Imperial College London

M2R Project Group 13
James Stein Estimator

Authors:
Joshua Rose (01844579)
Jing Hui Tan (01846103)
Jingyuan Wang(01883432)
Govind Bhachu (01860032)
Ervis Bucaj (01857968)
Jayden Barker (01927910)

Supervisor: Alastair Young

22 June 2022

Abstract

This paper analyses what the James-Stein Estimator is, why we introduce the James-Stein Estimator, and why it is a better method to estimate the true mean vector of a multivariate normal random variable under the assumption that it follows $N(\theta, I_p)$ with dimension $p \geq 3$ as compared to the naive estimator. To answer these questions, this paper analyses the performance of the James-Stein Estimator by considering its dominance over the naïve estimator and comparing it with naïve estimator through analysing the distribution of their loss functions. The limitation and potential improvements of the James-Stein Estimator are also discussed in this paper. To further evaluate the performance of the James-Stein Estimator, comparison between the loss function over absolute error norm and Euclidean norm is discussed as well. As for real-life implications, a football example which estimates the goal to shot ratios of football players is introduced to demonstrate the better performance of the James-Stein Estimator.

Keywords: James-Stein Estimator, naïve estimator, loss function, risk function, dominance, admissibility and inadmissibility, multivariate normal, shrinkage estimator, central and noncentral chi-square distribution, absolute error norm, Euclidean norm.

Contents

1 Introduction **3**

2 James Stein Estimator **4**
2.1 Motivation 4
2.2 Inadmissibility of the Naive Estimator 4
2.3 Explicit Determination of the Risk 5
2.4 Inadmissibility of the James Stein Estimator 8

3 The Distribution of the Loss Functions **9**
3.1 Motivation 9
3.2 The Naive Estimator 10
3.3 The James-Stein Estimator with zero mean 11
3.4 The General Case - Distribution of the Difference in Loss 13
3.5 Conclusion 18

4 Improving δ_{JS}^{+} **18**
4.1 Motivation 18
4.2 Stein's Lemma - reworked 20
4.3 Shrinkage Estimators 21
4.4 Improved estimator 22
4.4.1 Risk Calculation 22
4.4.2 Proof of dominance & finding optimal b 23
4.4.3 Analysis 24
4.4.4 Simulations 25
4.5 Conclusion & Performance Explanation 27

5 Norms **28**
5.1 Zero θ 28
5.2 Non zero θ 30
5.2.1 Conclusion 33

6 James-Stein Estimator Experiment **33**
6.1 James-Stein Estimator 33
6.2 Football Example 33
6.3 Results 35
6.4 Conclusion 36

A Code For Figures **38**

1 Introduction

Consider a simple estimation problem. Let $X_1, ..X_p$ be p independent normal random variables with unit variance, i.e. $X_i \sim N(\theta_i, 1)$ for $i = 1, .., p$. Naturally if $\theta = (\theta_1, ...\theta_p)^T$ is unknown, we would want to estimate the value of θ by some estimator $\hat{\theta}$. Whilst normally we may look to maximum likelihood estimation we introduce a more detailed environment which will give us a far better estimator than the MLE.

We measure how "close" an estimator $\hat{\theta}$ is to the true parameter through a **loss function**, $L(\hat{\theta}, \theta)$. We choose the Euclidean norm, so $L(\hat{\theta}, \theta) = ||\hat{\theta} - \theta||^2$.

Later in this paper we will examine the distribution of these $L(\hat{\theta}, \theta) = ||\hat{\theta} - \theta||^2$ for different estimators. Ideally we would like an *average loss* (over the probability distribution of $\hat{\theta}$) to summarise an estimators performance. Which leads to the introduction of the **risk function**

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

We compare the performance of two estimators $\hat{\theta}_1, \hat{\theta}_2$ through **dominance**. If $R(\hat{\theta}_1, \theta) \leq R(\hat{\theta}_2, \theta)$ for all values of θ, with a strict inequality for some value of θ, we say $\hat{\theta}_1$ **strictly dominates** $\hat{\theta}_2$.

Theoretically, the most powerful estimators are ones which are **not strictly dominated by any other estimator**, such estimators are called **admissible**. Now it must be noted that admissible estimators are not necessarily always the best.

