Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Biometrika*. Author manuscript; available in PMC 2022 March 08.

Published in final edited form as:

*Biometrika*. 2022 March ; 109(1): 265–272. doi:10.1093/biomet/asab016.

# Identifiability of causal effects with multiple causes and a binary outcome

**DEHAN KONG**,

Department of Statistical Sciences, University of Toronto,700 University Avenue, Toronto, Ontario M5G 1X6, Canada

**SHU YANG**,

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.

**LINBO WANG**

Department of Statistical Sciences, University of Toronto, 700 University Avenue, Toronto, Ontario M5G 1X6, Canada

## Summary

Unobserved confounding presents a major threat to causal inference in observational studies. Recently, several authors have suggested that this problem could be overcome in a shared confounding setting where multiple treatments are independent given a common latent confounder. It has been shown that under a linear Gaussian model for the treatments,the causal effect is not identifiable without parametric assumptions on the outcome model. In this note, we show that the causal effect is indeed identifiable if we assume a general binary choice model for the outcome with a non-probit link. Our identification approach is based on the incongruence between Gaussianity of the treatments and latent confounder and non-Gaussianity of a latent outcome variable. We further develop a two-step likelihood-based estimation procedure.

### Keywords

Binary choice model; Latent ignorability; Unmeasured confounding

## 1. Introduction

Unmeasured confounding poses a major challenge to causal inference in observational studies. Without further assumptions, it is often impossible to identify the causal effects of interest. Classical approaches to mitigating bias due to unmeasured confounding include instrumental variable methods (Angrist et al., 1996; Hernán & Robins, 2006; Wang & Tchetgen Tchetgen, 2018), causal structure learning (Drton & Maathuis, 2017), invariance prediction (Peters et al., 2016), negative controls (Kuroki & Pearl, 2014; Miao et al., 2018), and sensitivity analysis (Cornfield et al., 1959).

Supplementary material

Supplementary Material available at *Biometrika* online includes examples, simulation results, and two data illustrations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Several recent publications have suggested an alternative approaches to this problem that assume shared confounding between multiple treatments and independence of treatments given the confounder (Tran & Blei, 2017; Ranganath & Perotte, 2019; Wang & Blei, 2019a,b). These approaches leverage information in a potentially high-dimensional treatment to aid causal identification. Such settings are prevalent in many contemporary areas, such as genetics, recommendation systems and neuroimaging studies. Unfortunately, in general the shared confounding structure is not sufficient for causal identification. D'Amour (2019, Theorem 1) showed that under a linear Gaussian treatment model, except in trivial cases, the causal effects are not identifiable without parametric assumptions on the outcome model.To address this nonidentifiability problem,D'Amour (2019) and Imai & Jiang (2019) suggested collecting auxiliary variables such as negative controls or instrumental variables. Along these lines, Wang & Blei (2019b) showed that the deconfounder algorithm of Wang & Blei (2019a) is valid given a set of negative controls, and Veitch et al. (2019) further found a negative control in network settings.

The present work contributes to this discussion by establishing a new identifiability result for causal effects, assuming a general binary choice outcome model with a non-probit link in addition to a linear Gaussian treatment model. Our result provides a counterpart to the nonidentifiability result of D'Amour (2019, Theorem 1).We use parametric assumptions in place of auxiliary data for causal identification. This is similar in spirit to Heckman's selection model (Heckman, 1979) for correcting bias from nonignorable missing data. In contrast to the case with normally distributed treatments and outcome, in general the observed data distribution may contain information beyond the first two moments, thereby providing many more nontrivial constraints for causal identification (Bentler, 1983; Bollen, 2014). In particular, our approach leverages the incongruence between Gaussianity of the treatments and latent confounder and non-Gaussianity of a latent outcome variable to achieve causal identification. A referee pointed out that this is related to previous results of Peters et al. (2009) and Imai & Jiang (2019, §2.1) in other contexts of causal inference. Our identification approach is accompanied by a simple likelihood-based estimation procedure, and we illustrate the method through synthetic and real data analyses in the Supplementary Material.

