Shrinkage Estimation for Causal Inference and Experimental Design

Evan T. R. Rosenman[†], Guillaume Basse, Mike Baiocchi, Art B. Owen, Francesca Dominici, and Luke Miratrix

[†]Assistant Professor, Claremont McKenna College

February 7, 2024

Introductions

- Hi! I'm Evan Rosenman

Introductions

- Hi! I'm Evan Rosenman
- First-year Asst. Professor of Statistics at Claremont McKenna

Introductions

- Hi! I'm Evan Rosenman
- First-year Asst. Professor of Statistics at Claremont McKenna
- Research interests
  - Causal inference, experimental design (this talk)
  - Voting, elections, political methodology

Outline

Randomized Controlled Trials (RCTs)

- What is an RCT?

Randomized Controlled Trials (RCTs)

- What is an RCT?
  - Canonical "experiment" in medical sciences + public health

Randomized Controlled Trials (RCTs)

- What is an RCT?
    - Canonical "experiment" in medical sciences + public health
    - Individuals *randomized* to receive a treatment v. control

Randomized Controlled Trials (RCTs)

- What is an RCT?
  - Canonical "experiment" in medical sciences + public health
  - Individuals *randomized* to receive a treatment v. control
  - Outcomes measured + compared for treated vs. untreated

Randomized Controlled Trials (RCTs)

- What is an RCT?
  - Canonical "experiment" in medical sciences + public health
  - Individuals *randomized* to receive a treatment v. control
  - Outcomes measured + compared for treated vs. untreated
- Why are RCTs useful?

Randomized Controlled Trials (RCTs)

- What is an RCT?
  - Canonical "experiment" in medical sciences + public health
  - Individuals *randomized* to receive a treatment v. control
  - Outcomes measured + compared for treated vs. untreated
- Why are RCTs useful?
  - Researcher controls treatment assignment $\Longrightarrow$

Randomized Controlled Trials (RCTs)

- What is an RCT?
  - Canonical "experiment" in medical sciences + public health
  - Individuals *randomized* to receive a treatment v. control
  - Outcomes measured + compared for treated vs. untreated
- Why are RCTs useful?
  - Researcher controls treatment assignment $\Longrightarrow$
  - (Almost) no assumptions needed for unbiased treatment effect estimation!

RCTs for Everything?

Why not just use RCTs to answer every causal question?

RCTs for Everything?

Why not just use RCTs to answer every causal question?

A few reasons!

RCTs for Everything?

Why not just use RCTs to answer every causal question?

A few reasons!

- Sometimes, cannot (ethically) run an RCT, e.g. smoking

RCTs for Everything?

Why not just use RCTs to answer every causal question?

A few reasons!

- Sometimes, cannot (ethically) run an RCT, e.g. smoking
- Even when you can, RCTs are typically too small to get a precise estimate for every effect we want...

RCTs for Everything?

Why not just use RCTs to answer every causal question?

A few reasons!

- Sometimes, cannot (ethically) run an RCT, e.g. smoking
- Even when you can, RCTs are typically too small to get a precise estimate for every effect we want...
  - Expensive to run! $\Longrightarrow$ few individuals recruited
  - May need an answer rapidly
  - What about effects on small subgroups?

Observational Databases (ODBs)

- What is an ODB?

Observational Databases (ODBs)

- What is an ODB?
    - Passively collected data like electronic health records, insurance claims databases, etc.

Observational Databases (ODBs)

- What is an ODB?
  - Passively collected data like electronic health records, insurance claims databases, etc.
  - Individuals *not randomized* $\Longrightarrow$
    treatments assigned by some typically unknown procedure

Observational Databases (ODBs)

- What is an ODB?
  - Passively collected data like electronic health records, insurance claims databases, etc.
  - Individuals *not randomized* $\Longrightarrow$

    treatments assigned by some typically unknown procedure

- Can ODBs be useful?

Observational Databases (ODBs)

- What is an ODB?
  - Passively collected data like electronic health records, insurance claims databases, etc.
  - Individuals *not randomized* $\implies$
    treatments assigned by some typically unknown procedure
- Can ODBs be useful?
  - Often large, cheap, and representative! But...

Observational Databases (ODBs)

- What is an ODB?
  - Passively collected data like electronic health records, insurance claims databases, etc.
  - Individuals *not randomized* $\Longrightarrow$
    treatments assigned by some typically unknown procedure
- Can ODBs be useful?
  - Often large, cheap, and representative! But...
  - No randomization $\Longrightarrow$ treated and untreated units differ $\Longrightarrow$ confounding bias!

