# Estimating Causal Effects with Error-Prone Exposures Using Control Variates

ENAR 2024 Spring Meeting

**Keith Barnatchez**
Biostatistics PhD Student, Harvard University
Joint work with Kevin Josey, Rachel Nethery, Giovanni Parmigiani and Bryan Shepherd

March 12th, 2024

# Roadmap

Motivation

Problem setting

The control variates method

Simulation study: brief snapshot

Discussion

# Motivation

**Central to countless observational studies**: interest in some measure(s) of the causal effect of an exposure/treatment $A$ on an outcome $Y$

# Motivation

**Central to countless observational studies**: interest in some measure(s) of the causal effect of an exposure/treatment $A$ on an outcome $Y$

- Effect of early ART initiation on 1-year post-initiation risk of suffering an AIDs-defining event

# Motivation

**Central to countless observational studies**: interest in some measure(s) of the causal effect of an exposure/treatment $A$ on an outcome $Y$

- Effect of early ART initiation on 1-year post-initiation risk of suffering an AIDs-defining event

In observational studies, it is often difficult/expensive to obtain accurate measurements of the exposure $A$

- Time of initiation often transcribed/recalled incorrectly
    - Particularly when derived from electronic health records / self-reported

# Motivation

In practice, it's often more feasible to collect error-prone measurements of $A$, denoted $A^*$, for every subject

# Motivation

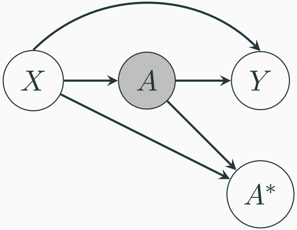
In practice, it's often more feasible to collect error-prone measurements of $A$, denoted $A^*$, for every subject

| $Y$ | $A$ | $A^*$ | $X$ |
|---|---|---|---|
| $Y_1$ | ? | $A_1^*$ | $X_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_n$ | ? | $A_n^*$ | $X_n$ |

# Motivation

In practice, it's often more feasible to collect error-prone measurements of $A$, denoted $A^*$, for every subject

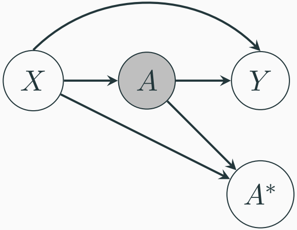
| $Y$ | $A$ | $A^*$ | $X$ |
|---|---|---|---|
| $Y_1$ | ? | $A_1^*$ | $X_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_n$ | ? | $A_n^*$ | $X_n$ |



# Motivation

In practice, it's often more feasible to collect error-prone measurements of $A$, denoted $A^*$, for every subject

| $Y$ | $A$ | $A^*$ | $X$ |
|---|---|---|---|
| $Y_1$ | ? | $A_1^*$ | $X_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_n$ | ? | $A_n^*$ | $X_n$ |



Growing literature on the perils of exposure measurement error in causal inference (Valeri 2021)

- Using $A^*$ in place of $A$ tends to produce biased effect estimates
- Difficult to correct for this bias without information on the measurement error mechanism
    - Requires *design-based* approaches that collect supplemental data

# Addressing measurement error via study design

One design-based workaround:

- Spend additional time + resources to obtain gold-standard measurements of $A$ for a **small subset** of the study data

# Addressing measurement error via study design

One design-based workaround:

- Spend additional time + resources to obtain gold-standard measurements of $A$ for a **small subset** of the study data
  - E.g. through manual chart review or follow-up interviews
  - Typically referred to as a *double sampling* study design (Hidiroglou 2001) or *validation study*
  - Usually infeasible to validate *every* subject (otherwise, would be no need for this talk)

# Addressing measurement error via study design

One design-based workaround:

- Spend additional time + resources to obtain gold-standard measurements of $A$ for a **small subset** of the study data
    - E.g. through manual chart review or follow-up interviews
    - Typically referred to as a *double sampling* study design (Hidiroglou 2001) or *validation study*
    - Usually infeasible to validate *every* subject (otherwise, would be no need for this talk)
- The subset of data with gold-standard measurements is typically referred to as the *validation data*
    - Intuition: provides complete information for a subset of data, and provides insight into measurement error mechanism

