STAT 215B Assignment 5

Xiaowei Zeng

March 14, 2024

Abstract

The main job in this assignment is to gain deep insights of James-Stein estimation and empirical Bayes. The first section gives the solutions for some mathematical proofs in empirical Bayes. The second section repeats the simulation in (Efron, 2012) and discusses the agreement between my results and Efron's. The third section applies empirical Bayes into the radon level real-world problem and explores the performance of MLE and JS estimator.

1 Math Stats

The following exercises are from (Efron, 2012).

1.1 Exercise 1.1

Suppose μ has a normal prior distribution with mean 0 and variance A, while z given μ is normal with mean μ and variance 1.

$$\mu \sim N(0, A) \quad \text{and} \quad z \mid \mu \sim N(\mu, 1).$$

Show that

$$\mu \mid z \sim N(Bz, B) \quad \text{where } B = A/(A+1).$$

Proof:

$$g(\mu) \propto \exp\left\{-\frac{1}{2A}\mu^2\right\} \quad \text{and} \quad f_\mu(z) \propto \exp\left\{-\frac{1}{2}(z-\mu)^2\right\}$$

By Bayes rule,

$$\begin{aligned} g(\mu \mid z) = \frac{g(\mu) f_\mu(z)}{f(z)} &\propto g(\mu) f_\mu(z) \\ &\propto \exp\left\{-\frac{1}{2A}\mu^2 - \frac{1}{2}(z-\mu)^2\right\} \\ &\propto \exp\left\{-\frac{1+A}{2A}(\mu - \frac{A}{A+1}z)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2B}(\mu - Bz)^2\right\}, \end{aligned}$$

where $B = A/(A+1)$, and $\exp\{-(\mu - Bz)^2/2B\}$ is the kernel of $N(Bz, B)$.

Thus, $\mu \mid z \sim N(Bz, B)$.

1.2 Exercise 1.2

Suppose we are dealing with large scale inference,

$$\boldsymbol{\mu} \sim N(0, A\boldsymbol{I}_N) \quad \text{and} \quad \boldsymbol{z} \mid \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_N),$$

where $\boldsymbol{\mu}, \boldsymbol{z}$ are $N \times 1$ vectors, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. Then Bayes rule gives us the posterior distribution

$$\boldsymbol{\mu} \mid \boldsymbol{z} \sim N(B\boldsymbol{z}, B\boldsymbol{I}_N) \quad \text{where } B = A/(A+1).$$

Verify that the Bayes estimator $\hat{\mu}^{Bayes}$ has risk

$$R^{(\text{Bayes})}(\boldsymbol{\mu}) = (1-B)^2||\boldsymbol{\mu}||^2 + NB^2,$$

and overall Bayes risk

$$R_A^{(\text{Bayes})} = \mathbb{E}_A\left[R^{(\text{Bayes})}(\boldsymbol{\mu})\right] = N\frac{A}{A+1}.$$

Proof:

$$\begin{aligned}
R^{(\text{Bayes})}(\boldsymbol{\mu}) &= \mathbb{E}_{\boldsymbol{\mu}}\left[L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}^{Bayes})\right] \\
&= \mathbb{E}_{\boldsymbol{\mu}}\left[||B\boldsymbol{z} - \boldsymbol{\mu}||^2\right] \\
&= \mathbb{E}_{\boldsymbol{\mu}}\left[(B\boldsymbol{z} - \boldsymbol{\mu})^T(B\boldsymbol{z} - \boldsymbol{\mu})\right] \\
&= \mathbb{E}_{\boldsymbol{\mu}}\left[B^2||\boldsymbol{z}||^2 - B\boldsymbol{z}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T B\boldsymbol{z} + ||\boldsymbol{\mu}||^2\right] \\
&= \mathbb{B}^2 E_{\boldsymbol{\mu}}\left[||\boldsymbol{z}||^2\right] - B\mathbb{E}_{\boldsymbol{\mu}}\left[\boldsymbol{z}^T\right]\boldsymbol{\mu} - B\boldsymbol{\mu}^T\mathbb{E}_{\boldsymbol{\mu}}\left[\boldsymbol{z}\right] + ||\boldsymbol{\mu}||^2 \\
&= B^2(\boldsymbol{\mu}^2 + N) - 2B||\boldsymbol{\mu}||^2 + ||\boldsymbol{\mu}||^2 \\
&= (1-B)^2||\boldsymbol{\mu}||^2 + NB^2 \\
R_A^{(\text{Bayes})} &= \mathbb{E}_A\left[R^{(\text{Bayes})}(\boldsymbol{\mu})\right] \\
&= \mathbb{E}_A\left[(1-B)^2||\boldsymbol{\mu}||^2 + NB^2\right] \\
&= (1-B)^2\mathbb{E}_A\left[||\boldsymbol{\mu}||^2\right] + NB^2 \\
&= (1-B)^2(0 + AN) + NB^2 \\
&= N\frac{A}{(A+1)^2} + N\frac{A^2}{(A+1)^2} \\
&= N\frac{A}{A+1}
\end{aligned}$$

