Covariate-Powered Empirical Bayes Estimation

Nikos Ignatiadis
Stanford University

Joint work with Stefan Wager

December 2019

NeurIPS pre-proceedings

This talk

1. What is empirical Bayes (EB)?
2. What is EB with covariates?
3. What is our method and what are its statistical guarantees?

What is empirical Bayes? The setup

1. We care about point estimation of parameters corresponding to units $i = 1, \ldots, n$.
2. Motivated by classical statistical theory, we reduce information about each unit to one number for which we understand the sampling distribution, say:

$$Z_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right) \text{ for all } i$$

3. We look at all parameters μ_i simultaneously: Burden and blessing of multiplicity

Empirical Bayes (Robbins [1956], Efron [2010]) presents a principled approach for **learning from others**.

What is empirical Bayes? The "EB principle"

- "Let us use a mixed model, even if it might not be appropriate" (van Houwelingen, 2014)

What is empirical Bayes? The "EB principle"

- "Let us use a mixed model, even if it might not be appropriate" (van Houwelingen, 2014)
- ... to derive procedures with frequentist guarantees.

Example of EB: James-Stein [1961], Efron-Morris [1973]

- Gaussian compound decision problem (known σ^2):

$$Z_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right) \text{ independently for } i = 1, \ldots, n$$

- "Posit" that $\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\nu, A)$.
- The Bayes rule is: $t^*(z) = \mathbb{E}\left[\mu_i \,\middle|\, Z_i = z\right] = \frac{A}{\sigma^2+A} z + \frac{\sigma^2}{\sigma^2+A}\nu$
- Observe that marginally $Z_i \sim \mathcal{N}\left(\nu, \sigma^2 + A\right)$ so can estimate ν by $\bar{Z}$ and A by $\widehat{A}_{\text{JS}}$.
- Estimate μ_i by estimated Bayes rule:

$$\hat{\mu}_i^{\text{JS}} = \frac{\widehat{A}_{\text{JS}}}{\sigma^2 + \widehat{A}_{\text{JS}}} Z_i + \frac{\sigma^2}{\sigma^2 + \widehat{A}_{\text{JS}}} \bar{Z}$$

- The James-Stein estimator has frequentist guarantees.

JS for predicting batting averages

- Efron and Morris [1975], Brown [2008]
- For player i, observe AB_i at-bats and H_i hits during first half of season.
- Goal: Predict batting average in second half of season.
- $H_i \sim \text{Binomial}(AB_i, p_i)$ where p_i true "skill" of player i.
- Then let:

$$Z_i = \arcsin\left(\sqrt{\frac{H_i + 1/4}{AB_i + 1/2}}\right) \dot\sim \mathcal{N}\left(\arcsin(\sqrt{p_i}), \frac{1}{4AB_i}\right)$$

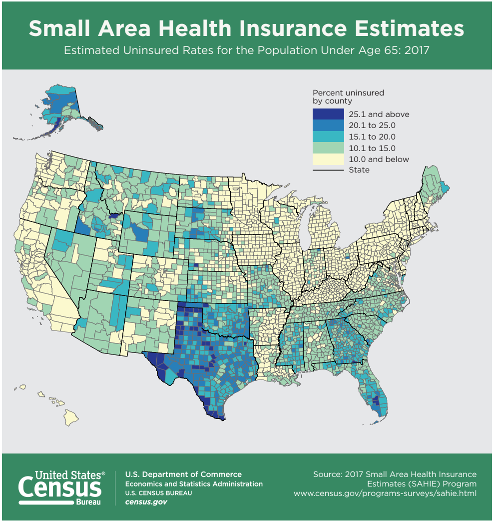
- Efron and Morris consider 18 players with 45 at-bats.
- Can then apply JS with $\sigma^2 = 1/(4 \cdot 45)$ to estimate $\arcsin(\sqrt{p_i})$.
- Then transform estimates back.

