The Frequentist Perspective of Shrinkage Estimation

Recall our question from the Empirical Bayes' Lecture:

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

We Reduced this problem to estimating the parameters of the following parameter model:

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

μ_i = latent i^{th} "true" quality
X_i = i^{th} batting average (observed)
σ_i^2 = known variance
μ, τ^2 = unknown hyperparameters

Which we solved using Empirical Bayes.
We now consider another perspective to understand why Empirical Bayes works.

Since σ_i^2 is known, we can divide by σ_i^2 to remove some parameters: $X_i \leftarrow X_i/\sigma_i^2$, $\theta_i \leftarrow \mu_i/\sigma_i^2$

and also suppose we are Frequentists and not Bayesians, i.e. we think it's Ridiculous to treat a parameter as a Random variable because it is more plausible to think of them as unknown fixed constants, i.e. Mookie's unknown "true" quality!

Then we are left with the model $X_i \overset{ind}{\sim} N(\theta_i, 1)$ $i=1,\dots,k$

and our task is to estimate the fixed unknown constants (the normal means) θ_i given the data $\{X_i\}_{i=1}^k$ to optimize the composite loss function $L(\theta, \hat{\theta}) = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2$ where $\theta = (\theta_1, \dots, \theta_k)^T$ and $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)^T$.

The performance of the joint estimator $\hat{\theta}$ is judged by the Risk function $R(\theta, \hat{\theta}) = \mathbb{E}\, L(\theta, \hat{\theta})$.

This simple setup leads to one of the most provocative results in mathematical statistics: Stein's paradox / shrinkage estimation / the James Stein estimator.

The estimation problem involves pairs of values (X_i, θ_i) $i = 1, \ldots, k$ where one element of each pair X_i is known and one θ_i is unknown.

The "obvious" or "ordinary" estimator is just $\hat{\theta}_i^{(MLE)} = X_i$.

This is the <u>maximum likelihood estimator (MLE)</u> or the parameters that maximize the probability of observing the data.

$$\hat{\theta}^{(MLE)} = \underset{\theta}{\operatorname{argmax}} \; \mathbb{P}(\text{data} \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \; \mathbb{P}(X_1, \ldots, X_k \mid \theta_1, \ldots, \theta_k)$$

$$= \underset{\theta}{\operatorname{argmax}} \; \prod_{i=1}^{k} \mathbb{P}(X_i \mid \theta_i) \quad \text{by independence}$$

$$= \underset{\theta}{\operatorname{argmax}} \; \log \prod_{i=1}^{k} \mathbb{P}(X_i \mid \theta_i) \quad \text{by monotonicity of log}$$

$$= \underset{\theta}{\operatorname{argmax}} \; \sum_{i=1}^{k} \log \mathbb{P}(X_i \mid \theta_i)$$

$$= \underset{\theta}{\operatorname{argmax}} \; \sum_{i=1}^{k} \log \mathcal{N}(X_i \mid \theta_i, 1) \quad \text{by our model}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{K} \log \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \theta_i)^2\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{K} -\frac{1}{2}(X_i - \theta_i)^2$$

$$= X.$$

In terms of our baseball example, just predict each player's mid season batting average.

Turns out these predictions are <u>terrible</u>.

We saw this in the empirical Bayes lecture.

But here the θ_i are unknown fixed constants and a prior is misspecified; so why does it work?

The estimation problem involves pairs of values (X_i, θ_i) $i=1,\dots,k$ where one element of each pair X_i is known and one θ_i is unknown. Since the θ_i's are unknown we cannot plot the pairs, but we imagine what such a plot would look like to help us understand the problem.