With the established theory behind us we look to a solution to the problem presented initially; it may seem obvious to choose the estimator $X = (X_1, ..., X_p)^T$. Since each X_i is normally distributed around θ_i each component of our X vector will naturally be "close" to our true mean θ.

However, it turns out this estimator is inadmissible for $p \geq 3$, and even more powerfully the James Stein estimator:

$$\delta_{JS} = \left(1 - \frac{p-2}{||X||^2}\right) X \tag{1}$$

strictly dominates the naive estimator, $\delta_0 = X$.

Examining δ_{JS} we notice *Stein's Paradox*; why do we get better performance on an estimator δ_{JS} whose ith component depends on every component of X, even though the means θ_i & θ_j are not related in any way? Somehow we have concluded that in an independent sample of p normal random variables, the best way to estimate the mean of any one of random variables is to create an estimator which is dependent on *all* of them!

In this paper we focus on analysing the performance of the James-Stein estimator, beginning with proving that we have dominance over the naive estimator X.

2 James Stein Estimator

2.1 Motivation

In 1956, Charles Stein proved that for $p \leq 2$ the sensible, naive estimator is **admissible** and surprisingly **inadmissible** for $p \geq 3$. However it took until 1961 for Willard James and Charles Stein to produce and prove that the James Stein estimator, δ_{JS} strictly dominates the naive estimator. The result seems paradoxical at first but the estimator is merely a shrinkage estimator, that shrinks our vector, X, towards the origin. Before proving the inadmissibility of the naive estimator we will provide a neat and useful result.

2.2 Inadmissibility of the Naive Estimator

Lemma 2.1 (Stein's Lemma)**.** *If $X \sim N(\theta, 1)$ and we have a well behaved real function, h(X) then*

$$E[(X-\theta)h(X)] = E[h'(X)].$$

Proof. By the definition of expectation

$$E[(X-\theta)h(X)] = \int_{-\infty}^{\infty} (x-\theta)h(x)\phi(x-\theta)\,dx,$$

where $\phi(x)$ is the probability density function of a standard normal distribution. By integration by parts let $u = h(x)$, $dv = (x-\theta)\phi(x-\theta)$, so $du = h'(x)$ and $v = -\phi(x-\theta)$. Then we have

$$\int_{-\infty}^{\infty} (x-\theta)h(x)\phi(x-\theta)\,dx = [-h(x)\phi(x-\theta)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} h'(x)\phi(x-\theta)\,dx.$$

The first term on the right hand side disappears and so we're left with

$$E[(X-\theta)h(X)] = \int_{-\infty}^{\infty} h'(x)\phi(x-\theta)\,dx = E[h'(X)].$$

□

Theorem 2.2. *For $p \geq 3$, the James Stein estimator, $\delta_{JS} = \left(1 - \frac{p-2}{||X||^2}\right) X$ dominates the naive estimator $\delta_0 = X$.*

Proof. [1] First we calculate the risk of the naive estimator,

$$R(\delta_0, \theta) = E\left[||\delta_0 - \theta||^2\right] = E\left[||X-\theta||^2\right].$$

Note that $||X||^2$ is just the sum of squared normal distributions (non-central chi-squared distribution). By subtracting θ our quantity has a central chi-squared

distribution that is $||X-\theta||^2 \sim \chi^2_p$. Therefore $R(\delta_0, \theta) = p$. Now we need to calculate the risk of the James Stein estimator,

$$\begin{aligned} R(\delta_{JS}, \theta) &= E\left[\left|\left|\left(1 - \frac{p-2}{||X||^2}\right)X - \theta\right|\right|^2\right] = E\left[\left|\left|X - \theta - \frac{(p-2)X}{||X||^2}\right|\right|^2\right] \\ &= E\left[||X-\theta||^2\right] + 2E\left[\frac{-(X-\theta)^T(p-2)X}{||X||^2}\right] + E\left[\frac{(p-2)^2 X^T X)}{||X||^4}\right] \\ &= E\left[||X-\theta||^2\right] - 2(p-2)E\left[\frac{(X-\theta)^T X}{||X||^2}\right] + (p-2)^2 E\left[\frac{1}{||X||^2}\right]. \end{aligned}$$

