James Stein estimator

Aratrika Mustafi, Steven Friedman

Columbia University in the City of New York

am5322@columbia.edu

April 10, 2020

Overview

Motivation

- A baseball player gets 7 hits out of 20 times AB[1]
- Batting Average $= 0.35$
- "Good" prediction suggests number of hits in next 100 times at bat $=$ Batting Average

[1]AB: At Bat- a batter's turn batting against a pitcher

Problem Set-up

- Aim : To predict probability of getting a hit on any given time at bat for each of 18 baseball players (for the 1970 season)
- Conventional Idea: Estimate the probability for each player by each of their individual batting averages.
- Better Idea: Use James-Stein Estimator (will explain)
- Why? On an average works better than using individual averages for predicting the probabilities.
- This is a paradox!!

Problem Set-up

- Taking averages is an easy and familiar way to estimate the probabilities.
- Why particularly an average?
- In most cases distribution of the random variable under study is assumed to be Gaussian. The MLE of the true mean is the sample mean itself.
- Why is the MLE good? Maximizes the probability of the observed data. It is also unbiased. No other unbiased function of the data (linear/nonlinear), can estimate true mean more accurately than the average, in terms of expected squared error.
- Now it makes sense why this is a paradox.

Before proceeding further, it will be helpful to brush up and clarify a few terms and definitions.

Loss, Risk, and MSE

- A **loss function** $L(\theta, \hat{\theta})$ penalizes prediction errors for some parameter θ.
- A common loss function and the one in this setting is **squared error loss**:

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

- (Frequentist) **risk** is expected loss:

$$R(\theta, \hat{\theta}) = E_\theta[L(\theta, \hat{\theta})]$$

- An estimator $\hat{\theta}$ is **inadmissible** if there exists another estimator θ^* such that

$$R(\theta, \theta^*) \leq R(\theta, \hat{\theta}) \text{ for all } \theta$$

  with strict inequality holding for atleast one θ
- Under squared error loss, we have the following expectation, known as **Mean Squared Error**.

$$R(\theta, \hat{\theta}) = E_\theta[(\theta - \hat{\theta})^2]$$

Bias-Variance Decomposition

- Recall bias and variance

$$Bias(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$$

$$Var(\hat{\theta}) = E_\theta[\hat{\theta} - E_\theta(\hat{\theta})]^2$$

- Mean Squared Error can be decomposed as follows. Suppose we have a model $y = f(x) + \epsilon$ with random component ϵ and functional component f we wish to model. For a given unobserved case (x_0, y_0) and corresponding prediction $\hat{y}_0 = \hat{f}(x_0)$:

$$MSE(x_0) = E[(y_0 - \hat{y}_0)^2] = Var(\hat{y}_0) + Bias^2(\hat{y}_0) + Var(\epsilon_0)$$

Bias-Variance Tradeoff

$$MSE(x_0) = E[(y_0 - \hat{y}_0)^2] = Var(\hat{y}_0) + Bias^2(\hat{y}_0) + Var(\epsilon_0)$$

Note that all terms in the decomposition are positive and that model error $Var(\epsilon_0)$ is irreducible. For variance to decrease, bias must increase. This is the **bias-variance tradeoff**.

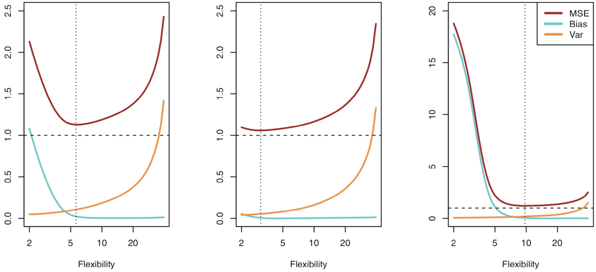

Figure: Figure 12.2, p.36. James, et al. An Introduction to Statistical Learning.

