STSCI 4780/5780
Shrinkage estimation and Hierarchical and empirical Bayes

Tom Loredo, CCAPS & SDS, Cornell University

Hierarchical (graphical) Bayesian modeling plan

Key idea: Sharing information across a population

- Today: Basic hierarchical modeling and shrinkage estimation for a *population* of scalars:
  - Background: Repeated sampling performance & bias/variance tradeoff
  - A population of baseball players
  - A population of coin tossers
- Measurement error models:
  - Lec21 — Density estimation with measurement error (population of scalars/vectors)
  - Lec22 — Regression with measurement error

Terminology

These terms are used for essentially the same classes of models:

- Hierarchical Bayesian model ("hierarchical Bayes," HB)
- Multilevel model (MLM)
- (Probabilistic) Graphical model (PGM)
- Bayesian network (Bayes net)

Empirical Bayes may be considered either an approximation to HB, or a frequentist alternative (multi-case inference with an estimated prior)

Repeated sampling performance of estimators

Setting

Consider a parametric model with sampling distribution $p(D|\theta)$ for data D, parameters θ

Construct a *point estimator* for θ (or some other quantity of interest), $\tilde{\Theta}(D)$

- Bayes: Posterior mode or mean or median...
- Frequentist: MLE, method of moments, best linear unbiased estimator (BLUE)...

How well do we expect $\tilde{\Theta}(D)$ to perform *on average*?

Address this via properties of the $\tilde{\Theta}(D)$ sampling distribution

$$p(\tilde{\theta}|\theta) = \int \mathrm{d}D \, p(D, \tilde{\theta}|\theta) = \int \mathrm{d}D \, p(D|\theta) \, \delta\left[\tilde{\theta} - \tilde{\Theta}(D)\right]$$

Note: In general, performance will depend on θ

Monte Carlo replication study

$$p(\tilde{\theta}|\theta) = \int \mathrm{d}D\, p(D, \tilde{\theta}|\theta) = \int \mathrm{d}D\, p(D|\theta)\, \delta\left[\tilde{\theta} - \tilde{\Theta}(D)\right]$$

Replicate the experiment:

1. Set θ to a fixed value
2. Draw a full dataset D from $p(D|\theta)$
3. Compute $\tilde{\Theta}(D)$
4. Repeat from (2) many times $\rightarrow$ sampling dist'n for $\tilde{\Theta}(D)$, $p(\tilde{\theta}|\theta)$ (e.g., as a histogram)
5. Repeat from (1), using a different θ

Viewpoints/motivations (cf. Lab10)

- Bayes: Pre-data comparison of choices of modeling choices and/or posterior summary (mode, mean, median. . .); the natural criteria average over the uncertain θ (using the prior)
- Frequentist:
  - Ideal: Seek estimator whose performance is *independent* of θ (not always possible—you need to be both lucky & clever!)
  - More commonly: Seek estimator with good *worst-case* performance

Error and bias

The *error* made if we use $\tilde{\Theta}(D)$ in place of θ is

$$e(D) = \tilde{\Theta}(D) - \theta$$

The *bias* of the estimator is the *expected error* (as a function of θ):

$$b(\theta) \equiv \mathbb{E}[\tilde{\Theta}(D) - \theta] \;=\; m(\theta) - \theta$$

where $m(\theta) = \mathbb{E}[\tilde{\Theta}(D)]$, and the expectation is WRT $p(D|\theta)$, averaging/integrating over D

An estimator with $b(\theta) = 0$ is an *unbiased estimator*

If $b(\theta) = b$ (a constant), we can subtract it off from the original $\tilde{\Theta}(D)$ to get an unbiased estimator, but usually the bias depends on θ (which we won't know in practice!)

