WSABI Summer Research Lab **Summer 2025**

Lab 12: Empirical Bayes

Instructor: Jonathan Pipping *Authors: RB, JP*

12.1 Rolling NBA Player Quality Estimates

Empirical Bayes is one of the bets ways to estimate player quality, especially as rolling estimates based on data from observed games. Today, we'll focus on estimating NBA players' "true" scoring talent or skill.

12.1.1 Data

We observe box score data from NBA games, with each row corresponding to a player-game combination (i, j). We have the following general information:

- n: the number of players in the dataset.
- G_i: the number of games player i plays.

And each row contains the following variables:

- i: the player index. $i = 1, \ldots, n$
- j: the game index for each player. $j = 1, \ldots, G_i$
- X_{ij}: Points scored by player i in their j^{th} game
- P_{ij}: The number of possessions player i was on the court in their j^{th} game

If we wnat to model the points scored by a player i in game j (X_{ij}), we can't simply assume an independent draw from player i's "true" scoring quality μ_i. This is because his quality may change over time, whether from aging, injury, or other factors that may cause him to improve or decline. How can we capture player i's quality μ_{ij} as it progresses across his career games $j \in 1, \ldots, G_i$?

12.1.2 Dynamic Bayesian Model

We do this by forming a dynamic Bayesian model, similar to the one published by Mark Glickman in [MEG]. The model is as follows:

$$X_{ij} \mid \mu_{ij} \sim \mathcal{N}(P_{ij}\mu_{ij}, P_{ij}\sigma^2) \tag{12.1}$$

$$\mu_{ij} \mid \mu_{i(j-1)} \sim \mathcal{N}(\mu_{i(j-1)}, \tau^2),\ j > 1 \tag{12.2}$$

$$\mu_{i1} \mid \mu \sim \mathcal{N}(\mu, \nu^2) \tag{12.3}$$

What do these parameters mean?

- μ_{ij}: Player i's *unobserved* "true" scoring quality in their j^{th} game.
- μ: The overal mean scoring quality of all players.
- σ^2: The variance in points scored on a possession given a player's true scoring quality.
- τ^2: The game-to-game variance in a player's true scoring quality.
- ν^2: The prior variance, or the variance in a player's true scoring quality before seeing any data.

We will use Empirical Bayes to estimate μ_{ij}, player i's latent scoring quality in their j^{th} game. The Bayesian estimate of μ_{ij} is the posterior mean, which is the mean of the posterior distribution of μ_{ij}, which depends on the observed data X_{ik} and P_{ik} for $k \leq j$.

$$\widehat{\mu}_{ij} = \mathbb{E}[\mu_{ij} \mid \{X_{ik}\}_{k\leq j}, \{P_{ik}\}_{k\leq j}, \mu, \sigma^2, \tau^2, \nu^2]$$

Computing this expectation is not straightforward, but we'll derive it step by step.

12.1.3 Deriving Posterior Means

To begin, we will derive the posterior mean of μ_{i1}. This is the most straightforward case, as it only depends on a single game's data (12.1) and the prior distribution (12.3). We express the posterior distribution of μ_{i1} as:

$$\mu_{i1} \mid X_{i1}, P_{i1}, \mu, \sigma^2, \nu^2 \propto \mathbb{P}(X_{i1} \mid \mu_{i1})\mathbb{P}(\mu_{i1} \mid \mu)$$

Task 1: Write a formula for the posterior mean of this normal-normal model, denoted

$$\widehat{\mu}_{i1} = \mathbb{E}[\mu_{i1} \mid X_{i1}, P_{i1}, \mu, \sigma^2, \nu^2]$$

Note that we derived the posterior distribution of a normal-normal model in Lecture 12. Feel free to use that result, but note that the parameter names are slightly different.

Now, we will derive the posterior mean of μ_{ij} for $j > 1$. This depends on the previous game's data (12.1) and the transition distribution defining how quality changes from game to game (12.2). We express the posterior distribution of μ_{ij} as:

$$\mu_{ij} \mid \{X_{ik}\}_{k\leq j}, \{P_{ik}\}_{k\leq j}, \mu, \sigma^2, \tau^2, \nu^2 \propto \mathbb{P}(X_{ij} \mid \mu_{ij})\mathbb{P}(\mu_{ij} \mid \mu_{i(j-1)})$$

Task 2: Write a formula for the posterior mean of this normal-normal model, denoted

$$\widehat{\mu}_{ij} = \mathbb{E}[\mu_{ij} \mid \{X_{ik}\}_{k\leq j}, \{P_{ik}\}_{k\leq j}, \mu, \sigma^2, \tau^2, \nu^2]$$

The formulas obtained for the posterior means $\{\widehat{\mu}_{ij}\}_{j=1}^{G_i}$ are functions of observed data $\{X_{ik}\}_{k\leq j}, \{P_{ik}\}_{k\leq j}$ and unobserved hyperparameters $\mu, \sigma^2, \tau^2, \nu^2$. We will use **Empirical Bayes** to estimate these hyperparameters using the data, then plug in these estimates to obtain the posterior means.

12.1.4 Empirical Bayes

In Lecture 12, we estimated the hyperparameters of our model by maximizing the marginal likelihood of the data. However, in this case the full likelihood is difficult to even write down, much less maximize! Instead, we'll pick some reasonable values and see how well they work.