## 2. Framework

Let $A = \left(A^{(1)}, A^{(2)}, \ldots, A^{(p)}\right)^{\mathrm{T}}$ be a $p$-vector of continuous treatments, $Y$ an outcome, and $X$ a $q$-vector of observed pre-treatment variables. The observed data $\{(X_i, A_i, Y_i): i = 1, \ldots, n\}$ are independent samples from a superpopulation. Under the potential outcomes framework, $Y$ is the potential outcome had the patient received treatment $a = \left(a^{(1)}, \ldots, a^{(p)}\right)^{\mathrm{T}}$. We are interested in identifying and estimating the mean potential outcome $E\{Y(a)\}$. We make the stable unit treatment value assumption, under which $Y(a)$ is well-defined and $Y = Y(a)$ if $A = a$.

We assume the shared confounding structure under which the treatments are conditionally independent given the baseline covariates $X$ and a scalar latent confounder $U$. Figure 1 provides a graphical illustration of the setting.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

***Assumption* 1 (Latent ignorablity).**

For all $a$, $A \perp\!\!\!\perp Y(a) \mid (X, U)$.

Under Assumption 1, we have

$$E\{Y(a)\} = E_{X,U}\{E(Y \mid A = a, X, U)\}. \quad (1)$$

We consider a latent factor model for the treatments:

$$U \sim N(0, 1), \quad A = \theta U + \epsilon_A, \quad (2)$$

where $\epsilon_A \sim N\{0, \text{diag}(\sigma^2_{A,1}, \ldots, \sigma^2_{A,p})\}$ and $\epsilon_A \perp\!\!\!\perp U$ Wang & Blei (2019a) suggested first constructing an estimate of $U$, the so-called deconfounder, and then using (1) to identify the mean potential outcomes and causalcontrasts.However,aspointedoutbyD'Amour(2019), Assumption 1 and model (2) are not sufficient for identification of $E\{Y(a)\}$. See also Example S1 in the Supplementary Material for a counterexample where $Y$ follows a Gaussian structural equation model.

## 3. Identification with a binary outcome

We now study the identification problem with a binary $Y$, thereby operating under a different set of assumptions from those in Example S1. To fix ideas, we first consider the case without measured covariates $X$ and later extend the results to the case with $X$. We assume that treatments $A$ follow the latent factor model (2). We also assume the following binary choice model:

$$Y = \mathbb{1}(T \leqslant \alpha + \beta^{\mathrm{T}} A + \gamma U), \quad (3)$$

where an auxiliary latent variable $T$, independent of $(A, U)$, has a known cumulative distribution function $G$. Equivalently, model (3) can be written as $\text{pr}(Y = 1 \mid A, U) = G(\alpha + \beta^{\mathrm{T}} A + \gamma U)$. This class of models is general and includes common models for the binary outcome. For example, when $T$ follows a logistic distribution with mean 0 and scale 1, model (3) becomes a logistic model; when $T$ follows a standard normal distribution, model (3) is a probit model; when $T$ follows a central $t$ distribution, model (3) is a robit model (Liu, 2004; Ding, 2014).

Our main identification result is summarized in Theorem 1.

**Theorem 1.**

*Suppose that* Assumption 1, *models* (2) *and* (3) *and the following conditions hold:*

- **i.** there exist at least three elements of $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ that are nonzero, and there exists at least one $j \in \{1, \ldots, p\}$ such that $\gamma\theta_j \neq 0$ and its sign is known a priori;
- **ii.** $\text{pr}(Y = 1 \mid A = a)$ is not a constant function of a.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Then the parameters $\theta, \Sigma_{AA}, \alpha, \beta, \gamma$ and hence $E\{Y(a)\}$ are identifiable if and only if T is not deterministic or normally distributed.

Theorem 1 entails that identifiability of causal effects is guaranteed as long as the outcome follows a nontrivial binary choice model with any link function other than the probit. Condition (i) of the theorem is plausible when the latent confounder $U$ affects at least three treatments, for at least one of which subjectspecific knowledge allows the signs of $\theta_j$ and $\gamma$ to be determined. Condition (ii) requires that the observed outcome means differ across treatment levels, and can be checked from the observed data.

