Confidently Comparing Estimates with the c-value

Brian Trippe, Sameer K. Deshpande, Tamara Broderick

CSAIL

MIT Computational & Systems Biology

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach

- Share strength across similar schools

- Region
- Enrollment size
- Type (Catholic, charter, public)
- ...

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach

- Share strength across similar schools
- Estimate school performances by the posterior mean

- Region
- Enrollment size
- Type (Catholic, charter, public)
- ...

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach

- Share strength across similar schools
- Estimate school performances by the posterior mean [Hoff, 2020]
  - 5-50 students tested at 676 schools

- Region
- Enrollment size
- Type (Catholic, charter, public)
- ...

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach [Lindley and Smith, 1972, Rubin, 1981, Gelman et al., 2013]

- Share strength across similar schools
- Estimate school performances by the posterior mean [Hoff, 2020]
  - 5-50 students tested at 676 schools

- Region
- Enrollment size
- Type (Catholic, charter, public)
- ...

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach [Lindley and Smith, 1972, Rubin, 1981, Gelman et al., 2013]

- Share strength across similar schools
- Estimate school performances by the posterior mean [Hoff, 2020]
  - 5-50 students tested at 676 schools
- Limitation: complexity, subjectivity of the prior

- Region
- Enrollment size
- Type (Catholic, charter, public)
- ...

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach [Lindley and Smith, 1972, Rubin, 1981, Gelman et al., 2013]

- Share strength across similar schools
- Estimate school performances by the posterior mean [Hoff, 2020]
  - 5-50 students tested at 676 schools
- Limitation: complexity, subjectivity of the prior
- **Question:** is this more accurate than simple averages?

- Region
- Enrollment size
- Type (Catholic, charter, public)
- . . .

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach [Lindley and Smith, 1972, Rubin, 1981, Gelman et al., 2013]

- Share strength across similar schools
- Estimate school performances by the posterior mean [Hoff, 2020]
  - 5-50 students tested at 676 schools
- Limitation: complexity, subjectivity of the prior
- **Question:** is this more accurate than simple averages?

- This question comes up in many new analyses!

- Region
- Enrollment size
- Type (Catholic, charter, public)
- . . .

Are Complex Statistical Methods Worth the Fuss?

Learning from Educational Testing Data

- National Center for Education Statistics gathers standardized tests from U.S. high schools
- Want to know school-level performances
- Standardized tests on small sample of students for each school – Noisy!

Hierarchical Bayesian Approach [Lindley and Smith, 1972, Rubin, 1981, Gelman et al., 2013]

- Share strength across similar schools
- Estimate school performances by the posterior mean [Hoff, 2020]
  - 5-50 students tested at 676 schools
- Limitation: complexity, subjectivity of the prior
- **Question:** is this more accurate than simple averages? **Yes (c=99.26%)!**
- This question comes up in many new analyses!

- Region
- Enrollment size
- Type (Catholic, charter, public)
- ...

Roadmap

- Justifying complexity
- Methods for choosing methods
- The c-value as a measure of confidence (our method)
- How and when we can compute c-values
- Application to educational testing
- Extensions to nonlinear models and estimates

Roadmap

- Justifying complexity
- **Methods for choosing methods**
- The c-value as a measure of confidence (our method)
- How and when we can compute c-values
- Application to educational testing
- Extensions to nonlinear models and estimates

Methods for choosing Methods

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot; \theta)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Standard criterion – Risk $R(\theta,\hat{\theta}(\cdot)) := \mathbb{E}_{y\sim p(\cdot;\theta)}\left[L(\theta,\hat{\theta}(y))\right]$

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta, \cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Standard criterion – Risk $R(\theta, \hat{\theta}(\cdot)) := \mathbb{E}_{y\sim p(\cdot;\theta)}\left[L(\theta, \hat{\theta}(y))\right]$

- Choose $\hat{\theta}(y)$ over $\theta^*(y)$ if $R(\theta, \hat{\theta}(\cdot)) < R(\theta, \theta^*(\cdot))$

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta, \cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Standard criterion – Risk $R(\theta, \hat{\theta}(\cdot)) := \mathbb{E}_{y\sim p(\cdot;\theta)}\left[L(\theta, \hat{\theta}(y))\right]$

- Choose $\hat{\theta}(y)$ over $\theta^*(y)$ if $R(\theta, \hat{\theta}(\cdot)) < R(\theta, \theta^*(\cdot))$
- Risk depends on θ!

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta, \cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Standard criterion – Risk $R(\theta, \hat{\theta}(\cdot)) := \mathbb{E}_{y \sim p(\cdot;\theta)} \left[L(\theta, \hat{\theta}(y)) \right]$

- Choose $\hat{\theta}(y)$ over $\theta^*(y)$ if $R(\theta, \hat{\theta}(\cdot)) < R(\theta, \theta^*(\cdot))$
- Risk depends on θ!
- I care about **my** y

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Standard criterion – Risk $R(\theta,\hat{\theta}(\cdot)) := \mathbb{E}_{y\sim p(\cdot;\theta)}\left[L(\theta,\hat{\theta}(y))\right]$

- Choose $\hat{\theta}(y)$ over $\theta^*(y)$ if $R(\theta,\hat{\theta}(\cdot)) < R(\theta,\theta^*(\cdot))$
- Risk depends on θ!
- I care about **my** y

Goal: Measure of confidence that $\theta^*(\cdot)$ has smaller loss than $\hat{\theta}(\cdot)$

Methods for choosing Methods

In Machine Learning

- Information criteria (AIC / BIC)
- Cross-validation

$\Big\}$ Apply to prediction, not parameter estimation!

