Lab: Empirical Bayes

1. Rolling Player Quality Estimates for NBA Players

Empirical Bayes is one of the best ways to estimate player quality, especially pre-game Rolling player quality which just uses data from previous games.

Today we'll consider an example: estimate an NBA player's "true" scoring talent or skill.

We observe box score data of the form

$$\begin{cases} X_{ij} = \text{pts scored by player } i \text{ in game } j \\ N_{ij} = \text{num possessions of player } i \text{ in game } j \\ N_i = \text{num games of player } i, \quad j = 1, \dots, N_i \\ N = \text{num players}, \quad i = 1, \dots, N \\ \text{data } \mathcal{D} = \{(X_{ij}, N_{ij}) : 1 \le i \le N, 1 \le j \le N_i\}. \end{cases}$$

It's not necessarily enough to model the points scored X_{ij} as a draw from player i's "true" scoring quality μ_i because his quality may change over time, for instance due to injury or aging or simply getting better/worse by practice or whatever reason (non-stationarity).

It would be great to capture player i's quality μ_{ij} in game j as j progresses across his career.

To do so, we form a dynamic Bayesian model, which in some form was seen in Mark Glickman's paper from the 1990s,

$$\begin{cases} X_{ij} \mid \mu_{ij} \sim \mathcal{N}(N_{ij}\mu_{ij}, N_{ij}\sigma^2) \\ \mu_{ij} \mid \mu_{i(j-1)} \sim \mathcal{N}(\mu_{i(j-1)}, \tau^2) & \text{if } j > 1 \\ \mu_{i1} \mid \mu \sim \mathcal{N}(\mu, \nu^2) \end{cases}$$

Think about what the parameters represent **before** looking at the answer (below).

μ_{ij} = latent (unobserved) "true" scoring talent of player i in game j

μ = overall mean player talent

σ^2 = variance in points scored on a possession given a player's true scoring talent

τ^2 = game-to-game variance in a player's scoring quality

ν^2 = variance in player talent before seeing any data (prior variance)

Use empirical Bayes to estimate μ_{ij}, player i's latent scoring talent during game j.

The Bayesian estimate of μ_{ij} is the posterior mean,

$$\hat{\mu}_{ij} = \mathbb{E}\left[\mu_{ij} \middle| \{X_{ik}\}_{k<j}, \{N_{ik}\}_{j<k}, \mu, \sigma^2, \tau^2, \nu^2\right].$$

Computing this expectation is hard.

To begin, estimate μ_{i1}.

The relevant part of the model is

$$\begin{cases} X_{i1} / \mu_{i1} \sim N(N_{i1}\mu_{i1}, N_{i1}\sigma^2) \\ \mu_{i1} / \mu \sim N(\mu, \nu^2). \end{cases}$$

Write a formula for the posterior mean of this normal-normal model $\hat{\mu}_{i1} = \mathbb{E}(\mu_{i1} | X_{i1}, N_{i1}, \mu, \sigma^2, \tau^2, \nu^2)$; we derived the posterior mean of a normal-normal model in class today, you can use that formula but the parameter names may be different.

Next, estimate μ_{ij} given $\mu_{i(j-1)}$ for $j>1$.

The relevant part of the model is

$$\begin{cases} X_{ij} \mid \mu_{ij} \sim \mathcal{N}(N_{ij}\mu_{ij},\ N_{ij}\sigma^2) \\ \mu_{ij} \mid \mu_{i(j-1)} \sim \mathcal{N}(\mu_{i(j-1)},\ \tau^2) \end{cases}$$

Again, write a formula for the posterior mean

$$\hat{\mu}_{ij} = \mathbb{E}[\mu_{ij} \mid \mu_{i(j-1)}, \tau^2, \sigma^2, X_{ij}, N_{ij}].$$

The formulas you obtained for $\{\mu_{ij}\}_{j=1}^{N_i}$ are functions of observed data $\{X_{ij}\}$ and $\{N_{ij}\}$ and unobserved hyperparameters $\mu, \sigma^2, \tau^2, v^2$.

The Empirical Bayes approach is to estimate the parameters and then plug in the estimates $\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2, \hat{v}^2$ into the formula for $\hat{\mu}_{ij}$.

So, we must estimate these hyperparameters from the data. It is too difficult to compute the MLE of these params because it is too difficult to even write down the full likelihood of the model.

Instead, lets just pick some somewhat reasonable values and see what happens.

* Estimate μ and v^2:

Player i's game-level PTS per possession averaged across all his games is

$$M_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{X_{ij}}{N_{ij}}.$$

Let the set $\mathcal{M}$ be the subset of $\{M_i\}_{i=1}^{N}$ who have high enough games (large enough N_i) and let $\hat{\mu} = \text{Mean}(\mathcal{M})$ and $\hat{v}^2 = \text{VAR}(\mathcal{M})$.

