MLE vs James Stein Estimator

Prepared by:

- Josh Levine (jl2108)
- Harsh Patel (hkp49)
- Jaini Patel (jp1891)
- Yifan Liao (yl1463)
- Aayush Shah (avs93)

Problem Statement:

Compare the risk of the James stein estimator vs the MLE for k =1 to 100 for random thetai, plot it, use 1000 samples at each k to estimate risk. k is dimension!Stein estimator is based on $X_i \sim N(\theta_i, 1)$, i = 1,. . . ,k.

Goal:

Here, we want to compare two estimators, the James-Stein estimator and the Maximum Likelihood estimator(MLE). We are comparing the estimators for θ, which is a k-dimensional vector.

To accomplish this goal, we compare the risk (estimated mean square errors) of both the estimators for the value of k ranging from 1 to 100. This k-dimensional vector is the mean vector of k-variate normal random variable. Here, the variance (σ^2) will be 1.

To implement this, first we generate a k-dimensional vector of mean θ as zero. Then, generate sample data of size 1000 for each random variable $\sim N(\theta_k, 1)$.

This way we estimate the MLE estimator, James-stein estimator and the corresponding risks from our observations.

James-Stein estimator:

Stein's example (or phenomenon or paradox), in decision theory and estimation theory, is the phenomenon that when three or more parameters are estimated simultaneously, there exist combined estimators more accurate on average (that is, having lower expected mean squared error) than any method that handles the parameters separately. It is named after Charles Stein of Stanford University, who discovered the phenomenon in 1955.

An intuitive explanation is that optimizing for the mean-squared error of a combined estimator is not the same as optimizing for the errors of separate estimators of the individual parameters. In practical terms, if the combined error is in fact of interest, then a combined estimator should be used, even if the underlying parameters are independent. If one is instead interested in estimating an individual parameter, then using a combined estimator does not help and is in fact worse.

The following is perhaps the simplest form of the paradox, the special case in which the number of observations is equal to the number of parameters to be estimated. Let θ be a vector consisting of $n \geq 3$ unknown parameters. To estimate these parameters, a single measurement X_i is performed for each parameter θ_i,

resulting in a vector X of length n. Suppose the measurements are known to be independent, Gaussian random variables, with mean θ and variance 1, i.e., $X_i \sim N(\theta_i, 1)$.

Thus, each parameter is estimated using a single noisy measurement, and each measurement is equally inaccurate.

If (σ^2) is known, for more than one vector observations, the James-Stein estimator is given by:

$$\widehat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2/n}{\|\bar{y}\|^2}\right)\bar{y}$$

where $\bar{y}$ is the m-length average of the n observations.

If one vector observation is available,

$$\widehat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|y\|^2}\right)y$$

Where m = k dimensions, y = single observation

(Here, the average of n observations is nothing but the estimation using MLE. The term getting multiplied by the MLE estimate has to be positive. JS estimated value is just the scaled down version of MLE estimated value.)

Maximum Likelihood estimator (MLE):

The maximum likelihood estimate for K-dimensional θ mean is the sample mean from the observations sampled from a multivariate Gaussian distribution.

Risk of an estimator:

The quality of such an estimator is measured by its risk function. A commonly used risk function is the mean squared error of the original parameter and the estimated parameter, defined as

$$E[\|\theta - \widehat{\theta}\|^2]$$

Code to compare MLE and James Stein Estimator:

Below chunk of code is the function to calculate the mle risk and james stein risk values.

