Covariate-Powered Empirical Bayes Estimation

Nikos Ignatiadis
Stanford University

Joint work with Stefan Wager

December 2019
NeurIPS pre-proceedings
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 $\mu_i$ 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 $\sigma^2$):

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

- "Posit" that $\mu_i \overset{iid}{\sim} \mathcal{N}(\nu, A)$.

- The Bayes rule is:

$$t^*(z) = \mathbb{E} [\mu_i \mid Z_i = z] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} \nu$$

- Observe that marginally $Z_i \sim \mathcal{N}(\nu, \sigma^2 + A)$ so can estimate $\nu$ by $\bar{Z}$ and $A$ by $\hat{A}_{JS}$.

- Estimate $\mu_i$ by estimated Bayes rule:

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

- 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:

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

- 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:

<table>
<thead>
<tr>
<th></th>
<th>All batters; $TSE^*$</th>
<th>All batters; $TSE_R^*$</th>
<th>All batters; $TWSE^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$ for estimation</td>
<td>567</td>
<td>567</td>
<td>567</td>
</tr>
<tr>
<td>$P$ for validation</td>
<td>499</td>
<td>499</td>
<td>499</td>
</tr>
<tr>
<td>Naive</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Group’s mean</td>
<td>0.852</td>
<td>0.887</td>
<td>1.120 (0.741$^1$)</td>
</tr>
<tr>
<td>EB(MM)</td>
<td>0.593</td>
<td>0.606</td>
<td>0.626</td>
</tr>
<tr>
<td>EB(ML)</td>
<td>0.902</td>
<td>0.925</td>
<td>0.607</td>
</tr>
<tr>
<td>NP EB</td>
<td>0.508</td>
<td>0.509</td>
<td>0.560</td>
</tr>
<tr>
<td>Harmonic prior</td>
<td>0.884</td>
<td>0.905</td>
<td>0.600</td>
</tr>
<tr>
<td>James–Stein</td>
<td>0.525</td>
<td>0.540</td>
<td>0.502</td>
</tr>
</tbody>
</table>

2. The best performing predictors in order are those corresponding to the non-parametric empirical Bayes method, the James–Stein method, and the parametric EB(MM) method. The performance of the parametric EB(ML) method and the true (formal) Bayes harmonic prior method is mediocre. They perform about equally poorly; indeed, the two estimators are numerically very similar, which is not surprising if one looks closely at the motivation for each.

3a. There are two explanations for the relatively poor performance of the EB(ML) and the HB estimators. First, Figure 3 contains the histogram for the values of $\{X_i\}$. Note that this histogram does not match an ordinary distribution. In fact, as suggested by the results in Table 1, it appears to be better modeled as a F(3).

Histogram and box-plot for $\{X_i: N_1 \geq 11\}$. 

---

PREDICTION OF BATTING A VERAGES 133
TABLE 2
Values for half-season predictions for all batters of $\hat{TSE}^*$, $\hat{TSE}_R^*$ and $\hat{TWSE}^*$ as defined in (5.1), below, and the discussion afterward.
Estimated Uninsured Rates for the Population Under Age 65: 2017

Each \( i \) could be a:
- state
- commuting zone
- county
- city or town

Source: 2017 Small Area Health Insurance Estimates (SAHIE) Program
www.census.gov/programs-surveys/sahie.html
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:

\[ Z_i \sim \mathcal{N} (\mu_i, \sigma^2), \ i = 1, \ldots, n \]

- 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:

\[ Z_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad i = 1, \ldots, n \]

- 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 \(\mu_i\)?
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 $\mu_i$ 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:
  \[
  \mu_i \mid X_i \sim \mathcal{N}(X_i^T \beta, A)
  \]
  \[
  Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2)
  \]
- Estimate $\beta, A$ through method of moments
Desiderata for a covariate-powered method


