James-Stein Estimator

RK

January 5, 2014

Abstract

This document contains some brief notes that might be helpful to understand *James-Stein estimator*.

Background

Many years ago, I stumbled on to "Steins's Pardox in Statistics", an article by Brad Efron and Carl Morris. Back then my understanding about statistics was cursory and could not properly appreciate the various aspects of the article. On a side note, there was a point in time when I had a chance to be a part of PhD program. For various reasons, I could not attend the program. However there was always an eagerness in me, to learn and apply statistical concepts. Having missed a chance to go over a formal education program, I took the next available option. Over the past few years, I have self-taught some aspects of statistics. So far I have been enjoying the *learning ride* and hope it continues in the future.

Last week I happened to refer to the article, "Steins's Pardox in Statistics". In my earlier date with the article, I was too blind to appreciate the beauty of it. This time my eyes were far more receptive in going through the key idea of the paper and its related developments. To fully appreciate the ideas mentioned in the paper, one needs to have some understanding about the following topics :

- MLE estimation of the parameters of a Multivariate Normal distrubution.
- Statistical Estimation theory
- Bayes Estimation.
- Empirical Bayes Estimation.

Also I firmly believe that if you can replicate the results from a paper/article, then your understanding is far more deeper than merely reading the paper. In this *Big Data* world, some amount statistical programming knowledge is becoming essential for learning, understanding and applying statistics. In this note, I will be using `R` code. Thanks to the wonderful **knitr** package, I have been able to combine code and content all in one document.

What does this note contain ?

This note contains some concepts that will help one understand JS estimator. I will go over some essential concepts of Statistical Decision Theory in Section 1. Post that I will explain Empirical Bayes in Section 2. In Section 3, I will dive in to "Steins's Pardox in Statistics" and highlight various points mentioned by the authors. I will also embed the `R` code that has been used to replicate some of the results.

1 Statistical Decision Theory

There are different ways to estimate the distribution parameters of a random variable such as MLE, posterior mean, method of moments etc. How does one choose among various estimators ? Decision theory gives the answer by laying down a formal theory for comparing statistical procedures.

Assume that we are interested in a parameter $\theta \in \Theta$ and we have $\hat{\theta}$ as the estimator for θ. One needs a way to compare the result of estimator or decision rule with that of the true parameter value. This discrepancy is measured using a function called **loss function**. Formally denoted by $L(\theta, \hat{\theta}) : \Theta \times \Theta \to \mathbb{R}$. A loss function is actually related to Utility function from the Utility theory. The relationship between Utility function and Loss function is

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

Some examples of loss functions are

$$\begin{aligned}
L(\theta, \hat{\theta}) &= (\theta - \hat{\theta})^2 \\
L(\theta, \hat{\theta}) &= |\theta - \hat{\theta}| \\
L(\theta, \hat{\theta}) &= |\theta - \hat{\theta}|^p \\
L(\theta, \hat{\theta}) &= 1 \text{ if } \theta = \hat{\theta} \text{ or } 0 \text{ otherwise} \\
L(\theta, \hat{\theta}) &= \int \log\left(\frac{f(x;\theta)}{f(x:\hat{\theta})}\right) f(x;\theta)dx
\end{aligned}$$

Loss function, alone is not sufficient for judging estimators. Except for the trivial case, the loss function usually fails to show that one estimator uniformly dominates another. The way out of this is to take its expected value. It goes by the name **risk function**.

$$R(\theta, \hat{\theta}) = E_\theta(L(\theta, \hat{\theta})) = \int L(\theta, \hat{\theta}) f(x;\theta)d\theta$$

It seems like the job's done but there are many cases where risk function is not uniformly better for an estimator versus others. Usually one comes across situations where for a certain θ range, the risk function of the estimator $\hat{\theta}_A$ is better than the estimator $\hat{\theta}_B$ and for a certain range, it is the other way around. To judge the performance of one estimator with the other, the decision theory literature gives one number summary of the risk function. At this juncture, the frequentists and Bayesians part ways.