James-Stein estimator

- Aim: Estimate single parameter μ from observation x in the Bayesian situation

$$\mu \sim N(M, A) \text{ and } x|\mu \sim N(\mu, 1) \tag{1}$$

- Then μ has posterior distribution

$$\mu|x \sim N(M + B(x - M), B) \text{ where } B = A/(A + 1) \tag{2}$$

- Bayes estimator of μ

$$\hat{\mu}^{Bayes} = M + B(x - M) \text{ with expected square loss } B \tag{3}$$

- MLE of μ

$$\hat{\mu}^{MLE} = x \text{ with expected square loss } 1 \tag{4}$$

James-Stein estimator

- Same calculation applies to situation where we have N independent versions of (1)

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)' \text{ and } \mathbf{x} = (x_1, x_2, \ldots, x_n)' \tag{5}$$

$$\text{with } \mu_i \sim N(M, A) \text{ and } x_i|\mu_i \sim N(\mu_i, 1) \tag{6}$$

individually for $i = 1, 2, \ldots, N$

- Vector of individual Bayes estimates

$$\hat{\mu}^{\mathbf{Bayes}} = (\hat{\mu_1}^{Bayes}, \hat{\mu_2}^{Bayes}, \ldots, \hat{\mu_n}^{Bayes})' \tag{7}$$

$$= \mathbf{M} + B(\mathbf{x} - \mathbf{M}) \tag{8}$$

where $\hat{\mu_i}^{Bayes} = M + B(x_i - M)$ and $\mathbf{M} = (M, M, \ldots, M)'$ with total squared error risk $N.B$

- $\hat{\mu}^{\mathbf{MLE}} = \mathbf{x}$ with total squared error risk N

James-Stein estimator

- If M and A (or M and B) is known all this is fine.
- If not, we estimate them from **x**. Marginally, (6) gives

$$x_i \overset{ind}{\sim} N(M, A+1) \tag{9}$$

- Then $\hat{M} = \bar{x}$ is an unbiased estimate of M. Moreover, for $N > 3$,

$$\hat{B} = 1 - \frac{N-3}{S} \text{ where } S = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{10}$$

unbiasedly estimates B.

James-Stein estimator

- The James-Stein estimator is the plugged-in version of (3)

$$\hat{\mu}_i^{JS} = \hat{M} + \hat{B}(x_i - \hat{M}) \text{ for i=1,2,...,N} \tag{11}$$

or equivalently $\hat{\mu}^{\mathbf{JS}} = \hat{\mathbf{M}} + \hat{B}(\mathbf{x} - \hat{\mathbf{M}})$ where $\hat{\mathbf{M}} = (\hat{M}, \hat{M}, \ldots, \hat{M})'$

- Expected squared risk is $N.B + 3(1 - B)$

Connection with Empirical Bayes

- Bayesian model (6) leads to the Bayes estimator (8), which itself is estimated empirically (i.e., frequentistically) from all the data $\mathbf{x}$, and then applied to the individual cases. Of course $\hat{\mu}^{\mathbf{JS}}$ cannot perform as well as the actual Bayes rule $\hat{\mu}^{\mathbf{Bayes}}$, but the increased risk is surprisingly modest.
- There is an empirical Bayes interpretation of the James-Stein estimator, where we place a prior $\mu \sim N\left(0, \tau^2 I\right)$ on the underlying mean, and estimate τ from the observed data X. Some people say that this perspective is misleading, since the prior encodes some similarity in the mean components (they share the same marginal variance) but the original paradox holds in a frequentist setting where the means are fixed and completely unrelated.

James-Stein estimator

The estimator has the tendency to shrink the estimates towards the observed sample mean $\hat{M}$ since $\hat{B}$ is less than 1 and acts as a shrinkage factor on the individual estimates x_i.

