Copula-based Sensitivity Analysis for Multi-Treatment Causal Inference with Unobserved Confounding

Jiajing Zheng Alexander D'Amour Alexander Franks
UCSB & Google Research

April 9, 2021

Presented by: Serge Assaad

Causal graphs & ignorability

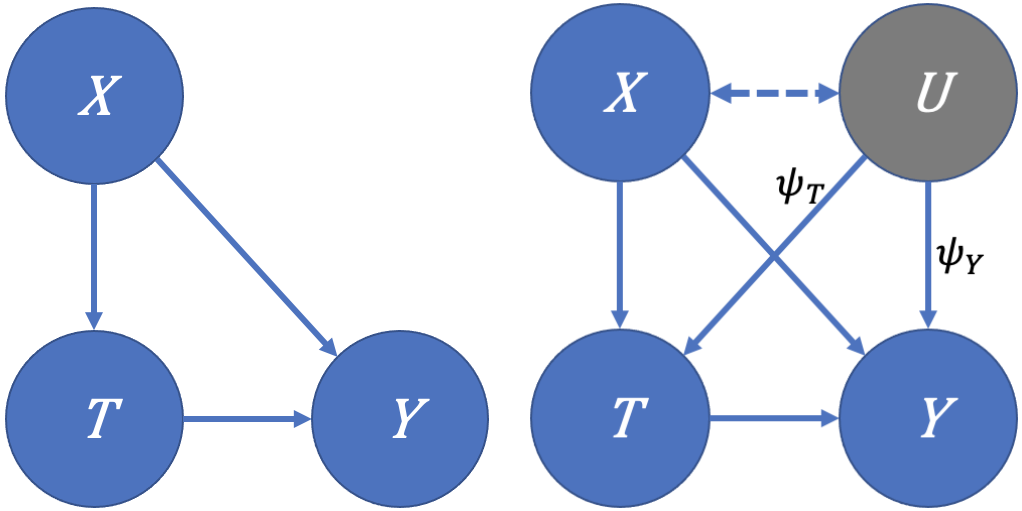


Figure: $U=$ unobserved confounders, $X=$ observed covariates, $T=$ treatment, $Y=$ outcome. (Left) Ignorable treatment assignment. (Right) Non-ignorable treatment assignment, with sensitivity parameters ψ_T, ψ_Y representing the assumed "strength" of $U \to T$ and $U \to Y$, respectively.

Ignorability, identification, sensitivity analysis

- Under ignorability (& other assumptions), we get **identification**, *i.e.*,

$$P(Y|X = x, do(T = t)) = P(Y|X = x, T = t),$$

  which allows us to estimate treatment effects (*e.g.*, $E(Y|X = x, do(T = 1)) - E(Y|X = x, do(T = 0))$).
- Ignorability is a strong (and untestable) assumption. To relax it, we assume U exists, and we can make assumptions about the strength of $U \to T$ (denoted ψ_T) and $U \to Y$ (denoted ψ_Y), and estimate the *range* of possible effects for different (reasonable) $\psi \triangleq \{\psi_T, \psi_Y\}$ pairs – this is known as **sensitivity analysis**.

Motivation

- "The blessings of multiple causes" (Wang and Blei, 2018) claimed to identify causal effects under unobserved confounding, owing to the presence of multiple treatments. The intuition is that, with multiple treatments, we can "reconstruct" the latent U (via, e.g., a matrix factorization method) and use the inferred U as a drop-in observed confounder we can control for.
- Many "counter-papers" (D'Amour, 2019a,b; Ogburn et al., 2019) show that identification is impossible in general, even in the presence of multiple causes.
- This paper is inspired from the above 2 points. We can't identify treatment effects with an unobserved U, but the intuition from Wang and Blei (2018) that multiple causes helps treatment effect estimation is correct – in this paper, the authors explore how the multi-cause setting is a fruitful one for sensitivity analysis.