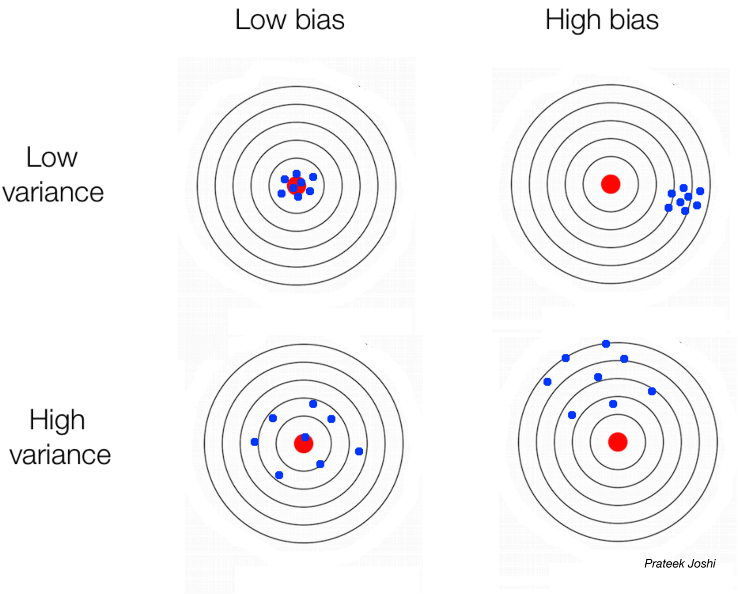
Other "typical" values of $\tilde{\Theta}(D)$ (measures of "central tendency") may be interesting—mode, median—but the bias is usually the easiest to analyze

Variability and variance

In repeated sampling with θ *fixed*, the estimator will give an ensemble of estimates

A measure of variability of the estimator is the variance *with respect to the mean*, i.e., the expected squared distance of an estimate from its expectation value:

$$\begin{aligned} v(\theta) &\equiv \mathbb{E}\left[\left(\tilde{\Theta}(D) - m(\theta)\right)^2\right] \\ &= \mathbb{E}\left[\tilde{\Theta}^2(D)\right] + m^2(\theta) - 2\mathbb{E}\left[\tilde{\Theta}(D)\, m(\theta)\right] \\ &= \mathbb{E}\left[\tilde{\Theta}^2(D)\right] - m^2(\theta) \end{aligned}$$


Low bias
High bias
Low variance
High variance
Prateek Joshi

Mean squared error (MSE)

Note that $v(\theta)$ is a measure of distance from $m(\theta)$, not from θ itself (the "true" value)

MSE is the average squared distance from $\tilde{\Theta}(D)$ to θ itself—a *"variance with respect to the truth"*:

$$\begin{aligned} \mathrm{MSE}(\theta) &\equiv \mathbb{E}\left[\left(\tilde{\Theta}(D) - \theta\right)^2\right] \\ &= \mathbb{E}\left[\tilde{\Theta}^2(D)\right] + \theta^2 - 2\theta m(\theta) \end{aligned}$$

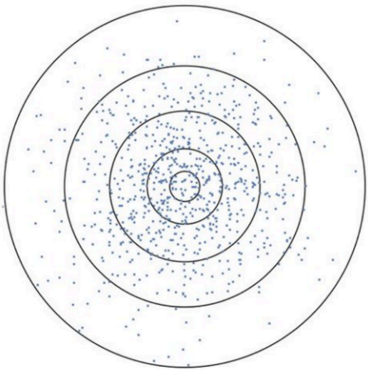
Recall that $b(\theta) = m(\theta) - \theta$, so that

$$b^2(\theta) = m^2(\theta) + \theta^2 - 2\theta m(\theta)$$

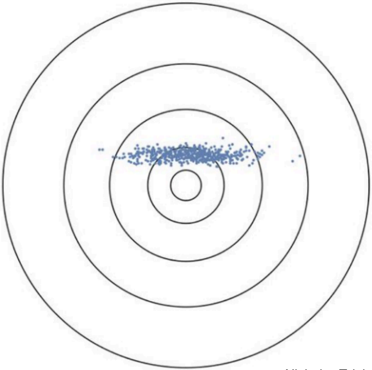
$$\begin{aligned} \Rightarrow\ \mathrm{MSE}(\theta) &= \mathbb{E}\left[\tilde{\Theta}^2(D)\right] - m^2(\theta) + b^2(\theta) \\ &= v(\theta) + b^2(\theta) \end{aligned}$$