James-Stein Theorem

Suppose that

$$x_i|\mu_i \sim N(\mu_i, 1) \tag{12}$$

independently for i$=1, 2, \ldots, N$ for $N \geq 4. Then$

$$E\|\hat{\mu}^{\mathbf{JS}} - \mu\|^2 < N = E\|\hat{\mu}^{\mathbf{MLE}} - \mu\|^2 \tag{13}$$

for all choices of $\mu \in \mathbb{R}^N$

Proof: http://www.stat.cmu.edu/ larry/=sml/stein.pdf

Implications of James-Stein Theorem

- From decision theoretic perspective, $\hat{\mu}^{\textbf{MLE}}$ is inadmissible.
- High dimensional situations (often arising in modern practice) requires shrinkage estimators

Why not Bayes Estimator?

Bayes estimator requires the knowledge of both M and A (or equivalently M and B).

Contrast to Gauss-Markov theorem

- Gauss-Markov theorem states that the Ordinary Least squares has the lowest sampling variance within the class of all linearly unbiased estimators
- If the condition of unbiasedness is dropped, the James-Stein theorem shows that there exists estimators with lower overall MSE than those given by the Gauss Markov theorem.

Simulation

1. Compare MLE, Bayes Estimator, and JS Estimator in one trial

Setup:
Model: $x_i \sim N(\mu_i, 1), \mu_i \sim N(M, A)$

```
library(tidyverse)

# params
M = 5  # mean of mu prior
A = 3  # variance of mu prior
N = 50

# generate data
set.seed(15)
mu = rnorm(mean=M, sd=sqrt(A), n=N)
x  = sapply(mu, function(mean) rnorm(mean=mean, sd=1, n=1))
```

1. Compare MLE, Bayes Estimator, and JS Estimator in one trial

Compute risk (expected total square error), estimators, and realized total square error

```
# expected total square error, i.e. risk
B = A/(A+1)
risk.mle   = N
risk.bayes = N * B
risk.js    = N * B + 3 * (1 - B)

# MLE of mu's
mu.hat.mle   = x
sse.mle = sum( (mu.hat.mle - mu)^2 )
```

1. Compare MLE, Bayes Estimator, and JS Estimator in one trial

```
# Bayes estimator
# can only do because we know M, A
mu.hat.bayes = M + B * (x - M)
sse.bayes    = sum( (mu.hat.bayes - mu)^2 )

# JS estimator of mu's
M.hat = mean(x) # unbiased estimator of M
S     = sum( (x - mean(x))^2 )
B.hat = 1 - (N - 3) / S

mu.hat.js = M.hat + B.hat * (x - M.hat)
sse.js    = sum( (mu.hat.js - mu)^2 )
```

1. Compare MLE, Bayes Estimator, and JS Estimator in one trial

Compare errors

```
summary = data.frame(
  "Estimator"     = c("MLE",    "Bayes",    "JS"),
  "Expected Total Sq Err (Risk)" = c(risk.mle, risk.bayes, ris
  "Realized Total Sq Err" = c(sse.mle,  sse.bayes,  sse.js)
)
knitr::kable(summary)
```

Estimator	Expected.Total.Sq.Err..Risk.	Realized.Total.Sq.Err
MLE	50.00	54.89232
Bayes	37.50	34.39785
JS	38.25	34.07457

2. Observe General Ability of Shrinkage to Improve MLE

James-Stein shrinks the MLE value by a factor of $\hat{B}$ toward the grand mean $\hat{M}$. Let's try shrinking to an arbitrary constant, 0, by a small amount and compare error to MLE.

```
arb.shrinkage.target = 0
arb.shrinkage.factor = 0.99

mu.hat.arb  = arb.shrinkage.target + arb.shrinkage.factor * (x
sse.arb     = sum( (mu.hat.arb - mu)^2 )
sse.arb

## [1] 54.01003
```

The total square error of the MLE is 54.8923201, while the total square error of the estimator with small shrinkage in an arbitrary direction is 54.0100269.

