Counterfactual Randomization: Rescuing Experimental Studies from Obscured Confounding

Appendix

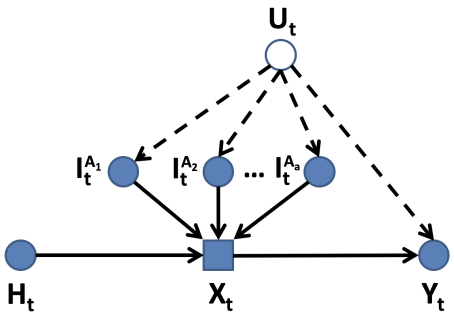

Figure 1: Graphical model of a prototypical HI-SCM M^A for a recommender agent viewing unit t, actor intents I_t^a, decision variable X_t, outcome Y_t, unobserved confounders U_t, and agent history H_t.

1 Proofs

To repeat the causal assumptions implicit in a confounded decision-making scenario with heterogeneous intents, we refer to the graphical model in Fig 1; the proofs that follow assume this structural causal model.

Theorem 4.1 (IEC-Specific Reward Superiority)**.** *Let X be a decision variable in a HI-SCM M^A with measured outcome Y, and let I^{ϕ_i} and I^{ϕ_j} be the heterogeneous intents of two distinct IECs ϕ_i, ϕ_j in the set of all IECs in the system, $\mathbf{\Phi}$. Maximized HI-specific rewards will always be at least as high as homogeneous, namely:*

$$\max_{x \in X} P(Y_x | I^{\phi_i}) \leq \max_{x \in X} P(Y_x | I^{\phi_i}, I^{\phi_j}) \; \forall \phi_i, \phi_j \in \Phi$$

Proof. Consider, without loss of generality, the case of binary treatment X (and therefore, binary intents). Let $x^* = \text{argmax}_{x \in X} P(Y_x | I^{\phi_i} = i^{\phi_i})$, we thus have:

$$P(Y_{x^*} | I^{\phi_i} = i^{\phi_i}) > P(Y_{x'} | I^{\phi_i} = i^{\phi_i}) \Rightarrow \tag{1}$$

$$\sum_{i^{\phi_j}} P(Y_{x^*} | I^{\phi_i} = i^{\phi_i}, I^{\phi_j} = i^{\phi_j}) P(I^{\phi_j} = i^{\phi_j} | I^{\phi_i} = i^{\phi_i})$$
$$> \sum_{i^{\phi_j}} P(Y_{x'} | I^{\phi_i} = i^{\phi_i}, I^{\phi_j} = i^{\phi_j}) P(I^{\phi_j} = i^{\phi_j} | I^{\phi_i} = i^{\phi_i}) \tag{2}$$

To simplify Eq. 2, we can write each summation as a sum of weighted HI-specific rewards across the possible values of I^{ϕ_j}, namely: let $p = P(I^{\phi_j} = i^{\phi_j} | I^{\phi_i} = i^{\phi_i})$. Re-writing Eq. 2, we have:

$$a(p) + b(p-1) > c(p) + d(p-1) \tag{3}$$

To exhaustively analyze the cases:

1. $p = 0 \Rightarrow b > d$, where, $b = P(Y_{x^*} | I^{\phi_i} = i^{\phi_i}, I^{\phi_j} = i^{\phi'_j}) = P(Y_{x^*} | I^{\phi_i} = i^{\phi_i})$
2. $p = 1 \Rightarrow a > c$, where, $a = P(Y_{x^*} | I^{\phi_i} = i^{\phi_i}, I^{\phi_j} = i^{\phi_j}) = P(Y_{x^*} | I^{\phi_i} = i^{\phi_i})$
3. $p \in (0,1) \Rightarrow max(a,b) \geq a(p) + b(p-1)$.

In all cases, the HI-specific rewards are greater than or equal to the homogeneous-intent-specific. □

Addendum: Note that in case (3), wherein $p \in (0,1)$, it is possible (for some choices of p) for Eq. 3 to hold but (when $I^{\phi_j} = i^{\phi_j}$) $a < c$ or (when $I^{\phi_j} = i^{\phi'_j}$) $b < d$, which is exploited by HI-RDC to improve treatment efficacy over RDC.

