

## Lecture 12: Empirical Bayes

*Instructor: Jonathan Pipping**Author: Ryan Brill*

## 12.1 Motivating Example: MLB Batting Averages

### 12.1.1 Single-Player Problem

Suppose player  $A$ 's batting average midway through the 2023 season is 0.300. Using no other information, we want to predict his end-of-season batting average. We set up a model to do this.

Let  $N$  and  $H$  represent the number of at-bats and hits for player  $A$  to this point in the season. We assume each at-bat  $\{X_i\}_{i=1}^N$  is an independent Bernoulli trial with success probability  $p$ . Then it follows that

$$H \sim \text{Binomial}(N, p)$$

and since  $H = \sum_{i=1}^N X_i$ , we model player  $A$ 's mid-season batting average as

$$BA = \frac{H}{N} \sim \frac{1}{N} \text{Binomial}(N, p)$$

As discussed previously, the MLE of a binomial random variable is the observed proportion of successes, which we write as

$$\hat{p}_{MLE} = \frac{H}{N}$$

We can use this for prediction, but we saw that this may not be the most stable estimator in small samples. In the previous lecture, we used a prior to stabilize the estimator, but without any other information, we may not know which prior to use. Let's consider a broader approach to the problem that will allow us to inform our choice of prior.

### 12.1.2 Multi-Player Problem

Suppose now we know each player's batting average midway through the 2023 season. Using no information from any previous season (i.e. using only the 2023 mid-season averages), predict each player's end-of-season batting average.

We can easily extend our single-player model to the multi-player case. We set up notation for the multi-player case as follows:

- $i$ : player index
- $n$ : total number of players
- $N_i$ : number of at-bats for player  $i$
- $H_i$ : number of hits for player  $i$

- $BA_i = \frac{H_i}{N_i}$ : batting average for player  $i$

We can then model each player's batting average as before:

$$BA_i = \frac{H_i}{N_i} \sim \frac{1}{N_i} \text{Binomial}(N_i, p_i)$$

If we use the MLE for each player's batting average, we arrive at a similar result as before:

$$\hat{p}_i^{(MLE)} = \frac{H_i}{N_i}$$

Our hope is to improve this guess with a prior, but how do we do that? We don't have any other information, so we can't use a prior from a previous season. What can/should we shrink to?

### 12.1.3 Idea: Shrinkage by Pooling

Since we have the midseason batting average of **each** baseball player, perhaps we can pool information across these players and shrink to the overall mean batting average. *The insight here is that each player  $i$  is a baseball player, and we can use that shared group information to improve our prediction.* This means we would predict player  $i$ 's end-of-season batting average as some mixture of their own mid-season batting average and the overall mean batting average. Let's modify our model to reflect this.

Let  $X_i = \frac{H_i}{N_i}$  be player  $i$ 's mid-season batting average. Then

$$X_i = \frac{H_i}{N_i} \sim \frac{1}{N_i} \text{Binomial}(N_i, p)$$

If we remove players  $i$  with a small number of at-bats (e.g.  $N_i < 30$ ), we can use the fact that  $N_i$  are large to apply the Central Limit Theorem from Lecture 8. That is,

$$X_i \overset{approx}{\sim} \mathcal{N}\left(p_i, \frac{p_i(1-p_i)}{N_i}\right) \text{ by the CLT}$$

Note that the proposed variance  $\sigma_i^2 = \frac{p_i(1-p_i)}{N_i}$  depends on the unknown parameter  $p_i$ , but it is much easier to work with **known** variance. We make the following *simplifying assumption* that

$$\sigma_i^2 = \frac{C}{N_i}$$

for some constant  $C$ . We'll use  $C = 0.035$  for our analysis, but we generally treat this as a hyperparameter to be tuned using data from the previous season. An alternative approach is to use a **variance-stabilizing transformation**  $h(X)$ , which transforms the batting average so that it has a known variance. We will discuss this at a later time.

In either case, we have that  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$  where  $\mu_i = p_i$  and  $\sigma_i^2$  is known. Now we can set up a parametric Bayesian model to predict each player's end-of-season batting average.

## 12.2 Parametric Bayesian Model

### 12.2.1 Model Setup

We implement the Bayesian approach, modeling the unknown parameter  $\mu_i = p_i$  as a random variable with its own distribution. We then set up our model as follows:

$$\begin{cases} X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i \sim \mathcal{N}(\mu, \tau^2) \end{cases}$$

Note that  $\mu$  represents the global mean of all baseball players' batting averages, and  $\tau^2$  represents the variance around this mean. Note also that  $\mu_i, p_i, \mu$ , and  $\tau$  are all unknown parameters, while  $\sigma_i = \sqrt{\frac{C}{N_i}}$  is assumed as known (from before).