3. Visualize Results of Parts 1 and 2

Observe that the JS Estimator reduces error generally but increases error for some individual cases.

```
mu.df = data.frame(
  "Index"           = 1:N,
  "Truth"           = mu,
  "JS"              = mu.hat.js,
  "Arb.Shrinkage"   = mu.hat.arb,
  "MLE"             = mu.hat.mle
)

# Plot ideas borrowed from https://bookdown.org/content/922/j
```

3. Visualize Results of Parts 1 and 2

```r
# Plot MLE, JS, and Truth
plot1 = mu.df %>%
gather(type, value, c(2,3,5)) %>%
mutate(type = factor(type, levels = c("Truth","JS","MLE"))) %>%
arrange(Index, type) %>%
ggplot(aes(x=value, y=type)) +
geom_point(color="black") +
geom_path(aes(group=Index),lty=2,color="grey") +
ggtitle("MLE vs. JS") +
xlab("Estimated/True Params") +
ylab("Estimator/Truth") +
theme_light() +
theme(plot.title = element_text(hjust = 0.5))
```

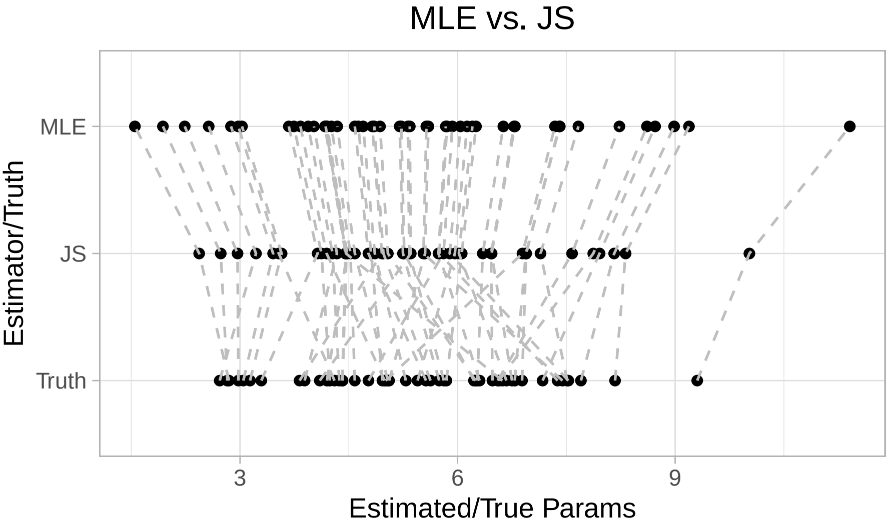

Figure: Plot of MLE, JS, and Truth

3. Visualize Results of Parts 1 and 2

```
# Plot MLE, Small Shrinkage to 0, and Truth
plot2 = mu.df %>%
gather(type, value, c(2,4,5)) %>%
mutate(type=factor(type,levels=c("Truth","Arb.Shrinkage","MLE")))%>%
arrange(Index, type) %>%
ggplot(aes(x=value, y=type)) +
geom_point(color="black") +
geom_path(aes(group=Index),lty=2,color="grey") +
ggtitle("MLE vs. Small Downward Shrinkage") +
xlab("Estimated/True Params") +
ylab("Estimator/Truth") +
theme_light() +
theme(plot.title = element_text(hjust = 0.5))
```

3. Visualize Results of Parts 1 and 2

```
set.seed(15)
runs = 1000
SSEs = as.data.frame(matrix(nrow=1000, ncol=2))
colnames(SSEs) = c("MLE", "JS")
for(i in 1:runs){
  # generate mu and sample x
  mu = rnorm(mean=M, sd=sqrt(A), n=N)
  x  = sapply(mu, function(mean) rnorm(mean=mean, sd=1, n=1))

  # MLE of mu's
  mu.hat.mle   = x
  sse.mle = sum( (mu.hat.mle - mu)^2 )

  # JS estimator of mu's
  M.hat = mean(x) # unbiased estimator of M
  S     = sum( (x - mean(x))^2 )
  B.hat = 1 - (N - 3) / S

  mu.hat.js = M.hat + B.hat * (x - M.hat)
  sse.js    = sum( (mu.hat.js - mu)^2 )

  # store
  SSEs[i,"MLE"] = sse.mle
  SSEs[i,"JS"] = sse.js
}
```