Brown [2008] batting results

Brown [2008] considers around 500 players:

	All batters; $\widehat{TSE}^*$	**All batters; $\widehat{TSE}^*_R$**	**All batters; $\widehat{TWSE}^*$**
$\mathcal{P}$ for estimation	567	567	567
$\mathcal{P}$ for validation	499	499	499
Naive	1	1	1
Group's mean	0.852	0.887	1.120 (0.741[1])
EB(MM)	0.593	0.606	0.626
EB(ML)	0.902	0.925	0.607
NP EB	0.508	0.509	0.560
Harmonic prior	0.884	0.905	0.600
James–Stein	0.525	0.540	0.502

Census data/ Small area estimation



Each i could be a:

- state
- commuting zone
- county
- city or town

Other application areas

- **Genomics:**
  - Gene expression profiling (each i is a gene)
  - Chemical compound screens (each i is a compound)
- **AB testing:**
  - Average treatment effects of multiple experiments or multiple treatment arms of the same experiment (Dimmery, Bakshy and Sekhon [2019])
  - Average treatment effects of one experiment on every advertiser

Empirical Bayes with side-information

- Gaussian compound decision problem:

$$Z_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right),\ i = 1, \ldots, n$$

- We know (Jiang and Cun-Hui Zhang [2009], Brown and Greenshtein [2009]) how to estimate $(\mu_1, \ldots, \mu_n)$ such that:

$$\mathbb{E}\left[\|\mu - \hat{\mu}\|^2\right] \text{ is small}$$

Empirical Bayes with side-information

- Gaussian compound decision problem:

$$Z_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right), \; i = 1, \ldots, n$$

- We know (Jiang and Cun-Hui Zhang [2009], Brown and Greenshtein [2009]) how to estimate $(\mu_1, \ldots, \mu_n)$ such that:

$$\mathbb{E}\left[\|\mu - \hat{\mu}\|^2\right] \text{ is small}$$

- What if we have side-information (covariates) X_i for each i, that may or may not be informative about μ_i?

Examples of side-information

- Batting: pitcher or non-pitcher, salary, team
- Genes: Ontologies
- AB tests: Percentage change in auxiliary metrics (Coey and Cunningham [2019])

Fay-Herriot model

- Census bureau in 1974
- Want to estimate per-capita income μ_i in small areas based on sample average Z_i.
- Covariates X_i: Per-capita income of whole county, value of owner-occupied housing, average adjusted gross income from older IRS returns
- Model:

$$\mu_i \mid X_i \sim \mathcal{N}\left(X_i^\top \beta, A\right)$$
$$Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$$

- Estimate β, A through method of moments

- Fay III, Robert E., and Roger A. Herriot. "Estimates of income for small places: an application of James-Stein procedures to census data." (JASA 1979)

Desiderata for a covariate-powered method

1. Analysis that allows for any black-box ML method, rather than tailored to specific predictor, e.g., linear regression as in Green and Strawderman (1991), Tan (2016), Kou and Yang (2017).
2. When covariates are non-informative: Come with similar guarantees as methods that do not use covariates.
3. When covariates are informative: Take advantage of additional information!

EB model with covariates

For a function $m(\cdot) : \mathcal{X} \to \mathbb{R}$ and $A, \sigma^2 > 0$:

$$\begin{aligned} X_i &\sim \mathbb{P}^X \\ \mu_i \mid X_i &\sim \mathcal{N}\left(m(X_i),\, A\right) \\ Z_i \mid \mu_i &\sim \mathcal{N}\left(\mu_i,\, \sigma^2\right) \end{aligned}$$

EB model with covariates

For a function $m(\cdot) : \mathcal{X} \to \mathbb{R}$ and $A, \sigma^2 > 0$:

$$\begin{aligned} X_i &\sim \mathbb{P}^X \\ \mu_i \mid X_i &\sim \mathcal{N}\left(m(X_i),\, A\right) \\ Z_i \mid \mu_i &\sim \mathcal{N}\left(\mu_i,\, \sigma^2\right) \end{aligned}$$

$$\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x)$$

EB model with covariates

For a function $m(\cdot) : \mathcal{X} \to \mathbb{R}$ and $A, \sigma^2 > 0$:

$$\begin{aligned} X_i &\sim \mathbb{P}^X \\ \mu_i \mid X_i &\sim \mathcal{N}\left(m(X_i),\, A\right) \\ Z_i \mid \mu_i &\sim \mathcal{N}\left(\mu_i,\, \sigma^2\right) \end{aligned}$$

$$\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x)$$

Goals: First understand EB shrinkage when model is true, then consider misspecification (for example deterministic μ_i).