# Roadmap

Motivation

Problem setting

The control variates method

Simulation study: brief snapshot

Discussion

# Exposure measurement error data structure

**Causal estimand**: $\tau \overset{\text{def}}{=} \mathbb{E}[Y(1) - Y(0)]$ (average treatment effect)

# Exposure measurement error data structure

**Causal estimand**: $\tau \overset{\text{def}}{=} \mathbb{E}[Y(1) - Y(0)]$ (average treatment effect)

| $Y$ | $A$ | $A^*$ | $X$ | $S$ |
|---|---|---|---|---|
| $Y_1$ | $A_1$ | $A_1^*$ | $X_1$ | 1 |
| $Y_2$ |  | $A_2^*$ | $X_2$ | 0 |
| $Y_3$ |  | $A_3^*$ | $X_3$ | 0 |
| $Y_4$ | $A_4$ | $A_4^*$ | $X_4$ | 1 |
| $Y_5$ |  | $A_5^*$ | $X_5$ | 0 |
| $Y_6$ | $A_6$ | $A_6^*$ | $X_6$ | 1 |

**(a)** Main dataset

# Exposure measurement error data structure

**Causal estimand**: $\tau \overset{\text{def}}{=} \mathbb{E}[Y(1) - Y(0)]$ (average treatment effect)

| $Y$ | $A$ | $A^*$ | $X$ | $S$ |
|---|---|---|---|---|
| $Y_1$ | $A_1$ | $A_1^*$ | $X_1$ | 1 |
| $Y_2$ |  | $A_2^*$ | $X_2$ | 0 |
| $Y_3$ |  | $A_3^*$ | $X_3$ | 0 |
| $Y_4$ | $A_4$ | $A_4^*$ | $X_4$ | 1 |
| $Y_5$ |  | $A_5^*$ | $X_5$ | 0 |
| $Y_6$ | $A_6$ | $A_6^*$ | $X_6$ | 1 |

**(a)** Main dataset

| $Y$ | $A$ | $A^*$ | $X$ | $S$ |
|---|---|---|---|---|
| $Y_1$ | $A_1$ | $A_1^*$ | $X_1$ | 1 |
| $Y_4$ | $A_4$ | $A_4^*$ | $X_4$ | 1 |
| $Y_6$ | $A_6$ | $A_6^*$ | $X_6$ | 1 |

**(b)** Validation dataset

# Assumptions

Make the standard causal inference assumptions: **consistency**, **positivity** and **unconfoundedness**

Additionally, assume the validation data is obtained completely at random: $S \perp\!\!\!\perp (Y, A, A^*, \boldsymbol{X})$

- Can be enforced by design in EHR data settings
- This can be relaxed to allow for more flexible validation sampling schemes/study designs

# Estimation wishlist

Ideally, we'd like an estimator that

- Is **(1)** unbiased, **(2)** model-agnostic, and **(3)** efficient

# Estimation wishlist

Ideally, we'd like an estimator that

- Is **(1)** unbiased, **(2)** model-agnostic, and **(3)** efficient

Notice we can achieve **(1)** and **(2)** by employing doubly-robust methods (e.g. AIPW) on only the validation data

- But this approach completely ignores the remaining observations!
  - Unbiased, but highly inefficient as validation samples are typically small

# Estimation wishlist

Ideally, we'd like an estimator that

- Is **(1)** unbiased, **(2)** model-agnostic, and **(3)** efficient

Notice we can achieve **(1)** and **(2)** by employing doubly-robust methods (e.g. AIPW) on only the validation data

- But this approach completely ignores the remaining observations!
  - Unbiased, but highly inefficient as validation samples are typically small

Our approach is adapted from Yang and Ding (2019)

- **Idea**: Improve the efficiency of an initial unbiased (but inefficient) estimator of $\tau$ by **augmenting** it with a variance reduction term formed from the full data

# Roadmap

Motivation

Problem setting

The control variates method

Simulation study: brief snapshot

Discussion

### Step 1: obtain validation data only estimator

| $Y$ | $A$ | $A^*$ | $\boldsymbol{X}$ |
|---|---|---|---|
| $Y_1$ | $A_1$ | $A_1^*$ | $X_1$ |
| $Y_3$ | $A_3$ | $A_3^*$ | $X_3$ |
| $Y_5$ | $A_5$ | $A_5^*$ | $X_5$ |

