Shrinkage Estimation for Causal Inference and Experimental Design

Evan Rosenman[†], Guillaume Basse, Mike Baiocchi, Art Owen, and Luke Miratrix

[†]Harvard University Data Science Initiative

December 4, 2022

Outline

Motivating Setting

Randomized Controlled Trials (RCT)

- Researcher controls assignment to treatment

Observational Databases (ODB)

- Treatment assignments observed, but not controlled

Motivating Setting

Randomized Controlled Trials (RCT)

- Researcher controls assignment to treatment
  - Relatively few assumptions for unbiasedness
  - Often costly, small

Observational Databases (ODB)

- Treatment assignments observed, but not controlled
  - Confounding $\implies$ unverifiable assumptions for unbiasedness
  - Large, often inexpensive.

Motivating Setting

Randomized Controlled Trials (RCT)

- Researcher controls assignment to treatment
  - Relatively few assumptions for unbiasedness
  - Often costly, small
- "Unbiased but imprecise"

Observational Databases (ODB)

- Treatment assignments observed, but not controlled
  - Confounding $\implies$ unverifiable assumptions for unbiasedness
  - Large, often inexpensive.
- "Precise, but biased"

Why should we care?

- Ubiquity of observational data in modern era
  - Electronic health records, disease surveillance
  - Fitness trackers, wearable devices, "internet of things"
  - E-commerce data, online behavior

Why should we care?

- Ubiquity of observational data in modern era
  - Electronic health records, disease surveillance
  - Fitness trackers, wearable devices, “internet of things”
  - E-commerce data, online behavior
- Two major utilities to these data

Why should we care?

- Ubiquity of observational data in modern era
  - Electronic health records, disease surveillance
  - Fitness trackers, wearable devices, "internet of things"
  - E-commerce data, online behavior
- Two major utilities to these data
  - ODB measures same treatment as RCT $\Longrightarrow$ more precise causal estimates

Why should we care?

- Ubiquity of observational data in modern era
  - Electronic health records, disease surveillance
  - Fitness trackers, wearable devices, "internet of things"
  - E-commerce data, online behavior
- Two major utilities to these data
  - ODB measures same treatment as RCT $\Longrightarrow$ more precise causal estimates
  - ODB completed and designing a prospective RCT $\Longrightarrow$ design focused more on understudied subgroups

Combining these data: Locating the Problem

How do we combine evidence from an RCT and an ODB?

This problem relates to several areas of research:

- Meta-analysis (Mueller et al., 2018; Prevost et al., 2000; Thompson et al., 2011)
- Transportability/generalizability (Stuart et al., 2011; Hartman et al., 2015; Bareinboim and Pearl, 2016)
- Causal inference (Kallus et al., 2018; Ghassami et al., 2022; Mooij et al., 2016)

Our Approach

We consider two problems:

- **How to design shrinkage estimators to merge ODB and RCT data?**
- **How to improve experimental design using shrinkers?**

Work in a stratified setting, arising from:

- Subject matter knowledge
- Modern machine learning technique (Wager and Athey, 2018; Hill, 2011)

Outline

Potential Outcomes Framework

- Have a sample of units $i = 1, \ldots, n$. We are interested in some outcome measure Y

Potential Outcomes Framework

- Have a sample of units $i = 1, \ldots, n$. We are interested in some outcome measure Y
- For each unit, i, we suppose there are two associated values
  - $Y_i(1)$: outcome if unit i receives the treatment
  - $Y_i(0)$: outcome if unit i receives placebo

Potential Outcomes Framework

- Have a sample of units $i = 1, \ldots, n$. We are interested in some outcome measure Y
- For each unit, i, we suppose there are two associated values
  - $Y_i(1)$: outcome if unit i receives the treatment
  - $Y_i(0)$: outcome if unit i receives placebo
- Causal quantity we are interested in is

$$\tau_i = Y_i(1) - Y_i(0)$$

Causal Estimands

- **Fundamental Problem of Causal Inference**
  - Each unit has a treatment status $Z_i \in \{0, 1\}$, and we observe

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

  - Hence: cannot observe both $Y_i(0)$ and $Y_i(1)$ simultaneously!
- Typically settle for:
  - **Average treatment effect (ATE):**

$$\mathbb{E}(Y(1) - Y(0)), \qquad \text{or}$$

  - **Conditional average treatment effect (CATE):**

$$\mathbb{E}(Y(1) - Y(0) \mid X \in \mathcal{X}).$$

Our Problem: Notation

- Observational data: n_o units sampled from

$$(\underbrace{Y_i(0), Y_i(1)}_{\text{potential outcomes}}, \underbrace{X_i}_{\text{covariates}}, \underbrace{Z_i}_{\text{treatment indicators}}) \overset{\text{iid}}{\sim} F_O.$$

- Experimental data: sample n_r units via

$$(Y_i(0), Y_i(1), X_i, Z_i) \overset{\text{iid}}{\sim} F_R.$$

- Assume strata $k = 1, \ldots, K$. Stratum k defined by set of covariates values $\mathcal{X}_k$. Define indicators:

$$S_i = k \iff X_i \in \mathcal{X}_k \, .$$

Assumptions and Non-Assumptions

1. Under F_O,

$$Y_i(1), Y_i(0) \mid X_i \not\perp\!\!\!\perp Z_i$$

No unconfoundedness assumption for observational study.

2. Under F_R,

$$Y_i(1), Y_i(0) \mid X_i \perp\!\!\!\perp Z_i \, .$$

3. For $k = 1, \ldots, K$, have

$$\tau_k \equiv \mathbb{E}_R \left(Y_i(1) - Y_i(0) \mid S_i = k \right) = \mathbb{E}_O \left(Y_i(1) - Y_i(0) \mid S_i = k \right)$$

Assume **transportability** of CATEs across datasets.
Denote as $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ the vector of CATEs

Setup

- Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = \left(\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}\right), \quad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \left(\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}\right),$$

Setup

- Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = \left(\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}\right), \quad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \left(\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}\right),$$

- Under mild conditions, we have

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \sim N\left(\boldsymbol{\tau}, \Sigma_r\right), \quad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \sim \left(\boldsymbol{\tau} + \boldsymbol{\xi}, \Sigma_o\right)$$

  for bias $\boldsymbol{\xi}$ and diagonal covariance matrices Σ_r and Σ_o
  - $\Sigma_r = \mathrm{diag}(\sigma_{r1}^2, \ldots, \sigma_{rK}^2)$ is estimable from the data
  - $\boldsymbol{\xi}$ cannot be estimated, and estimates of Σ_o will be biased

