Covariate-Powered Empirical Bayes Estimation

Nikolaos Ignatiadis[1] and Stefan Wager[2]

[1] Department of Statistics, Stanford University (ignat@stanford.edu)
[2] Graduate School of Business, Stanford University (swager@stanford.edu)

Setup

Model I: We observe ($i = 1, \dots, n$)

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

where $m(\cdot) : \mathcal{X} \to \mathbb{R}$ and $A > 0$ unknown, $\sigma^2 > 0$ known.

Goal: Estimate μ_i by $\hat{\mu}_i$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(\mu_i - \hat{\mu}_i)^2] \text{ is small}$$

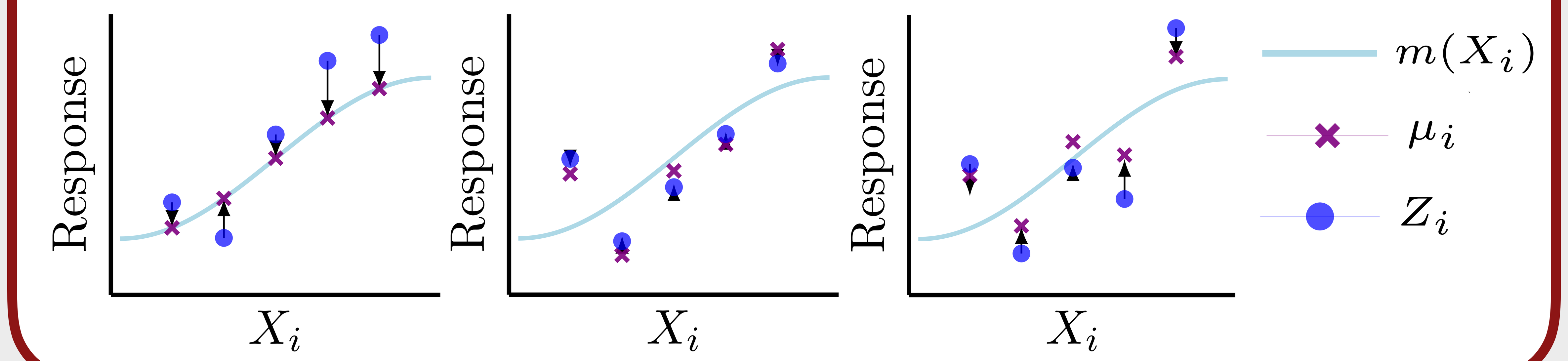
Benchmark: The Bayes rule

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

$A = 0$: $t^*_{m,A}(x, z) = m(x)$

$A \gg \sigma^2$: $t^*_{m,A}(x, z) \approx z$

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



Empirical Bayes with Cross-Fitting

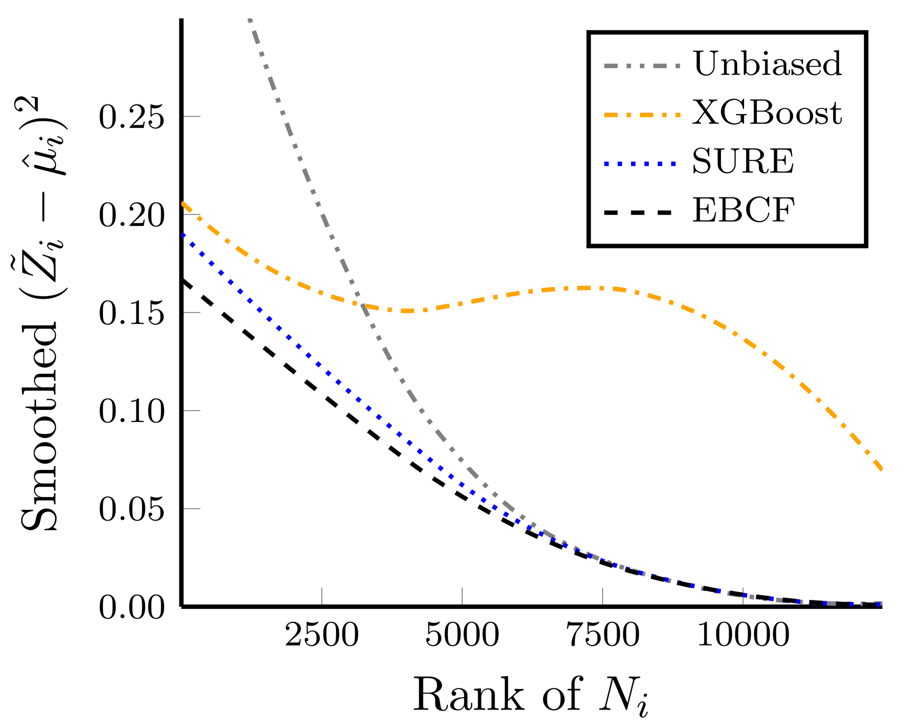
1. Form a partition of $\{1, \dots, n\}$ into two folds I_1 and I_2.
2. Use observations in I_1, to estimate the regression $m(x) = \mathbb{E}\left[Z_i \mid X_i = x\right]$ by $\hat{m}_{I_1}(\cdot)$.
3. Use observations in I_2, to estimate A, through the formula
$$\hat{A}_{I_2} = \left(\frac{1}{|I_2|}\sum_{i \in I_2}\left(\hat{m}_{I_1}(X_i) - Z_i\right)^2 - \sigma^2\right)_+$$
4. For $i \in I_2$, estimate μ_i by $\hat{\mu}_i^{\text{EBCF}} = t^*_{\hat{m}_{I_1}, \hat{A}_{I_2}}(X_i, Z_i)$.
5. Repeat with folds I_1 and I_2 flipped.