Measuring the effect of a treatment

Authors use "marginal contrast estimands", namely:

$$MCE_{t_1,t_2} \triangleq \tau(E[v(y)|do(t_1)], E[v(y)|do(t_2)]) \quad (1)$$

for some functions v and τ.
Special cases include:

- the population average treatment effect

$$PATE_{t_1,t_2} \triangleq E(Y|do(t_1)) - E(Y|do(t_2))$$

  (set v =identity and $\tau(a,b) = a - b$).

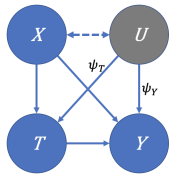
- Relative risk for binary outcomes

$$RR_{t_1,t_2} \triangleq P(Y = 1|do(t_1))/P(Y = 1|do(t_2))$$

  (set $v(y) = \mathbb{1}(y = 1)$ and $\tau(a,b) = a/b$)

We can also write all of the above conditioned on covariate value $X = x$.

Framework for sensitivity analysis



- We posit densities $f_{\psi_T}(u|t)$ and $f_{\psi_Y}(y|u,t)$ describing relationship between latent U and the observed T and Y, respectively
- We require that the marginals of these densities evaluate to the *observed* density as:

$$f(y|T=t) = \int_{\mathcal{U}} f_{\psi_Y}(y|u,t) f_{\psi_T}(u|t) du$$

Note that the LHS does not depend on the sensitivity parameters ψ (it shouldn't).

- We can write the *interventional distribution*

$$f_\psi(y|do(t)) = \int_{\mathcal{U}} f_{\psi_Y}(y|t,u) f(u) du = \int_{\mathcal{U}} f_{\psi_Y}(y|t,u) \left[\int_{\mathcal{T}} f_{\psi_T}(u|\tilde{t}) f(\tilde{t}) dt \right] du$$

Fundamental challenge in sensitivity analysis

- A fundamental challenge is to specify a class of densities $f_{\psi_Y}(y|t,u)$ and $f_{\psi_T}(u|t)$ for which the *interventional* distribution $f_{\psi}(y|do(t))$ changes with ψ, but the *observational* distribution $f(y|t)$ does not.
- A useful tool to do this is the *copula*, introduced next.

Background on copulas

- Suppose we have a set of n RVs $X_1, ..., X_n$, with respective c.d.f.'s $F_1, ..., F_n$ (and p.d.f.'s $f_1, ..., f_n$).
- It can be readily shown that the RVs $U_i \triangleq F_i(X_i) \sim \text{Uniform}(0,1), \forall i \in \{1, .., n\}$.
- A copula c.d.f. is the joint distribution $C(u_1, ..., u_2) \triangleq \Pr(U_1 \leq u_1, ..., U_n \leq u_n) = \Pr(F_1(X_1) \leq u_1, ..., F_n(X_n) \leq u_n)$. Note that $\text{supp}(C) = [0,1]^n$. We can also define a copula p.d.f. $c(u_1, ..., u_n) = \frac{\partial^n C(u_1,...,u_n)}{\partial u_1 ... \partial u_n}$.

Theorem (Sklar's Theorem)

The joint distribution $f(x_1, ..., x_n)$ can be written as:

$$f(x_1, ..., x_n) = f_1(x_1) \cdot ... \cdot f_n(x_n) \cdot c(F_1(x_1), ..., F_n(x_n)) \quad (2)$$

Intuitively, the copula contains all the information about the dependency between the RV's.

Using copulas for sensitivity analysis

A straightforward application of Sklar's theorem gives:

$$f_\psi(y|u,t) = f(y|t)c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u)|t) \tag{3}$$

Note the copula's dependence on ψ_Y since it encodes the dependency between U and Y.
We can plug (3) into the interventional distribution $f_\psi(y|do(t))$ to get:

$$f_\psi(y|do(t)) = f(y|t)\int c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u)|t)\left[\int_{\mathcal{T}} f_{\psi_T}(u|\tilde{t})f(\tilde{t})dt\right]du \tag{4}$$

this can be thought of as a "tilt" of the observational distribution $f(y|t)$ to get the interventional distribution $f_\psi(y|do(t))$, where the tilting function is $w_\psi(y,t) \triangleq \int c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u)|t)f_{\psi_T}(u)du$.