Note that the second term on the right hand side can be written as

$$E\left[\frac{(X-\theta)^T X}{||X||^2}\right] = E\left[\sum_{i=1}^p \frac{X_i(X_i-\theta_i)}{\sum_{j=1}^p X_j^2}\right] = \sum_{i=1}^p E\left[\frac{X_i(X_i-\theta_i)}{\sum_{j=1}^p X_j^2}\right].$$

Letting $h(X) = X_i/(\sum_{j=1}^p X_j)$ and applying Lemma 2.1 (Stein's Lemma) we have

$$\sum_{i=1}^p E\left[\frac{\partial}{\partial X_i}\left(\frac{X_i}{\sum_{j=1}^p X_j^2}\right)\right] = \sum_{i=1}^p E\left[\frac{\sum_{j=1}^p X_j^2 - 2X_i^2}{\sum_{j=1}^p (X_j^2)^2}\right] = E\left[\frac{p-2}{||X||^2}\right].$$

So our expression becomes

$$\begin{aligned} &E\left[||X-\theta||^2\right] - 2(p-2)E\left[\frac{(X-\theta)^T X}{||X||^2}\right] + (p-2)^2 E\left[\frac{1}{||X||^2}\right] \\ &= p - 2(p-2)(p-2)E\left[\frac{1}{||X||^2}\right] + (p-2)^2 E\left[\frac{1}{||X||^2}\right] \\ &= p - (p-2)^2 E\left[\frac{1}{||X||^2}\right]. \end{aligned}$$

We realise that $R(\delta_{JS}, \theta) = p - (p-2)^2 E[1/||X||^2] < p = R(\delta_0, \theta)$ and so the James Stein estimator strictly dominates the naive estimator. □

2.3 Explicit Determination of the Risk

We have the James Stein estimator, $\delta_{JS} = \left(1 - \frac{p-2}{||X||^2}\right)X$ and proved that it dominates the naive estimator, $\delta_0 = X$. Even though we have an expression for the risk,

$$R(\delta_{JS}, \theta) = p - (p-2)^2 E\left[\frac{1}{||X||^2}\right],$$

how do we evaluate $E[1/||X||^2]$? Focusing on the case where $X \sim N(0, I_p)$, then $||X||^2 = X_1^2 + \ldots + X_p^2$ is simply the sum of squared standard normal distributions. Hence, $||X||^2 \sim \chi^2_p$, a central chi-squared distribution. In the central case we simply have that $1/||X||^2 \sim$ Inv-χ^2_p (inverse chi-squared distribution) and by

using the formula for the expectation of the inverse chi-squared distribution, our expectation reads as $E[1/||X||^2] = 1/(p-2)$. So we have an explicit expression for the risk of the James Stein estimator

$$R(\delta_{JS}, 0) = p - (p-2)^2 \cdot \frac{1}{p-2} = 2,$$

provided $p \geq 3$. We see that when sampling from a standard normal distribution that the risk of the James Stein Estimator is constant regardless of the dimension. However what happens when $X \sim N(\theta, I_p)$? Instead, $||X||^2$ ends up having a non-central chi-squared distribution namely, $||X||^2 \sim \chi^2_p(\lambda)$ where $\lambda = ||\theta||^2$. The non-central chi-squared distribution has probability density function

$$f(x; p, \lambda) = e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} f(x; p+2i, 0),$$

where $f(x; p+2i, 0)$ is the probability density function of a central chi-squared distribution with $p+2i$ degrees of freedom. Adaptation from Paolella 2007 [2], let $Y = ||X||^2$ and so by the definition of expectation

$$\begin{aligned}
E\left[\frac{1}{Y}\right] &= \int_0^{\infty} \frac{1}{y} e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} f(y; p+2i, 0)\, dy \\
&= e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} \int_0^{\infty} \frac{1}{y} \frac{1}{2^{\frac{p+2i}{2}} \Gamma\left(\frac{p+2i}{2}\right)} y^{\frac{p+2i}{2}-1} e^{-\frac{y}{2}}\, dy \\
&= e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} \frac{1}{2^{\frac{p}{2}+i} \Gamma\left(\frac{p}{2}+i\right)} \int_0^{\infty} y^{\frac{p}{2}+i-2} e^{-\frac{y}{2}}\, dy.
\end{aligned}$$

