Biometrika (2022), **109**, 1, *pp.* 265–272
Printed in Great Britain

doi: 10.1093/biomet/asab016
Advance Access publication 15 May 2021

Identifiability of causal effects with multiple causes and a binary outcome

BY DEHAN KONG

Department of Statistical Sciences, University of Toronto,
700 University Avenue, Toronto, Ontario M5G 1X6, Canada
kongdehan@utstat.toronto.edu

SHU YANG

Department of Statistics, North Carolina State University,
2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.
syang24@ncsu.edu

AND LINBO WANG

Department of Statistical Sciences, University of Toronto,
700 University Avenue, Toronto, Ontario M5G 1X6, Canada
linbo.wang@utoronto.ca

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.

Some key words: 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).

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,

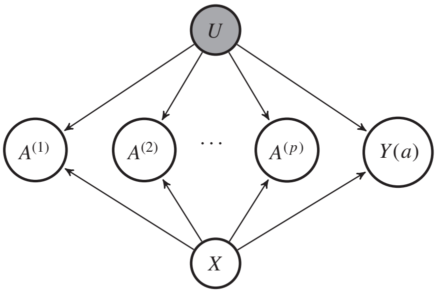


Fig. 1. A graphical illustration of the shared confounding setting. The latent ignorability assumption is encoded by the absence of arrows between $A^{(j)}$ and $Y(a)$ for $j = 1, \ldots, p$. The grey node indicates that U is unobserved.

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 $A = (A^{(1)}, A^{(2)}, \ldots, A^{(p)})^{\mathrm{T}}$ 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, $Y(a)$ is the potential outcome had the patient received treatment $a = (a^{(1)}, \ldots, a^{(p)})^{\mathrm{T}}$. 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.

Assumption 1 (Latent ignorablity). For all a, $A \perp\!\!\!\perp Y(a) \mid (X, U)$.

Under Assumption 1, we have

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

We consider a latent factor model for the treatments:

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

where $\epsilon_A \sim N\{0, \mathrm{diag}(\sigma^2_{A,1}, \ldots, \sigma^2_{A,p})\}$ 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 causal contrasts. However, as pointed out by D'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:

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

where an auxiliary latent variable T, independent of (A, U), has a known cumulative distribution function G. Equivalently, model (3) can be written as $\mathrm{pr}(Y = 1 \mid A, U) = G(\alpha + \beta^{\mathrm{T}} A + \gamma U)$. 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) $\mathrm{pr}(Y = 1 \mid A = a)$ *is not a constant function of a.*

Then the parameters θ, Σ_{AA}, α, β, γ 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 subject-specific knowledge allows the signs of θ_j and γ 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), $(U, A^{\mathrm{T}})^{\mathrm{T}}$ follows a joint multivariate normal distribution

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

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

The starting point of our identification approach is the following orthogonalization of $(U, A^{\mathrm{T}})^{\mathrm{T}}$. Let $Z = (U - \mu_{U|A})/\sigma_{U|A}$ 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

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

where $c_1 = \alpha$, $c_2 = (c_2^{(1)}, \ldots, c_2^{(p)})^{\mathrm{T}} = \beta + \gamma\theta^{\mathrm{T}}\Sigma_{AA}^{-1}$, $c_3 = \gamma\sigma_{U|A}$ 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 θ and Σ_{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 α, β, γ and hence $E\{Y(a)\}$ from c_1, c_2, c_3, θ and Σ_{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.*

(I) *There exist $(\tilde{C}, \tilde{c}_1, |\tilde{c}_3|) \neq (C, c_1, |c_3|)$, $\tilde{T} \overset{\mathcal{D}}{=} T$ 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, θ is identified up to its sign from the factor model, and it may be possible to identify the sign of θ from subject-matter knowledge. However, if U is a multi-dimensional vector, then the factor model (2) becomes $A = \Theta U + \epsilon_A$, where Θ is the loading matrix. In this case, Θ 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