We now present an outline of our identification strategy leading to Theorem 1. Under model (2), $\left(U, A^{\mathrm{T}}\right)^{\mathrm{T}}$ follows a joint multivariate normal distribution

$$\begin{pmatrix} U \\ A \end{pmatrix} \sim N_{p+1}(0, \Sigma_J), \qquad \Sigma_J = \begin{pmatrix} 1 & \theta^{\mathrm{T}} \\ \theta & \Sigma_{AA} \end{pmatrix},$$

where $\Sigma_{AA} = \theta\theta^{\mathrm{T}} + \mathrm{diag}\left(\sigma^2_{A,1}, \ldots, \sigma^2_{A,p}\right)$. Therefore $U \mid A$ follows a univariate normal distribution with mean $\mu_{U \mid A} = \theta^{\mathrm{T}}\Sigma^{-1}_{AA}A$ and variance $\sigma^2_{U \mid A} = 1 - \theta^{\mathrm{T}}\Sigma^{-1}_{AA}\theta$.

The starting point of our identification approach is the following orthogonalization of $\left(U, A^{\mathrm{T}}\right)^{\mathrm{T}}$. Let $Z = \left(U - \mu_{U \mid A}\right)/\sigma_{U \mid A}$ be the standardized latent confounder conditional on $A$. Then $Z \perp\!\!\!\perp A$ and $Z$ follows a standard normal distribution. Model (3) then implies that

$$Y = \mathbb{1}\left(T \leqslant c_1 + c_2^{\mathrm{T}}A + c_3 Z\right), \tag{4}$$

where $c_1 = \alpha, c_2 = \left(c_2^{(1)}, \ldots, c_2^{(p)}\right)^{\mathrm{T}} = \beta + \gamma\theta^{\mathrm{T}}\Sigma^{-1}_{AA}, c_3 = \gamma\sigma_{U \mid A}$ and $(A,T,Z)$ are jointly independent.

The unknown parameters can then be identified in three steps.In the first step,we prove the identifiability of $\theta$ and $\Sigma_{AA}$ using standard results from factor analysis (Anderson & Rubin, 1956). In the second step, we study the binary choice model (4), and show that both $c_2$ and the distribution of $T - c_1 - c_3 Z$ are identifiable up to a positive scale parameter. In the third step, we show that when the distribution of $T$ is nondeterministic and non-Gaussian, one can leverage the incongruence between the Gaussianity of $Z$ and the non-Gaussianity of $T$ to identify $c_1$, $c_3$ and the scale parameter in the second step. The key to this step is the following lemma. Finally, we identify $\alpha,\beta,\gamma$ and hence $E\{Y(a)\}$ from $c_1$, $c_2$, $c_3$, $\theta$ and $\Sigma_{AA}$.

**Lemma 1.**

Suppose $T_1 = T - c_1 - c_3 Z$ and that T is independent of Z, where Z follows a standard normal distribution and $c_1$ and $c_3$ are constants. The following statements are equivalent.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**I.** There exist $(\tilde{C}, \tilde{c_1}, |\tilde{c_3}|) \ddagger (C, c_1, |c_3|)$, $\tilde{T} \overset{\mathcal{D}}{=} T$ and $\tilde{Z} \overset{\mathcal{D}}{=} Z$ such that $C\tilde{C} > 0$, $\tilde{T} \perp\!\!\!\perp \tilde{Z}$ and $CT_1 \overset{\mathcal{D}}{=} \tilde{C}(\tilde{T} - \tilde{c_1} - \tilde{c_3}\tilde{Z})$, where $E \overset{\mathcal{D}}{=} F$ means that the random variables E and F have the same distribution.

**II.** The random variable T is either deterministic or normally distributed.

***Remark* 1.**

In this paper we only allow $U$ to be a scalar. In this case, $\theta$ is identified up to its sign from the factor model, and it may be possible to identify the sign of $\theta$ from subject-matter knowledge. However, if $U$ is a multi-dimensional vector, then the factor model (2) becomes $A = \Theta U + \epsilon_A$, where $\Theta$ is the loading matrix. In this case, $\Theta$ is only identifiable up to a rotation. Consequently, in general, there are infinitely many causal effect parameters that are compatible with the observed data distribution; see Miao et al. (2020) for related discussions.