Computing marginal contrasts

We can use the previous expression for $f_\psi(y|do(t))$ to compute any marginal contrast via:

$$E[v(Y)|do(t)] = \int v(y) w_\psi(y,t) f(y|t) dy \tag{5}$$

where $w_\psi(y,t)$ is the tilt to get $f_\psi(y|do(t))$ from $f(y|t)$:

$$w_\psi(y,t) \triangleq \int c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u)|t) \left[\int f_{\psi_T}(u|\tilde{t}) f(\tilde{t}) d\tilde{t} \right] du \tag{6}$$

$$\approx \frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \left[\int c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u)|t) f_{\psi_T}(u|t_i) du \right] \tag{7}$$

where $\mathcal{T}$ is the set of all sampled treatment vectors.

Multiple treatments, causal equivalence class

- The previously described method is fully general, and allows us to compute marginal contrasts in many settings (single treatment, multi-treatment, multi-outcome...), but doesn't take advantage of the multi-treatment setting to make sensitivity analysis "easier"
- This paper focuses on inference with latent variable methods where the sensitivity parameter ψ_T is identifiable *up to a causal equivalence class*.

Definition (Causal equivalence class)

$[\psi_T]$ is a causal equivalence class of ψ_T if and only if for any $\tilde{\psi}_T$ in $[\psi_T]$, then, for every ψ_Y there exists a $\tilde{\psi}_Y$ such that $f_{\psi_Y,\psi_T}(y \mid do(T=t)) = f_{\tilde{\psi}_Y,\tilde{\psi}_T}(y \mid do(T=t))$ for all y, t.

Otherwise stated, $\psi_T^1 \sim \psi_T^2$ ($\sim$ here is an equivalence) if $\forall \psi_Y^1, \exists \psi_Y^2$ s.t. $f_{\psi_T^1,\psi_Y^1}(y|do(t)) = f_{\psi_T^2,\psi_Y^2}(y|do(t))$

Causal equivalence class

- Identification of ψ_T up to a causal equivalence class is generally not possible in the single-treatment setting
- Identification holds in the multi-treatment setting (under additional assumptions).
- If ψ_T is identified (up to a causal equivalence class), the only remaining degree of freedom is the copula c_{ψ_Y}.

Gaussian copula

Assumption (Copula invariance)

The conditional copula does not depend on the value of t, that is, the conditional dependence between Y and U is invariant to the level of T.

Assumption (Gaussian copula)

The conditional copula between the outcome and m-dimensional latent confounders given treatments, $c_\psi(F_{Y|t}(y), F_{U|t}(u) \mid t)$, is a Gaussian copula.

Generative model under a Gaussian copula

Under the previously stated assumptions, we have the following generative model:

$$T \sim F_T \tag{8}$$

$$f(u \mid t) \sim N(\mu_{u|t}, \Sigma_{u|t}) \tag{9}$$

$$\tilde{Y} = \gamma'(U - \mu_{u|t}) + \epsilon_{\tilde{y}|t,u}, \quad \epsilon_{\tilde{y}|t,u} \sim N(0, \sigma^2_{\tilde{y}|t,u}), \quad \gamma^T \Sigma_{u|t} \gamma + \sigma^2_{\tilde{y}|t,u} = 1 \tag{10}$$

$$Y = F^{-1}_{Y|t}(\Phi(\tilde{Y})) \tag{11}$$

Under this generative model, the sensitivity parameters are $\psi_T = \{\mu_{u|t}, \Sigma_{u|t}\}$ and $\psi_Y = \{\gamma\}$. In general, $\mu_{u|t}$ and $\Sigma_{u|t}$ will not be point identified, although under many latent variable models they can be identified up to invertible linear transformation of U.

Causal equivalence class for the Gaussian copula

The following theorem establishes that the class of ψ_T defined by all invertible linear transformations of U is a causal equivalance class.