Shrinkage versus non-parametric regression

$$X_i \sim \mathbb{P}^X, \;\; \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A\right), \;\; Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i,\, \sigma^2\right)$$

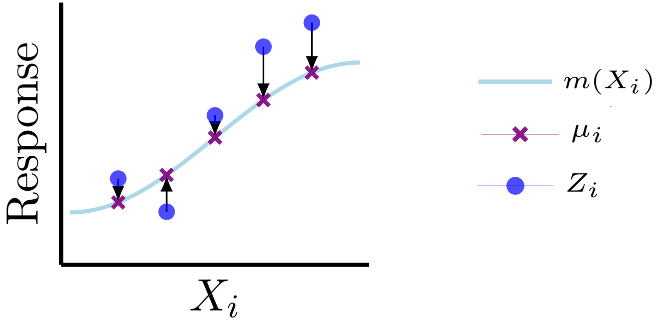
$$\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x)$$

Shrinkage versus non-parametric regression

$$X_i \sim \mathbb{P}^X, \quad \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A\right), \quad Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i,\, \sigma^2\right)$$

$$\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x)$$

If $A = 0$: $\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] = m(x)$

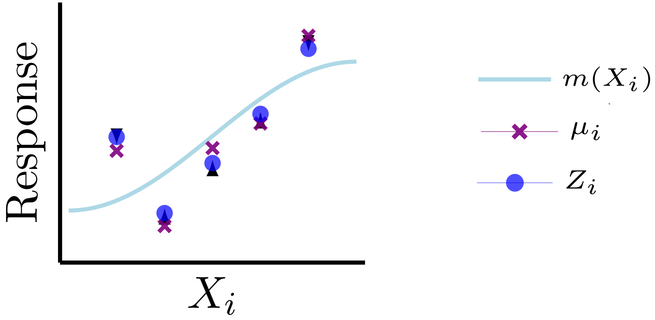


Shrinkage versus non-parametric regression

$$X_i \sim \mathbb{P}^X, \quad \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A\right), \quad Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i,\, \sigma^2\right)$$

$$\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x)$$

If $A \gg \sigma^2$: $\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x,\, Z_i = z\right] \approx z$

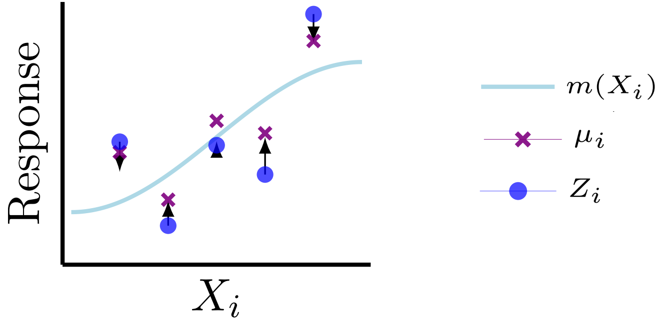


Shrinkage versus non-parametric regression

$$X_i \sim \mathbb{P}^X, \;\; \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i), A\right), \;\; Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$$

$$\mathbb{E}_{m,A}\left[\mu_i \mid X_i = x, Z_i = z\right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x)$$

If $A \approx \sigma^2$: Convex combination



The EB benchmark (Robbins [1964])

$$X_i \sim \mathbb{P}^X, \;\; \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A\right), \;\; Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i,\, \sigma^2\right),$$

- We observe n i.i.d. pairs (X_i, Z_i), not μ_i.

The EB benchmark (Robbins [1964])

$$X_i \sim \mathbb{P}^X, \;\; \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A\right), \;\; Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i,\, \sigma^2\right),$$

- We observe n i.i.d. pairs (X_i, Z_i), not μ_i.
- The task is to construct a function $\hat{t}_n(\cdot,\cdot) : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and we will use it to estimate μ_{n+1} by $\hat{t}_n(X_{n+1}, Z_{n+1})$ for a future draw $(\mu_{n+1}, X_{n+1}, Z_{n+1})$.