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

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

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

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

where $\Sigma_{A|X} = \Sigma_{AA} - B\Sigma_{XX}B^{\mathrm{T}}$ with Σ_{AA} and Σ_{XX} being the covariances of A and X, respectively. Then $U \mid X = x, A = a$ follows a univariate normal distribution with mean $\mu_{U|x,a} = \theta^{\mathrm{T}}\Sigma_{A|X}^{-1}(a - Bx)$ and variance $\sigma_{U|x,a}^2 = 1 - \theta^{\mathrm{T}}\Sigma_{A|X}^{-1}\theta$. 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*) $\mathrm{pr}(Y = 1 \mid A = a, X = x)$ depends on a or x or both. Furthermore, if $\mathrm{pr}(Y = 1 \mid A = a, X = x)$ depends only on a subset of x, say $\{x_{j_1}, x_{j_2}, \ldots, x_{j_k}, 1 \leqslant j_1 < \cdots < j_k \leqslant q\}$, then at least one of $\{X_{j_1}, X_{j_2}, \ldots, X_{j_k}\}$ 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 θ, Σ_{AA}, α, β, γ, η 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 can use 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 $\hat{\Sigma}_{A|X}$ 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 $\prod_{i=1}^{n}[\tilde{r}_i(\alpha,\beta,\gamma,\eta)^{Y_i}\{1 - \tilde{r}_i(\alpha,\beta,\gamma,\eta)\}^{1-Y_i}]$, where $\tilde{r}_i(\alpha,\beta,\gamma,\eta) = \mathrm{pr}(Y = 1 \mid A = A_i, X = X_i; \alpha,\beta,\gamma,\eta,\hat{\theta},\hat{\Sigma}_{A|X})$.

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.

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.

Supplementary material

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

Appendix

Proof of Theorem 1

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

We first establish the identifiability results for θ and Σ_{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 θ up to sign and uniquely identify σ_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 θ_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 θ_j and γ to be their respective negatives. It follows that both θ and Σ_{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

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

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

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

where the second equality holds because $A \perp\!\!\!\perp T_1$. Since A follows a multivariate Gaussian distribution, (A2) holds for any $a \in \mathbb{R}^p$. Setting $a^{(-1)} = 0$ in (A2), we can identify $\mathrm{pr}(T_1 \leqslant c_2^{(1)} a^{(1)})$ for any $a^{(1)} \in \mathbb{R}$. Condition (ii) and (A2) guarantee that this is a monotone nonconstant function of $a^{(1)}$. It is easy to see that $c_2^{(1)} > 0$ if and only if $\mathrm{pr}(T_1 \leqslant c_2^{(1)} a^{(1)})$ 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 $T_2 = [T_1 - \{c_2^{(-1)}\}^{\mathrm{T}} A^{(-1)}]/c_2^{(1)}$, then (A2) implies that for any $a^{(-1)} \in \mathbb{R}^{p-1}$,

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

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

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

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 $c_3 = \gamma \sigma_{U|A}$ can then be determined from the sign of γ, as $\sigma_{U|A} \geqslant 0$. Thus, the parameters θ, Σ_{AA}, α, β, γ 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 $\tilde{T}_1 = \tilde{T} - \tilde{c}_1 - \tilde{c}_3 \tilde{Z}$.

We first show that (II) implies (I). Suppose $T \sim N(\mu_T, \sigma_T^2)$, where $\sigma_T^2 > 0$ if T is normally distributed and $\sigma_T^2 = 0$ if T is deterministic. Then $T_1 \sim N(\mu_T - c_1,\, \sigma_T^2 + c_3^2)$ and $\tilde{C}\tilde{T}_1 \sim N\{\tilde{C}(\mu_T - \tilde{c}_1),\, \tilde{C}^2(\sigma_T^2 + \tilde{c}_3{}^2)\}$. It is easy to verify that if $\tilde{C} = 2$, $\tilde{c}_1 = (\mu_T + c_1)/2$ and $\tilde{c}_3{}^2 = c_3^2/4 - 3\sigma_T^2/4$, then $C T_1 \overset{\mathcal{D}}{=} \tilde{C}\tilde{T}_1$.