Decision Theory

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta, \cdot)$

Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Standard criterion – Risk $R(\theta, \hat{\theta}(\cdot)) := \mathbb{E}_{y\sim p(\cdot;\theta)}\left[L(\theta, \hat{\theta}(y))\right]$

- Choose $\hat{\theta}(y)$ over $\theta^*(y)$ if $R(\theta, \hat{\theta}(\cdot)) < R(\theta, \theta^*(\cdot))$
- Risk depends on θ!
- I care about **my** y

Goal: Measure of confidence that $\theta^*(\cdot)$ has smaller loss than $\hat{\theta}(\cdot)$

1. On the observed dataset
2. Without needing subjective assumptions about θ

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$
Estimates: $\hat{\theta}(y)$ vs. $\theta^*(y)$

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta,y)}_{\text{Win}} := L(\theta,\hat{\theta}(y)) - L(\theta,\theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta,y) > 0$

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta, \cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta, y)}_{\text{Win}} := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta, y) > 0$

Challenge: identify if $W(\theta, y) > 0$

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta,y)}_{\text{Win}} := L(\theta,\hat{\theta}(y)) - L(\theta,\theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta,y) > 0$

Challenge: identify if $W(\theta,y) > 0$

Our approach

Introduce high probability lower bound on the win, $b(y,\alpha)$

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta,y)}_{\text{Win}} := L(\theta,\hat{\theta}(y)) - L(\theta,\theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta,y) > 0$

Challenge: identify if $W(\theta,y) > 0$

Our approach

Introduce high probability lower bound on the win, $b(y,\alpha)$

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_{y\sim p(\cdot;\theta)}\left[W(\theta,y) > b(y,\alpha)\right] \geq \alpha$

Introducing the c-value

Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta, \cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta, y)}_{\text{Win}} := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta, y) > 0$

Challenge: identify if $W(\theta, y) > 0$

Our approach

Introduce high probability lower bound on the win, $b(y, \alpha)$

Condition: For any θ and $\alpha \in [0, 1]$, $\mathbb{P}_{y\sim p(\cdot;\theta)}\left[W(\theta, y) > b(y, \alpha)\right] \geq \alpha$

- 95% confidence in $\theta^*(y)$ if $b(y, 0.95) > 0$

Introducing the c-value

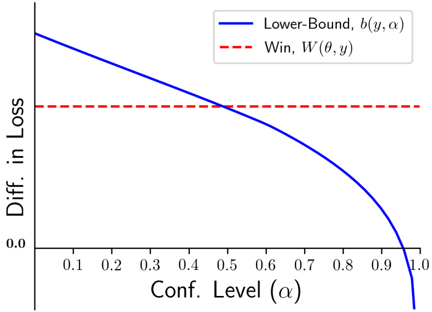
Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta,y)}_{\text{Win}} := L(\theta,\hat{\theta}(y)) - L(\theta,\theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta,y) > 0$

Challenge: identify if $W(\theta,y) > 0$



Our approach

Introduce high probability lower bound on the win, $b(y,\alpha)$

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_{y\sim p(\cdot;\theta)}\left[W(\theta,y) > b(y,\alpha)\right] \geq \alpha$

- 95% confidence in $\theta^*(y)$ if $b(y,0.95) > 0$

Introducing the c-value

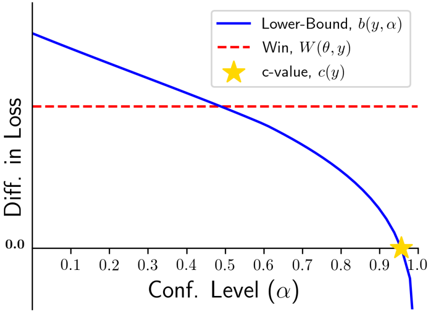
Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat\theta(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta,y)}_{\text{Win}} := L(\theta,\hat\theta(y)) - L(\theta,\theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta,y) > 0$

Challenge: identify if $W(\theta,y) > 0$



Our approach

Introduce high probability lower bound on the win, $b(y,\alpha)$

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_{y\sim p(\cdot;\theta)}\left[W(\theta,y) > b(y,\alpha)\right] \geq \alpha$

- 95% confidence in $\theta^*(y)$ if $b(y,0.95) > 0$

Introducing the c-value

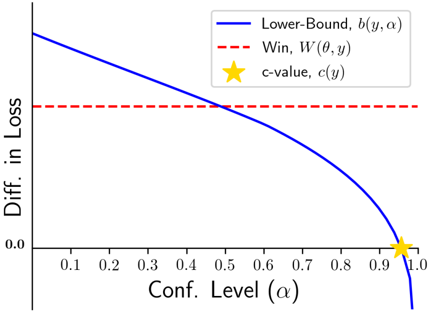
Model: $y \sim p(\cdot;\theta)$ Loss: $L(\theta,\cdot)$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$

$$\underbrace{W(\theta,y)}_{\text{Win}} := L(\theta,\hat{\theta}(y)) - L(\theta,\theta^*(y))$$

- Prefer $\theta^*(y)$ iff $W(\theta,y) > 0$

Challenge: identify if $W(\theta,y) > 0$



Our approach

Introduce high probability lower bound on the win, $b(y,\alpha)$

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_{y\sim p(\cdot;\theta)}\left[W(\theta,y) > b(y,\alpha)\right] \geq \alpha$

- 95% confidence in $\theta^*(y)$ if $b(y,0.95) > 0$

Define $\mathbf{c}(\mathbf{y}) := \inf_{\alpha\in[\mathbf{0},\mathbf{1}]}\left\{\alpha|\mathbf{b}(\mathbf{y},\alpha) \leq \mathbf{0}\right\}$