Setup

- Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = \left(\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}\right), \qquad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \left(\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}\right),$$

- Under mild conditions, we have

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \sim N\left(\boldsymbol{\tau}, \Sigma_r\right), \qquad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \sim \left(\boldsymbol{\tau} + \boldsymbol{\xi}, \Sigma_o\right)$$

  for bias $\boldsymbol{\xi}$ and diagonal covariance matrices Σ_r and Σ_o
  - $\Sigma_r = \text{diag}(\sigma_{r1}^2, \ldots, \sigma_{rK}^2)$ is estimable from the data
  - $\boldsymbol{\xi}$ cannot be estimated, and estimates of Σ_o will be biased
- Seek to design shrinkage estimator $\hat{\boldsymbol{\tau}} = f(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$ to minimize expected L_2 loss (optionally weighted by $\boldsymbol{W}$),

$$\mathcal{L}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) = \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right)^{\mathsf{T}} \boldsymbol{W} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right).$$

Useful Prior Work

- **Shrinkage estimation**: "learn weights from the data" $\Longrightarrow$ a rich literature stretching back to multivariate normal mean estimation via the **James-Stein estimator** (Stein, 1956)
- Green and Strawderman (1991) and Green et al. (2005) propose estimators δ_1, δ_2 for shrinkage between ...
  - a normal, unbiased estimator (like $\hat{\tau}_r$), and
  - a biased estimator (like $\hat{\tau}_o$)

- **Key ideas**
  - Take convex combinations of components of $\hat{\tau}_r$ and $\hat{\tau}_o$.
  - Bias-variance tradeoff: estimators can stabilize high-variance $\hat{\tau}_r$ by introducing some bias with shrinkage toward $\hat{\tau}_o$

Outline

1. Problem Background
2. Assumptions and Set-Up
3. Inference
   - A Recipe for Estimators
   - Application to the WHI
4. Design
   - Problem Framework
   - Design Heuristics

Outline

A Generalized Unbiased Risk Estimate (I)

Theorem (Estimator Risk)

Suppose we have $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, *random* $\boldsymbol{B}$, *and* $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{W} (\boldsymbol{v} - \boldsymbol{\theta})$ *where* $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2)$ *and* $\boldsymbol{W} = 1/K \cdot \mathrm{diag}(w_1, \ldots, w_K)$ *is a diagonal weight matrix.*

A Generalized Unbiased Risk Estimate (I)

Theorem (Estimator Risk)

Suppose we have $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, *random* $\boldsymbol{B}$, *and* $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{W} (\boldsymbol{v} - \boldsymbol{\theta})$ *where* $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_k^2)$ *and* $\boldsymbol{W} = 1/K \cdot diag(w_1, \ldots, w_K)$ *is a diagonal weight matrix. Then for*

$$\boldsymbol{\kappa}(\boldsymbol{U}, \boldsymbol{B}) = \boldsymbol{U} - \boldsymbol{\Sigma} \boldsymbol{g}(\boldsymbol{U}, \boldsymbol{B})$$

where $\boldsymbol{g}(\boldsymbol{U}, \boldsymbol{B})$ *is a function of* $\boldsymbol{U}$ *and* $\boldsymbol{B}$ *that is differentiable, satisfying* $E(||\boldsymbol{g}||^2) < \infty$,

A Generalized Unbiased Risk Estimate (I)

Theorem (Estimator Risk)

Suppose we have $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, *random* $\boldsymbol{B}$, *and* $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{\theta})^\mathsf{T} \boldsymbol{W} (\boldsymbol{v} - \boldsymbol{\theta})$ *where* $\boldsymbol{\Sigma} = diag(\sigma_1^2, , \dots, \sigma_k^2)$ *and* $\boldsymbol{W} = 1/K \cdot diag(w_1, \dots, w_K)$ *is a diagonal weight matrix. Then for*

$$\boldsymbol{\kappa}(\boldsymbol{U}, \boldsymbol{B}) = \boldsymbol{U} - \Sigma \boldsymbol{g}(\boldsymbol{U}, \boldsymbol{B})$$

where $\boldsymbol{g}(\boldsymbol{U}, \boldsymbol{B})$ *is a function of* $\boldsymbol{U}$ *and* $\boldsymbol{B}$ *that is differentiable, satisfying* $E(||\boldsymbol{g}||^2) < \infty$, *we have*

$$R(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{U}, \boldsymbol{B})) = \mathbb{E}\left(\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{U}, \boldsymbol{B}))\right)$$

$$= \frac{1}{K}\left(Tr(\boldsymbol{\Sigma W}) + \mathbb{E}\left(\sum_{k=1}^{K} \sigma_k^4 w_k \left(g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\frac{\partial g_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}\right)\right)\right).$$

A Generalized Unbiased Risk Estimate (II)

From Theorem 1, obtain a generalization of Stein's Unbiased Risk Estimate (Stein, 1981),

$$\mathrm{URE}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{Z}, \boldsymbol{Y})) =$$

$$\frac{1}{K}\left(\mathrm{Tr}\left(\Sigma \boldsymbol{W}\right) + \sum_{k=1}^{K} \sigma_{rk}^{4} w_k \left(g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\frac{\partial \boldsymbol{g}_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}\right)\right) .$$

A Generalized Unbiased Risk Estimate (II)

From Theorem 1, obtain a generalization of Stein's Unbiased Risk Estimate (Stein, 1981),

$$\mathrm{URE}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{Z}, \boldsymbol{Y})) =$$
$$\frac{1}{K}\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{W}\right) + \sum_{k=1}^{K} \sigma_{rk}^4 w_k \left(g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\frac{\partial \boldsymbol{g}_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}\right)\right).$$

Common tactic: minimize URE over a hyperparameter (Li et al., 1985; Xie et al., 2012).