For an *unbiased* estimator, $v(\theta)$ measures the average scale of the error, but for a *biased* estimator, we have to worry about the $b^2(\theta)$ contribution $\rightarrow$ *bias-variance tradeoff*

No bias, but large MSE



Biased, but smaller MSE

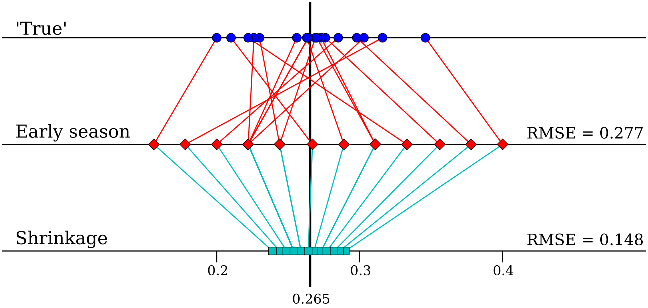


Nicholas Taleb

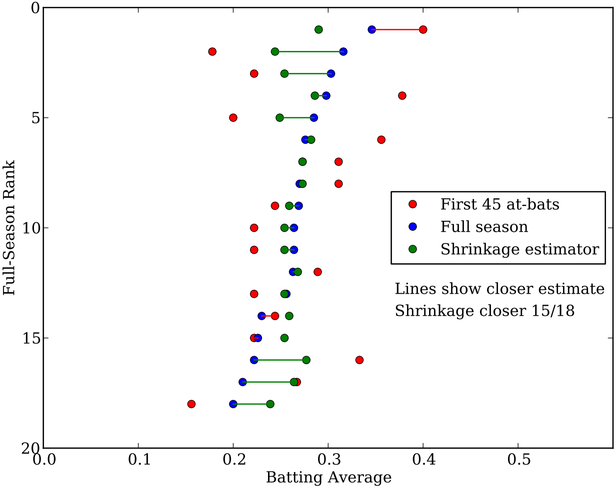
1970 baseball batting averages

Efron & Morris looked at batting averages of all-star baseball players who had $N = 45$ at-bats in May 1970 — 'large' N & includes Roberto Clemente (outlier!)

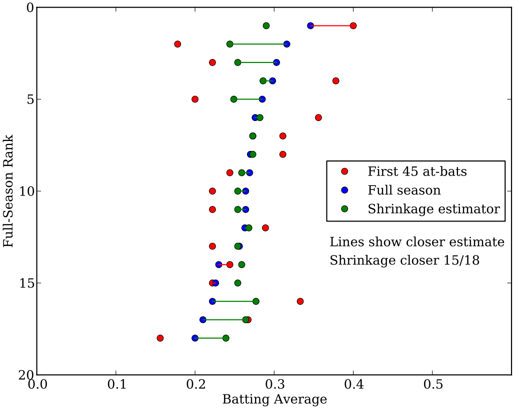
Red $= n/N$ maximum likelihood estimates of true averages
Blue $=$ Remainder of season, $N_{\mathrm{rmdr}} \approx 9N$ (surrogate for truth)



Cyan $=$ James-Stein estimator: nonlinear, correlated, biased
But *better*!



Theorem (independent Gaussian setting): When estimating *three or more normal means*, *shrinkage estimators always beat independent MLEs* in terms of expected RMS error (over the ensemble)



"The single most striking result of post-World War II statistical theory."
— Brad Efron

"Probably the most startling statistical discovery of the past century."
— Lawrence Brown

"Stunned with disbelief."
— Erich Lehmann's reaction

Some shrinkage estimators

In place of batting average f_i (a Bernoulli probability parameter for team member i), use a *variance stabilizing transform* to get x_i that each have an approximately normal distribution with $\sigma = 1$:

$$x_i = \sqrt{45}\,\arcsin(2f_i - 1)$$

Compute the squared magnitude of the x vector:

$$s^2 = \sum_{i=1}^{N} x_i^2$$

The *James-Stein estimator* is

$$\hat{\theta}_i^{\mathrm{JS}} = \left(1 - \frac{C}{s^2}\right) x_i$$

The best value of C is $C = N - 2$

Stein, and then James & Stein, motivated this from the *geometry of multivariate normal distributions*