### 12.2.2 MLE vs Posterior Mean

Recall from before that the MLE ignores all information from other players, instead only using the information relevant to player  $i$ . Specifically, we know that the MLE of  $\mu_i$  is the observed batting average  $X_i$ .

$$\hat{\mu}_i^{(MLE)} = X_i$$

The Bayesian analog of this quantity is the **posterior mean**, which we define below.

**Definition 12.1** (Posterior Mean). *The posterior mean of a random variable  $X$  is defined as*

$$\hat{\theta}_{Bayes} = \mathbb{E}[\theta|X] = \int \theta p(\theta|X) d\theta$$

In our case, the posterior mean of  $\mu_i$  from the data  $\{X_i\}_{i=1}^n$  is

$$\hat{\mu}_i^{(Bayes)} = \mathbb{E}[\mu_i|X_i]$$

To go further than this, we need the posterior distribution  $\mathbb{P}(\mu_i|X_i)$ . We can derive this using Bayes' rule:

$$\begin{aligned} \mathbb{P}(\mu_i|X_i) &= \frac{\mathbb{P}(X_i|\mu_i)\mathbb{P}(\mu_i)}{\mathbb{P}(X_i)} \text{ by Bayes' Rule} \\ &\propto \mathbb{P}(X_i|\mu_i)\mathbb{P}(\mu_i) \text{ since } \mathbb{P}(X_i) \text{ is constant w.r.t. } \mu_i \end{aligned}$$

From our model setup, we re-express the likelihood and prior as

$$\begin{aligned} \mathbb{P}(\mu_i|X_i) &= \mathbb{P}(\mathcal{N}(\mu_i, \sigma_i^2) = X_i) \mathbb{P}(\mathcal{N}(\mu, \tau^2) = \mu_i) \\ &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(X_i - \mu_i)^2}{2\sigma_i^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu_i - \mu)^2}{2\tau^2}\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{X_i^2}{\sigma_i^2} - 2\frac{X_i\mu_i}{\sigma_i^2} + \frac{\mu_i^2}{\sigma_i^2} + \frac{\mu_i^2}{\tau^2} - 2\frac{\mu_i\mu}{\tau^2} + \frac{\mu^2}{\tau^2}\right)\right] \end{aligned}$$

Since the first and last terms are constant w.r.t.  $\mu_i$ , we omit them and factor to get

$$\begin{aligned}\mathbb{P}(\mu_i|X_i) &\propto \exp \left\{ -\frac{1}{2} \left[ \mu_i^2 \left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right) - 2\mu_i \left( \frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2} \right) \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right) \left[ \mu_i^2 - 2\mu_i \frac{\left( \frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2} \right)}{\left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right) \left[ \mu_i - \frac{\left( \frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2} \right)}{\left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)} \right]^2 \right\}\end{aligned}$$

We recognize this as a normal distribution, and so we can write

$$\mathbb{P}(\mu_i|X_i) \sim \mathcal{N} \left( \frac{\left( \frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2} \right)}{\left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)}, \frac{1}{\left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)} \right)$$

Since we have a posterior distribution with a recognizable form, we can extract the posterior mean directly from the mean of the normal distribution. From Definition 12.1, we have that

$$\begin{aligned}\hat{\mu}_i^{(\text{Bayes})} &= \mathbb{E}[\mu_i|X_i] \\ &= \frac{\left( \frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2} \right)}{\left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)} \\ &= \mu + \frac{\tau^2}{\tau^2 + \sigma_i^2} (X_i - \mu)\end{aligned}\tag{12.1}$$

And since we know that  $X_i = \frac{H_i}{N_i}$  and  $\sigma_i^2 = \frac{C}{N_i}$ , we can write

$$\hat{\mu}_i^{(\text{Bayes})} = \frac{\frac{H_i}{C} + \frac{\mu}{\tau^2}}{\frac{N_i}{C} + \frac{1}{\tau^2}}$$

This looks a lot like the formulation we saw in the last lecture:

$$\frac{(W + W')}{(W + W') + (L + L')}$$

There's still a problem though: we don't know what  $\mu$  or  $\tau$  are! How do we estimate these parameters? Empirical Bayes offers a solution to this problem.