A Generalized Unbiased Risk Estimate (II)

From Theorem 1, obtain a generalization of Stein's Unbiased Risk Estimate (Stein, 1981),

$$\mathrm{URE}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{Z}, \boldsymbol{Y})) =$$
$$\frac{1}{K}\left(\mathrm{Tr}\left(\Sigma \boldsymbol{W}\right) + \sum_{k=1}^{K} \sigma_{rk}^4 w_k \left(g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\frac{\partial \boldsymbol{g}_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}\right)\right).$$

Common tactic: minimize URE over a hyperparameter (Li et al., 1985; Xie et al., 2012).

Points us toward a simple procedure:

1. Posit a structure for the shrinkage estimator
2. Derive a functional form by minimizing URE

Case 1: Common Shrinkage Factor

We consider shrinkage estimators which share a common shrinkage λ factor across components. Denote a generic estimator as

$$\boldsymbol{\kappa}(\lambda, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})\,.$$

Case 1: Common Shrinkage Factor

We consider shrinkage estimators which share a common shrinkage λ factor across components. Denote a generic estimator as

$$\boldsymbol{\kappa}(\lambda, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) .$$

Then the URE evaluates to

$$\text{URE}(\lambda) = \text{Tr}\left(\boldsymbol{\Sigma}_r \boldsymbol{W}\right) + \lambda^2 \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \boldsymbol{W} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right) - 2\lambda \text{Tr}(\boldsymbol{\Sigma}_r \boldsymbol{W})$$

Case 1: Common Shrinkage Factor

We consider shrinkage estimators which share a common shrinkage λ factor across components. Denote a generic estimator as

$$\boldsymbol{\kappa}(\lambda, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \,.$$

Then the URE evaluates to

$$\text{URE}(\lambda) = \text{Tr}\left(\boldsymbol{\Sigma}_r \boldsymbol{W}\right) + \lambda^2 \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \boldsymbol{W} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right) - 2\lambda \text{Tr}(\boldsymbol{\Sigma}_r \boldsymbol{W})$$

which has minimizer in λ,

$$\lambda_1^{\text{URE}} = \frac{\text{Tr}(\boldsymbol{\Sigma}_r \boldsymbol{W})}{\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \boldsymbol{W} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)} \,.$$

A Note on λ_1^{URE}

The true risk-minimizing shrinkage weight is given by

$$\lambda_{\text{opt}} = \frac{\operatorname{Tr}(\Sigma_r \boldsymbol{W})}{\operatorname{Tr}(\Sigma_r \boldsymbol{W}) + \operatorname{Tr}(\Sigma_o \boldsymbol{W}) + \underbrace{\xi^{\mathsf{T}} \boldsymbol{W} \xi}_{\text{Not estimable from data}}},$$

but observe that

$$E\left((\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})\right) = \operatorname{Tr}(\Sigma_r \boldsymbol{W}) + \operatorname{Tr}(\Sigma_o \boldsymbol{W}) + \xi^{\mathsf{T}} \boldsymbol{W} \xi \,.$$

λ_1^{URE} substitutes the quadratic form for its expectation,

$$\lambda_1^{\text{URE}} = \frac{\operatorname{Tr}(\Sigma_r \boldsymbol{W})}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \,.$$

Useful Properties of λ_1^{URE} (I)

1. Define

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda_1^{\text{URE}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

$\boldsymbol{\kappa}_1$ admits a testable condition under which it is guaranteed to reduce risk relative to $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$.

Lemma ($\boldsymbol{\kappa}_1$ Risk Guarantee)

Suppose $4 \max_k w_k \sigma_{rk}^2 < \sum_k w_k \sigma_{rk}^2$. Then $\boldsymbol{\kappa}_1$ has risk strictly less than that of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$.

- Requires a dimension of at least $K = 4$.
- May require substantially larger K if high heteroscedasticity or non-uniform weights.

Useful Properties of λ_1^{URE} (II)

2. Its positive part analogue,

$$\boldsymbol{\kappa}_{1+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{\lambda_1^{\mathrm{URE}}\right\}_{[0,1]} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}\right) ,$$

where

$$\{u\}_{[0,1]} = \min(\max(u, 0), 1) ,$$

satisfies the following notion of optimality:

Useful Properties of λ_1^{URE} (III)

Theorem ($\boldsymbol{\kappa}_{1+}$ Asymptotic Risk)

Suppose

$$\limsup_{K\to\infty} \frac{1}{K}\sum_k d_k^2\sigma_{rk}^2\xi_k^2 < \infty\,, \quad \limsup_{K\to\infty} \frac{1}{K}\sum_k d_k^2\sigma_{rk}^2\sigma_{ok}^2 < \infty\,,$$

$$\text{and} \quad \limsup_{K\to\infty} \frac{1}{K}\sum_k d_k^2\sigma_{rk}^4 < \infty\,.$$

Then, in the limit $K \to \infty$, $\boldsymbol{\kappa}_{1+}$ has the lowest risk among all estimators with a shared shrinkage factor across components.

Case 2: Variance-Weighted Shrinkage Factor

This procedure is general purpose. For example, may instead want an estimator that shrinks each component proportionally to σ^2_{rk}.

Easy to solve for

$$\boldsymbol{\kappa}_2 = \boldsymbol{\kappa}(\lambda_2^{\text{URE}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \frac{\text{Tr}(\boldsymbol{\Sigma}_r^2 \boldsymbol{W})\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

and its positive-part improvement,

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\text{Tr}(\boldsymbol{\Sigma}_r^2 \boldsymbol{W})\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ .$$

Simulated Data Visualization

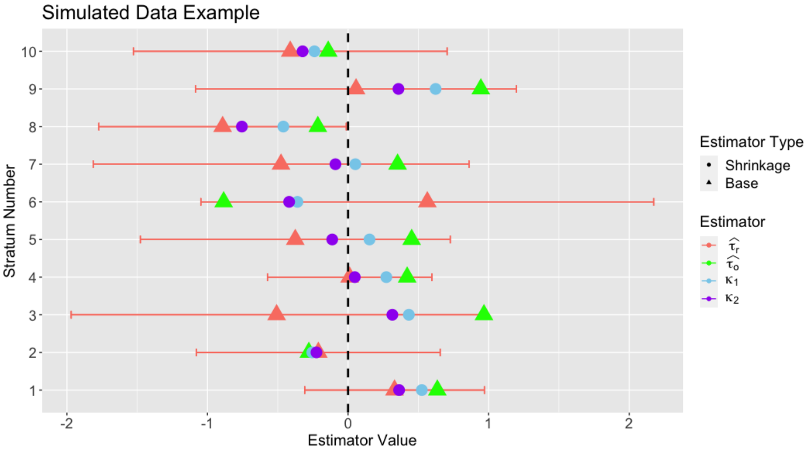


Figure 1: Simulated shrinkage between $\hat{\tau}_r$ and $\hat{\tau}_o$ with ten strata. 90% confidence intervals for $\hat{\tau}_r$ in red, with κ_{1+} and κ_{2+} shown in circles.