We next show that (I) implies (II). We start by showing that $\tilde{C} \neq 1$. Suppose otherwise; then $T - c_1 - c_3 Z \overset{\mathcal{D}}{=} \tilde{T} - \tilde{c}_1 - \tilde{c}_3 \tilde{Z}$. We then have that for all $t \in \mathbb{R}$, $\phi_T(t)\phi_{c_1 + c_3 Z}(t) = \phi_T(t)\phi_{\tilde{c}_1 + \tilde{c}_3 Z}(t)$ and hence $\phi_{c_1 + c_3 Z}(t) = \phi_{\tilde{c}_1 + \tilde{c}_3 Z}(t)$, where $\phi_T(t)$ is the characteristic function of T. As a result, $c_1 + c_3 Z \overset{\mathcal{D}}{=} \tilde{c}_1 + \tilde{c}_3 Z$, which implies $(c_1, |c_3|) = (\tilde{c}_1, |\tilde{c}_3|)$. This is a contradiction.

We now let $c_1^* = \tilde{C}\tilde{c}_1$ and $c_3^* = \tilde{C}\tilde{c}_3$ so that $\tilde{C}\tilde{T} - c_1^* - c_3^* \tilde{Z} \overset{\mathcal{D}}{=} T - c_1 - c_3 Z$. 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

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

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

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

Consequently,

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

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 $\|\phi_T(t)\| \leqslant \|\phi_V(\tilde{C}^{K-1}t)\| \to 0$ 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 $\sum_{k=1}^{K}\tilde{C}^{k-1}V_k \to N\{(1 - \tilde{C})^{-1}\mu_V, (1 - \tilde{C}^2)^{-1}\sigma_V^2\}$ in distribution. Therefore, $T \sim N\{(1 - \tilde{C})^{-1}\mu_V, (1 - \tilde{C}^2)^{-1}\sigma_V^2\}$. Thus the proof is complete.

References

Anderson, T. W. & 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, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–55.

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

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.

Cosslett, S. R. (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, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Statist. Appl.* **4**, 365–93.

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

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

Hernán, M. A. & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17**, 360–72.

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, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105**, 987–93.

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

Tran, D. & Blei, D. M. (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, D. M. (2019a). The blessings of multiple causes. *J. Am. Statist. Assoc.* **114**, 1574–96.

Wang, Y. & Blei, D. M. (2019b). Multiple causes: A causal graphical view. *arXiv:* 1905.12793.

[*Received on* 31 *July* 2019. *Editorial decision on* 19 *January* 2021]

Biometrika (2018), **103**, 1, *pp.* 1–9
Printed in Great Britain

Advance Access publication on 31 July 2018

Supplementary Material for "Identifiability of causal effects with multiple causes and a binary outcome"

BY DEHAN KONG

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5G 1X6, Canada
kongdehan@utstat.toronto.edu

SHU YANG

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
syang24@ncsu.edu

AND LINBO WANG

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5G 1X6, Canada
linbo.wang@utoronto.ca

FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE [*]

SUMMARY

The supplementary material is organized as follows. In § S1, we provide a counterexample when the identification is not possible for multi-cause causal inference with a continuous outcome. § S2 reports simulation results. For illustration, we include a data application using data from Alzheimer's Disease Neuroimaging Initiative in § S3. Finally, in § S4, we perform sensitivity analysis of the proposed method using the data from the National Health and Nutrition Examination Survey.

S1. A COUNTEREXAMPLE

*Example S*1. Assume model (2) and the following model

$$Y(a) = \beta^{\mathrm{T}} a + \gamma U + \epsilon_Y, \tag{S1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ and $\epsilon_Y \perp\!\!\!\perp (A, U)$. Under models (2) and (S1), $E\{Y(a)\} = E\{E(Y|A = a, U)\} = \beta^{\mathrm{T}} a$; however, β is not identifiable.