## 12.3 Empirical Bayes

Empirical Bayes is a way to estimate the parameters of a Bayesian model (in our case  $\mu$  and  $\tau$ ) using the data itself. Specifically, this involves plugging in the MLEs of  $\mu$  and  $\tau$  into Equation 12.1, which we do to

define the **Parametric Empirical Bayes Estimator** for our model.

$$\begin{aligned}\hat{\mu}_i^{(EB)} &= \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2} (X_i - \hat{\mu}) \\ &= \frac{\frac{H_i}{C} + \frac{\hat{\mu}}{\hat{\tau}^2}}{\frac{N_i}{C} + \frac{1}{\hat{\tau}^2}}\end{aligned}\tag{12.2}$$

where  $\hat{\mu}$  and  $\hat{\tau}^2$  are the MLEs of  $\mu$  and  $\tau^2$ . Before we derive these estimate, let's consider the implications of different values of  $\tau^2$  on the posterior mean.

### 12.3.1 Implications of Different $\tau^2$ Values on Posterior Mean

If  $\tau^2 = 0$ , then  $\mu_i \sim \mathcal{N}(\mu, \tau^2) \stackrel{d}{=} \mathcal{N}(\mu, 0) = \mu$ .

If  $\tau^2 = \infty$ , then  $\mu_i \sim \mathcal{N}(\mu, \tau^2) \stackrel{d}{=} \text{Uniform}(-\infty, \infty)$  and  $\hat{\mu}_i = X_i = \frac{H_i}{N_i} = \hat{\mu}_i^{(MLE)}$ .

Otherwise, we have that  $\hat{\mu}_i = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2} (X_i - \hat{\mu})$ .

- Is closer to  $\hat{\mu}$  if  $\frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}$  is small (meaning  $\sigma_i^2$  is large, or  $N_i$  is small).
- Is closer to  $X_i$  if  $\frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}$  is large (meaning  $\sigma_i^2$  is small, or  $N_i$  is large).

Now we will proceed by deriving  $\hat{\mu}_{MLE}$  and  $\hat{\tau}_{MLE}^2$ .

### 12.3.2 Deriving $\hat{\mu}_{MLE}$ and $\hat{\tau}_{MLE}^2$

Recall that our model is

$$\begin{cases} X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i \sim \mathcal{N}(\mu, \tau^2) \end{cases}$$

Since we want to find the MLE of  $\mu$  and  $\tau^2$ , we need to maximize the likelihood function  $\mathcal{L}(\mu, \tau^2 \mid X_i)$ . Like we did in Lecture 10, we will instead maximize the log-likelihood function  $\ell(\mu, \tau^2 \mid X_i)$ .

$$\ell(\mu, \tau^2 \mid X_i) = \log \mathbb{P}(X_1, \dots, X_n \mid \mu, \tau^2)$$

Since the  $X_i$  are independent, we have that

$$\begin{aligned}\ell(\mu, \tau^2 \mid X_i) &= \log \prod_{i=1}^n \mathbb{P}(X_i \mid \mu, \tau^2) \\ &= \sum_{i=1}^n \log \mathbb{P}(X_i \mid \mu, \tau^2)\end{aligned}$$

From Bayes' Rules, we have that  $\mathbb{P}(X_i | \mu, \tau^2) \sim \mathcal{N}(\mu, \tau^2 + \sigma_i^2)$ . So we have that

$$\begin{aligned}\ell(\mu, \tau^2 | X_i) &= \sum_{i=1}^n \log \mathbb{P}(\mathcal{N}(\mu, \tau^2 + \sigma_i^2) = X_i) \\ &= \sum_{i=1}^n \log \left[ \frac{1}{\sqrt{2\pi(\tau^2 + \sigma_i^2)}} \exp \left( -\frac{(X_i - \mu)^2}{2(\tau^2 + \sigma_i^2)} \right) \right] \\ &= \sum_{i=1}^n \left\{ \log \left( \frac{1}{\sqrt{2\pi(\tau^2 + \sigma_i^2)}} \right) - \frac{(X_i - \mu)^2}{2(\tau^2 + \sigma_i^2)} \right\} \\ &\propto -\frac{1}{2} \sum_{i=1}^n \log(\tau^2 + \sigma_i^2) - \frac{1}{2} \sum_{i=1}^n \frac{(X_i - \mu)^2}{\tau^2 + \sigma_i^2}\end{aligned}$$