Outline

WHI Overview

Dataset Overview

- Study of postmenopausal women initiated in 1991
- RCT of hormone therapy (estrogen and progestin) w/ 16k enrollees
- ODB w/ 50k comparable enrollees

Application to the WHI

- Compute "true" causal effect of **hormone therapy** on **coronary heart disease** using entire RCT (16k units)

Application to the WHI

- Compute "true" causal effect of **hormone therapy** on **coronary heart disease** using entire RCT (16k units)
- Repeat 500 times:
  - Draw bootstrap samples:
    - $1,000$ RCT units
    - Observational sample (50k units)
  - Compute L_2 loss for $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_{1+}, \boldsymbol{\kappa}_{2+}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2$.

Application to the WHI

- Compute "true" causal effect of **hormone therapy** on **coronary heart disease** using entire RCT (16k units)
- Repeat 500 times:
  - Draw bootstrap samples:
    - $1,000$ RCT units
    - Observational sample (50k units)
  - Compute L_2 loss for $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_{1+}, \boldsymbol{\kappa}_{2+}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2$.
- Average loss over draws

Choice of Stratification Variables

Stratify on:

- two variables from WHI protocol:
  age + history of cardiovascular disease (Roehm, 2015).
- a variable unassociated with treatment effect:
  solar irradiance ("Langley") $\Longrightarrow$ uncorrelated with outcome

Results

Subgroup Variable(s)	# of Strata	Loss as % of $\hat{\tau}_r$ Loss			
		κ_{1+}	κ_{2+}	δ_1	δ_2
CVD	2	37.6%	**36.9%**	100.0%	100.0%
Age	3	37.3%	**30.1%**	61.5%	72.8%
Langley	5	29.4%	**23.5%**	40.0%	52.2%
CVD, Age	6	**38.0%**	38.2%	38.3%	82.4%
CVD, Langley	10	30.6%	32.5%	**30.0%**	87.2%
Age, Langley	15	**22.4%**	23.0%	22.5%	43.1%
Age, CVD, Langley	30	50.3%	**50.3%**	50.3%	78.4%

Outline

Outline

A New Setting: Design

Can these insights inform the design of a **prospective** RCT?

- Observational study already completed, $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ obtained.
- Designing a prospective RCT of n_r units
- Want to use a shrinker to combine $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ with $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. Design experiment to better complement ODB

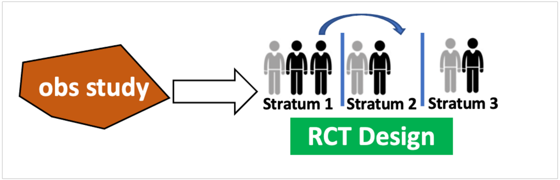
A New Setting: Design

Can these insights inform the design of a **prospective** RCT?

- Observational study already completed, $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ obtained.
- Designing a prospective RCT of n_r units
- Want to use a shrinker to combine $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ with $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. Design experiment to better complement ODB

Goal: choose an RCT allocation of treated and control counts per stratum, $\boldsymbol{d} = \{(n_{rkt}, n_{rkc})\}_{k=1}^{K}$, s.t. $\sum_k n_{rkt} + n_{rkc} = n_r$:

- implies how to *recruit* ...
- and *assign* treatment



Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_r^2 \boldsymbol{W})\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_r^2 \boldsymbol{W})\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \xi)$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_r^2 \boldsymbol{W})\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \xi)$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^{K}$

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\operatorname{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^K$
- bias vector $\boldsymbol{\xi}$.

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_r^2 \boldsymbol{W})\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^2 \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^K$
- bias vector $\boldsymbol{\xi}$.

Can compute this efficiently via numerical integration (Bao and Kan, 2013), as long as $\boldsymbol{V}$ and $\boldsymbol{\xi}$ are known.

Outline

1. Neyman Allocation

Can estimate $\boldsymbol{V}$ using pilot estimates obtained from ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\text{var}}\left(Y(1) \mid S = k\right) \quad \text{and} \quad \hat{\sigma}^2_{kc} = \widehat{\text{var}}\left(Y(0) \mid S = k\right) .$$

Simplest design heuristic: use a Neyman allocation, e.g.

$$n_{rkt} = \frac{n_r \cdot \hat{\sigma}^2_{kt}}{\sum_k \hat{\sigma}^2_{kt} + \hat{\sigma}^2_{kc}} \quad \text{and} \quad n_{rkc} = \frac{n_r \cdot \hat{\sigma}^2_{kc}}{\sum_k \hat{\sigma}^2_{kt} + \hat{\sigma}^2_{kc}} .$$

Optimizes over only the non-shrinkage portion of the risk, but reasonable in many practical settings.

2. Naïve Optimization Assuming $\xi = 0$ (I)

Use. a simple heuristic: assume $\boldsymbol{\xi} = 0$. Then solve:

$$\begin{aligned} \text{minimize} \quad & \mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi}) \\ \text{subject to} \quad & \boldsymbol{\xi} = 0, \boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^{K}, \\ & 0 < n_{rkt}, n_{rkc}, , \quad k = 1, \ldots, K, \\ & n_r = \sum_k n_{rkt} + n_{rkc} \, . \end{aligned} \tag{1}$$

But $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$ is not convex in the design $\boldsymbol{d}$...

