The Galtonian 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, predict each player's 2nd-half-of-season batting average.

We reduced this problem to estimating parameters of this model:

$$\begin{cases} X_i \sim N\left(P_i, \underbrace{\frac{c}{N_i}}_{\sigma_i^2 \text{ Known}}\right) \\ P_i \sim N(\mu, \tau^2) \end{cases}$$

X_i = mid-season batting average

We used Empirical Bayes to estimate $\{P_i\}$.

Today: another perspective on why this works.

$X_i \leftarrow \frac{X_i}{\sigma_i}$ / $\theta_i \leftarrow \frac{P_i}{\sigma_i}$ / Kill the prior

New Model

$$\boxed{X_i \overset{ind}{\sim} N(\theta_i, 1)}$$

θ_i = "true" latent batter quality parameter

X_i = data, transformed batting average

i = batter index

Today we are Frequentists, not Bayesians. We want to estimate each player's θ_i from the data $\{X_i\}$ but we think it's weird to treat θ_i as a Random Variable, so we instead think of it as an unknown fixed constant.

Our task is to estimate the fixed unknown constants, i.e. the Normal means, $\{\theta_i\}$ given the data $\{X_i\}$, which is to create an estimator $\{\hat{\theta}_i\}$, and we want this loss function

$$L(\theta, \hat{\theta}) = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2$$

MSE

to be small.

θ: fixed unknown constants $\theta = (\theta_1, \dots, \theta_n)$

$\hat{\theta}$: an estimator, a function of the data, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ is a Random Variable

Consider Risk Function $R(\theta, \hat{\theta}) = \mathbb{E}\, L(\theta, \hat{\theta})$.

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

This estimation problem involves pairs of values $\{(X_i, \theta_i)\}_{i=1}^{n}$ where one element of each pair X_i is known and one, θ_i, is unknown.

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

$$\boxed{X_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1)}$$

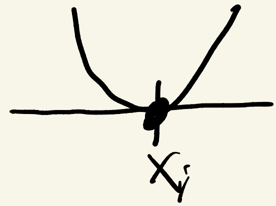
$$\hat{\theta}_i^{(MLE)} = \underset{\theta_i}{\operatorname{argmax}}\ \mathbb{P}(\text{data} \mid \text{param } \theta)$$

$$= \underset{\theta_i}{\operatorname{argmax}}\ \mathbb{P}(X_i \mid \theta_i)$$

$$= \underset{\theta_i}{\operatorname{argmax}}\ \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(X_i - \theta_i)^2\right)$$

$$= \underset{\theta_i}{\arg\max} \; -\frac{1}{2}(X_i-\theta_i)^2$$

$$= \underset{\theta_i}{\arg\min} \; (X_i-\theta_i)^2$$



$$= X_i$$

MLE: predict the 2^{nd}-half-of-season batting avg to be just the midseason batting average.

↳ Not the best the we can do

With Empirical Bayes we saw that we can do better by shrinking the estimates towards a common mean,
but here we are Frequentists, a prior is misspecified, and why does shrinkage work?

Pairs $\{(X_i, \theta_i)\}_{i=1}^{n}$

X_i = known
θ_i = unknown
estimate θ_i

Since $\{\theta_i\}$ unknown we cannot plot the pair (X_i, θ_i), but for a second imagine what such a plot would look like.

$X_i \overset{ind}{\sim} N(\theta_i, 1)$

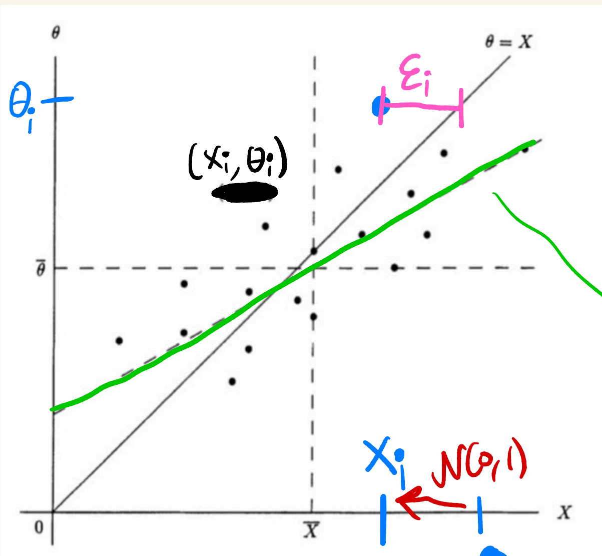


FIG. 1. Hypothetical bivariate plot of θ_i vs. X_i, for $i = 1, \cdots, k$.

$X_i = \theta_i + \varepsilon_i$
$\varepsilon_i \sim N(0,1)$

best fit line of the points,

Regression line of θ on X

the MLE is based on the black line

Shrinkage estimators are based on the green line

$X_i \sim \mathcal{N}(\theta_i, 1)$

Let's consider these 2 lines and build estimators from them.

Black line: $\mathbb{E}(X|\theta) = \theta \longrightarrow X = \theta$

Green line: $\mathbb{E}(\theta|X)$

* $\hat{\theta}_i^{(MLE)} = X_i$ is a linear function of X_i

Inspired by the plot, despite its weirdness, consider a "better" linear estimator of the form $\hat{\theta}_i = a + bX_i$ so as to minimize $L(\theta, \hat{\theta}) = \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2$.

