Empirical Bayes When Estimation Precision Predicts Parameters

with applications to economic mobility

Jiafeng (Kevin) Chen

Harvard Economics | Harvard Business School

April 26, 2024

arXiv:2212.14444

R package: `github.com/jiafengkevinchen/close`

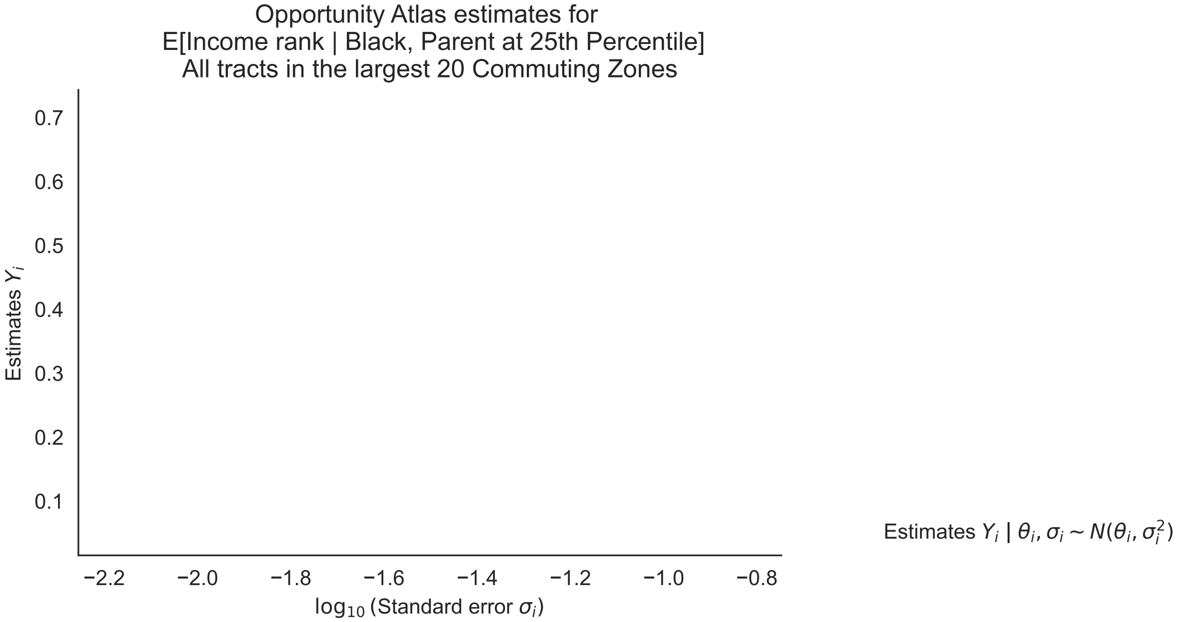
Empirical Bayes (EB) when estimation precision predicts parameters

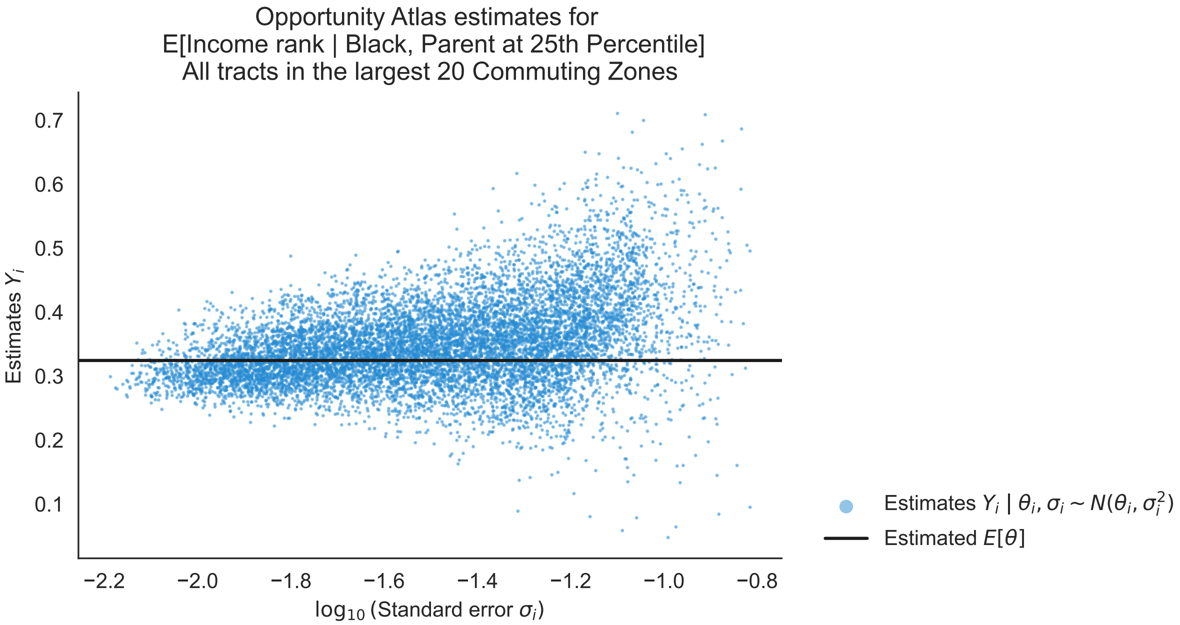
- **Empirical Bayes** methods are popular for **improving** data-driven **economic decisions**
  - → Want to make decisions, but only have noisy estimates
  - → EB methods better recover true parameters from noisy estimates
  - → **Ex**: better learn true neighborhood quality, true teacher value-added, firm-level discrimination, . . .
  - → Leading special case is shrinkage (shrink noisy estimates to the mean of estimates)
  - → Shrinkage estimates = posterior mean under an estimated prior (hence empirical Bayes)
- Conventional EB methods embed a **prior independence** assumption
  - → Precision of estimates does not predict true parameters
  - → Economically questionable, statistically rejected
- Imposing **prior independence can harm EB methods**
  - → Shrinks to the wrong target and sometimes worse than doing nothing
- **Contributions**: (1) new EB methods that generalize; (2) prove theoretical guarantees
  - → Normalize away potential dependence and apply best-in-class existing methods

Motivating empirical example (neighborhood characteristics)

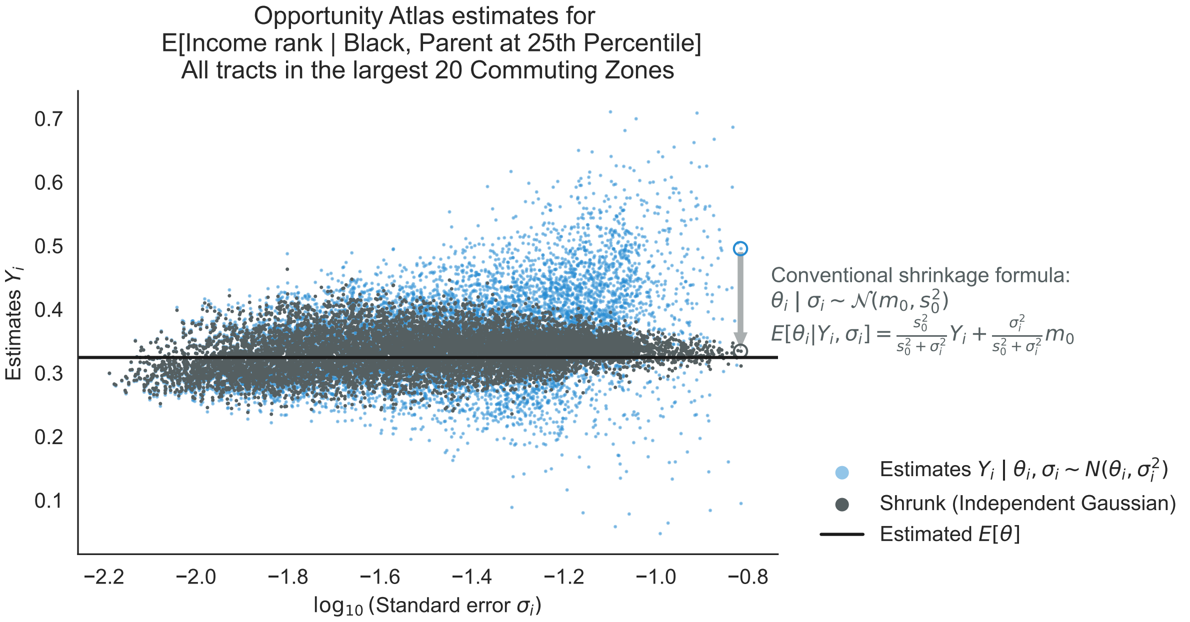
- To illustrate, the Opportunity Atlas (Chetty et al., 2020) produces economic mobility $\underbrace{\text{estimates}}_{Y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)}$ with $\underbrace{\text{standard errors}}_{\sigma_i}$ for $\underbrace{\text{true economic mobility}}_{\theta_i}$ at the $\underbrace{\text{Census tract}}_{i}$ level
- **Example decision problem**: select high mobility Census tracts
  - → Bergman et al. (2024) selected top $\frac{1}{3}$, nudged low-income households to move
  - → Mathematically: Observe (Y_i, σ_i), want to pick out the high θ_i's
  - → Pick units with high shrinkage estimates (EB posterior means) $\rightsquigarrow$ better selections
- For this problem, $\overbrace{\textbf{prior independence}}^{\theta_i \perp\!\!\!\perp \sigma_i}$ might cause bad selections: Consider $\theta_i = \mathbb{E}\,[\text{Income rank} \mid \text{Black, Parents@P25, Tract } i]$ (race-spec. version of Bergman et al.'s target)
  - → **Why** $\theta_i \not\perp\!\!\!\perp \sigma_i$**:** Lower $\sigma_i \leftrightsquigarrow$ More poor Black families in i (sample size) $\leftrightsquigarrow$ Lower mobility
    - $\rightsquigarrow$ Predicts positive correlation between σ_i and θ_i

Bergman et al's target also violates PI (mildly after covariates) | Empirical evidence for (% black or % poor)-on-mobility correlation

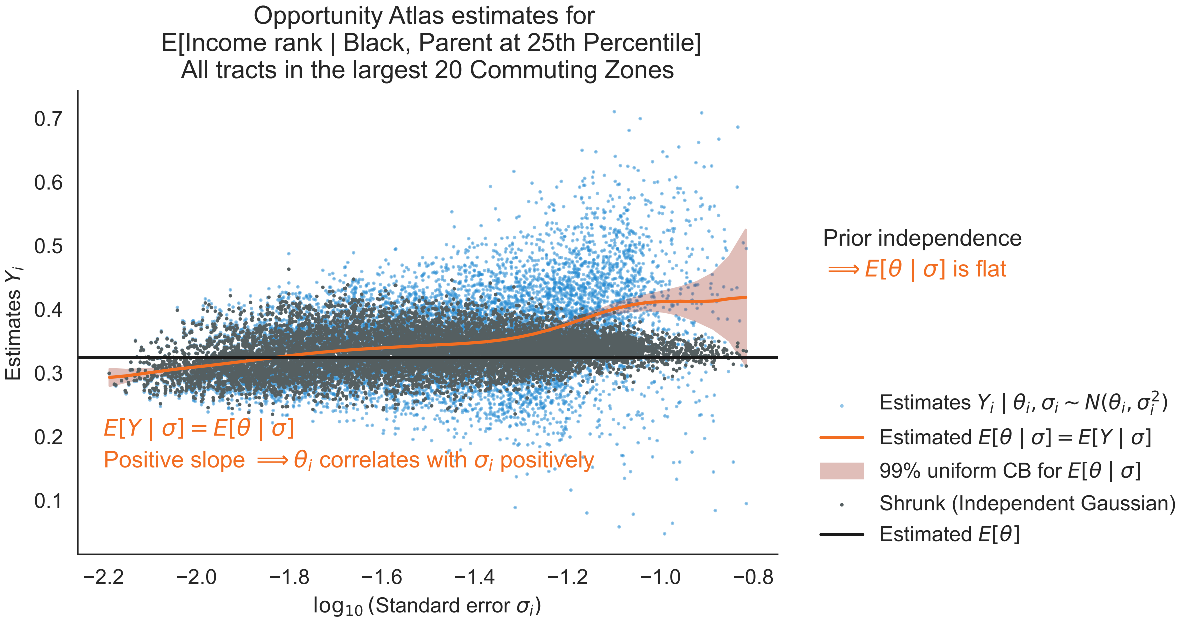

Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
Estimates Y_i
0.7
0.6
0.5
0.4
0.3
0.2
0.1
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
$\log_{10}$(Standard error σ_i)
Estimates $Y_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2)$



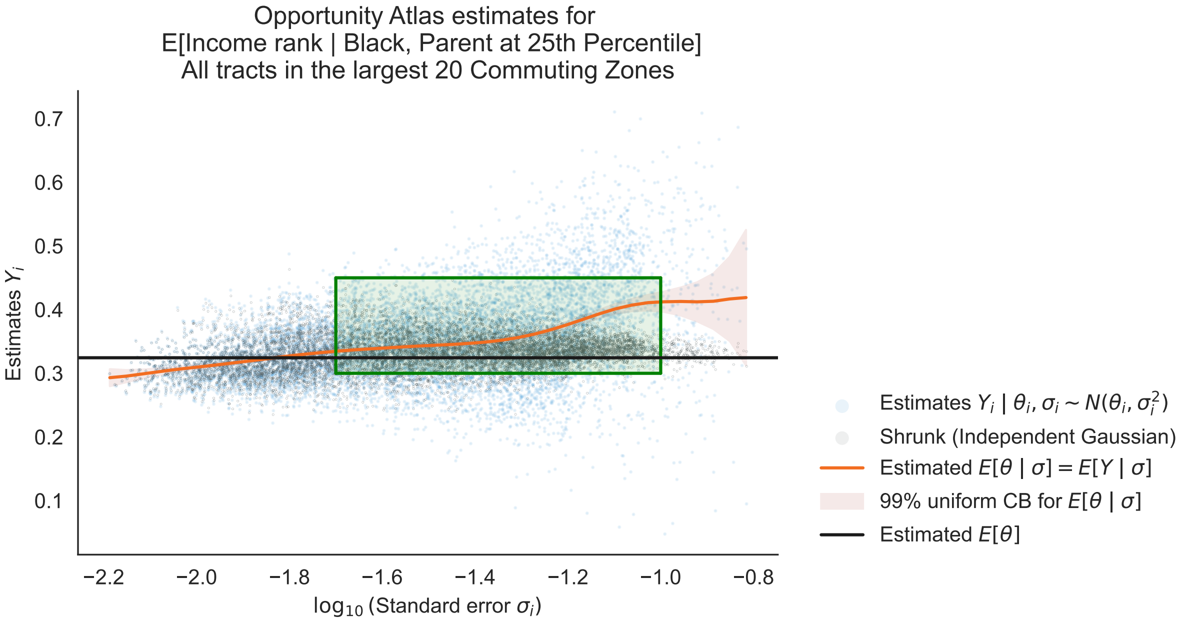
Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
Estimates Y_i
0.7
0.6
0.5
0.4
0.3
0.2
0.1
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
$\log_{10}$ (Standard error σ_i)
Estimates $Y_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2)$
Estimated $E[\theta]$