Then we have that

$$MLE(\mu, \tau^2) = \arg \max_{\mu, \tau^2} \ell(\mu, \tau^2 | X_i)$$

which we can solve by taking the derivative with respect to  $\mu$  and  $\tau^2$  and setting them equal to 0. First, for  $\mu$ :

$$\begin{aligned}\frac{\partial}{\partial \mu} [\ell(\mu, \tau^2 | X_i)] &= \frac{1}{2} \sum_{i=1}^n \frac{2(X_i - \mu)}{\tau^2 + \sigma_i^2} = 0 \\ \implies \hat{\mu}_{MLE} &= \frac{\sum_{i=1}^n \frac{X_i}{\tau^2 + \sigma_i^2}}{\sum_{i=1}^n \frac{1}{\tau^2 + \sigma_i^2}}\end{aligned}$$

Now we take the derivative with respect to  $\tau^2$ :

$$\begin{aligned}\frac{\partial}{\partial \tau^2} [\ell(\mu, \tau^2 | X_i)] &= -\frac{1}{2} \sum_{i=1}^n \frac{1}{\tau^2 + \sigma_i^2} + \frac{1}{2} \sum_{i=1}^n \frac{(X_i - \mu)^2}{(\tau^2 + \sigma_i^2)^2} = 0 \\ \implies \sum_{i=1}^n \frac{(X_i - \mu)^2}{(\hat{\tau}_{MLE}^2 + \sigma_i^2)^2} &= \sum_{i=1}^n \frac{1}{\hat{\tau}_{MLE}^2 + \sigma_i^2}\end{aligned}$$

Now, we have a slight problem: our expression for  $\hat{\mu}_{MLE}$  depends on  $\tau^2$ , and our expression for  $\hat{\tau}_{MLE}^2$  depends on  $\hat{\mu}$ . We can solve this by iterating.

1. Make initial guesses  $\mu_{(0)}$  and  $\tau_{(0)}^2$ .
2. While  $|\mu_{(t)} - \mu_{(t-1)}| \leq \delta$  and  $|\tau_{(t)}^2 - \tau_{(t-1)}^2| \leq \delta$ , do the following in R:

$$\text{Set } \hat{\mu}_{(t)} = \frac{\sum_{i=1}^n \frac{X_i}{\tau_{(t-1)}^2 + \sigma_i^2}}{\sum_{i=1}^n \frac{1}{\tau_{(t-1)}^2 + \sigma_i^2}}$$

$$\text{Set } \hat{\tau}_{(t)}^2 \text{ to the solution of } \sum_{i=1}^n \frac{(X_i - \hat{\mu}_{(t)})^2}{(\hat{\tau}_{(t-1)}^2 + \sigma_i^2)^2} = \sum_{i=1}^n \frac{1}{\hat{\tau}_{(t-1)}^2 + \sigma_i^2} \text{ using the } \texttt{uniroot} \text{ function}$$

3. Iterate until convergence to  $\hat{\mu}_{MLE}$  and  $\hat{\tau}_{MLE}^2$ .

### 12.3.3 Results

We plot the predicted batting averages for the 2023 season ( $\hat{\mu}_i$ ) in Figure 12.1, with the MLE on the x-axis and the Empirical Bayes estimate on the y-axis. Note that the size of the points is proportional to the number of at-bats ( $N_i$ ) for each player.

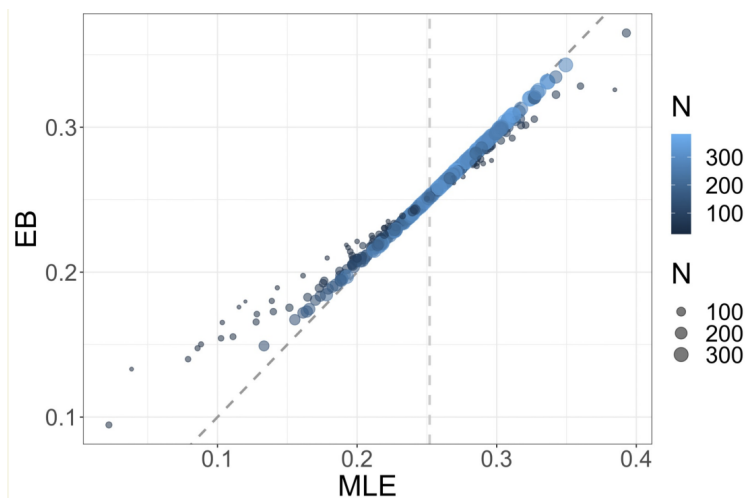


Figure 12.1: Comparison of MLE and EB estimates of 2023 end-of-season batting averages. The diagonal represents the identity line  $EB = MLE$ .