A frequentist defines **maximum risk** as

$$\overline{R}(\hat{\theta}) = \sup_\theta R(\theta, \hat{\theta})$$

and a Bayesian defines **Bayes risk** as

$$r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta)d\theta$$

With these one number summaries, one can compare estimators in the frequentist world and Bayesian world. The two summaries of risk function suggest two different methods of devising estimators : choosing $\hat{\theta}$ to minimize the maximum risk leading to **minimax rule** , choosing $\hat{\theta}$ to minimize Bayes risk leading to **Bayes rule**.

How does on go about finding a Bayes estimator from the Bayes risk. Here is the crucial connection that makes it easy to find Bayes estimator. To make the connection, one needs to first define **posterior risk** of an estimator $\hat{\theta}(x)$ by

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta.$$

Once posterior risk is defined, the connection between Bayes risk and Posterior risk is as follows:

$$r(\hat{\theta}|x) = \int r(\hat{\theta}|x) f(\theta|x)\, m(x)\, dx$$

where $m(x) = \int f(x, \theta) d\theta$, the marginal distribution of X.

So, how does the above connection help in computing Bayes estimator ? If one ends up minimizing posterior risk function for every x , one achieves the equivalent rule of minimizing the integrand at every x and thus minimize Bayes risk . The way to find a Bayes estimator for a prior f and a loss function L is to find $\hat{\theta}$ that minimizes $r(\hat{\theta}|x)$. For tractable loss functions, one can differentiate the posterior risk function to find the Bayes rule. Ok, this is all good for a Bayesian ; Why should a frequentist bother about Bayes risk ? The main reason (sometimes gets lost in all the math) is that, finding minimax estimators is difficult and complicated. If an estimator has constant risk function, then automatically it is minimax.

One last concept from the decision theory that is useful for understanding the article, is admissibility. Minimax estimators and Bayes estimators are good estimators in the sense that they have small risk. To characterize bad estimators, one needs to understand admissibility.

An estimator $\hat{\theta}$ is inadmissible if there exists another rule $\hat{\theta}'$ such that

$$\begin{aligned} &R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta}) \quad \forall \theta \\ &R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta}) \quad \text{for atleast one } \theta \end{aligned}$$

Otherwise, $\hat{\theta}$ is admissible. This concept of inadmissible is used in the paper to show that MLE estimators are inadmissible. So, one needs to carefully understand this aspect of bad estimators. The advantage of working in Bayes world is that Bayes rules are admissible.

Let me summarize the key terms that one needs to understand :

- Loss function
- Risk function
- Total Risk
- Bayes Risk
- Bayesian estimator/rule
- Minimax rule
- Posterior risk
- Inadmissible estimator
- Admissible estimator

2 Empirical Bayes

Bayes has grown in to prominence in the last decade or so, mainly because of the tools and technology available for one and all to "do bayes". There are three things that you need to do Bayesian analysis.

1. Likelihood function, $f(x|\theta)$
2. Prior, π
3. Loss function, $L(\theta, \hat{\theta})$

The discussion of Empirical Bayes in the context of this paper is apt as one might say that Efron and Morris are the founders of Modern Empirical Bayes. Efron and Morris wrote a series of papers that is considered to be the first major work in the field of empirical bayes. There's an excellent paper by George Casella, "An Introduction to Empirical Bayes Data Analysis", that is the source of this section.