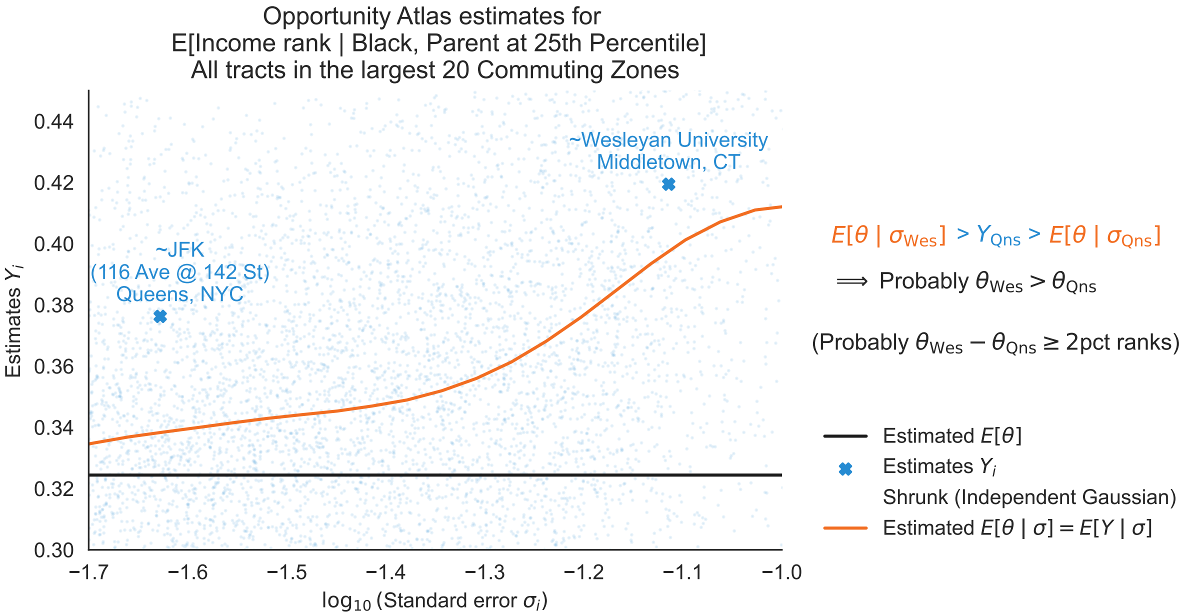


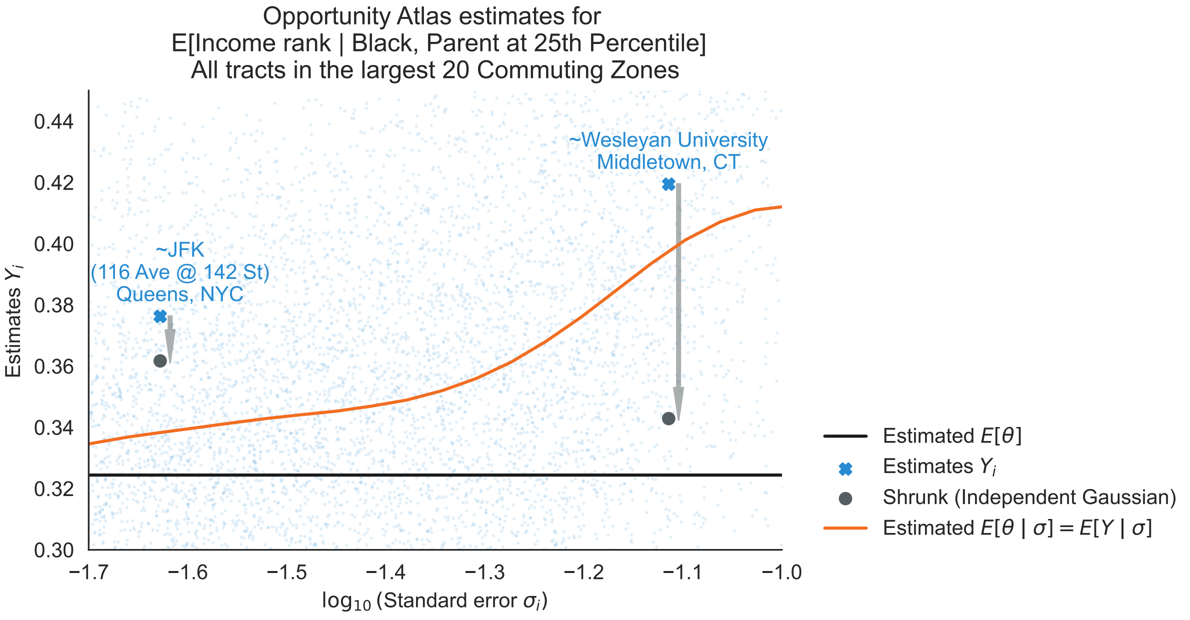
Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
Estimates Y_i
0.7
0.6
0.5
0.4
0.3
0.2
0.1
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
$\log_{10}$ (Standard error σ_i)
Conventional shrinkage formula:
$\theta_i \mid \sigma_i \sim \mathcal{N}(m_0, s_0^2)$
$E[\theta_i|Y_i, \sigma_i] = \frac{s_0^2}{s_0^2 + \sigma_i^2} Y_i + \frac{\sigma_i^2}{s_0^2 + \sigma_i^2} m_0$
Estimates $Y_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2)$
Shrunk (Independent Gaussian)
Estimated $E[\theta]$



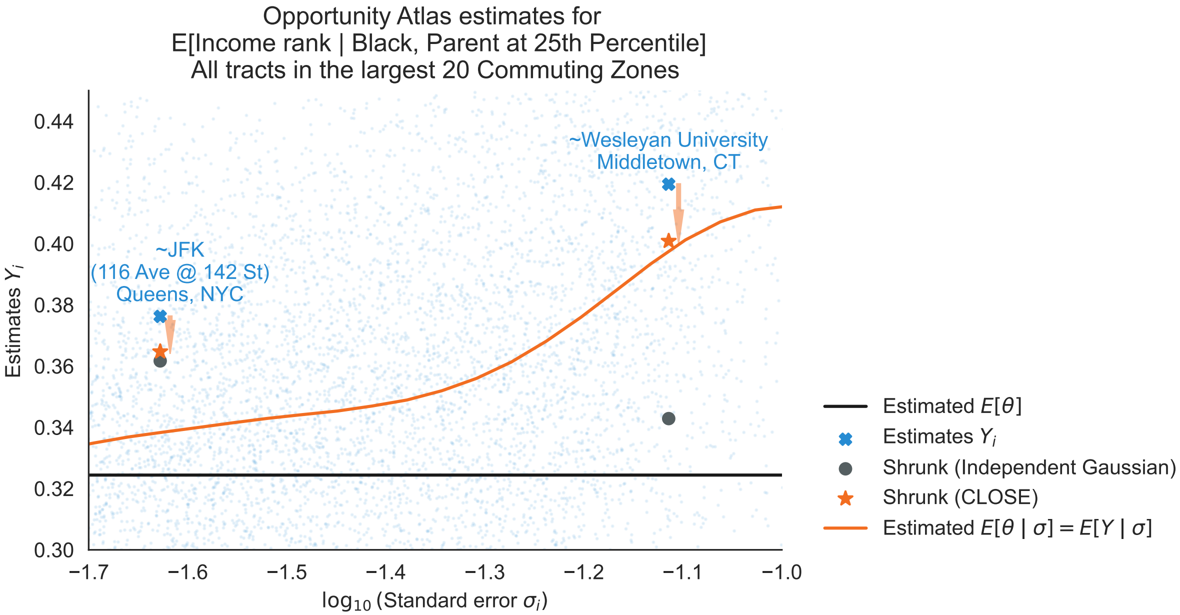
Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
Estimates Y_i
0.7
0.6
0.5
0.4
0.3
0.2
0.1
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
$\log_{10}$ (Standard error σ_i)
$E[Y \mid \sigma] = E[\theta \mid \sigma]$
Positive slope $\implies \theta_i$ correlates with σ_i positively
Prior independence
$\implies E[\theta \mid \sigma]$ is flat
Estimates $Y_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2)$
Estimated $E[\theta \mid \sigma] = E[Y \mid \sigma]$
99% uniform CB for $E[\theta \mid \sigma]$
Shrunk (Independent Gaussian)
Estimated $E[\theta]$


Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
Estimates Y_i
0.7
0.6
0.5
0.4
0.3
0.2
0.1
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
$\log_{10}$ (Standard error σ_i)
Estimates $Y_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2)$
Shrunk (Independent Gaussian)
Estimated $E[\theta \mid \sigma] = E[Y \mid \sigma]$
99% uniform CB for $E[\theta \mid \sigma]$
Estimated $E[\theta]$


Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
~Wesleyan University
Middletown, CT
~JFK
(116 Ave @ 142 St)
Queens, NYC
Estimates Y_i
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30
−1.7
−1.6
−1.5
−1.4
−1.3
−1.2
−1.1
−1.0
$\log_{10}$ (Standard error σ_i)
$E[\theta \mid \sigma_{\text{Wes}}] > Y_{\text{Qns}} > E[\theta \mid \sigma_{\text{Qns}}]$
$\implies$ Probably $\theta_{\text{Wes}} > \theta_{\text{Qns}}$
(Probably $\theta_{\text{Wes}} - \theta_{\text{Qns}} \geq$ 2pct ranks)
Estimated $E[\theta]$
Estimates Y_i
Shrunk (Independent Gaussian)
Estimated $E[\theta \mid \sigma] = E[Y \mid \sigma]$

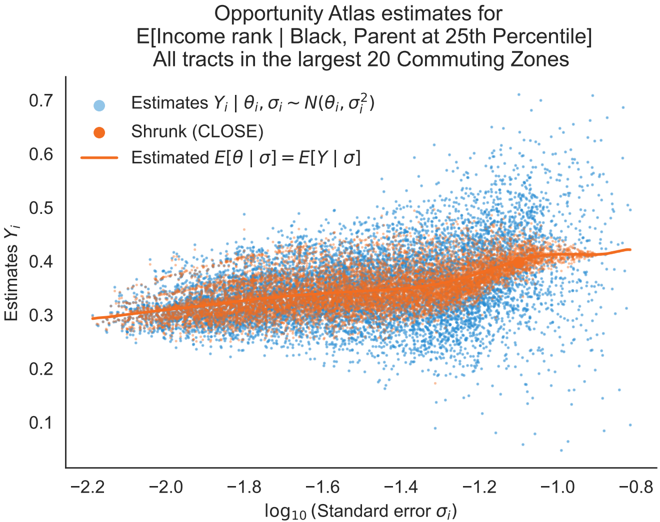


Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
~Wesleyan University
Middletown, CT
~JFK
(116 Ave @ 142 St)
Queens, NYC
Estimates Y_i
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30
−1.7
−1.6
−1.5
−1.4
−1.3
−1.2
−1.1
−1.0
$\log_{10}$ (Standard error σ_i)
Estimated $E[\theta]$
Estimates Y_i
Shrunk (Independent Gaussian)
Estimated $E[\theta \mid \sigma] = E[Y \mid \sigma]$



Opportunity Atlas estimates for
E[Income rank | Black, Parent at 25th Percentile]
All tracts in the largest 20 Commuting Zones
~Wesleyan University
Middletown, CT
~JFK
(116 Ave @ 142 St)
Queens, NYC
Estimates Y_i
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30
−1.7
−1.6
−1.5
−1.4
−1.3
−1.2
−1.1
−1.0
$\log_{10}$ (Standard error σ_i)
Estimated $E[\theta]$
Estimates Y_i
Shrunk (Independent Gaussian)
Shrunk (CLOSE)
Estimated $E[\theta \mid \sigma] = E[Y \mid \sigma]$

A preview of new empirical Bayes method (CLOSE)



- For selecting top $\frac{1}{3}$, large difference in performance (**average θ among selected tracts**)
- CLOSE selects neighborhoods that are higher by **0.5 percentile ranks** [$\approx \$500$ in annual income]
- $\frac{(\text{CLOSE}) - (\text{Conventional})}{(\text{Conventional}) - (\text{No shrinkage})} =$ **320%**
- These performance calculations adjust for covariates; if not, *conventional EB performs worse than no shrinkage*

Residualize by covariates | More nonlinear conditional mean

1. **Empirical Bayes framework**
2. New empirical Bayes method (CLOSE)
3. Theoretical guarantees for CLOSE
4. Empirical application

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates
  - → Sufficiently general for many empirical contexts beyond neighborhood mobility

    Kane and Staiger (2008), Deming (2014), Chandra et al. (2016b), Aaronson, Barrow, and Sander (2007), Arnold, Dobbie, and Hull (2022), Bloom et al. (2019), Kline and Walters (2021), Kline, Rose, and Walters (2023), Abadie et al. (2023), Diamond and Moretti (2021), Azevedo et al. (2020), Stock and Watson (2012), and Finkelstein, Gentzkow, and Williams (2021)
  - → NB: We assume Y_i are credible estimates of θ_i through (natural) experiments and/or structural models
  - → Economic intuition suggests failure of prior independence
  - → e.g., θ is hospital value-added and more patients select into better hospitals other examples

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates
- Empirical Bayes works by **estimating the distribution of θ_i and use it as a prior**

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates

- We maintain **standard empirical Bayes assumptions**:
  1. **Gaussian sequence model**: Y_i is Gaussian with variance σ_i^2: known vs. estimated σ_i

     $$Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2).$$

     (motivated by CLT on $\sqrt{n_i}(Y_i - \theta_i)$)
  2. **Random effects**: Parameters are random $(\theta_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^{(\text{joint})}$ EB interpretation without iid
     - Minor: we primarily work with the conditional distribution P_0 of $\theta \mid \sigma$

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates

- We maintain **standard empirical Bayes assumptions**:
  1. **Gaussian sequence model**: Y_i is Gaussian with variance σ_i^2: known vs. estimated σ_i
  2. **Random effects**: Parameters are random $(\theta_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^{(\text{joint})}$ EB interpretation without iid

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates

- We maintain **standard empirical Bayes assumptions**:
  1. **Gaussian sequence model**: Y_i is Gaussian with variance σ_i^2: known vs. estimated σ_i
  2. **Random effects**: Parameters are random $(\theta_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^{(\text{joint})}$ EB interpretation without iid
- **Oracle Bayes (optimal)**: If we *knew* the distribution P_0 of $(\theta \mid \sigma)$, we can obtain a posterior distribution $\theta_i \mid Y_i, \sigma_i$ under P_0

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates

- We maintain **standard empirical Bayes assumptions**:
  1. **Gaussian sequence model**: Y_i is Gaussian with variance σ_i^2: known vs. estimated σ_i
  2. **Random effects**: Parameters are random $(\theta_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^{(\text{joint})}$ EB interpretation without iid
- **Oracle Bayes (optimal)**: If we *knew* the distribution P_0 of $(\theta \mid \sigma)$, we can obtain a posterior distribution $\theta_i \mid Y_i, \sigma_i$ under P_0

  $\longrightarrow$ Policy decisions with respect to this oracle posterior are optimal

How does EB work and why can I estimate a prior?

