Sensitivity Analysis with the R²-Calculus

Tobias Freidling Qingyuan Zhao
Statistical Laboratory, University of Cambridge

Sensitivity Analysis as Optimisation Problem

Model
- Model for observed and unobserved variables, O and U: $(O, U) \sim \mathbb{P}_{\theta,\psi}$
- θ parametrises the observable and ψ the unobservable aspects of $\mathbb{P}_{\theta,\psi}$.
- Quantity of interest: $g_n(\hat{\theta}, \psi)$, e.g. a point estimate or confidence interval for a causal effect.

Sensitivity Analysis
- Causal assumptions like 'no unmeasured confounding' correspond to fixing the value of ψ.
- Sensitivity analysis considers a set of plausible values Ψ instead.
- The range of values of g_n is the solution of

$$\max / \min \; g_n(\hat{\theta}, \psi) \qquad \text{subject to } \psi \in \Psi.$$

Challenges
- Identification of $g_n(\hat{\theta}, \psi)$ in terms of the sensitivity parameters ψ
- Translating domain knowledge into the constraints $\psi \in \Psi$

Using interpretable R^2-values helps practitioners to express their beliefs about the unmeasured confounder. The rules of the R^2-calculus translate these into Ψ.

R²-Calculus

The R^2-calculus assembles algebraic rules for R^2-values and correlations as a coherent system of their own.

Definitions Let $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times q}$ be n i.i.d. samples.
- R^2-value and partial R^2-value:

$$R^2_{Y\sim X} := 1 - \frac{\operatorname{var}(Y - X\hat{\beta}_X)}{\operatorname{var}(Y)}, \qquad R^2_{Y\sim X|W} := \frac{R^2_{Y\sim X+W} - R^2_{Y\sim W}}{1 - R^2_{Y\sim W}}.$$

- R-value: $R_{Y\sim X} := \operatorname{corr}(Y, X)$ for $X \in \mathbb{R}^n$.
- f-value: $f_{Y\sim X} := R_{Y\sim X}/\sqrt{1 - R^2_{Y\sim X}}$.

Some Calculation Rules
- Decomposition of unexplained variance:

$$1 - R^2_{Y\sim X+W} = (1 - R^2_{Y\sim X|W})(1 - R^2_{Y\sim W})$$

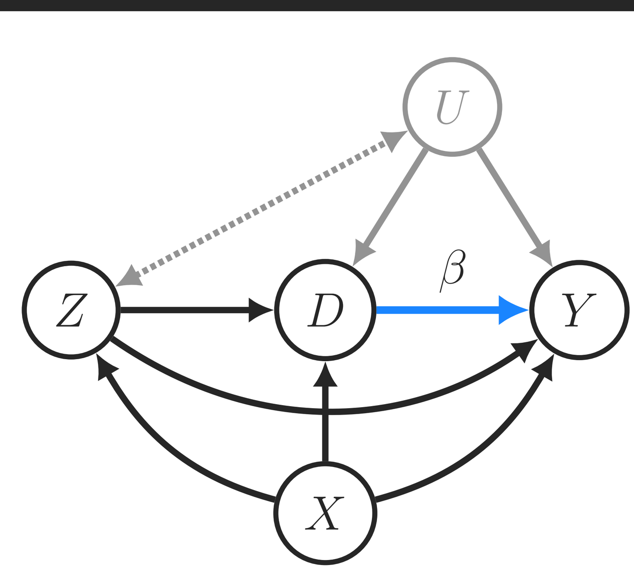
- Recursive partial correlation formula:

$$R_{Y\sim X|W} = \frac{R_{Y\sim X} - R_{Y\sim W}R_{X\sim W}}{\sqrt{1 - R^2_{Y\sim W}}\sqrt{1 - R^2_{X\sim W}}}, \qquad \text{for } X, W \in \mathbb{R}^n$$

Bias of the k-class estimator

Linear model with continuous outcome $Y \in \mathbb{R}^n$, observed variables $D \in \mathbb{R}^n$, $Z \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and unmeasured confounder $U \in \mathbb{R}^n$:

$$Y := \nu + D\beta + Z\delta + X\xi + U\lambda + \varepsilon.$$



We estimate β, the causal effect of D on Y, with the k-class estimator

$$\hat{\beta}_k = \frac{\operatorname{cov}(D^{\perp X}, Y^{\perp X}) - k\operatorname{cov}(D^{\perp Z,X}, Y^{\perp Z,X})}{\operatorname{var}(D^{\perp X}) - k\operatorname{var}(D^{\perp Z,X})}.$$

It interpolates between the OLS-estimator ($k \to -\infty$) and the IV-estimator ($k = 1$) with instrumental variable Z.

