Appendix for "On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives"[1]

Alexander D'Amour
Google AI

A LARGE-m ASYMPTOTICS FOR LINEAR GAUSSIAN EXAMPLE

In the linear Gaussian example in Section 5, the ignorance region does not in general disappear in the large treatment number (large-m) limit. Here, we extend our example to an asymptotic frame where the ignorance region maintains the same (multiplicative) size even as m goes to infinity. Consider a sequence of problems where the number of treatments analyzed in each problem is increasing in the sequence. Each problem has its own data generating process, with some structural parameters indexed by m: $(\alpha_m, \beta_m, \gamma, \sigma_U^2, \sigma_A^2, \sigma_Y^2)$. We keep the scalar parameters not indexed by m fixed.

Importantly, we expect the marginal variance of Y to be relatively stable, no matter how many treatments we choose to analyze. Given our setup, this means that if the number of treatments is large, the effect of each individual treatment on average needs to become smaller as m grows large, or else the variance of Y would increase in m (this is clear from the specification of Σ_{YY}). To handle this, we fix some constant scalars a_0 and b_0 and assume that, for problem m,

$$\alpha_m = \mathbf{1}_{m\times 1} \cdot a_0/\sqrt{m}; \quad \beta_m = \mathbf{1}_{m\times 1} \cdot b_0/\sqrt{m}.$$

Thus, as $m \to \infty$, the norms of α_m and β_m, as well as their inner product $\alpha_m^\top \beta_m$, which appears in the expression for Σ_{YY}, remain fixed. [2]

Under this setup, the interval of valid values for the latent scaling factor c remains fixed for any value of m.For a fixed c in this interval, we examine how the corresponding shift vector $\Delta_{\beta,m}(c) = \beta_1(c) - \beta$ behaves as m grows large. The components of the shift $\Delta_{\beta,m}(c)$ scale as $m^{-1/2}$. Specifically, applying the Sherman-Morrison formula,

$$\begin{aligned}\Delta_{\beta,m}(c) &= \Sigma_{AA}^{-1}\alpha_m \cdot \gamma\sigma_U^2 \left(1 - \frac{1}{c}\right) \\ &= m^{-1/2} \cdot \mathbf{1}_{m\times 1} \cdot \frac{a_0}{\sigma_A^2 + \sigma_U^2 a_0^2} \cdot \gamma\sigma_U^2 \left(1 - \frac{1}{c}\right).\end{aligned}$$

Thus, for each k, the ratio of the kth component of the shift vector relative to the kth component of the true parameters remains fixed in m:

$$\frac{\Delta_{\beta,m}^{(k)}(c)}{\beta_m^{(k)}} = \frac{a_0}{b_0(\sigma_A^2 + \sigma_U^2 a_0^2)} \cdot \gamma\sigma_U^2 \left(1 - \frac{1}{c}\right).$$

Thus, even asymptotically as $m \to \infty$ there is no identification.

B PROOFS OF RESULTS

For convenience, we restate each result in addition to providing its proof.

B.1 Proof of Proposition 2

Proposition 2. *In the setting of Theorem 1, suppose that $P(U \mid A)$ is almost surely non-degenerate. Then, the following are true*

1. *The copula density $c(Y, U \mid A)$ is not identified.*
2. *Either $P(Y \mid do(A)) = P(Y \mid A)$, or $P(Y \mid do(A))$ is not identified.*

Proof. For the first statement, the joint distribution $P(U, Y, A)$ can be written

$$P(U, A, Y) = P(A)P(Y \mid A)P(U \mid A)c(Y, U \mid A),$$

By assumption, $P(Y, A)$ and $P(U, A)$ are identified, but the copula density $c(Y, U \mid A)$ remains unspecified because there are no restrictions on $P(Y \mid U, A)$.

For the second statement, note that the independence copula $c(Y, U \mid A) = 1$ is compatible with the observed data, as a result of the first statement. Under

[2] The asymptotic frame in this section is not the only way to maintain stable variance in Y as m increases.

the independence copula, $P(Y \mid do(A)) = P(Y \mid A)$. If this causal hypothesis is not true, then the true $P(Y \mid do(A))$ is also compatible with the observed data, so multiple causal hypotheses are compatible with the observed data, and $P(Y \mid do(A = a))$ is not identified. □

B.2 Proof of Proposition 3

Proposition 3. *Suppose that Assumption 1 holds, that $P(U)$ is not degenerate, and that there exists a consistent estimator $\hat{U}(A_m)$ of U as m grows large. Then positivity is violated as m grows large.*

Proof. Because $P(U)$ is non-degenerate, U takes on multiple values with positive probability. In this case, we establish that A_m concentrates in different sets of $\mathcal{A}_m$ depending on the value of U. For each m and each latent variable value u in the support of U, define the set

$$E_m(u) = \{a_m : \hat{U}(a_m) \neq u\}.$$

$E_m(u)$ is the set of cause vector values $a_m \in \mathcal{A}_m$ that $\hat{U}(\cdot)$ would map to a value *other* than u. Because $\hat{U}(A_m)$ is consistent, for each u in the support of U, as m grows large,

$$P(A_m \in E_m(u) \mid U = u) = P(\hat{U}(A_m) \neq u \mid U = u) \to 0.$$

Likewise, for any $u' \neq u$ in the support of U,

$$P(A \in E_m(u) \mid U = u') = P(\hat{U}(A_m) \neq u \mid U = u') \to 1.$$

Thus, positivity is violated. □

B.3 Proof of Theorem 1

Theorem 1. *Suppose that Assumption 1 holds, that $P(U, A)$ is identified, and that the model for $P(Y \mid U, A)$ is not subject to parametric restrictions.*

Then either $P(Y \mid do(A)) = P(Y \mid A)$ almost everywhere, or $P(Y \mid do(A))$ is not identified. ‘

Proof. One of two cases must hold: $P(U \mid A)$ is either degenerate almost everywhere, or not. In the non-degenerate case, Proposition 2 proves non-identification, except in the trivial case.

In the degnerate case, there are again two cases: either $P(U)$ is degenerate, or not. If $P(U)$ is not degenerate, Proposition 3 shows that the positivity assumption fails, and because $P(Y \mid U, A)$ is nonparametric by assumption, $P(Y \mid U, A)$ is inestimable for some $(u, a) \in \mathcal{U} \times \mathcal{A}$, and (1) is not identified. If $P(U)$ is degenerate, then the latent variable does not induce any confounding, and $P(Y \mid do(A)) = P(Y \mid A)$. □