To see this, let $\Sigma_{AA} \in \mathbb{R}^{p \times p}$ be the covariance of A, $\Sigma_{AY} \in \mathbb{R}^p$ the covariance between A and Y, and $\Sigma_{YY} \in \mathbb{R}$ the variance of Y. By linking the population covariance matrices of the

[*] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

observed variables and model parameters, we construct the following equations:

$$\Sigma_{AA} = \theta\theta^{\mathrm{T}} + \mathrm{diag}(\sigma_{A,1}^2, \ldots, \sigma_{A,p}^2), \tag{S2}$$

$$\Sigma_{AY} = \Sigma_{AA}\beta + \gamma\theta, \tag{S3}$$

$$\Sigma_{YY} = (\beta^{\mathrm{T}}\theta + \gamma)^2 + \beta^{\mathrm{T}}\mathrm{diag}(\sigma_{A,1}^2, \ldots, \sigma_{A,p}^2)\beta + \sigma_Y^2. \tag{S4}$$

In this setting, $(A^{\mathrm{T}}, Y)^{\mathrm{T}}$ follows a multivariate normal distribution, for which the first and second moments are sufficient statistics. Because the first moments are all zero, they do not provide information for identification of the model parameters. Thus, (S2), (S3) and (S4), from the second moment conditions, are the full set of equations for identifying the model parameters.

We now show that β is not identifiable through equations (S2), (S3) and (S4). First, equation (S2) can be used to identify θ and $\sigma_A = (\sigma_{A,1}^2, \ldots, \sigma_{A,p}^2)^{\mathrm{T}}$. In particular, when $p \geq 3$, if there exist at least three non-zero elements of $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, by Theorem 5.5 of Anderson and Rubin (1956), one can identify θ up to a sign flip and uniquely identify σ_A^2 through equation (S2). Second, we show that even if θ is identifiable, i.e., the sign can be determined, β is still not identifiable. Since equation (S2) only involves θ and σ_A, to identify $(\beta^{\mathrm{T}}, \gamma, \sigma_Y^2)^{\mathrm{T}} \in \mathbb{R}^{p+2}$, one needs to use equations (S3) and (S4). However, (S3) gives p equations and (S4) gives 1 equation, resulting in $p + 1$ equations in total. Consequently, without additional assumptions, we cannot identify the $p + 2$ dimensional parameters $(\beta^{\mathrm{T}}, \gamma, \sigma_Y^2)^{\mathrm{T}}$ from the $p + 1$ equations. Thus the causal effects cannot be identified.

S2. SIMULATIONS

We now evaluate the finite sample performance of the proposed estimators via simulation. In our simulations, we first generate a latent confounder U, an observed common confounder X and an additional observed covariate X^* from independent standard normal distributions. The treatments and outcome are then generated from the following linear and logistic structural equation models:

$$A = \theta U + \beta X + \beta^* X^* + \epsilon_A,$$

$$\mathrm{pr}(Y = 1 | A, U, X, X^*) = \mathrm{expit}(\alpha + \beta^{\mathrm{T}} A + \gamma U + \eta X + \eta^* X^*), \tag{S5}$$

where $\theta = (1, -1, 0.5)^{\mathrm{T}}$, $\alpha = 0$, $\beta = (1, 1, 1)^{\mathrm{T}}$, $\gamma = 1$, $\epsilon_A \sim \mathcal{N}_3(0, 0.25 I_3)$ and I_3 denotes the 3×3 identify matrix. We consider three settings. In setting 1, $\beta = \beta^* = (0, 0, 0)^{\mathrm{T}}$, $\eta = \eta^* = 0$ so that there are no observed confounders. In setting 2, $\beta = (1, -1, 1)^{\mathrm{T}}, \beta^* = (0, 0, 0)^{\mathrm{T}}, \eta = 1, \eta^* = 0$ so that there is a common confounder X. In setting 3, $\beta = (1, -1, 1)^{\mathrm{T}}, \beta^* = (1, 0, 0)^{\mathrm{T}}, \eta = \eta^* = 1$ so that there is a common confounder X and a so-called single-cause confounder X^*. When applying the proposed method, we assume that T follows a logistic distribution with mean zero and scale one, and the sign of $\theta_1\gamma$ is known. We compare the proposed method to a naive method where we only adjust for the observed confounders.

Tables S1 summarizes the simulation results. From Table S1, the estimates obtained from the naive method is subject to unmeasured confounding bias. In contrast, the bias of our proposed estimator is small for all model parameters and mean potential outcomes even with sample size 200, which further reduces as the sample size grows.