Via the R^2-calculus and a technical result in [1], we get:

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

Similarly, we can identify the ends of a $1-\alpha$ confidence interval. This extends previous work on the OLS-estimator [2].

We choose the sensitivity parameters $\psi = (R_{D\sim U|X,Z}, R_{Y\sim U|X,Z,D})$.

References

[1] Carrie A. Hosman, Ben B. Hansen, and Paul W. Holland. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 4(2):849 – 870, 2010.

[2] 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.

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

[4] Tobias Freidling and Qingyuan Zhao. Sensitivity Analysis with the R²-calculus. *Working Paper (available upon request)*, 2022.

Specifying Bounds Ψ

Assume that $X = (\tilde{X}, \dot{X})$ can be separated such that $\dot{X} \perp\!\!\!\perp U | \tilde{X}$ holds.

Linear Regression
- Choose constraints from the $U \to D$ and $U \to Y$ block of the table below.
- Range bounds: direct specification of ψ, e.g. $R_{D\sim U|X,Z} \in [-0.3, 0.2]$.
- Comparative bounds: benchmarking against observed covariates, e.g. $R^2_{D\sim U|\tilde{X},\dot{X}_{-j},Z} \le 0.5\, R^2_{D\sim \dot{X}_j|\tilde{X},\dot{X}_{-j},Z}$ means that U can explain at most half as much variability in D than $\dot{X}_j$ can after partialling out $(\tilde{X}, \dot{X}_{-j}, Z)$.

Comparative bounds are translated into constraints on ψ with the R^2-calculus.

Instrumental Variable
- Choose any constraints from the table.
- R-values that parametrise the direct effect of Z on Y and the correlation between Z and U are connected to ψ via

$$f_{Y\sim Z|X,U,D}\sqrt{1 - R^2_{Y\sim U|X,D,Z}} = f_{Y\sim Z|X,D}\sqrt{1 - R^2_{Z\sim U|X,D}} - R_{Y\sim U|X,D,Z}\, R_{Z\sim U|X,D},$$

$$f_{Z\sim U|X,D}\sqrt{1 - R^2_{D\sim U|X,Z}} = f_{Z\sim U|X}\sqrt{1 - R^2_{D\sim Z|X}} - R_{D\sim Z|X}\, R_{D\sim U|X,Z}.$$

- Add these equations as constraints to the optimisation problem.

