$ for all $i = 1, \ldots, n$, $j = 1, \ldots, p$. In other words, if $\boldsymbol{e}_{m\times p}$, $\boldsymbol{e}_{n\times p}$ are matrices of all ones, then we take $[\boldsymbol{\theta}_{u,l}, \boldsymbol{\theta}_{r,l}] = [\theta_l \boldsymbol{e}_{m\times p}, \theta_r \boldsymbol{e}_{m\times p}]$, $[\boldsymbol{\mu}_{u,l}, \boldsymbol{\mu}_{r,l}] = [\mu_l \boldsymbol{e}_{n\times p}, \mu_r \boldsymbol{e}_{n\times p}]$.

To inform our choice of $\theta_l, \theta_r, \mu_l, \mu_r$, we run the following linear regressions (with intercept):

- `fit1`: regress Z onto (A, X),
- `fit2`: regress Z onto (A, X, W),
- `fit3`: regress W onto X,
- `fit4`: regress W onto (A, X, Z).

Assuming our LSEM, `fit2` yields estimates for θ_a (the coefficient of A) and $\theta_u^T(\mu_u^T)^\dagger$ (the coefficients of W). Moreover, `fit4` yields estimates for $\mu_u^T(\theta_u^T)^\dagger$ (the coefficients of Z) and $\mu_u^T(\theta_u^T)^\dagger\theta_a$ (the coefficient of A). The 95% CIs for the coefficients are included in $(-25, 25)$ for $\mu_u^T(\theta_u^T)^\dagger$ and $(-1, 1)$ for $\theta_u^T(\mu_u^T)^\dagger$. We then choose endpoints $[\theta_l, \theta_u]$, $[\mu_l, \mu_u]$ to ensure resulting samples $\theta_u \in [\theta_l \boldsymbol{e}_{m\times p}, \theta_r \boldsymbol{e}_{m\times p}]$, $\theta_u \in [\theta_l \boldsymbol{e}_{n\times p}, \theta_r \boldsymbol{e}_{n\times p}]$ are included in $(-25, 25)$ and $(-1, 1)$, respectively.

In addition, the coefficients of X in `fit1` and `fit3` represent biased estimates of θ_u and μ_u, respectively. Assuming these coefficients are at least informative for the magnitude (in terms of powers of 10) and not values of θ_u and μ_u, we choose $[\theta_l, \theta_r] = [-10, 10]$, $[\mu_l, \mu_r] = [-1.5, 1.5]$.

In fact, our experiments show that the distribution of biases does not change significantly for fixed $[\mu_l, \mu_r] = [-1.5, 1.5]$ and different choices of interval magnitudes $[\theta_l, \theta_r] \in \{[-1.5, 1.5], [-5, 5], [-10, 10]\}$, so we keep $[\theta_l, \theta_r] = [-1.5, 1.5]$ for the slight runtime improvement in the sampling process.

D.2.4 Bootstrapping strategy for setting $\mathbb{E}[\boldsymbol{AU}]$, $\boldsymbol{\gamma_u}$

We compute 500 bootstrap estimates of the covariance matrices and draw $\mathbb{E}[AU]$ and γ_u from the resulting distribution of (26) and (27) evaluated at the bootstrap covariance estimates.