Observational Databases (ODBs)

- What is an ODB?
  - Passively collected data like electronic health records, insurance claims databases, etc.
  - Individuals *not randomized* $\Longrightarrow$
    treatments assigned by some typically unknown procedure

- Can ODBs be useful?
  - Often large, cheap, and representative! But...
  - No randomization $\Longrightarrow$ treated and untreated units differ $\Longrightarrow$ confounding bias!
    - Ex: Doctors give a drug to sicker patients
    - Ex: Healthier and wealthier people opt into a vaccine

The Data-Combination Problem

RCTs are the gold standard, but yield estimates that are *imprecise*

The Data-Combination Problem

RCTs are the gold standard, but yield estimates that are *imprecise*

ODBs yield *biased* estimates... but cheap and ubiquitous!

- Electronic health records, disease surveillance
- Fitness trackers, wearable devices, “internet of things”
- E-commerce data, online behavior

The Data-Combination Problem

RCTs are the gold standard, but yield estimates that are *imprecise*

ODBs yield *biased* estimates... but cheap and ubiquitous!

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

Overall goals

Can we use ODBs to...

- *obtain better causal estimates by combining ODB causal estimates with those obtained from RCTs?*
- *design prospective experiments to be more accurate?*

A Modern Challenge

- Considerable interest in RCT-ODB evidence synthesis methods from both the FDA (2018) and the European Medicines Agency (2022)

A Modern Challenge

- Considerable interest in RCT-ODB evidence synthesis methods from both the FDA (2018) and the European Medicines Agency (2022)
- Problem relates to several active areas of research:

A Modern Challenge

- Considerable interest in RCT-ODB evidence synthesis methods from both the FDA (2018) and the European Medicines Agency (2022)
- Problem relates to several active areas of research:
  - Meta-analysis (Mueller et al., 2018; Prevost et al., 2000; Thompson et al., 2011)

A Modern Challenge

- Considerable interest in RCT-ODB evidence synthesis methods from both the FDA (2018) and the European Medicines Agency (2022)
- Problem relates to several active 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)

A Modern Challenge

- Considerable interest in RCT-ODB evidence synthesis methods from both the FDA (2018) and the European Medicines Agency (2022)
- Problem relates to several active 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)

A Modern Challenge

- Considerable interest in RCT-ODB evidence synthesis methods from both the FDA (2018) and the European Medicines Agency (2022)
- Problem relates to several active 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)
- Notable uptick in methodological work since roughly 2020 (Oberst et al., 2022; Yang et al., 2023; Cheng and Cai, 2021; Chen et al., 2021; Lin and Evans, 2023).

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!

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

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

- Experimental data: sample n_r units via

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

Stratification

Assume strata $k = 1, \ldots, K$. Stratum k defined by set of covariates values $\mathcal{X}_k$. Define variables: $S_i = k \iff X_i \in \mathcal{X}_k$.

Stratification

Assume strata $k = 1, \dots, K$. Stratum k defined by set of covariates values $\mathcal{X}_k$. Define variables: $S_i = k \iff X_i \in \mathcal{X}_k$.

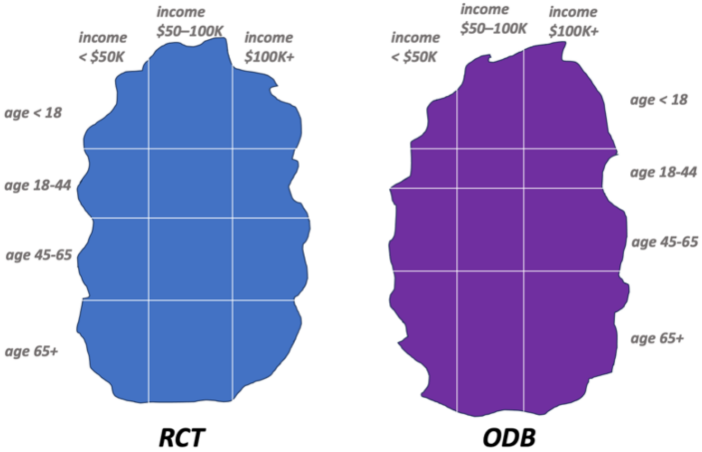


Figure 1: Example stratification of RCT and ODB with 12 strata.

Assumptions and Non-Assumptions

1. Under F_O,

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

No unconfoundedness assumption for observational study.

Assumptions and Non-Assumptions

1. Under F_O,

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

   No unconfoundedness assumption for observational study.

2. Under F_R,

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

Assumptions and Non-Assumptions

1. Under F_O,

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

   No unconfoundedness assumption for observational study.