- Recall: we have estimates (Y_i) and standard errors (σ_i) for parameters (θ_i) Covariates

- We maintain **standard empirical Bayes assumptions**:
  1. **Gaussian sequence model**: Y_i is Gaussian with variance σ_i^2: known vs. estimated σ_i
  2. **Random effects**: Parameters are random $(\theta_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^{(\text{joint})}$ EB interpretation without iid
- **Oracle Bayes (optimal)**: If we *knew* the distribution P_0 of $(\theta \mid \sigma)$, we can obtain a posterior distribution $\theta_i \mid Y_i, \sigma_i$ under P_0
  - → Policy decisions with respect to this oracle posterior are optimal
  - → Selecting high-mobility neighborhoods (Bergman et al., 2024) $\rightsquigarrow$ rank on $\mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$

How does EB work and why can I estimate a prior?

- Recall: we have estimates and standard errors for parameters Covariates
  - estimates: Y_i; standard errors: σ_i; parameters: θ_i
- We maintain **standard empirical Bayes assumptions**:
  1. **Gaussian sequence model**: Y_i is Gaussian with variance σ_i^2: known vs. estimated σ_i
  2. **Random effects**: Parameters are random $(\theta_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^{(\text{joint})}$ EB interpretation without iid
- **Oracle Bayes (optimal)**: If we *knew* the distribution P_0 of $(\theta \mid \sigma)$, we can obtain a posterior distribution $\theta_i \mid Y_i, \sigma_i$ under P_0
  - → Policy decisions with respect to this oracle posterior are optimal
  - → Selecting high-mobility neighborhoods (Bergman et al., 2024) ⇝ rank on $\mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$
- **Empirical Bayes (feasible)**: Approximate these infeasible decisions by estimating the distribution P_0 of $\theta_i \mid \sigma_i$ from data (Y_i, σ_i)
  - → "Shrinkage estimates" are empirical Bayes estimates of the posterior mean, $\mathbb{E}_{\hat{P}}[\theta_i \mid Y_i, \sigma_i]$

Prior independence

- Empirical Bayes imitates the oracle by estimating the oracle prior P_0
- Prior independence ($\theta_i \perp\!\!\!\perp \sigma_i$) simplifies this estimation:
  - → **Independent Gaussian** (Morris, 1983): $\theta_i \mid \sigma_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_0, s_0^2)$
    - Conventional shrinkage formula = posterior mean under this model

$$\mathbb{E}[\theta_i \mid Y_i, \sigma_i] = \frac{\sigma_i^2}{s_0^2 + \sigma_i^2} m_0 + \frac{s_0^2}{s_0^2 + \sigma_i^2} Y_i$$

  - → **Independent NPMLE** (Gilraine, Gu, and McMillan, 2020): $\theta_i \mid \sigma_i \overset{\text{i.i.d.}}{\sim} G_{(0)}$ Same shrinkage issue
    - Nonparametric maximum likelihood has good theoretical and computational properties under prior independence
    - **New method (CLOSE) builds on these properties**
- Economic reasoning suggests implicit sample size, at least, predicts θ_i:
  - → **Selection**: The sample size n_i (used to compute Y_i) selects on θ_i (Chandra et al., 2016a)
  - → **Congestion**: The sample size n_i causes an increase/decrease in θ_i (Derenoncourt, 2022)

1. Empirical Bayes works by imitating an oracle
2. **New empirical Bayes method (CLOSE)**
   - → Relax prior independence, but making use of NPMLE
   - → Normalize away the dependence in the first two conditional moments (location and scale)
   - → **C**onditional **lo**cation-**s**cale **e**mpirical Bayes
3. Theoretical guarantees for CLOSE
4. Empirical application

Normalizing prior dependence away

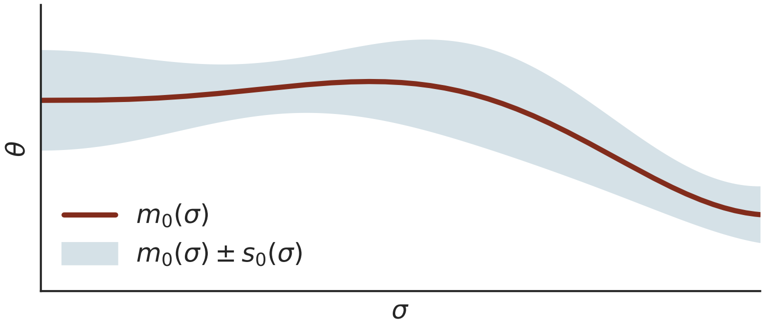
- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$

Normalizing prior dependence away

- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$
- Assume: For some G_0 with mean 0 and variance 1,

$$P_0(\theta \leq q \mid \sigma) = G_0\left(\frac{q - m_0(\sigma)}{s_0(\sigma)}\right)$$

(**Conditional Location-Scale**)

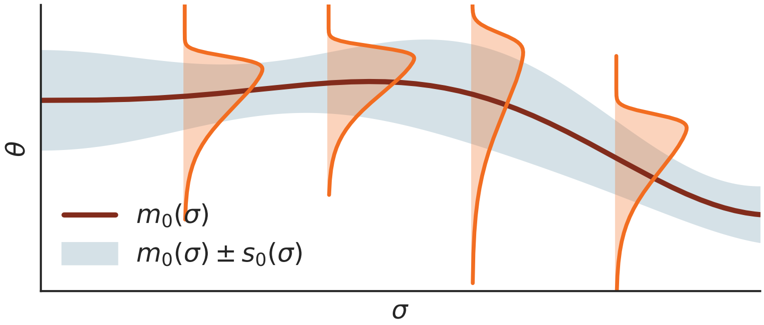


Normalizing prior dependence away

- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$
- Assume: For some G_0 with mean 0 and variance 1,

$$P_0(\theta \leq q \mid \sigma) = G_0\left(\frac{q - m_0(\sigma)}{s_0(\sigma)}\right)$$

(**Conditional Location-Scale**)



Normalizing prior dependence away

- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$
- Assume: For some G_0 with mean 0 and variance 1,

$$P_0(\theta \le q \mid \sigma) = G_0\left(\frac{q - m_0(\sigma)}{s_0(\sigma)}\right)$$

(**Conditional Location-Scale**)

- CLOSE estimates the distribution of $\overbrace{\theta_{1:n} \mid \sigma_{1:n}}^{P_0}$ by estimating the $\overbrace{\textbf{hyperparameters}}^{m_0(\cdot), s_0(\cdot), G_0(\cdot)}$

Normalizing prior dependence away

- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$
- Assume: For some G_0 with mean 0 and variance 1,

$$P_0(\theta \leq q \mid \sigma) = G_0\left(\frac{q - m_0(\sigma)}{s_0(\sigma)}\right)$$

(**Conditional Location-Scale**)

- CLOSE estimates the distribution of $\overbrace{\theta_{1:n} \mid \sigma_{1:n}}^{P_0}$ by estimating the $\overbrace{\textbf{hyperparameters}}^{m_0(\cdot), s_0(\cdot), G_0(\cdot)}$
- **Estimating** (m_0, s_0): $m_0(\sigma) = \mathbb{E}[Y \mid \sigma]$, $\quad s_0^2(\sigma) = \mathrm{Var}(Y \mid \sigma) - \sigma^2$

Normalizing prior dependence away

- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$
- Assume: For some G_0 with mean 0 and variance 1,

$$P_0(\theta \le q \mid \sigma) = G_0\left(\frac{q - m_0(\sigma)}{s_0(\sigma)}\right)$$ (**Conditional Location-Scale**)

- CLOSE estimates the distribution of $\underbrace{\theta_{1:n} \mid \sigma_{1:n}}_{P_0}$ by estimating the $\underbrace{\textbf{hyperparameters}}_{m_0(\cdot), s_0(\cdot), G_0(\cdot)}$
- **Estimating** (m_0, s_0): $m_0(\sigma) = \mathbb{E}[Y \mid \sigma]$, $s_0^2(\sigma) = \mathrm{Var}(Y \mid \sigma) - \sigma^2$
- **Estimating** G_0: Thanks to location-scale ass'n, we can **normalize dependence away**:

$$\tau_i = \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)} \quad Z_i = \frac{Y_i - m_0(\sigma_i)}{s_0(\sigma_i)} \quad \nu_i = \frac{\sigma_i}{s_0(\sigma_i)}$$

In transformed space, Z_i is a Gaussian signal on τ_i where **prior independence holds**:

$$\underbrace{Z_i \mid \tau_i, \nu_i \sim \mathcal{N}(\tau_i, \nu_i^2)}_{\text{cf. } Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)} \qquad \underbrace{\tau_i \mid \nu_i \overset{\text{i.i.d.}}{\sim} G_0}_{\text{Prior independence}}$$

Normalizing prior dependence away

- Relaxing prior independence $\rightsquigarrow$ posit more flexible distribution for $\theta_i \mid \sigma_i$
- Assume: For some G_0 with mean 0 and variance 1,

$$P_0(\theta \leq q \mid \sigma) = G_0\left(\frac{q - m_0(\sigma)}{s_0(\sigma)}\right) \qquad \text{(Conditional Location-Scale)}$$

- CLOSE estimates the distribution of $\overbrace{\theta_{1:n} \mid \sigma_{1:n}}^{P_0}$ by estimating the $\overbrace{\textbf{hyperparameters}}^{m_0(\cdot), s_0(\cdot), G_0(\cdot)}$
- **Estimating** (m_0, s_0): $m_0(\sigma) = \mathbb{E}[Y \mid \sigma]$, $\quad s_0^2(\sigma) = \mathrm{Var}(Y \mid \sigma) - \sigma^2$
- **Estimating** G_0: Thanks to location-scale ass'n, we can **normalize dependence away**:

$$\tau_i = \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)} \quad Z_i = \frac{Y_i - m_0(\sigma_i)}{s_0(\sigma_i)} \quad \nu_i = \frac{\sigma_i}{s_0(\sigma_i)}$$

In transformed space, Z_i is a Gaussian signal on τ_i where **prior independence holds**:

$$\underbrace{Z_i \mid \tau_i, \nu_i \sim \mathcal{N}(\tau_i, \nu_i^2)}_{\text{cf. } Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)} \qquad \underbrace{\tau_i \mid \nu_i \overset{\text{i.i.d.}}{\sim} G_0}_{\text{Prior independence}}$$

$\rightsquigarrow$ Use **NPMLE** on the transformed model

Proposed method: CLOSE (GitHub @jiafengkevinchen/close)

- We propose a natural strategy: **C**onditional **lo**cation-**s**cale **e**mpirical Bayes

Why CLOSE? | Alternatives don't dominate CLOSE | Robustness to conditional location-scale assumption

Proposed method: CLOSE (GitHub @jiafengkevinchen/close)

- We propose a natural strategy: **C**onditional **lo**cation-**s**cale **e**mpirical Bayes

$\hat{m}(\sigma), \hat{s}(\sigma)$ **for** m_0, s_0

1. **Estimate the conditional moments** (e.g., local polynomial regression)

Proposed method: CLOSE (GitHub @jiafengkevinchen/close)

- We propose a natural strategy: **C**onditional **lo**cation-**s**cale **e**mpirical Bayes

$\hat{m}(\sigma), \hat{s}(\sigma)$ **for** m_0, s_0

1. **Estimate the conditional moments** (e.g., local polynomial regression)
2. **Estimate** G_0 **on** $(\hat{Z}_i, \hat{\nu}_i)$ **via Independent NPMLE**

$$(Y, \sigma), \theta \xrightarrow{\text{Transform}} \underbrace{\left(\hat{Z} = \frac{Y - \hat{m}(\sigma)}{\hat{s}(\sigma)}, \hat{\nu} = \frac{\sigma}{\hat{s}(\sigma)}\right), \tau}_{\text{Approx. satisfies prior independence if } \hat{m} \approx m_0 \text{ and } \hat{s} \approx s_0} \xrightarrow{\text{Estimate } G_0} \hat{G}$$

Jiang (2020), Koenker and Gu (2017), Koenker and Mizera (2014), Jiang and Zhang (2009), Soloff, Guntuboyina, and Sen (2021), Kiefer and Wolfowitz (1956), Gilraine, Gu, and McMillan (2020), Saha and Guntuboyina (2020), and Polyanskiy and Wu (2020)

Proposed method: CLOSE (GitHub @jiafengkevinchen/close)

- We propose a natural strategy: **C**onditional **lo**cation-**s**cale **e**mpirical Bayes

We use **NPMLE** to compute $\hat{G}$ More on NPMLE:

$$\hat{G} \in \arg\max_{G \in \mathcal{P}(\mathbb{R})} \sum_{i=1}^{n} \log \underbrace{\left(\int_{-\infty}^{\infty} \frac{1}{\hat{\nu}_i} \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i} \right) G(d\tau) \right)}_{\text{density of } \hat{Z} \sim \mathcal{N}(0, \hat{\nu}^2) \star G \text{ [}G\text{-Gaussian mixture]}}$$

where $\mathcal{P}(\mathbb{R})$ is the set of all distributions on $\mathbb{R}$.