Letting $u = y/2$ then $du = dy/2$ our integral becomes

$$\begin{aligned}
E\left[\frac{1}{Y}\right] &= e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} \frac{1}{2^{\frac{p}{2}+i} \Gamma\left(\frac{p}{2}+i\right)} \int_0^{\infty} (2u)^{\frac{p}{2}+i-2} e^{-u} \cdot 2\, du \\
&= e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} \frac{1}{\Gamma\left(\frac{p}{2}+i\right)} \frac{2^{\frac{p}{2}+i-2} \cdot 2}{2^{\frac{p}{2}+i}} \int_0^{\infty} u^{\frac{p}{2}+i-2} e^{-u}\, du \\
&= e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} \frac{1}{\Gamma\left(\frac{p}{2}+i\right)} \cdot 2^{-1} \cdot \Gamma\left(\frac{p}{2}+i-1\right) \\
&= 2^{-1} e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^i}{i!} \frac{\Gamma\left(\frac{p}{2}+i-1\right)}{\Gamma\left(\frac{p}{2}+i\right)},
\end{aligned}$$

where we have used the integral definition of the gamma function. Our expression can be simplified further by using the identity $\Gamma(x+n) = x^{(n)}\Gamma(x)$ where

$x^{(n)}$ denotes the Pochhammer function (rising factorial). We then have

$$E\left[\frac{1}{Y}\right] = 2^{-1}e^{-\frac{\lambda}{2}}\sum_{i=0}^{\infty}\frac{\left(\frac{\lambda}{2}\right)^i}{i!}\frac{(\frac{p}{2}-1)^{(i)}\Gamma\left(\frac{p}{2}-1\right)}{(\frac{p}{2})^{(i)}\Gamma\left(\frac{p}{2}\right)}$$
$$= 2^{-1}e^{-\frac{\lambda}{2}}\frac{\Gamma\left(\frac{p}{2}-1\right)}{\Gamma\left(\frac{p}{2}\right)}\sum_{i=0}^{\infty}\frac{\left(\frac{\lambda}{2}\right)^i}{i!}\frac{(\frac{p}{2}-1)^{(i)}}{(\frac{p}{2})^{(i)}}.$$

Upon noticing that our sum is in the form of the confluent hypergeometric function we obtain a neater form of our expectation

$$E\left[\frac{1}{Y}\right] = 2^{-1}e^{-\frac{\lambda}{2}}\frac{\Gamma\left(\frac{p}{2}-1\right)}{\Gamma\left(\frac{p}{2}\right)}{}_1F_1\left(\frac{p}{2}-1;\frac{p}{2};\frac{\lambda}{2}\right)$$
$$= 2^{-1}e^{-\frac{\lambda}{2}}\int_0^1 w^{\frac{p}{2}-2}e^{\frac{\lambda}{2}w}\,dw$$

where the last line is obtained by the integral representation of the confluent hypergeometric function. We now have a nice closed form for $E\left[1/||X||^2\right]$ and hence the risk of the James Stein estimator given any θ reads as

$$R(\delta_{JS},\theta) = p - \frac{(p-2)^2}{2e^{\frac{\lambda}{2}}}\int_0^1 w^{\frac{p}{2}-2}e^{\frac{\lambda}{2}w}\,dw,$$

for $p \geq 3$. Figure 1 shows simulations of the risk function as a way to check our expression for the risk. We note that our risk function for the James Stein estimator only depends on the value of θ. Upon taking derivatives of the risk function with respect to λ we see that our function is in fact increasing and also concave, as shown in figure 2. Also note upon increasing our dimension, p, our risk function increases too.