Validation data only estimate: $\hat{\tau}_{\text{val}}$

### Step 2: construct the control variate

| $Y$ | $A$ | $A^*$ | $\boldsymbol{X}$ |
|---|---|---|---|
| $Y_1$ | $A_1$ | $A_1^*$ | $X_1$ |
| $Y_2$ |  | $A_2^*$ | $X_2$ |
| $Y_3$ | $A_3$ | $A_3^*$ | $X_3$ |
| $Y_4$ |  | $A_4^*$ | $X_4$ |
| $Y_5$ | $A_5$ | $A_5^*$ | $X_5$ |
| $Y_6$ |  | $A_6^*$ | $X_6$ |

Error-prone estimate: $\hat{\tau}^{\text{e.p.}}_{\text{main}}$

| $Y$ | $A$ | $A^*$ | $\boldsymbol{X}$ |
|---|---|---|---|
| $Y_1$ | $A_1$ | $A_1^*$ | $X_1$ |
| $Y_3$ | $A_3$ | $A_3^*$ | $X_3$ |
| $Y_5$ | $A_5$ | $A_5^*$ | $X_5$ |

Error-prone estimate: $\hat{\tau}^{\text{e.p.}}_{\text{val}}$

### Step 3: Compute the variance reduction term

Obtain $\hat{\Gamma} = \widehat{\text{Cov}}(\hat{\tau}_{\text{val}}, \hat{\tau}^{\text{e.p.}}_{\text{main}} - \hat{\tau}^{\text{e.p.}}_{\text{val}})$ and $\hat{V} = \widehat{\text{Var}}(\hat{\tau}^{\text{e.p.}}_{\text{main}} - \hat{\tau}^{\text{e.p.}}_{\text{val}})$

### Step 4: Form the final estimator

Obtain final estimate: $\hat{\tau}_{\text{CV}} = \hat{\tau}_{\text{val}} - \hat{\Gamma}\hat{V}^{-1}(\hat{\tau}^{\text{e.p.}}_{\text{main}} - \hat{\tau}^{\text{e.p.}}_{\text{val}})$

# Control variates method: properties

**Efficiency gain**: $\mathsf{Var}(\hat{\tau}_{\mathsf{CV}}) = \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) - \Gamma^2/V$, where

- $\Gamma$ is the covariance between $\hat{\tau}_{\mathsf{val}}$ and the control variate
- $V$ is the variance of the control variate

# Control variates method: properties

**Efficiency gain**: $\mathsf{Var}(\hat{\tau}_{\mathsf{CV}}) = \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) - \Gamma^2/V$, where

- $\Gamma$ is the covariance between $\hat{\tau}_{\mathsf{val}}$ and the control variate
- $V$ is the variance of the control variate

**Double robustness**: Consistent if either the outcome model or propensity score model is correctly specified

# Control variates method: properties

**Efficiency gain**: $\mathsf{Var}(\hat{\tau}_{\mathsf{CV}}) = \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) - \Gamma^2/V$, where

- $\Gamma$ is the covariance between $\hat{\tau}_{\mathsf{val}}$ and the control variate
- $V$ is the variance of the control variate

**Double robustness**: Consistent if either the outcome model or propensity score model is correctly specified

**Flexible estimation of nuisance models**:

- $\hat{\tau}_{\mathsf{CV}} \xrightarrow{p} \tau$ at $\sqrt{n}$ rates

# Control variates method: properties

**Efficiency gain**: $\text{Var}(\hat{\tau}_{\text{CV}}) = \text{Var}(\hat{\tau}_{\text{val}}) - \Gamma^2/V$, where

- $\Gamma$ is the covariance between $\hat{\tau}_{\text{val}}$ and the control variate
- $V$ is the variance of the control variate

**Double robustness**: Consistent if either the outcome model or propensity score model is correctly specified

**Flexible estimation of nuisance models**:

- $\hat{\tau}_{\text{CV}} \xrightarrow{p} \tau$ at $\sqrt{n}$ rates
- Even if the nuisance models are estimated with ML methods that themselves have slower rates of convergence