Efron & Morris estimator: "An astute follower of baseball might be aware that just as each player's batting ability can be represented by a Gaussian curve, so too the true batting abilities of all major-league players have an approximately normal distribution.... With this valuable extra information, which statisticians call a 'prior distribution,' it is possible to construct a superior estimate of each player's true batting ability."

$$\bar{x} = \frac{1}{N}\sum_i x_i; \qquad r^2 = \sum_i (x_i - \bar{x})^2 \qquad \text{(population averages)}$$

$$\begin{aligned}\hat{\theta}_i^{\mathrm{EM}} &= \bar{x} + \left(1 - \frac{K}{r^2}\right)(x_i - \bar{x}) \\ &= \bar{x}\left[1 - \left(1 - \frac{K}{r^2}\right)\right] + \left(1 - \frac{K}{r^2}\right) x_i\end{aligned}$$

The best K is $K = N - 3$

Dennis Lindley: This looks like Bayesian inference using a conjugate "prior," but with μ_0 *determined by the data* (JS resembles $\mu_0 = 0$)

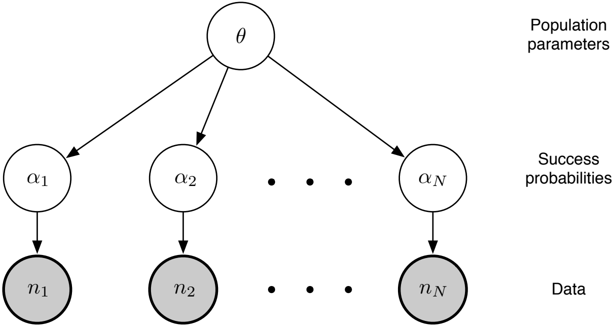
A population of coins/flippers





Each flipper+coin flips different number of times

- What do we learn about the *population* of coins—the distribution of αs?
- How does population membership affect inference for a single coin's α?



$$p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta)\, p(n_i|\alpha_i)$$
$$= \pi(\theta) \prod_i p(\alpha_i|\theta)\, \ell_i(\alpha_i)$$

Terminology:

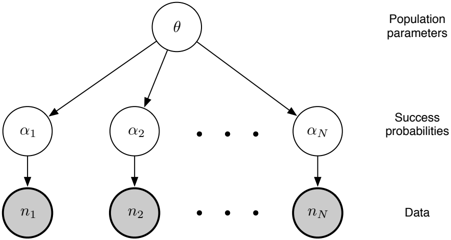
- θ are *hyperparameters*
- $\pi(\theta)$ is the *hyperprior*
- $\ell_i(\alpha_i) \equiv p(n_i|\alpha_i)$ is a member or item likelihood function

A simple multilevel model: beta-binomial

Goals:

- Learn a population-level "prior" by pooling data
- Account for population membership in member inferences

Qualitative



$$p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta)\, p(n_i|\alpha_i)$$
$$= \pi(\theta) \prod_i p(\alpha_i|\theta)\, \ell_i(\alpha_i)$$

Quantitative

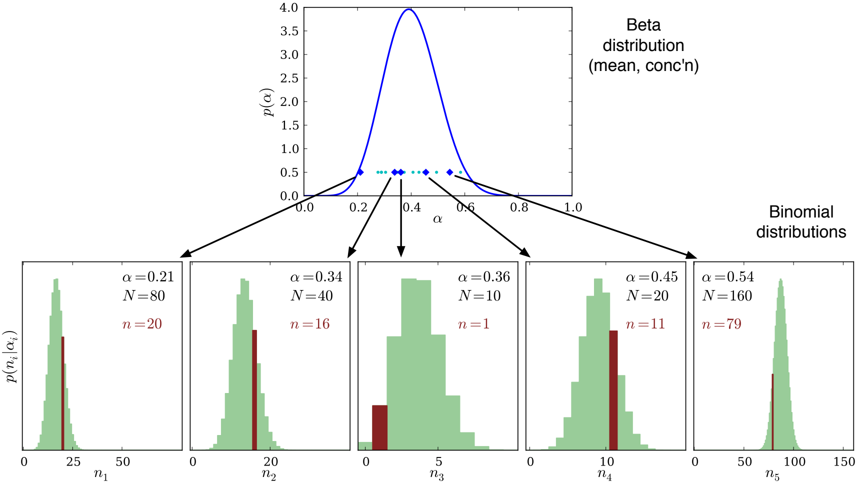
$$\theta = (a, b) \text{ or } (\mu, \sigma)$$