What's Empirical Bayes ?. Empirical Bayes is one where you don't have to struggle to define the prior. It can be culled from the data. The empirical Bayes technique relies on the observations (and the marginal distribution) to estimate the parameters of the prior distribution; it is used by frequentists more often than by Bayesians because it does not belong to the Bayesian paradigm.
An example is better to illustrate Empirical Bayes. Let's consider a p variate Multivariate Normal Random variable,

$$X_i \sim N(\theta_i, \sigma^2), \quad i = 1, 2, \dots p$$

and let the prior on θ_i be

$$\theta_i \sim N(\mu, \tau^2), \quad i = 1, 2, \dots p$$

Posterior distribution of θ_i can be obtained by going through the Bayes grind

$$\pi(\theta_i|X) \sim N\left(\frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}X_i, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$$

In the Bayes world, your selection of prior affects the posterior. The empirical Bayesian agrees with the Bayes model but refuses to specify values for μ and τ^2. Instead he estimates these parameters from the data. All of the information for the prior parameters is contained in the marginal distribution of X_i and another standard calculation shows that this marginal distribution, $f(X_i)$, is given by

$$f(X_i) \sim N(\mu, \sigma^2 + \tau^2), \quad i = 1, 2, \dots, p$$

Thus empirically one can deterimine the parameters via the following relations

$$E(\overline{X}) = \mu, E\left[\frac{(p-3)\sigma^2}{\sum(X_i - \overline{X})^2}\right] = \frac{\sigma^2}{\sigma^2+\tau^2}$$

Substituting the value of prior parameter values in the Bayes estimator, one obtains the following expression as an estimator for θ_i

$$\left[\frac{(p-3)\sigma^2}{\sum(X_i - \overline{X})^2}\right]\overline{X} + \left[1 - \frac{(p-3)\sigma^2}{\sum(X_i - \overline{X})^2}\right]X_i$$

Thus the estimator for θ_i uses information from all the X_i's. This takes advantage of what has come to be known as Stein effect.The Stein effect asserts that estimates can be improved by using information from all coordinates when estimating each coordinate.

3 Stein's Paradox

With all the prerequisites covered, let me dive in to this article and highlight some of the important points. This article is termed as a fairly nontechnical account by many leading statisticians, but somehow I failed to grasp most of the content on the first go. This time around I think I have understood the paper to a large extent. In this note, I will attempt to verbalize the same.
The subtitle of the paper is pretty eye catching

> The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average.

In 1955, Charles Stein of Stanford University discovered a mathematical result that is strikingly contrary to the generally held belief. His result was contradictory to a century and half of work on estimator theory. Naturally there was a stiff resistance. But soon when everyone realized the logical soundness of the proof, Stein's methods were adopted everywhere. Counting and Averaging are the two most basic processes in statistics. An estimator that challenges average was bound to face a lot of resistance. What is so paradoxical about Stein's statment ? Let's say you have three cities and you have a sample from each of the three cities. To estimate the incidence rate in each of the city, one typically goes about taking the average of people affected in each sample as an estimate for incidence rate of each city. Stein's paradox says that this estimator is inadmissible. Given the understanding from Section 1, it means that there is another estimator whose risk is less than the risk associated with sample average.
The authors use baseball data to illustrate the working's of Stein's method.

```
pathname            <- file.path(".", "/data/baseball.txt")
data                <- read.table(pathname, header=T,stringsAsFactors=F)
data$name           <- paste(data[,2],data[,1],sep=",")
data                <- subset(data, select  =c(name, BattingAverage,RemainingAverage))
colnames(data)      <- c("Player","InSamp","OutSamp")
```

Table 1 shows the dataset used in the paper.
The InSamp column is the batting average of the players based on 45 times at bat. The OutSamp column is the batting average for the rest of the season. If one goes by MLE, then the best estimate for the rest of the season is in fact the performance on the forst 45 sample points for each of the player. James-Stein estimator for the rest of the season is based on empirical bayes analysis.
First let us see the results before we go on to the theory part.

```
n                   <- 45
k                   <- 18
data$InSampT        <- sqrt(n)*asin(2*data[,2]-1)
data$OutSampT       <- sqrt(n)*asin(2*data[,3]-1)
global.avg          <- mean(data$InSampT)
dev                 <- data$InSampT - global.avg
data$jst            <- global.avg + (1 - (k-3)/sum((data$InSampT -global.avg)^2 ))*dev
data$js             <- 0.5*(sin(data$js/sqrt(n))+1)
performance         <- sum((data$InSampT - data$OutSampT)^2)/
```