2. Under F_R,

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

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 $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K) \in \mathbb{R}^K$ the vector of CATEs

Outline

1. Problem Background
2. Assumptions and Set-Up
3. Estimators to Combine Data
   - SURE-Based Procedures
   - Using a Hierarchical Model
4. Application to the WHI

Setup

Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = (\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}), \quad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = (\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}) \in \mathbb{R}^K$$

Setup

Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = (\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}), \qquad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = (\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}) \in \mathbb{R}^K$$

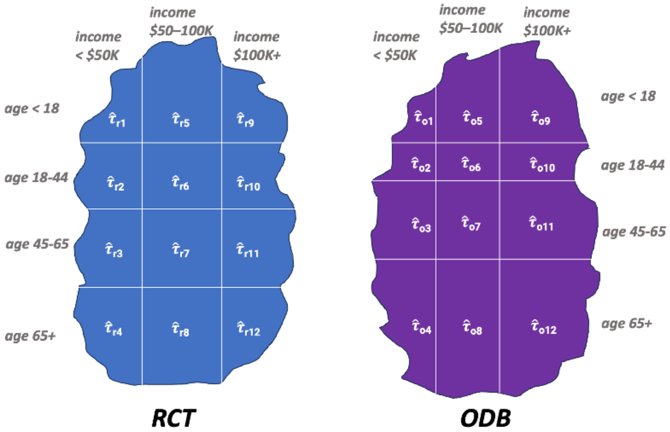


Figure 2: Causal estimates by stratum.

- Under mild conditions, we have

$$\hat{\tau}_r \sim N\left(\tau, \Sigma_r\right), \qquad \hat{\tau}_o \sim \left(\tau + \xi, \Sigma_o\right)$$

for bias ξ and covariance matrices Σ_r and Σ_o

- 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 covariance matrices Σ_r and Σ_o
  - $\Sigma_r = \text{diag}(\sigma^2_{r1}, \ldots, \sigma^2_{rK})$ is estimable from the data
  - $\boldsymbol{\xi}$ cannot be estimated using obs data alone

- 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 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 using obs data alone

- Seek to design estimator $\hat{\boldsymbol{\tau}} = f(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$ to minimize expected squared error loss:

$$\mathcal{L}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) = \sum_{k=1}^{K} (\hat{\tau}_k - \tau_k)^2.$$

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)

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 $\hat{\tau}_o$)

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. Estimators to Combine Data
   - SURE-Based Procedures
   - Using a Hierarchical Model
4. Application to the WHI

Overall Idea

Stein's Unbiased Risk Estimate (SURE): foundational result in the shrinkage estimation literature.

Overall Idea

Stein's Unbiased Risk Estimate (SURE): foundational result in the shrinkage estimation literature.

Upshot: when weighting between between (normal) estimator $\hat{\theta}_1$ and another estimator $\hat{\theta}_2$: SURE is an unbiased estimator of the *estimation error*, even if parameter θ is unknown!

Overall Idea

Stein's Unbiased Risk Estimate (SURE): foundational result in the shrinkage estimation literature.

Upshot: when weighting between between (normal) estimator $\hat{\theta}_1$ and another estimator $\hat{\theta}_2$: SURE is an unbiased estimator of the *estimation error*, even if parameter θ is unknown!

Utility: gives us an objective function! To design estimators, a common tactic (Li et al., 1985; Xie et al., 2012) is to

1. posit a method to do the weighting
2. derive exact functional form by minimizing SURE

A Generalized Version of SURE (I)

Theorem (Estimator Risk)

Suppose we have $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$, *random* $\boldsymbol{B}$, *and* $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{\theta})^{\mathsf{T}}(\boldsymbol{v} - \boldsymbol{\theta})$ *where* $\Sigma = diag(\sigma_1^2, , \ldots, \sigma_k^2)$.

A Generalized Version of SURE (I)

Theorem (Estimator Risk)

Suppose we have $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$, *random* $\boldsymbol{B}$, *and* $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{\theta})^{\mathsf{T}}(\boldsymbol{v} - \boldsymbol{\theta})$ *where* $\Sigma = diag(\sigma_1^2, , \ldots, \sigma_k^2)$. *Then for*

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

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

A Generalized Version of SURE (I)

Theorem (Estimator Risk)

Suppose we have $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$, *random* $\boldsymbol{B}$, *and* $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{\theta})^{\mathsf{T}}(\boldsymbol{v} - \boldsymbol{\theta})$ *where* $\Sigma = diag(\sigma_1^2, , \ldots, \sigma_k^2)$. *Then for*