Theorem

Assume model 9-11. Let $[\psi_T] = \{\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} : A \in \mathcal{S}^+\}$ *where* $\mathcal{S}^+$ *is the space of symmetric positive definite matrices. Then* $[\psi_T]$ *is a causal equivalence class.*

Summary of results for the Linear-Gaussian Model

Let's assume (Y, T, U) are jointly multivariate Gaussian. At a high level, we have the following results:

- For a single treatment, the confounding bias for PATE_{t_1,t_2} is unbounded.
- For multiple treatments which can be used to identify (up to a causal equivalence class) the conditional confounder distribution $f(u|t)$, **the magnitude of the confounding bias for PATE_{t_1,t_2} is bounded**.

Linear-Gaussian Model

Suppose we have the following generative model for U, T, Y:

$$U = \epsilon_u, \quad \epsilon_u \sim N_m(0, \Sigma_u), \tag{12}$$

$$T = BU + \epsilon_{t|u}, \quad \epsilon_{t|u} \sim N_k(0, \sigma^2_{t|u} I_k), \tag{13}$$

$$Y = \tau' T + \gamma' U + \epsilon_{y|t,u}, \quad \epsilon_{y|t,u} \sim N_m(0, \sigma^2_{y|t,u}), \tag{14}$$

Remarks:

- When either $B = 0$ or $\gamma = 0$, there is no confounding.
- Note that, under this model, we can write the density $f(u|t) \sim \mathcal{N}(\mu_{u|t}, \Sigma_{u|t})$, where $\mu_{u|t}$ and $\Sigma_{u|t}$ are known functions of B and $\sigma^2_{t|u}$

Confounding bias in the Linear-Gaussian model

We can easily show that the interventional distribution is:

$$f(y|do(T=t)) \sim \mathcal{N}(\tau' t, \sigma^2_{y|u,t} + \gamma' \Sigma_u \gamma) \tag{15}$$

We can also show that the observational distribution is:

$$f(y|T=t) \sim \mathcal{N}(\tau'_{\text{naive}} t, \sigma^2_{y|t}) \tag{16}$$

where

$$\tau_{\text{naive}} = \tau + (B\Sigma_u B' + \sigma^2_{t|u} I_k)^{-1} B\Sigma_u \gamma \tag{17}$$

$$\sigma^2_{y|t} = \sigma^2_{y|t,u} + \gamma'(\Sigma_u - \Sigma_u B'(B\Sigma_u B' + \sigma^2_{t|u} I_k)^{-1} B\Sigma_u)\gamma \tag{18}$$

$$= \sigma^2_{y|t,u} + \gamma' \Sigma_{u|t} \gamma \tag{19}$$

We can express the PATE as:

$$\text{PATE}_{\Delta t} = \tau'(t_1 - t_2) := \tau' \Delta t, \tag{20}$$

The confounding bias, denoted $\text{Bias}_{\Delta t} = \tau'_{\text{naive}} \Delta t - \text{PATE}_{\Delta t}$, can then be expressed as

$$\text{Bias}_{\Delta t} = \gamma' \Sigma_u B' (B\Sigma_u B' + \sigma^2_{t|u} I_k)^{-1} \Delta t = \gamma'(E(U \mid T = t_1) - E(U \mid T = t_2))$$

$$\triangleq \gamma' \mu_{u|\Delta t}, \tag{21}$$

Identification of ψ_T in Linear-Gaussian model

Theorem

Suppose that the observed data is generated by model (12)-(14). *When there* k *treatments with* $1 < m < k$, *then* ψ_T *is identified up to the causal equivalence class* $[\psi_T] = \{\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} : A \in \mathcal{S}^+\}$. *When there is a single treatment* ($k = 1$) *or at least* $m = k$ *confounders, then* ψ_T *is not identifiable up to causal equivalence class.*