***Remark* 2.**

Example S1 in the Supplementary Material shows that when the continuous outcome $Y$ follows a Gaussian structural model, $E\{Y(a)\}$ is not identifiable. Intuitively, the binary outcome in a probit regression can be obtained by dichotomizing a continuous outcome following a Gaussian distribution, and there is no reason to believe that dichotomization improves identifiability. So it should not be surprising that $E\{Y(a)\}$ is not identifiable in the probit case.

In the presence of baseline covariates $X$, we assume that

$$A = \theta U + BX + \epsilon_A, \tag{5}$$

$$\text{pr}\{Y(a) = 1 \mid U, X\} = G(\alpha + \beta^{\mathrm{T}} a + \gamma U + \eta^{\mathrm{T}} X), \tag{6}$$

where $X \perp\!\!\!\perp (U, \epsilon_A)$. We also assume that

$$\begin{pmatrix} U \\ A \end{pmatrix} \mid X \sim N_{p+1}\left\{ \begin{pmatrix} 0 \\ BX \end{pmatrix}, \Sigma_J^* \right\}, \qquad \Sigma_J^* = \begin{pmatrix} 1 & \theta^{\mathrm{T}} \\ \theta & \Sigma_{A \mid X} \end{pmatrix}, \tag{7}$$

where $\Sigma_{A \mid X} = \Sigma_{AA} - B\Sigma_{XX}B^{\mathrm{T}}$ with $\Sigma_{AA}$ and $\Sigma_{XX}$ being the covariances of $A$ and $X$, respectively. Then $U \mid X = x, A = a$ follows a univariate normal distribution with mean $\mu_{U \mid x,a} = \theta^{\mathrm{T}}\Sigma_{A \mid X}^{-1}(a - Bx)$ and variance $\sigma^2_{U \mid x,a} = 1 - \theta^{\mathrm{T}}\Sigma_{A \mid X}^{-1}\theta$. Identifiability of $E\{Y(a)\}$ can then be obtained as in Theorem 1, except that now we replace (ii) of Theorem 1 with the following weaker condition:

(ii*) $\text{pr}(Y = 1 \mid A = a, X = x)$ depends on $a$ or $x$ or both. Furthermore, if $\text{pr}(Y = 1 \mid A = a, X = x)$ depends only on a subset of $x$, say

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\{x_{j1}, x_{j2}, \ldots, x_{jk}, 1 \leqslant j_1 < \cdots < j_k \leqslant q\}$, then at least one of $\{X_{j1}, X_{j2}, \ldots, X_{jk}\}$ has full support in $\mathbb{R}$.

**Theorem 2.**

*Suppose that* Assumption 1, (5)–(7), *and conditions* (i) *and* (ii) *of* Theorem 1 *hold. Then the parameters* $\theta, \Sigma_{AA}, \alpha, \beta, \gamma, \eta$ *and hence* $E\{Y(a)\}$ *are identifiable if and only if T is not deterministic or normally distributed.*

The proof of Theorem 2 is similar to that of Theorem 1 and hence omitted.

## 4. Discussion

When the causal effects are identifiable,one canuse the following likelihood-based procedure to estimate the model parameters. Asymptotic normality and the resulting inference procedures follow directly from standard M-estimation theory.

*Step* 1. Let$A^*$ be the residual of a linear regression of $A$ on $X$. Obtain the maximum likelihood estimators $\hat{\theta}$ and $\hat{\Sigma}_{A \mid X}$ based on a factor analysis on $A^*$, using an off-the-shelf package such as the factanal function in R (R Development Core Team, 2022).When there are no observed confounders $X$, one can use $A$ instead of $A^*$ and perform factor analysis.

*Step* 2. Estimate $(\alpha, \beta^{\mathrm{T}}, \gamma, \eta)$ by maximizing the conditional likelihood $\prod_{i=1}^{n} \left[\tilde{r}_i(\alpha, \beta, \gamma, \eta)^{Y_i} \{1 - \tilde{r}_i(\alpha, \beta, \gamma, \eta)\}^{1 - Y_i}\right]$, where $\tilde{r}_i(\alpha, \beta, \gamma, \eta) = \mathrm{pr}(Y = 1 \mid A = A_i, X = X_i; \alpha, \beta, \gamma, \eta, \hat{\theta}, \hat{\Sigma}_{A \mid X})$..