The EB benchmark (Robbins [1964])

$$X_i \sim \mathbb{P}^X, \;\; \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i), A\right), \;\; Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right),$$

- We observe n i.i.d. pairs (X_i, Z_i), not μ_i.
- The task is to construct a function $\hat{t}_n(\cdot,\cdot) : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and we will use it to estimate μ_{n+1} by $\hat{t}_n(X_{n+1}, Z_{n+1})$ for a future draw $(\mu_{n+1}, X_{n+1}, Z_{n+1})$.
- Benchmark in terms of regret. For a function $t : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ define:

$$L(t; m, A) := \mathbb{E}_{m,A}\left[\left(t(X_{n+1}, Z_{n+1}) - \mu_{n+1}\right)^2\right] - \frac{A\sigma^2}{A+\sigma^2}$$

The EB benchmark (Robbins [1964])

$$X_i \sim \mathbb{P}^X, \;\; \mu_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A\right), \;\; Z_i \mid \mu_i \sim \mathcal{N}\left(\mu_i,\, \sigma^2\right),$$

- We observe n i.i.d. pairs (X_i, Z_i), not μ_i.
- The task is to construct a function $\hat{t}_n(\cdot,\cdot) : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and we will use it to estimate μ_{n+1} by $\hat{t}_n(X_{n+1}, Z_{n+1})$ for a future draw $(\mu_{n+1}, X_{n+1}, Z_{n+1})$.
- Benchmark in terms of regret. For a function $t : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ define:

$$L(t; m, A) := \mathbb{E}_{m,A}\left[\left(t(X_{n+1}, Z_{n+1}) - \mu_{n+1}\right)^2\right] - \frac{A\sigma^2}{A + \sigma^2}$$

- We want $\mathbb{E}\left[L(\hat{t}_n; m, A)\right]$ to be small and close to 0.

Minimax EB regret

- A known, $\sigma^2 > 0$ known, regret incurred by not knowing $m(\cdot)$, but only that $m(\cdot) \in \mathcal{C}$.

Minimax EB regret

- A known, $\sigma^2 > 0$ known, regret incurred by not knowing $m(\cdot)$, but only that $m(\cdot) \in \mathcal{C}$.
- Minimax expected regret:

$$\mathfrak{M}_n^{\text{EB}}\left(\mathcal{C}; A, \sigma^2\right) := \inf_{\hat{t}_n} \max_{m \in \mathcal{C}} \left\{ \mathbb{E}_{m,A}\left[L(\hat{t}_n; m, A) \right] \right\}$$

Minimax EB regret

- A known, $\sigma^2 > 0$ known, regret incurred by not knowing $m(\cdot)$, but only that $m(\cdot) \in \mathcal{C}$.
- Minimax expected regret:

$$\mathfrak{M}_n^{\mathrm{EB}}\left(\mathcal{C}; A, \sigma^2\right) := \inf_{\hat{t}_n} \max_{m \in \mathcal{C}} \left\{ \mathbb{E}_{m,A}\left[L(\hat{t}_n; m, A) \right] \right\}$$

- We also have the minimax risk in the regression problem where we observe $X_i \sim \mathbb{P}^X, Z_i \mid X_i \sim \mathcal{N}\left(m(X_i), A + \sigma^2\right)$ and want to estimate $m(\cdot)$ w.r.t. $L^2(\mathbb{P}^X)$:

$$\mathfrak{M}_n^{\mathrm{Reg}}\left(\mathcal{C}; A + \sigma^2\right) := \inf_{\hat{m}_n} \max_{m \in \mathcal{C}} \mathbb{E}_{m,A}\left[\int \left(\hat{m}_n(x) - m(x)\right)^2 d\mathbb{P}^X \right]$$

Minimax EB regret

- A known, $\sigma^2 > 0$ known, regret incurred by not knowing $m(\cdot)$, but only that $m(\cdot) \in \mathcal{C}$.
- Minimax expected regret:

$$\mathfrak{M}_n^{\mathrm{EB}}\left(\mathcal{C}; A, \sigma^2\right) := \inf_{\hat{t}_n} \max_{m \in \mathcal{C}} \left\{ \mathbb{E}_{m,A}\left[L(\hat{t}_n; m, A) \right] \right\}$$