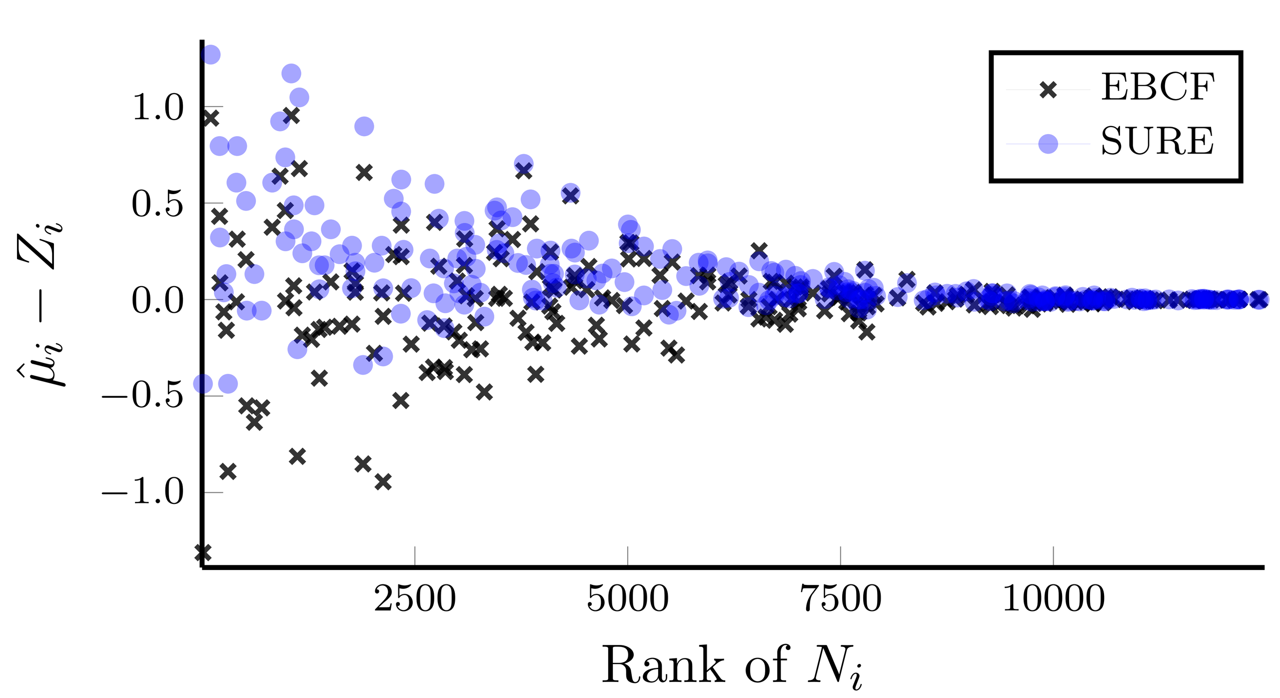
MovieLens [2] data analysis

- 20 million ratings in $\{0, 0.5, \dots, 5\}$ from 138,000 users applied to 27,000 movies ($n \geq 10{,}000$ after filtering).
- Keep 10% of users, calculate average rating Z_i for each movie based on N_i users.
- X_i includes N_i, year of release, genres.
- μ_i is "true" movie rating.
- Posit that $Z_i \mid \mu_i, X_i \sim (\mu_i, \sigma^2/N_i)$.
- "Ground-truth": $\widetilde{Z}_i$, the average movie rating based on other 90% of users.
- Evaluation by mean-squared error: $\sum_{i=1}^n (\widetilde{Z}_i - \hat{\mu}_i)^2/n$

$\sum_{i=1}^n (\widetilde{Z}_i - \hat{\mu}_i)^2/n$

		All	Sci-Fi & Horror
Z_i →	Unbiased	0.098 (± 0.005)	0.098 (± 0.032)
[1] →	XGBoost	0.150 (± 0.005)	0.210 (± 0.036)
[4] →	SURE	0.061 (± 0.002)	0.064 (± 0.018)
This work →	EBCF (with XGBoost)	**0.055** (± 0.002)	**0.051** (± 0.012)





EBCF is minimax optimal

Model I: $\implies X_i \overset{\text{iid}}{\sim} \mathbb{P}^X, \quad Z_i \mid X_i \sim \mathcal{N}\left(m(X_i),\, A + \sigma^2\right)$

Minimax regression error over $\mathcal{C} \subset \{f : \mathcal{X} \to \mathbb{R}\}$

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

Minimax empirical Bayes excess risk [3] over $\mathcal{C}$, with $A > 0$ fixed (but unknown)

$$\mathfrak{M}_n^{\text{EB}}\left(\mathcal{C}; A, \sigma^2\right) \mathrel{\widehat{=}} \inf_{\hat{t}_n} \max_{m \in \mathcal{C}} \left\{\text{Expected risk of } \hat{t}_n \; - \; \text{Bayes risk}\right\}$$