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{align*}
X_i & \sim \mathcal{P}^X \\
\mu_i \mid X_i & \sim \mathcal{N}(m(X_i), A) \\
Z_i \mid \mu_i & \sim \mathcal{N}(\mu_i, \sigma^2)
\end{align*}
EB model with covariates

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

$$X_i \sim \mathbb{P}^X$$

$$\mu_i \mid X_i \sim \mathcal{N}(m(X_i), A)$$

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

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

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

\[
X_i \sim \mathbb{P}^X \\
\mu_i \mid X_i \sim \mathcal{N}(m(X_i), A) \\
Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2)
\]

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

Goals: First understand EB shrinkage when model is true, then consider misspecification (for example deterministic \( \mu_i \)).
Shrinkage versus non-parametric regression

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

\[ \mathbb{E}_{m,A} \left[ \mu_i \mid X_i = x, Z_i = z \right] = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} m(x) \]
Shrinkage versus non-parametric regression

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

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

If \( A = 0 \):
\[ \mathbb{E}_{m,A}[\mu_i \mid X_i = x, Z_i = z] = m(x) \]
Shrinkage versus non-parametric regression

\[ X_i \sim \mathcal{P}^X, \quad \mu_i \mid X_i \sim \mathcal{N}(m(X_i), A), \quad Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2) \]

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

If \( A \gg \sigma^2 \): \( \mathbb{E}_{m,A}[\mu_i \mid X_i = x, Z_i = z] \approx z \)
Shrinkage versus non-parametric regression

\[ X_i \sim P^X, \quad \mu_i \mid X_i \sim \mathcal{N}(m(X_i), A), \quad Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2) \]

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

If \( A \approx \sigma^2 \): Convex combination
The EB benchmark (Robbins [1964])

\[ X_i \sim \mathcal{P}^X, \quad \mu_i \mid X_i \sim \mathcal{N}(m(X_i), A), \quad Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2), \]

- We observe \( n \) i.i.d. pairs \((X_i, Z_i)\), not \( \mu_i \).
The EB benchmark (Robbins [1964])

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

- We observe \( n \) i.i.d. pairs \((X_i, Z_i)\), not \( \mu_i \).
- The task is to construct a function \( \hat{t}_n(\cdot, \cdot) : \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R} \) and we will use it to estimate \( \mu_{n+1} \) 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])

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

- We observe \( n \) i.i.d. pairs \((X_i, Z_i)\), not \( \mu_i \).
- 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 \( \mu_{n+1} \) 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])

\[ X_i \sim P^X, \quad \mu_i \mid X_i \sim \mathcal{N}(m(X_i), A), \quad Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2), \]

- We observe \( n \) i.i.d. pairs \((X_i, Z_i)\), not \( \mu_i \).
- 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 \( \mu_{n+1} \) 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:

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

- 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 $m(\cdot) \in C$. 

Minimax expected regret:

$$M_{EB} n(C; A, \sigma^2) := \inf \hat{m} n \max m \in C \left\{ \mathbb{E} m, A \left[ L(\hat{m} n; m, A) \right] \right\}$$

We also have the minimax risk in the regression problem where we observe $X_i \sim P_X$, $Z_i | X_i \sim N(m(X_i), A + \sigma^2)$ and want to estimate $m(\cdot)$ w.r.t. $L_2(P_X)$:

$$M_{Reg} n(C; A + \sigma^2) := \inf \hat{m} n \max m \in C \mathbb{E} m, A \left[ \int (\hat{m} n(x) - m(x))^2 dP_X \right]$$

Claim: EB Regret often satisfies

$$M_{EB} n(C; A, \sigma^2) \approx \sigma^4 (\sigma^2 + A)^2 M_{Reg} n(C; A + \sigma^2)$$
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:

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

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

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

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

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

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

$$\mathcal{M}_{\text{Reg}}^n(\mathcal{C}; A + \sigma^2) := \inf \max \mathbb{E}_{m,A} \left[ \int (\hat{m}_n(x) - m(x))^2 d\mathbb{P}X \right]$$