- We also have the minimax risk in the regression problem where we observe $X_i \sim \mathbb{P}^X, Z_i \mid X_i \sim \mathcal{N}\left(m(X_i), A + \sigma^2\right)$ and want to estimate $m(\cdot)$ w.r.t. $L^2(\mathbb{P}^X)$:

$$\mathfrak{M}_n^{\mathrm{Reg}}\left(\mathcal{C}; A + \sigma^2\right) := \inf_{\hat{m}_n} \max_{m \in \mathcal{C}} \mathbb{E}_{m,A}\left[\int (\hat{m}_n(x) - m(x))^2 \, d\mathbb{P}^X \right]$$

- Claim: EB Regret often satisfies

$$\mathfrak{M}_n^{\mathrm{EB}}\left(\mathcal{C}; A, \sigma^2\right) \asymp \frac{\sigma^4}{(\sigma^2 + A)^2} \mathfrak{M}_n^{\mathrm{Reg}}\left(\mathcal{C}; A + \sigma^2\right)$$

Minimax results: One example

- $\mathcal{X} = [0,1]^d$ with density f^X such that $\eta \leq f^X(u) \leq 1/\eta$, $\eta > 0$.
- Lipschitz class:

$$\mathrm{Lip}([0,1]^d, L) := \left\{ m : [0,1]^d \to \mathbb{R} : \left| m(x) - m(x') \right| \leq L \left\| x - x' \right\|_2 \right\}$$

- Then (I., Wager 2019):

$$\lim_{n\to\infty} \left| \log \left(\mathfrak{M}_n^{\mathrm{EB}} \left(\mathrm{Lip}([0,1]^d, L); A, \sigma^2 \right) \Big/ \frac{\sigma^4}{(\sigma^2 + A)^2} \cdot \left(\frac{L^d \left(\sigma^2 + A\right)}{n} \right)^{\frac{2}{2+d}} \right) \right|$$

$$\leq C_{\mathrm{Lip}}(d, \eta)$$

Minimax estimator: Known prior variance A

- Let $\widehat{m}(\cdot)$ achieve the minimax rate for estimating $m(\cdot)$ over $\mathcal{C}$.
- Then the following plug-in estimator achieves the Empirical Bayes minimax benchmark:

$$t^*_{\widehat{m},A}(x,z) = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} \widehat{m}(x)$$

Minimax estimator: Unknown prior variance A

What if A is unknown? Ansatz: Plug-in $\widehat{A}, \widehat{m}$

$$t^*_{\widehat{m},\widehat{A}}(x,z) = \frac{\widehat{A}}{\sigma^2 + \widehat{A}} z + \frac{\sigma^2}{\sigma^2 + \widehat{A}} \widehat{m}(x)$$

- Marginally $Z_i \mid X_i \sim \mathcal{N}\left(m(X_i), \sigma^2 + A\right)$.

Minimax estimator: Unknown prior variance A

What if A is unknown? Ansatz: Plug-in $\widehat{A}, \widehat{m}$

$$t^*_{\widehat{m},\widehat{A}}(x,z) = \frac{\widehat{A}}{\sigma^2 + \widehat{A}} z + \frac{\sigma^2}{\sigma^2 + \widehat{A}} \widehat{m}(x)$$

- Marginally $Z_i \mid X_i \sim \mathcal{N}\left(m(X_i), \sigma^2 + A\right)$.
- Idea 1: Estimate $\mathrm{Var}\left[Z_i \mid X_i\right] = \sigma^2 + A$ to get $\widehat{A + \sigma^2}$ and then $\widehat{A}$.

Minimax estimator: Unknown prior variance A

What if A is unknown? Ansatz: Plug-in $\widehat{A}, \widehat{m}$

$$t^*_{\widehat{m},\widehat{A}}(x,z) = \frac{\widehat{A}}{\sigma^2 + \widehat{A}} z + \frac{\sigma^2}{\sigma^2 + \widehat{A}} \widehat{m}(x)$$