To assess the sensitivity of the proposed estimator to model misspecification, in setting 1 for $n = 200$, we fit the robit regression model (Liu; 2004; Ding; 2014)

$$\mathrm{pr}\{Y(a) = 1 \mid U\} = \mathrm{pr}\{Y = 1 \mid A = a, U\} = F_\nu(\alpha + \beta^{\mathrm{T}} a + \gamma U), \tag{S6}$$

Table S1. *Simulation results based on 1000 Monte Carlo repetitions: bias (standard deviation) of $\widehat{\beta}^{(1)}$, $\widehat{\beta}^{(2)}$, $\widehat{\beta}^{(3)}$ and the mean potential outcomes evaluated at $a_{(1)} = (1,1,0)^{\mathrm{T}}$ and $a_{(2)} = (0,0,1)^{\mathrm{T}}$, are reported. In the naive method we fit the model ignoring the unobserved confounder U*

Setting	Method	$\widehat{\beta}^{(1)}$	$\widehat{\beta}^{(2)}$	$\widehat{\beta}^{(3)}$	$\widehat{E}\{Y(a_{(1)})\}$	$\widehat{E}\{Y(a_{(2)})\}$
Sample size = 200						
1	Proposed	0.024(0.33)	0.019(0.30)	0.023(0.38)	−0.006(0.07)	0.001(0.07)
	Naive	0.435(0.32)	−0.381(0.29)	0.214(0.37)	0.029(0.06)	0.066(0.07)
2	Proposed	0.009(0.38)	0.015(0.37)	0.037(0.44)	−0.013(0.08)	−0.002(0.08)
	Naive	0.430(0.36)	−0.403(0.33)	0.238(0.45)	0.044(0.07)	0.081(0.08)
3	Proposed	0.065(0.42)	0.034(0.42)	0.066(0.48)	−0.009(0.09)	−0.001(0.08)
	Naive	0.468(0.43)	−0.377(0.38)	0.26(0.49)	0.049(0.09)	0.078(0.09)
Sample size = 500						
1	Proposed	−0.007(0.20)	−0.004(0.19)	−0.002(0.24)	−0.005(0.05)	−0.001(0.05)
	Naive	0.389(0.20)	−0.408(0.18)	0.193(0.24)	0.028(0.04)	0.067(0.05)
2	Proposed	0.002(0.23)	0.001(0.22)	−0.013(0.25)	−0.005(0.05)	−0.003(0.05)
	Naive	0.404(0.22)	−0.403(0.20)	0.182(0.25)	0.056(0.05)	0.084(0.05)
3	Proposed	0.021(0.25)	0.011(0.25)	0.029(0.29)	−0.006(0.06)	−0.002(0.05)
	Naive	0.426(0.26)	−0.389(0.23)	0.224(0.29)	0.061(0.06)	0.084(0.05)
Sample size = 1000						
1	Proposed	−0.010(0.14)	−0.001(0.13)	−0.016(0.16)	−0.003(0.03)	−0.003(0.03)
	Naive	0.386(0.14)	−0.403(0.12)	0.180(0.16)	0.031(0.03)	0.066(0.03)
2	Proposed	−0.009(0.17)	0.006(0.16)	0.000(0.17)	−0.004(0.04)	−0.001(0.03)
	Naive	0.392(0.16)	−0.399(0.14)	0.197(0.18)	0.058(0.03)	0.088(0.04)
3	Proposed	−0.003(0.17)	0.008(0.17)	0.003(0.19)	−0.005(0.04)	−0.003(0.03)
	Naive	0.397(0.17)	−0.400(0.15)	0.202(0.19)	0.064(0.04)	0.086(0.04)

where $F_\nu(\cdot)$ denotes the cumulative distribution function of the central-t random variable with scale one and degrees of freedom ν. We vary ν in the range $\{3,7,20\}$. Under model misspecification, the parameter β is no longer well-defined. So we only report $\widehat{E}\{Y(a_{(1)})\}$ and $\widehat{E}\{Y(a_{(2)})\}$ in the sensitivity analysis. Results in Table S2 show that when $\nu = 7$, the performance of our estimator is close to that under the logistic regression model. This is because when $\nu = 7$, the robit regression model is close to the logistic regression model (Liu; 2004).