Theorem: For many $\mathcal{C}$, e.g., Lipschitz class in $\mathbb{R}^d$

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

EBCF is robust to misspecification

Model II: Non-Gaussian, equal variances

$$(X_i, \mu_i, Z_i) \sim \mathbb{P}^{(X_i, \mu_i, Z_i)}, \;\; \mathbb{E}\left[Z_i \mid \mu_i, X_i\right] = \mu_i, \;\; \text{Var}\left[Z_i \mid \mu_i, X_i\right] = \sigma^2$$

Guarantees for EBCF in fold I_2 (under bounded $\mathbb{E}\left[Z_i^4 \mid \mu_i, X_i\right], \mu_i$):

$$\frac{1}{|I_2|}\sum_{i \in I_2} \mathbb{E}\left[\left(\mu_i - \hat{\mu}_i^{\text{EBCF}}\right)^2\right] \leq \left\{\frac{\sigma^2}{\frac{1}{|I_2|}\sum_{i \in I_2} \mathbb{E}\left[\left(\mu_i - \hat{m}_{I_1}(X_i)\right)^2\right]}\right\} + O\left(\frac{1}{\sqrt{|I_2|}}\right)$$

EBCF can be extended (with similar guarantees) to

Model III: Non-Gaussian, unequal variances

$$(X_i, \mu_i, Z_i) \sim \mathbb{P}^{(X_i, \mu_i, Z_i)}, \;\; \mathbb{E}\left[Z_i \mid \mu_i, X_i\right] = \mu_i, \;\; \text{Var}\left[Z_i \mid \mu_i, X_i\right] = \sigma_i^2$$

Resources

Code availability

Software: https://github.com/nignatiadis/EBayes.jl

Reproducibility: https://github.com/nignatiadis/EBCrossFitPaper

References

[1] Tianqi Chen and Carlos Guestrin. *ACM SIGKDD* 22:785–794, 2016.
[2] F Maxwell Harper and Joseph A Konstan. *ACM TIIS*, 5(4):19, 2016.
[3] Herbert Robbins. *Annals of Mathematical Statistics*, 35:1–20, 1964.
[4] Xianchao Xie, SC Kou, and Lawrence D Brown. *JASA*, 107(500):1465–1479, 2012.

Acknowledgments

This research was funded by a gift from Google.