2. Naïve Optimization Assuming $\xi = 0$ (II)

A practical approach: **greedy algorithm**. Define $\boldsymbol{d_j}$ as design on j^{th} iteration, and define

$$\mathcal{D}_j = \{\boldsymbol{d}' \mid \boldsymbol{d}' \text{ changes one unit across strata/treatment level from } \boldsymbol{d}_j\}.$$

Run Algorithm 2 from several values of $\boldsymbol{d}_0$ and take minimum:

$$
\begin{array}{l}
\texttt{Start with design } \boldsymbol{d}_0 = \{(n_{rkt}^{(0)}, n_{rkc}^{(0)})\}_k. \\
\texttt{For iteration } j = 1, 2, \ldots: \\
\quad \texttt{For each design } \boldsymbol{d}' \texttt{ in } \mathcal{D}_{j-1}: \\
\quad\quad \texttt{Compute } \mathcal{R}_2(\boldsymbol{d'}, \boldsymbol{V}, 0)). \\
\quad \texttt{Set } \boldsymbol{d_j} = \operatorname*{argmin}_{\boldsymbol{d'} \in \mathcal{D}_{j-1}} \mathcal{R}_2(\boldsymbol{d'}, \boldsymbol{V}, 0) \\
\quad \texttt{If } \mathcal{R}_2(\boldsymbol{d_j}, \boldsymbol{V}, 0) >= \mathcal{R}_2(\boldsymbol{d_{j-1}}, \boldsymbol{V}, 0) \\
\quad\quad \texttt{Return } d_{j-1}.
\end{array}
\tag{2}
$$

3. Heuristic Optimization Assuming Worst-Case Error Under Γ-Level Unmeasured Confounding

- Can take a more pessimistic approach again using marginal sensitivity model of Tan (2006)
- For a user-chosen value of $\Gamma \geq 1$:
    - can obtain worst-case $\xi_k(\Gamma)$ using Zhao et al. (2019), and...
    - if outcome $Y_i \in \{0, 1\}$, can obtain associated $\hat{\sigma}^2_{kt}$ and $\hat{\sigma}^2_{kc}$.

3. Heuristic Optimization Assuming Worst-Case Error Under Γ-Level Unmeasured Confounding

- Can take a more pessimistic approach again using marginal sensitivity model of Tan (2006)
- Recall: for a user-chosen value of $\Gamma \geq 1$:
  - can obtain worst-case $\xi_k(\Gamma)$ using Zhao et al. (2019), and...
  - if outcome $Y_i \in \{0, 1\}$, can obtain associated $\hat{\sigma}^2_{kt}$ and $\hat{\sigma}^2_{kc}$.

posit a value of $\Gamma \implies$
collect results into $\boldsymbol{V}(\Gamma)$ and $\boldsymbol{\xi}(\Gamma) \implies$
run Algorithm 2 using $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}(\Gamma), \boldsymbol{\xi}(\Gamma))$ instead

Stratified WHI Study Design of $n_r = 1,000$ units

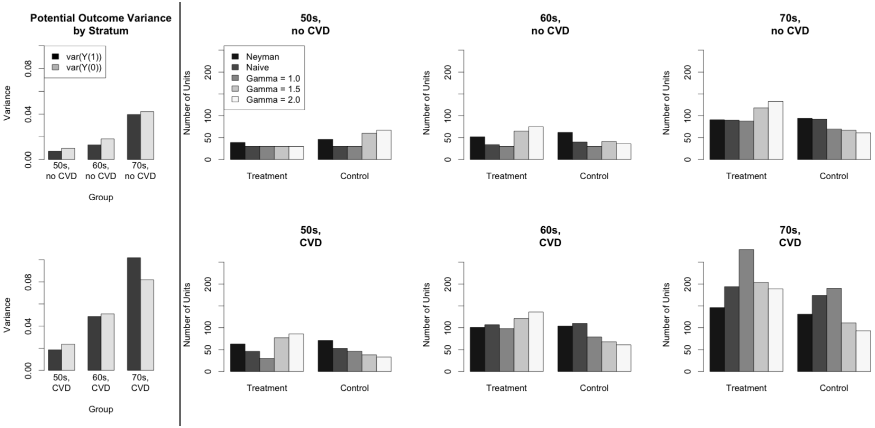


Figure 2: Allocations in WHI with strata defined by history of CVD and age, under different design heuristics.

Future Work

Some areas I'm excited about pursuing:

- **Applied project**: air pollution and mortality (with Francesca Dominici & Luke Miratrix)
  - Combining Medicare ("observational database") database with Medicare Current Beneficiary Survey ("close to" RCT)
  - Approach via *double shrinkage*:

$$\psi_k = a_k \left(\lambda_k \hat{\tau}_{rk} + (1 - \lambda_k) \hat{\tau}_{ok} \right)$$

    where a_k, λ_k are data-driven EB shrinkage parameters

- **ML approaches**
  - Move beyond stratification
  - Flexible shrinkage between CATE functions $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}(x)$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}(x)$

Acknowledgments

Thank you to my collaborators on this work:

- Guillaume Basse
- Mike Baiocchi
- Art Owen
- Luke Miratrix

Inference paper available at arXiv:2002.06708
Design paper available at arXiv:2204.06687

Thanks!

References (I)

Bao, Y. and Kan, R. (2013). On the moments of ratios of quadratic forms in normal random variables. *Journal of Multivariate Analysis*, 117:229–245.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Ghassami, A., Shpitser, I., and Tchetgen, E. T. (2022). Combining experimental and observational data for identification of long-term causal effects. *arXiv preprint arXiv:2201.10743*.

Green, E. J. and Strawderman, W. E. (1991). A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006.

Green, E. J., Strawderman, W. E., Amateis, R. L., and Reams, G. A. (2005). Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1):1–6.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From sate to patt: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 10:1111.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Kallus, N., Puli, A. M., and Shalit, U. (2018). Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pages 10888–10897.

Li, K.-C. et al. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, 13(4):1352–1377.

Mooij, J. M., Magliacane, S., and Claassen, T. (2016). Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*.

Mueller, M., D'Addario, M., Egger, M., Cevallos, M., Dekkers, O., Mugglin, C., and Scott, P. (2018). Methods to systematically review and meta-analyse observational studies: a systematic scoping review of recommendations. *BMC Medical Research Methodology*, 18(1):44.

Prevost, T. C., Abrams, K. R., and Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine*, 19(24):3359–3376.

References (II)

Roehm, E. (2015). A reappraisal of Women's Health Initiative estrogen-alone trial: long-term outcomes in women 50–59 years of age. *Obstetrics and Gynecology International*, 2015.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Shalit, U. (2020). Can we learn individual-level treatment policies from clinical data? *Biostatistics*, 21(2):359–362.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386.