Player	InSamp	OutSamp
Clemente,Roberto	0.400	0.346
Robinson,Frank	0.378	0.298
Howard,Frank	0.356	0.276
Johnstone,Jay	0.333	0.222
Berry,Ken	0.311	0.273
Spencer,Jim	0.311	0.270
Kessinger,Don	0.289	0.265
Alvarado,Luis	0.267	0.210
Santo,Ron	0.244	0.269
Swaboda,Ron	0.244	0.230
Petrocelli,Rico	0.222	0.264
Rodriguez,Ellie	0.222	0.226
Scott,George	0.222	0.303
Unser,Del	0.222	0.264
Williams,Billy	0.222	0.330
Campaneris,Bert	0.200	0.285
Munson,Thurman	0.178	0.316
Alvis,Max	0.156	0.200

Table 1: Baseball data

```
                    sum((data$jst - data$OutSampT)^2)
temp                <- subset(data, select = c(Player, OutSamp,InSamp,js))
colnames(temp)      <- c("Player","TrueValue", "MLE","Stein")
countperf           <- sum(abs(temp$TrueValue - temp$Stein) <
                           abs(temp$TrueValue - temp$MLE))
```

Table 2 shows MLE and Stein's estimate. On a MSE criterion, the efficiency of Stein is 3.18. On a count basis, Stein is closer to true parameter value for 15 out of 18 players.The paper shows it as 16 but I guess the slight discrepancy is due to numerical approximation.

Player	TrueValue	MLE	Stein
Clemente,Roberto	0.346	0.400	0.290
Robinson,Frank	0.298	0.378	0.286
Howard,Frank	0.276	0.356	0.282
Johnstone,Jay	0.222	0.333	0.277
Berry,Ken	0.273	0.311	0.273
Spencer,Jim	0.270	0.311	0.273
Kessinger,Don	0.265	0.289	0.268
Alvarado,Luis	0.210	0.267	0.264
Santo,Ron	0.269	0.244	0.259
Swaboda,Ron	0.230	0.244	0.259
Petrocelli,Rico	0.264	0.222	0.254
Rodriguez,Ellie	0.226	0.222	0.254
Scott,George	0.303	0.222	0.254
Unser,Del	0.264	0.222	0.254
Williams,Billy	0.330	0.222	0.254
Campaneris,Bert	0.285	0.200	0.249
Munson,Thurman	0.316	0.178	0.244
Alvis,Max	0.200	0.156	0.239

Table 2: Batting average and their estimates

```
temp2               <- subset(data, select = c(Player, OutSamp,InSamp,js))
colnames(temp2)     <- c("Player","TrueValue", "MLE","Stein")
temp2               <- melt(temp2, id=c("Player"))
colnames(temp2)     <- c("Player","Type", "Average")
ggplot(temp2, aes(x=Player,y = Average,fill=Type))+
  geom_bar(stat="identity",position="dodge")+coord_flip()
```