→ Approximate $\mathcal{P}(\mathbb{R})$ with a grid $\rightsquigarrow$ Highly computationally tractable concave program (Koenker and Mizera, 2014; Koenker and Gu, 2017)

→ We can think of NPMLE as a "deluxe" version of CLOSE

→ For a "lite" version, can model $G_0 \sim \mathcal{N}(0, 1)$ directly (Weinstein et al., 2018)

Why CLOSE? Alternatives don't dominate CLOSE Robustness to conditional location-scale assumption

Proposed method: CLOSE (GitHub @jiafengkevinchen/close)

- We propose a natural strategy: **C**onditional **lo**cation-**s**cale **e**mpirical Bayes

$\hat{m}(\sigma), \hat{s}(\sigma)$ **for** m_0, s_0

1. **Estimate the conditional moments** (e.g., local polynomial regression)
2. **Estimate** G_0 **on** $(\hat{Z}_i, \hat{\nu}_i)$ **via Independent NPMLE**

$$(Y, \sigma), \theta \xrightarrow{\text{Transform}} \underbrace{\left(\hat{Z} = \frac{Y - \hat{m}(\sigma)}{\hat{s}(\sigma)}, \hat{\nu} = \frac{\sigma}{\hat{s}(\sigma)}\right)}_{\text{Approx. satisfies prior independence if } \hat{m} \approx m_0 \text{ and } \hat{s} \approx s_0}, \tau \xrightarrow{\text{Estimate } G_0} \hat{G}$$

Jiang (2020), Koenker and Gu (2017), Koenker and Mizera (2014), Jiang and Zhang (2009), Soloff, Guntuboyina, and Sen (2021), Kiefer and Wolfowitz (1956), Gilraine, Gu, and McMillan (2020), Saha and Guntuboyina (2020), and Polyanskiy and Wu (2020)

Proposed method: CLOSE (GitHub @jiafengkevinchen/close)

- We propose a natural strategy: **C**onditional **lo**cation-**s**cale **e**mpirical Bayes

$\hat{m}(\sigma), \hat{s}(\sigma)$ **for** m_0, s_0

1. **Estimate the conditional moments** (e.g., local polynomial regression)
2. **Estimate G_0 on $(\hat{Z}_i, \hat{\nu}_i)$ via Independent NPMLE**

$$(Y, \sigma), \theta \xrightarrow{\text{Transform}} \left(\hat{Z} = \frac{Y - \hat{m}(\sigma)}{\hat{s}(\sigma)}, \hat{\nu} = \frac{\sigma}{\hat{s}(\sigma)}\right), \tau \xrightarrow{\text{Estimate } G_0} \hat{G}$$

Approx. satisfies prior independence if $\hat{m} \approx m_0$ and $\hat{s} \approx s_0$

Jiang (2020), Koenker and Gu (2017), Koenker and Mizera (2014), Jiang and Zhang (2009), Soloff, Guntuboyina, and Sen (2021), Kiefer and Wolfowitz (1956), Gilraine, Gu, and McMillan (2020), Saha and Guntuboyina (2020), and Polyanskiy and Wu (2020)

3. **Plug in prior estimates to decision rules $\hat{\delta}_{\text{EB}}\left(\hat{m}, \hat{s}, \hat{G}\right)$** (e.g. take the posterior mean)

Why CLOSE? | Alternatives don't dominate CLOSE | Robustness to conditional location-scale assumption

1. Empirical Bayes works by imitating an oracle
2. CLOSE works by normalizing and applying NPMLE
3. **Theoretical guarantees for CLOSE**
   - → Measuring performance by **regret**
   - → Prove regret bounds
4. Empirical application

Regret

- **Goal**: Characterize the performance of CLOSE for various decision problems
- Recall the empirical Bayes logic: "emulate the **oracle Bayesian** by estimating P_0"
  - → This is sensible b/c oracle decision $\boldsymbol{\delta}^\star$ is optimal for expected loss (**Bayes risk**)

$$\boldsymbol{\delta}^\star \in \arg\min_{\boldsymbol{\delta}} \underbrace{R_n(\boldsymbol{\delta})}_{\text{Bayes risk}} = \arg\min_{\boldsymbol{\delta}} \underbrace{\mathbb{E}_{P_0}\left[L(\boldsymbol{\delta}(Y_{1:n}, \sigma_{1:n}), \theta_{1:n})\right]}_{\text{Expected loss (over } \theta, Y \mid \sigma)}$$

  - → With large n, *hopefully* $\hat{P} \approx P_0$ and $R_n(\hat{\boldsymbol{\delta}}_{\mathrm{EB}}) \approx R_n(\boldsymbol{\delta}^\star)$—but how close exactly?
- Natural to consider **regret** (Jiang and Zhang, 2009), which is the **suboptimality of EB**:

$$\text{Regret} = \overbrace{R_n(\hat{\boldsymbol{\delta}}_{\text{EB}})}^{\text{Bayes risk of EB rule}} - \overbrace{R_n(\boldsymbol{\delta}^\star)}^{\text{Bayes risk of oracle}}$$

- For estimating θ in mean-squared error (here the posterior mean $\theta_i^\star$ is optimal)

$$\text{Regret} = \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\delta}_i(Y_i, \sigma_i) - \theta_i)^2\right] - \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\theta_i^\star - \theta_i)^2\right]$$

Main regret rate result (preview)

$$\text{Regret} = \underbrace{R_n(\hat{\boldsymbol{\delta}}_{\text{EB}})}_{\text{Bayes risk of EB rule}} - \underbrace{R_n(\boldsymbol{\delta}^\star)}_{\text{Bayes risk of oracle}}$$

- **Main result**: For estimation in squared error, under the location-scale model, CLOSE attains

$$\text{Regret} \le C_0 (\log n)^{C_1} \cdot \max\left(\underbrace{\mathbb{E} \sup_\sigma |\hat{m}(\sigma) - m_0(\sigma)|^2, \mathbb{E} \sup_\sigma |\hat{s}(\sigma) - s_0(\sigma)|^2}_{\text{How fast the (error)}^2 \text{ of estimating } \eta_0 = (m_0, s_0) \text{ shrinks as fn of } n} \right)$$

$\to$ MSE Regret $\le$ (How poorly we estimate η_0 via nonparametric regression)^2

We are CLOSE to the oracle

Theorem (MSE regret control, informal)

1. *Assume the* ***location-scale assumption*** *holds* Robustness to CLS
2. *Assume* G_0 *is* $\exp(-c_0|x|^{c_1})$*-**tailed** and **variances*** $(s_0^2(\sigma), \sigma^2)$ ***are bounded away*** *from* 0 *and* ∞
   As'ns
3. ***(Hölder-p smoothness)*** *Assume* $m_0(\cdot), s_0(\cdot)$ *have* p *bounded derivatives*
4. ***(Good estimators)*** *Assume estimators* $\hat{m}(\cdot), \hat{s}(\cdot)$ *are suitably smooth and rate-optimal in* $\|\cdot\|_\infty$

Then, there exists constants $C_0, C_1 > 0$ *such that, uniformly over* $(P_0, \sigma_{1:n})$,

MSE regret of CLOSE

(Error)^2-rate for estimating Hölder-smooth η_0

$$\underbrace{R_n(\delta_{\mathrm{EB}}(\hat{m}, \hat{s}, \hat{G})) - R_n(\delta^\star)}_{\text{MSE regret of CLOSE}} \le C_0 (\log n)^{C_1} \underbrace{\left(n^{-\frac{p}{2p+1}}\right)^2}_{\text{(Error)}^2\text{-rate}} .$$

Proof ideas

Controlling MSE regret for CLOSE is not much harder than estimating 1D functions with p derivatives

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  $\rightarrow$ $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  → $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:
  → Jiang and Zhang (2009) (homoskedastic), Saha and Guntuboyina (2020) (homoskedastic + multivariate), Jiang (2020) (heteroskedastic + **prior indep.**), Soloff, Guntuboyina, and Sen (2021) (heteroskedastic + multivariate + **prior indep.**), Polyanskiy and Wu (2021) (homoskedastic + lower bounds)

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  → $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:
  → Without *any* assumption on P_0: the worst-case regret $\not\to 0$

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  - → $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:
  - → Without *any* assumption on P_0: the worst-case regret $\not\to 0$
  - → With prior independence: the regret is $\tilde{O}(n^{-1})$ (Soloff, Guntuboyina, and Sen, 2021; Jiang, 2020)

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  - $\to$ $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:
  - $\to$ Without *any* assumption on P_0: the worst-case regret $\not\to 0$
  - $\to$ **This regret upper bound**: Smoothness on $\sigma \mapsto (\theta \mid \sigma) \implies$ rate in between
  - $\to$ With prior independence: the regret is $\tilde{O}(n^{-1})$ (Soloff, Guntuboyina, and Sen, 2021; Jiang, 2020)

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  - $\rightarrow$ $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:
  - $\rightarrow$ Without *any* assumption on P_0: the worst-case regret $\not\to 0$
  - $\rightarrow$ **This regret upper bound**: Smoothness on $\sigma \mapsto (\theta \mid \sigma) \implies$ rate in between
  - $\rightarrow$ With prior independence: the regret is $\tilde{O}(n^{-1})$ (Soloff, Guntuboyina, and Sen, 2021; Jiang, 2020)
- **Minimax lower bound** Statement+intuition: Under the location-scale model, the worst-case regret of any procedure $\gtrsim n^{-2p/(2p+1)}$

$\implies$ Upper bound is not improvable in the worst-case up to logs

Regret upper bound: context and fundamental limits

- Result recap: Squared error regret rate is $n^{-2p/(2p+1)}$ up to logs
  - → $n^{-p/(2p+1)}$ is the fundamental difficulty of estimating 1d functions w/ p derivatives
- This rate **smoothly extrapolates** from existing regret rates to accommodate prior dependence:
  - → Without *any* assumption on P_0: the worst-case regret $\not\to 0$
  - → **This regret upper bound**: Smoothness on $\sigma \mapsto (\theta \mid \sigma) \implies$ rate in between
  - → With prior independence: the regret is $\tilde{O}(n^{-1})$ (Soloff, Guntuboyina, and Sen, 2021; Jiang, 2020)
- **Minimax lower bound** Statement+intuition: Under the location-scale model, the worst-case regret of any procedure $\gtrsim n^{-2p/(2p+1)}$

$\implies$ Upper bound is not improvable in the worst-case up to logs

- MSE regret upper bound, natural extension of literature, and not improvable
- Next, how is MSE result useful for other economic decisions?
  - → **Ranking/classification-type problems** (Bergman et al., 2024)

Two ranking/classification problems

1. (**Utility maximization by selection**) The utility function is

$$\overbrace{-L(\boldsymbol{\delta}, \theta_{1:n})}^{\text{Utility function}} = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\delta_i(Y_{1:n}, \sigma_{1:n})}^{\text{decision rule (binary)}} \; (\theta_i - \overbrace{c_i}^{\text{known}})$$

→ The oracle Bayes rule **thresholds** on the oracle posterior means $\theta_i^\star$: $\delta_i^\star = \mathbb{1}(\theta_i^\star \geq c_i)$

- Treatment choice: Manski (2004), Kitagawa and Tetenov (2018), Athey and Wager (2021), ...
- Classification: Audibert and Tsybakov (2007), Bonvini, Kennedy, and Keele (2023), ...