Sutton, A. J., Cooper, N. J., Lambert, P. C., Jones, D. R., Abrams, K. R., and Sweeting, M. J. (2002). Meta-analysis of rare and adverse event data. *Expert review of pharmacoeconomics & outcomes research*, 2(4):367–379.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A. K., Mander, A., Sharp, S., Turner, R., and Wilks, D. (2011). A proposed method of bias adjustment for meta-analyses of published observational studies. *International journal of epidemiology*, 40(3):765–777.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Xie, X., Kou, S., and Brown, L. D. (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Appendices

Practical Considerations

- **Variance estimation:** In practice, Σ_r not known. Must be estimated from data.

Practical Considerations

- **Variance estimation:** In practice, Σ_r not known. Must be estimated from data.
- **Propensity score adjustment**
  - No unconfoundedness $\implies$ propensity score adjustment can't remove all bias
  - If ODB is large, adjusting will typically be good practice. We suggest stabilized IPTW adjustments.

Combining these data: an Ongoing Challenge

"[A model] addressing how data from...randomized and observational evidence can be synthesized, while acknowledging the data [are] susceptible to different types and degrees of biases, is needed." (Sutton et al., 2002)

"We should develop new methods for jointly analyzing experimental and observational data, trying to make the best of both" (Shalit, 2020)

Computing Shrinker Risk

Goal is to optimize experimental design over $\mathcal{R}(\boldsymbol{\kappa}_2)$.

Define $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$ as risk of $\boldsymbol{\kappa}_2$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^{K}$
- bias vector $\boldsymbol{\xi}$.

Computing Shrinker Risk

Goal is to optimize experimental design over $\mathcal{R}(\boldsymbol{\kappa}_2)$.

Define $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$ as risk of $\boldsymbol{\kappa}_2$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^{K}$
- bias vector $\boldsymbol{\xi}$.

Reduces to a ratio of Gaussian quadratic forms! $\Longrightarrow$
solvable via numerical integral of Bao and Kan (2013)

Upshot: can efficiently compute the risk of any design if we have values for $\boldsymbol{V}$ and $\boldsymbol{\xi}$.

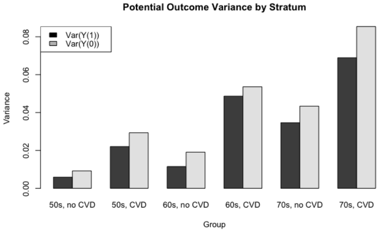
Estimating V: Updated Assumptions

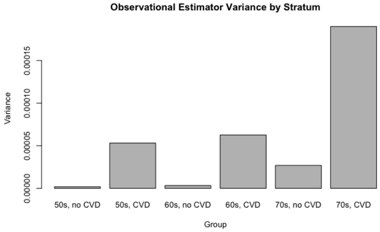
Same assumptions, but a stronger form of **transportability**:

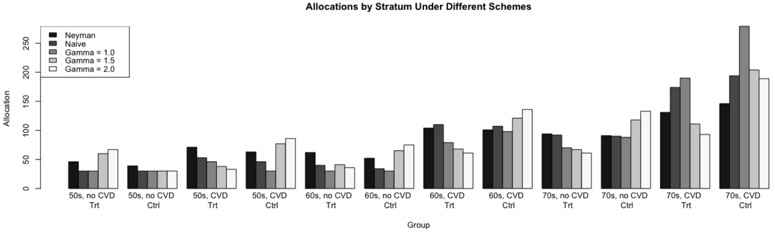
3. For $k = 1, \ldots, K$ and $w \in \{0, 1\}$:

$$\mathbb{E}_O\left(Y(w) \mid S = k\right) = \mathbb{E}_R\left(Y(w) \mid S = k\right) \text{ and}$$
$$\mathrm{var}_O\left(Y(w) \mid S = k\right) = \mathrm{var}_R\left(Y(w) \mid S = k\right) .$$

Sample Designs







Guardrails

Simplicity of Algorithm 2 makes it easy to impose guardrails $\Longrightarrow$ for any invalid design, just set objective value to ∞.

Recommend simple guardrails for designs:

1. **Sample size**: to retain CLT, enforce

$$\min_k n_{rkt} \geq SS_{\min}, \quad \min_k n_{rkc} \geq SS_{\min}$$

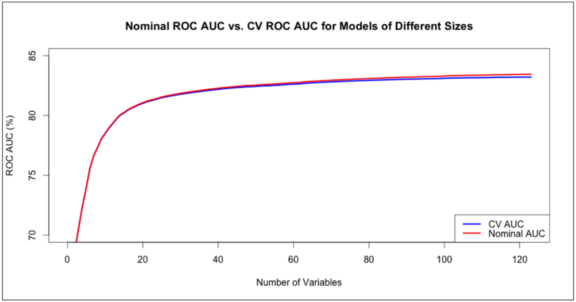
2. **Detachability**: for default design $\tilde{\boldsymbol{d}} = \{\tilde{n}_{rkt}, \tilde{n}_{rkc}\}_k$ and tolerance parameter $\delta_d \geq 1$, enforce

$$\sum_k \frac{\hat{\sigma}^2_{kt}}{n'_{rkt}} + \frac{\hat{\sigma}^2_{kc}}{n'_{rkc}} \geq \delta_d \sum_k \frac{\hat{\sigma}^2_{kt}}{\tilde{n}_{rkt}} + \frac{\hat{\sigma}^2_{kc}}{\tilde{n}_{rkc}},$$

for any proposed design $\boldsymbol{d'} = \{n'_{rkt}, n'_{rkc}\}_k$.

Propensity Modeling

- Rich data set. Consider 684 covariates: demographics, medical history, diet, etc.
- Fit $\hat{e}(\boldsymbol{x}) = \hat{\mathbb{E}}(W \mid \boldsymbol{x})$ by stepwise logistic regression w/ cross-validation. 53 variables chosen.



Covariate Balance (I)

Table 1: Standardized differences (SD) between treated and control populations in the observational dataset, before and after stratification on the propensity score, for clinical risk factors for coronary heart disease.

	Unweighted			**Stratified**		
	Test	**Ctrl**	**SD**	**Test**	**Ctrl**	**SD**
Age	60.78	64.72	−0.56	63.06	63.33	−0.04
BMI	25.55	27.11	−0.25	26.71	26.62	0.00
Physical functioning	85.23	79.58	0.26	81.15	81.23	0.03
Age at menopause	50.49	50.19	0.06	50.35	50.33	0.02

Covariate Balance (II)

Table 2: Standardized differences (SD) between treated and control populations in the observational database, before and after stratification on the propensity score, for ethnicity category.