- Claim: EB Regret often satisfies

$$\mathcal{M}_{\text{EB}}^n(\mathcal{C}; A, \sigma^2) \asymp \frac{\sigma^4}{(\sigma^2 + A)^2} \mathcal{M}_{\text{Reg}}^n(\mathcal{C}; A + \sigma^2)$$
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:

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

- Then (I., Wager 2019):

\[
\lim_{n \to \infty} \left| \log \left( \mathcal{M}_n^\text{EB} \left( \text{Lip}([0,1]^d, L); A, \sigma^2 \right) \right) \right| \geq \frac{\sigma^4}{(\sigma^2 + A)^2} \cdot \left( \frac{L^d (\sigma^2 + A)}{n} \right)^{\frac{2}{2+d}} \leq C_{\text{Lip}}(d, \eta)
\]
Minimax estimator: Known prior variance $A$

Let $\hat{m}(\cdot)$ achieve the minimax rate for estimating $m(\cdot)$ over $\mathcal{C}$.

Then the following plug-in estimator achieves the Empirical Bayes minimax benchmark:

$$t_{\hat{m},A}^*(x, z) = \frac{A}{\sigma^2 + A} z + \frac{\sigma^2}{\sigma^2 + A} \hat{m}(x)$$
Minimax estimator: Unknown prior variance $A$

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

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

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

Minimax estimator: Unknown prior variance $A$

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

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

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

Minimax estimator: Unknown prior variance $A$

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

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

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

$$A_{\tilde{m}} = \mathbb{E} \left[ (\tilde{m}(X_{n+1}) - Z_{n+1})^2 \right] - \sigma^2 = A + \mathbb{E} \left[ (\tilde{m}(X_{n+1}) - m(X_{n+1}))^2 \right]$$
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
   \(m(x) = \mathbb{E} [Z_i \mid X_i = x]\) by \(\hat{m}_{I_1}(\cdot)\).

3. Use observations in \(I_2\), to estimate \(A\), through the formula:

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

4. The estimated denoiser is then \(\hat{t}_n^{\text{EBCF}}(\cdot, \cdot) = t^*_{\hat{m}_{I_1}, \hat{A}_{I_2}}(\cdot, \cdot)\).
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 \( m(x) = \mathbb{E} [Z_i \mid X_i = x] \) by \( \hat{m}_{I_1}(\cdot) \).

3. Use observations in \( I_2 \), to estimate \( A \), through the formula:

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

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

Still achieves minimax rates without knowledge of \( A \).
A small simulation

Simulate from:

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

\[ m(x) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5 \text{ [Friedman (1991)]} \]

\[ \sigma^2 = 4, A \in \{0, 4, 9\} \]

\[ \hat{m} \text{ cross-validated XGBoost} \]
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:

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

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

\[ Z_i \mid X_i, \mu_i \sim \mathcal{N}(\mu_i, \sigma^2) \]

Then if \(|l_1|, |l_2| \geq 5|:
James-Stein property

Assume independence and that:

\[ Z_i \mid X_i, \mu_i \sim \mathcal{N} (\mu_i, \sigma^2) \]

Then if \( |l_1|, |l_2| \geq 5\):

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

Now only assume that (and 4th moment condition on $Z_i$, bounds on $\mu_i$)

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

Then:
Further misspecification result

Now only assume that (and 4th moment condition on $Z_i$, bounds on $\mu_i$)

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

Then:

$$\frac{1}{|l_2|} \sum_{i \in l_2} \mathbb{E} \left[ \left( \mu_i - \hat{\mu}_i^{\text{EBCF}} \right)^2 \mid X_{l_2}, \mu_{l_2} \right] \leq \sigma^2 + O \left( \frac{1}{\sqrt{|l_2|}} \right)$$

$$\frac{1}{|l_2|} \sum_{i \in l_2} \mathbb{E} \left[ \left( \mu_i - \hat{\mu}_i \right)^2 \mid X_{l_2}, \mu_{l_2} \right] \leq \frac{1}{|l_2|} \sum_{i \in l_2} \mathbb{E} \left[ \left( \mu_i - \hat{m}_1(X_i) \right)^2 \mid X_{l_2}, \mu_{l_2} \right] + O \left( \frac{1}{\sqrt{|l_2|}} \right)$$
SURE: Stein’s Unbiased Risk Estimate (Stein [1981])