# Control variates method: properties

**Efficiency gain**: $\mathsf{Var}(\hat{\tau}_{\mathsf{CV}}) = \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) - \Gamma^2/V$, where

- $\Gamma$ is the covariance between $\hat{\tau}_{\mathsf{val}}$ and the control variate
- $V$ is the variance of the control variate

**Double robustness**: Consistent if either the outcome model or propensity score model is correctly specified

**Flexible estimation of nuisance models**:

- $\hat{\tau}_{\mathsf{CV}} \xrightarrow{p} \tau$ at $\sqrt{n}$ rates
- Even if the nuisance models are estimated with ML methods that themselves have slower rates of convergence
    - Common property of “doubly-robust” estimators

# Control variates method: extensions

The control variates method is **flexible** – can account for...

- More general validation data sampling schemes / account for multiple study sites
- Simultaneous error in the outcome of interest

# Control variates method: extensions

The control variates method is **flexible** – can account for...

- More general validation data sampling schemes / account for multiple study sites
- Simultaneous error in the outcome of interest
- Other causal estimands
    - E.g. **local average treatment effects** if one has access to an instrumental variable

# Roadmap

Motivation

Problem setting

The control variates method

Simulation study: brief snapshot

Discussion

# Brief snapshot of simulation study

Compared the control variates estimator to

- An oracle estimator (know $A$ for the entire dataset, estimate $\tau$ with AIPW) and a naive estimator that uses $A^*$ in place of $A$
- A validation data only estimator
- Multiple imputation
  - Standard method for performing causal inference with error-prone exposures

# Brief snapshot of simulation study

Compared performances of each estimator under:

# Brief snapshot of simulation study

Compared performances of each estimator under:

- Varying degrees of **exposure misclassification**

# Brief snapshot of simulation study

Compared performances of each estimator under:

- Varying degrees of **exposure misclassification**

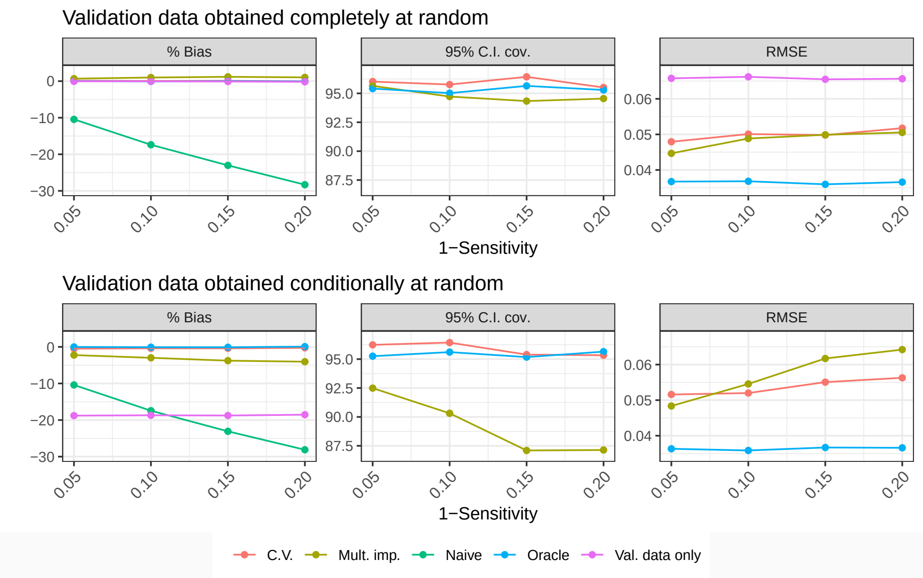
- Two different **sampling schemes** for the validation data
  1. Obtained completely at random
  2. Obtained conditionally (on $\boldsymbol{X}$) at random

# Brief snapshot of simulation study

Compared performances of each estimator under:

- Varying degrees of **exposure misclassification**
- Two different **sampling schemes** for the validation data
  1. Obtained completely at random
  2. Obtained conditionally (on $\boldsymbol{X}$) at random
- Varying the **relative size** of the validation data, $\mathbb{P}(S = 1)$
  - For this talk, $\mathbb{P}(S = 1) = 0.3$

# Brief snapshot of simulation study



# Roadmap

Motivation

Problem setting

The control variates method

Simulation study: brief snapshot

Discussion

# Discussion

- Growing set of methods for accounting for measurement error in causal inference

# Discussion

- Growing set of methods for accounting for measurement error in causal inference
- Control variates method enjoys the **flexibility** of methods like multiple imputation / regression calibration...