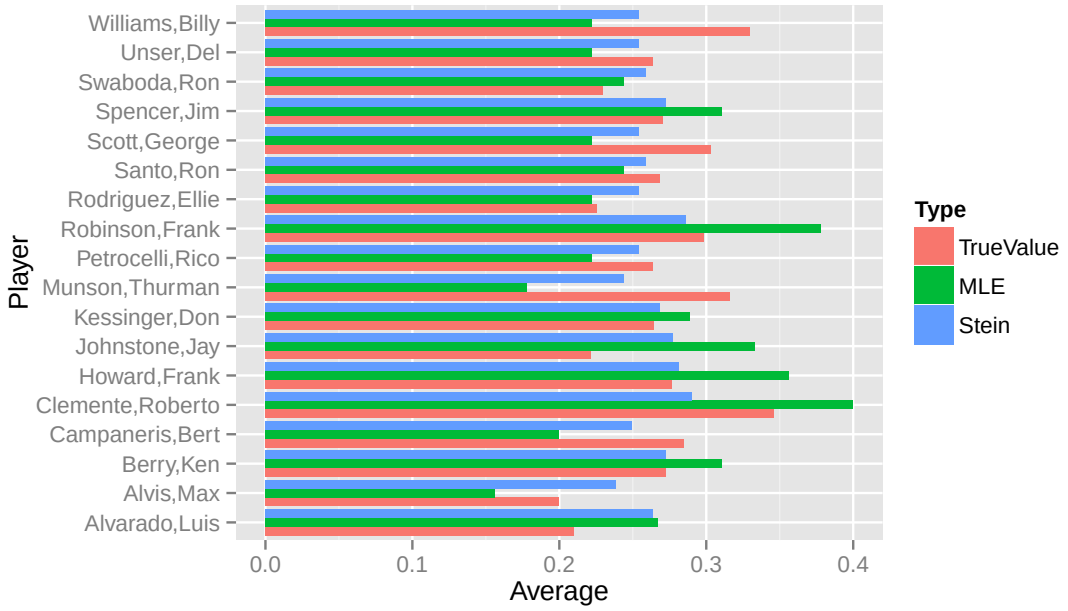

Figure 3.1: Batting Average comparison

```
temp3               <- subset(data, select = c(Player, OutSamp,InSamp,js))
colnames(temp3)     <- c("Player","TrueValue", "MLE","Stein")
temp3$MLE.errorsq   <- (temp3$TrueValue-temp3$MLE)^2
temp3$Stein.errorsq <- (temp3$TrueValue-temp3$Stein)^2
temp3               <- subset(temp3, select = c(Player, MLE.errorsq,Stein.errorsq))
temp3               <- melt(temp3, id=c("Player"))
colnames(temp3)     <- c("Player","Type", "ErrorSquared")
ggplot(temp3, aes(x=Player,y = ErrorSquared,fill=Type))+
  geom_bar(stat="identity",position="dodge")+coord_flip()
```

Figure 3.2: Error Squared comparison

The visuals and the numbers show that Stein estimation beats MLE estimates comprehensively. Now here is where things start to become paradoxical. If you want to estimate the fraction of foreign made automobiles in Chicago, there is nothing that stops from mixing that sample along with baseball average sample and reporting a Stein estimator. At the first go, it sounds crazy. How can one mix automobile sample with that of base ball sample. Why should a batter's underperformance or over performance have any effect on the fraction of automobiles in Chicago. Such type of questions have been raised by critics and in order to fully explain the method, the paper takes on a brief detour in to estimation theory.

The paper explains why we are so biased towards averaging. For any random sample from a normal distribution, Gauss showed that the average is the best unbiased estimator. No linear or non linear function of data is as good as the average. "Best" meant expected squared error of estimate for average is less than any other estimator. In 1930's , a mathematically more rigorous approach to statistical inference was undertaken by Neyman and Pearson. They discarded the idea of requirement of unbiased estimation and examined all functions of data that could serve as estimators. Let's say you have a sample of data points from a normal distribution with an unknown mean and known standard deviation. From the observed data, one could form many estimators. Let's say the estimators chosen are