1.3 Exercise 1.4

The expected total squared error (TSE) of $\hat{\boldsymbol{\mu}}$ is

$$\mathbb{E}_{\boldsymbol{\mu}}\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2\right] = \mathbb{E}_{\boldsymbol{\mu}}\left[||\boldsymbol{z} - \hat{\boldsymbol{\mu}}||^2\right] - N + 2\sum_{i=1}^{N}\text{cov}_{\boldsymbol{\mu}}(\hat{\mu}_i, z_i), \tag{1}$$

where $\text{cov}_{\boldsymbol{\mu}}$ indicates covariance under $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_N)$ Integration by parts involving the multivariate normal density function $f_{\boldsymbol{\mu}}(\boldsymbol{z}) = (2\pi)^{-N/2}\exp\{-\sum_i(z_i - \mu_i)^2/2\}$ shows that

$$\text{cov}_{\boldsymbol{\mu}}(\hat{\mu}_i, z_i) = E_{\boldsymbol{\mu}}\left[\frac{\partial\hat{\mu}_i}{\partial z_i}\right],$$

as long as μ_i is continuously differentiable in $\boldsymbol{z}$. This reduces (1) to

$$\mathbb{E}_{\boldsymbol{\mu}}\left[||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||^2\right]=\mathbb{E}_{\boldsymbol{\mu}}\left[||\boldsymbol{z}-\hat{\boldsymbol{\mu}}||^2\right]-N+2\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{\partial\hat{\mu}_i}{\partial z_i}\right]. \tag{2}$$

The James–Stein estimator is defined as

$$\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}=\left(1-\frac{N-2}{S}\right)\boldsymbol{z}, \tag{3}$$

where S is the sum of squares $||\boldsymbol{z}||^2$. Apply (2) in (3) to show that

$$\mathbb{E}_{\boldsymbol{\mu}}\left[||\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}-\boldsymbol{\mu}||^2\right]=N-\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{(N-2)^2}{S}\right], \tag{4}$$

and then use (4) to show that the overall Bayes risk of $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$ is

$$R_A^{(\mathrm{JS})}=N\frac{A}{A+1}+\frac{2}{A+1}. \tag{5}$$

Proof:

$$\begin{aligned}\hat{\mu}_i^{(\mathrm{JS})}&=\left(1-\frac{N-2}{\sum_i z_i^2}\right)z_i\\ \frac{\partial\hat{\mu}_i^{(\mathrm{JS})}}{\partial z_i}&=1-\frac{N-2}{\sum_i z_i^2}+\frac{N-2}{(\sum_i z_i^2)^2}\cdot 2z_i\cdot z_i\\ &=1-\frac{N-2}{S}+\frac{2(N-2)z_i^2}{S^2}\end{aligned}$$