- Loosely, largest level α below which $b(y,\alpha) > 0$

The c-value as an Informative Measure of Confidence

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0, 1]$, $\mathbb{P}_\theta \left[W(\theta, y) > b(y, \alpha) \right] \geq \alpha$

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[W(\theta, y) > b(y, \alpha)\right] \geq \alpha$

The c-value as a quantification of confidence

Thm 1: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[c(y) > \alpha \text{ and } W(\theta, y) \leq 0\right] \leq 1-\alpha$

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[W(\theta, y) > b(y,\alpha)\right] \geq \alpha$

The c-value as a quantification of confidence

Thm 1: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[c(y) > \alpha \text{ and } W(\theta, y) \leq 0\right] \leq 1-\alpha$

- Unlikely that both (A.) c-value is close to 1 **and** (B.) $\theta^*(y)$ is not more accurate than $\hat{\theta}(y)$

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[W(\theta, y) > b(y, \alpha)\right] \geq \alpha$

The c-value as a quantification of confidence

Thm 1: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[c(y) > \alpha \text{ and } W(\theta, y) \leq 0\right] \leq 1 - \alpha$

- Unlikely that both (A.) c-value is close to 1 **and** (B.) $\theta^*(y)$ is not more accurate than $\hat{\theta}(y)$

Using $c(y)$ to choose between $\hat{\theta}(y)$ and $\theta^*(y)$

Define two-stage estimator

$$\theta^\dagger(y, \alpha) := \mathbb{1}[c(y) > \alpha]\theta^*(y) + \mathbb{1}[c(y) \leq \alpha]\hat{\theta}(y)$$

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[W(\theta, y) > b(y, \alpha)\right] \geq \alpha$

The c-value as a quantification of confidence

Thm 1: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[c(y) > \alpha \text{ and } W(\theta, y) \leq 0\right] \leq 1-\alpha$

- Unlikely that both (A.) c-value is close to 1 **and** (B.) $\theta^*(y)$ is not more accurate than $\hat{\theta}(y)$

Using $c(y)$ to choose between $\hat{\theta}(y)$ and $\theta^*(y)$

Define two-stage estimator

$$\theta^\dagger(y, \alpha) := \mathbb{1}[c(y) > \alpha]\theta^*(y) + \mathbb{1}[c(y) \leq \alpha]\hat{\theta}(y)$$

- Deviates from default only when confident in alternative

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[W(\theta,y) > b(y,\alpha)\right] \geq \alpha$

The c-value as a quantification of confidence

Thm 1: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[c(y) > \alpha \text{ and } W(\theta,y) \leq 0\right] \leq 1-\alpha$

- Unlikely that both (A.) c-value is close to 1 **and** (B.) $\theta^*(y)$ is not more accurate than $\hat{\theta}(y)$

Using $c(y)$ to choose between $\hat{\theta}(y)$ and $\theta^*(y)$

Define two-stage estimator

$$\theta^\dagger(y,\alpha) := \mathbb{1}[c(y) > \alpha]\theta^*(y) + \mathbb{1}[c(y) \leq \alpha]\hat{\theta}(y)$$

- Deviates from default only when confident in alternative

Thm 2: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[L(\theta,\theta^\dagger(y,\alpha)) > L(\theta,\hat{\theta}(y))\right] \leq 1-\alpha$

The c-value as an Informative Measure of Confidence

Condition: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[W(\theta, y) > b(y, \alpha)\right] \geq \alpha$

The c-value as a quantification of confidence

Thm 1: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[c(y) > \alpha \text{ and } W(\theta, y) \leq 0\right] \leq 1 - \alpha$

- Unlikely that both (A.) c-value is close to 1 **and** (B.) $\theta^*(y)$ is not more accurate than $\hat{\theta}(y)$

Using $c(y)$ to choose between $\hat{\theta}(y)$ and $\theta^*(y)$

Define two-stage estimator

$$\theta^\dagger(y, \alpha) := \mathbb{1}[c(y) > \alpha]\theta^*(y) + \mathbb{1}[c(y) \leq \alpha]\hat{\theta}(y)$$

- Deviates from default only when confident in alternative

Thm 2: For any θ and $\alpha \in [0,1]$, $\mathbb{P}_\theta\left[L(\theta, \theta^\dagger(y, \alpha)) > L(\theta, \hat{\theta}(y))\right] \leq 1 - \alpha$

- If we report $\theta^*(y)$ only when $c(y) > 0.95$, we do worse than $\hat{\theta}(\cdot)$ at

Roadmap

- Justifying complexity
- Methods for choosing methods
- The c-value as a measure of confidence
- How and when we can compute c-values
- Application to educational testing
- Extensions to nonlinear models and estimates

Roadmap

- Justifying complexity
- Methods for choosing methods
- The c-value as a measure of confidence
- **How and when we can compute c-values**
- Application to educational testing
- Extensions to nonlinear models and estimates

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot;\theta)$ $\qquad$ $L(\theta, \theta'(y))$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\qquad$ $W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot; \theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot; \theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot; \theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

$$W(\theta, y) = \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2$$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot; \theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown $\boldsymbol{\theta}$**, and **$\boldsymbol{L(\theta, \hat{\theta})}$ and $\boldsymbol{L(\theta, \theta^*)}$ are dependent**

$$W(\theta, y) = \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2$$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot; \theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown $\boldsymbol{\theta}$**, and **$\boldsymbol{L(\theta, \hat{\theta})}$ and $\boldsymbol{L(\theta, \theta^*)}$ are dependent**

$$
\begin{aligned}
W(\theta, y) &= \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\
&= \|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2 + 2\langle \hat{\theta}(y) - \theta^*(y), y - \theta \rangle
\end{aligned}
$$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot;\theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown $\boldsymbol{\theta}$**, and **$\boldsymbol{L(\theta, \hat{\theta})}$ and $\boldsymbol{L(\theta, \theta^*)}$ are dependent**