Bounded confounding bias for the Linear-Gaussian model

As a reminder, in the Linear-Gaussian model we have $Y = \tau' T + \gamma' U + \epsilon_{y|t,u}$, which gives $\sigma^2_{y|t} = \sigma^2_{y|t,u} + \gamma' \Sigma_{u|t} \gamma$. The fraction of residual outcome variance explained by U is:

$$0 \leq R^2_{Y \sim U|T} = \frac{\gamma^T \Sigma_{u|t} \gamma}{\sigma^2_{y|t}} \leq 1 \tag{22}$$

Theorem

Suppose that the observed data is generated by model 12-14 with $\sigma^2_{t|u} > 0$. Then, $\forall \gamma$ satisfying Assumptions 1 and 2,

$$\gamma^T \Sigma_{u|t} \gamma \leq \sigma^2_{y|t} \tag{23}$$

For any given Δt, we have

$$Bias^2_{\Delta t} \leq \sigma^2_{y|t} R^2_{Y \sim U|T} \| \Sigma^{-1/2}_{u|t} \mu_{u|\Delta t} \|^2_2, \tag{24}$$

The bound is achieved when γ is colinear with $\Sigma^{-1}_{u|t} \mu_{u|\Delta t}$.

Bounded confounding bias for the Linear-Gaussian model

- The implication of the previous theorem is that the true treatment effect lies in the interval $\tau'_{naive}\Delta t \pm \sqrt{\sigma^2_{y|t} R^2_{Y\sim U|T}} \|\Sigma^{-1/2}_{u|t} \mu_{u|\Delta t}\|_2$.

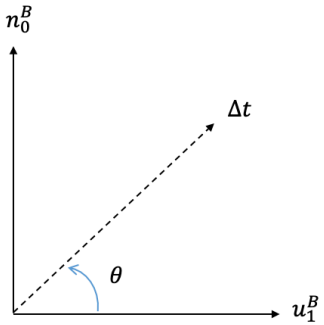
Corollary

Let d_1 be the largest singular value of B. For all Δt with $\| \Delta t \|_2 = 1$, the squared bias is bounded by

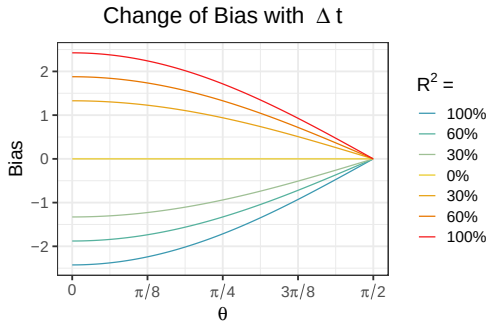
$$Bias^2_{\Delta t} \leq \frac{d_1^2}{(d_1^2 + \sigma^2_{t|u})} \frac{\sigma^2_{y|t}}{\sigma^2_{t|u}} R^2_{Y\sim U|T}, \tag{25}$$

with equality when $\Delta t = u_1^B$, the first left singular vector of B. When $\Delta t \in Null(B')$, the naive estimate is unbiased, that is, $PATE_{\Delta t} = \tau'_{naive}\Delta t$.

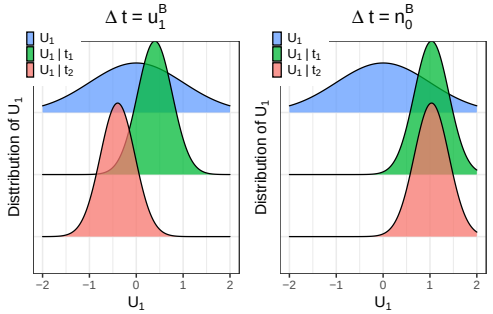
Illustration of corollary



(a)
$\Delta t := t_1 - t_2$



(b) Estimation Bias of $\text{PATE}_{\Delta t}$



(c) Distribution of U_1

Figure: (a) θ is the angle between Δt and the first left singular vector of B, denoted u_1^B. (b) Confounding bias of naive estimates of $\text{PATE}_{\Delta t}$ changes with θ and depends on $R^2_{Y \sim U|T}$. (c) Confounder densities in different populations. The blue, green, red densities denote distributions of U_1 in the observed population, the subpopulation receiving t_1 and the subpopulation receiving treatment t_2 respectively. Observed data estimates of $\text{PATE}_{\Delta t}$ are unbiased when $\Delta t = n_0^B$.