1. sample average
2. half of sample average
3. median

How does one decide which one of the estimators is good. This is where the Statistical Decision theory of Section 1 comes in to picture. One can compute the risk function for various values of true parameter and see how the estimator performs. Let us simulate some data to draw the risk functions for each of the above estimators.

```
risk   <- function(theta){
  n    <- 50
  b    <- 1000
  avg  <- replicate(b,{data <- rnorm(n,theta,1)
      (mean(data)-theta)^2
        })

  half.avg <- replicate(b,{data <- rnorm(n,theta,1)
      (mean(data)/2-theta)^2
        })

  median <- replicate(b,{data <- rnorm(n,theta,1)
      (quantile(data,prob=0.5)-theta)^2
        })
   c(mean(avg),mean(half.avg),mean(median))
}
thetas     <- seq(-1,1,length.out = 100)
results    <- sapply(thetas,function(z)risk(z))
```

```
plot(thetas, results[1,], type="l", col = "green", ylim = range(results),
     ylab="Risk",xlab=expression(theta))
par(new=T)
plot(thetas, results[2,], type="l", col = "blue", ylim = range(results),
     ylab="",xlab="")
par(new=T)
plot(thetas, results[3,], type="l", col = "red", ylim = range(results),
     ylab="",xlab = "")
legend("center",legend=c("average","average/2","median"),
       col=c("green","blue","red"),lty=c(1,1,1),cex=0.8)
```

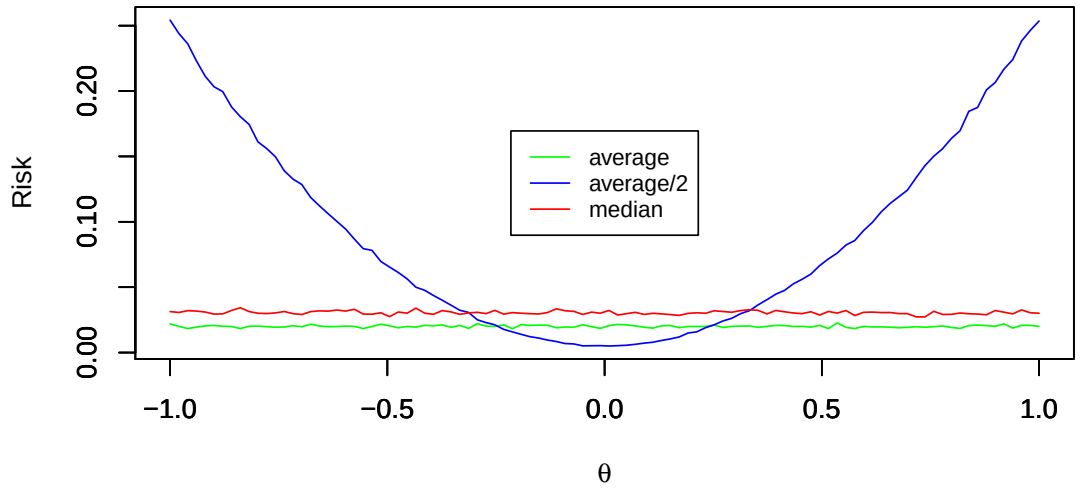

Figure 3.3: Single Mean Estimator comparison

As one can see the average is uniformly better than median. The risk function for half of average is actually a parabola. At specific values of the θ, the risk function is better than the average but soon average starts dominating. So, is there a better estimator for a single sample data ? Answer : No. This was conclusively proven in 1950. In the language of decision theory, the average is an **admissible estimator**.

Stein's theorem is concerned with the estimation of several unknown means. No relation between the means need to be assumed. They can be batting abilities or proportions of imported cars. One can simulate data and verify Stein's result by comparing the total risk of an MLE estimate vs. total risk of Stein estimate.
The following code simulates data from a multivariate normal and uses James-Stein Estimator. Subsequently the total risk is computed for varying levels of σ^2.

```
total.risk <- function(means){
  sig      <- 1
  b        <- 1000
  results  <- replicate(b,{
```

```
    x        <- rmvnorm(1, mean = means)
    grd.avg  <- mean(x)
    c        <- 1 - 7/sum((x-grd.avg)^2)
    estimate <- grd.avg + c*( x - grd.avg)
    sum((estimate - means)^2)
    })
   mean(results)
}
mus       <- cbind(sqrt(seq(0,10,by=0.20)),-sqrt(seq(0,10,by=0.20)))
risk.val  <- apply(mus,1,function(z){
             mu.temp <- c(rep(z[1],5),rep(z[2],5))
             total.risk(mu.temp)
             })
ssq       <- apply(mus,1,function(z)sum(z^2))
```

```
plot(ssq, risk.val, type="l", col = "blue", ylim = c(0,12),lwd=2,
     ylab="Total Risk",
     xlab="Total Squared Deviation of means from their average",
     cex.lab=0.8)
abline(h=10,lwd= 2, col = "red")
legend("center",legend=c("James-Stein Estimators","Observed Averages"),
       col=c("blue","red"),lty=c(1,1,1),lwd =2, cex=0.8)
```