Applying (2), we have

$$\begin{aligned}R^{(\mathrm{JS})}(\boldsymbol{\mu})&=\mathbb{E}_{\boldsymbol{\mu}}\left[||\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}-\boldsymbol{\mu}||^2\right]\\ &=\mathbb{E}_{\boldsymbol{\mu}}\left[\left|\left|\boldsymbol{z}-\left(1-\frac{N-2}{S}\right)\boldsymbol{z}\right|\right|^2\right]-N+2\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\mu}}\left[1-\frac{N-2}{S}+\frac{2(N-2)z_i^2}{S^2}\right]\\ &=\mathbb{E}_{\boldsymbol{\mu}}\left[\left(\frac{N-2}{S}\right)^2S\right]+N-2N\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{N-2}{S}\right]+4\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{N-2}{S^2}S\right]\\ &=\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{(N-2)^2-2N(N-2)+4(N-2)}{S}\right]+N\\ &=N-\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{(N-2)^2}{S}\right]\end{aligned}$$

The marginal distribution of $\boldsymbol{z}$ is

$$\boldsymbol{z}\sim N(0,(A+1)\boldsymbol{I}_N),$$

and thus S has a scaled chi-square distribution with N degrees of freedom,

$$S\sim(A+1)\mathcal{X}_N^2,$$

so that

$$\mathbb{E}\left[\frac{N-2}{S}\right] = \frac{1}{A+1}.$$

Therefore,

$$\begin{aligned}
R_A^{(\mathrm{JS})} &= \mathbb{E}_A\left[R^{(\mathrm{JS})}(\boldsymbol{\mu})\right] \\
&= \mathbb{E}_A\left[N - (N-2)\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{(N-2)}{S}\right]\right] \\
&= N - (N-2)\frac{1}{A+1} \\
&= N\frac{A}{A+1} + \frac{2}{A+1}.
\end{aligned}$$

1.4 Exercise 1.5

If we assume that the μ_i values in Table 1.2 below were obtained from $\mu_i \overset{\text{ind}}{\sim} N(0, A)$, is the total error 8.13 about right?

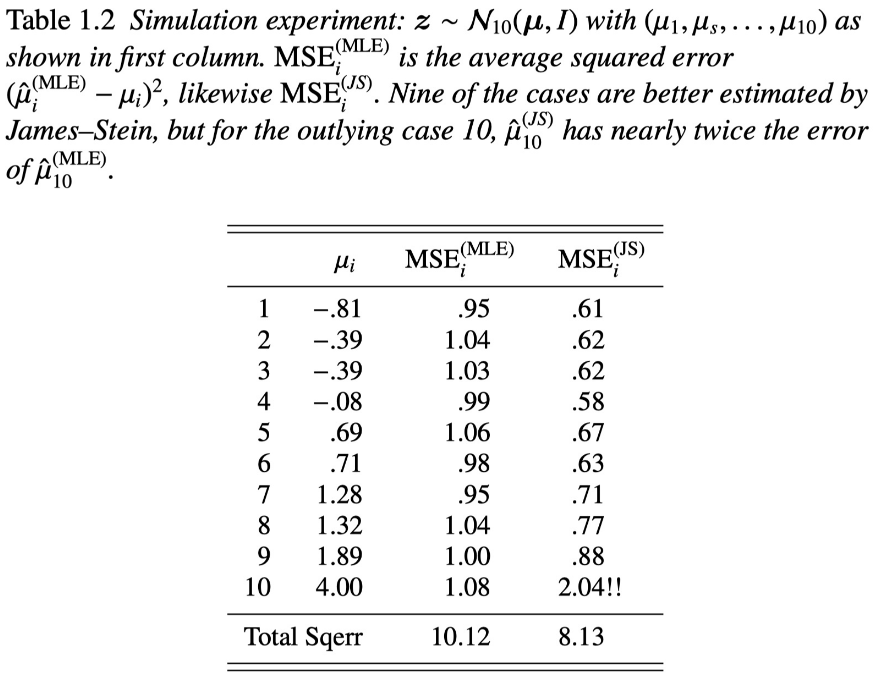

Table 1.2 *Simulation experiment:* $\boldsymbol{z} \sim \mathcal{N}_{10}(\boldsymbol{\mu}, I)$ *with* $(\mu_1, \mu_s, \ldots, \mu_{10})$ *as shown in first column.* $\mathrm{MSE}_i^{(\mathrm{MLE})}$ *is the average squared error* $(\hat{\mu}_i^{(\mathrm{MLE})} - \mu_i)^2$*, likewise* $\mathrm{MSE}_i^{(JS)}$*. Nine of the cases are better estimated by James–Stein, but for the outlying case 10,* $\hat{\mu}_{10}^{(JS)}$ *has nearly twice the error of* $\hat{\mu}_{10}^{(\mathrm{MLE})}$.