2. (**Top-m selection**) $-L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{m}\sum_{i=1}^{n} \overbrace{\delta_i(Y_{1:n}, \sigma_{1:n})}^{\text{binary, sum to } m} \theta_i$

→ In Bergman et al. (2024), $m = n/3$. We can think of $-L$ as the expected mobility that a mover experiences, if the mover moves uniformly at random to one of the recommended tracts

→ The oracle Bayes rule **ranks** the oracle Bayes posterior means: Set $\delta_i^\star = 1$ iff $\theta_i^\star$ is in the top m Generalization to weighted version

MSE regret rate implies bounds for ranking-type decisions

- **Preview**: Regret for ranking $\leq$ (MSE regret)$^{1/2} = \tilde{O}(n^{-p/(2p+1)})$
- In all three decision problems, the oracle Bayes rule is a function of the oracle PMs $\theta_i^\star$
- The empirical Bayes recipe says we should plug in certain estimates of the oracle PM $\hat{\theta}_i$
- **Intuition**: When EB makes a selection mistake, if MSE regret is low, the mistake isn't costly

Theorem

1. *For* ***utility maximization by selection****,*

$$\underbrace{R_n^{\text{(UM)}}(\hat{\boldsymbol{\theta}})}_{\text{Bayes risk of plug-in } \hat{\theta}} - \underbrace{R_n^{\text{(UM)}}(\boldsymbol{\theta}^\star)}_{\text{Bayes risk of oracle}} \leq \Bigg(\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(\theta_i^\star - \hat{\theta}_i\right)^2\right]}_{\text{Bayes regret in squared error}} \Bigg)^{1/2}$$

2. *For* ***top-m selection****,*

$$R_n^{\text{(Top)}}(\hat{\boldsymbol{\theta}}) - R_n^{\text{(Top)}}(\boldsymbol{\theta}^\star) \leq 2\sqrt{\frac{n}{m}} \underbrace{\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(\theta_i^\star - \hat{\theta}_i\right)^2\right]\right)^{1/2}}_{\lesssim \|\hat{\eta}-\eta_0\|_\infty (\log n)^{C_1/2} \text{ by regret bound}}$$

Remarks

Theory summary

- CLOSE attains squared error regret **upper bound** under location-scale model
- This upper bound is approximately tight: Matches regret **lower bound**
- This upper bound is useful: **Regret for ranking-type problems** dominated by regret in squared error
- Maybe the upper bound is too optimistic. In the paper: Without location-scale,
  - → (Interpretation under misspecification) CLOSE brings the conditional distributions $\theta_i \mid \sigma_i$ closer to each other, so that prior independence is a plausibly better approximation, and NPMLE has a better shot at succeeding
  - → (Bounded badness under misspecification) a version of CLOSE achieves risk within a constant multiple of a notion of minimax risk Robustness to CLS

1. Empirical Bayes works by imitating an oracle
2. CLOSE works by normalizing and applying NPMLE
3. CLOSE is regret rate-optimal
4. **Empirical applications**
    - → Simulation
    - → Empirical application to selecting high-mobility neighborhoods

Opportunity Atlas (Chetty et al., 2020)

(Y_i, σ_i) θ_i i

- Recall: The OA produces estimates for economic mobility at the Census tract level
  - → Causal evidence that neighborhoods matter for upward mobility (Chetty and Hendren, 2018; Chetty, Hendren, and Katz, 2016; Chyn and Katz, 2021; Laliberté, 2021)
  - → Chetty et al. (2020): Observational measures predict these causal effects
  - → Bergman et al. (2024): Incentivizes poor households to move $\rightsquigarrow$ find significant take-up
  - → A program that identifies and recommends high-mobility areas can have real gains
- For our purposes, OA measures of mobility take the following form

  θ_i is the population mean `outcome` for individuals of `race` growing up in Census tract i, whose parents are at the 25th pct of nat'l income

  - → $\theta_i = \mathbb{E}[\text{Income rank} \mid \text{Black}, \text{Parents@P25}, \text{Census tract } i]$.
- EB applied to **residual** of (Y_i, θ_i) against covariates; fitted values $\hat{\gamma}' X$ are added back
- **Empirical ex. today**: Perform **calibrated simulation** and **empirical application**

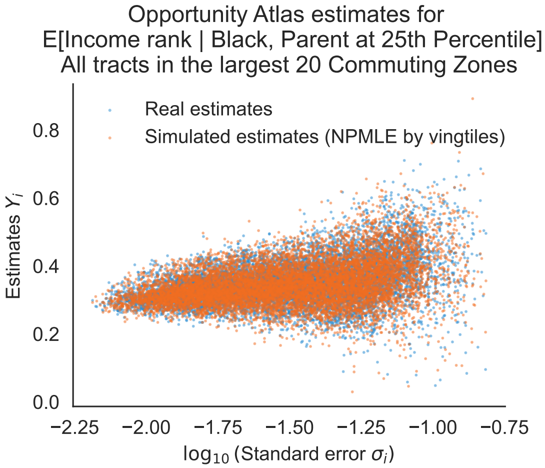
  calibrated simulation: I draw from a known DGP constructed based on real data

  empirical application: I evaluate out-of-sample on the real data

Residualized against covariates | OA estimation details | Empirical application

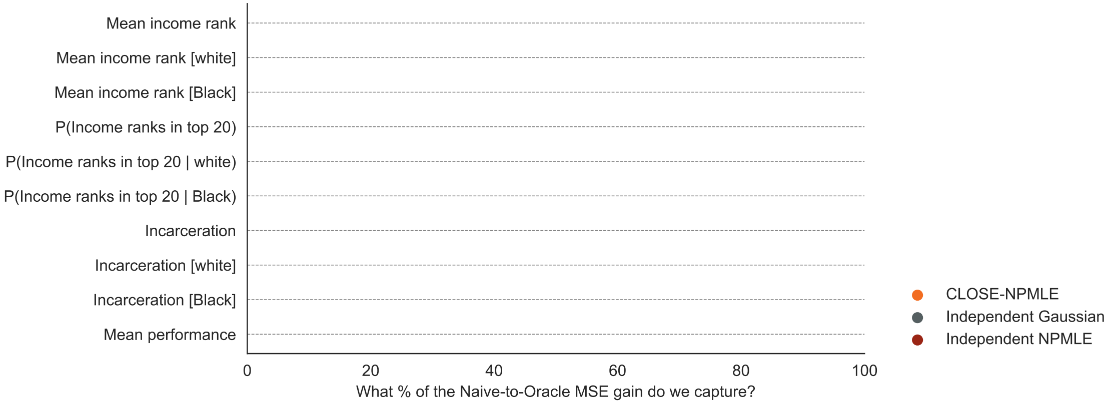
Calibrated simulation (location-scale model misspecified)

Details on calibrated DGP



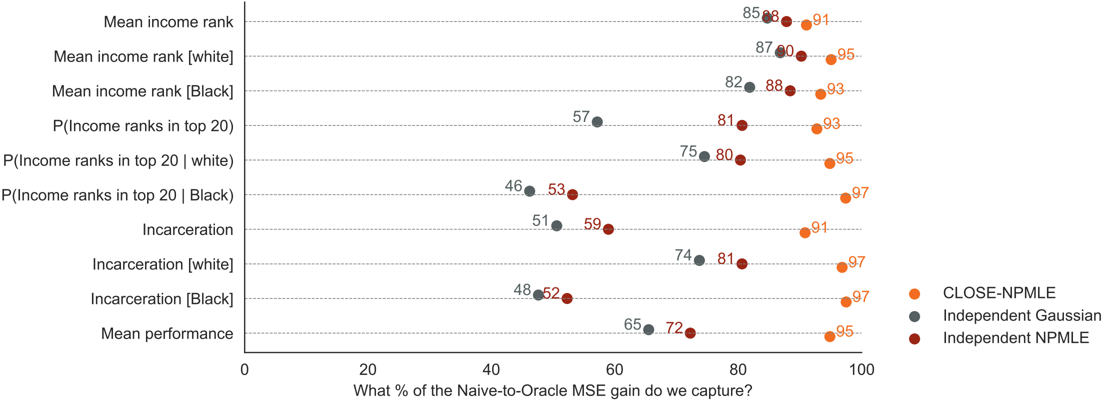
- Estimate $\tilde{P}$ for P_0 (NPMLE within vingtiles of σ, without imposing location-scale model)
- On repeated draws from $\tilde{P}$, compute various EB procedures
- $\tilde{P} \approx P_0$ in terms of the implied distribution of Y_i

Calibrated simulation (location-scale model misspecified)



(Naive = using Y_i directly)

Calibrated simulation (location-scale model misspecified)



(Naive = using Y_i directly)

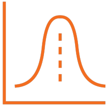
More comprehensive results

1. Empirical Bayes works by imitating an oracle
2. CLOSE works by normalizing and applying NPMLE
3. CLOSE is regret rate-optimal
4. **Empirical application**
   → CLOSE has near oracle performance in simulations
   → **Empirical application to selecting high-mobility neighborhoods** Back

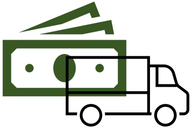
Empirical application to Creating Moves to Opportunity (Bergman et al., 2024)

OPPORTUNITY INSIGHTS

Opportunity Atlas (Chetty et al., 2018) estimates of economic mobility for Seattle



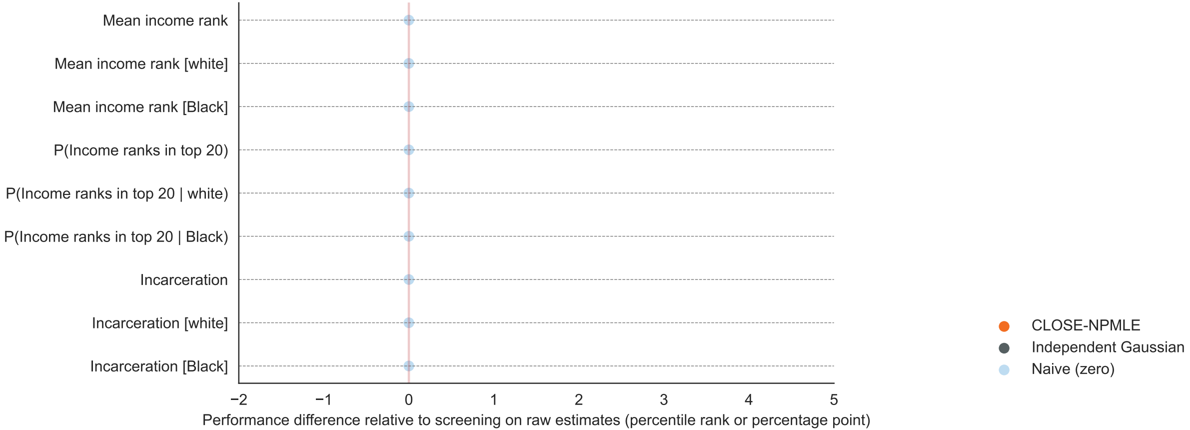
Select **top third** of Census tracts via empirical Bayes shrinkage methods

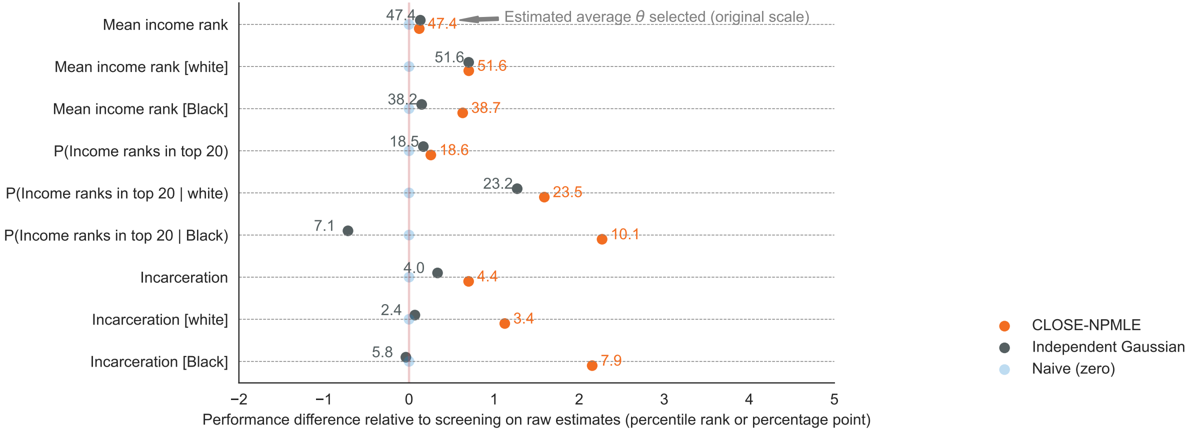


Provide resources to Housing Choice Voucher recipients to move to selected high-mobility areas

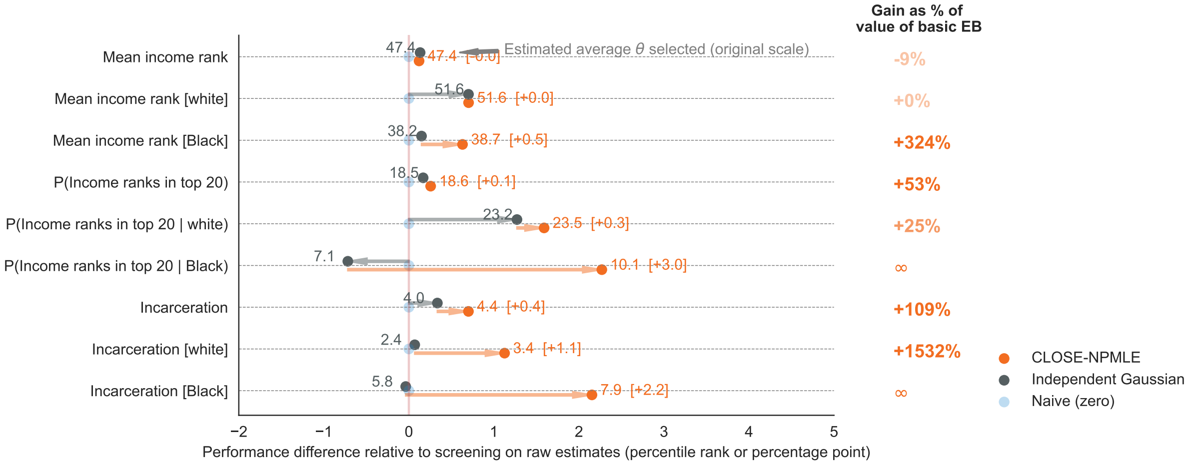
- **Question**: Can we select more economically mobile tracts (on average) with CLOSE?
- We validate performance out-of-sample (i.e. unbiased loss estimates)
  - → Ideally, sample-split the micro-data $\rightsquigarrow$ Obtain $Y_i^{\text{train}}, Y_i^{\text{test}}$ (cond. independent given θ_i)
  - → Use Y_i^{train} to estimate decision rules, evaluate with Y_i^{test} (**Avg.** θ **among selected**)
  - → Don't have the micro-data, but can emulate the splitting via coupled bootstrap (**90/10-split**)

Oliveira, Lei, and Tibshirani (2021)

