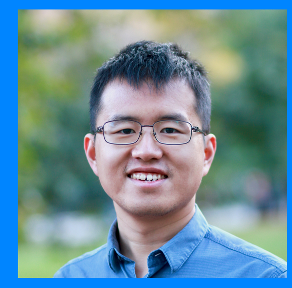
Sensitivity Analysis with the R^2-Calculus

Tobias Freidling Qingyuan Zhao
Statistical Laboratory, University of Cambridge

Sensitivity Analysis is an Optimization Problem

Problem

- Plethora of methods for sensitivity analysis, but hardly used in practice
- Focus on closed-form solutions → simplistic sensitivity models and unintuitive sensitivity parameters

Our Framework

- $(V_i, U_i)_{i=1}^n \sim \mathbb{P}_{V,U}$, but only $(V_i)_{i=1}^n$ is observed
- Objective: $\beta(\theta, \psi)$, e.g. causal effect
- θ are estimable parameters, i.e. only depend on $\mathbb{P}_V \rightarrow \hat{\theta}$
- ψ are sensitivity parameters, i.e. depend on $\mathbb{P}_{V,U}$
- Specifying domain knowledge about ψ as constraints:

$$g(\theta, \psi) \leq 0 \qquad h(\theta, \psi) = 0$$

- Solving the plug-in optimization problem:

$$\min_\psi / \max_\psi \beta(\hat{\theta}, \psi) \quad \text{subject to } g(\hat{\theta}, \psi) \leq 0, h(\hat{\theta}, \psi) = 0$$

R^2-Calculus

Let Y be a random variable and X, W, Z be random vectors. $Y^{\perp X}$ is the residual of Y after regressing out X.

Definitions

- R^2-value: $R^2_{Y\sim X} := 1 - \frac{\text{var}(Y^{\perp X})}{\text{var}(Y)}$
- Partial R^2-value: $R^2_{Y\sim X|Z} := \frac{R^2_{Y\sim X+Z} - R^2_{Y\sim Z}}{1 - R^2_{Y\sim Z}}$
- Partial R-value: $R_{Y\sim X|Z} := \text{corr}(Y^{\perp Z}, X^{\perp Z})$
- Partial f-value: $f_{Y\sim X|Z} := \frac{R_{Y\sim X|Z}}{\sqrt{1 - R^2_{Y\sim X|Z}}}$

R^2-Calculus[2]

System of algebraic rules for (partial) R^2- and R-values, e.g.

(i) Independence: If $Y \perp\!\!\!\perp X$, then $R^2_{Y\sim X} = 0$

(iv) Recursion of partial correlation:

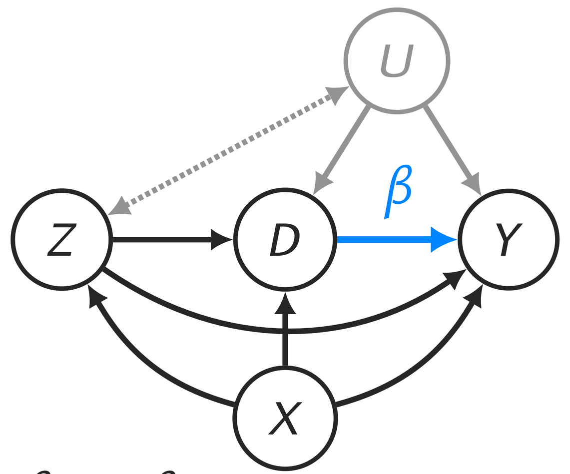
$$R_{Y\sim X|W} = \frac{R_{Y\sim X} - R_{Y\sim W} R_{X\sim W}}{\sqrt{1 - R^2_{Y\sim W}}\sqrt{1 - R^2_{X\sim W}}}$$

(v) Reduction of partial correlation: If $Y \perp\!\!\!\perp W$, then

$$R_{Y\sim X|W} = \frac{R_{Y\sim X}}{\sqrt{1 - R^2_{X\sim W}}}$$

Sensitivity Analysis in a Linear Model

Key Idea: Using R- and R^2-values as sensitivity parameters ψ because they are easy to interpret. [3]



- Causal effect: $\beta = \beta_{Y\sim D|X,Z,U}$
- Family of k-class estimands β_k, where $k \leq 1$:

$$\beta_1 = \beta_{\text{TSLS}}, \quad \beta_0 = \beta_{Y\sim D|X}, \quad \lim_{k\to-\infty} \beta_k = \beta_{Y\sim D|X,Z}$$