Table S2. *Sensitivity analysis results for setting 1 based on 1000 Monte Carlo repetitions with 200 samples each: bias (standard deviation) of the mean potential outcomes evaluated at* $a_{(1)} = (1, 1, 0)^{\mathrm{T}}$ *and* $a_{(2)} = (0, 0, 1)^{\mathrm{T}}$ *are reported. We apply our method with the correctly specified logistic model* (S5) *and misspecified robit models* (S6) *with* $\nu = 3, 7, 20$

	Logistic	$\nu = 3$	$\nu = 7$	$\nu = 20$
$\widehat{E}\{Y(a_{(1)})\}$	−0.006(0.07)	−0.015(0.07)	−0.007(0.07)	−0.006(0.07)
$\widehat{E}\{Y(a_{(2)})\}$	0.001(0.07)	−0.005(0.08)	−0.001(0.07)	−0.002(0.07)

S3. ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE DATA APPLICATION

S3.1. *Data Usage Acknowledgement*

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

S3.2. *Data Analysis*

For illustration, we apply the proposed estimator to data from the Alzheimer's Disease Neuroimaging Initiative, a large-scale observational study launched in 2003 through a $60 million, 5-year public-private partnership. The study recruited adults aged between 55 and 90 years old. The 800 participants include cognitively normal individuals, as well as subjects with mild cognitive impairments and early Alzheimer's disease.

In our analysis, the treatments $A^{(1)}, A^{(2)}$ and $A^{(3)}$ are defined as the relative volumes of three brain regions: the frontal, cingulate cortex and hippocampal regions; the relative volume is defined as the ratio between the volume of a specific brain region and the total volume of the brain. The outcome is an indicator that the Mini Mental State Examination score is smaller than 24, a measure of cognitive decline that has been commonly used in diagnosis of Alzheimer's disease (O'Bryant et al.; 2008). The observed confounders include age, gender and years of education. For illustrative purpose, we include 674 subjects with complete covariate information in our analysis. Among these subjects, the average age is 75.3 (SD = 6.8), the average years of education is 15.7 (SD = 2.9), and 40.9% of the subjects are females.

The proposed approach makes the following assumptions: (1) there are no other confounding factors beyond a latent scalar U representing progression of Alzheimer's and an observed X representing age, gender and education length; (2) conditional on age, gender and education length, the relative volumes of the three brain regions and the latent disease progression follow a multivariate normal distribution; (3) age, gender and education length contribute linearly to the expected relative volumes; (4) the relative volumes of the three brain regions, the latent disease progression, age, gender and education length all contribute linearly to the log-odds of having a low cognitive score; (5) disease progression contributes to both hippocampal atrophy and a lower Mini Mental State Examination score (Sabuncu et al.; 2011), so that $\gamma\theta_3 < 0$.

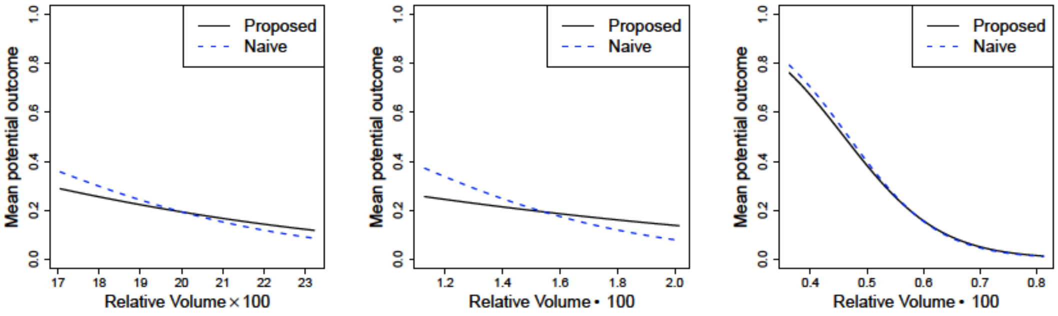
Analysis results show that conditional on age, gender, education length and the latent disease progression, each one percent of shrinkage in the relative volume of the frontal, cingulate cortex and hippocampal regions increases the odds of having a low cognitive score by 0.2% (95% CI: [-0.3%, 0.7%]), 0.9% (95% CI: [-5.1%, 7.0%]) and 13.2% (95% CI: [9.3%, 17.2%]), respectively. The directions of causal effect estimates are consistent with associations reported in the literature, suggesting that the bias from latent confounding is not large enough to distort the qualitative conclusions. Our results show that shrinkage of the hippocampal region has a stronger effect on the cognitive score, which aligns with the common belief that hippocampal atrophy is among the most significant structural biomarkers of Alzheimer's disease imaging (Henneman et al.; 2009).

