WSABI Summer Research Lab **Summer 2025**

Lecture 13: Shrinkage Estimation

Instructor: Jonathan Pipping *Author: Ryan Brill*

13.1 Problem Setup

13.1.1 Revisiting the Problem

Recall our question from the previous lecture:

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

We reduced this problem to estimating the parameters of the following model:

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

where X_i is the observed mid-season BA for player i, μ_i is the latent "true quality" of player i, σ_i^2 is a known variance which depends on N_i, the number of at-bats for player i, and μ and τ^2 are unknown hyperparameters representing the overall mean and variance of the population of players.

We solved this previously with Empirical Bayes, and will now consider another perspective to understand why this works.

13.1.2 Revising the Model

Since σ_i^2 is known, we can divide by σ_i^2 to remove some parameters:

- $X_i \leftarrow X_i/\sigma_i^2$
- $\theta_i \leftarrow \mu_i/\sigma_i^2$

Now suppose we are taking a frequentist approach, that is, we will think of each parameter as unknown fixed constants rather than random variables (i.e. the unknown "true" quality of each player). Then we are left with the model:

$$X_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1), \quad i = 1, \dots, k$$

and our task is to estimate the fixed unknown constants θ_i (the normal means) given the data $\{X_i\}_{i=1}^k$ to optimize the composite loss function

$$L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{k} \left(\theta_i - \widehat{\theta}_i\right)^2$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ are the true parameters to be estimated, and $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)$ are our estimates.

The performance of the joint estimator $\widehat{\boldsymbol{\theta}}$ is judged by the risk function, or the expected loss:

$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \mathbb{E}L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$$

This simple setup leads to one of the most significant results in mathematical statistics: Stein's Paradox and Shrinkage Estimation.

13.2 Shrinkage Estimation

13.2.1 Starting Point: The MLE

The estimation problem above involves pairs of values X_i, θ_i, $i = 1, \ldots, k$ where one element of each pair is known (X_i) and one (θ_i) is unknown. The "obvious" or ordinary estimator is just $\widehat{\theta}_i = X_i$, which is the **Maximum Likelihood Estimator (MLE)** we're already familiar with! This estimator maximizes the probability of observing the data we did.

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(MLE)} &= \arg\max_{\boldsymbol{\theta}} \mathbb{P}(\text{data} \mid \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \mathbb{P}(X_1, \ldots, X_k \mid \theta_1, \ldots, \theta_k)
\end{aligned}$$

By assumed independence between the X_i and the monotonicity of the log function, we have

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(MLE)} &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{k} \mathbb{P}(X_i \mid \theta_i) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{k} \log \mathbb{P}(X_i \mid \theta_i)
\end{aligned}$$

Then from our model, we have

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(MLE)} &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{k} \log \mathcal{N}(X_i; \theta_i, 1) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{k} \log \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \theta_i)^2 \right) \right] \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{k} -\frac{1}{2}(X_i - \theta_i)^2 \\
&= \boldsymbol{X}
\end{aligned}$$

where $\boldsymbol{X} = (X_1, \ldots, X_k)$.

In terms of our baseball example, this means just predict using each player's mid-season batting average. However, as we saw in the last lecture, these predictions are **terrible**. But here $\widehat{\boldsymbol{\theta}}$ are unknown fixed constants and a prior is mis-specified. Why does it work?

13.2.2 Visualizing the Problem

Recall that this estimation problem involves pairs of values X_i, θ_i, $i = 1, \ldots, k$ where one element of each pair is known (X_i) and one (θ_i) is unknown. Since the θ_i's are unknown, we cannot plot the pairs, but we

can imagine what such a plot would look like to understand the problem. Such a plot is included in Figure 13.1.

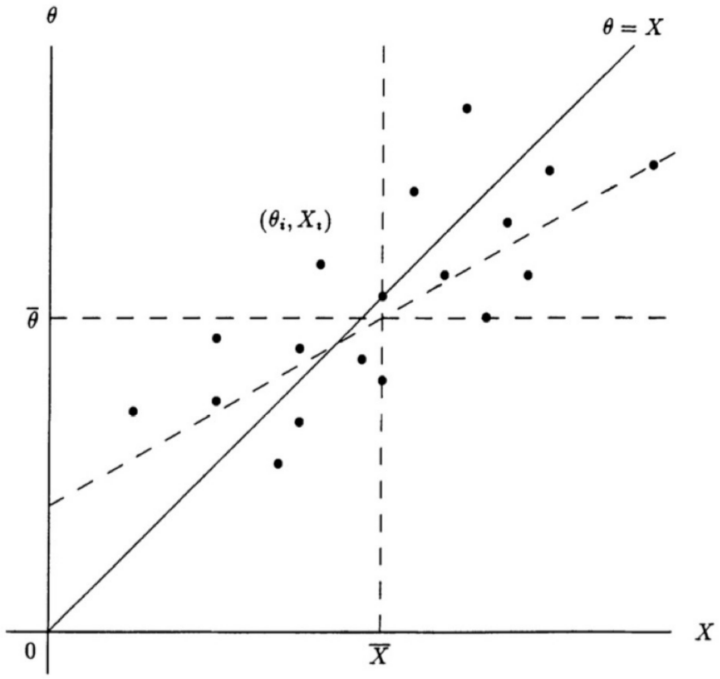

Figure 13.1: A hypothetical bivariate plot of the pairs X_i, θ_i, $i = 1, \ldots, k$.