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

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

$$\mathbb{E}\left(||\boldsymbol{\theta} - \boldsymbol{\kappa}(\boldsymbol{U}, \boldsymbol{B})||_2^2\right) = Tr(\Sigma) + \mathbb{E}\left(\sum_{k=1}^{K} g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\sigma_k^2 \frac{\partial g_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}\right).$$

A Generalized Version of SURE (I)

From this theorem, obtain a generalization of Stein's Unbiased Risk Estimate (Stein, 1981),

$$\text{SURE}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{Z}, \boldsymbol{Y})) = \text{Tr}\,(\Sigma) + \sum_{k=1}^{K} g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\sigma_k^2 \frac{\partial g_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}\,.$$

A Generalized Version of SURE (I)

From this theorem, obtain a generalization of Stein's Unbiased Risk Estimate (Stein, 1981),

$$\mathrm{SURE}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{Z}, \boldsymbol{Y})) = \mathrm{Tr}\,(\Sigma) + \sum_{k=1}^{K} g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\sigma_k^2 \frac{\partial g_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}.$$

A Generalized Version of SURE (I)

From this theorem, obtain a generalization of Stein's Unbiased Risk Estimate (Stein, 1981),

$$\mathrm{SURE}(\boldsymbol{\theta}, \boldsymbol{\kappa}(\boldsymbol{Z}, \boldsymbol{Y})) = \mathrm{Tr}\,(\Sigma) + \sum_{k=1}^{K} g_k^2(\boldsymbol{U}, \boldsymbol{B}) - 2\sigma_k^2 \frac{\partial g_k(\boldsymbol{U}, \boldsymbol{B})}{\partial U_k}.$$

In keeping with the literature, a simple procedure:

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

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 SURE evaluates to

$$\mathrm{SURE}(\lambda) = \mathrm{Tr}\left(\Sigma_r\right) + \lambda^2 \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right) - 2\lambda \mathrm{Tr}(\Sigma_r)$$

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 SURE evaluates to

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

which has minimizer in λ,

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

A Note on λ_1^{SURE}

The true risk-minimizing shrinkage weight is given by

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

A Note on λ_1^{SURE}

The true risk-minimizing shrinkage weight is given by

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

but observe that

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

A Note on λ_1^{SURE}

The true risk-minimizing shrinkage weight is given by

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

but observe that

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

λ_1^{SURE} substitutes $(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^\mathsf{T} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})$ for its own expectation,

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

Useful Property of λ_1^{SURE}

Define

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

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

Useful Property of λ_1^{SURE}

Define

$$\boldsymbol{\kappa}_{1+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \{\lambda_1^{\text{SURE}}\}_{[0,1]} \, (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

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

$\boldsymbol{\kappa}_1$ 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 $4 \max_k \sigma_{rk}^2 < \sum_k \sigma_{rk}^2$. 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.

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

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

and its positive-part improvement,

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

Simulated Data Visualization

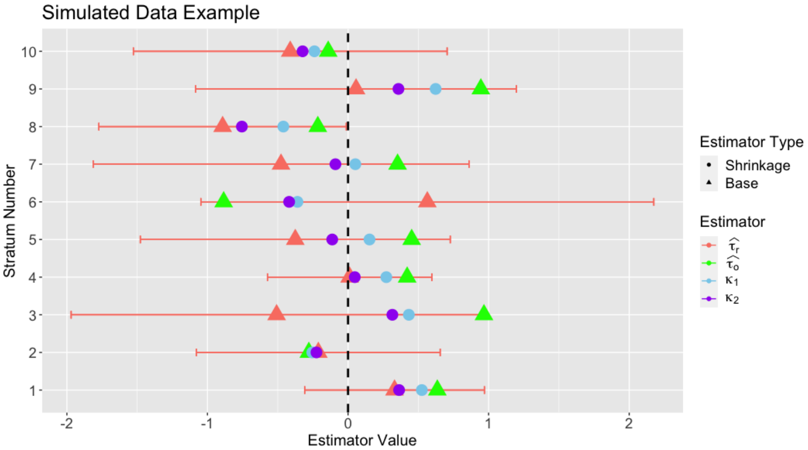


Figure 3: Simulated shrinkage between $\hat{\tau}_r$ and $\hat{\tau}_o$ with ten strata. 90% confidence intervals for $\hat{\tau}_r$ in red, with κ_{1+} and κ_{2+} shown in circles.