In the Supplementary Material, we report numerical results from analyses of synthetic data and real datasets. In a recent note, Grimmer et al. (2020) showed that the deconfounder algorithm of Wang & Blei (2019a) may not consistently outperform naive regression, ignoring the unmeasured confounder when the outcome and treatments follow Gaussian models. In constrast, our numerical results suggest that under our identification conditions, the likelihood-based estimates outperform naive regression estimates. Furthermore, these estimates exhibit some robustness against violations of the binary choice model specification. Nevertheless, we end with a cautionary remark that our results show that identification of causal effects in the multi-cause setting requires additional parametric structural assumptions, including the linear Gaussian treatment model, the binary choice outcome model, and a scalar confounder.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgement

The authors thank the editor, associate editor and referees for their helpful comments and suggestions.The authors also thank Jiaying Gu, Stanislav Volgushev and Ying Zhou for insightful discussions that have improved the main identification theorem. Kong and Wang were partially supported by the Natural Sciences and Engineering Research Council of Canada.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

# Appendix

## *Proof of* Theorem 1

We use the following notation. Let $A^{(-1)} = \left(A^{(k)}: k \neq 1\right) \in \mathbb{R}^{p-1}$ and define $a^{(-1)} \in \mathbb{R}^{p-1}$ and $c_2^{(-1)} \in \mathbb{R}^{p-1}$ analogously. Also write $A^{(-1,-j)} = \left(A^{(k)}: k \notin \left\{1, j\right\}\right) \in \mathbb{R}^{p-2}$.

We first establish the identifiability results for $\theta$ and $\Sigma_{AA}$. When $p \geqslant 3$, by condition (i) of Theorem 1 there exist at least three nonzero elements of $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$. By Anderson & Rubin (1956, Theorem 5.5) one can identify $\theta$ up to sign and uniquely identify $\sigma_A^2$. As $U$ is latent with a symmetric distribution around zero, without loss of generality we may assume we know $\gamma > 0$ so that the sign of $\theta_j$ in condition (i) is determined accordingly; otherwise, we may redefine $U$ to be its negative, and all the assumptions in Theorem 1 then hold if we also redefine $\theta_j$ and $\gamma$ to be their respective negatives. It follows that both $\theta$ and $\Sigma_{AA}$ are identifiable.

We now study the binary choice model (4). This is a nontraditional binary choice model as the right-hand side of the inequality involves a latent variable $Z$. We therefore let $T_1 = T - c_1 - c_3 Z$ so that $A \perp\!\!\!\perp T_1$, and model (4) becomes

$$Y = \mathbb{1}\left(T_1 \leqslant c_2^{\mathrm{T}} A\right). \tag{A1}$$

This is a binary choice model that was first introduced in economics (e.g., Cosslett, 1983; Gu & Koenker, 2020) and recently studied in statistics (e.g., Tchetgen Tchetgen et al., 2018). Condition (ii) of Theorem 1 implies that there exists $j$ such that $c_2^{(j)} \neq 0$. Without loss of generality we assume $c_2^{(1)} \neq 0$.

To identify the sign of $c_2^{(1)}$ and the distribution of $T_1/c_2^{(1)}$ observe that (A1) implies

$$\mathrm{pr}(Y = 1 \mid A = a) = \mathrm{pr}\left(T_1 \leqslant c_2^{\mathrm{T}} A \mid A = a\right) = \mathrm{pr}\left(T_1 \leqslant c_2^{\mathrm{T}} a\right), \tag{A2}$$

where the second equality holds because $A \perp\!\!\!\perp T_1$. Since $A$ follows a multivariate Gaussian distribution, (1) (A2) holds for any $a \in \mathbb{R}^p$. Setting $a^{(-1)} = 0$ in (A2), we can identify $\mathrm{pr}\left(T_1 \leqslant c_2^{(1)} a^{(1)}\right)$ for any $a^{(1)} \in \mathbb{R}$. Condition (ii) and (A2) guarantee that this is a monotone nonconstant function of $a$. It is easy to see that $c_2^{(1)} > 0$ if and only if $\mathrm{pr}\left(T_1 \leqslant c_2^{(1)} a^{(1)}\right)$ is an increasing function of $a^{(1)}$ so that the sign of $c_2^{(1)}$ is identifiable. Thus the distribution of $T_1/c_2^{(1)}$ is identifiable.