Since $\boldsymbol{X}$ is $\mathcal{N}(\boldsymbol{\theta}, 1)$, we can think of the X_i as being generated by $\mathcal{N}(0,1)$ "errors" to the given θ_i's. So the horizontal deviates of the X_i's from the 45° line $\theta = X$ are independent $\mathcal{N}(0,1)$ random variables. Our goal is to estimate all 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.

13.2.3 The Galtonian Perspective

Q: Why should we expect that the ordinary estimator $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}$ 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 $\widehat{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}]$ and use this, the theoretical regression function of $\boldsymbol{\theta}$ on $\boldsymbol{X}$, to generate estimates of the θ_i's by evaluating it for each X_i. This is an unattainable ideal though, because we do **not** know the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$.

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 relying purely on frequentist principles. We **do** however know the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{\theta}$, $\mathcal{N}(\boldsymbol{\theta}, 1)$, and we can calculate $\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{\theta}]$. Indeed, this theoretical regression line corresponds to the 45° line $\theta = X$, and this line yields the ordinary estimators $\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$.

So the ordinary estimator may be viewed as being based on the "wrong" regression line, on $\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{\theta}]$ instead of $\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}]$.

As Francis Galton knew in the 1880s, the regressions of $\boldsymbol{X}$ on $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ on $\boldsymbol{X}$ can be markedly different, as

you'll show in the Homework . He suggested that this ordinary estimator could be improved upon, and even suggested a method for doing so: by attempting to approximate $\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}]$ in a setting where the θ_i'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 $\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}]$. Instead we note that $\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$ is a linear function of $\boldsymbol{X}$, and we can look for a best linear estimtor of the form $\widehat{\boldsymbol{\theta}} = a + b\boldsymbol{X}$ so as to minimize the composite loss function

$$\begin{aligned} L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) &= \sum_{i=1}^{k} \left(\theta_i - \widehat{\theta}_i\right)^2 \\ &= \sum_{i=1}^{k} \left(\theta_i - (a + bX_i)\right)^2 \end{aligned}$$

If the θ_i's are known, we would have a standard simple linear regression problem with the best linear estimator given by the regression line

$$\widehat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}} + \widehat{\beta}(\boldsymbol{X} - \bar{\boldsymbol{X}})$$

where

$$\widehat{\beta} = \frac{\sum_{i=1}^{k}(X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum_{i=1}^{k}(X_i - \bar{X})^2}$$

These θ_i's are unknown, but if we could estimate the functions of these unknown parameters ($\bar{\theta}$ and $\widehat{\beta}$), we could estimate the regression line of $\boldsymbol{\theta}$ on $\boldsymbol{X}$. What can we use to estimate these functions?

13.2.4 Deriving the Best Linear Shrinkage Estimator

First, we can use the sample mean $\bar{X}$ to estimate $\bar{\theta}$. Then for the slope $\widehat{\beta}$:

$$\begin{aligned} \widehat{\beta} &= \frac{\sum_{i=1}^{k}(X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum_{i=1}^{k}(X_i - \bar{X})^2} \\ &= \frac{S_{X\theta}}{S_X^2} \end{aligned}$$

where

$$S_X^2 = \sum_{i=1}^{k}(X_i - \bar{X})^2$$

is the sample variance of the X_i's, an unbiased estimator of $\mathrm{Var}(\boldsymbol{X})$ and

$$S_{X\theta} = \sum_{i=1}^{k}(X_i - \bar{X})(\theta_i - \bar{\theta})$$

is the sample covariance of $\boldsymbol{X}$ and $\boldsymbol{\theta}$, an unbiased estimate of $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{\theta})$. Of these two, $S_{X\theta}$ is unknown, so we'll need to estimate it from the data. Since $\boldsymbol{X} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ and $\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{\theta}) + \mathrm{Var}(\boldsymbol{\epsilon}) = \mathrm{Var}(\boldsymbol{\theta}) + 1$, we have that

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{\theta}) &= \mathrm{Cov}(\boldsymbol{\theta} + \boldsymbol{\epsilon}, \boldsymbol{\theta}) \\ &= \mathrm{Var}(\boldsymbol{\theta}) + \mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\theta}) \end{aligned}$$