Task 3: Estimate μ and ν^2

Player i's average game-level points per possession across all his games can be written as

$$M_i = \frac{1}{G_i}\sum_{j=1}^{G_i}\frac{X_{ij}}{P_{ij}}$$

Define the set

$$\mathcal{M} = \{M_i\}_{i:G_i>G^*}$$

as the set of averages for players with at least G^* games played (where G^* ensures that you've observed enough data to estimate μ and ν^2 accurately). Then, estimate μ and ν^2 as the mean and variance of the set $\mathcal{M}$.

$$\widehat{\mu} = \text{Mean}(\mathcal{M})$$
$$\widehat{\nu}^2 = \text{Var}(\mathcal{M})$$

Task 4: Estimate σ^2

Let the variance of player i's game-level points per possession be written as

$$V_i = \text{Var}\left\{\frac{X_{ij}}{P_{ij}}\right\}_{j=1}^{G_i}$$

And we define the set $\mathcal{V}$ as a subset of $\{V_i\}_{i=1}^n$ consisting of just average players, or players whose M_i is very close to the overall mean $\widehat{\mu}$. Then define $\widehat{\sigma}^2$ as the mean of $\mathcal{V}$.

$$\widehat{\sigma}^2 = \text{Mean}(\mathcal{V})$$

This represents an approximation of the variance of the points per possession for average players.

Task 5: Estimate τ^2

We will treat τ^2 as a **tuning parameter**. For each player i with $G_i > G^*$, split their career into two. Let

$$\mathcal{D}_{1i} = \{X_{ij}\}_{j=1}^{G_i/2}$$
$$\mathcal{D}_{2i} = \{X_{ij}\}_{j=G_i/2+1}^{G_i}$$

Then set a sequence of small values for τ^2 (ex: from 10^{-7} to 10^{-3}). For each value of τ^2 we are considering, estimate μ_{ik}, where $k = G_i/2$ represents the last game in the first half of their career. Then the accuracy of this predictor for each player i on the out-of sample data $\mathcal{D}_{2i}$ is given by the root mean squared error (RMSE) of the predictions, defined as

$$\text{RMSE}_i(\tau^2) = \sqrt{\frac{1}{|\mathcal{D}_{2i}|}\sum_{j\in\mathcal{D}_{2i}}(\mu_{ik}-\mu_{ij})^2}$$

And the mean error across all players is given by

$$\frac{1}{n}\sum_{i=1}^{n}\text{RMSE}_i(\tau^2)$$

Search over the grid of values for τ^2 and find the value that minimizes the mean error.

12.1.5 Implementing the Model

Finally, we have hyperparameters $\widehat{\mu}, \widehat{\nu}^2, \widehat{\sigma}^2, \widehat{\tau}^2$, as well as data $\{X_{ij}, P_{ij}\}_{j=1}^{G_i}$ for each player i, and are ready to compute the posterior means $\widehat{\mu}_{ij}$ for each player i and game j.

Task 6: Compute the posterior means $\widehat{\mu}_{ij}$ for each player i and game j. Plot the trajectory of the posterior means for some sample of your players as a function of j, the game number in their career. What do you notice about players' scoring quality over time?

12.2 Rolling NFL Kicker Quality Estimates

The Empirical Bayes model in the previous question can be written down fairly easily, but it is still a lot of work to solve for hyperparameters and the posterior mean. Sometimes, a formal model like this is overkill, and you can estimate a player quality trajectory of similar quantity in a much simpler way. To illustrate this, we will estimate kicker quality trajectories for NFL kickers by using a **weighted sum** of their field goal probabilities added over all previous kicks in their career.

12.2.1 Simple Model

Task 1: Fit a simple field goal probability model $\mathbb{P}_{FG}^{(0)}$ as a function of just yard line ydl. Like in Lab 4, you can use logistic regression and a spline. For now, ignore the intrinsic selection bias.

12.2.2 FGPA & Kicker Quality

We define the field goal probability added of the j^{th} field goal as

$$\text{FGPA}_j = \mathbb{1}\{FG_j\} - \mathbb{P}_{FG}^{(0)}(ydl_j)$$

where $\mathbb{1}\{FG_j\}$ is an indicator function that is 1 if the j^{th} field goal was made and 0 if not, and ydl_j is the yard line of the j^{th} field goal.

Furthermore, define a kicker i's quality prior to field goal j as

$$\text{KQ}_{ij} = \alpha \cdot \text{KQ}_{i(j-1)} + \text{FGPA}_{i(j-1)}, \ \alpha \in (0,1) \tag{12.4}$$

where $\text{KQ}_{i0} = 0$ and $\text{FGPA}_{i0} = 0$ prior to kicker i's first field goal.

Task 2: Find a formula for KQ_{ij} that depends only on α and $\{\text{FGPA}_k\}_{k<j}$. You'll notice that α plays a similar role as τ^2 in the Empirical Bayes model from the previous section.

12.2.3 The Role of α

α is a tunable hyperparameter that acts as an **exponential decay weight**: upweighting more recent field goals and downweighting older ones. This accounts for non-stationarity in a kicker's quality! For instance, $\alpha = 0.995$ weighs the the previous field goal twice as much as the 138^{th} most-recent field goal, since $\alpha^{138} \approx 0.5$.

Our estimator of kicker quality is equivalent to that of an Empirical Bayes model, just without writing down the full model. What's the prior this model implies?

12.2.4 Implementing the Model

We are now ready to implement a final model for kicker quality.

Task 3: Write a function that, given α, estimates each kicker i's quality prior to each kick j. Use a for loop with the formula you defined in Equation 12.4 for computational efficiency. Plot the trajectory of some sample kicker's quality as a function of j, the kick number in their career. What do you notice about the quality of kickers over time?

References

[MEG] Glickman, M. E., *Parameter Estimation in Large Dynamic Paired Comparison Experiments*, Journal of the Royal Statistical Society. Series C (Applied Statistics), 1999.