We may write \( \hat{A}_{l_2} \) as:

\[
\hat{A}_{l_2} = \left( \frac{1}{|l_2|} \sum_{i \in l_2} (\hat{m}_{l_1}(X_i) - Z_i)^2 - \sigma^2 \right) \quad \iff \quad \hat{A}_{l_2} = \arg\min_{A \geq 0} \{ \text{SURE}_{l_2}(A) \},
\]

\[
\text{SURE}_{l_2}(A) := \frac{1}{|l_2|} \sum_{i \in l_2} \left( \sigma^2 + \frac{\sigma^4}{(A + \sigma^2)^2} (Z_i - \hat{m}_{l_1}(X_i))^2 - 2 \frac{\sigma^4}{A + \sigma^2} \right).
\]

SURE satisfies:

\[
\mathbb{E} [\text{SURE}_{l_2}(A) \mid X_{1:n}, \mu_{1:n}] = \frac{1}{|l_2|} \sum_{i \in l_2} \mathbb{E} \left[ \left( \mu_i - t^*_{\hat{m}_{l_1}, A}(X_i, Z_i) \right)^2 \mid X_{1:n}, \mu_{1:n} \right]
\]
Heteroskedastic case

- In heteroskedastic setting, $\text{Var} \left[ Z_i \mid X_i, \mu_i \right] = \sigma_i^2$.
- 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:

$$\hat{A}_{l_2} = \arg\min_{A \geq 0} \{ \text{SURE}_{l_2}(A) \},$$

$$\text{SURE}_{l_2}(A) := \frac{1}{|l_2|} \sum_{i \in l_2} \left( \sigma_i^2 + \frac{\sigma_i^4}{(A + \sigma_i^2)^2} (Z_i - \hat{m}_{l_1}(X_i))^2 - 2 \frac{\sigma_i^4}{A + \sigma_i^2} \right)$$
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”: $\tilde{Z}_i$ average movie rating based on other 90% of users. Benchmark based on $\sum_{i=1}^n \left( \tilde{Z}_i - \hat{\mu}_i \right)^2 / n$.
- Compare: Unbiased estimator $Z_i$, XGBoost predictor, EB without covariates (SURE) (Xie, Kou and Brown [2012]) and EBCF with XGBoost.
MovieLens results

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>Sci-Fi &amp; Horror</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>0.098</td>
<td>0.098</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.145</td>
<td>0.183</td>
</tr>
<tr>
<td>SURE</td>
<td>0.061</td>
<td>0.064</td>
</tr>
<tr>
<td>EBCF</td>
<td><strong>0.055</strong></td>
<td><strong>0.052</strong></td>
</tr>
</tbody>
</table>
### MovieLens results

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Sci-Fi &amp; Horror</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>0.098</td>
<td>0.098</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.145</td>
<td>0.183</td>
</tr>
<tr>
<td>SURE</td>
<td>0.061</td>
<td>0.064</td>
</tr>
<tr>
<td>EBCF</td>
<td><strong>0.055</strong></td>
<td><strong>0.052</strong></td>
</tr>
</tbody>
</table>
MovieLens results

\[ \hat{\mu}_i - Z_i \]

Rank of \( N_i \)

\( \hat{\mu}_i - Z_i \) vs. Rank of \( N_i \)

- EBCF
- SURE
Future work: Variance modulation

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

$$\mathbb{E} [\mu_i | X_i = x] \approx 0$$

- Instead model covariates as modulating:

$$\mathbb{P} [\mu_i = 0 | X_i = x] \quad \text{or} \quad \text{Var} [\mu_i | X_i = x]$$
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

Thank you for your attention!