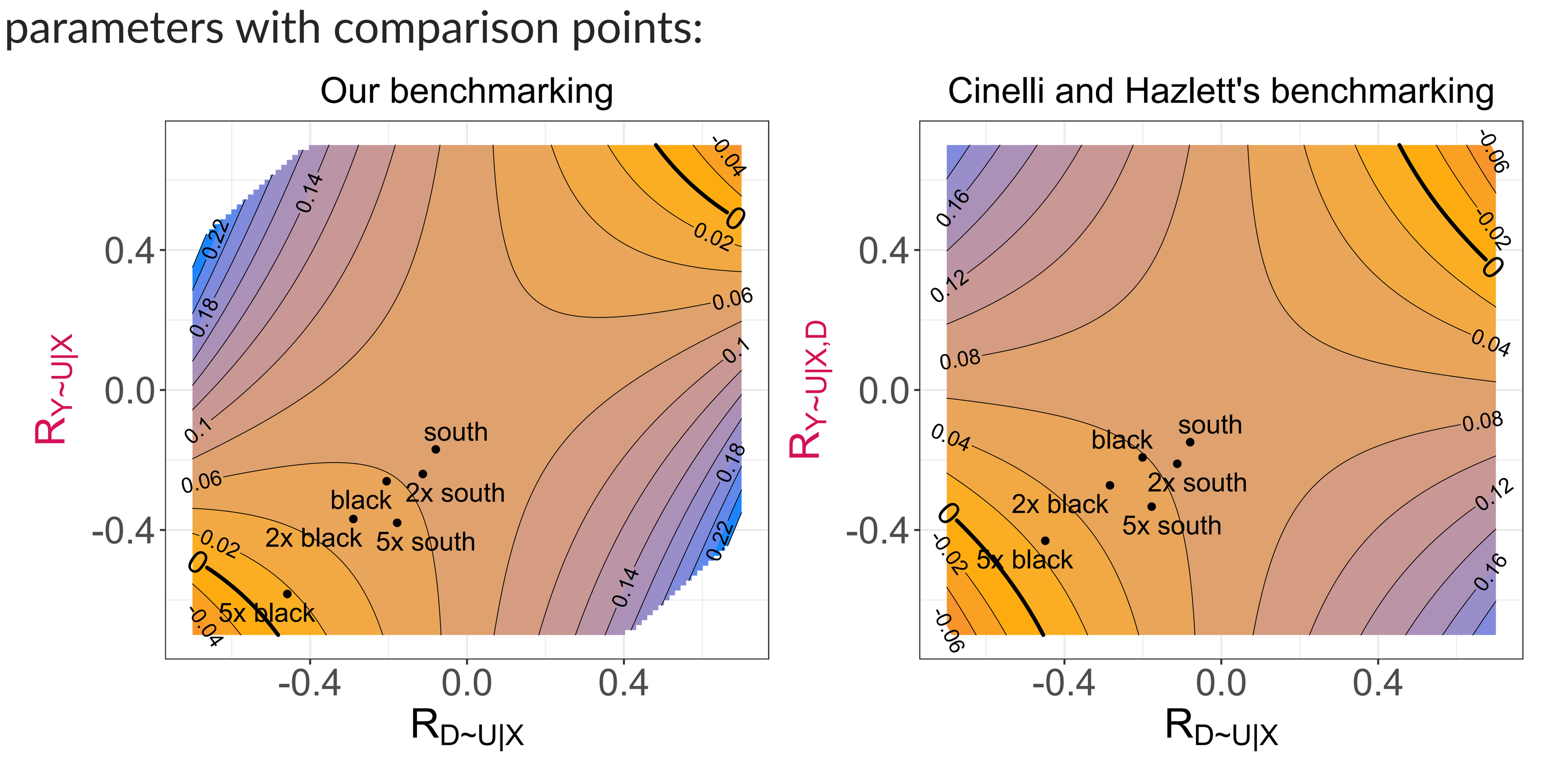
$U \to D$	1. $R_{D\sim U|X,Z} \in [B_l^D, B_u^D]$ 2. $R^2_{D\sim U|\tilde{X},\dot{X}_I,Z} \le \eta_D\, R^2_{D\sim \dot{X}_J|\tilde{X},\dot{X}_I,Z}$
$U \to Y$	1. $R_{Y\sim U|X,Z,D} \in [B_l^Y, B_u^Y]$ 2. $R^2_{Y\sim U|\tilde{X},\dot{X}_I,Z} \le \eta_Y\, R^2_{Y\sim \dot{X}_J|\tilde{X},\dot{X}_I,Z}$ 3. $R^2_{Y\sim U|\tilde{X},\dot{X}_I,Z,D} \le \eta_Y\, R^2_{Y\sim \dot{X}_J|\tilde{X},\dot{X}_I,Z,D}$
$U \leftrightarrow Z$	1. $R_{Z\sim U|X} \in [B_l^Z, B_u^Z]$ 2. $R^2_{Z\sim U|\tilde{X},\dot{X}_{-j}} \le \eta_Z\, R^2_{Z\sim \dot{X}_j|\tilde{X},\dot{X}_{-j}}$
$Z \to Y$	1. $R_{Y\sim Z|X,U,D} \in [B_l^{YZ}, B_u^{YZ}]$ 2. $R^2_{Y\sim Z|X,U,D} \le \eta_{YZ}\, R^2_{Y\sim \dot{X}_j|\tilde{X},\dot{X}_{-j},U,D}$

Table Specify interpretable bounds and use the R^2-calculus to translate them into the constraints Ψ of the optimisation problem.

Insights

Data example: Inference on the causal effect of education on earnings, cf. [3].

Linear Regression
Lower end of the 95% confidence interval for different values of the sensitivity parameters with comparison points:



Instrumental Variable
We use college proximity as instrument for education and set the bounds

$$R_{D\sim U|X,Z} \in [-0.9, 0.9], \qquad R^2_{Y\sim U|X_{-j},Z,D} \le 5\, R^2_{Y\sim X_j|X_{-j},Z,D},$$

where X_j is an indicator for being black. For different IV-related bounds, we get