$$
\begin{aligned}
W(\theta, y) =& \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\
=& \underbrace{\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2}_{\text{observed and computable}} + \underbrace{2\langle \hat{\theta}(y) - \theta^*(y), y - \theta \rangle}_{\text{unobserved}(\ddagger)}
\end{aligned}
$$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot;\theta)$ with $\theta, y \in \mathbb{R}^N$ $L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown $\boldsymbol{\theta}$**, and **$\boldsymbol{L(\theta, \hat{\theta})}$ and $\boldsymbol{L(\theta, \theta^*)}$ are dependent**

$$
\begin{aligned}
W(\theta, y) =& \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\
=& \underbrace{\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2}_{\text{observed and computable}} + \underbrace{2\langle \hat{\theta}(y) - \theta^*(y), y - \theta\rangle}_{\text{unobserved}(\ddagger)} \\
\geq& \text{ observed} + F_{\ddagger}^{-1}(1 - \alpha; \theta), \text{with prob. } \alpha
\end{aligned}
$$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot;\theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown** $\boldsymbol{\theta}$, and $\boldsymbol{L(\theta, \hat{\theta})}$ **and** $\boldsymbol{L(\theta, \theta^*)}$ **are dependent**

$$
\begin{aligned}
W(\theta, y) =& \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\
=& \underbrace{\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2}_{\text{observed and computable}} + \underbrace{2\langle \hat{\theta}(y) - \theta^*(y), y - \theta \rangle}_{\text{unobserved}(\ddagger)} \\
\geq& \text{ observed} + F_{\ddagger}^{-1}(1 - \alpha; g(\theta)), \text{with prob. } \alpha
\end{aligned}
$$

- Key observation: $F_{\ddagger}$ depends on θ only through a scalar $g(\theta)$

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot; \theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1 - \alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown $\boldsymbol{\theta}$**, and **$\boldsymbol{L(\theta, \hat{\theta})}$ and $\boldsymbol{L(\theta, \theta^*)}$ are dependent**

$$
\begin{aligned}
W(\theta, y) &= \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\
&= \underbrace{\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2}_{\text{observed and computable}} + \underbrace{2\langle \hat{\theta}(y) - \theta^*(y), y - \theta \rangle}_{\text{unobserved}(\ddagger)} \\
&\geq \text{observed} + F_{\ddagger}^{-1}(1 - \alpha; g(\theta)), \text{with prob. } \alpha \\
&\geq \text{observed} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right), \text{with prob. } \alpha
\end{aligned}
$$

- Key observation: $F_{\ddagger}$ depends on θ only through a scalar $g(\theta)$
- Split excess α across interval and quantile (union bound)

Constructing $b(y, \alpha)$, the High-Confidence Lower Bound

Model: $y \sim p(\cdot;\theta)$ with $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y)}_{\text{Default}}$ vs. $\underbrace{\theta^*(y)}_{\text{Alternative}}$ $\quad W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$

Idea: Use $1-\alpha$ lower quantile of $W(\theta, y)$

- Relies on **unknown $\boldsymbol{\theta}$**, and **$\boldsymbol{L(\theta, \hat{\theta})}$ and $\boldsymbol{L(\theta, \theta^*)}$ are dependent**

$$
\begin{aligned}
W(\theta, y) =& \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\
=& \underbrace{\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2}_{\text{observed and computable}} + \underbrace{2\langle \hat{\theta}(y) - \theta^*(y), y - \theta\rangle}_{\text{unobserved}(\ddagger)} \\
\geq& \text{ observed} + F_{\ddagger}^{-1}(1-\alpha; g(\theta)), \text{with prob. } \alpha \\
\geq& \underbrace{\text{observed} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)}_{b(y,\alpha)}, \text{with prob. } \alpha
\end{aligned}
$$

- Key observation: $F_{\ddagger}$ depends on θ only through a scalar $g(\theta)$
- Split excess α across interval and quantile (union bound)

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \text{observed} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger =$ unobserved

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^N y_n$

Filling out the details of the bound

$$b(y, \alpha) = \|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2 + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger =$ unobserved

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1}\sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = -\left\|\frac{y + \tau^{-2}\mathbf{1}_N\bar{y}}{1 + \tau^{-2}} - y\right\|^2 + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger =$ unobserved

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \dfrac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1}\sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1+\tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger =$ unobserved

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1 + \tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger =$ **unobserved**

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1 + \tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger = 2\langle \hat{\theta}(y) - \theta^*(y), y - \theta \rangle$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1 + \tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger = \frac{2}{1+\tau^2}\langle y - \bar{y}\mathbf{1}_N, y - \theta\rangle$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1}\sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1 + \tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(\frac{1}{4}\|\theta - \bar{\theta}\mathbf{1}_N\|^2) - \frac{1}{4}\|\theta - \bar{\theta}\mathbf{1}_N\|^2\right]$

- $\chi^2_{N-1}(\lambda)$ is non-central χ^2 with $N - 1$ degrees of freedom, non-centrality λ

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1+\tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1}\sum_{n=1}^N y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1+\tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(g(\theta)) - g(\theta)\right]$, $g(\theta) = \frac{1}{4}\|\theta - \bar{\theta}\mathbf{1}_N\|^2$

- $\chi^2_{N-1}(\lambda)$ is non-central χ^2 with $N-1$ degrees of freedom, non-centrality λ

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1 + \tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(g(\theta)) - g(\theta)\right]$, $g(\theta) = \frac{1}{4}\|\theta - \bar{\theta}\mathbf{1}_N\|^2$