		White	Black	Latino	AAPI	Native American	Missing/ Other	SD
Before Strat.	Treated	89.0%	2.7%	2.9%	4.0%	0.2%	1.1%	0.26
	Control	83.1%	8.1%	3.9%	2.8%	0.4%	1.5%	
After Strat.	Treated	83.4%	6.9%	4.3%	3.6%	0.5%	1.4%	0.05
	Control	84.8%	6.4%	3.6%	3.4%	0.4%	1.4%	

Covariate Balance (III)

Table 3: Standardized differences (SD) between treated and control populations in the observational database, before and after stratification on the propensity score, for smoking category.

		Never Smoked	**Past Smoker**	**Current Smoker**	**SD**
Before Stratifying	Treated	48.7%	46.2%	5.1%	0.11
	Control	52.3%	41.1%	6.6%	
After Stratifying	Treated	50.9%	42.5%	6.6%	0.01
	Control	51.0%	42.7%	6.3%	

Useful Prior Results (I)

- **Green and Strawderman (1991)** consider the $\Sigma_o = \gamma^2 I_K, \Sigma_r = \sigma_r^2 I_K$ case. Show that estimator

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(1 - \frac{(K-2)\sigma_r^2}{||\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}||^2}\right)_+ (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

  dominates $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ under squared error loss, and has bounded risk as ξ grows large

Useful Prior Results (II)

- **Green et al. (2005)**: Generalize results to heteroscedastic case and propose modified estimators

$$\delta_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(1 - \frac{(K-2)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^{-1} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}\right)_+ (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

$$\delta_2 = \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(1 - \frac{(K-2)\boldsymbol{\Sigma}_r^{-1}}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^{\mathsf{T}} \boldsymbol{\Sigma}_r^{-2} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}\right)_+ (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Fewer theoretical guarantees.
δ_1 is designed for precision-weighted loss, but outperforms δ_2 under regular L_2 loss in simulation.

Integral Expressions

Bao and Kan (2013) give a method for computing these ratios exactly via numerical integrals:

$$\mathbb{E}_r\left(\frac{\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^5\boldsymbol{\nu}}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^3\boldsymbol{\nu})^2}\right) = \int_0^\infty \det(\boldsymbol{I} + 2t\boldsymbol{\Sigma}_r^3)^{-1/2} \cdot \exp\left(\frac{1}{2}\left(\boldsymbol{\xi}^\mathsf{T}(\boldsymbol{I} + 2t\boldsymbol{\Sigma}_r^3)^{-1}\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi}\right)\right) \left(\mathrm{Tr}(\boldsymbol{R}) + (\boldsymbol{L}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi})^\mathsf{T}\boldsymbol{R}(\boldsymbol{L}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi})\right) t dt$$

$$\mathbb{E}_r\left(\frac{1}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^3\boldsymbol{\nu})}\right) = \int_0^\infty \det(\boldsymbol{I} + 2t\boldsymbol{\Sigma}_r^3)^{-1/2} \cdot \exp\left(\frac{1}{2}\left(\boldsymbol{\xi}^\mathsf{T}(\boldsymbol{I} + 2t\boldsymbol{\Sigma}_r^3)^{-1}\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi}\right)\right)$$

$$\text{where} \qquad \boldsymbol{L} = (\boldsymbol{I} + 2t\boldsymbol{\Sigma}_r^3)^{-1/2} \qquad \text{and} \qquad \boldsymbol{R} = \boldsymbol{L}^\mathsf{T}\boldsymbol{\Sigma}_r^5\boldsymbol{L}.$$

This gives us a way to efficiently compute the risk of any design, under a set of assumptions about the values of $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\xi}$.

Improving Interpretability of $\boldsymbol{\kappa}_{1+}$

- Recall: λ_1^{URE} can be interpreted as an estimate of

$$\lambda_{\text{opt}} = \frac{\text{Tr}(\Sigma_r \boldsymbol{W})}{\text{Tr}(\Sigma_r \boldsymbol{W}) + \text{Tr}(\Sigma_o \boldsymbol{W}) + \boldsymbol{\xi}^\mathsf{T} \boldsymbol{W}^2 \boldsymbol{\xi}},$$

  true MSE-minimizing weight on $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ in a convex combination
- We can use this idea to improve interpretability of $\boldsymbol{\kappa}_{1+}$!
- **Key idea**: frame in context of sensitivity model of Tan (2006)

Prior Work

- Marginal sensitivity model of Tan (2006) summarizes degree of unmeasured confounding by a single value, $\Gamma \geq 1$
  - Γ bounds odds ratio of treatment prob. conditional on potential outcomes + covariates vs. covariates only
  - Related to the famous model of Rosenbaum (1987), but extends to the setting of inverse probability weighting
- Zhao et al. (2019) derive valid confidence intervals for causal estimates under the set of models indexed by any choice of Γ
  - Implicitly maps Γ to a worst-case bias $\xi(\Gamma)$ and variance $\Sigma_O(\Gamma)$
  - Under some assumptions, allows us to obtain worst-case estimate of λ_{opt} as a function of Γ, which we call $\lambda(\Gamma)$

Relating the Models

- **Intuition**: larger Γ (confounding parameter) $\implies$ optimal weight λ_{opt} is smaller
- Let $\Gamma_{\text{imp}} = \sup\{\Gamma : \lambda(\Gamma) > \lambda_1^{\text{URE}}\}$
  - Largest value Γ for which the optimal shrinkage factor $\lambda(\Gamma)$ is greater than our shrinkage parameter λ_1^{URE}.
- Γ_{imp} can be used to evaluate level of shrinkage
  - If we believe true confounding level $\Gamma < \Gamma_{\text{imp}}$, then

    $$\lambda_1^{\text{URE}} \approx \lambda(\Gamma_{\text{imp}}) \leq \lambda_{\text{opt}} = \lambda(\Gamma)$$

    Hence the shrinkage level is conservative. ✓
  - If we believe $\Gamma > \Gamma_{\text{imp}}$, then estimator is overshrinking, relies too much on the observational estimate. X