Outline

1. Problem Background
2. Assumptions and Set-Up
3. Estimators to Combine Data
   - SURE-Based Procedures
   - Using a Hierarchical Model
4. Application to the WHI

Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters

Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters
- An alternative approach is to derive the functional form from a **hierarchical model**

Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters
- An alternative approach is to derive the functional form from a **hierarchical model**

Simple model generalizing one introduced in Green and Strawderman (1991):

$$\boldsymbol{\tau} \sim \mathcal{N}\left(0, \eta^2 \boldsymbol{I}_K\right),$$
$$\boldsymbol{\xi} \sim \mathcal{N}\left(0, \gamma^2 \boldsymbol{I}_K\right),$$

Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters
- An alternative approach is to derive the functional form from a **hierarchical model**

Simple model generalizing one introduced in Green and Strawderman (1991):

$$\begin{aligned}
\boldsymbol{\tau} &\sim \mathcal{N}\left(0, \eta^2 \boldsymbol{I}_K\right), \\
\boldsymbol{\xi} &\sim \mathcal{N}\left(0, \gamma^2 \boldsymbol{I}_K\right), \\
\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \mid \boldsymbol{\tau} &\sim \mathcal{N}\left(\boldsymbol{\tau}, \Sigma_r\right), \text{ and} \\
\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \mid \boldsymbol{\tau}, \boldsymbol{\xi} &\sim \mathcal{N}\left(\boldsymbol{\tau} + \boldsymbol{\xi}, \Sigma_o\right).
\end{aligned} \tag{1}$$

for **unknown** hyperparameters η^2 and γ^2, but **known** covariance matrices Σ_r, Σ_o.

Estimator Form

Bayesian stats: compute **posterior mean** of $\boldsymbol{\tau}$ under Model 1:

$$\psi_k(\eta^2,\gamma^2) = \underbrace{\left(\frac{\eta^2\left(\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2\right)}{\sigma_{rk}^2\left(\gamma^2+\sigma_{ok}^2\right)+\eta^2\left(\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2\right)}\right)}_{\boldsymbol{a_k}(\boldsymbol{\eta}^2,\boldsymbol{\gamma}^2)\text{: aggregate shrinkage toward zero}} \times \left(\underbrace{\frac{\left(\gamma^2+\sigma_{ok}^2\right)}{\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2}}_{\substack{\boldsymbol{\lambda_k}(\boldsymbol{\eta}^2,\boldsymbol{\gamma}^2)\text{:}\\ \text{data-driven weight}}}\hat{\tau}_{rk} + \underbrace{\frac{\sigma_{rk}^2}{\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2}}_{1-\boldsymbol{\lambda_k}(\boldsymbol{\eta}^2,\boldsymbol{\gamma}^2)}\hat{\tau}_{ok}\right). \tag{2}$$

Estimator Form

Bayesian stats: compute **posterior mean** of $\boldsymbol{\tau}$ under Model 1:

$$\psi_k(\eta^2,\gamma^2) = \underbrace{\left(\frac{\eta^2\left(\gamma^2+\sigma^2_{ok}+\sigma^2_{rk}\right)}{\sigma^2_{rk}\left(\gamma^2+\sigma^2_{ok}\right)+\eta^2\left(\gamma^2+\sigma^2_{ok}+\sigma^2_{rk}\right)}\right)}_{\boldsymbol{a_k(\eta^2,\gamma^2)}\text{: aggregate shrinkage toward zero}} \times \left(\underbrace{\frac{\left(\gamma^2+\sigma^2_{ok}\right)}{\gamma^2+\sigma^2_{ok}+\sigma^2_{rk}}}_{\substack{\boldsymbol{\lambda_k(\eta^2,\gamma^2)}\text{:}\\ \text{data-driven weight}}}\hat{\tau}_{rk} + \underbrace{\frac{\sigma^2_{rk}}{\gamma^2+\sigma^2_{ok}+\sigma^2_{rk}}}_{1-\boldsymbol{\lambda_k(\eta^2,\gamma^2)}}\hat{\tau}_{ok}\right). \tag{2}$$

This is the **double-shrinkage** property: take a data-driven convex combo of $\boldsymbol{\hat{\tau}_r}$ and $\boldsymbol{\hat{\tau}_o}$ and then a Stein-like shrinakge toward zero.

MLE Version of the Estimator

To construct a usable estimator, need estimates of η^2, γ^2.

MLE Version of the Estimator

To construct a usable estimator, need estimates of η^2, γ^2.
An approach from Xie et al. (2012)...

Maximum Likelihood: Observing that

$$\mathcal{L}(\eta^2, \gamma^2) \propto \prod_k \left(\eta^2 + \sigma_{rk}^2\right)^{-1/2} e^{-\frac{\hat{\tau}_{rk}^2}{2\left(\eta^2+\sigma_{rk}^2\right)}} \times \prod_k \left(\eta^2 + \gamma^2 + \sigma_{ok}^2\right)^{-1/2} e^{-\frac{\hat{\tau}_{ok}^2}{2\left(\eta^2+\gamma^2+\sigma_{ok}^2\right)}}.$$

MLE Version of the Estimator

To construct a usable estimator, need estimates of η^2, γ^2.
An approach from Xie et al. (2012)...