- $\chi^2_{N-1}(\lambda)$ is non-central χ^2 with $N - 1$ degrees of freedom, non-centrality λ
- Interval $C(y, 1 - \alpha)$ for $g(\theta)$ from $\|y - \bar{y}\mathbf{1}_N\|^2 \sim \chi^2_{N-1}(4g(\theta))$

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1} \sum_{n=1}^{N} y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1 + \tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(g(\theta)) - g(\theta)\right]$, $g(\theta) = \frac{1}{4}\|\theta - \bar{\theta}\mathbf{1}_N\|^2$

- $\chi^2_{N-1}(\lambda)$ is non-central χ^2 with $N-1$ degrees of freedom, non-centrality λ
- Interval $C(y, 1-\alpha)$ for $g(\theta)$ from $\|y - \bar{y}\mathbf{1}_N\|^2 \sim \chi^2_{N-1}(4g(\theta))$

Take-aways: Correct coverage by construction

Example Bound – The Lindley and Smith [1972] Estimator

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, I_N)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\underbrace{\hat{\theta}(y) = y}_{\text{Default (MLE)}}$ vs. $\underbrace{\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1+\tau^{-2}}}_{\text{Alternative (Lindley and Smith)}}$

- $\theta^*(y)$ is a classic Bayes estimate, shrinks towards mean
- $\tau > 0$ and $\bar{y} := N^{-1}\sum_{n=1}^N y_n$

Filling out the details of the bound

$$b(y, \alpha) = \frac{-\|y - \bar{y}\mathbf{1}_N\|^2}{(1+\tau^2)^2} + \inf_{\lambda \in C(y, \frac{1-\alpha}{2})} F_{\ddagger}^{-1}\left(\frac{1-\alpha}{2}; g(\theta) = \lambda\right)$$

where $\ddagger \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(g(\theta)) - g(\theta)\right]$, $g(\theta) = \frac{1}{4}\|\theta - \bar{\theta}\mathbf{1}_N\|^2$

- $\chi^2_{N-1}(\lambda)$ is non-central χ^2 with $N-1$ degrees of freedom, non-centrality λ
- Interval $C(y, 1-\alpha)$ for $g(\theta)$ from $\|y - \bar{y}\mathbf{1}_N\|^2 \sim \chi^2_{N-1}(4g(\theta))$

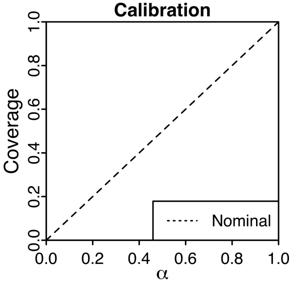
Take-aways: Correct coverage by construction, Computable

Example Bound – Simulation Results

- Use simulated data for calibration, power, and risk

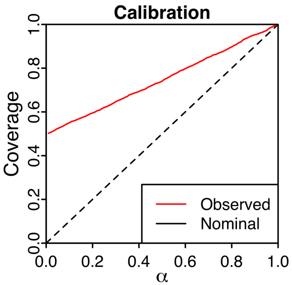
Example Bound – Simulation Results

- Use simulated data for calibration, power, and risk



Example Bound – Simulation Results

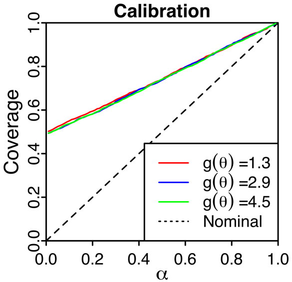
- Use simulated data for calibration, power, and risk



- $b(y, \alpha)$ is conservative across levels α and θ

Example Bound – Simulation Results

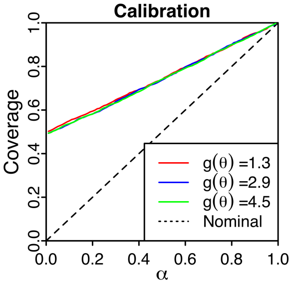
- Use simulated data for calibration, power, and risk



- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$

Example Bound – Simulation Results

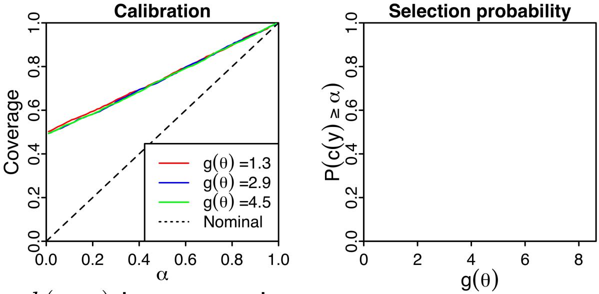
- Use simulated data for calibration, power, and risk



- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence

Example Bound – Simulation Results

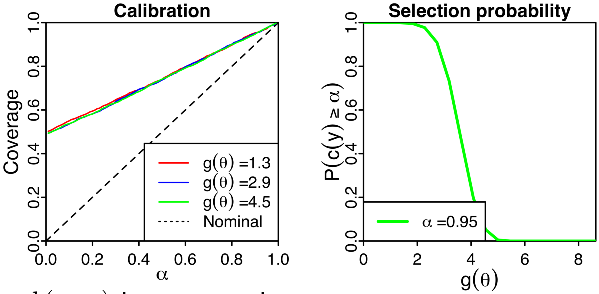
- Use simulated data for calibration, power, and risk



- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence

Example Bound – Simulation Results

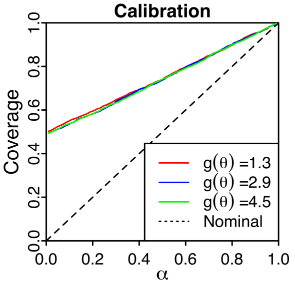
- Use simulated data for calibration, power, and risk