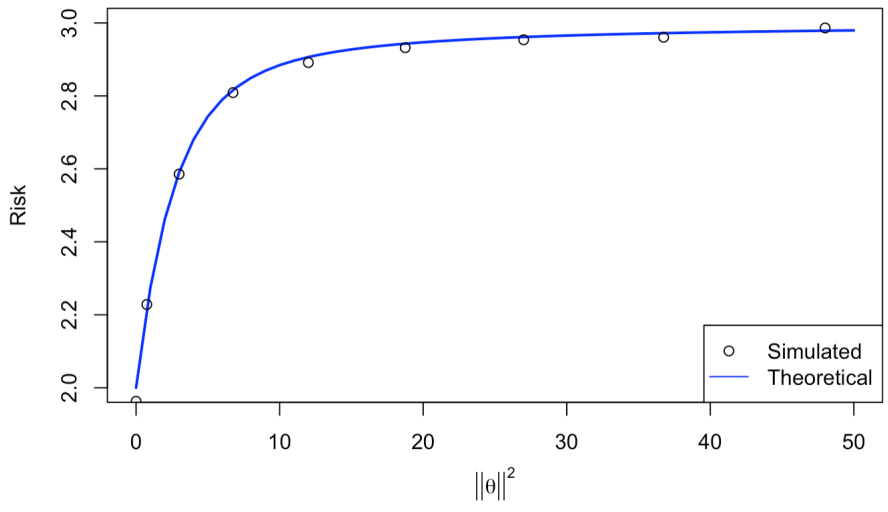


Figure 1: Comparison of risk function and simulations for $p = 3$

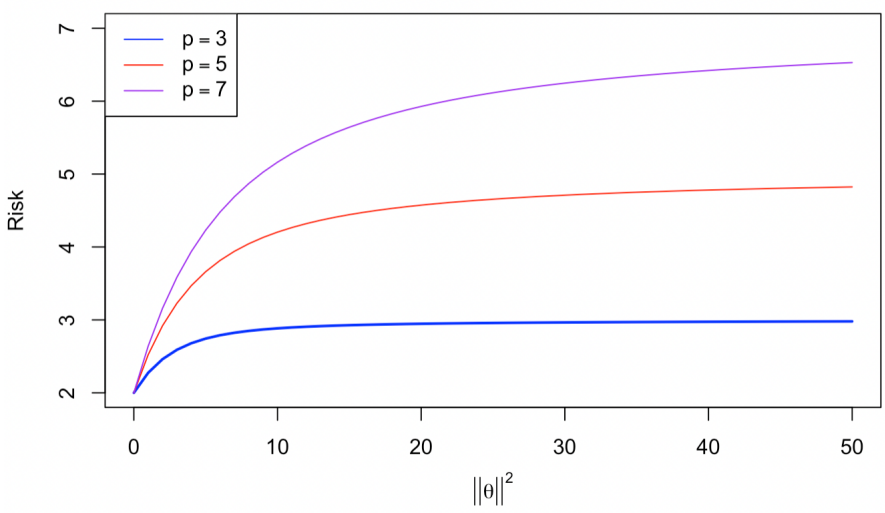


Figure 2: Graph of risk for different dimensions, p

2.4 Inadmissibility of the James Stein Estimator

As seen previously, the James Stein Estimator $\delta_{JS} = (1-(p-2)/||X||^2)X$ which dominates the naive estimator, $\delta_0 = X$ and can be seen as a shrinkage estimator, shrinking the values of X towards the origin. However when $||X||^2 < p-2$ our estimate becomes negative and begins to deviate from the origin. To fix this we introduce the positive-part James Stein estimator

$$\delta_{JS}^{+} = \max\{0, \delta_{JS}\}.$$

It turns out that the James Stein estimator is itself inadmissible and in fact dominated by the positive-part James Stein estimator.

Consider the scenario when $X \sim N(\theta, I_p)$ where $X = (X_1, \ldots, X_p)^T$ and $\theta = (\theta_1, \ldots, \theta p)^T$ and consider estimators of the form $T(||X||)X$ and $T^+(||X||)X$ where $T^+(X) = \max\{0, T(X)\}$.