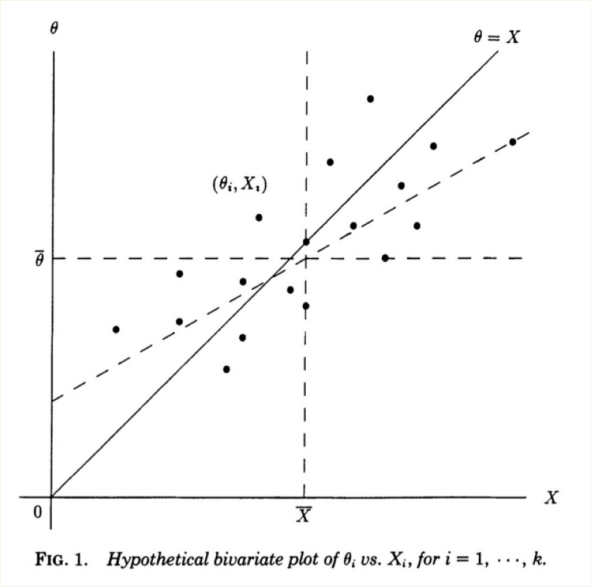

FIG. 1. *Hypothetical bivariate plot of θ_i vs. X_i, for $i = 1, \cdots, k$.*

Since X is $N(\theta, 1)$ we can think of the X's as being generated by $N(0,1)$ "errors" to the given θ_i's. So the horizontal deviations of the X_i's from the 45° line $\theta = X$ are independent $N(0,1)$.

Our goal is to estimate all of the θ_i's given all of the X_i's with no assumptions about a possible distributional structure for the θ_i's — they are simply to be viewed as unknown constants.

Q. Why should we expect that the ordinary estimator $\hat{\theta}_i^{(MLE)} = X_i$ can be improved upon?

Well, if the θ_i's, and hence the pair (X_i, θ_i), had a known joint distribution, a natural method of proceeding is $\hat{\theta}(x) = \mathbb{E}(\theta|X)$ and use this, the theoretical Regression function of θ on X, to generate estimates of the θ_i's by evaluating it for each X_i. This is an unattainable ideal because we do not know the conditional distribution of θ given X.

Moreover, we don't assume that our uncertainty about the unknown constants θ_i can be described by a probability distribution at all; we are not Bayesian.

We do know the conditional distribution of X given θ, $N(\theta, 1)$, and we can calculate $\mathbb{E}(X|\theta) = \theta$.

Indeed this theoretical regression line corresponds to the 45° line $\theta = x$, and this line yields the ordinary estimators $\hat{\theta}_i^{(MLE)} = X_i$. Thus the ordinary estimator may be viewed as being based on the "wrong" regression line, on $\mathbb{E}(X|\theta)$ rather than on $\mathbb{E}(\theta|X)$.

As Francis Galton knew in the 1880s, the regression of X on θ and θ on X can be markedly different (HW), suggesting that the ordinary estimator can be improved upon and even suggesting how this might be done: by attempting to approximate $\mathbb{E}(\theta|x)$ or whatever that might mean in a setting where the θ's do not have a distribution.

With no distributional assumptions on the θ_i's, we are of course prevented from looking at an optimal estimator of $\mathbb{E}(\theta|x)$.

Instead we note that $\hat{\theta}_i^{(MLE)} = X_i$ is a linear function of X_i and we can look for a best linear estimator of the form $\hat{\theta}_i = a + bX_i$ so as to minimize the composite loss function $L(\theta, \hat{\theta}) = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2$.

If the θ_i's were actually available to us, we would have a standard simple linear regression problem, with best linear estimator given by the regression line $\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X})$ where $\hat{\beta} = \frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum (X_i - \bar{X})^2}$.

The θ_i are not available, but if we can estimate the functions $\bar{\theta}$ and $\hat{\beta}$ of these unknown parameters we can estimate the regression line of θ on X.

$\bar{X}$ is the obvious estimator of $\bar{\theta}$.

$$\hat{\beta} = \frac{\sum(X_i-\bar{X})(\theta_i-\bar{\theta})}{\sum(X_i-\bar{X})^2} = \frac{S_{X\theta}}{S^2}$$

where $S^2 = \frac{1}{K-1}\sum_{i=1}^{K}(X_i-\bar{X})^2$ is the sample variance, an unbiased estimate of $\mathrm{var}(X)$, and $S_{X\theta} = \frac{1}{K-1}\sum_{i=1}^{K-1}(X_i-\bar{X})(\theta_i-\bar{\theta})$ is the sample covariance, an unbiased estimate of $\mathrm{cov}(X,\theta)$. But $\{\theta_i\}$ are unknown so $S_{X\theta}$ is unknown. We need to estimate it from the data $\{X_i\}$.