$$\pi(\theta) = \mathrm{Flat}(\mu, \sigma)$$

$$p(\alpha_i|\theta) = \mathrm{Beta}(\alpha_i|\theta)$$

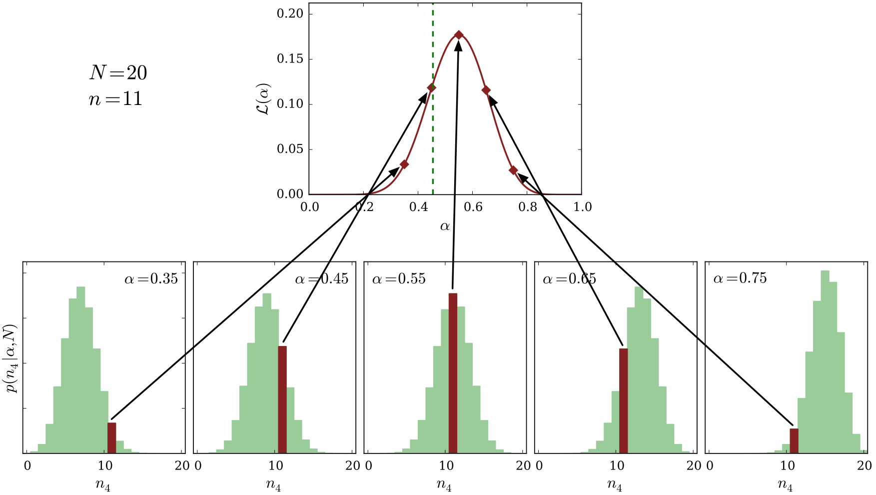
$$p(n_i|\alpha_i) = \binom{N_i}{n_i} \alpha_i^{n_i} (1 - \alpha_i)^{N_i - n_i}$$