Lemma 2.3. *[3] Given the conditions above we have the following inequality*

$$R(T^+(||X||)X, \theta) \leq R(T(||X||)X, \theta).$$

Proof. Consider the quantity, $R(T(||X||)X, \theta) - R(T^+(||X||)X, \theta) \geq 0$. By def-

inition

$$
\begin{aligned}
&R(T(||X||)X,\theta) - R(T^+(||X||)X,\theta) \\
&= E\left[||T(||X||)X-\theta||^2\right] - E\left[||T^+(||X||)X-\theta||^2\right] \\
&= E\left[||T(||X||)X-\theta||^2 - ||T^+(||X||)X-\theta||^2\right] \\
&= E\left[T(||X||)^2||X||^2 - 2\theta^T XT(||X||) + \theta^T\theta - T^+(||X||))^2||X||^2 + 2\theta^T XT^+(||X||) - \theta^T\theta\right] \\
&= E\left[T(||X||)^2||X||^2 - T^+(||X)||)^2||X||^2 + 2\theta^T X(T^+(||X||) - T(||X||)\right] \geq 0.
\end{aligned}
$$

By the definition and linearity of expectation observe the quantity,

$$
\begin{aligned}
&E\left[\theta^T X(T^+(||X||) - T(||X||)\right] \\
&= ||\theta|| \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_1[(T^+(||X||) - T(||X||)] \cdot \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\left[\sum_{i=1}^p x_i^2 - 2x_1||\theta|| + ||\theta||^2\right]} dx_1 \dots dx_p \\
&\quad = ||\theta|| e^{-\frac{1}{2}||\theta||^2} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_1 \left[(T^+(||X||) - T(||X||)\right] \left[e^{x_1||\theta||} - e^{-x_1||\theta||}\right] \\
&\qquad\qquad \cdot \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\sum_{i=1}^p x_i^2} dx_1 \dots dx_p \geq 0
\end{aligned}
$$

with the inequality coming from the property of the positive-part function and also $E\left[T(||X||)^2||X||^2 - T^+(||X)||)^2||X||^2\right] \geq 0$. □

Theorem 2.4. *Positive-part James Stein estimator dominates the James Stein estimator*

Proof. Consider $T(x) = \left(1 - \frac{p-2}{x^2}\right)$ then apply Lemma 2.3. □

We've proven that the James Stein estimator is inadmissible and in fact dominated by its own positive part. It turns out that the positive-part James Stein estimator is in fact inadmissible itself. However the construction of such estimators that dominate the positive-part James Stein estimator become more complex; in the sense that no 'nice' forms exist and do not provide that much of an improvement as discussed in Chapter 4.

3 The Distribution of the Loss Functions

3.1 Motivation

Until now, we have explored the James-Stein estimator and compared its 'usefulness' to the naive estimator through the notion of admissibility. In this, we mainly focused on the risk function (expected loss) of the estimator and through this, we showed that the JS estimator dominates the naive estimator. Although the expected loss of the estimators provide us insight into their 'usefulness', it only highlights one aspect of the loss incurred by these estimators. A question that naturally arises is whether there are other features and characteristics of

their loss functions that might provide a more complete picture of their 'usefulness'. This chapter will delve into this question by exploring the distribution of their loss functions.

We begin our investigation by considering the distribution of the naive estimator. Similar to previous chapters, let $X = (X_1, ..., X_p)$ be normally distributed with mean vector θ and covariance matrix I_p where I_p is the p-dimensional identity matrix, i.e. $X_i \sim N(\theta_i, 1)$ independently for $i = 1, ..., p$.

3.2 The Naive Estimator

In the case of the naive estimator, We note that the loss function, $\mathcal{L}$, of the naive estimator is given by:

$$\mathcal{L}(\delta_0, \theta) = ||X - \theta||^2 = \sum_{n=1}^{p} (X_i - \theta_i)^2 \tag{2}$$

Importantly, we observe that $Z_i = X_i - \theta_i \sim N(0, 1)$. Then, the loss function, $\mathcal{L}$, is just the sum of p standard normal random variables. Hence,

$$\mathcal{L}(\delta_0, \theta) = \sum_{n=1}^{p} Z_i^2 \sim \chi_p^2$$

This agrees with our previous findings, where $\mathcal{R}(\delta_0, \theta) = p$. This emphasises the fact that the loss function of the naive estimator depends only on the length of our vector X.