- Marginally $Z_i \mid X_i \sim \mathcal{N}\left(m(X_i), \sigma^2 + A\right)$.
- Idea 1: Estimate $\mathrm{Var}\left[Z_i \mid X_i\right] = \sigma^2 + A$ to get $\widehat{A + \sigma^2}$ and then $\widehat{A}$.
- Idea 2: Say we use (deterministic) $\tilde{m}(\cdot) \neq m(\cdot)$, then even if we knew true A we would not want to use it, instead

$$A_{\tilde{m}} = \mathbb{E}\left[\left(\tilde{m}(X_{n+1}) - Z_{n+1}\right)^2\right] - \sigma^2 = A + \mathbb{E}\left[\left(\tilde{m}(X_{n+1}) - m(X_{n+1})\right)^2\right]$$

Sample-split EB

1. Form a partition of $\{1, \ldots, n\}$ into two folds I_1 and I_2.
2. Use observations in I_1, to estimate the regression $m(x) = \mathbb{E}\left[Z_i \mid X_i = x\right]$ by $\hat{m}_{I_1}(\cdot)$.
3. Use observations in I_2, to estimate A, through the formula:

$$\hat{A}_{I_2} = \left(\frac{1}{|I_2|} \sum_{i \in I_2} \left(\hat{m}_{I_1}(X_i) - Z_i\right)^2 - \sigma^2\right)_+$$

4. The estimated denoiser is then $\hat{t}_n^{\mathrm{EBCF}}(\cdot, \cdot) = t^*_{\hat{m}_{I_1}, \hat{A}_{I_2}}(\cdot, \cdot)$.

Sample-split EB

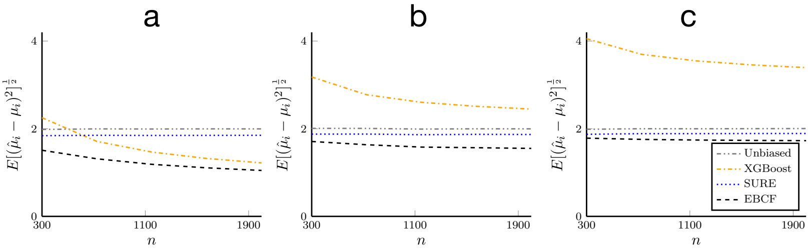
1. Form a partition of $\{1, \ldots, n\}$ into two folds I_1 and I_2.
2. Use observations in I_1, to estimate the regression $m(x) = \mathbb{E}\left[Z_i \mid X_i = x\right]$ by $\hat{m}_{I_1}(\cdot)$.
3. Use observations in I_2, to estimate A, through the formula:

$$\hat{A}_{I_2} = \left(\frac{1}{|I_2|} \sum_{i \in I_2} \left(\hat{m}_{I_1}(X_i) - Z_i\right)^2 - \sigma^2 \right)_+$$

4. The estimated denoiser is then $\hat{t}_n^{\mathrm{EBCF}}(\cdot, \cdot) = t^*_{\hat{m}_{I_1}, \hat{A}_{I_2}}(\cdot, \cdot)$.

Still achieves minimax rates without knowledge of A.

A small simulation



- Simulate from:

$$
\begin{aligned}
X_i &\sim U[0,1]^{15} \\
\mu_i \mid X_i &\sim \mathcal{N}\left(m(X_i),\ A\right) \\
Z_i \mid \mu_i &\sim \mathcal{N}\left(\mu_i,\ \sigma^2\right)
\end{aligned}
$$

- $m(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5$ [Friedman (1991)]
- $\sigma^2 = 4, A \in \{0, 4, 9\}$
- $\hat{m}$ cross-validated XGBoost

Empirical Bayes with Cross-Fitting (EBCF)

If we want to predict $\mu_1, \ldots, \mu_n$:

1. Form a partition of $\{1, \ldots, n\}$ into two folds I_1 and I_2.
2. Use observations in I_1, to estimate the regression $m(x) = \mathbb{E}[Z_i \mid X_i = x]$ by $\hat{m}_{I_1}(\cdot)$.
3. Use observations in I_2, to estimate A, through the formula:

$$\hat{A}_{I_2} = \left(\frac{1}{|I_2|} \sum_{i \in I_2} \left(\hat{m}_{I_1}(X_i) - Z_i \right)^2 - \sigma^2 \right)_+$$

4. The estimated denoiser is then $\hat{t}_n^{\text{EBCF}}(\cdot, \cdot) = t^*_{\hat{m}_{I_1}, \hat{A}_{I_2}}(\cdot, \cdot)$.
5. Estimate $\hat{\mu}_i^{\text{EBCF}} = t^*_{\hat{m}_{I_1}, \hat{A}_{I_2}}(X_i, Z_i)$ for $i \in I_2$
6. Repeat with folds I_1 and I_2 flipped.

