# Homework 2 - Solution

Instructor: Hedibert Freitas Lopes
Course: STP 598 Advanced Bayesian Statistical Learning (Class # 31199)
Semester: Spring 2022.

## 1 Risk analysis

Recall that the risk of an estimator $\hat{\theta}$ is given by

$$R(\theta, \hat{\theta}) = E_\theta[L(\theta, \hat{\theta})] = \int L(\theta, \hat{\theta}) p(x|\theta) dx,$$

while the maximum risk is $\overline{R}(\hat{\theta}) = \sup_\theta R(\theta, \hat{\theta})$, and the Bayes risk is

$$r(\pi, \hat{\theta}) = \int_\Theta R(\theta, \hat{\theta}) \pi(\theta) d\theta,$$

where $\pi$ is a prior for $\theta$. Assume that the loss function is squared error, so the risk is just the mean squared error (MSE):

$$R(\theta, \hat{\theta}) \quad = \quad E_\theta[(\hat{\theta} - \theta)^2] = E_\theta[(\hat{\theta} - E_\theta(\hat{\theta}))^2 + (E_\theta(\hat{\theta}) - \theta)^2] = V_\theta(\hat{\theta}) + \text{bias}^2_\theta(\hat{\theta}).$$

Now, let $X_1, \ldots, X_n$ be, conditionally on $\theta$, independent Bernoulli($\theta$), for $\theta \in (0, 1)$. Consider squared error loss and two estimators of $\theta$:

$$\hat{\theta}_1 = \frac{X_1 + \cdots + X_n}{n} \quad \text{and} \quad \hat{\theta}_2 = \frac{X_1 + \cdots + X_n + \alpha}{\alpha + \beta + n},$$

where $\alpha$ and $\beta$ are positive constants.

a) Show that

$$R(\theta, \hat{\theta}_1) = \frac{\theta(1-\theta)}{n}$$

Since $\hat{\theta}_1 = (1/n)s_n$ and $s_n|\theta \sim Binomial(n, \theta)$, for $s_n = X_1 + \cdots + X_n$, it follows that $E(\hat{\theta}_1|\theta) = \theta$ and

$$R(\theta, \hat{\theta}_1) = Var(\hat{\theta}_1|\theta) = \frac{1}{n^2} n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}.$$

b) Show that

$$R(\theta, \hat{\theta}_2) = \frac{n\theta(1-\theta)}{(\alpha + \beta + n)^2} + \left(\frac{n\theta + \alpha}{\alpha + \beta + n} - \theta\right)^2$$

Similarly,

$$\hat{\theta}_2 = \frac{n}{\alpha+\beta+n}\hat{\theta}_1 + \frac{\alpha}{\alpha+\beta+n},$$

so

$$E(\hat{\theta}_2|\theta) = \frac{n\theta+\alpha}{\alpha+\beta+n} \quad \text{and} \quad V(\hat{\theta}_2|\theta) = \frac{n^2}{(\alpha+\beta+n)^2}V(\hat{\theta}_1|\theta) = \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2}.$$

For c), d) and e), assume that $\alpha = \beta = \sqrt{n/4}$.

c) Graphically show that neither estimator uniformly dominates the other. Try $n = 1, 10, 50$ to see how the risk functions behave as $n$ increases.

```
R1 = function(theta,n){
  theta*(1-theta)/n
}
R2 = function(theta,n,alpha,beta){
 n*theta*(1-theta)/(alpha+beta+n)^2+
 ((n*theta+alpha)/(alpha+beta+n)-theta)^2
}

thetas = seq(0,1,length=1000)
par(mfrow=c(2,2))
for (n in c(1,10,100,1000)){
  alpha = round(sqrt(n/4),1)
  beta  = round(sqrt(n/4),1)
  r1 = R1(thetas,n)
  r2 = R2(thetas,n,alpha,beta)
  plot(thetas,r1/r2,xlab=expression(theta),
       ylab="Relative risk - R1/R2",type="l",lwd=2)
  title(paste("n=",n,"\n (alpha,beta)=(",alpha,",",beta,")",sep=""))
  abline(h=1,lty=2)
}
```

d) Show that the maximum risks are $\overline{R}(\hat{\theta}_1) = \frac{1}{4n}$ and $\overline{R}(\hat{\theta}_2) = \frac{n}{4(n+\sqrt{n})^2}$, so, based on the maximum risk, $\hat{\theta}_2$ is a better estimator. However, when $n$ is large, $R(\hat{\theta}_1)$ has smaller risk except for a small region in the parameter space near $\theta = 1/2$, where the risk of $\hat{\theta}_1$ is maximum.