Simulations Set-Up (I)

- ODB has 20K units ($j \in \mathcal{O}$). RCT has 1,000 ($i \in \mathcal{E}$)
- Untreated potential outcomes $Y_\ell \in \{0,1\}$ for $\ell \in \mathcal{O} \cup \mathcal{E}$ sampled as indep. Bernoullis with

$$\Pr(Y_\ell(0) = 1 \mid \boldsymbol{x}_\ell) = \frac{1}{1 + e^{-\alpha - \beta^\mathsf{T} \boldsymbol{x}_\ell + \varepsilon_\ell}}, \quad \text{for } \beta = (1,1,1,1,1)^\mathsf{T}$$

  for covariates $X_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{I}_5)$, α chosen s.t. mean is 10%.
- Treatment variables W_j for $j \in \mathcal{O}$ sampled via

$$\Pr(W_j = 1 \mid \boldsymbol{x}_j) = \frac{1}{1 + e^{-\gamma^\mathsf{T} \boldsymbol{x}_j}}, \quad \text{for } \gamma = (\sqrt{2}, \sqrt{2}, \sqrt{2}, 0, 0)^\mathsf{T}.$$

Simulations Set-Up (II)

- Treatment effects
  - Define $k = 1, \ldots, 12$ strata based on first + second covariate
  - Assign τ_k, stratum CATEs, via 3 treatment effect models:

$$\tau_k = T, \quad \tau_k = -T \times \frac{k}{K}, \quad \text{and} \quad \tau_k = T \times \left(\frac{k}{K}\right)^2$$

  - T chosen so that Cohen's D in ODB equals 0.5
- Simulation structure
  - Sample ODB data a single time. Correct via SIPW.
  - Compute RCT designs under different heuristics
  - Resample RCT units 5,000 times. For each sample, compute L_2 error in estimating $\boldsymbol{\tau}$ using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_2$, and $\boldsymbol{\kappa}_{2+}$

Simulations Set-Up (I)

- ODB has 20K units ($j \in \mathcal{O}$). RCT has 1,000 ($i \in \mathcal{E}$)
- Untreated potential outcomes $Y_\ell \in \{0,1\}$ for $\ell \in \mathcal{O} \cup \mathcal{E}$ sampled as indep. Bernoullis with

$$\Pr(Y_\ell(0) = 1 \mid \boldsymbol{x}_\ell) = \frac{1}{1 + e^{-\alpha - \beta^\mathsf{T} \boldsymbol{x}_\ell + \varepsilon_\ell}}, \quad \text{for } \beta = (1,1,1,1,1)^\mathsf{T}$$

  for covariates $X_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{I}_5)$, α chosen s.t. mean is 10%.
- Treatment variables W_j for $j \in \mathcal{O}$ sampled via

$$\Pr(W_j = 1 \mid \boldsymbol{x}_j) = \frac{1}{1 + e^{-\gamma^\mathsf{T} \boldsymbol{x}_j}}, \quad \text{for } \gamma = (\sqrt{2}, \sqrt{2}, \sqrt{2}, 0, 0)^\mathsf{T}.$$

Simulations Set-Up (II)

- Treatment effects
  - Define $k = 1, \ldots, 12$ strata based on first + second covariate
  - Assign τ_k, stratum CATEs, via 3 treatment effect models:

$$\tau_k = T, \quad \tau_k = -T \times \frac{k}{K}, \quad \text{and} \quad \tau_k = T \times \left(\frac{k}{K}\right)^2$$

  - T chosen so that Cohen's D in ODB equals 0.5
- Simulation structure
  - Sample ODB data a single time. Correct via SIPW.
  - Compute RCT designs under different heuristics
  - Resample RCT units 5,000 times. For each sample, compute L_2 error in estimating $\boldsymbol{\tau}$ using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_2$, and $\boldsymbol{\kappa}_{2+}$

Idealized Case: All Covariates Measured

Est	Trt	Eq.	Ney.	Naïve	Max Bias, Γ Value 1.0	1.1	1.2	1.5	Oracle
$\hat{\tau}_r$		100%	**87%**	91%	100%	96%	94%	94%	96%
κ_2	c	82%	48%	**44%**	52%	48%	47%	50%	42%
κ_{2+}		38%	28%	**26%**	26%	26%	26%	28%	23%
$\hat{\tau}_r$		100%	**89%**	92%	95%	94%	95%	97%	104%
κ_2	ℓ	93%	66%	58%	58%	**57%**	60%	64%	50%
κ_{2+}		59%	51%	45%	**43%**	45%	47%	49%	33%
$\hat{\tau}_r$		100%	**86%**	91%	95%	98%	94%	92%	91%
κ_2	q	81%	47%	**45%**	52%	52%	50%	48%	41%
κ_{2+}		37%	29%	**27%**	28%	28%	30%	29%	25%

Table 4: Risk over 5,000 iterations of $\hat{\tau}_r, \kappa_2$, and κ_{2+} in the case of no unmeasured confounding in the observational study. Risks are expressed as a percentage of the risk of $\hat{\tau}_r$ using an equally allocated experiment, for each of the three treatment effect models.

Simulations: Third Covariate Unmeasured

Est	Trt	Eq.	Ney.	Naïve	Worst Case, Γ Value 1.0	1.1	1.2	1.5	Oracle
$\hat{\tau}_r$		100%	**90%**	90%	90%	92%	93%	95%	102%
κ_2	c	102%	81%	74%	**72%**	72%	72%	77%	69%
κ_{2+}		96%	80%	74%	**71%**	72%	72%	76%	67%
$\hat{\tau}_r$		100%	93%	**93%**	94%	95%	96%	96%	104%
κ_2	ℓ	102%	85%	77%	**75%**	76%	77%	79%	73%
κ_{2+}		98%	84%	77%	**75%**	76%	76%	79%	71%
$\hat{\tau}_r$		100%	**89%**	90%	93%	92%	91%	96%	96%
κ_2	q	101%	74%	69%	68%	68%	**67%**	73%	66%
κ_{2+}		88%	72%	67%	66%	66%	**65%**	71%	63%

Table 5: Risk over 5,000 iterations of $\hat{\tau}_r$, κ_2, and κ_{2+} under various experimental designs, in the case of unmeasured confounding in the observational study via failure to measure the third covariate.