	μ_i	$\mathrm{MSE}_i^{(\mathrm{MLE})}$	$\mathrm{MSE}_i^{(\mathrm{JS})}$
1	−.81	.95	.61
2	−.39	1.04	.62
3	−.39	1.03	.62
4	−.08	.99	.58
5	.69	1.06	.67
6	.71	.98	.63
7	1.28	.95	.71
8	1.32	1.04	.77
9	1.89	1.00	.88
10	4.00	1.08	2.04!!
Total Sqerr		10.12	8.13

Solution:

In this scenario, $N = 10$. Rewrite the expectation of TSE into

$$\begin{aligned}
\mathbb{E}\left[\mathrm{TSE}\right] &= \mathbb{E}_A\left[\mathbb{E}_{\boldsymbol{\mu}}\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2\right]\right] \\
&= \mathbb{E}_A\left[\mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{i=1}^{10}(\hat{\mu}_i - \mu_i)^2\right]\right] \\
&= \sum_{i=1}^{10}\mathbb{E}\left[(\hat{\mu}_i - \mu_i)^2\right].
\end{aligned} \tag{6}$$

We can expect the summation of $\mathrm{MSE}_i^{(\mathrm{exp})}$ obtained in the simulation experiment to approach $\mathbb{E}\left[\mathrm{TSE}\right]$ because by Law of Large Number (LLN),

$$\mathrm{MSE}_i^{(\mathrm{exp})} = \frac{1}{1000}\sum_{b=1}^{1000}(\hat{\mu}_i^{(b)} - \mu_i)^2 \to \mathbb{E}\left[(\hat{\mu}_i - \mu_i)^2\right], i = 1, ..., 10,$$

and thus by (6), we have

$$\sum_{i=1}^{10}\left[\frac{1}{1000}\sum_{b=1}^{1000}(\hat{\mu}_i^{(b)} - \mu_i)^2\right] \to \mathbb{E}\left[\mathrm{TSE}\right], i = 1, ..., 10.$$

By (5), we have

$$\mathbb{E}\left[\mathrm{TSE}^{(\mathrm{JS})}\right] = \frac{10A + 2}{A + 1}$$

By plugging in the moment estimator of A

$$\hat{A} = \frac{1}{9}\sum_{i=1}^{9}(\mu_i - \bar{\mu})^2 \approx 2.016,$$

we can estimate TSE as

$$\widehat{\mathrm{TSE}} \approx \frac{10 \times 2.016 + 2}{2.016 + 1} \approx 7.347.$$

Therefore, if we assume that the μ_i's were obtained from $\mu_i \overset{\mathrm{ind}}{\sim} N(0, A)$, $\mathrm{TSE}^{(\mathrm{exp})}$ should be close to 7.347. The experiment result in Table 1.2 gives $\mathrm{TSE}^{(\mathrm{exp})} = 8.13$, 0.78 larger than $\widehat{\mathrm{TSE}}$. However, we cannot determine whether 0.78 is a substantial difference or not.

We need to conduct a simulation experiment with

$$A' = \hat{A} = 2.016,$$

$$\boldsymbol{\mu} \sim N(0, A'\boldsymbol{I}_N) \text{ and } \boldsymbol{z} \mid \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_N),$$

and obtain a distribution for $\mathrm{TSE}^{(\mathrm{exp})}$ to see whether the possibility of $\mathrm{TSE}^{(\mathrm{exp})} = 8.13$ is statistically significantly small, say, smaller than 0.05.