Since $X = \theta + \varepsilon$ where $\varepsilon \sim N(0,1)$

and $\mathrm{var}(X) = \mathrm{var}(\theta) + \mathrm{var}(\varepsilon) = \mathrm{var}(\theta) + 1$,

$\mathrm{cov}(X,\theta) = \mathrm{cov}(\theta+\varepsilon, \theta) = \mathrm{var}(\theta) + \cancel{\mathrm{cov}(\varepsilon,\theta)}^{0}$

$= \mathrm{var}(\theta) = \mathrm{var}(X) - 1$,

so $\frac{1}{K-1}\sum(X_i-\bar{X})^2 - 1$ and $\frac{1}{K-1}\sum_{i=1}^{K-1}(X_i-\bar{X})(\theta_i-\bar{\theta})$ have the same expectation.

Hence $\hat{\beta} = \frac{\sum(X_i-\bar{X})(\theta_i-\bar{\theta})}{\sum(X_i-\bar{X})^2} \approx \frac{\sum(X_i-\bar{X})^2-(K-1)}{\sum(X_i-\bar{X})^2} = 1 - \frac{(K-1)}{S^2}$.

This leads to the Efron-Morris estimated least squares line

$$\hat{\theta}_i^{(EM)} = \bar{X} + \left(1 - \frac{K-1}{S^2}\right)(X_i - \bar{X}).$$

James Stein's original shrinkage estimator can be derived by considering the class of estimators that are linear in X with 0 intercept, $\hat{\theta}_i = bX_i$. The least squares estimate has $\hat{\beta} = \frac{\sum \theta_i X_i}{\sum X_i^2}$ and $\theta_i X_i$ has the same expectation as $\sum X_i^2 - K$, yield the James Stein estimator

$$\hat{\theta}^{(JS)} = \left(1 - \frac{K}{S^2}\right) X_i.$$

So, the ordinary estimators $\hat{\theta}_i^{(MLE)} = X_i$ are derived from the theoretical regression line of X on θ, which is useful if our goal is to predict X from θ. But, our goal is the reverse, to predict θ from X using the sum of squares criterion $\sum(\theta_i - \hat{\theta}_i)^2$, so the optimal estimator is the least squares regression line of θ on X, and the James Stein and Efron Morris estimators are themselves approximations of this Right regression line.

It turns out that both of these estimators are better (have less Risk) than the ordinary estimate (the MLE), but we need to show this.

Let $\hat{\theta}^b = bX_i$ represent the class of linear estimators with zero intercept. The Risk is

$$R(\theta, \hat{\theta}^b) = \mathbb{E}\, L(\theta, \hat{\theta}^b)$$

$$= \mathbb{E} \sum (\theta_i - \hat{\theta}_i^b)^2$$

$$= \mathbb{E} \sum (\theta_i - \hat{\theta}_i^{LS} + \hat{\theta}_i^{LS} - \hat{\theta}_i^b)^2$$

where $\hat{\theta}_i^{LS} = \hat{\beta} X_i$ where $\hat{\beta} = \Sigma \theta_i X_i / \Sigma X_i^2$

$$= \mathbb{E}\left[\sum (\theta_i - \hat{\theta}_i^{LS})^2 + 2 \sum (\theta_i - \hat{\theta}_i^{LS})(\hat{\theta}_i^{LS} - \hat{\theta}_i^b) + \sum (\hat{\theta}_i^{LS} - \hat{\theta}_i^b)^2\right]$$