Generating the population & data

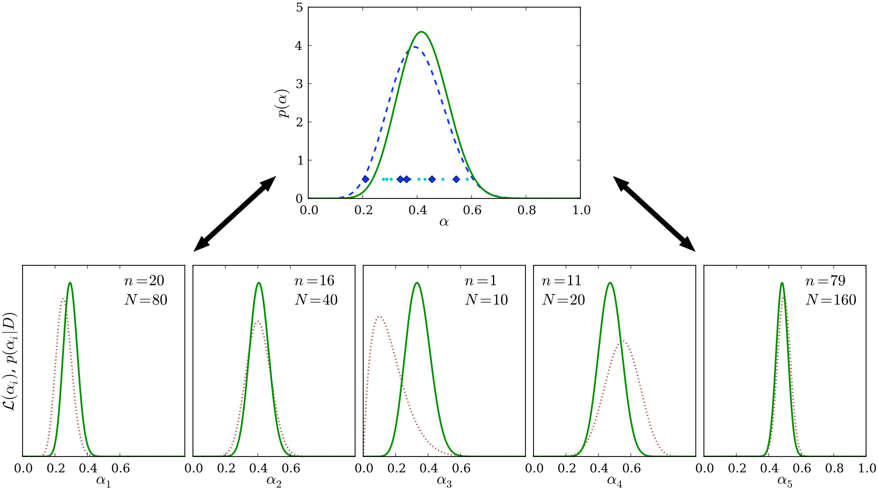


Likelihood function for one member's α

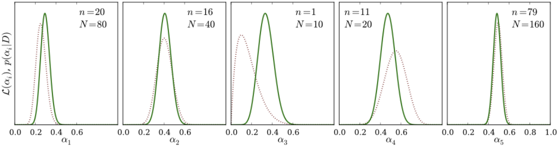
$N = 20$
$n = 11$

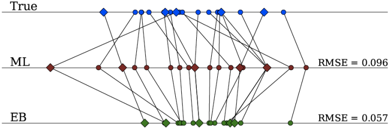


Learning the population distribution



Lower level estimates





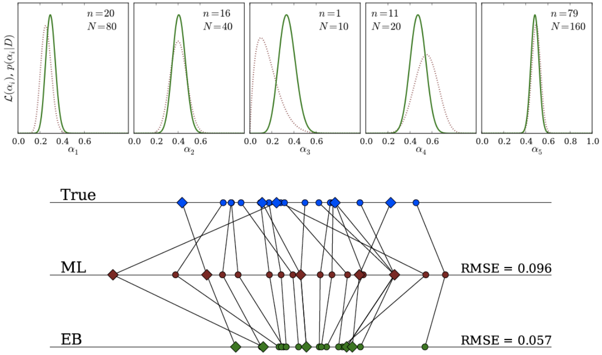
Two approaches to member/item estimation

- **Hierarchical Bayes (HB):** Calculate marginals

$$p(\alpha_j|\{n_i\}) \propto \ell_j(\alpha_j) \int d\theta\, \pi(\theta)\, p(\alpha_j|\theta) \left[\prod_{i\neq j} \int d\alpha_i\, p(\alpha_i|\theta)\, \ell_i(\alpha_i) \right]$$

- **Empirical Bayes (EB):** Plug in an optimum $\hat{\theta}$ and estimate $\{\alpha_i\}$
  View as approximation to HB, or a frequentist procedure that estimates a prior from the data (in multi-case settings)

Lower level estimates



Bayesian outlook — Information sharing across the pop'n

- Marginal posteriors are *narrower* than likelihoods
- Point estimates tend to be closer to true values than MLEs (averaged across the population)
- Joint distribution for $\{\alpha_i\}$ is *dependent*

Frequentist outlook

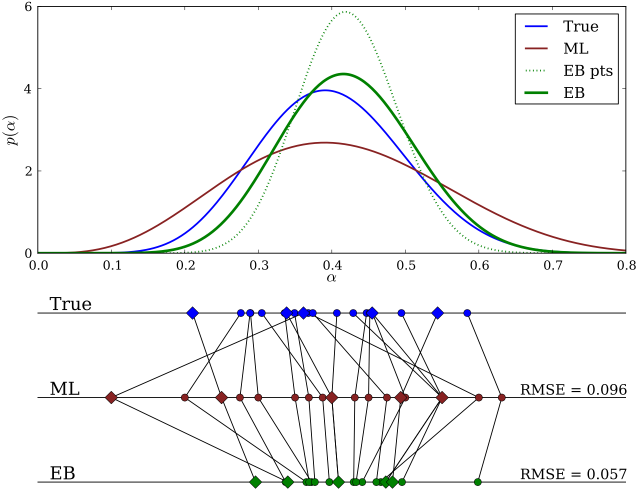
- Point estimates are biased
- Reduced variance $\rightarrow$ estimates are closer to truth on average (lower MSE in repeated sampling)
- Bias for one member estimate depends on data for all other members

Lingo

- Estimates *shrink* toward prior/population mean
- Estimates "muster and *borrow strength*" across population (Tukey's phrase); increases accuracy and precision of estimates
- Efron* describes shrinkage as a consequence of accounting for *indirect evidence*

*Bradley Efron (2010): "The Future of Indirect Evidence"

Population and member estimates



Competing data analysis goals

"Shrunken" member estimates provide improved & reliable estimate for population member properties

But they are *under-dispersed* in comparison to the true values → not optimal for estimating *population* properties*

No point estimates of member properties are good for all tasks!

We should view population data tables/catalogs as providing *descriptions of member likelihood functions*, not "estimates with errors"

*Louis (1984); Eddington noted this in 1940!