Via the R^2-calculus, we get:

$$\beta_k - \beta = \left[\frac{f_{Y\sim Z|X,D}\, R_{D\sim Z|X}}{1 - k + k R^2_{D\sim Z|X}} + R_{Y\sim U|X,Z,D}\, f_{D\sim U|X,Z}\right] \frac{\text{sd}(Y^{\perp X,Z,D})}{\text{sd}(D^{\perp X,Z})}$$

Choose the sensitivity parameters:

$$\psi = (R_{Y\sim U|X,Z,D}, R_{D\sim U|X,Z})$$

Specifying Domain Knowledge

- Experts specify their knowledge as constraints on ψ.
- Assume that $X = (\tilde{X}, \dot{X})$ can be split such that $\dot{X} \perp\!\!\!\perp U | \tilde{X}, Z$.
- Direct bounds and comparative bounds, e.g.

$$R_{D\sim U|X,Z} \in [-0.2, 0.4] \qquad R^2_{Y\sim U|\tilde{X},\dot{X}_{-j},Z} \leq 2 R^2_{Y\sim \dot{X}_j|\tilde{X},\dot{X}_{-j},Z} \tag{1}$$

"U can explain at most 2 times as much variance in Y than $\dot{X}_j$ does - after accounting for $(\tilde{X}, \dot{X}_{-j}, Z)$."

- Additional equality constraints to connect (1) to initial sensitivity parameters:

$$R^2_{Y\sim U|X,Z} \overset{(v)}{=} \frac{R^2_{Y\sim U|\tilde{X},\dot{X}_{-j},Z}}{1 - R^2_{Y\sim \dot{X}_j|\tilde{X},\dot{X}_{-j},Z}} \tag{2}$$

$$R_{Y\sim U|X,Z,D} \overset{(iv)}{=} \frac{R_{Y\sim U|X,Z} - R_{Y\sim D|X,Z}\, R_{D\sim U|X,Z}}{\sqrt{1 - R^2_{Y\sim D|X,Z}}\sqrt{1 - R^2_{D\sim U|X,Z}}} \tag{3}$$

- Optimization problem becomes

$$\min/\max \beta_{Y\sim D|X,Z} - R_{Y\sim U|X,Z,D}\, f_{D\sim U|X,Z} \frac{\text{sd}(Y^{\perp X,Z,D})}{\text{sd}(D^{\perp X,Z})}$$

subject to (1), (2), (3)

- Whole suite of constraints for the relationships:

$$U \to D \qquad U \to Y \qquad U \leftrightarrow Z \qquad Z \to Y$$

Practical Application

- National Longitudinal Survey of Young Men [4]
- Causal effect β of education D on salary Y
- Potential instrument Z: growing up close to a college
- Baseline covariates X: $\dot{X}$ indicator for being black, $\tilde{X}$ remaining covariates
- Unmeasured confounder U: motivation

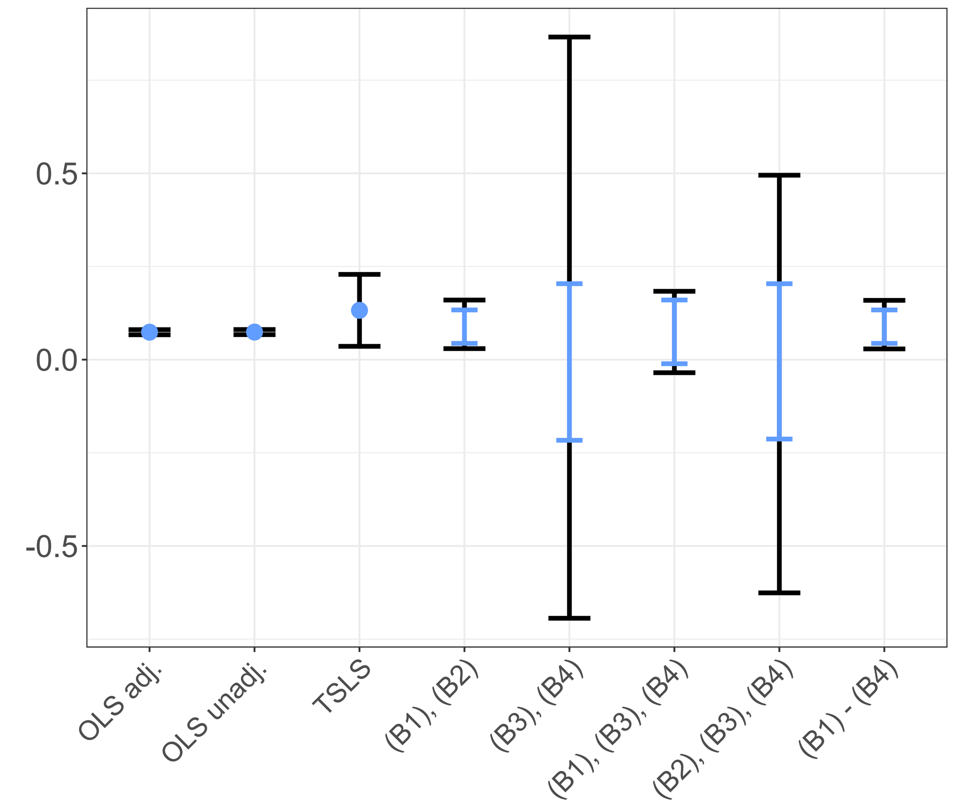
Sensitivity analysis with comparative bounds:

(B1) $R^2_{D\sim U|\tilde{X},Z} \leq 4 R^2_{D\sim \dot{X}|\tilde{X},Z}$ (B2) $R^2_{Y\sim U|\tilde{X},Z,D} \leq 5 R^2_{Y\sim \dot{X}|\tilde{X},Z,D}$

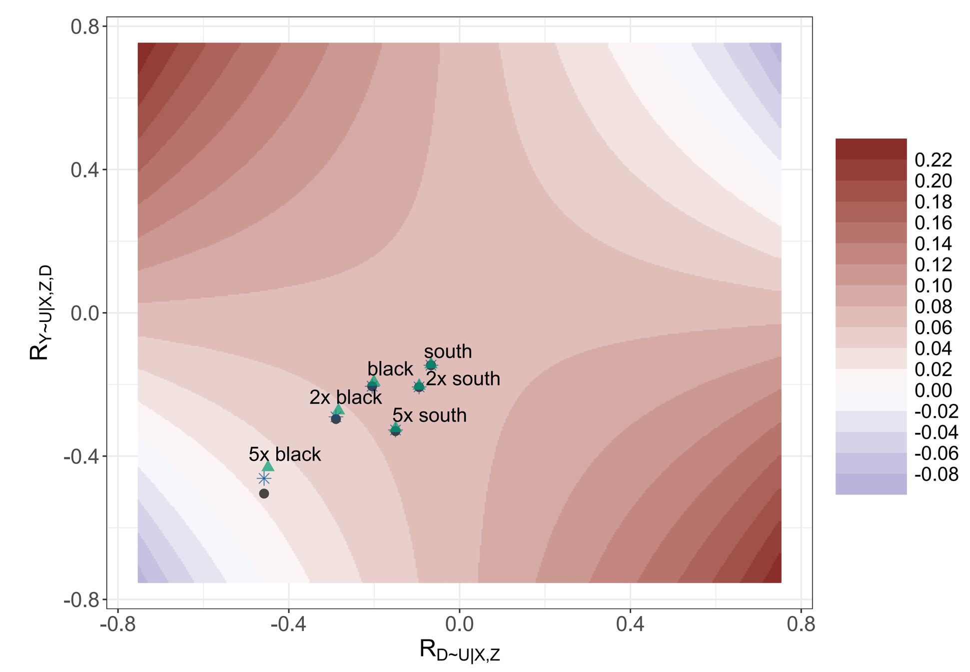
(B3) $R^2_{Z\sim U|\tilde{X}} \leq 0.5 R^2_{Z\sim \dot{X}|\tilde{X}}$ (B4) $R^2_{Y\sim Z|X,U,D} \leq 0.1 R^2_{Y\sim \dot{X}|\tilde{X},Z,U,D}$

Range of estimates (blue)

95% confidence intervals (black)



Strength of U in comparison with observed covariates



References

[1] Tobias Freidling and Qingyuan Zhao. Sensitivity analysis with the R^2-calculus. *arXiv 2301.00040*, 2023.

[2] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Publications in Mathematical Statistics, New York; London, 1st edition, 1958.

[3] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

[4] David Card. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Technical Report w4483, National Bureau of Economic Research, 1993.

E-mail: taf40@cam.ac.uk

tobias-freidling.onrender.com