We also compare the proposed and naive methods in terms of the mean potential outcomes. One can see from Figure S1 that compared to the proposed method, the naive method suggests a stronger association between the shrinkage of brain regions and odds of having a low cognitive score, due to strong confounding by the latent disease progress.

In addition, we have performed sensitivity analysis of our method by assuming the underlying models are robit regression models (S6) with degrees of freedom $\nu = 3, 7, 20$ respectively. Results in Figure S2 show that the estimates are fairly robust to model misspecifications.

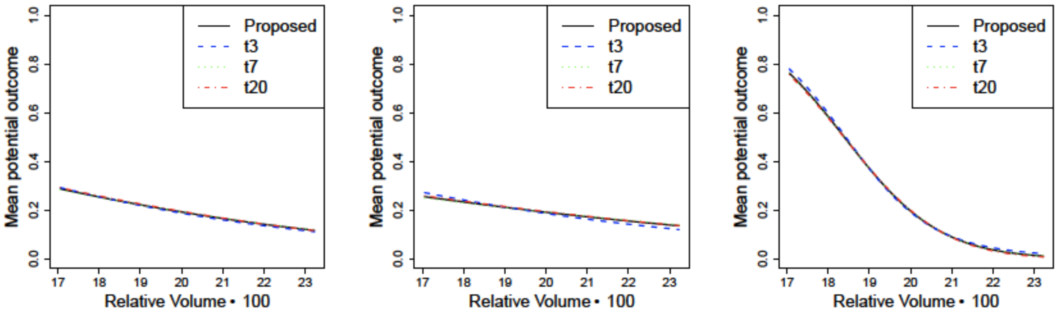
S4. Sensitivity Analysis based on the National Health and Nutrition Examination Survey

We now conduct a sensitivity analysis comparing the proposed method with an important confounder intentionally left out of the analysis, to a standard causal estimator adjusting for a full set of confounders. Our data come from the 2005-2006 cycle of the National Health and Nutrition Examination Survey, a program of studies designed to assess the health and nutritional status of



(a) The frontal region (b) The cingulate cortex region (c) The hippocampal region

Fig. S1. Probability of having a low Mini Mental State Examination score as a function of the relative volume of three brain regions (in percentages) estimated by the proposed and naive methods. In each plot, the relative volume of the other two regions are fixed at their sample medians.

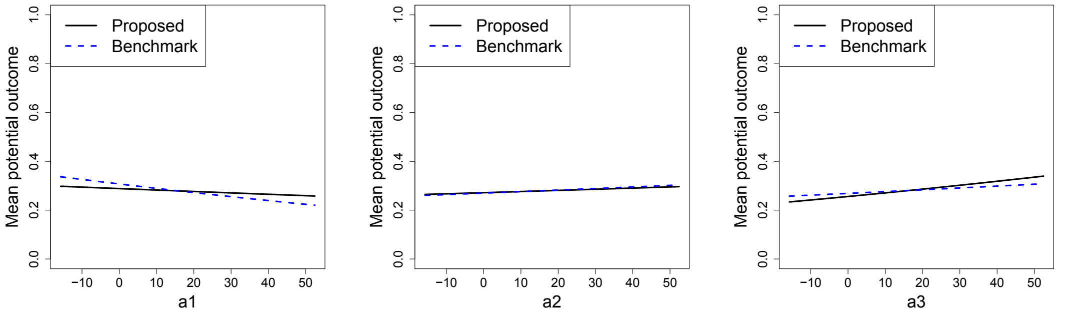


(a) The frontal region (b) The cingulate cortex region (c) The hippocampal region

Fig. S2. Probability of having a low Mini Mental State Examination score as a function of the relative volume of three brain regions (in percentages) estimated under a range of binary choice methods. In each plot, the relative volume of the other two regions are fixed at their sample medians.

adults and children in the U.S.. The outcome Y is an indicator that a person is overweight. The multiple treatments include intake of fiber (gm/day), intake of fat (gm/day) and intake of cholesterol (mg/day). We consider three confounders age, gender and family income. The family income is measured as the ratio of family income and poverty guidelines.