And since $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\theta}$, $\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\theta}) = 0$ and we have that

$$\text{Cov}(\boldsymbol{X}, \boldsymbol{\theta}) = \text{Var}(\boldsymbol{\theta}) = \text{Var}(\boldsymbol{X}) - 1$$

so $S_X^2 - 1$ and $S_{X\theta}$ have the same expectation, and

$$\begin{aligned}
\widehat{\beta} &= \frac{S_{X\theta}}{S_X^2} \\
&= \frac{\sum_{i=1}^k (X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum_{i=1}^k (X_i - \bar{X})^2} \\
&\approx \frac{\sum_{i=1}^k (X_i - \bar{X})^2 - (k-1)}{\sum_{i=1}^k (X_i - \bar{X})^2} \\
&= 1 - \frac{k-1}{\sum_{i=1}^k (X_i - \bar{X})^2}
\end{aligned}$$

This leads to the **Efron-Morris estimated least-squares line**:

$$\widehat{\boldsymbol{\theta}}^{(EM)} = \bar{\boldsymbol{X}} + \left(1 - \frac{k-1}{S_X^2}\right)(\boldsymbol{X} - \bar{\boldsymbol{X}}), \text{ where } S_X^2 = \sum_{i=1}^k (X_i - \bar{X})^2$$

James-Stein's original shrinkage estimator can be derived by considering the class of estimators that are linear in $\boldsymbol{X}$ with 0 intercept,

$$\widehat{\boldsymbol{\theta}} = b\boldsymbol{X}$$

where b is a constant. The least squares estimator has

$$\widehat{\beta} = \frac{\sum_{i=1}^k \theta_i X_i}{\sum_{i=1}^k X_i^2}$$

and $\theta_i X_i$ has the same expectation as $\sum_{i=1}^k X_i^2 - k$, which yields the **James-Stein shrinkage estimator**:

$$\widehat{\boldsymbol{\theta}}^{(JS)} = \left(1 - \frac{k}{S_X^2}\right)\boldsymbol{X}, \text{ where } S_X^2 = \sum_{i=1}^k (X_i - \bar{X})^2$$

13.2.5 Comparing the Estimators

So the ordinary estimators $\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$ are derived from the theoretical regression line of $\boldsymbol{X}$ on $\boldsymbol{\theta}$, which is useful if our goal is to predict $\boldsymbol{X}$ from $\boldsymbol{\theta}$. But our goal is the opposite, to predict $\boldsymbol{\theta}$ from $\boldsymbol{X}$ with the sum of squares criterion:

$$\sum_{i=1}^k \left(\theta_i - \widehat{\theta}_i\right)^2$$

So the optimal estimator is the least squares regression line of $\boldsymbol{\theta}$ on $\boldsymbol{X}$, and the James-Stein and Efron-Morris estimators are themselves approximations of this correct regression line. It turns out that both of these estimates are better (have less risk) than the ordinary estimate (the MLE), but we need to **show this**.

Let $\widehat{\boldsymbol{\theta}}^{(b)} = b\boldsymbol{X}$ represent the class of linear estimators with zero intercept. The Risk is:

$$\begin{aligned} R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(b)}) &= \mathbb{E}L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(b)}) \\ &= \mathbb{E}\sum_{i=1}^{k}\left(\theta_i - \widehat{\theta}_i\right)^2 \\ &= \mathbb{E}\sum_{i=1}^{k}\left(\theta_i - \widehat{\theta}_i^{(LS)} + \widehat{\theta}_i^{(LS)} - \widehat{\theta}_i^{(b)}\right)^2 \end{aligned}$$

where $\widehat{\theta}_i^{(LS)} = \widehat{\beta}X_i$ and $\widehat{\beta} = \frac{\sum_{i=1}^{k}\theta_i X_i}{\sum_{i=1}^{k} X_i^2}$. Then

$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(b)}) = \mathbb{E}\left[\sum_{i=1}^{k}\left(\theta_i - \widehat{\theta}_i^{(LS)}\right)^2 + 2\sum_{i=1}^{k}\left(\theta_i - \widehat{\theta}_i^{(LS)}\right)\left(\widehat{\theta}_i^{(LS)} - \widehat{\theta}_i^{(b)}\right) + \sum_{i=1}^{k}\left(\widehat{\theta}_i^{(LS)} - \widehat{\theta}_i^{(b)}\right)^2\right]$$