The p-value for $\mathrm{TSE}^{(\mathrm{exp})} = 8.13$ is $0.19 > 0.05$, indicating that it is about right.

2 Simulation

I produce my own version of Table 1.2 in (Efron, 2012) by repeating the simulation study described on Page 7-9 and using the same μ_i's as Efron. To analyze how many decimal places of agreement one would expect to see between my results and Efron's, I use both mathematical and simulation techniques for MLE (Section 2.1), and use only simulation technique for the JS estimator (Section 2.2). My simulation results and how well it aligns with this expectation of agreement are shown in Section 2.3.

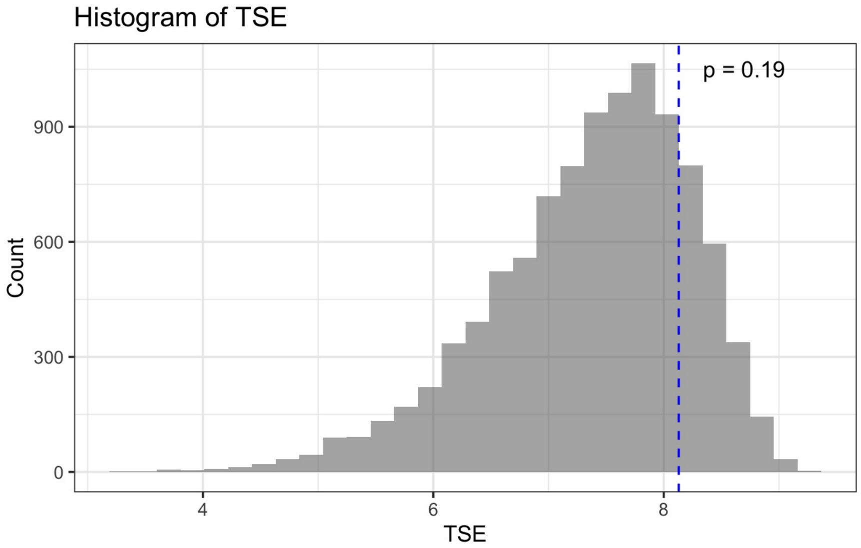

Figure 1: Histogram of $\text{TSE}^{(\text{exp})}$ in 10000 Simulations

2.1 MLE

The MLE for μ_i is z_i. We have

$$z_i - \mu_i \mid \mu_i \sim N(0,1) \text{ and } (z_i - \mu_i)^2 \mid \mu_i \sim \mathcal{X}_1^2,$$

$$\sum_{b=1}^{B} (z_i^{(b)} - \mu_i)^2 \mid \mu_i \sim \mathcal{X}_B^2,$$

where B is the simulation times 1000. The $\text{MSE}_i^{(\text{MLE})}$ in Table 1.2 is defined as

$$\frac{1}{B}\sum_{b=1}^{B} (z_i^{(b)} - \mu_i)^2 := \text{MSE}_i^{(\text{MLE})}.$$

Since the sample size $N = 1000$ is really large, by Central Limit Theorem (CLT), $\text{MSE}_i^{(\text{MLE})}$ is approximately normally distributed. We can expect my results and Efron's results are within the range ± 1.96 (97.5% quantile of standard normal) times standard deviation, centered at the expectation.

$$\mathbb{E}_\mu\left[\text{MSE}_i^{(\text{MLE})}\right] = \frac{1}{B}\mathbb{E}_\mu\left[(z_i - \mu_i)^2\right] = B/B = 1,$$

$$\text{SD}_\mu\left[\text{MSE}_i^{(\text{MLE})}\right] = \frac{1}{B}\sqrt{\text{Var}_\mu\left[(z_i - \mu_i)^2\right]} = \frac{1}{B}\cdot\sqrt{2B} = \sqrt{\frac{2}{B}} = \frac{1}{10\sqrt{5}} \approx 0.0447,$$

I also conduct 1,000 simulations, obtain 1,000 × 10 $\text{MSE}_i^{(\text{MLE})}$'s and estimate the standard deviation of $\text{MSE}_i^{(\text{MLE})}$ by pooling all of them together because they have the same variance and are independent of one another. The simulation also gives 0.0447 (only different with $1/10\sqrt{5}$ at the sixth decimal place), verifying our mathematical derivation.