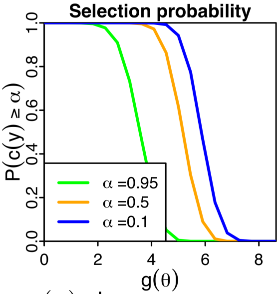


- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence
- $c(y)$ detects $W(\theta, y) > 0$

Example Bound – Simulation Results

- Use simulated data for calibration, power, and risk



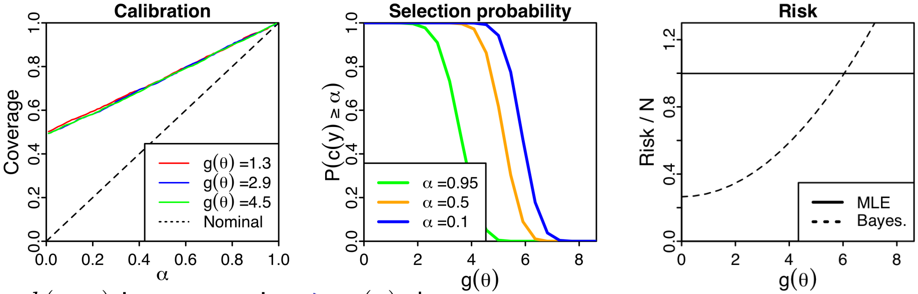


- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence

- $c(y)$ detects $W(\theta, y) > 0$
- More frequent selection of θ^* for smaller α

Example Bound – Simulation Results

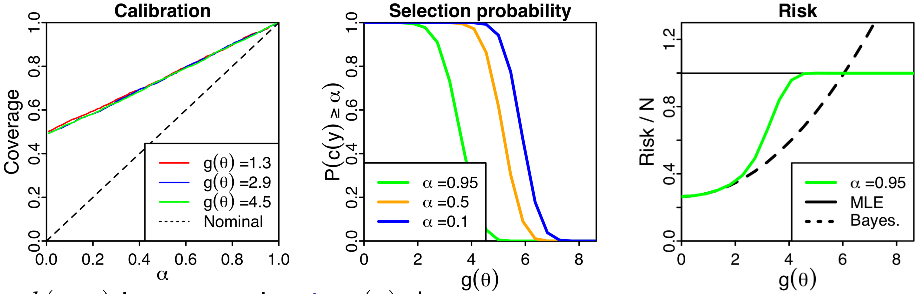
- Use simulated data for calibration, power, and risk



- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence
- $c(y)$ detects $W(\theta, y) > 0$
- More frequent selection of θ^* for smaller α

Example Bound – Simulation Results

- Use simulated data for calibration, power, and risk



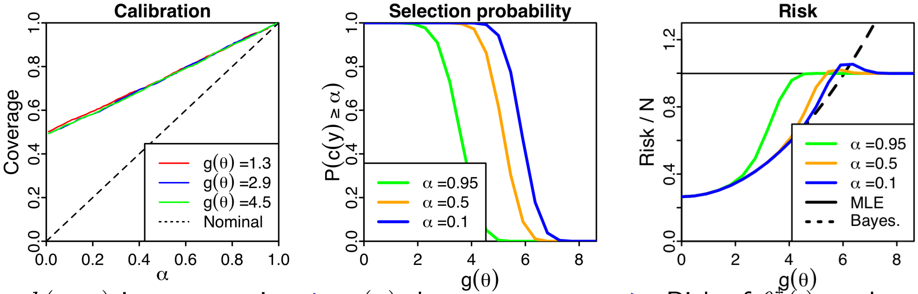
- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence

- $c(y)$ detects $W(\theta, y) > 0$
- More frequent selection of θ^* for smaller α

- Risk of $\theta^\dagger(\cdot)$ trades off between $\hat{\theta}(\cdot)$ and $\theta^*(\cdot)$

Example Bound – Simulation Results

- Use simulated data for calibration, power, and risk



- $b(y, \alpha)$ is conservative across levels α and θ
- Distribution of $W(\theta, y)$ and $b(y, \alpha)$ depend only on $g(\theta)$
- Coverage has little θ dependence

- $c(y)$ detects $W(\theta, y) > 0$
- More frequent selection of θ^* for smaller α

- Risk of $\theta^\dagger(\cdot)$ trades off between $\hat{\theta}(\cdot)$ and $\theta^*(\cdot)$
- For small α, $\theta^\dagger$ can do worse

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y_n \overset{indep}{\sim} \mathcal{N}\left(\theta_n, \sigma^2/\text{size}_n\right)$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\mathsf{size}_1, \ldots, \sigma^2/\mathsf{size}_N)$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \dots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\text{size}_1, \dots, \sigma^2/\text{size}_N)$
  - Estimate τ, β, σ by empirical Bayes (`lme4`)

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\mathsf{size}_1, \ldots, \sigma^2/\mathsf{size}_N)$
  - Estimate τ, β, σ by empirical Bayes (`lme4`)
- Alternative $\theta^*(y) = [I_N + \tau^{-2}\Sigma]^{-1} y + [I_N + \tau^2 \Sigma^{-1}]^{-1} X\beta$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\mathsf{size}_1, \ldots, \sigma^2/\mathsf{size}_N)$
  - Estimate τ, β, σ by empirical Bayes (`lme4`)
- Alternative $\theta^*(y) = [I_N + \tau^{-2}\Sigma]^{-1} y + [I_N + \tau^2 \Sigma^{-1}]^{-1} X\beta$
  - $\theta^*(y)$ is an affine transformation of y

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$ for $A, C \in \mathbb{R}^{N \times N}$ $\; k, \ell \in \mathbb{R}^N$