As we can see, players with smaller  $N_i$  have  $\hat{\mu}_i^{(EB)}$  shrunk towards the overall mean, while players with larger  $N_i$  have  $\hat{\mu}_i^{(EB)} \approx \hat{\mu}_i^{(MLE)}$ , their mid-season batting averages. When we plot both estimates against the end-of-season batting averages in Figure 12.2, we don't see much of a difference between the two. However, considering their respective RMSE's in Figure 12.3, we see that the Empirical Bayes estimate achieves better prediction than the actual mid-season batting averages (MLE).

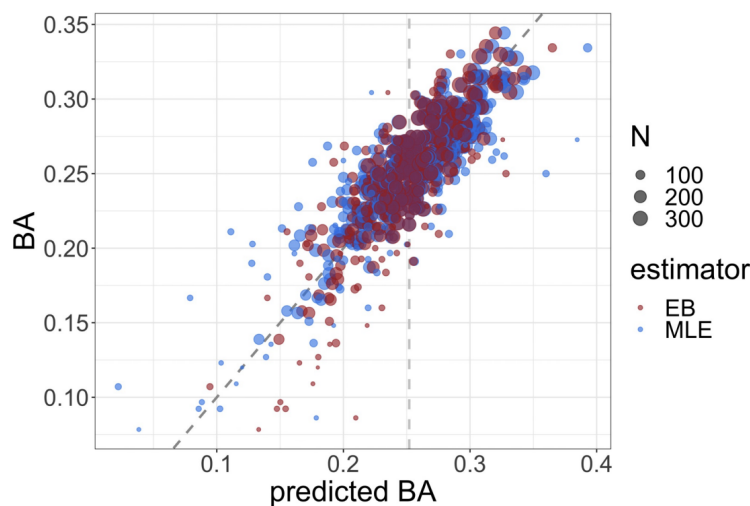


Figure 12.2: Comparison of MLE and EB estimates of end-of-season batting averages against the actual 2023 end-of-season batting averages. The diagonal represents the identity line  $BA = \bar{BA}$ .

rmse_MLE	rmse_EB
0.02629828	0.02383808

Figure 12.3: RMSE of the MLE and EB estimates of 2023 end-of-season batting averages.

### 12.3.4 Takeaways

- Shrinkage towards the overall mean helps prediction, especially when we have small sample sizes.
- Sharing information across players helps, and Empirical Bayes allows us to estimate these global parameters directly from the data.

### 12.3.5 If We Had Access to Previous Seasons

What if we had access to previous seasons' data? Consider our Parametric Empirical Bayes Estimator from Equation 12.2:

$$\hat{\mu}_i^{(EB)} = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2} (X_i - \hat{\mu})$$

where  $\hat{\mu}$  and  $\hat{\tau}^2$  are the MLEs of  $\mu$  and  $\tau^2$ . This is one example of a class of shrinkage estimators, which we will define generally now.

**Definition 12.2** (General Shrinkage Estimator). *A general shrinkage estimator  $\hat{\theta}_i$  takes the form*

$$\hat{\theta}_i = \hat{\theta} + \beta(X_i - \hat{\theta})$$

where  $\hat{\theta}$  is the baseline (often the MLE or mean),  $\beta \in [0, 1]$  is the shrinkage parameter, and  $X_i$  is the observed data.

The procedure to incorporate data from previous seasons is as follows:

1. Let  $\{X_i^{(2022)}\}_{i=1}^n$  be the observed mid-season batting averages from 2022, let  $\hat{\mu}^{(2022)}$  be the overall mean batting average from that season, and let  $\hat{\mu}_i^{(2022)}$  be the known end-of-season batting averages from that year.
2. Estimate  $\beta$  using the following regression:

$$\begin{aligned} \hat{\mu}_i^{(2022)} &= \hat{\mu}^{(2022)} + \beta \left( X_i^{(2022)} - \hat{\mu}^{(2022)} \right) + \epsilon_i \\ \implies \hat{\mu}_i^{(2022)} - \hat{\mu}^{(2022)} &= \beta \left( X_i^{(2022)} - \hat{\mu}^{(2022)} \right) + \epsilon_i \\ \implies Y_i &= \beta X_i + \epsilon_i \end{aligned}$$

where  $Y_i = \hat{\mu}_i^{(2022)} - \hat{\mu}^{(2022)}$  and  $X_i = X_i^{(2022)} - \hat{\mu}^{(2022)}$ . Call this estimate  $\hat{\beta}$ .

3. Predict the end-of-season batting averages for the current season as follows:

$$\hat{\mu}_i^{(2023)} = \hat{\mu}^{(2023)} + \hat{\beta} \left( X_i^{(2023)} - \hat{\mu}^{(2023)} \right)$$