The length of the 95% MSE interval is

$$2 \times 1.96 \times \text{SD}_\mu\left[\text{MSE}_i^{(\text{MLE})}\right] \approx 0.1753,$$

and thus we can expect the difference between $\text{MSE}_i^{(\text{MLE})}$'s is within 0.1753.

For $\text{TSE}^{(\text{MLE})}$, the estimated standard deviation in the simulation is 0.1395. So the length of the 95% TSE interval is

$$2 \times 1.96 \times \hat{\text{SD}}_\mu \left[\text{TSE}^{(\text{MLE})}\right] \approx 0.5467,$$

and thus we can expect the difference between $\text{TSE}^{(\text{MLE})}$'s is within 0.5467.

2.2 J-S Estimator

The J-S estimator for μ_i is $(1 - \frac{N-2}{S})z_i$. However, unlike $\hat{\mu}^{(\text{MLE})}$, we cannot derive the distribution for $(\hat{\mu}_i^{(\text{J-S})} - \mu_i)$ or $(\hat{\mu}_i^{(\text{J-S})} - \mu_i)^2$, so I have to use simulation to approximate standard deviation for $(\hat{\mu}_i^{(\text{J-S})} - \mu_i)^2$ and employ the same trick in Section 2.1 to obtain the expected agreement for J-S estimator.

I conduct 1,000 simulations, obtain 1,000 × 10 $\text{MSE}_i^{(\text{JS})}$'s, but estimate the standard deviation of $\text{MSE}_i^{(\text{JS})}$ respectively for $i = 1, ..., 10$ (Table 1). The expected difference column is the length of 95% MSE interval,

$$2 \times 1.96 \times \hat{\text{SD}}_\mu \left[\text{MSE}_i^{(\text{JS})}\right],$$

referring to the expected difference between my $\text{MSE}_i^{(\text{JS})}$'s and Efron's for each i.

Table 1: Simulation for Standard Deviation of $\text{MSE}_i^{(\text{JS})}$

i	μ_i	$\hat{\text{SD}}_\mu[\text{MSE}_i^{(\text{JS})}]$	Expected Difference
1	−.81	.0285	.1118
2	−.39	.0286	.1121
3	−.39	.0283	.1110
4	−.08	.0278	.1089
5	.69	.0295	.1158
6	.71	.0285	.1119
7	1.28	.0310	.1214
8	1.32	.0310	.1216
9	1.89	.0359	.1408
10	4.00	.0723	.2833

For $\text{TSE}^{(\text{JS})}$, the estimated standard deviation in the simulation is 0.1102. So the length of the 95% TSE interval is

$$2 \times 1.96 \times \hat{\text{SD}}_\mu \left[\text{TSE}^{(\text{JS})}\right] \approx 0.4321,$$

and thus we can expect the difference between $\text{TSE}^{(\text{JS})}$'s is within 0.4321.

2.3 My Results

My own version of simulation is shown in Table 2. Similar to the conclusion in (Efron, 2012), nine of the cases are better estimated by JS estimator, but for the outlying case 10, $\hat{\mu}_{10}^{(\text{JS})}$ has nearly twice the error of $\hat{\mu}_{10}^{(\text{MLE})}$ (1.99/1.01 ≈ 1.97).

The differences between my results and Efron's results are given in the "(...)", which perfectly meets the expectation analyzed in Section 2.1, 2.2.