The data set has 5325 subjects with complete information on the multiple treatments, outcome, and three confounders. We shall use the proposed method to estimate the causal effects, treating gender and family income as the observed confounders X, while age as a latent confounder. We then compare our results to the benchmark method that estimates the causal effects based on models (6) and (7), using all three confounders. We assume the binary choice model has a logistic link in both the proposed and benchmark methods.



(a) The transformed intake of fiber (b) The transformed intake of fat (c) The transformed intake of cholesterol

Fig. S3. Estimated potential overweight proportion as a function of the transformed intakes of fiber, fat and cholesterol. In each plot, the transformed intakes of the other two types of nutrition are fixed at their sample medians.

Recall that our method requires the latent confounder to be independent of the observed confounders X, and that the three treatments and the unobserved confounder follow a multivariate normal distribution conditional on the observed confounders X. To make these assumptions plausible, we preprocess the data as follows. For age, we regress it on the observed confounders X, and obtain the residuals. We further perform rank-based inverse normal transform using `R` function `rankNorm`. We define our latent confounder U to be the transformed residuals, which preserve the sample mean and standard deviation as the residuals before the inverse normal transformation. We perform a similar regression of each of the three treatments on X, obtain their residuals and fitted values, and apply the inverse normal transformation to these residuals. The treatments $A^{(1)}, A^{(2)}, A^{(3)}$ are then defined as the sum of the transformed residuals and their respective fitted values. The sign of $\theta_1\gamma$ in our framework is determined by the sign of the corresponding estimate obtained by the benchmark method.

From the proposed method, $\hat{\beta}_1 = -0.0029$ (95% CI: [-0.0100, 0.0041]), $\hat{\beta}_2 = 0.0005$ (95% CI: [-0.0289, 0.0299]), and $\hat{\beta}_3 = 0.0003$ (95% CI: [0.0000, 0.0007]). In comparison, from the benchmark method, $\hat{\beta}^b_1 = -0.0088$ (95% CI: [-0.0161, -0.0016]), $\hat{\beta}^b_2 = 0.0007$ (95% CI: [-0.0012 , 0.0026]) and $\hat{\beta}^b_3 = 0.0002$ (95% CI: [-0.0002, 0.0005]). The causal effect estimates obtained from our method have the same directions and similar magnitudes as their corresponding estimates via the benchmark method.

We also compare the proposed and benchmark methods for estimating the mean potential outcomes. From Figure S3, our method yields similar estimates for the causal effect of the transformed intakes of fiber, fat and cholesterol, on the odds of overweight, compared to the benchmark method.

References

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis, *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. 5, pp. 111–150.

Ding, P. (2014). Bayesian robust inference of sample selection using selection-t models, *J. Multivar. Anal.* **124**: 451–464.

Henneman, W., Sluimer, J., Barnes, J., Van Der Flier, W., Sluimer, I., Fox, N., Scheltens, P., Vrenken, H. and Barkhof, F. (2009). Hippocampal atrophy rates in alzheimer disease: added value over whole brain volume measures, *Neurology* **72**(11): 999–1007.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression, *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives* pp. 227–238.

O'Bryant, S. E., Humphreys, J. D., Smith, G. E., Ivnik, R. J., Graff-Radford, N. R., Petersen, R. C. and Lucas, J. A. (2008). Detecting dementia with the mini-mental state examination in highly educated individuals, *Archives of Neurology* **65**(7): 963–967.

Sabuncu, M. R., Desikan, R. S., Sepulcre, J., Yeo, B. T. T., Liu, H., Schmansky, N. J., Reuter, M., Weiner, M. W., Buckner, R. L., Sperling, R. A. et al. (2011). The dynamics of cortical and hippocampal atrophy in alzheimer disease, *Archives of Neurology* **68**(8): 1040–1048.

[*Received on* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]