Maximum Likelihood: Observing that

$$\mathcal{L}(\eta^2, \gamma^2) \propto \prod_k \left(\eta^2 + \sigma_{rk}^2\right)^{-1/2} e^{-\frac{\hat{\tau}_{rk}^2}{2\left(\eta^2+\sigma_{rk}^2\right)}} \times \prod_k \left(\eta^2 + \gamma^2 + \sigma_{ok}^2\right)^{-1/2} e^{-\frac{\hat{\tau}_{ok}^2}{2\left(\eta^2+\gamma^2+\sigma_{ok}^2\right)}} .$$

We can numerically optimize to obtain the estimates

$$(\hat{\eta}^2_{\text{mle}}, \hat{\gamma}^2_{\text{mle}}) = \max_{\eta^2, \gamma^2 \geq 0} \log\left(\mathcal{L}(\eta^2, \gamma^2)\right) .$$

Confidence Intervals

- Advantage of hierarchical model: straightforward to build confidence intervals

Confidence Intervals

- Advantage of hierarchical model: straightforward to build confidence intervals
- Intervals have Empirical Bayes coverage guarantee *without* enforcing parametric assumptions on distribution of τ and ξ

Confidence Intervals

- Advantage of hierarchical model: straightforward to build confidence intervals
- Intervals have Empirical Bayes coverage guarantee *without* enforcing parametric assumptions on distribution of $\boldsymbol{\tau}$ and $\boldsymbol{\xi}$

Definition (Robust EB Confidence Intervals (EBCIs))

The robust EBCI for ψ_k, the causal effect estimate obtained from any version of double-shrinkage estimators, is

$$\psi_k \pm cva(c_k)\hat{a}_k \sqrt{\left(\hat{\lambda}_k^2 \sigma_{rk}^2 + (1-\hat{\lambda}_k)^2 \sigma_{ok}^2\right)},$$

where $\hat{a}_k$ and $\hat{\lambda}_k$ are the shrinkage factors, and $cva(c_k)$ is an inflation factor whose form is given in Armstrong et al. (2020).

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 squared error loss for $\hat{\tau}_r, \hat{\psi}_{\text{mle}}, \kappa_{1+}, \kappa_{2+}, \delta_1, \delta_2$.

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 squared error loss for $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\psi}}_{\textbf{mle}}, \boldsymbol{\kappa}_{1+}, \boldsymbol{\kappa}_{2+}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2$.
- 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 ("sun") $\Longrightarrow$ uncorrelated with outcome

Results

Subgroup Variable(s)	# of Strata	Loss as % of $\hat{\tau}_r$ Loss				
		$\hat{\psi}_{\text{mle}}$	κ_{1+}	κ_{2+}	δ_1	δ_2
CVD	2	**16%**	36%	36%	100%	100%
Age	3	**16%**	37%	30%	62%	73%
Sun	5	**9%**	28%	22%	40%	52%
CVD, Age	6	**21%**	39%	42%	38%	82%
CVD, Sun	10	**17%**	34%	36%	30%	87%
Age, Sun	15	**8%**	22%	21%	23%	43%
Age, CVD, Sun	30	**20%**	51%	51%	50%	78%

Further Work: Design

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

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

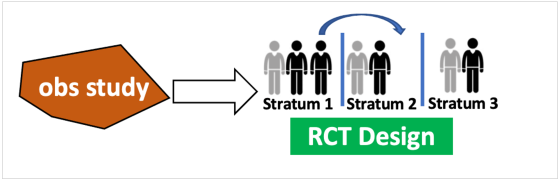
Further Work: Design

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

- Observational study already completed, $\hat{\tau}_o$ obtained.
- Designing a prospective RCT of n_r units
- Want to use a shrinker to combine $\hat{\tau}_r$ with $\hat{\tau}_o$. 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



Current & Future Work

- Current work
  - **Applied project**: air pollution and mortality. Synthesizing evidence from Medicare claims database with Medicare Current Beneficiary Survey using these estimators.
  - **Design methods**: extending to online & adaptive designs

Current & Future Work

- Current work
  - **Applied project**: air pollution and mortality. Synthesizing evidence from Medicare claims database with Medicare Current Beneficiary Survey using these estimators.
  - **Design methods**: extending to online & adaptive designs
- Future work
  - **ML approaches**: shrinkage between flexible *functional* estimates of CATEs $\hat{\tau}_{\boldsymbol{r}}(x)$ and $\hat{\tau}_{\boldsymbol{o}}(x)$
  - **Inference:** are shorter confidence intervals possible?