($2 \sum (\theta_i - \hat{\theta}_i^{LS})(\hat{\theta}_i^{LS} - \hat{\theta}_i^b) \to 0$)

$$= R(\theta, \hat{\theta}^{LS}) + \mathbb{E} \sum (\hat{\beta} X_i - b X_i)^2$$

$$= R(\theta, \hat{\theta}^{LS}) + \mathbb{E}\left[(\hat{\beta} - b)^2 S^2\right] \quad \text{where } S^2 = \Sigma X_i^2$$

so a James Stein estimator will improve an ordinary estimator iff $\mathbb{E}((\hat{\beta} - b)^2 S^2) < \mathbb{E}((\hat{\beta} - 1) S^2)$ for all θ. Since $\hat{b}$ is a "reasonable" estimator of β but the constant 1 is not, we can expect James Stein to dominate the MLE.

Theorem (Stein's "Paradox").

Suppose $\{X_i\}_{i=1}^{k}$ are drawn independently by $X_i \sim N(\theta_i, 1)$.

Then the James Stein estimator of θ,

$$\hat{\theta}_i^{(JS)} = \left(1 - \frac{c}{S^2}\right) X_i,$$

where $S^2 = \sum X_i^2$, $k \geq 3$, and $0 < c \leq 2(k-2)$

and the Efron Morris Estimator of θ,

$$\hat{\theta}_i^{(EM)} = \bar{X} + \left(1 - \frac{c}{S^2}\right)(X_i - \bar{X}),$$

where $S^2 = \Sigma(X_i - \bar{X})^2$, $k \geq 4$, and $0 < c \leq 2(k-3)$

has uniformly lower squared error risk than

(i.e., uniformly dominates)

the "obvious" maximum likelihood estimator of θ,

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

i.e. $R(\theta, \hat{\theta}^{(JS)}) < R(\theta, \hat{\theta}^{(MLE)}) \quad \forall \theta$

and $R(\theta, \hat{\theta}^{(EM)}) < R(\theta, \hat{\theta}^{(MLE)}) \quad \forall \theta.$

Proof See Stigler (1981) Page 150

Shrinkage

$\hat{\theta}^{(JS)}$ is the weighted average of 0 and X_i and so shrinks the ordinary estimator $\hat{\theta}^{(MLE)}$ towards 0.

$\hat{\theta}^{(EB)}$ is the weighted average of $\bar{X}$ and X_i and so shrinks the ordinary estimator $\hat{\theta}^{(MLE)}$ towards $\bar{X}$.

These shrinkage estimators dominate the ordinary estimates as long as $K \geq 4$.

Connection to Empirical Bayes

$$\hat{\theta}_i^{(EB)} = \bar{X} + \left(\frac{\tau^2}{\tau^2+1}\right)(X_i - \bar{X})$$

has the same form as the Efron Morris estimator, shrinking towards the overall mean $\bar{X}$.

The prior variance τ^2 (where $\theta_i \sim N(0, \tau^2)$) determines how much we shrink.

Pardox?

To estimate θ_i it is optimal to use information from all the other observations $\{X_j\}_{j \neq i}$, via $\bar{X}$ and S^2, even though the X_i are drawn independently and are unrelated in the sense that each has its own mean θ_i.

This seems preposterous!

How can information about Mookie Betts Batting average and Shohei Ohtani's batting average be used to improve an estimate of Freddie Freeman's batting average?

How can information about the price of apples in Washington and about the price of oranges in Florida be used to improve an estimate of the true value of French wine, when it is assumed they are unrelated?

Consultant's Dilemma

In the middle of the season Billy Beane asks you to predict Mookie Betts' end-of-season performance (using data available to you only during that season).

The estimator that is best on average across all players (a shrinkage estimator) is different from the estimator that is best for one specific individual player (the MLE).

Optimizing for the squared error aggregated across all players is Not the same as optimizing for the errors of separate estimators of the individual parameters. A combined shrinkage estimator should be used to optimize a combined loss, but this combined estimator is worse if we want to estimate just one individual parameter.

Which estimator should we use for Mookie?