Generalizing the Linear-Gaussian case

Assuming the Gaussian copula parametrization discussed earlier (but not the Linear-Gaussian model), we have the following theorem:

Theorem

Assume the Gaussian copula model with Gaussian outcomes. If $\Sigma_{u|t}$ *is non-invertible, then* Bias_{t_1,t_2} *is bounded if and only if* $\mu_{u|t_1} - \mu_{u|t_2}$ *is in the row space of* $\Sigma_{u|t}$*. When bounded,* $\text{Bias}^2_{t_1,t_2} \leq \sigma^2_{y|t} R^2_{Y \sim U|T} \|(\Sigma^{\dagger}_{u|t})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2})\|_2^2$*, where* $\Sigma^{\dagger}_{u|t}$ *is the pseudo-inverse of* $\Sigma_{u|t}$*.*

Calibration

- This paper operates under the assumption that ψ_T is identified up to a causal equivalence class, so there is no need to calibrate $\psi_T = \{\mu_{u|t}, \Sigma_{u|t}\}$ to lie in a "reasonable range"
- However, we still need to calibrate $\psi_Y = \gamma$ to be in a reasonable range.
- To calibrate γ, the authors reparametrize γ in terms of a magnitude and a direction:

$$\gamma = \sqrt{R^2_{Y \sim U|T}} \Sigma_{u|t}^{-1/2} d, \tag{26}$$

where $d \in \mathbb{S}^{m-1}$ is an m-dimensional unit vector. They then calibrate the magnitude and direction separately.

Calibrating the magnitude and direction of γ

- Authors compute the fraction of residual outcome variance explained by a specific treatment (or set of treatments) T_j, conditioned on all other treatments T_{-j}, as:

$$R^2_{Y\sim T_j|T_{-j}} := \frac{R^2_{Y\sim T} - R^2_{Y\sim T_{-j}}}{1 - R^2_{Y\sim T_{-j}}}, \tag{27}$$

- To set the direction of γ, authors are conservative: $Bias_{t_1,t_2}$ is maximized when γ is colinear with $\Sigma^{-1/2}_{u|t}(\mu_{u|t_1} - \mu_{u|t_2})$, so we can set d in that direction.