3. Visualize Results of Parts 1 and 2

```r
# plot
plot3 = SSEs %>%
        gather(type, value) %>%
        mutate(type=factor(type,levels=c("MLE","JS")))%>%
        ggplot(aes(x=value, color=type, fill=type)) +
        geom_histogram(position="dodge") +
        ggtitle("plot of MLE and JS SSEs Over 1000 Trials") +
        xlab("Total Square Error") +
        ylab("Count") +
        labs(color = "Estimator", fill="Estimator") +
        theme_light() +
        theme(plot.title = element_text(hjust = 0.5))
```

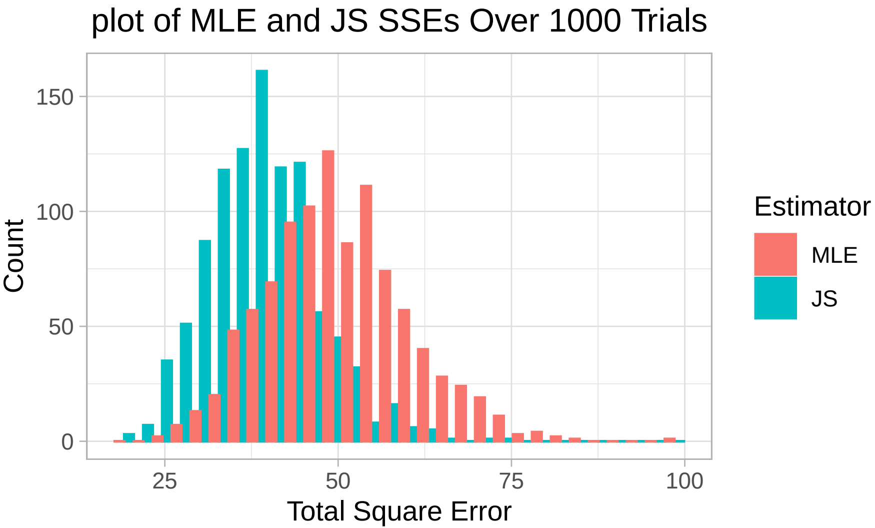

Figure: Plot of MLE and JS SSEs Over 1000 Trials

Takeaways

- "Shrinkage estimation, as exemplified by the James–Stein rule, has become a necessity in the high-dimensional situations of modern practice." (CASI, p. 94)
- "Shrinkage estimation tends to produce better results *in general*, at the possible expense of extreme cases." (CASI, p. 103)
- The James Stein estimator is often used to demonstrate the inadequacy of MLE but is a good estimator in its own right - Have Bayesian properties and also dominate the MLE, rendering it inadmissible. (Ref. 3)

References

- Efron, B., and Morris, C. (1977), "Stein's Paradox in Statistics," Scientific American, Springer Science and Business Media LLC, 236, 119–127. https://doi.org/10.1038/scientificamerican0577-119.
- Ijiri, Y., and Leitch, R. A. (1980), "Stein's Paradox and Audit Sampling," Journal of Accounting Research, JSTOR, 18, 91. https://doi.org/10.2307/2490394.
- Efron, B., and Morris, C. (1973), "Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach," Journal of the American Statistical Association, JSTOR, 68, 117. https://doi.org/10.2307/2284155.
- http://www.stat.cmu.edu/ larry/=sml/stein.pdf
- http://statweb.stanford.edu/ ckirby/brad/LSI/chapter1.pdf

Thank you

Questions?
Slides available at- https://github.com/Aratrika-cs/James-Stein