James-Stein property

Assume indepedence and that:

$$Z_i \mid X_i, \mu_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$$

Then if $|I_1|, |I_2| \geq 5$:

James-Stein property

Assume indepedence and that:

$$Z_i \mid X_i, \mu_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$$

Then if $|I_1|, |I_2| \geq 5$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\mu_i - \hat{\mu}_i^{\mathrm{EBCF}})^2 \mid X_{1:n}, \mu_{1:n}\right] < \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\mu_i - Z_i)^2 \mid X_{1:n}, \mu_{1:n}\right] = \sigma^2$$

Further misspecification result

Now only assume that (and 4th moment condition on Z_i, bounds on μ_i)

$$\mathbb{E}\left[Z_i \mid \mu_i, X_i\right] = \mu_i, \;\; \mathrm{Var}\left[Z_i \mid \mu_i, X_i\right] = \sigma^2$$

Then:

Further misspecification result

Now only assume that (and 4th moment condition on Z_i, bounds on μ_i)

$$\mathbb{E}\left[Z_i \mid \mu_i, X_i\right] = \mu_i, \quad \mathrm{Var}\left[Z_i \mid \mu_i, X_i\right] = \sigma^2$$

Then:

$$\frac{1}{|I_2|}\sum_{i\in I_2}\mathbb{E}\left[(\mu_i - \hat{\mu}_i^{\mathrm{EBCF}})^2 \mid X_{I_2}, \mu_{I_2}\right] \leq \sigma^2 + O\left(\frac{1}{\sqrt{|I_2|}}\right)$$

$$\begin{aligned}\frac{1}{|I_2|}\sum_{i\in I_2}\mathbb{E}\left[(\mu_i - \hat{\mu}_i^{\mathrm{EBCF}})^2 \mid X_{I_2}, \mu_{I_2}\right] &\leq \frac{1}{|I_2|}\sum_{i\in I_2}\mathbb{E}\left[(\mu_i - \hat{m}_{I_1}(X_i))^2 \mid X_{I_2}, \mu_{I_2}\right] \\ &\quad + O\left(\frac{1}{\sqrt{|I_2|}}\right)\end{aligned}$$

Why does this work? SURE

- SURE: Stein's Unbiased Risk Estimate (Stein [1981])
- We may write $\hat{A}_{I_2}$ as:

$$\hat{A}_{I_2} = \left(\frac{1}{|I_2|} \sum_{i \in I_2} (\hat{m}_{I_1}(X_i) - Z_i)^2 - \sigma^2 \right)_+ \iff \hat{A}_{I_2} = \underset{A \geq 0}{\text{argmin}} \left\{ \text{SURE}_{I_2}(A) \right\},$$

$$\text{SURE}_{I_2}(A) := \frac{1}{|I_2|} \sum_{i \in I_2} \left(\sigma^2 + \frac{\sigma^4}{(A + \sigma^2)^2} (Z_i - \hat{m}_{I_1}(X_i))^2 - 2 \frac{\sigma^4}{A + \sigma^2} \right).$$

- SURE satisfies:

$$\mathbb{E}\left[\text{SURE}_{I_2}(A) \mid X_{1:n}, \mu_{1:n}\right] = \frac{1}{|I_2|} \sum_{i \in I_2} \mathbb{E}\left[\left(\mu_i - t^*_{\hat{m}_{I_1}, A}(X_i, Z_i) \right)^2 \mid X_{1:n}, \mu_{1:n} \right]$$

Heteroskedastic case

- In heteroskedastic setting, $\text{Var}\left[Z_i \,\middle|\, X_i, \mu_i\right] = \sigma_i^2$.
- Then (following Xie, Kou, Brown [2012] in setting without covariates): Consider estimators

$$t^*_{m,A}(X_i, Z_i, \sigma_i) = \frac{A}{\sigma_i^2 + A} Z_i + \frac{\sigma_i^2}{\sigma_i^2 + A} m(x)$$