Acknowledgments

Thank you to my collaborators on this work:

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

The papers...

- *SURE-based procedure* paper available in Biometrics.
- *Hierarchical model* paper available at arXiv:2309.06727.
- *Design* paper available in Electronic Journal of Statistics.

Thanks!

References (I)

Armstrong, T. B., Kolesár, M., and Plagborg-Moller, M. (2020). Robust empirical bayes confidence intervals. *arXiv preprint arXiv:2004.03448*.

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.

Chen, S., Zhang, B., and Ye, T. (2021). Minimax rates and adaptivity in combining experimental and observational data. *arXiv preprint arXiv:2109.10522*.

Cheng, D. and Cai, T. (2021). Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*.

European Medicines Agency (2022). Qualification opinion for prognostic covariate adjustment (procova).

FDA, U. (2018). Framework for fda's real-world evidence program. *Silver Spring, MD: US Department of Health and Human Services Food and Drug Administration*.

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.

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.

References (II)

Lin, X. and Evans, R. J. (2023). Many data: Combine experimental and observational data through a power likelihood. *arXiv preprint arXiv:2304.02339*.

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.

Oberst, M., D'Amour, A., Chen, M., Wang, Y., Sontag, D., and Yadlowsky, S. (2022). Bias-robust integration of observational and experimental estimators. *arXiv preprint arXiv:2205.10467*.

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.

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.

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.

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.

References (III)

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.

Yang, S., Gao, C., Zeng, D., and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596.

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

A New Setting: Design

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

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

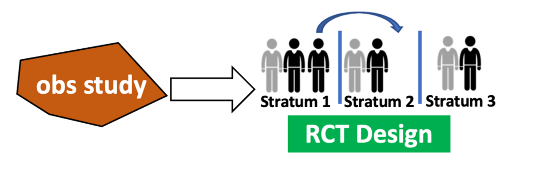
A New Setting: Design

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

- Observational study already completed, $\hat{\tau}_o$ obtained.
- Designing a prospective RCT of n_r units
- Want to use a shrinker to combine $\hat{\tau}_r$ with $\hat{\tau}_o$. 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 $\boldsymbol{\kappa}_1$ from the prior section:

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\mathrm{Tr}(\Sigma_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ ,$$

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_1$ from the prior section:

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\mathrm{Tr}(\Sigma_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ ,$$

Optimize experimental design over $\mathcal{R}_1(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_1$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_1$ from the prior section:

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\mathrm{Tr}(\Sigma_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ ,$$

Optimize experimental design over $\mathcal{R}_1(\boldsymbol{d}, \boldsymbol{V}, \xi)$, the risk of $\boldsymbol{\kappa}_1$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$

Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_1$ from the prior section:

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\mathrm{Tr}(\Sigma_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ ,$$

Optimize experimental design over $\mathcal{R}_1(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_1$ 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 $\boldsymbol{\kappa}_1$ from the prior section:

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\mathrm{Tr}(\Sigma_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ ,$$

Optimize experimental design over $\mathcal{R}_1(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_1$ 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 $\boldsymbol{\kappa}_1$ from the prior section:

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\mathrm{Tr}(\Sigma_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) ,$$

Optimize experimental design over $\mathcal{R}_1(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_1$ 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.

Design Heuristics

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

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

Design Heuristics

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

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

Design heuristics:

1. **Naïve Optimization**: Assume $\xi = 0$ and minimize $\mathcal{R}_1(\boldsymbol{d}, \hat{\boldsymbol{V}}, \xi = 0)$ over $\boldsymbol{d}$, via **greedy swap algorithm**.
2. **Robust Optimization**: Under model of Tan (2006) and a user-chosen value of sensitivity $\Gamma \geq 1$, optimize the design $\boldsymbol{d}$ under worst-case bias

1. Neyman Allocation

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

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

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

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

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 $\xi = 0$. Then solve:

$$
\begin{aligned}
\text{minimize} \quad & \mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \xi) \\
\text{subject to} \quad & \xi = 0, \boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^{K}, \\
& 0 < n_{rkt}, n_{rkc}, , \quad k = 1, \ldots, K, \\
& n_r = \sum_k n_{rkt} + n_{rkc} \, .
\end{aligned} \tag{3}
$$

