Empirical Bayes/ Shrinkage

Chris Conlon

April 23, 2021

NYU Stern

Empirical Bayes

A (famous) Baseball Example

Suppose we want to estimate batting averages (AVG) for some baseball players

- $AVG = \frac{\#\text{hits}}{\#AtBats}$
- Use data on the first $n = 45$ at bats and hits x_i for the 1970 season.
- Predict the batting average μ_i for the end of the season ($n = 400 - 500$ at bats).
- Obvious estimate is batting average after 45 at bats: $\widehat{\mu}_i^{MLE} = x_i/45$.
- Is there a better estimate?

A Baseball Example

Table 1.1: Batting averages $z_i = \hat{\mu}_i^{(\text{MLE})}$ for 18 major league players early in the 1970 season; μ_i values are averages over the remainder of the season. The James–Stein estimates $\hat{\mu}_i^{(\text{JS})}$ (1.35) based on the z_i values provide much more accurate overall predictions for the μ_i values. (By coincidence, $\hat{\mu}_i$ and μ_i both average 0.265; the average of $\hat{\mu}_i^{(\text{JS})}$ must equal that of $\hat{\mu}_i^{(\text{MLE})}$.)

Name	hits/AB	$\hat{\mu}_i^{(\text{MLE})}$	μ_i	$\hat{\mu}_i^{(\text{JS})}$
Clemente	18/45	.400	**.346**	.294
F Robinson	17/45	.378	**.298**	.289
F Howard	16/45	.356	**.276**	.285
Johnstone	15/45	.333	**.222**	.280
Berry	14/45	.311	**.273**	.275
Spencer	14/45	.311	**.270**	.275
Kessinger	13/45	.289	**.263**	.270
L Alvarado	12/45	.267	**.210**	.266
Santo	11/45	.244	**.269**	.261
Swoboda	11/45	.244	**.230**	.261
Unser	10/45	.222	**.264**	.256
Williams	10/45	.222	**.256**	.256
Scott	10/45	.222	**.303**	.256
Petrocelli	10/45	.222	**.264**	.256
E Rodriguez	10/45	.222	**.226**	.256
Campaneris	9/45	.200	**.286**	.252
Munson	8/45	.178	**.316**	.247
Alvis	7/45	.156	**.200**	.242
Grand Average		.265	**.265**	.265

A (famous) Baseball Example

Probably we can do better than the MLE here:

- Thurman Munson wins Rookie of the Year and ends up batting $\mu_i = .316$. If he batted .178 all year, his career would not have lasted long.
- Clemente's $.400$ seems unlikely to hold up. Last player to hit $> .400$ was Ted Williams $.406$ in 1941.
- But how?

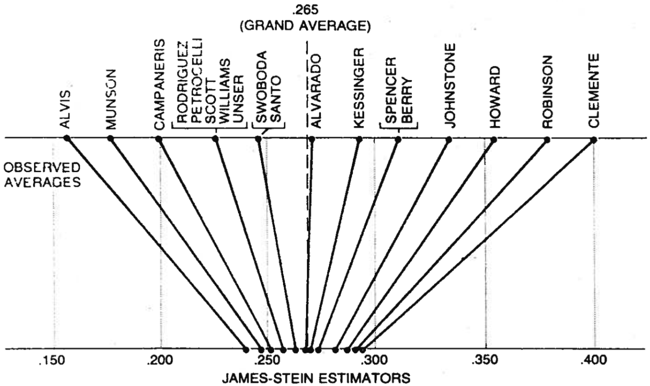
Bayesian Shrinkage

Idea is to take an average between the observed average y_i and the overall mean $\overline{y}$:

$$\widehat{\mu}_i^{JS} = (1-\lambda)\cdot\overline{y} + \lambda\cdot y_i, \quad \lambda = 1 - \frac{(k-3)\sigma^2}{\sum_i (y_i - \overline{y})^2}$$

- This has the effect of shrinking y_i towards the prior mean $\overline{y}$.
- In this case the prior mean is just $\overline{y}$ the grand-mean of all players
- How can information about unrelated players inform us about μ_i?
- Also consider proportion of foreign cars in Chicago as an additional y_i, can this help too?
- The shrinkage factor λ depends on sample size and variance, but how is it chosen?

A Baseball Example



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Aside: James-Stein Estimator

This is a famous (and confusing) result from statistics:

- For normally distributed $Y \sim N(\theta, \sigma^2 I)$ with unknown means $[\theta_1, \theta_2, \ldots, \theta_k]$
- Why would using information from Y_2 tell us anything about Y_1?
- And yet the James-Stein or (pooled) shrinkage estimator is biased, but has lower MSE than the naive estimator.
- Why does Clemente's batting average tell us anything about Munson's?

See `https://statweb.stanford.edu/~ckirby/brad/LSI/chapter1.pdf` for formal results.

What is Emprical Bayes?

- Priors can be an important modeling choice
- But what makes a good prior?
  - Sufficiently diffuse
  - As non-informative as possible
  - Don't tip the scales
  - Don't rule out the truth
- Idea: can we use the data itself to construct a prior?
  - If everything is a function of data, are we back in frequentist paradigm?
  - Can we get benefits of Bayes estimation without unpalatable assumptions?