We now show that $c_2/c_2^{(1)}$ is identifiable. Without loss of generality we assume $c_2^{(1)} > 0$. If we let $T_2 = \left[T_1 - \left\{c_2^{(-1)}\right\}^{\mathrm{T}} A^{(-1)}\right]/c_2^{(1)}$, then (A2) implies that for any $a^{(-1)} \in \mathbb{R}^{p-1}$,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathrm{pr}(Y=1 \mid A=a)=\mathrm{pr}\left(T_2 \leqslant A^{(1)} \mid A=a\right)=\mathrm{pr}\left(T_2 \leqslant a^{(1)} \mid A^{(-1)}=a^{(-1)}\right) \quad \forall a^{(1)} \in \mathbb{R}.$$

Consequently, the distribution, and hence the expectation, of $T_2 \mid A^{(-1)}=a^{(-1)}$ is identifiable. It follows that for $j=2,\ldots,p$ we can also identify

$$c_2^{(j)}/c_2^{(1)}=E\left(T_2 \mid A^{(-1)}=0\right)-E\left(T_2 \mid A^{(-1,\,-j)}=0, A^{(j)}=1\right),$$

where the equality holds because $A \perp\!\!\!\perp T_1$.

We now turn to the third step of the proof. Lemma 1 implies that $c_2^{(1)}$, $c_1$ and $c_3^2$ are all identifiable if and only if $T$ is not deterministic or normally distributed. The sign of $c_3=\gamma\sigma_{U \mid A}$ can then be determined from the sign of $\gamma$, as $\sigma_{U \mid A} \geqslant 0$. Thus, the parameters $\theta, \Sigma_{AA}, \alpha, \beta, \gamma$ and hence $E\{Y(a)\}$ are identifiable if and only if $T$ is not deterministic or normally distributed, which finishes the proof.

## *Proof of* Lemma 1

Without loss of generality we assume $C=1$. Let $\tilde{T}_1=\tilde{T}-\widetilde{c_1}-\widetilde{c_3}\tilde{Z}$.

We first show that (II) implies (I). Suppose $T \sim N\left(\mu_T, \sigma_T^2\right)$, where $\sigma_T^2>0$ if $T$ is normally distributed and 0 if $T$ is deterministic. Then $T_1 \sim N\left(\mu_T-c_1, \sigma_T^2+c_3^2\right)$ and $\tilde{C}\tilde{T}_1 \sim N\left\{\tilde{C}(\mu_T-\tilde{c}_1), \tilde{C}^2\left(\sigma_T^2+\widetilde{c_3}^2\right)\right\}$. It is easy to verify that if $\tilde{C}=2, \widetilde{C_1}=(\mu_T+c_1)/2$ and $\widetilde{c_3}^2=c_3^2/4-3\sigma_T^2/4$, then $CT_1 \overset{\mathcal{D}}{=} \tilde{C}\tilde{T}_1$.

We next show that (I) implies (II). We start by showing that $\tilde{C} \neq 1$.
Suppose otherwise; then $T-c_1-c_3Z \overset{\mathcal{D}}{=} \tilde{T}-\widetilde{c_1}-\widetilde{c_3}\tilde{Z}$. We then have that for all $t \in \mathbb{R}, \phi_T(t)\phi_{c_1+c_3Z}(t)=\phi_T(t)\phi_{\widetilde{c_1}+\widetilde{c_3}Z}(t)$ and hence $\phi_{c_1+c_3Z}(t)=\phi_{\widetilde{c_1}+\widetilde{c_3}Z}(t)$, where $\phi_T(t)$ is the characteristic function of $T$. As a result, $c_1+c_3Z \overset{\mathcal{D}}{=} \widetilde{c_1}+\widetilde{c_3}Z$, which implies $(c_1, |c_3|)=(\widetilde{c_1}, |\widetilde{c_3}|)$ This is a contradiction.