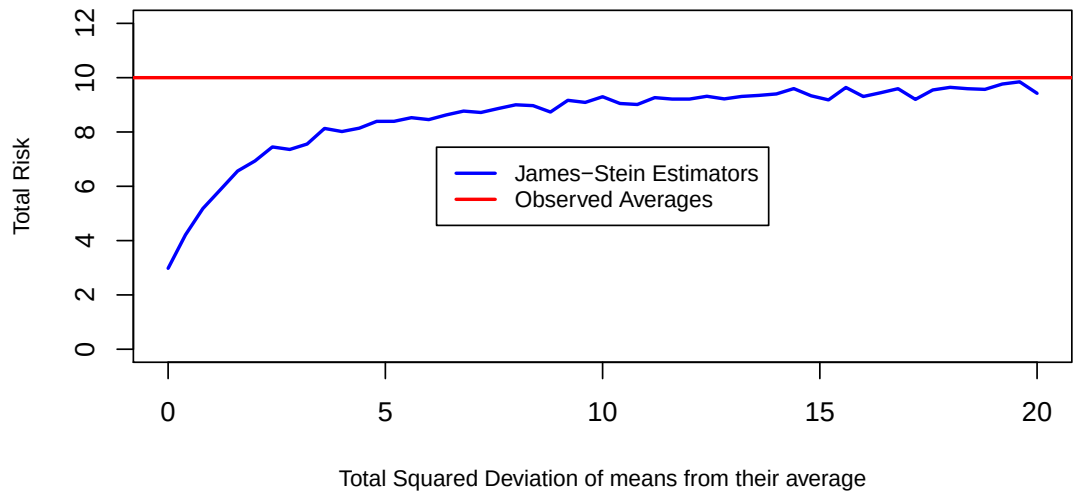

Figure 3.4: Total Risk Function

The simulation results show that total risk for Stein estimator is always less than MLE estimator, no matter what the variation with in the true means of the 10 dimensional random variable.

Word of Caution:

The authors subsequently point out an important aspect of JS. The formula is in such a way that it shrinks the individual estimates towards a grand average and in doing so is more efficient than the MLE. However the individual estimates are sometimes way off. This brings back the discussion on," whether it is sensible to add chicago automobile data to base ball data" ? JS estimation says its fine. But something is not correct here.

Also what if you have a decent idea of the prior ? Obviously it makes sense to apply that prior instead of the prior from Empirical Bayes. These are issues that one must keep in mind while using JS estimator. One must be cautious about not using "atypical" data. In this context, the paper talks about another case study that involves estimation of a certain endemic disease. The point the author want to make via the other case study is that one must not blindly apply JS estimation to a multivariate mean estimation. It is possible that the data is clustered and each of clusters are completely different. By treating different clusters as one cluster, JS estimation does not give good results.

The paper ends by saying that JS estimator is not the only one that is known to be better than the sample averages. JS estimator is itself inadmissible. Its failure lies in the fact that the shrinkage factor can sometimes be negative. This can be remedied by taking the shrinkage factor only when it is positive. Even this type of tweak does not make James-Stein estimator an admissible one.

The search for new estimators continues....