My Own Example: Conlon and Mortimer

- We remove Snickers Δq_{jt} and measure change in sales of substitutes Δq_{kt}.
    - We use nearest neighbor matching for each machine-week t.
- We are interested in the average diversion ratio $D_{jk} = \frac{\sum_t \Delta q_{kt}}{\sum_t \Delta q_{jt}}$
- Several consumers switch to no-purchase option D_{j0}.
- Problems:
    - Some products are rarely available (small Δq_{jt}) and we measure huge D_{jk} for them.
    - Some products have sales decline $\Delta q_{kt} < 0$ even though they are (weak) substitutes.
    - Mostly this is just that q_{kt} and q_{jt} are very noisy.
    - We ran the experiment for almost a month – we can't run it forever.

My Own Example: Conlon and Mortimer

Idea:

- We know that $\sum_k D_{jk} = 1$ and $D_{jk} \geq 0$ and would like to impose this.
- We have lots of information about certain substitutes but not others.

Assume that $\mathbf{D_{j,\cdot}} \sim \text{Dirichlet}(m, p_1, \ldots, p_K, p_0)$.

- This is like having m observations from a "multinomial" prior distribution.
- In enforces that probabilities are positive and sum to one.
- Now we have something like Δq_{jt} observations for each (k,t) so that the more information we have, the less shrinkage.

We also try a beta prior so that $\hat{D}_{jk} = (1-\lambda)p_k + \lambda \cdot \frac{\Delta q_k}{\Delta q_j}$ where $\lambda = \frac{\Delta q_j}{\Delta q_j + m}$.

Candy Bars

Mfg	Product	Treated Machine Weeks	Δq_k Subst Sales	Δq_j Focal Sales	$\Delta q_k / \lvert\Delta q_j\rvert$ Div	Assn 3 Diversion $(m = K)$	Assn 3 Diversion $(m = 300)$	Assn 4 Diversion $(m = 4.15)$
			Snickers Removal					
Mars	M&M Peanut	176	375.5	-954.3	39.4	37.0	30.8	18.4
Mars	Twix Caramel	134	289.6	-702.4	41.2	37.9	29.5	15.9
Pepsi	Rold Gold (Con)	174	161.4	-900.1	17.9	16.8	13.9	7.5
Nestle	Butterfinger	61	72.9	-362.8	20.1	17.1	11.2	4.5
Mars	M&M Milk Chocolate	97	71.8	-457.4	15.7	13.8	9.8	4.1
Kraft	Planters (Con)	136	78.0	-759.9	10.3	9.6	7.8	3.8
Kellogg	Zoo Animal Cracker	177	65.7	-970.2	6.8	6.5	5.7	2.9
Pepsi	Sun Chip	159	45.3	-866.1	5.2	5.0	4.3	2.1
Hershey	Choc Hershey (Con)	41	29.8	-179.6	16.6	12.2	6.3	2.0
Kellogg	Rice Krispies Treats	17	17.7	-66.5	26.7	13.5	5.0	1.3
Misc	Farleys (Con)	18	14.9	-114.2	13.0	8.3	3.7	1.0
Nestle	Nonchoc Nestle (Con)	3	9.4	-10.5	89.5	12.4	3.1	0.7
Mars	Choc Mars (Con)	11	6.4	-32.7	19.7	6.5	2.0	0.4
Hershey	Payday	2	1.1	-9.8	10.9	1.4	0.4	0.1
Mars	3-Musketeers	2	0.0	0.0				
Misc	BroKan (Con)	3	0.0	0.0				
	Outside Good	180	460.9	-970.2	47.5			23.1

Fully Hierarchical Models: What's the point

Suppose we want to estimate a lognormal distribution for income in different places

- Fully Pooled: estimate a single (μ, σ)
- Non-Pooled: estimate a separate (μ_k, σ_k) for each zip code
- Paritally-Pooled: some combination of the two?
  - Allow $\mu_k \sim F(\mu, \sigma)$.
  - estimate both the common (hyper) parameter and the group specific one.
  - Limit high variance from small groups.

Fully Hierarchical Models

Suppose we have several groups, each with their own mean:

$$\begin{aligned} b_i &\sim \mathcal{N}\left(0, \sigma_b^2\right) \\ \mu_i &= \mu + b_i \\ y_{ij} &\sim \mathcal{N}\left(\mu_i, \sigma_y^2\right) \end{aligned}$$

Or we could have written:

$$y_{ij} = \mu + \underbrace{b_i}_{\sim\mathcal{N}\left(0,\sigma_b^2\right)} + \underbrace{\epsilon_{ij}}_{\sim\mathcal{N}\left(0,\sigma_y^2\right)}$$

That is there is a common mean μ and a group specific mean b_i.

Fully Hierarchical Models

$$y_{ij} = \mu + \underbrace{b_i}_{\sim\mathcal{N}\left(0,\sigma_b^2\right)} + \underbrace{\epsilon_{ij}}_{\sim\mathcal{N}\left(0,\sigma_y^2\right)}$$

- Sometimes we interpret b_i as a random effect
- In any case we will get some shrinkage to the overall mean μ
- We could estimate by MLE if we know which b_i corresponded to which y_{ij} otherwise via Bayesian methods.

Thanks!