Theorem 4.2 (Empirical IEC Clustering Criteria)**.** *Let A_i, A_j be two actors modeled by a HI-SCM, and I^{A_i}, I^{A_j} their intents for some decision. Actors A_i, A_j are clustered into the same IEC, $\{A_i, A_j\} \in \phi_r$, whenever their intended actions over the same units correlate, as their intent-specific treatment outcomes will agree. Formally:*

$$\rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow \{A_i, A_j\} \in \phi_r \in \mathbf{\Phi}$$
$$\Rightarrow P(Y_x | I^{A_i}) = P(Y_x | I^{A_i}, I^{A_j})$$

Proof. We begin with some premises:

Premise 1

$$\rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow P(I^{A_i}) = P(I^{A_j}) = P(I^{A_i}, I^{A_j}) \tag{4}$$

The equivalence of each actors' intent priors $P(I^{A_i}) = P(I^{A_j})$ follows trivially from the fact that each actors responses are the same in each trial, and thus generate the same prior distribution over intents. However, to show that these quantities are the same as the joint requires a small insight in what values of i are realizable, notably, that:

$$\rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow P(I^{A_i} = i^i, I^{A_j} = i^j) = 0 \; \forall \; i^i \neq i^j \tag{5}$$

$$\begin{aligned} P(I^{A_i} = i^i, I^{A_j} = i^j) &= P(I^{A_i} = i^i) \\ &= P(I^{A_j} = i^j) \; \forall \; i^i = i^j \end{aligned} \tag{6}$$

Premise 2

$$\begin{aligned} \rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow P(I^{A_i}|U) &= P(I^{A_j}|U) \\ &= P(I^{A_i}, I^{A_j}|U) \end{aligned} \tag{7}$$

This result stems from the *canonical partitioning* of the latent space implicit within a SCM (Pearl 2000, Ch. 8). Fig. 2 demonstrates that, for binary treatment, the individual partitions r^{I^i}, r^{I^j} over each actors' intents (top model) can be equivalently represented as a single intent partitioning r^{I^*} whenever $\rho(I^{A_i}, I^{A_j}) = 1$. Put differently, for any instantiation of $U = u$, $I^{A_i} = I^{A_j} \Rightarrow r^{I^i} = r^{I^j} = r^{I^*}$, the latent space maps to the same values of each intent. As such, we can state that under any context $U, P(I^{A_i}|U) = P(I^{A_j}|U)$. To then demonstrate equivalence to $P(I^{A_i}, I^{A_j}|U)$, we recall that:

$$\begin{aligned} \rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow P(I^{A_i} = i^i, I^{A_j} = i^j|U) &= 0 \; \forall i^i \neq i^j \\ \Rightarrow P(I^{A_i} = i^i|U) &= P(I^{A_j} = i^j|U) \\ = P(I^{A_i} = i^i, I^{A_j} = i^j|U) & \end{aligned} \tag{8}$$

Premise 3

$$\begin{aligned} \rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow P(U|I^{A_i}) &= P(U|I^{A_j}) \\ &= P(U|I^{A_i}, I^{A_j}) \end{aligned} \tag{9}$$

Using Premises 1 and 2, we have:

$$\begin{aligned} P(I^{A_i}|U) = P(I^{A_j}|U) = P(I^{A_i}, I^{A_j}|U) = \\ \frac{P(U|I^{A_i})P(I^{A_i})}{P(U)} = \frac{P(U|I^{A_j})P(I^{A_j})}{P(U)} \\ = \frac{P(U|I^{A_i}, I^{A_j})P(I^{A_i}, I^{A_j})}{P(U)} = \\ P(U|I^{A_i}) = P(U|I^{A_j}) = P(U|I^{A_i}, I^{A_j}) \end{aligned} \tag{10}$$