Table 2: My Simulation Experiment

i	μ_i	$\mathrm{MSE}_i^{(\mathrm{MLE})}$ (diff)	$\mathrm{MSE}_i^{(\mathrm{JS})}$ (diff)
1	$-.81$	.96 (+.01)	.62 (+.01)
2	$-.39$	1.01 ($-$.03)	.61 ($-$.01)
3	$-.39$	.96 ($-$.07)	.56 ($-$.06)
4	$-.08$	.97 ($-$.02)	.57 ($-$.01)
5	.69	.99 ($-$.07)	.61 ($-$.06)
6	.71	.99 (+.01)	.62 ($-$.01)
7	1.28	1.02 (+.07)	.72 (+.01)
8	1.32	1.06 (+.02)	.78 (+.01)
9	1.89	.93 ($-$.07)	.85 ($-$.03)
10	4.00	1.01 ($-$.07)	1.99 ($-$.05)
Total Sqerr		9.90 ($-$.22)	7.93 ($-$.20)

3 Shrinking Radon

The file `srrs2.dat` contains 12,777 observed radon levels from households throughout the United States. This data file comes from Andrew Gelman's website, `http://www.stat.columbia.edu/~gelman/arm/software/`. We will focus on the 766 measurements taken in the basements of the Minnesota homes. These homes are spread across 85 counties in Minnesota; the data set tells us which observations came from which counties.

3.1 Data Preparing

Load the dataset `srrs2.dat` into `R`. Remove the redundant spaces in numbers and strings. Extract the subset of observations taken in Minnesota basements: (1) keep only the observations with the `state` variable valuing "MN" (the abbreviation of Minnesota State); (2) keep only the observations with the `floor` variable valuing 0 (referring to basement). Next, reduce the dataset further: keep only the data for counties with at least 10 observations. Now the dataset contains 17 such counties, with a total of 511 observations.

Then, split the data into two sets: a training set with 5 randomly chosen observations from each county, and a test set with the other observations.

3.2 Estimation

Compute $\boldsymbol{\mu}$, the vector of mean radon levels by county in the test data. Radon levels are given in the variable `activity`. From now on we will treat $\boldsymbol{\mu}$ as a population-level parameter to be estimated.

Make the standard James-Stein independent-normals assumption: the five observations $z_{i,k}, k = 1, ..., 5$ in county i are i.i.d. draws from a $N(\mu_i, \tau^2)$ distribution; these five draws are independent of the draws from every other county. Compute $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$, the maximum-likelihood estimate of $\boldsymbol{\mu}$ based on the training data, where

$$\hat{\mu}_i^{(\mathrm{MLE})} = \frac{1}{5}\sum_{k=1}^{5} z_{i,k} = \bar{z}_i.$$

We are assuming that the components of $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$, share a common standard error (SE). Using the same number of observations in each county tends to aid this assumption.

$$z_{i,k} \sim N(\mu_i, \tau^2) \quad \rightarrow \quad \hat{\mu}_i^{(\mathrm{MLE})} = \bar{z}_i \sim N(\mu_i, \frac{\tau^2}{5})$$

To estimate this shared SE, we estimate τ^2 using the pooled-variance technique: add up all the within-county squared residuals, and divide by the total degrees of freedom,

$$\hat{\tau}^2 = S_p^2 = \frac{\sum_{i=1}^{17}(5-1)S_i^2}{\sum_{i=1}^{17}(5-1)} = \frac{\sum_{i=1}^{17} S_i^2}{17},$$

where

$$S_i^2 = \frac{1}{5-1}\sum_{k=1}^{5}(z_{i,k} - \bar{z}_i)^2.$$

The SE estimate of $\hat{\mu}_i^{(\mathrm{MLE})}$ should be

$$\hat{\mathrm{se}} = \sqrt{\frac{\hat{\tau}^2}{5}}$$

rather than $\hat{\tau}$, otherwise an over-shrink will occur.

Now compute $\hat{\mu}_i^{(\mathrm{JS})}$, the James-Stein estimator, using the average value $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$ as the shrinkage target.

$$\hat{\mu}_i^{(\mathrm{JS})} = \bar{z} + \left(1 - \frac{(17-3)\hat{\mathrm{se}}^2}{S}\right)(\hat{\mu}_i^{(\mathrm{MLE})} - \bar{z}),$$

where

$$S = \sum_{i=1}^{17}(\hat{\mu}^{(\mathrm{MLE})} - \bar{z})^2,$$

$$\bar{z} = \frac{1}{17}\sum_{i=1}^{17}\hat{\mu}_i^{(\mathrm{MLE})} = \frac{1}{17}\sum_{i=1}^{17}\frac{1}{5}\sum_{k=1}^{5} z_{i,k} = \frac{1}{17\times 5}\sum_{i,k} z_{i,k}.$$

Finally, compute the TSE of $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$ and $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$. All the results are shown in Table 3. The TSE of MLE is 2.17 times larger than the TSE of JS estimator, indicating a tremendous advantage for the Stein shrinkage in this random split of training and test set.