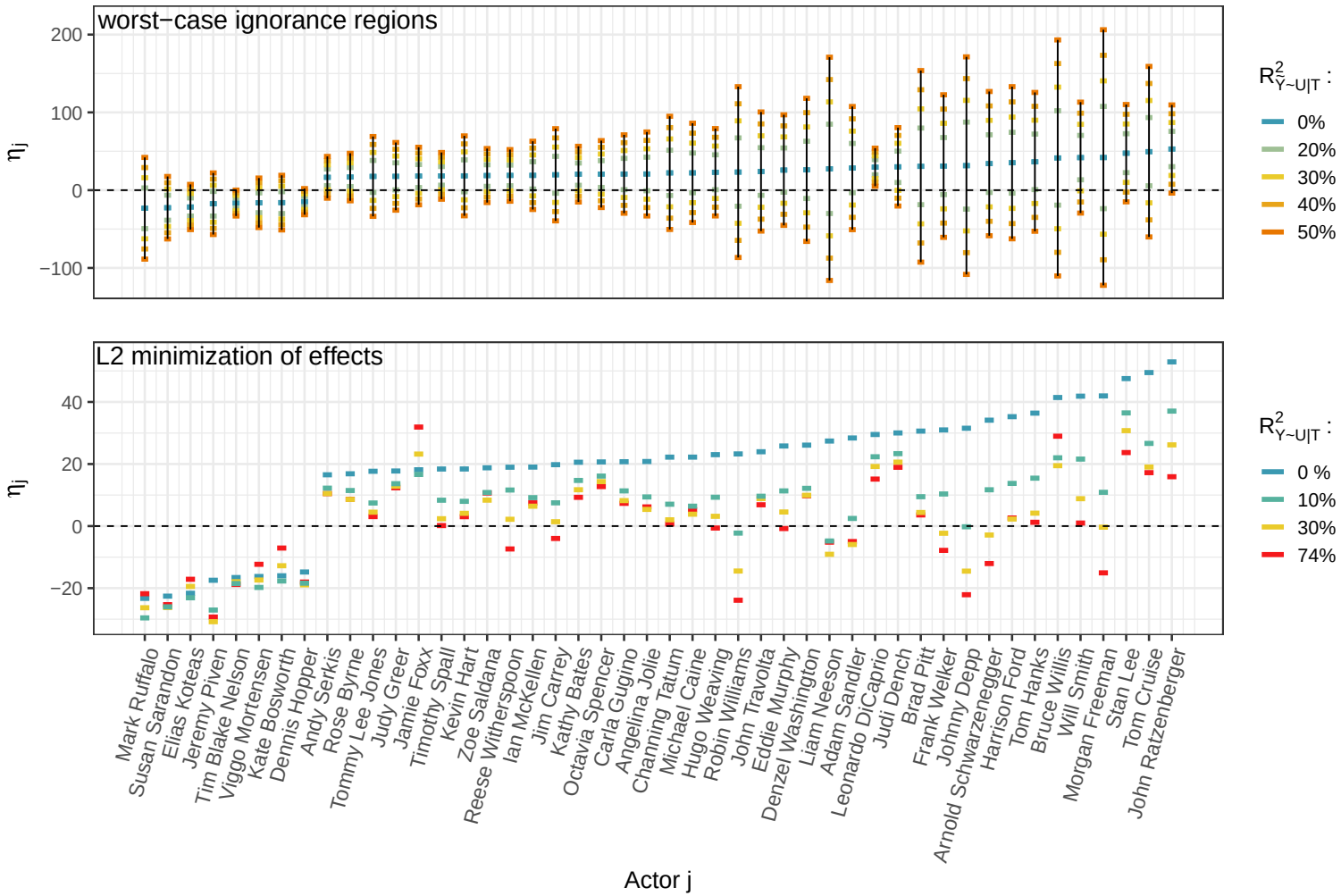
Experiments: Actor Case Study

- Use the TMDB 5000 Movie Dataset
- Estimate causal effect of an actor's presence on the movie's log revenue.
- Let Y be the log revenue and $T_i = (T_{i1}, \cdots, T_{ik})$ be the movie cast, where the binary random variable $T_{ij} \in \{0, 1\}$ indicates whether actor j appeared in the movie i and $T_i \in \mathcal{T} = \{T_1, ..., T_n\}$. Also let $\mathcal{T}^j$ denote the set of all movies T_i for which $T_{ij} = 1$. We define the estimand of interest, η_j, as the the total log revenue contributed by actor j:

$$\eta_j := \sum_{t_i \in \mathcal{T}^j} \text{PATE}_{t_i, t_i^j} \tag{28}$$

where t_i^j corresponds to the observed treatment vector for movie i excluding actor j.

Results: Actor Case Study



Conclusion

- Authors propose a sensitivity analysis method in the multi-cause setting
- Bounds on confounding bias in the Linear-Gaussian case and under Gaussian copula assumption
- Identification up to a causal equivalence class of sensitivity parameter ψ_T
- Calibration strategy to set the sensitivity parameters ψ_Y
- Results on the Actor Case Study & other simulated datasets

D'Amour, A. (2019a). Comment: Reflections on the deconfounder.

D'Amour, A. (2019b). On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3478–3486.

Ogburn, E. L., I. Shpitser, and E. J. T. Tchetgen (2019). Comment on "blessings of multiple causes". *Journal of the American Statistical Association 114*(528), 1611–1615.

Wang, Y. and D. M. Blei (2018, May). The Blessings of Multiple Causes. *arXiv e-prints*, arXiv:1805.06826.