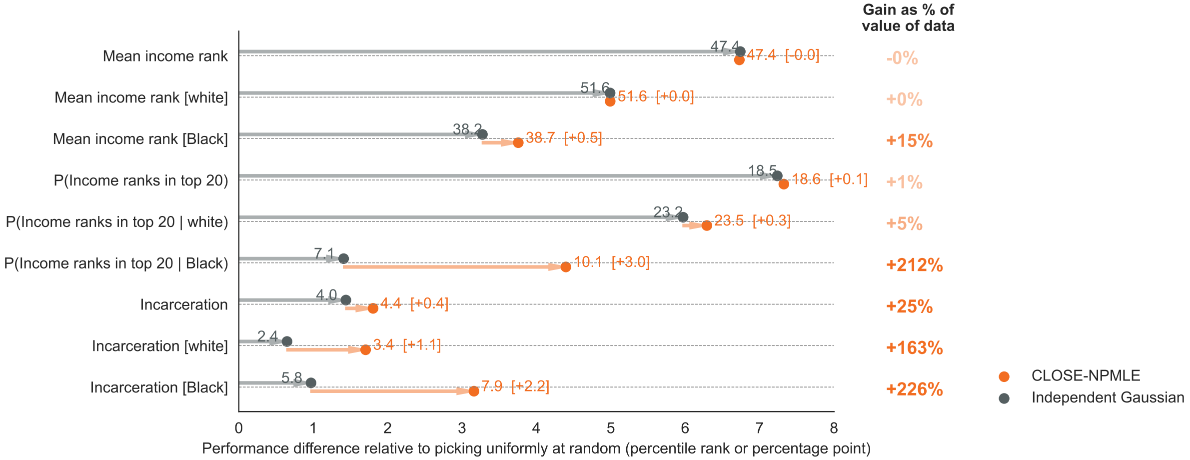

Mean income rank
Mean income rank [white]
Mean income rank [Black]
P(Income ranks in top 20)
P(Income ranks in top 20 | white)
P(Income ranks in top 20 | Black)
Incarceration
Incarceration [white]
Incarceration [Black]
−2
−1
0
1
2
3
4
5
Performance difference relative to screening on raw estimates (percentile rank or percentage point)
CLOSE-NPMLE
Independent Gaussian
Naive (zero)



Mean income rank
Mean income rank [white]
Mean income rank [Black]
P(Income ranks in top 20)
P(Income ranks in top 20 | white)
P(Income ranks in top 20 | Black)
Incarceration
Incarceration [white]
Incarceration [Black]
47.4
47.4
Estimated average θ selected (original scale)
51.6
51.6
38.2
38.7
18.5
18.6
23.2
23.5
7.1
10.1
4.0
4.4
2.4
3.4
5.8
7.9
−2
−1
0
1
2
3
4
5
Performance difference relative to screening on raw estimates (percentile rank or percentage point)
CLOSE-NPMLE
Independent Gaussian
Naive (zero)



Gain as % of value of basic EB
Estimated average θ selected (original scale)
Mean income rank
Mean income rank [white]
Mean income rank [Black]
P(Income ranks in top 20)
P(Income ranks in top 20 | white)
P(Income ranks in top 20 | Black)
Incarceration
Incarceration [white]
Incarceration [Black]
47.4
47.4 [-0.0]
-9%
51.6
51.6 [+0.0]
+0%
38.2
38.7 [+0.5]
+324%
18.5
18.6 [+0.1]
+53%
23.2
23.5 [+0.3]
+25%
7.1
10.1 [+3.0]
∞
4.0
4.4 [+0.4]
+109%
2.4
3.4 [+1.1]
+1532%
5.8
7.9 [+2.2]
∞
−2
−1
0
1
2
3
4
5
Performance difference relative to screening on raw estimates (percentile rank or percentage point)
CLOSE-NPMLE
Independent Gaussian
Naive (zero)



Gain as % of value of data
Mean income rank
Mean income rank [white]
Mean income rank [Black]
P(Income ranks in top 20)
P(Income ranks in top 20 | white)
P(Income ranks in top 20 | Black)
Incarceration
Incarceration [white]
Incarceration [Black]
47.4
47.4 [-0.0]
-0%
51.6
51.6 [+0.0]
+0%
38.2
38.7 [+0.5]
+15%
18.5
18.6 [+0.1]
+1%
23.2
23.5 [+0.3]
+5%
7.1
10.1 [+3.0]
+212%
4.0
4.4 [+0.4]
+25%
2.4
3.4 [+1.1]
+163%
5.8
7.9 [+2.2]
+226%
CLOSE-NPMLE
Independent Gaussian
0
1
2
3
4
5
6
7
8
Performance difference relative to picking uniformly at random (percentile rank or percentage point)

Recap/Conclusion Back

- Conventional empirical Bayes methods perform badly when prior independence fails
  - → Shrinks to the wrong target and makes unreasonable selections
- New procedure (CLOSE) **normalizes dependence away** and applies SotA methods
  - → Relaxes prior independence, but still taking advantage of NPMLE
- **Theoretical contributions**
  - → We prove that CLOSE's squared error risk is close to the oracle at optimal rates
  - → This result implies regret rates for two ranking-type decision problems
- In calibrated sims, **near-oracle MSE performance**
- **Significantly improves selection decisions** for selecting the top third in the OA data, relative to standard methods
  - → Mean income rank for Black individuals: 0.5 percentile rank gain (**15%** of the value of data, 320% of the value of basic EB) Targeting minority-focused outcomes vs. targeting pooled outcomes
  - → P(income ranks in the top 20) for Black individuals: 3pp gain (**220%** the value of data)

Thank you!

jiafengchen@g.harvard.edu

Appendix

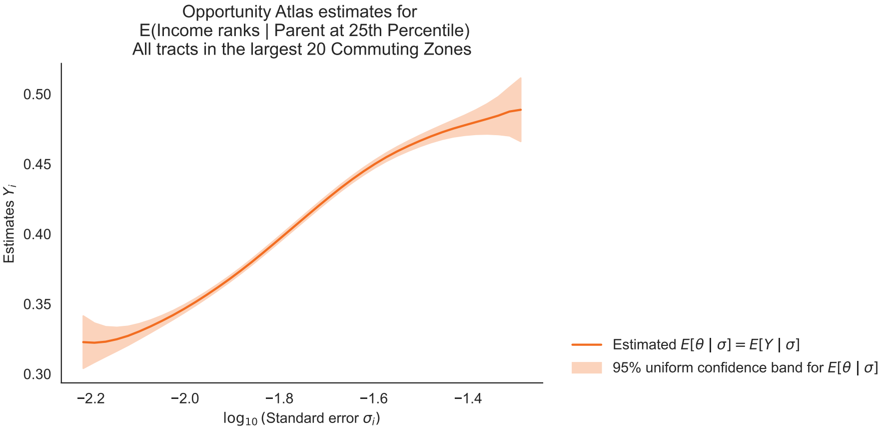
Non-iid Back

- Consider $\boldsymbol{Y} \mid \boldsymbol{\theta}, \Sigma \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$ where $(\boldsymbol{\theta}, \Sigma)$ has some joint distribution
- For squared error loss, consider separable decision rules

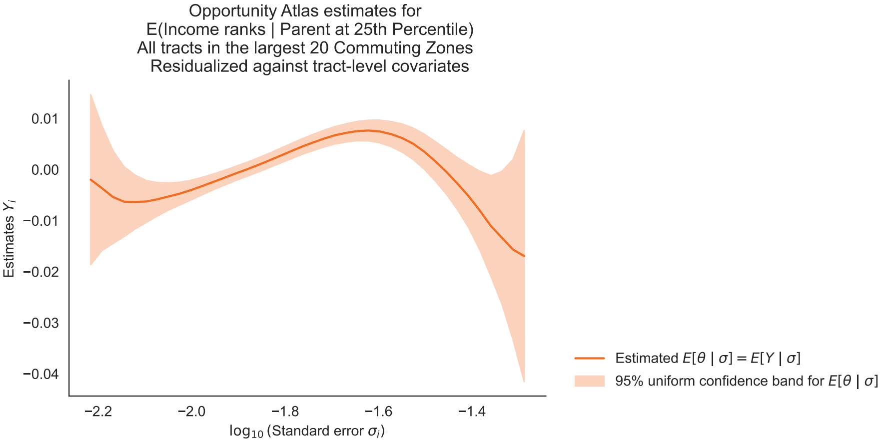
$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\theta_i - \delta_i(Y_i))^2 \mid \Sigma\right]$$

- Optimal decision rule is $\delta_i^\star(Y_i) = \mathbb{E}[\theta_i \mid Y_i, \Sigma]$
- This decision rule depends on the marginal distribution of θ_i $(\theta_i \mid \Sigma)$
- Empirical Bayes methods assuming iid data can be viewed as attempting to learn $\theta_i \mid \Sigma$
- Statistical guarantees for EB procedures might not extend, depending on the correlation structure of $\boldsymbol{\theta} \mid \Sigma$

Mean income rank (unresidualized by covariates) Back

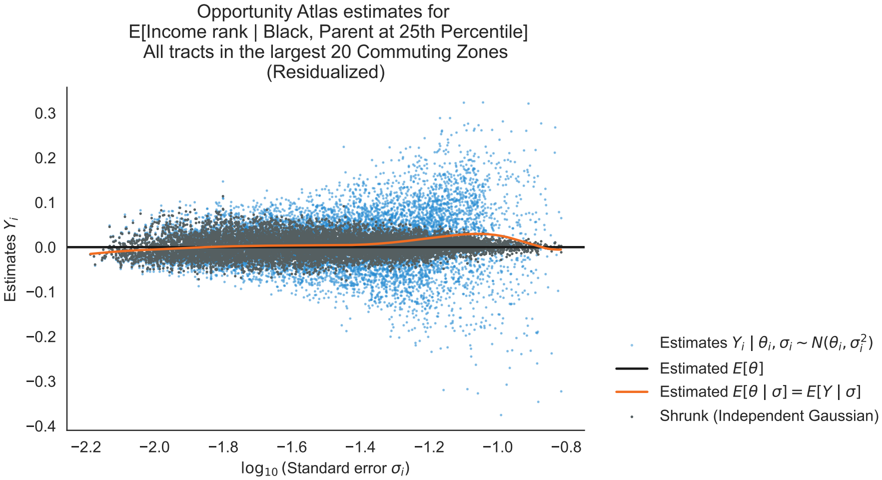


Mean income rank (residualized by covariates) Back



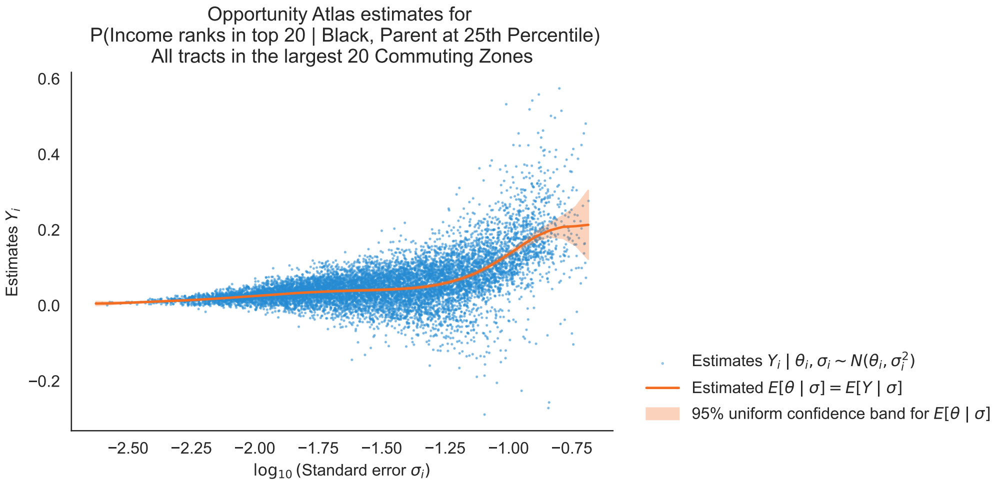
Residualize by covariates

Back

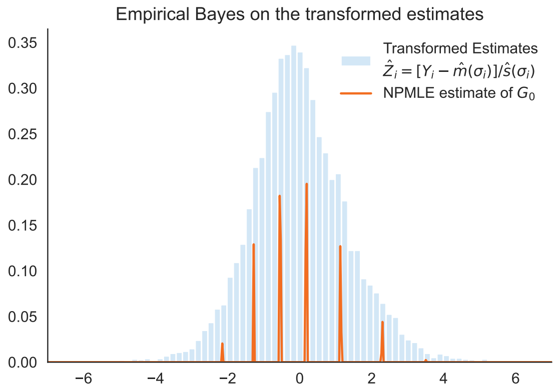


Nonlinear conditional mean

Back



More on NPMLE Back



$$\hat{G} \in \underset{G \in \mathcal{P}(\mathbb{R})}{\arg\max} \sum_{i=1}^{n} \log \left(\int_{-\infty}^{\infty} \frac{1}{\hat{\nu}_i} \varphi \left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i} \right) G(d\tau) \right)$$

- Tuning free statistical objective
- Approximate computationally with fine grid over an interval, which results in a concave optimization problem
  - → Theoretically, no bias-variance tradeoff b/c objective is "self-regularized"

Known vs. estimated σ_i^2 Back

- We assume that the asymptotic approximation $\sigma_i^{-1}(Y_i - \theta_i) \xrightarrow{d} \mathcal{N}(0,1)$ holds **exactly**
  - → If we're calling σ_i a "standard error," we would rely on $Y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ for inference—in this case exactly ignoring the asymptotics
  - → Still, this is a theoretical limitation, but imposed by much of the EB literature
  - → Behavior of empirical Bayes methods when the asymptotic approximation is poor is an important consideration for future work
- The asymptotic approximation is valid regardless if we use the true σ_i^* or some estimated σ_i, provided that

$$(\sigma_i^*)^2 = \sigma_{0i}^2/n \quad \sigma_i^2 = \hat{\sigma}_{0i}^2/n \quad \hat{\sigma}_{0i}^2 = \sigma_{0i}^2 + o_p(1)$$

- Thus, if happy with ignoring asymptotics (on a $\sqrt{n}$-scale), also happy with ignoring the difference between σ_i^* and σ_i

Alternatives to CLOSE Back

∄ free-lunch improvement of our assumptions

- Not the first to consider the non-independence problem
- For example,
  - → 2D empirical Bayes with $\boldsymbol{Y}_i = (Y_i, n_i\sigma_i^2)$ (Gu and Koenker, 2017; Banerjee et al., 2020) has $\boxed{1}$ and $\boxed{4}$
  - → SURE-based procedures (Xie, Kou, and Brown, 2012; Kwon, 2021) has $\boxed{2}$
  - → Working off t-statistics Y/σ has $\boxed{2}$ and $\boxed{3}$
  - → Variance-stabilizing transforms with binary-means data has $\boxed{1}$ and $\boxed{3}$
- Generally speaking, existing alternatives have some of the following features

$\boxed{1}$ Still assumes θ_i is independent from some known nuisance parameter (e.g. sample size)

$\boxed{2}$ Limit optimality consideration to a restricted class of procedures

$\boxed{3}$ Change the objective function

$\boxed{4}$ Require underlying microdata

Generalization of Top-m Selection Back

- Suppose each position of $1, \ldots, n$ is associated with a weight w_k where $\sum_i w_i = m$, such that $w_n \geq w_{n-1} \geq \ldots$.
- The DM outputs a ranking of $i = 1, \ldots, n$ denoted by a permutation $\sigma(i)$, where $i = \sigma(n)$ is the most favorable element.
- The utility of the DM is

$$\frac{1}{n} \sum_{k=1}^{n} w_k \theta_{\sigma(k)}$$

- The oracle Bayes rule is rank according to posterior mean $\theta_i^\star$
- When $w_k \in \{0, 1\}$, this problem is top-m selection
- If people are more likely to move to places where we place a higher recommendation in ways that depend solely on rank, then this corresponds to a reasonable objective in CMTO.

