Empirical Bayes: Fundamentals (Part I)

ECON 2400: Applied Econometrics II

Juan C. Yamin

Brown University

Spring 2026

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

Empirical Bayes Applications

Applied work increasingly studies many unit-specific parameters:

- teacher and school value-added
- firm wage premia and manager effects
- neighborhood effects
- judges, hospitals, doctors, police officers, and more

In settings with many unit-specific parameters, **empirical Bayes (EB)** methods are useful for

- learning about the distribution of parameters across units
- improving estimates for individual units (*borrowing strength*)
- making decisions and reporting results
  Policy: what to do? Scientific: what to report?

A motivating example: teacher value-added

Suppose student j test outcomes $\tilde{Y}_{ij}$ taught by teacher i satisfy:

$$\tilde{Y}_{ij} = \alpha_i + X_{ij}'\beta + u_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, J_i.$$

- X_{ij} captures observed determinants of achievement (e.g., family background).
- α_i is the value-added of teacher i.
- We are interested in estimating α_i for many teachers.

Let $\hat{\alpha}_i$ denote a first-step estimate of teacher i's value-added. A useful approximation is

$$\hat{\alpha}_i \mid \alpha_i, \sigma_i \approx N(\alpha_i, \sigma_i^2), \qquad \text{where } \sigma_i \text{ is the standard error of } \hat{\alpha}_i.$$

Key feature

Different teachers are estimated with different precision: some σ_i's are small, others are large.

What questions do we care about?

In this setting, at least three questions arise.

- *Individual effects:* what is the value-added of a particular teacher?
- *Features of the distribution:* how much do teacher effects vary across teachers?
- *Decisions and reporting:* how should we rank, select, or summarize teachers when estimates are noisy?

Examples:

- Which teachers appear especially effective or ineffective?
- How much of the observed dispersion in $\hat{\alpha}_i$ reflects real differences rather than sampling noise?
- Which teachers should receive support, intervention, or closer review?

Why empirical Bayes?

Empirical Bayes is useful for all three: learning the distribution of effects, improving unit-level estimates, and guiding decisions or reporting.

Roadmap for today

Today's goal

Build the core **theoretical foundations** of empirical Bayes.

Today we will focus on:

- the statistical decision-theoretic setup
- the normal-normal empirical Bayes model
- James-Stein shrinkage and its empirical Bayes interpretation
- a brief introduction to nonparametric empirical Bayes

Next class: empirical applications

Next class will be **entirely applied**: school value-added, firm effects, place effects, discrimination, and how empirical Bayes is used in practice.

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

Why do we need statistical decision theory?

Suppose we want to estimate the value-added of a teacher.

From the data, we could use many different estimators:

- the raw estimate, a conservative estimate, or something else.

Core question

How do we decide which estimator is *better*?

Statistical decision theory gives a formal way to answer this question. It is useful to think of estimation as a simple game:

- **Nature** chooses the true teacher effect,
- **data** give us noisy information about it,
- **we** choose what to report,
- and then we evaluate how costly that choice was.

The goal is to choose an estimator that performs well in this game.

A statistical decision problem: one teacher

- Parameter: the unknown object we care about. In this case, the teacher's true value-added is α.
- Statistical model: how the observed data are generated given the truth. Our model is $\hat{\alpha} \mid \alpha \sim \mathcal{N}(\alpha, \sigma^2)$
- Action: a possible choice we could make. Here, the action is a reported teacher effect $a \in \mathbb{R}$.
- Loss function: how we measure the cost of a decision. For example, under squared error loss,
$$\ell(a, \alpha) = (a - \alpha)^2.$$
- Decision rule: a rule/function that maps the observed data into an action. After observing $\hat{\alpha}$, we report $a = \delta(\hat{\alpha})$.

Objective

Choose a decision rule δ that leads to low average loss.

Frequentist view: the teacher effect is fixed

Suppose the teacher's true value-added α is *fixed but unknown*.

We still observe a noisy estimate $\hat{\alpha} \mid \alpha \sim \mathcal{N}(\alpha, \sigma^2)$ and use a decision rule (estimator) $\delta(\hat{\alpha})$.

The frequentist risk of δ at α is

$$R(\delta, \alpha) = \mathbb{E}_\alpha[\ell(\delta(\hat{\alpha}), \alpha)] .$$

Under squared error loss,

$$R(\delta, \alpha) = \mathbb{E}_\alpha\left[(\delta(\hat{\alpha}) - \alpha)^2\right] .$$

Intuition

Fix the teacher's true value-added at α. Then imagine repeatedly drawing new noisy estimates $\hat{\alpha}$ from the same teacher.
The risk is the *average loss* of the decision rule across those repeated samples.

Bayesian view: the teacher effect is drawn from a prior

Now suppose the teacher's true value-added is itself random:

$$\alpha \sim \pi, \qquad \hat{\alpha} \mid \alpha \sim \mathcal{N}(\alpha, \sigma^2).$$

For a given decision rule δ, the Bayes risk is $r(\delta, \pi) = \mathbb{E}_\pi[R(\delta, \alpha)]$.
Under squared error loss, this becomes $r(\delta, \pi) = \mathbb{E}_\pi\left[\mathbb{E}_\alpha\left[(\delta(\hat{\alpha}) - \alpha)^2\right]\right]$.