- Pick A again by cross-fitting and SURE:

$$\hat{A}_{I_2} = \underset{A \geq 0}{\operatorname{argmin}} \left\{ \text{SURE}_{I_2}(A) \right\},$$

$$\text{SURE}_{I_2}(A) := \frac{1}{|I_2|} \sum_{i \in I_2} \left(\sigma_i^2 + \frac{\sigma_i^4}{(A + \sigma_i^2)^2} (Z_i - \hat{m}_{I_1}(X_i))^2 - 2 \frac{\sigma_i^4}{A + \sigma_i^2} \right)$$

MovieLens 20M (Harper and Konstan [2016])

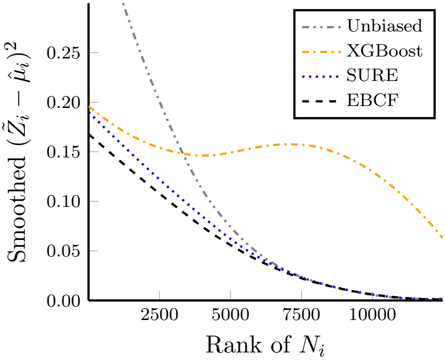
- 20 million ratings in $\{0, 0.5, \ldots, 5\}$ from 138,000 users applied to 27,000 movies.
- Keep 10% of users, calculate average rating Z_i for each movie based on N_i users.
- X_i include N_i, year of release, genres...
- Posit that $Z_i \mid \mu_i, X_i \sim (\mu_i, \sigma^2/N_i)$.
- "Ground-truth": $\widetilde{Z}_i$ average movie rating based on other 90% of users. Benchmark based on $\sum_{i=1}^{n} \left(\widetilde{Z}_i - \hat{\mu}_i\right)^2 / n$.
- Compare: Unbiased estimator Z_i, XGBoost predictor, EB without covariates (SURE) (Xie, Kou and Brown [2012]) and EBCF with XGBoost.

MovieLens results

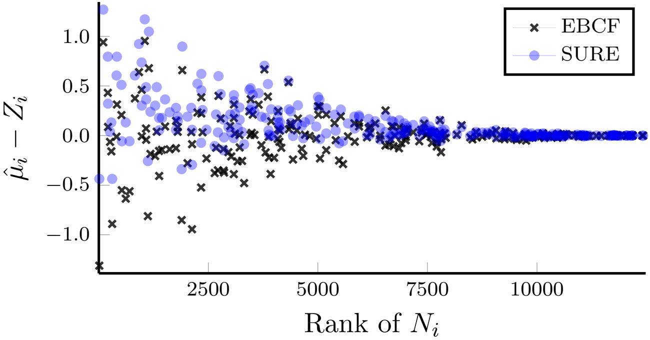
	All	Sci-Fi & Horror
Unbiased	0.098	0.098
XGBoost	0.145	0.183
SURE	0.061	0.064
EBCF	**0.055**	**0.052**

MovieLens results

	All	Sci-Fi & Horror
Unbiased	0.098	0.098
XGBoost	0.145	0.183
SURE	0.061	0.064
EBCF	**0.055**	**0.052**



MovieLens results



Future work: Variance modulation

- So far, the covariates have been modulating the prior mean $\mathbb{E}\left[\mu_i \mid X_i = x\right]$.
- For differential gene expression studies, often μ_i is the log-fold change of gene expression between two conditions:

$$\mathbb{E}\left[\mu_i \mid X_i = x\right] \approx 0$$

- Instead model covariates as modulating:

$$\mathbb{P}\left[\mu_i = 0 \mid X_i = x\right] \quad \text{or} \quad \mathrm{Var}\left[\mu_i \mid X_i = x\right]$$

Conclusion

- As argued in a series of papers by Efron and co-authors, Empirical Bayes presents a powerful framework for learning from others.
- In this work: How can we apply EB in the presence of rich side-information about each unit?
- Such side-information is ubiquitous and may be leveraged also in other setting, e.g., in Multiple Testing (Lei and Fithian [2016], I. and Huber [2017]).
- Key ideas: Cross-fitting, Stein's Unbiased Risk estimate

- Manuscript: `https://arxiv.org/abs/1906.01611` and NeurIPS 2019

Thank you for your attention!