Regret control of different decision problem Back

- Can replace the L_2 norm for **utility maximization by selection** with L_1 norm, but worst-case L_1 and L_2 risks are the same.
- These bounds are possibly not tight. However, the plug-in procedures considered is natural, and so its performance may be better than the bounds imply.
- Coey and Hung (2022) study top-m selection. Their bound is in terms of error in estimating $G_{(0)}$, which is logarithmic in nonparametric settings. Their bound in parametric settings is tighter than ours.
- For the generalization of top-m selection, the bound is

$$2\frac{\|w\|}{\sqrt{n}} \cdot (\mathbb{E}[\mathrm{MSE}_n])^{1/2}$$

Controlled tails Back

- If we only have $\hat{\eta} = (\hat{m}, \hat{s})$ being $O_P(r_n)$-consistent, we can show that there is some probability $(1 - \delta)$ event $A_n = \{\|\hat{\eta} - \eta_0\|_\infty \leq C(\delta) r_n\}$ such that

$$\mathbb{E}[\text{Regret}_n \mid A_n] \leq C_0 (\log n)^{C_1} r_n$$

- Turns out, due to the data being thin tailed, there exists some $C(q), C_2$ s.t.

$$\mathrm{P}(\|\hat{\eta} - \eta_0\|_\infty > C(q)(\log n)^{C_2} r_n) \leq \frac{1}{n^q} \qquad \text{(Controlled tails)}$$

- This allows us to also control

$$\mathbb{E}[\text{Regret}_n \mathbb{1}(A_n^C)].$$

Relaxing the requirement on $\hat{s}$ Back

- Since we assumed $\hat{s}$ is sup-norm consistent at rate $r_n \to 0$ and s_0 is bounded away from zero, for all sufficiently large n, with probability tending to 1
- Hence at minimum, we can say that on some event A_n w.p. $\to 1$ and for all $n > N_0$ (so that $Cr_{N_0} \ll (\inf s_0)$)

$$\mathbb{E}[\mathrm{Regret}_n \mid A_n] \leq C_0(\log n)^{C_1} r_n$$

- To show that

$$\mathbb{E}[\mathrm{Regret}_n \mathbb{1}(A_n^C)] \leq C_0(\log n)^{C_1} r_n$$

we need that A_n^C is sufficiently unlikely ("controlled tails"), and that Regret_n isn't too large. The latter is satisfied when $\hat{s} \geq \frac{c}{n}$, which is satisfied with our truncation rule

Opportunity atlas estimation details Back

- Let y_j be the underlying microdata for individual j: e.g. Income rank for individual j.
- Restrict to a particular (race, sex) cell, consider a nonparametric estimate $\hat{f}$ of the function

$$\mathbb{E}[y_j \mid \text{Family income rank, Sex, Race}]$$

- Then, consider the regression for the particular (sex, race) cell:

$$y_j = \alpha_{i(j)} + \beta_{i(j)} \hat{f}(r_j) + U_j$$

where $\alpha_{i(j)}, \beta_{i(j)}$ are tract-level fixed effects

- The estimate Y_i is a fitted value of this regression:

$$\hat{\alpha}_{i(j)} + \hat{\beta}_{i(j)} \frac{1}{n_i} \sum_{j \in i} \hat{f}(r_j)$$

where σ_i is the corresponding standard error

Other examples Back

- θ_i can be a value-added for some teacher (Kane and Staiger, 2008)
  [more experienced teachers have higher value-added, $\text{Corr}(\theta_i, \sigma_i) < 0$]
- treatment effect for some intervention in a metaanalysis (Azevedo et al., 2020)
  [given fixed power, larger effect sizes correlate with smaller experiments, $\text{Corr}(\theta_i, \sigma_i) > 0$]
- racial contact gap for some firm (Kline, Rose, and Walters, 2023)
  [empirically, firms with more precise estimates have less bias against Black names, $\text{Corr}(\theta_i, \sigma_i) < 0$]

Robustness to location-scale Back

For simplicity, let's assume now m_0, s_0 are known, but suppose $\tau_i = \frac{\theta_i - m_0}{s_0}$ are not identically distributed across i, and adversary chooses the shape $\tau_i \mid \sigma_i$

- In CLOSE, if we assume $G_0 \sim \mathcal{N}(0,1)$, the resulting decision rule

$$\delta^\star_{\text{CLOSE-Gaussian}} = m_0(\sigma_i) + \frac{s_0^2(\sigma_i)}{s_0^2(\sigma_i) + \sigma_i^2}(Y_i - m_0(\sigma_i))$$

  is the best linear-in-Y decision rule for MSE (Weinstein et al., 2018)

- $\delta^\star_{\text{CLOSE-Gaussian}}$ is also minimax in this game, with worst-case risk equal to

$$\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + s_0^2(\sigma_i)} s_0^2(\sigma_i) \geq c\left(\frac{1}{n}\sum_{i=1}^n s_0^2(\sigma_i)\right)$$

- How bad can $\hat{G}$ estimated by NPMLE mess up? If $\tilde{G}$ has mean 0 and variance 1, then

$$\text{Worst-case Bayes Risk}(\tilde{G}) \leq C\left(\frac{1}{n}\sum_{i=1}^n s_0^2(\sigma_i)\right)$$

How mild are these assumptions for regret upper bound? Back

Assume the following holds uniformly in n. The constants C and $(\log n)^\beta$ are tied to constants in the following assumptions.

- **(Approximate NPMLE)** $\hat{G}$ is an approximate NPMLE on $(\hat{Z}_i, \hat{\nu}_i)$ supported inside the range of the data
- **(Prior thin-tailed)** The prior shape G_0 has tails $\le c_1 \exp\left(-c_2|\tau|^\alpha\right)$ for $\alpha \in (0, 2]$:
  - → All moments exist $\supset$ Exponential-power tails $\supset$ Subexponential $\supset$ Subgaussian
- **(Bounded, positive variances)** The variances and conditional variances $\sigma^2_{1:n}, s_0^2(\cdot)$ are bounded away from zero and ∞
- **(Good estimators)** The estimators $\hat{\eta} = (\hat{m}, \hat{s})$ are:
  - → $\|\hat{\eta} - \eta_0\|_\infty = O_P(n^{-p/(2p+1)}(\log n)^{C_2})$ with controlled tails Controlled tails
  - → Reside in some function class $\mathcal{V}$ with metric entropy bound (e.g. Hölder; relaxed if X-fitting)
  - → $\hat{s}$ is bounded away from zero and infinity uniformly in n Can be relaxed

Proof ideas for regret upper bound (1) Back

- (Z, ν, τ) satisfies prior independence. Problem: only have $\hat{Z}, \hat{\nu}$, which depends on $(\hat{m}, \hat{s})$
- The logic of the previous literature (Jiang, 2020; Jiang and Zhang, 2009; Soloff, Guntuboyina, and Sen, 2021)
  1. The infeasible NPMLE $\tilde{G}_n$ approximately maximizes the infeasible likelihood $G \mapsto \Psi_n(m_0, s_0, G) = \frac{1}{n} \sum_i \log f_{\mathcal{N}(0,\nu_i)\star G}(Z_i)$
  2. With high probability, approximate maximizers of the infeasible likelihood is close to G_0 in average Hellinger distance for the induced distribution of Z_i
  3. Any $\tilde{G}$ that is close in average Hellinger distance to G_0 produces posterior means that are close to those produced by G_0
- **Key component of our argument**: Given good $\hat{m}, \hat{s}$, the feasible NPMLE $\hat{G}_n$ also approximately maximizes the infeasible likelihood (Modifying 1.)

  → **Side effect**: Our lower bound for $\Psi_n(m_0, s_0, \hat{G}_n)$ requires according modifications of the Hellinger large-deviation inequality (Modifying 2.)

Proof ideas for regret upper bound (2) Back

- **Key component of our argument**: Given good $\hat{m}, \hat{s}$, the feasible NPMLE $\hat{G}_n$ also approximately maximizes the infeasible likelihood
  - $\rightarrow$ **Side effect**: Our lower bound for $\Psi_n(m_0, s_0, \hat{G}_n)$ requires according modifications of the Hellinger large-deviation inequality
- Linearization:

$$\Psi_n(\hat{m}, \hat{s}, \hat{G}_n) - \Psi_n(m_0, s_0, \hat{G}_n) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f_i}{\partial \eta} \underbrace{(\hat{\eta}(\sigma_i) - \eta(\sigma_i))}_{\lesssim \text{sup-norm rate}}$$

- Without bounding $\frac{\partial \log f_i}{\partial \eta}$, the resulting rate is only $\tilde{O}(n^{-p/(2p+1)})$
- Key observation is that

$$\left| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \eta} \log f_{\hat{G}_n \star \mathcal{N}(0,\nu_i^2)}(Z_i) \right| \lesssim (\log n)^\gamma \left(\text{Average Hellinger distance between } \hat{G}_n \text{ and } G_0\right)$$

- Side effect: bound for likelihood depends on Hellinger distance

CLOSE is rate-optimal (up to logs) Back

Goal for LB: Given any procedure, what's its worst-case regret under location-scale?

- Imagine an adversary picks m_0, s_0, G_0 (value of game = difficulty of statistical problem)
- This is a harder problem than if we knew $G_0 = \mathcal{N}(0,1)$ and $s_0 = 1$
- Here, can show that **good posterior mean estimates** $\hat{\theta}_i$ **imply a good estimate** $\hat{m}$
- But $\hat{m}$ cannot be too good (Stone, 1980) $\implies$ $\hat{\theta}_i$ cannot be too good

Theorem (Regret lower-bound, informal)

Suppose m_0, s_0 belong to a Hölder class of order p. Then,

$$\underbrace{\inf_{\hat{\theta}} \sup_{(m_0,s_0),\sigma_{1:n}\in[C_3,C_4],G_0} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^\star)^2\right]}_{\text{worst-case Bayes regret}} \gtrsim (\text{Minimax IMSE for } m_0) \gtrsim n^{-\frac{2p}{2p+1}}$$

Ignatiadis and Wager (2019)

Covariates Back

There are additional covariates in the OA data. In keeping with Bergman et al. (2024), by default, we residualize against the covariates linearly. For data $(\sigma_i, \tilde{Y}_i, \tilde{\theta}_i)$, we perform EB on $(\sigma_i, Y_i = \tilde{Y}_i - X_i'\hat{\beta}, \tilde{\theta}_i = \tilde{\theta}_i - X_i'\hat{\beta})$ and ignore uncertainty in $\hat{\beta}$.

- If $\tilde{Y}_i \mid \sigma_i^2, \tilde{\theta}_i, X_i \sim \mathcal{N}(\tilde{\theta}_i, \sigma_i^2)$ ("X_i only predicts mobility and does not predict noise in estimating mobility"), then we do not need to adjust σ_i for residualized variables.
- Residualization against covariate mimics an oracle Bayesian who has access to the residuals (though the location-scale assumptions are different for different residualization schemes)

Covariates used

The covariates used are poverty rate in 2010, share of Black individuals in 2010, mean household income in 2000, log wage growth for high school graduates, mean family income rank of parents, mean family income rank of Black parents, the fraction with college or post-graduate degrees in 2010, and the number of children—and the number of Black children—under 18 living in the given tract with parents whose household income was below the national median.