We now let $c_1^*=\tilde{C}\widetilde{c_1}$ and $c_3^*=\tilde{C}\widetilde{c_3}$ so that $\tilde{C}\tilde{T}-c_1^*-c_3^*\tilde{Z} \overset{\mathcal{D}}{=} T-c_1-c_3Z$. We first consider the case where $|c_3^*|=|c_3|$. By a similar characteristic function argument to that above, $\tilde{C}T-c_1^* \overset{\mathcal{D}}{=} T-c_1$, so $T$ is a constant almost surely. We next consider the case where $|c_3^*| \neq |c_3|$. Without loss of generality we assume $|c_3^*|>|c_3|$. By a similar characteristic function argument to that above, we have

$$T \overset{\mathcal{D}}{=} \tilde{C}T+V, \tag{A3}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $V \perp\!\!\!\perp T$ and $V \sim N\left(\mu_V, \sigma_V^2\right)$ with $\mu_V = c_1 - c_1^*$ and $\sigma_V^2 = \left(c_3^*\right)^2 - c_3^2$. Equation (A3) implies that

$$\phi_T(t) = \phi_T(\tilde{C}t)\phi_V(t) = \phi_T\left(\tilde{C}^2 t\right)\phi_V(\tilde{C}t)\phi_V(t) = \cdots = \phi_T\left(\tilde{C}^K t\right)\prod_{k=1}^{K}\phi_V\left(\tilde{C}^{k-1}t\right) = \cdots. \tag{A4}$$

Consequently,

$$T \overset{\mathcal{D}}{=} \tilde{C}T + V_1 \overset{\mathcal{D}}{=} \tilde{C}\left(\tilde{C}T + V_2\right) + V_1 \overset{\mathcal{D}}{=} \cdots \overset{\mathcal{D}}{=} \tilde{C}^K T + \sum_{k=1}^{K}\tilde{C}^{k-1}V_k \overset{\mathcal{D}}{=} \cdots, \tag{A5}$$

where $V_k (k = 1, \ldots, K, \ldots)$ are independent and identically distributed and are independent of $T$. We will now show that $\tilde{C} < 1$. Suppose otherwise; then $\tilde{C} > 1$. Let $\| \cdot \|$ denote the modulus of a complex number. For any $t > 0$, by (A4) and the property of a normal distribution we have that $\|\phi_T(t)\| \leqslant \left\|\phi_V\left(\tilde{C}^{K-1}t\right)\right\| \to 0$ as $K \to \infty$. This is a contradiction, as by the continuity of the characteristic function we have $\lim_{t \to 0}\phi_T(t) = 1$.

We can now see that in (A5), as $K \to \infty$, $\tilde{C}^K T \to 0$ in probability and $\sum_{k=1}^{K}\tilde{C}^{k-1}V_k \to N\left\{(1-\tilde{C})^{-1}\mu_V, \left(1-\tilde{C}^2\right)^{-1}\sigma_V^2\right\}$ in distribution. Therefore, $T \sim N\left\{(1-\tilde{C})^{-1}\mu_V, \left(1-\tilde{C}^2\right)^{-1}\sigma_V^2\right\}$. Thus the proof is complete.

## References


Anderson TW & Rubin H. (1956). Statistical inference in factor analysis. In Proc. 3rd Berkeley Sympos. Mathematical Statistics and Probability, vol. 5. Berkeley, California: University of California Press, pp. 111–50.

Angrist JD, Imbens GW & Rubin DB (1996). Identification of causal effects using instrumental variables. J. Am. Statist. Assoc 91, 444–55.

Bentle PM. (1983). Simultaneous equation systems as moment structure models: With an introduction to latent variable models. J. Economet 22, 13–42.

Bollen KA (2014). Structural Equations with Latent Variables. NewYork: John Wiley & Sons.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB & Wynder EL (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. J. Nat. Cancer Inst 22, 173–203. [PubMed: 13621204]

Cosslett SR (1983). Distribution-free maximum likelihood estimator of the binary choice model. Econometrica 51, 765–82.

D'Amour A. (2019). On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. Proc. Mach. Learn. Res 89, 3478–86.

Ding P. (2014). Bayesian robust inference of sample selection using selection-*t* models. J. Mult. Anal 124, 451–64.

Drton M. & Maathuis MH (2017). Structure learning in graphical modeling. Annu. Rev. Statist. Appl 4, 365–93.