# Discussion

- Growing set of methods for accounting for measurement error in causal inference
- Control variates method enjoys the **flexibility** of methods like multiple imputation / regression calibration...
    - with some additional theoretical properties commonly associated with traditional “doubly-robust” estimators

# Discussion

- Growing set of methods for accounting for measurement error in causal inference
- Control variates method enjoys the **flexibility** of methods like multiple imputation / regression calibration...
    - with some additional theoretical properties commonly associated with traditional “doubly-robust” estimators

⚠ The control variates method isn’t appropriate for every measurement error problem

- Requires the availability of /ability to obtain a validation dataset
- As always, care should be taken to assess the plausibility of the causal assumptions

Thank you!

keithbarnatchez@g.harvard.edu

Working paper coming soon!

# References

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694.

Hidiroglou, M. (2001). Double sampling. *Survey methodology*, 27(2):143–154.

Valeri, L. (2021). Measurement error in causal inference. In *Handbook of Measurement Error Models*, pages 453–480. Chapman and Hall/CRC.

Yang, S. and Ding, P. (2019). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*.

Zeng, Z., Kennedy, E. H., Bodnar, L. M., and Naimi, A. I. (2023). Efficient generalization and transportation. *arXiv preprint arXiv:2302.00092*.

# Variance Reduction

Notice

$$\begin{aligned}\mathsf{Var}(\hat{\tau}_{\mathsf{CV}}) &= \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) + b^2\mathsf{Var}(\hat{\tau}_{\mathsf{main,ep}} - \hat{\tau}_{\mathsf{val,ep}}) - 2b\mathsf{Cov}(\hat{\tau}_{\mathsf{val}}, \hat{\tau}_{\mathsf{main,ep}} - \hat{\tau}_{\mathsf{val,ep}}) \\ &= \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) + b^2 V - 2b\Gamma\end{aligned}$$

Minimzing with respect to $b$ yields

$$b = \Gamma V^{-1}$$

Implying with this choice of $b$,

$$\mathsf{Var}(\hat{\tau}_{\mathsf{CV}}) = \mathsf{Var}(\hat{\tau}_{\mathsf{val}}) - \Gamma^2 V^{-1}$$

Back to control variates slide

# Covariate-dependent selection

Back to main CV method

In many realistic scenarios, validation data won't just be a random draw from main dataset

- When that's the case, naively implementing C.V. method will actually *add* bias
- Intuition is that $\hat{\tau}_{\text{val,e.p.}} \overset{p}{\not\to} \hat{\tau}_{\text{main,e.p.}}$ when $\boldsymbol{X} \not\perp\!\!\!\perp S$ (distributions of effect modifiers are different)
- On top of that, our validation-data only estimator $\hat{\tau}_{\text{val}}$ will be subject to *external validity bias*

This implies we need to explore ways to

- Adjust $\hat{\tau}_{\text{val}}$ so that it targets the ATE in the population of interest
- Adjust $\hat{\tau}_{\text{val,e.p.}}$ in the same manner

# Covariate-dependent selection

In the 2 study *generalizability* setting, Dahabreh et al. (2019) and Zeng et al. (2023) have proposed the following doubly-robust estimator for $\tau$:

$$\hat{\psi}_a = \sum_{i=1}^{n} \left[ \frac{I(A_i = a, S_i = 1)(Y_i - \hat{\mu}_a(\boldsymbol{X}_i))}{\hat{\rho}(\boldsymbol{X}_i)\hat{\pi}_a(\boldsymbol{X}_i)} + \hat{\mu}_a(\boldsymbol{X}_i) \right]$$