We do this so that $\hat{\mu}$ and $\hat{v}^2$ are more reflective of a PRIOR distribution on player talent.

* Estimate σ^2:

let the variance of player i's game level pts per possession be $V_i = VAR\left\{\frac{X_{ij}}{N_{ij}}\right\}_{j=1}^{N_i}$

Now form the set $\mathcal{V}$, a subset of $\{V_i\}_{i=1}^{N}$, consisting of just average players (i.e. players i whose M_i values are very close to the overall mean).

Let $\hat{\sigma}^2 = mean(\mathcal{V})$.

* Estimate τ^2:

We will treat τ^2 as a tuning parameter.

Consider a fixed value of τ^2.

Consider player i (assuming N_i is large enough).

Split his career in half, forming datasets

$$\mathcal{D}_{1i} = \{ (X_{ij}, N_{ij}) : j \in \text{games in the 1st half of his career} \}$$

$$\mathcal{D}_{2i} = \{ (X_{ij}, N_{ij}) : j \in \text{games in the 2nd half of his career} \}$$

From $\mathcal{D}_{1i}$, $\hat{\mu}$, $\hat{v}^2$, $\hat{\delta}^2$, and the τ^2 we are considering, estimate $\hat{\mu}_{iK}$ in the final game K of $\mathcal{D}_{1i}$.

Then the accuracy of this predictor on the out-of-sample data $\mathcal{D}_{2i}$ is

$$RMSE_i(\tau^2) = \sqrt{\sum_{j \in \mathcal{D}_{2i}} \left(\frac{X_{ij}}{N_{ij}} - \hat{\mu}_{iK} \right)^2}$$

and the mean accuracy is $\frac{1}{N} \sum_{i=1}^{N} RMSE_i(\tau^2)$.

Now, searching over values of τ^2 that are small (say 10^{-7} to 10^{-3}), choose the value of τ^2 that has lowest RMSE.

* Finally, given $\hat{\mu}, \hat{v}^2, \hat{\sigma}^2, \hat{\tau}^2$, and $\{(X_{ij}, N_{ij}) : \substack{1 \le i \le N \\ 1 \le j \le N_i}\}$, compute $\hat{\mu}_{ij}$ for each i, j.

Make a spaghetti plot:

Plot estimated scoring talent $\hat{\mu}_{ij}$ (y axis) vs. time (x axis) for some players of your choosing (color).

Play around with different values of v^2, τ^2.

2. A simpler version of dynamic Empirical Bayes illustrated via Kicker Quality

The Empirical Bayes model in the previous question isn't even that complex to write down, but it is a pain to solve for the posterior mean and especially the hyperparameters.

Sometimes, such a formal model is overkill and we can estimate a player quality trajectory of similar quality but in a much easier way. We'll illustrate this by estimating kicker quality trajectories for NFL kickers.

We'll define kicker quality by a weighted sum of his field goal probability added over all his previous kicks in his career.

To begin, fit a field goal probability model $P^{(0)}_{FG}$ as a function of just yard line.

Use logistic regression and a spline.
Ignore the selection bias.
Define the field goal probability added of the j^{th} field goal by

$$FGPA_j = \mathbb{1}\{\text{n^{th} field goal made}\} - \hat{p}^{(0)}_{FG}(\text{yardline of n^{th} field goal}).$$

Note $\mathbb{1}\{x\} = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{else} \end{cases}$ is the indicator function.

Now we define Kicker quality using Raven's Kicker Justin Tucker for concreteness. Index all his field goals in order by j. We define his quality prior to field goal j by

$$(*) \quad KQ_j := \alpha \cdot KQ_{j-1} + FGPA_{j-1},$$

$KQ_0 = 0$ and $FGPA_0 = 0$.

Find a formula for KQ_j purely in terms of α and $\{FGPA_\ell\}_{\ell<j}$. You'll notice that α plays a similar role as τ^2 in the Empirical Bayes model of the previous question.

$\alpha < 1$ is an exponential decay weight that upweighs more recent field goals, thus accounting for non-stationarity. For instance, $\alpha = 0.995$ weighs the 138^{th} field goal in the past half as much as the previous field goal, $\alpha^{138} = 0.50$.

Our estimator of kicker quality is equivalent to that of an Empirical Bayes model! We just never wrote out the model. What's the prior?

Write a function that, given α, estimates each player i's kicker quality prior to each kick j, $\{KQ_{ij} : 1 \le i \le N,\ 1 \le j \le N_i\}$. Use a for loop with formula (*), not the closed form formula, to make it fast.

Finally, make a spaghetti plot for $\alpha = 0.995$: plot kicker quality KQ_{ij} (y axis) vs. time (x axis) for some kickers of your choosing (color).