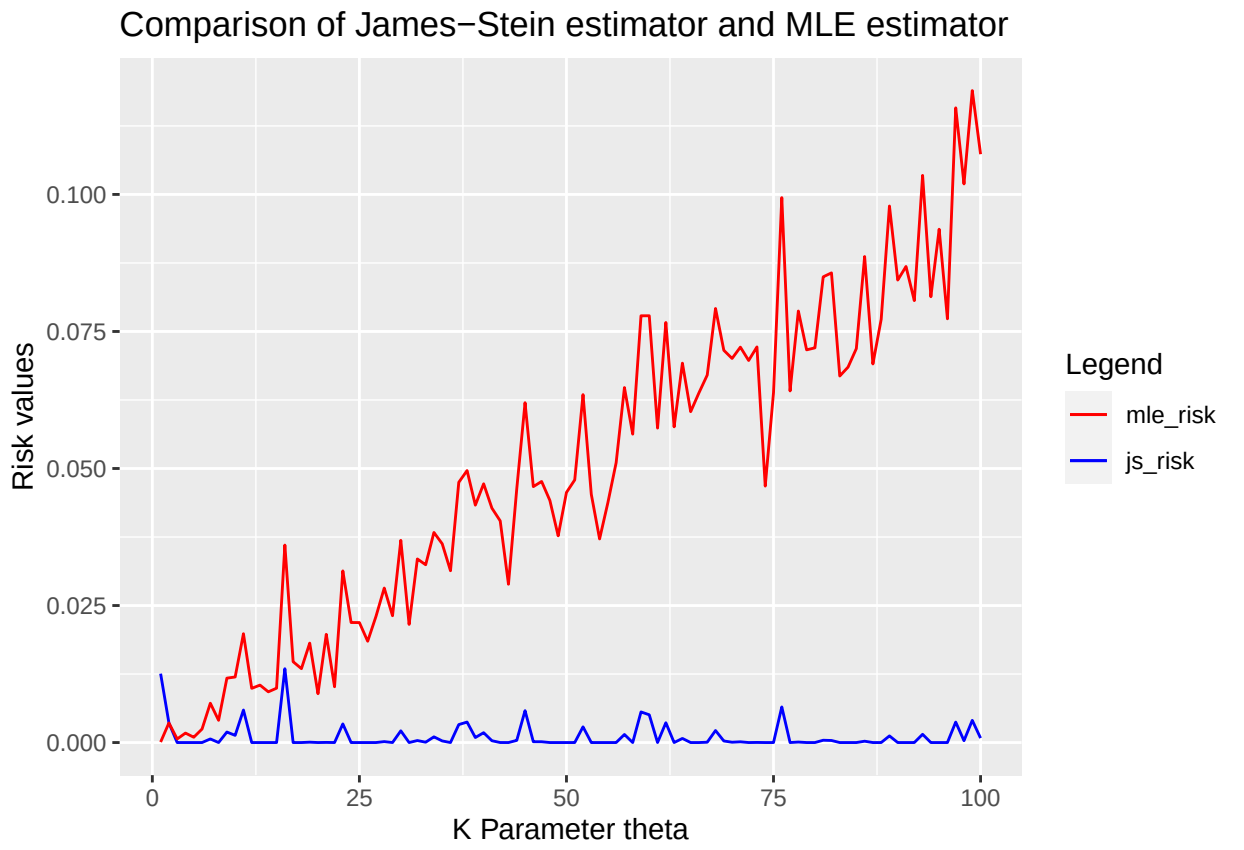
```
pos <- function(x){
  if (x>0){
    return(x)
  }
  else{
    return(0)
  }
}

risk_test_function <-function(k,n)
{
    #initializing mean values
    mu <- 0
    for (i in 1:k){
```

```
      mu[i] <- 0
    }
    #initializing random data
    input_data = mvrnorm(n, mu, diag(k))

    #calculating mle of data
    theta_mle = colMeans(input_data)

    #mle risk calculation
    risk_mle = sum((mu - theta_mle)**2)
    #print(paste("MLE theta",theta_mle))
    #print(paste("Risk MLE",risk_mle))

    #calculating js estimates
    theta_mle_norm = norm(theta_mle, "2")
    theta_js = pos(1 - ((k-2)/(n*(theta_mle_norm**2)))) * theta_mle
    #print(paste("MLE norm",theta_mle_norm))
    #print(paste("JS theta",theta_js))

    #js risk calculation
    risk_js = sum((mu-theta_js)**2)
    #print(paste("risk JS", risk_js))

    return(c(risk_js,risk_mle))

}
```

Following chunk of code is the stimulation function that calls the risk function and plots the comparison.

```
risk_calculation <- function(){
  n = 1000
  js_risk <- NULL
  mle_risk <- NULL
  for (k in (1:100)){
    out <- risk_test_function(k,n)
    js_risk <- c(js_risk,out[1])
    mle_risk <- c(mle_risk,out[2])

  }

  #converting lists into dataframe
  df_js <- data.frame(x = 1:100, js_risk) #dataframe for js risk estimates
  df_mle <- data.frame(x = 1:100, mle_risk) #dataframe for mle risk estimates
  #print(df_js)
  #print(out)
  #print(paste("JS Risk", js_risk))
  #print(paste("MLE risk", mle_risk))

  #plotting both the risk estimates.
  plot_risk <- ggplot() +
    geom_line(data = df_js, aes(x = x, y = js_risk,color = "blue")) +
    geom_line(data = df_mle, aes(x= x, y = mle_risk, color = "red")) +
    labs(x = "K Parameter theta", y = "Risk values" ) +
```

```
    ggtitle("Comparison of James-Stein estimator and MLE estimator") +
    scale_color_identity(name = "Legend", breaks = c("red","blue"),
                         labels = c("mle_risk", "js_risk"), guide = "legend")
  plot_risk
}


risk_calculation()
```



Conclusion:

Why is this toy problem a good model for many problems?

For k $>$ 2, the James-Stein estimator dominates the Maximum likelihood estimator, as the risk of James-stein estimator is lower than the risk of MLE.

An estimator $\hat{\theta}_1$ is said to dominate another estimator $\hat{\theta}_1$ if, for all values of θ, the risk of $\hat{\theta}_1$ is lower than, or equal to, the risk of $\hat{\theta}_2$, and if the inequality is strict for some θ. An estimator is said to be admissible if no other estimator dominates it, otherwise it is inadmissible.

Thus, Stein's example can be simply stated as follows: The ordinary decision rule for estimating the mean of a multivariate Gaussian distribution is inadmissible under mean squared error risk.

From our observations, we can see that the least squares estimators are inadmissible when $k \geq 3$. Thus, when three or more unrelated parameters are measured, their total MSE can be reduced by using a combined

estimator such as the James–Stein estimator; whereas when each parameter is estimated separately, the least squares (LS) estimator is admissible.

References:

1) https://en.wikipedia.org/wiki/James%E2%80%93Stein_estimator
2) https://en.wikipedia.org/wiki/Stein's_example