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$ for $A, C \in \mathbb{R}^{N \times N}$ $k, \ell \in \mathbb{R}^N$

- Applications: Gaussian process kernel selection, shrinkage estimation, linear regression

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$ for $A, C \in \mathbb{R}^{N \times N}$ $\; k, \ell \in \mathbb{R}^N$

- Applications: Gaussian process kernel selection, shrinkage estimation, linear regression

$$b(y,\alpha) := \|\hat{\theta} - y\|^2 - \|\theta^* - y\|^2 + 2\mathrm{tr}[(A-C)\Sigma] +$$

$$2z_{\frac{1-\alpha}{2}} \sqrt{U\left(\frac{1-\alpha}{2}\right) + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2}$$

where

$$U(1-\alpha) := \inf_{\delta > 0}\left\{\delta \,\middle|\, \|\hat{\theta}(y) - \theta^*(y)\|_\Sigma^2 \le (\delta + \|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_F^2) + \right.$$

$$\left. z_{1-\alpha}\sqrt{2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\delta}\right\}$$

is a high confidence upper bound on $g(\theta) := \left\|(A-C)\theta + (k-\ell)\right\|_\Sigma^2$

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$ for $A, C \in \mathbb{R}^{N \times N}$ $\; k, \ell \in \mathbb{R}^N$

- Applications: Gaussian process kernel selection, shrinkage estimation, linear regression

$$b(y, \alpha) := \|\hat{\theta} - y\|^2 - \|\theta^* - y\|^2 + 2\mathrm{tr}[(A - C)\Sigma] +$$

$$2z_{\frac{1-\alpha}{2}} \sqrt{U\left(\frac{1-\alpha}{2}\right) + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2}$$

where

$$U(1-\alpha) := \inf_{\delta > 0} \left\{ \delta \;\middle|\; \|\hat{\theta}(y) - \theta^*(y)\|_\Sigma^2 \le (\delta + \|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_F^2) + \right.$$

$$\left. z_{1-\alpha}\sqrt{2\|\Sigma^{\frac{1}{2}}(A - C)\Sigma(A - C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\delta} \right\}$$

is a high confidence upper bound on $g(\theta) := \left\|(A - C)\theta + (k - \ell)\right\|_\Sigma^2$

- Computable : `c(y) = c_value(y, Σ, A, k, C, l)`

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$ for $A, C \in \mathbb{R}^{N\times N}$ $\; k, \ell \in \mathbb{R}^N$

- Applications: Gaussian process kernel selection, shrinkage estimation, linear regression

$$b(y,\alpha) := \|\hat{\theta} - y\|^2 - \|\theta^* - y\|^2 + 2\mathrm{tr}[(A-C)\Sigma] +$$

$$2z_{\frac{1-\alpha}{2}}\sqrt{U\left(\frac{1-\alpha}{2}\right) + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2}$$

where

$$U(1-\alpha) := \inf_{\delta>0}\left\{\delta \,\middle|\, \|\hat{\theta}(y) - \theta^*(y)\|_\Sigma^2 \le (\delta + \|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_F^2) + \right.$$

$$\left. z_{1-\alpha}\sqrt{2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\delta}\right\}$$

is a high confidence upper bound on $g(\theta) := \left\|(A-C)\theta + (k-\ell)\right\|_\Sigma^2$

- Computable : `c(y) = c_value(y, Σ, A, k, C, l)`
- Some analytical challenges:
  - Non-asymptotic error control [Berry, 1941]

Example Bound 2 – Affine Estimates & Correlated Noise

Model: $\theta, y \in \mathbb{R}^N$ with $y \sim \mathcal{N}(\theta, \Sigma)$ $\quad L(\theta, \theta'(y)) = \|\theta'(y) - \theta\|^2$

Estimates: $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$ for $A, C \in \mathbb{R}^{N \times N}$ $\ k, \ell \in \mathbb{R}^N$

- Applications: Gaussian process kernel selection, shrinkage estimation, linear regression

$$b(y, \alpha) := \|\hat{\theta} - y\|^2 - \|\theta^* - y\|^2 + 2\mathrm{tr}[(A - C)\Sigma] +$$

$$2z_{\frac{1-\alpha}{2}} \sqrt{U\left(\frac{1-\alpha}{2}\right) + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2}$$

where

$$U(1-\alpha) := \inf_{\delta > 0} \left\{ \delta \;\middle|\; \|\hat{\theta}(y) - \theta^*(y)\|_\Sigma^2 \leq (\delta + \|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_F^2) + \right.$$

$$\left. z_{1-\alpha} \sqrt{2\|\Sigma^{\frac{1}{2}}(A - C)\Sigma(A - C)^\top \Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2 \delta} \right\}$$

is a high confidence upper bound on $g(\theta) := \left\|(A - C)\theta + (k - \ell)\right\|_\Sigma^2$

- Computable : `c(y) = c_value(y, Σ, A, k, C, l)`
- Some analytical challenges:
  - Non-asymptotic error control [Berry, 1941]
  - Conservatism

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\text{size}_1, \ldots, \sigma^2/\text{size}_N)$
- Alternative $\theta^*(y) = [I_N + \tau^{-2}\Sigma]^{-1}\, y + [I_N + \tau^2\Sigma^{-1}]^{-1} X\beta$

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\text{size}_1, \ldots, \sigma^2/\text{size}_N)$
- Alternative $\theta^*(y) = [I_N + \tau^{-2}\Sigma]^{-1}\, y + [I_N + \tau^2\Sigma^{-1}]^{-1} X\beta$