But $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \xi)$ 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 4 from several values of $\boldsymbol{d}_0$ and take minimum:

$$\begin{aligned}
&\texttt{Start with design } \boldsymbol{d}_0 = \{(n^{(0)}_{rkt}, n^{(0)}_{rkc})\}_k\,. \\
&\texttt{For iteration } j = 1, 2, \ldots : \\
&\quad \texttt{For each design } \boldsymbol{d}' \texttt{ in } \mathcal{D}_{j-1}\texttt{:} \\
&\qquad \texttt{Compute } \mathcal{R}_2(\boldsymbol{d'}, \boldsymbol{V}, 0))\,. \\
&\quad \texttt{Set } \boldsymbol{d_j} = \operatorname*{argmin}_{\boldsymbol{d'} \in \mathcal{D}_{j-1}} \mathcal{R}_2(\boldsymbol{d'}, \boldsymbol{V}, 0) \\
&\quad \texttt{If } \mathcal{R}_2(\boldsymbol{d_j}, \boldsymbol{V}, 0) >= \mathcal{R}_2(\boldsymbol{d_{j-1}}, \boldsymbol{V}, 0) \\
&\qquad \texttt{Return } d_{j-1}\,.
\end{aligned} \tag{4}$$

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 $\hat{\sigma}^2_{kc}$.

$$\Gamma \implies$$

$$\boldsymbol{\xi}(\Gamma) \qquad \boldsymbol{V}(\Gamma) \implies$$

$$\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}_\Gamma, \boldsymbol{\xi}_\Gamma) \qquad \mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, 0)$$

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 $\Gamma \Longrightarrow$

collect results into $\boldsymbol{V}(\Gamma)$ and $\boldsymbol{\xi}(\Gamma) \Longrightarrow$

run Algorithm 4 using $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}(\Gamma), \boldsymbol{\xi}(\Gamma))$ instead

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

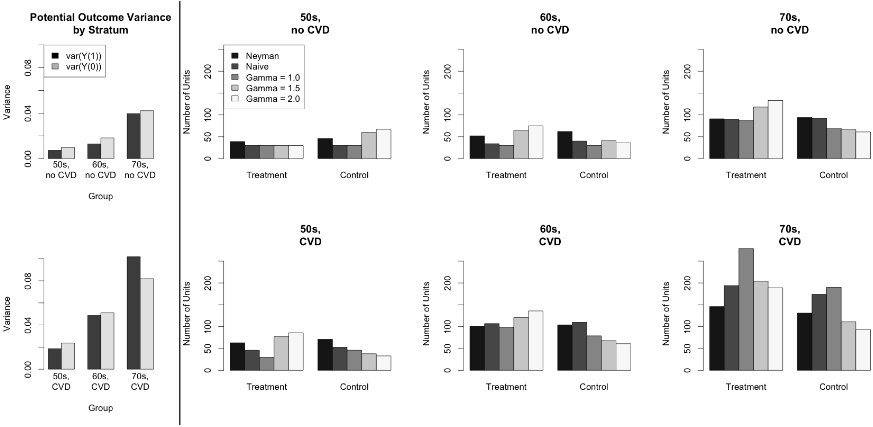


Figure 4: Allocations in WHI with strata defined by history of CVD and age, under different design heuristics.

Useful Properties of λ_1^{SURE} (I)

1. Define

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_r - \lambda_1^{\text{SURE}} (\hat{\boldsymbol{\tau}}_r - \hat{\boldsymbol{\tau}}_o)$$

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

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 $\hat{\boldsymbol{\tau}}_r$.

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

Useful Properties of $\lambda_1^{\mathrm{SURE}}$ (II)

2. Its positive part analogue,

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

where

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

satisfies the following notion of optimality:

Useful Properties of λ_1^{SURE} (III)

Theorem (κ_{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 $K \to \infty$, κ_{1+} has the lowest risk among all estimators with a shared shrinkage factor across components.

EB Coverage

- Valid confidence interval construction for shrinkage estimators is an open area of research (?)

EB Coverage

- Valid confidence interval construction for shrinkage estimators is an open area of research (?)
- Frequentist intervals shorter than standard CIs about $\hat{\tau}_r$ are impossible order-wise and difficult to obtain in practice (Chen et al., 2021).

EB Coverage

- Valid confidence interval construction for shrinkage estimators is an open area of research (?)
- Frequentist intervals shorter than standard CIs about $\hat{\tau}_r$ are impossible order-wise and difficult to obtain in practice (Chen et al., 2021).
- **EB coverage** is a frequently-used weaker condition
  - Implies **average coverage**: under fixed τ, a $1-\alpha$ fraction of effects are covered with high probability in large samples
  - However, some outlying effects may <u>not</u> be covered with $1-\alpha$ probability across repeated samples of the data