Intuition

The Bayesian view adds one more layer of averaging.

- First, for each fixed α, evaluate the rule using its frequentist risk.
- Then average that risk over the prior distribution π.

So a good Bayesian rule is one that performs well *on average over the prior* π.

Bayes rule under squared error loss

Suppose

$$\alpha \sim \pi, \qquad \hat{\alpha} \mid \alpha \sim \mathcal{N}(\alpha, \sigma^2).$$

Result

Under squared error loss,

$$\ell(a, \alpha) = (a - \alpha)^2,$$

the Bayes-optimal decision rule is

$$\delta^B(\hat{\alpha}) = \mathbb{E}[\alpha \mid \hat{\alpha}].$$

Posterior distribution: the distribution of the unknown teacher effect α *after* observing the noisy estimate $\hat{\alpha}$.

So under squared error loss, the optimal Bayesian estimator is the **posterior mean**.

(Under absolute loss, the optimal rule is the posterior median.)

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

From one teacher to many teachers

Let us go back to the problem we actually care about.

Now we want to estimate the value-added of *many* teachers: $\alpha = (\alpha_1, \ldots, \alpha_n)$.

For each teacher i, we observe a noisy estimate $\hat{\alpha}_i \mid \alpha_i \sim \mathcal{N}(\alpha_i, \sigma_i^2)$, $i = 1, \ldots, n$.

- Parameters: the vector of true teacher effects $(\alpha_1, \ldots, \alpha_n)$
- Actions: reported estimates $(a_1, \ldots, a_n)$
- Decision rule: a map from $(\hat{\alpha}_1, \ldots, \hat{\alpha}_n)$ into reported values: $\delta_i(\hat{\alpha}_1, \ldots, \hat{\alpha}_n)$
- Loss: how costly the whole vector of reported estimates is $L(a, \alpha) = \frac{1}{n} \sum_{i=1}^{n} (a_i - \alpha_i)^2$.

Key difference from before

We no longer care only about one teacher in isolation. We care about how well our rule performs *across many teachers at once*.

Normal-normal Bayes

Suppose that for each teacher i,

$$\hat{\alpha}_i \mid \alpha_i, \sigma_i \sim \mathcal{N}(\alpha_i, \sigma_i^2), \qquad \alpha_i \sim \mathcal{N}(\mu, \tau^2).$$

Interpretation:

- $\hat{\alpha}_i$ is the noisy estimate we observe from the data
- σ_i^2 is the sampling variance of teacher i's estimate
- μ is the average teacher effect in the population
- τ^2 is the true heterogeneity across teachers

Posterior mean

Under squared error loss, the Bayes rule is

$$\mathbb{E}[\alpha_i \mid \hat{\alpha}_i] = \frac{\tau^2}{\tau^2 + \sigma_i^2}\hat{\alpha}_i + \frac{\sigma_i^2}{\tau^2 + \sigma_i^2}\mu.$$

What is still missing?

So far, we have solved a **Bayesian** normal-normal model:

$$\hat{\alpha}_i \mid \alpha_i, \sigma_i \sim \mathcal{N}(\alpha_i, \sigma_i^2), \qquad \alpha_i \sim \mathcal{N}(\mu, \tau^2).$$

The posterior mean is

$$\mathbb{E}[\alpha_i \mid \hat{\alpha}_i] = \frac{\tau^2}{\tau^2 + \sigma_i^2}\hat{\alpha}_i + \frac{\sigma_i^2}{\tau^2 + \sigma_i^2}\mu.$$

But there is a problem

In practice, we usually do *not* know μ or τ^2.

So the key question becomes:

How can we learn μ and τ^2 from the data?

That is the **empirical** part of empirical Bayes.

Estimating the hyperparameters

So far, the posterior mean depends on two unknown quantities:

$$\mu \qquad \text{and} \qquad \tau^2.$$

A simple empirical Bayes approach is to estimate them from the ensemble of teacher estimates:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i,$$

$$\hat{\tau}^2 = \frac{1}{n} \sum_{i=1}^{n} \left[(\hat{\alpha}_i - \hat{\mu})^2 - \sigma_i^2 \right].$$

Interpretation

- $\hat{\mu}$ estimates the average teacher effect.
- $\hat{\tau}^2$ estimates the true heterogeneity across teachers.

The empirical Bayes estimator

Plugging $\hat{\mu}$ and $\hat{\tau}^2$ into the posterior mean gives

$$\hat{\alpha}_i^{EB} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}\hat{\alpha}_i + \frac{\sigma_i^2}{\hat{\tau}^2 + \sigma_i^2}\hat{\mu}.$$

This is the **empirical Bayes** estimator:

- **Bayes:** it has the same form as the posterior mean in the normal-normal model
- **Empirical:** the hyperparameters are estimated from the data

- "Borrow strength" from the ensemble through $(\hat{\mu}, \hat{\sigma}^2)$.
- This is the workhorse parametric EB estimator in value-added studies.