It is easy to see that $R(\theta, \hat{\theta}_1)$ is maximized when $\theta = 1/2$ (quadratic function!). From (c), we graphically see that $R(\theta, \hat{\theta}_2)$ is constant for all values of $\theta$. Therefore, for $\theta = 0$, the results follows easily and directly.

e) Show that the Bayes risks are $r(\pi, \hat{\theta}_1) = \frac{1}{6n}$ and $r(\pi, \hat{\theta}_2) = \frac{n}{4(n+\sqrt{n})^2}$, when $\pi$ is the uniform prior in the interval $(0, 1)$. For large $n$ (larger than or equal to 20), $\hat{\theta}_1$ is a better estimator. This corroborates with the graphical inspection obtained in c).

$$r(\pi, \hat{\theta}_1) = \int_0^1 \frac{\theta(1-\theta)}{n} d\theta = \frac{1}{n} \int_0^1 \underbrace{\theta^{2-1}(1-\theta)^{2-1}}_{\text{Kernel of a } Beta(2,2)} d\theta = \frac{1}{n} \frac{\Gamma(2)\Gamma(2)}{\Gamma(4)} = \frac{1}{6n}.$$

Similarly,

$$r(\pi, \hat{\theta}_2) = \int_0^1 \frac{n}{4(n+\sqrt{n})^2} d\theta = \frac{n}{4(n+\sqrt{n})^2}.$$

## 2 Stein's Paradox

Suppose that $X \sim N(\theta, 1)$ and consider estimating $\theta$ with squared error loss. We know that $\hat{\theta}(X) = X$ is admissible. Now consider estimating two, unrelated quantities $\theta = (\theta_1, \theta_2)$ and supposed that $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ independently, with loss

$$L(\theta, \hat{\theta}) = (\theta_1 - \hat{\theta}_1)^2 + (\theta_2 - \hat{\theta}_2)^2.$$

Not surprisingly, $\hat{\theta}(X) = X$ is again admissible where $X = (X_1, X_2)$. Now consider the generalization to $k$ normal means. Let $\theta = (\theta_1, \ldots, \theta_k)$, $X = (X_1, \ldots, X_k)$ with $X_i \sim N(\theta_i, 1)$ (independent) and loss

$$L(\theta, \hat{\theta}) = (\theta_1 - \hat{\theta}_1)^2 + \cdots + (\theta_k - \hat{\theta}_k)^2.$$

Stein astounded everyone when he proved that if $k \geq 3$, then $\hat{\theta}(X) = X$ is inadmissible. It can be shown that the **James-Stein estimator**

$$\hat{\theta}^S = (\hat{\theta}_1^S, \ldots, \hat{\theta}_k^S)$$

has smaller risk, where

$$\hat{\theta}_i^S(X) = \left(1 - \frac{k-2}{X_1^2 + \cdots + X_k^2}\right)^+ X_i,$$

where $(z)^+ = max\{0, z\}$. This estimator shrinks the $X_i$'s towards 0. The message is that, when estimating many parameters, there is great value in shrinking the estimates.

**Computer Experiment:** Compare the risk of the MLE and the James-Stein estimator by simulation. Try various values of $k$ and various vectors $\theta$. Summarize your results.

```
ks   = c(seq(4,20,by=1),seq(30,300,by=10))
Rep  = 1000
Risk = matrix(0,Rep,2)
nk   = length(ks)
RR   = rep(0,nk)
for (l in 1:nk){
  theta = rnorm(ks[l])
  for (r in 1:Rep){
    x      = rnorm(ks[l],theta,1)
    th.mle = x
    th.js  = max((1-(ks[l]-2)/sum(x^2)),0)*x
    Risk[r,1] = sum((th.mle-theta)^2)
    Risk[r,2] = sum((th.js-theta)^2)
  }
  RR[l] = mean(Risk[,1])/mean(Risk[,2])
}

plot(ks,RR,xlab="Dimesion of x vector",ylab="Relative Risk",main="")
title("MLE vs James-Stein",type="b")
```