Table 3: Estimation Results

County	μ_i	$\hat{\mu}_i^{(\mathrm{MLE})}$	$\hat{\mu}_i^{(\mathrm{JS})}$
ANOKA	3.22	1.86	4.36
BLUE EARTH	7.30	7.18	6.85
CLAY	7.20	14.92	10.47
DAKOTA	4.93	4.78	5.73
GOODHUE	6.36	12.26	9.22
HENNEPIN	4.83	5.08	5.87
ITASCA	3.23	2.76	4.78
MOWER	8.45	7.82	7.15
OLMSTED	3.91	7.50	7.00
RAMSEY	3.48	3.66	5.20
RICE	5.22	9.56	7.96
ST LOUIS	3.12	6.28	6.43
STEARNS	5.26	6.36	6.47
STEELE	5.02	6.20	6.39
WASHINGTON	5.06	5.68	6.15
WINONA	9.58	5.74	6.18
WRIGHT	7.70	3.84	5.29
Total Sqerr		171.44	79.14

3.3 Simulation

However, the effectiveness of the estimators may greatly depend on the split of training and test set. The JS estimator may be worse than MLE if the variances of $z_{i,k}$'s in the training set have a substantial difference due to the randomness of splitting.

Therefore, to get a more convincing conclusion, we need to conduct a sensitivity analysis on the dataset splitting. 10,000 simulations with different splits of the dataset are conducted. For each, we compute the ratio of $\mathrm{TSE}^{(\mathrm{MLE})}$ to $\mathrm{TSE}^{(\mathrm{JS})}$.

The distribution of the TSE ratio (Figure 2) shows that MLE outperforms JS estimator in approximately 14.4% of cases, namely JS estimator outperforms MLE in approximately 85.6% of cases, indicating a tremendous advantage for the empirical Bayes estimates.

References

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

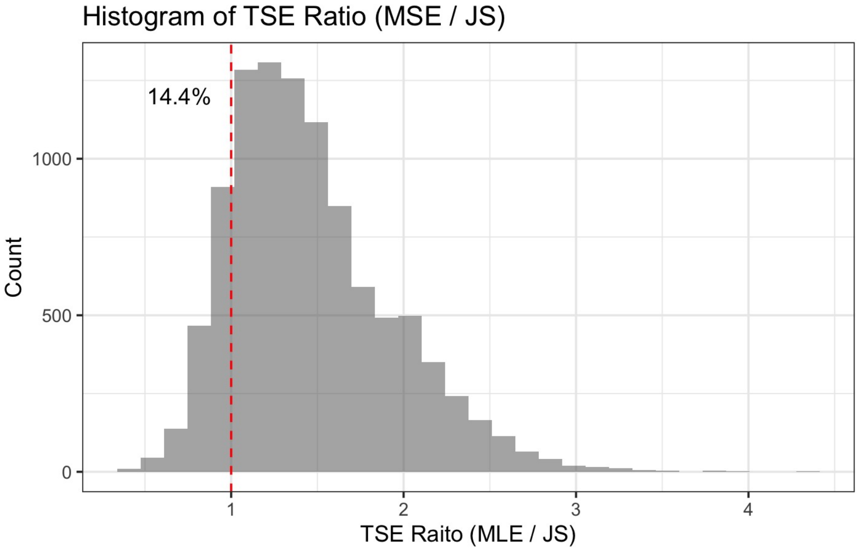

Figure 2: Histogram of $\text{TSE}^{(\text{MLE})}/\text{TSE}^{(\text{JS})}$ in 10000 Simulations