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

A frequentist payoff of empirical Bayes shrinkage

So far, shrinkage came from a normal-normal empirical Bayes model.

Now consider the homoskedastic many-teacher problem $\hat{\alpha}_i \mid \alpha_i \sim \mathcal{N}(\alpha_i, 1)$, $i = 1, \ldots, n$, and evaluate estimators using **compound squared-error risk**

$$R(\delta, \alpha) = \mathbb{E}_\alpha \left[\frac{1}{n} \sum_{i=1}^{n} \left(\delta_i(\hat{\alpha}) - \alpha_i \right)^2 \right].$$

James-Stein phenomenon

Let the plug-in rule be the vector $\delta^{plug}(\hat{\alpha}) = \hat{\alpha}$. When $n \geq 3$,

$$R(\delta^{EB}, \alpha) \leq R(\delta^{plug}, \alpha) \qquad \text{for all } \alpha \in \mathbb{R}^n.$$

- The improvement is for the **whole vector**, not necessarily for each teacher one by one.
- So shrinkage is not only a Bayesian modeling trick: it also comes with a strong **frequentist** guarantee.

Why can empirical Bayes lower compound risk?

Under squared error loss, risk depends on both **bias** and **variance**.

In the homoskedastic case, the empirical Bayes rule takes the form

$$\delta_i^{EB}(\hat{\alpha}) = \mu + B(\hat{\alpha}_i - \mu), \qquad B = \frac{\tau^2}{\tau^2 + 1} \in (0, 1).$$

Key intuition

- The plug-in rule has no bias, but it can have a lot of variance.
- The empirical Bayes rule introduces some bias by pulling estimates toward μ.
- But it also reduces variance, because only a fraction B of the noise remains.

More precisely, for each coordinate i, $\mathrm{Var}_\alpha(\delta_i^{EB}(\hat{\alpha})) = B^2 < 1 = \mathrm{Var}_\alpha(\delta_i^{plug}(\hat{\alpha}))$. If the reduction in variance is large enough, it can outweigh the added bias and lower **average risk across coordinates**.

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

What have we assumed so far?

Our parametric empirical Bayes analysis relied on four key assumptions:

- Approximate normality of estimates: $\hat{\alpha}_i \mid \alpha_i, \sigma_i \sim \mathcal{N}(\alpha_i, \sigma_i^2)$.
- Known or estimable precision: we observe the standard errors σ_i.
- A parametric prior: $\alpha_i \sim \mathcal{N}(\mu, \tau^2)$.
- Independent units: teacher effects are treated as independent draws from the same population distribution.

Why this is useful

These assumptions make the model transparent and fully tractable.

Why this may be restrictive

In practice, the distribution of teacher effects may be skewed, fat-tailed, or even multi-modal.

Why go beyond the normal prior?

The parametric EB rule assumes that the cross-teacher distribution is normal:

$$G = \mathcal{N}(\mu, \tau^2).$$

That is convenient, but it may be wrong.

- The true distribution of teacher effects may be asymmetric.
- There may be unusually strong or unusually weak teachers in the tails.
- A single normal distribution may smooth away important features of heterogeneity.

Frontier idea

Instead of estimating only μ and τ^2, try to estimate the *entire* mixing distribution G.

That is the basic idea of **nonparametric empirical Bayes**.

Nonparametric empirical Bayes via G-modeling

Keep the same first-step sampling model:

$$\hat{\alpha}_i \mid \alpha_i, \sigma_i \sim \mathcal{N}(\alpha_i, \sigma_i^2), \qquad \alpha_i \sim G.$$

The difference is that now we do *not* assume $G = \mathcal{N}(\mu, \tau^2)$.

Instead, we estimate G flexibly from the ensemble of noisy estimates

$$(\hat{\alpha}_1, \sigma_1), \ldots, (\hat{\alpha}_n, \sigma_n).$$

G-modeling

Estimate the whole distribution G directly, then plug the estimate $\hat{G}$ into the posterior:

$$\hat{\alpha}_i^{NPEB} = \mathbb{E}_{\hat{G}}[\alpha_i \mid \hat{\alpha}_i, \sigma_i].$$

- **Parametric EB:** estimate a few hyperparameters.
- **Nonparametric EB:** estimate the full distribution of effects.

Outline

1. Motivation: Why Empirical Bayes in applied work?
2. Statistical Decision Theory Crash Course
3. Parametric Empirical Bayes
4. The James-Stein surprise (shrinkage can improve overall accuracy)
5. Nonparametric Empirical Bayes
6. Conclusion

Takeaways and next class

What we learned today

- EB is a way to improve estimation when we have many noisy unit-level effects.
- The Bayes rule under squared error loss is the posterior mean.
- The empirical Bayes estimator plugs estimated hyperparameters into that posterior mean and produces shrinkage.
- Shrinkage can improve *compound* performance by trading bias for a reduction in variance.
- It also has a frequentist payoff in many-parameter problems.

Next class: applications and beyond L_2

- empirical applications of EB in economics
- why the relevant loss may depend on the policy problem
- what changes when we care about ranking, selection, or other policy-relevant objectives instead of only L_2