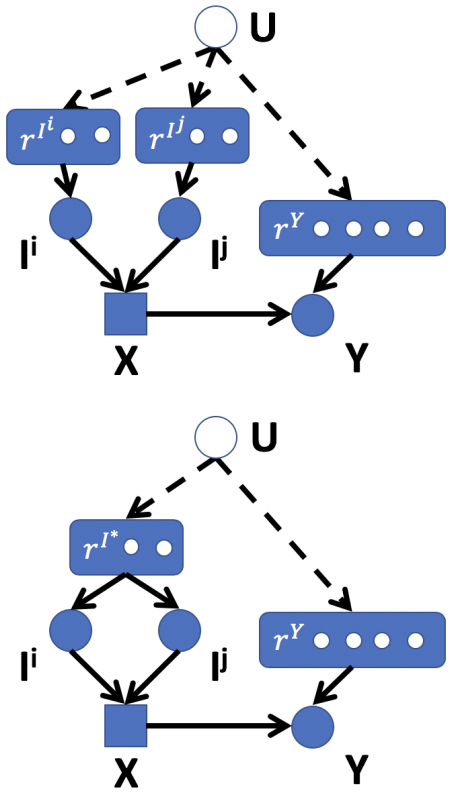

Figure 2: Canonical partitioning over latent space, U, such that equivalence classes for actors with $\rho(I^{A_i}, I^{A_j}) = 1$ are shown to be equivalently partitioned with separate r^{I^i}, r^{I^j} (top) as with a single r^{I^*}.

Finally, putting these premises together, we have:

$$\begin{aligned} P(Y_x|I^{A_i}) = P(Y_x|I^{A_j}) = P(Y_x|I^{A_i}, I^{A_j}) = \\ \sum_u P(Y_x|u, I^{A_i})P(u|I^{A_i}) = \sum_u P(Y_x|u, I^{A_j})P(u|I^{A_j}) \\ = \sum_u P(Y_x|u, I^{A_i}, I^{A_j})P(u|I^{A_i}, I^{A_j}) \\ = \sum_u P(Y_x|u)P(u|I^{A_i}) = \sum_u P(Y_x|u)P(u|I^{A_j}) \\ = \sum_u P(Y_x|u)P(u|I^{A_i}, I^{A_j}) \end{aligned} \tag{11}$$

Thus, by Premise 3, we see that these final 3 summations are equivalent, and can conclude that correlated intents yield no information about the UC state when concerted, yielding our empirical IEC measurement criteria because:

$$\rho(I^{A_i}, I^{A_j}) = 1 \Rightarrow P(Y_x|I^{A_i}) = P(Y_x|I^{A_i}, I^{A_j})$$

□

Theorem 4.3. ***(HI-RCT Confounding Criteria)*** *In a CDM scenario modeled by an HI-SCM M^A with treatment X, outcome Y, actor intended treatments I^{A_i}, and set of actor IECs $\mathbf{\Phi} = \{\phi_1, ..., \phi_m\}$, there exists some unobserved*[1] *U such that $X \leftarrow U \rightarrow Y$ whenever $\exists\ x \in X, i^{\Phi} \in I^{\Phi}$: $P(Y_x) \neq P(Y_x|i^{\mathbf{\Phi}})$.*

Proof. The proof follows trivially from the Counterfactual Interpretation of the Back-Door theorem (Pearl, Glymour, and Jewell 2016); viz., referrencing the SCM in Fig. 1, if there is *no* confounding path between $I^{\Phi} \leftarrow U \rightarrow Y$, then $Y_x \perp\!\!\!\perp I^{\Phi} \Rightarrow P(Y_x|i^{\mathbf{\Phi}}) = P(Y_x)$. □

2 Simulations

Information regarding simulation source and analysis will be made available here upon paper acceptance (inclusion would compromise author anonymity).

References

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. Second ed., 2009.

[1] Assuming that any observed confounders have been controlled for (see back-door criterion, (Pearl 2000, Ch. 3)).