Emulated sample-splitting Back

- We will rely on an emulated hold-out set for our first exercises
- Idea: add and subtract noise to estimates $\to$ independent noised-up estimates $\sim$ sample splitting the microdata
- For $(Y_i, \sigma_i, \theta_i)$, let $W \sim \mathcal{N}(0,1)$. Observe that

$$\begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} = \begin{bmatrix} Y_i + c\sigma_i W \\ Y_i - \frac{1}{c}\sigma_i W \end{bmatrix} \mid \theta_i, \sigma_i \sim \mathcal{N}\left(\begin{bmatrix} \theta_i \\ \theta_i \end{bmatrix}, \begin{bmatrix} (1+c^2)\sigma_i^2 & 0 \\ 0 & (1+1/c^2)\sigma_i^2 \end{bmatrix} \right)$$

- Observe that unbiased loss estimators and associated standard errors are available

$$\mathbb{E}\left[\sum_{i=1}^n \delta_i(Y_{1,1:n}) Y_{2i} \mid \theta_{1:n}, Y_{1,1:n} \right] = \sum_{i=1}^n \delta_i(Y_{1,1:n})\theta_i$$

$$\mathrm{Var}\left(\sum_{i=1}^n \delta_i(Y_{1,1:n}) Y_{2i} \mid \theta_{1:n}, Y_{1,1:n} \right) = \sum_{i=1}^n \delta_i(Y_{1,1:n})\sigma_{2i}^2$$

- Here, we will emulate a 90-10 split

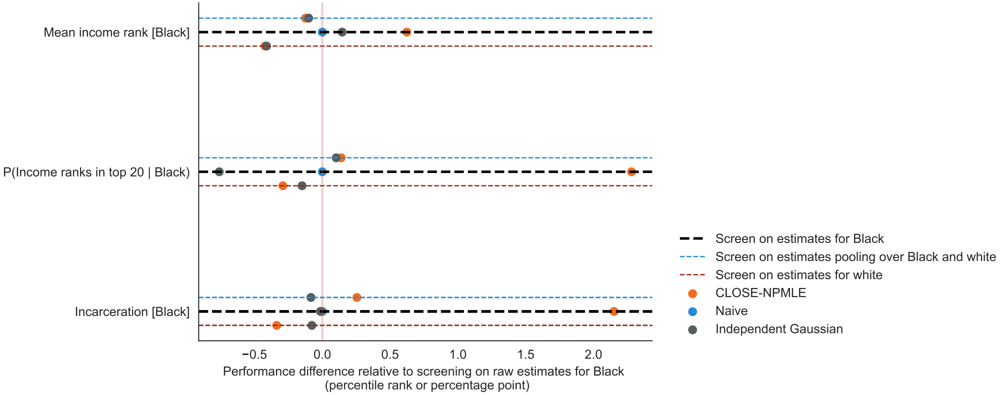
Calibrated simulation Back

- Regress raw $\tilde{Y}_i$ on tract-level covariates X_i to obtain $Y_i = X_i'\beta + Y_i$
- Estimate $m(\sigma) = \mathbb{E}[Y_i \mid \sigma]$ and $s^2(\sigma) = \mathrm{Var}(Y_i \mid \sigma) - \sigma^2$
- Take $Z_i = (Y_i - m(\sigma))/s(\sigma)$
- Estimate $G_1, \ldots, G_{20}$ via NPMLE for each vingtile of σ_i
- The sampling process for new observations is:
  - → Sample τ_i^* from one of the estimated G_k's, depending on σ_i
  - → Set $\theta_i^* = \tau_i^* s(\sigma_i) + m(\sigma_i)$
  - → Sample $Y_i^* = \theta_i^* + \mathcal{N}(0, \sigma_i^2)$
  - → Set $\tilde{Y}_i^* = Y_i^* + X_i'\beta$
  - → Return $(\tilde{Y}_i^*, X_i, \sigma_i)$ as the new data

Why this particular method? Back

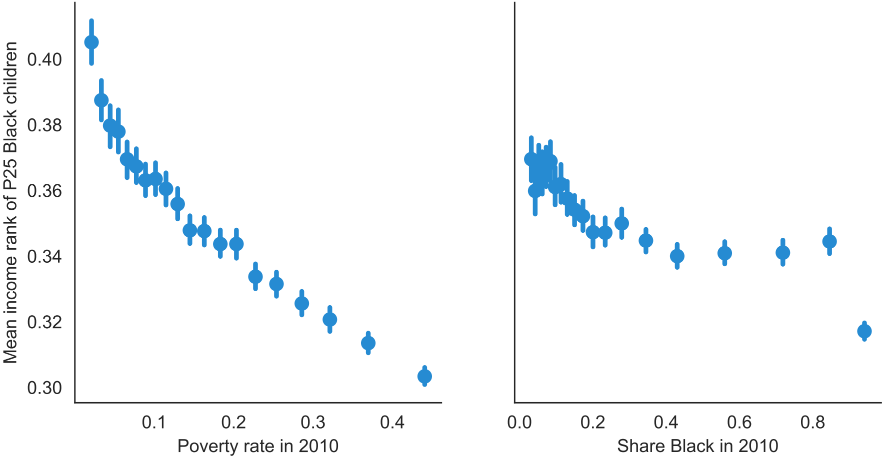
- Many potential models of the joint distribution of (θ_i, σ_i): Among them,
  - → CLOSE is particularly computationally tractable ($\sim$5 seconds with 10,000 estimates)
  - → Takes advantage of computational and theoretical results in the $\theta_i \perp\!\!\!\perp \sigma_i$ case, since NPMLE is the state-of-the-art under prior independence (Koenker and Mizera, 2014; Jiang and Zhang, 2009; Jiang, 2020; Soloff, Guntuboyina, and Sen, 2021)
  - → **Most flexible in the class of "transform data then apply NPMLE"**
- Even when the conditional location-scale assumption fails, CLOSE enjoys a certain degree of robustness (worst-case Bayes risk over choices of shape $G_{(i)}$ within a finite multiple of minimax Bayes risk) Robustness
- Empirical results do not impose location-scale assumption, and CLOSE appears to perform well

Tradeoff between accurate targeting and estimation noise Back

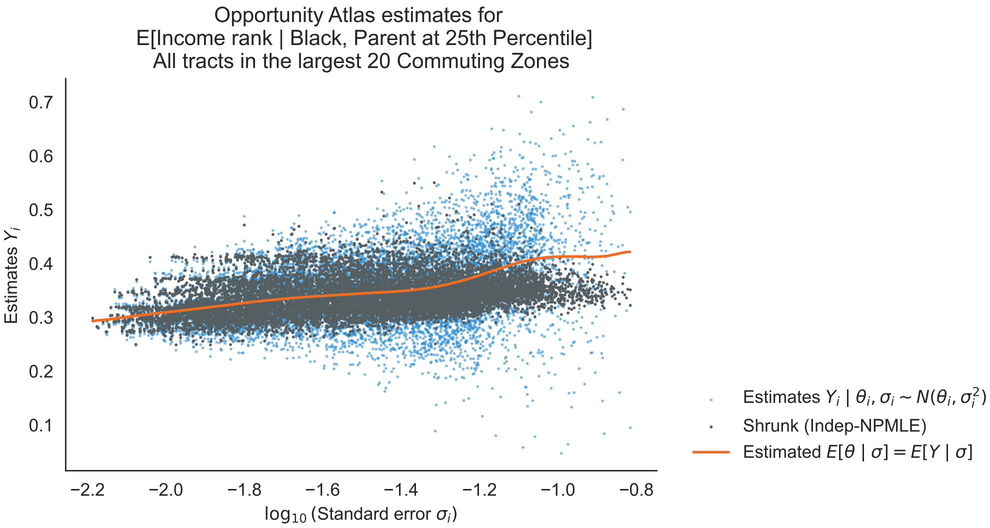


Poorer places and places with more minorities tend to be less mobile

Back



NPMLE shrinkage
Back



MSE results from simulation (complete) Back

What % of Naive-to-Oracle MSE gain do we capture?

	Indep-Gauss (No residualization)	Indep-NPMLE (No residualization)	CLOSE-Gauss (No residualization)	CLOSE-NPMLE (No residualization)	Indep-Gauss	Indep-NPMLE	CLOSE-Gauss	Oracle-Gauss	CLOSE-NPMLE
Mean income rank	-4	25	49	50	85	88	91	91	91
Mean income rank [white]	55	60	66	66	87	90	94	95	95
Mean income rank [Black]	30	61	87	87	82	88	93	94	93
Mean income rank [white male]	63	69	74	75	89	92	93	94	95
Mean income rank [Black male]	32	54	86	87	83	86	93	93	94
P(Income ranks in top 20)	-160	9	67	67	57	81	91	93	93
P(Income ranks in top 20 | white)	31	51	65	65	75	80	94	97	95
P(Income ranks in top 20 | Black)	-6	24	93	95	46	53	95	97	97
P(Income ranks in top 20 | white male)	23	46	71	72	70	76	90	94	94
P(Income ranks in top 20 | Black male)	-8	21	94	96	37	45	95	97	97
Incarceration	-5	32	68	68	51	59	88	95	91
Incarceration [white]	61	72	90	96	74	81	91	93	97
Incarceration [Black]	42	51	94	95	48	52	96	98	97
Incarceration [white male]	43	53	92	96	60	64	93	95	98
Incarceration [Black male]	25	42	90	90	42	49	96	99	96
Column median	30	51	86	87	70	80	93	95	95

References I

Aaronson, Daniel, Lisa Barrow, and William Sander (2007). "Teachers and student achievement in the Chicago public high schools". *Journal of labor Economics* 25.1, pp. 95–135.

Abadie, Alberto et al. (2023). "Estimating the Value of Evidence-Based Decision Making". *arXiv preprint arXiv:2306.13681*.

Arnold, David, Will Dobbie, and Peter Hull (2022). "Measuring racial discrimination in bail decisions". *American Economic Review* 112.9, pp. 2992–3038.

Athey, Susan and Stefan Wager (2021). "Policy learning with observational data". *Econometrica* 89.1, pp. 133–161.

Audibert, Jean-Yves and Alexandre Tsybakov (2007). "Fast learning rates for plug-in classifiers". *Annals of Statistics*.

Azevedo, Eduardo M et al. (2020). "A/b testing with fat tails". *Journal of Political Economy* 128.12, pp. 4614–000.

References II

Banerjee, Trambak et al. (2020). "Nonparametric empirical bayes estimation on heterogeneous data". *arXiv preprint arXiv:2002.12586*.

Bergman, Peter et al. (2024). "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice". *American Economic Review*.

Bloom, Nicholas et al. (2019). "What drives differences in management practices?" *American Economic Review* 109.5, pp. 1648–1683.

Bonvini, Matteo, Edward H Kennedy, and Luke J Keele (2023). "Minimax optimal subgroup identification". *arXiv preprint arXiv:2306.17464*.

Chandra, Amitabh et al. (2016a). "Health care exceptionalism? Performance and allocation in the US health care sector". *American Economic Review* 106.8, pp. 2110–44.

— (2016b). "Productivity dispersion in medicine and manufacturing". *American Economic Review* 106.5, pp. 99–103.

References III

Chetty, Raj and Nathaniel Hendren (2018). "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects". *The Quarterly Journal of Economics* 133.3, pp. 1107–1162.

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz (2016). "The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment". *American Economic Review* 106.4, pp. 855–902.

Chetty, Raj et al. (2020). *The opportunity atlas: Mapping the childhood roots of social mobility*. Tech. rep.

Chyn, Eric and Lawrence F Katz (2021). "Neighborhoods matter: Assessing the evidence for place effects". *Journal of Economic Perspectives* 35.4, pp. 197–222.

Coey, Dominic and Kenneth Hung (2022). "Empirical Bayesian Selection for Value Maximization". *arXiv preprint arXiv:2210.03905*.

Deming, David J (2014). "Using school choice lotteries to test measures of school effectiveness". *American Economic Review* 104.5, pp. 406–411.

References IV

Derenoncourt, Ellora (2022). "Can you move to opportunity? Evidence from the Great Migration". *American Economic Review* 112.2, pp. 369–408.

Diamond, Rebecca and Enrico Moretti (2021). *Where is standard of living the highest? Local prices and the geography of consumption*. Tech. rep. National Bureau of Economic Research.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams (2021). "Place-based drivers of mortality: Evidence from migration". *American Economic Review* 111.8, pp. 2697–2735.

Gilraine, Michael, Jiaying Gu, and Robert McMillan (2020). *A new method for estimating teacher value-added*. Tech. rep. National Bureau of Economic Research.

Gu, Jiaying and Roger Koenker (2017). "Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data". *Journal of Applied Econometrics* 32.3, pp. 575–599.

Ignatiadis, Nikolaos and Stefan Wager (2019). "Covariate-powered empirical Bayes estimation". *Advances in Neural Information Processing Systems* 32.

References V

Jiang, Wenhua (2020). "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means". *Electronic Journal of Statistics* 14.1, pp. 2272–2297.

Jiang, Wenhua and Cun-Hui Zhang (2009). "General maximum likelihood empirical Bayes estimation of normal means". *The Annals of Statistics* 37.4, pp. 1647–1684.

Kane, Thomas J and Douglas O Staiger (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Tech. rep. National Bureau of Economic Research.

Kiefer, Jack and Jacob Wolfowitz (1956). "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". *The Annals of Mathematical Statistics*, pp. 887–906.

Kitagawa, Toru and Aleksey Tetenov (2018). "Who should be treated? empirical welfare maximization methods for treatment choice". *Econometrica* 86.2, pp. 591–616.

Kline, Patrick, Evan K Rose, and Christopher Walters (2023). "A Discrimination Report Card". *arXiv preprint arXiv:2306.13005*.

References VI

Kline, Patrick and Christopher Walters (2021). "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination". *Econometrica* 89.2, pp. 765–792.

Koenker, Roger and Jiaying Gu (2017). "REBayes: an R package for empirical Bayes mixture methods". *Journal of Statistical Software* 82, pp. 1–26.

Koenker, Roger and Ivan Mizera (2014). "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules". *Journal of the American Statistical Association* 109.506, pp. 674–685.

Kwon, Soonwoo (2021). "Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models". *EliScholar–A Digital Platform for Scholarly Publishing at Yal e*, p. 1.

Laliberté, Jean-William (2021). "Long-term contextual effects in education: Schools and neighborhoods". *American Economic Journal: Economic Policy* 13.2, pp. 336–377.

Manski, Charles F (2004). "Statistical treatment rules for heterogeneous populations". *Econometrica* 72.4, pp. 1221–1246.

References VII

Morris, Carl N (1983). "Parametric empirical Bayes inference: theory and applications". *Journal of the American statistical Association* 78.381, pp. 47–55.

Oliveira, Natalia L, Jing Lei, and Ryan J Tibshirani (2021). "Unbiased risk estimation in the normal means problem via coupled bootstrap techniques". *arXiv preprint arXiv:2111.09447*.

Polyanskiy, Yury and Yihong Wu (2020). "Self-regularizing property of nonparametric maximum likelihood estimator in mixture models". *arXiv preprint arXiv:2008.08244*.

— (2021). "Sharp regret bounds for empirical Bayes and compound decision problems". *arXiv preprint arXiv:2109.03943*.

Saha, Sujayam and Adityanand Guntuboyina (2020). "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising". *The Annals of Statistics* 48.2, pp. 738–762.

References VIII

Soloff, Jake A, Adityanand Guntuboyina, and Bodhisattva Sen (2021). "Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood". *arXiv preprint arXiv:2109.03466*.

Stock, James H and Mark W Watson (2012). "Generalized shrinkage methods for forecasting using many predictors". *Journal of Business & Economic Statistics* 30.4, pp. 481–493.

Stone, Charles J (1980). "Optimal rates of convergence for nonparametric estimators". *The annals of Statistics*, pp. 1348–1360.

Weinstein, Asaf et al. (2018). "Group-linear empirical Bayes estimates for a heteroscedastic normal mean". *Journal of the American Statistical Association* 113.522, pp. 698–710.

Xie, Xianchao, SC Kou, and Lawrence D Brown (2012). "SURE estimates for a heteroscedastic hierarchical model". *Journal of the American Statistical Association* 107.500, pp. 1465–1479.