- $\hat{\rho}(\boldsymbol{X}_i)$ is estimated probability of "selection" into val. data
- $\hat{\mu}_a(\boldsymbol{X}_i) = \hat{\mathbb{E}}(Y|\boldsymbol{X}_i, A = a, S = 1)$, $\hat{\pi}_a(\boldsymbol{X}_i) = \hat{\mathbb{P}}(A_i = a|\boldsymbol{X}_i, S_i = 1)$
- $\hat{\tau}$ is obtained by taking the difference $\hat{\psi}_1 - \hat{\psi}_0$

# Control variates method

1. Using the validation dataset, obtain an estimate of $\tau$, denoted $\hat{\tau}_{\text{val}}$

# Control variates method

1. Using the validation dataset, obtain an estimate of $\tau$, denoted $\hat{\tau}_{\text{val}}$
2. Using $A^*$, obtain *error-prone* estimates of $\tau$ in the main and validation datasets, denoted $\hat{\tau}_{\text{val,ep}}$ and $\hat{\tau}_{\text{main,ep}}$ respectively

# Control variates method

1. Using the validation dataset, obtain an estimate of $\tau$, denoted $\hat{\tau}_{\text{val}}$
2. Using $A^*$, obtain *error-prone* estimates of $\tau$ in the main and validation datasets, denoted $\hat{\tau}_{\text{val,ep}}$ and $\hat{\tau}_{\text{main,ep}}$ respectively
3. Under the earlier causal assumptions, will have

$$\sqrt{n_{\text{val}}}\begin{pmatrix} \hat{\tau}_{\text{val}} - \tau \\ \hat{\tau}_{\text{main,ep}} - \hat{\tau}_{\text{val,ep}} \end{pmatrix} \xrightarrow{D} N\left(\mathbf{0}, \mathbf{\Sigma}\right), \quad \mathbf{\Sigma} = \begin{pmatrix} v & \Gamma \\ \Gamma & V \end{pmatrix}$$

# Control variates method

1. Using the validation dataset, obtain an estimate of $\tau$, denoted $\hat{\tau}_{\text{val}}$
2. Using $A^*$, obtain *error-prone* estimates of $\tau$ in the main and validation datasets, denoted $\hat{\tau}_{\text{val,ep}}$ and $\hat{\tau}_{\text{main,ep}}$ respectively
3. Under the earlier causal assumptions, will have

$$\sqrt{n_{\text{val}}}\begin{pmatrix} \hat{\tau}_{\text{val}} - \tau \\ \hat{\tau}_{\text{main,ep}} - \hat{\tau}_{\text{val,ep}} \end{pmatrix} \xrightarrow{D} N\left(\mathbf{0}, \mathbf{\Sigma}\right), \quad \mathbf{\Sigma} = \begin{pmatrix} v & \Gamma \\ \Gamma & V \end{pmatrix}$$

where we can construct estimators of the form

$$\hat{\tau}_{\text{CV}} = \hat{\tau}_{\text{val}} - b(\hat{\tau}_{\text{main,ep}} - \hat{\tau}_{\text{val,ep}}),$$

# Control variates method

1. Using the validation dataset, obtain an estimate of $\tau$, denoted $\hat{\tau}_{\text{val}}$
2. Using $A^*$, obtain *error-prone* estimates of $\tau$ in the main and validation datasets, denoted $\hat{\tau}_{\text{val,ep}}$ and $\hat{\tau}_{\text{main,ep}}$ respectively
3. Under the earlier causal assumptions, will have

$$\sqrt{n_{\text{val}}}\begin{pmatrix} \hat{\tau}_{\text{val}} - \tau \\ \hat{\tau}_{\text{main,ep}} - \hat{\tau}_{\text{val,ep}} \end{pmatrix} \overset{D}{\to} N\left(\mathbf{0}, \mathbf{\Sigma}\right), \quad \mathbf{\Sigma} = \begin{pmatrix} v & \Gamma \\ \Gamma & V \end{pmatrix}$$

where we can construct estimators of the form

$$\hat{\tau}_{\text{CV}} = \hat{\tau}_{\text{val}} - b(\hat{\tau}_{\text{main,ep}} - \hat{\tau}_{\text{val,ep}}),$$

setting $b = \Gamma V^{-1}$ so that $\text{Var}(\hat{\tau}_{\text{CV}}) \leq \text{Var}(\hat{\tau}_{\text{val}})$ Finding $b$