Compute `c_value(y, Σ, A, k, C, l)`

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \dots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$ $[\ A = I_N, k = 0\ ]$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\mathsf{size}_1, \dots, \sigma^2/\mathsf{size}_N)$
- Alternative $\theta^*(y) = \underbrace{[I_N + \tau^{-2}\Sigma]^{-1}}_{C} y + \underbrace{[I_N + \tau^2\Sigma^{-1}]^{-1} X\beta}_{\ell}$

Compute `c_value(y, Σ, A, k, C, l)`

Educational Testing – Estimate from [Hoff, 2020]

Educational Longitudinal Study (2002–2012)

- Standardized test of reading ability in 10th grade students
- Sample of 5-50 students at $N = 676$ schools ($y = [y_1, \ldots, y_N]$)
- **Goal:** estimate school-specific means, $\theta \in \mathbb{R}^N$
- Default: $\hat{\theta}(y) = y$ $[\ A = I_N, k = 0\ ]$

Small area inference [Fay and Herriot, 1979, Hoff, 2020]

- $D = 8$ features of each school (region, school type, enrollment, ...)
- Prior: $\theta_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \beta, \tau^2)$ for $\beta \in \mathbb{R}^D$
- Likelihood: $y \sim \mathcal{N}\big(\theta, \Sigma\big)$ for $\Sigma = \mathrm{diag}(\sigma^2/\mathsf{size}_1, \ldots, \sigma^2/\mathsf{size}_N)$
- Alternative $\theta^*(y) = \underbrace{[I_N + \tau^{-2}\Sigma]^{-1}}_{C} y + \underbrace{[I_N + \tau^2 \Sigma^{-1}]^{-1} X\beta}_{\ell}$

Compute `c_value(y, Σ, A, k, C, l)`**=0.9926**

Beyond Affine Estimates & Gaussian Noise

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood
- **We show:** our bounds provide nominal coverage as sample size $\rightarrow \infty$

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood
- **We show:** our bounds provide nominal coverage as sample size $\rightarrow \infty$

Many estimates are approximately affine

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood
- **We show:** our bounds provide nominal coverage as sample size $\rightarrow \infty$

Many estimates are approximately affine

- Empirical Bayes

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood
- **We show:** our bounds provide nominal coverage as sample size $\rightarrow \infty$

Many estimates are approximately affine

- Empirical Bayes
- E.g. James–Stein estimator: $\theta^*_{\mathsf{JS}}(y) = \left(1 - \frac{N-2}{\|y\|^2}\right) y$

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood
- **We show:** our bounds provide nominal coverage as sample size $\rightarrow \infty$

Many estimates are approximately affine

- Empirical Bayes
- E.g. James–Stein estimator: $\theta^*_{\mathsf{JS}}(y) = \left(1 - \frac{N-2}{\|y\|^2}\right) y$
- **We show:** our bounds provide nominal coverage as dimension $N \rightarrow \infty$

Beyond Affine Estimates & Gaussian Noise

Many likelihoods are approximately Gaussian

- E.g. Logistic Regression
- Asymptotic normality of MLE $\rightarrow$ Gaussian approximation to likelihood
- **We show:** our bounds provide nominal coverage as sample size $\rightarrow \infty$

Many estimates are approximately affine

- Empirical Bayes
- E.g. James–Stein estimator: $\theta^*_{\mathsf{JS}}(y) = \left(1 - \frac{N-2}{\|y\|^2}\right) y$
- **We show:** our bounds provide nominal coverage as dimension $N \rightarrow \infty$

Open Directions:

1. Different losses - L1, zero-one
2. Different models - sparse regression
3. Tighter bounds – overly conservative

Summary

- We proposed c-values to frequentist confidence in new estimates
  - on the observed dataset
  - without assumptions on θ
- Our bounds cover a range of models & estimates for squared error
- We demonstrate conclusive evaluations on real problems

Further Information

Trippe, Brian L., Sameer K. Deshpande, and Tamara Broderick. "Confidently Comparing Estimators with the c-value." *Journal of the American Statistical Association* (2023).

Code Available: github.com/blt2114/c_values

Contact me: btrippe@mit.edu

References

Andrew C Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

Robert E Fay and Roger A Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

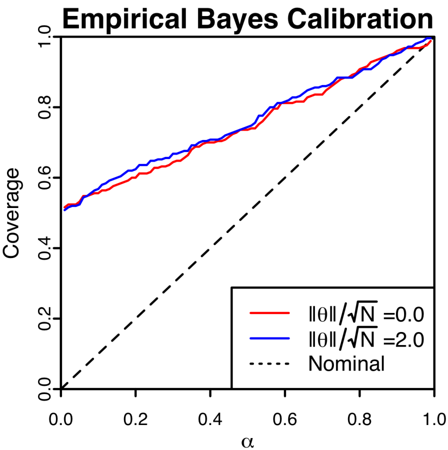
Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

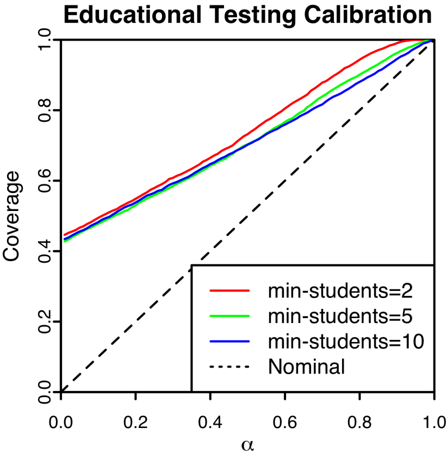
Peter D Hoff. Smaller p-values via indirect information. *Journal of the American Statistical Association*, pages 1–35, 2020.

Dennis V Lindley and Adrian FM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B*, 34(1):1–18, 1972.

Donald B Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

Coverage of Empirical Bayes





- James-Stein estimator vs. MLE coverage

- Educational testing application, coverage in simulation with empirical Bayes step