Grimmer J, Knox D. & Stewart BM (2020). Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. arXiv: 2007.12702.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Gu J. & Koenker R. (2020). Nonparametric maximum likelihood methods for binary response models with random coefficients. J. Am. Statist. Assoc to appear, DOI: 10.1080/01621459.2020.1802284.

Heckman JJ (1979). Sample selection bias as a specification error. Econometrica 47, 153–61.

Hernán MA & Robins JM (2006). Instruments for causal inference: An epidemiologist's dream? Epidemiology 17, 360–72. [PubMed: 16755261]

Imai K. & Jiang Z. (2019). Discussion of 'The blessings of multiple causes' by Wang and Blei. arXiv: 1910.06991.

Kuroki M. & Pearl J. (2014). Measurement bias and effect restoration in causal inference. Biometrika 101, 423–37.

Liu C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives. London: Wiley, pp. 227–38.

Miao W, Geng Z. & Tchetgen Tchetgen EJ (2018). Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika 105, 987–93. [PubMed: 33343006]

Miao W, Hu W, Ogburn E. & Zhou X. (2020). Identifying effects of multiple treatments in the presence of unmeasured confounding. arXiv: 2011.04504v3.

Peters J, Bühlmann P. & Meinshausen N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. J. R. Statist. Soc. B 78, 947–1012.

Peters J, Janzing D, Gretton A. & Schölkopf B. (2009). Detecting the direction of causal time series. In Proc. 26th Annu. Int. Conf. Machine Learning. NewYork: Association for Computing Machinery, pp. 801–8.

R Development Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Ranganath R. & Perotte A. (2019). Multiple causal inference with latent confounding. arXiv: 1805.08273v3.

Tchetgen Tchetgen EJ, Wang L. & Sun B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statist. Sinica 28, 2069–88.

Tran D. & Blei DM (2017). Implicit causal models for genome-wide association studies. arXiv: 1710.10742.

Veitch V, Wang Y. & Blei D. (2019). Using embeddings to correct for unobserved confounding in networks. In Advances in Neural Information Processing Systems 32 (NeurIPS 2019). La Jolla, California: Neural Information Processing Systems Foundation, pp. 13792–802.

Wang L. & Tchetgen Tchetgen E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. J. R. Statist. Soc. B 80, 531–50.

Wang Y. & Blei DM (2019a). The blessings of multiple causes. J. Am. Statist. Assoc 114, 1574–96.

Wang Y. & Blei DM (2019b). Multiple causes: A causal graphical view. arXiv: 1905.12793.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

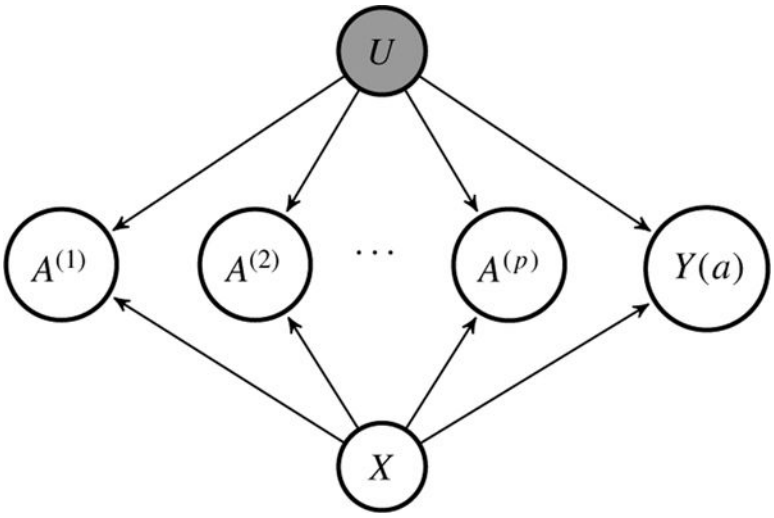


**Fig.1.**

A graphical illustration of the shared confounding setting. The latent ignorability assumption is encoded by the absence of arrows between $A^{(j)}$ and $Y(a)$ for $j = 1, \ldots, p$. The grey node indicates that $U$ is unobserved.