Since the residuals $\theta_i - \widehat{\theta}_i^{(LS)}$ are orthogonal to any linear function of X_i, the second term equals 0 by the orthogonality principle, and

$$\begin{aligned} R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(b)}) &= \mathbb{E}\left[\sum_{i=1}^{k}\left(\theta_i - \widehat{\theta}_i^{(LS)}\right)^2 + \sum_{i=1}^{k}\left(\widehat{\theta}_i^{(LS)} - \widehat{\theta}_i^{(b)}\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{k}\left(\theta_i - \widehat{\theta}_i^{(LS)}\right)^2\right] + \mathbb{E}\left[\sum_{i=1}^{k}\left(\widehat{\theta}_i^{(LS)} - \widehat{\theta}_i^{(b)}\right)^2\right] \\ &= R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(LS)}) + \mathbb{E}\left[\sum_{i=1}^{k}\left(\widehat{\beta}X_i - bX_i\right)^2\right] \\ &= R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(LS)}) + \mathbb{E}\left[\left(\widehat{\beta} - b\right)^2 S_X^2\right] \text{ where } S_X^2 = \sum_{i=1}^{k} X_i^2 \end{aligned}$$

So a James-Stein estimator will improve on the ordinary estimator if and only if

$$\mathbb{E}\left[\left(\widehat{\beta} - b\right)^2 S_X^2\right] < \mathbb{E}\left[\left(\widehat{\beta} - 1\right)^2 S_X^2\right] \text{ for all } \boldsymbol{\theta}$$

Since $\widehat{b}$ is a "reasonable" estimator of β but the constant 1 is not, we can expect the James-Stein estimator to dominate the MLE. This leads us into a landmark result in statistics: Stein's Paradox.

13.3 Stein's Paradox

Theorem 13.1 (Stein's Paradox)**.** *Suppose $\{X_i\}_{i=1}^{k}$ are drawn independently by $X_i \sim \mathcal{N}(\theta_i, 1)$. Then the James-Stein estimator of $\boldsymbol{\theta}$,*

$$\widehat{\boldsymbol{\theta}}^{(JS)} = \left(1 - \frac{C}{S_X^2}\right)\boldsymbol{X}, \text{ where } S_X^2 = \sum_{i=1}^{k}(X_i - \bar{X})^2, k \geq 3, \text{ and } 0 < C \leq 2(k-2)$$

and the Efron-Morris estimator of $\boldsymbol{\theta}$*,*

$$\widehat{\boldsymbol{\theta}}^{(EM)} = \bar{\boldsymbol{X}} + \left(1 - \frac{C}{S_X^2}\right)(\boldsymbol{X} - \bar{\boldsymbol{X}}), \text{ where } S_X^2 = \sum_{i=1}^{k}(X_i - \bar{X})^2, k \geq 4, \text{ and } 0 < C \leq 2(k-3)$$

both uniformly dominate (i.e. have uniformly-lower squared error risk) the "obvious" maximum likelihood estimator of $\boldsymbol{\theta}$*,*

$$\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$$

Concisely,

$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(JS)}) < R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(MLE)}) \ \forall \ \boldsymbol{\theta}$$
$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(EM)}) < R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(MLE)}) \ \forall \ \boldsymbol{\theta}$$

See [SS] for a formal proof.

13.3.1 Shrinkage

Why do we call these estimators "shrinkage" estimators?

- The James-Stein estimator $\widehat{\boldsymbol{\theta}}^{(JS)}$ is the weighted average of 0 and $\boldsymbol{X}$, and so **shrinks** the ordinary estimator $\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$ towards 0.
- The Efron-Morris estimator $\widehat{\boldsymbol{\theta}}^{(EM)}$ is the weighted average of $\bar{X}$ and $\boldsymbol{X}$, and so **shrinks** the ordinary estimator $\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$ towards $\bar{X}$.

These shrinkage estimators dominate the ordinary estimator $\widehat{\boldsymbol{\theta}}^{(MLE)} = \boldsymbol{X}$ as long as $k \geq 3$ for the James-Stein estimator and $k \geq 4$ for the Efron-Morris estimator.

13.3.2 Connection to Empirical Bayes

Recall the Empirical Bayes estimator of $\boldsymbol{\theta}$ is given by

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

This estimator has the same form as the Efron-Morris estimator, shrinking towards the overall mean $\bar{X}$. Here, the prior variance τ^2 (where $\theta_i \sim \mathcal{N}(\theta, \tau^2)$) determines how much the estimator shrinks towards the overall mean.

13.3.3 The Paradox

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

How can information about player A's batting average and player B's batting average help us improve our estimate of player C's batting average? How could information about the price of apples in Washington and the price of oranges in Florida help us improve our estimate of the price of French wine, when it's assumed they're unrelated? Herein lies the paradox.

13.3.4 Consultant's Dilemma

Suppose that in the middle of the season, an MLB general manager asks you to predict the end-of-season batting average of one player on his team, player A, using only that season's available data.

The estimator that is best on average across all players (a shrinkage estimator) is different than 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.

So... which estimator should we use?

References

[SS] Stigler, S. M., *The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators*, Statistical Science, 1990.