$\boxed{X_i \overset{ind}{\sim} N(\theta_i, 1)}$

$\{\theta_i\}$ unknown. If they were known:

Simple linear regression

$\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X})$, $\hat{\beta} = \dfrac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum (X_i - \bar{X})^2}$

doesn't make sense.

Can we instead estimate $\bar{\theta}, \hat{\beta}$ term and plug those in?

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

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(\theta_i - \bar{\theta})}{\sum (x_i - \bar{x})^2} = \frac{S_{x\theta}}{S_x^2}$$

← Sample covariance b/t X and θ

← Sample variance of X

θ is not observed so $S_{x\theta}$ is not available to us

$$X_i = \theta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,1)$$

$$Var(X_i) = Var(\theta_i) + Var(\varepsilon_i) = Var(\theta_i) + 1$$

$$\begin{aligned} Cov(X_i, \theta_i) &= Cov(\theta_i + \varepsilon_i, \theta_i) \\ &= Cov(\theta_i, \theta_i) + \cancelto{0}{Cov(\varepsilon_i, \theta)} \\ &= Var(\theta_i) \\ &= Var(X_i) - 1 \end{aligned}$$

$$Cov(x_i, \theta_i) = var(x_i) - 1$$

We can approximate $(x_i - \bar{x})(\theta_i - \bar{\theta})$ using $(x_i - \bar{x})^2 - 1$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\theta_i-\bar{\theta})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \approx \cancel{\frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})^2-1\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$1-\frac{n}{S_x^2}$$

$$= 1-\frac{n-1}{S_x^2}$$

If we knew (θ_i), which we don't, then

$$\hat{\theta}_i = \bar{\theta}+\hat{\beta}(x_i-\bar{x})$$

estimate $\hat{\beta}$ using • and estimate $\bar{\theta}$ by $\bar{x}$,

yields Efron Morris estimator

$$\boxed{\hat{\theta}_i^{(EM)} = \bar{x} + \underbrace{\left(1-\frac{n-1}{S_x^2}\right)}_{<1\ \text{positive (shrinkage factor)}}(x_i-\bar{x})}$$

$$\boxed{\hat{\theta}_i^{(MLE)} = x_i}$$

EM estimator is a shrinkage estimator because $\hat{\theta}_i^{(EM)}$ will lie b/t $\bar{X}$ and X_i

We got a formula similar to Empirical Bayes with no prior & no Bayesian model. We pretended to do a regression of θ on X, $\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X})$, and then did an empirical Bayes style estimation of $\bar{\theta}$ and $\hat{\beta}$.

Contrast with MLE: $\hat{\theta}_i^{(MLE)} = \beta_0 + \beta_1 X_i$, $\beta_0 = 0$

$= 0 + 1 \cdot X_i$ $\quad \beta_1 = 1$

James Stein $\hat{\theta}_i^{(JS)} = 0 + \beta_1 \cdot X_i$

↳ 0 intercept

but estimate slope smartly

$$\boxed{\hat{\theta}_i^{(JS)} = \left(1 - \frac{n}{S_x^2}\right) \cdot X_i}$$

$$\boxed{\hat{\theta}_i^{(EM)} = \bar{X} + \left(1 - \frac{n-1}{S_x^2}\right)(X_i - \bar{X})}$$

Both JS & EM estimators are ~~better~~ better than MLE: they have less Risk

$$R(\theta, \hat{\theta}) = \mathbb{E} L(\theta, \hat{\theta}) = \mathbb{E}\left[\sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2\right].$$

Let $\hat{\theta}^b = b \cdot X_i$ represent the class of linear estimators with zero intercept.

The risk of this estimator is

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

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

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

$\hat{\theta}_i^{LS} = \hat{\beta} X_i$, $\hat{\beta} = \sum \theta_i X_i / \sum X_i^2$

LS: least squares of θ on X

$$= \mathbb{E}\left[\begin{array}{l} \sum (\theta_i - \hat{\theta}_i^{LS})^2 \\ + \cancel{2 \cdot \sum (\theta_i - \hat{\theta}_i^{LS})(\hat{\theta}_i^{LS} - \hat{\theta}^b)} \rightarrow 0 \\ + \sum (\hat{\theta}_i^{LS} - \hat{\theta}_i^b)^2 \end{array} \right] \quad (*)$$

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

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

a James Stein style estimator $\hat{\theta}^{b} = b \cdot X_i$ improves upon the MLE $(b=1)$

if $\mathbb{E}\left[(\hat{\beta}-b)^2 \sum X_i^2\right] < \mathbb{E}\left[(\hat{\beta}-1)^2 \sum X_i^2\right]$

Stein's Paradox

James Stein & Efron morris estimators $\hat{\theta}$ have less Risk $R(\theta, \hat{\theta}) = \mathbb{E}\left[\sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2\right]$ than $\hat{\theta}^{(MLE)}$

no matter the choices of θ! $(\forall \theta)$

assuming $X_i \overset{ind}{\sim} N(\theta_i, 1)$.

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

  this seems preposterous!

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

- How can info about the weight of apples in Washington and about oranges in Florida be used to improve the estimate of the "true" weight of a pear in Cali?

Consultant's Dilemma

In the middle of the season Billy Beane asks you to predict Mookie Betts' 2^{nd} half of season performance.

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

$$L(\theta,\hat{\theta}) = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2$$

minimizes $(\hat{\theta}_i - \theta_i)^2$

Do you optimize aggregated loss or individual loss?
Optimizing for squared error aggregated across all players is not the same as optimizing for the errors of separate estimates of individual parameters.

Which to use for Mookie?