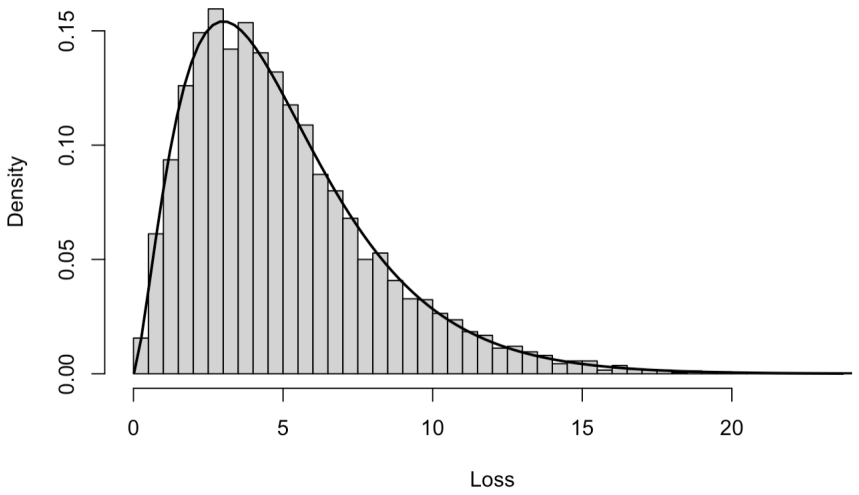


Figure 3: The distribution of the loss function of the naive estimator when $p = 5$

By running simulations (code can be found in Appendix A) where we sample from a multivariate normal distribution with a fixed random mean vector and covariance matrix I_5, we can plot a histogram of the calculated loss. We observe from the figure above that the histogram does fit a χ^2_5 distribution.

Given that we now know the loss function follows a χ^2_p distribution, we know that the variance of the loss function is $2p$. So, we note that as the length of our vector X, given by p, increases, both the expected value and the variance of the loss incurred increases.

3.3 The James-Stein Estimator with zero mean

We first consider the case where the mean vector, θ, is the zero vector. Then, the loss function for the James-Stein estimator in this particular case, $\mathcal{L}(\delta_{JS}, 0)$ denoted L, is given by:

$$\begin{aligned} L &= \left|\left|\left(1 - \frac{p-2}{||X||^2}\right) X\right|\right|^2 \\ &= \left(1 - \frac{2(p-2)}{||X||^2} + \frac{(p-2)^2}{(||X||^2)^2}\right) ||X||^2 \end{aligned}$$

where $||X||^2 \sim \chi^2_p$. Let $V = ||X||^2$. Then the loss function, L, is simply a function of the continuous random variable V,

$$L = g(V) = V - 2(p-2) + \frac{(p-2)^2}{V}$$

In order to find the distribution of L, we will consider a transformation of the random variable V where V has the support $R_V = (0, \infty]$. By differentiating g, we find that g is strictly decreasing in the interval $(0, p-2]$ and strictly increasing in $(p-2, \infty]$. By partitioning R_V into the two intervals, we find that the cdf of L, F_L, is given by:

$$\begin{aligned} F_L(l) &= \mathbb{P}(L \leq l) \\ &= \mathbb{P}(v_1 \leq v \leq v_2) \\ &= F_V(v_2) - F_V(v_1) \end{aligned}$$

where F_V is the cdf of V and

$$\begin{aligned} v_1 &= \frac{1}{2}(-\sqrt{l(4p+l-8)} + 2p + l - 4) \\ v_2 &= \frac{1}{2}(\sqrt{l(4p+l-8)} + 2p + l - 4) \end{aligned}$$

are solutions to $g(v) = l$, which we have obtained by inverting g. Then, by differentiating the cdf of L, we get that the pdf of L is

$$\begin{aligned} f_L(l) &= \frac{f_V(v_1)}{|g'(v_1)|} + \frac{f_V(v_2)}{|g'(v_2)|} \\ &= \frac{1}{2^{p/2}\Gamma(\frac{p}{2})} \left(\frac{(v_2)^{\frac{p}{2}-1} e^{-\frac{v_2}{2}} v_2^2}{v_2^2 - (p-2)^2} - \frac{(v_1)^{\frac{p}{2}-1} e^{-\frac{v_1}{2}} v_1^2}{v_1^2 - (p-2)^2} \right) \quad 0 < y \leq \infty \end{aligned}$$

where f_V is the pdf of V. It can be numerically verified that $\mathbb{E}[L] = 2$ for any value of p, as expected.

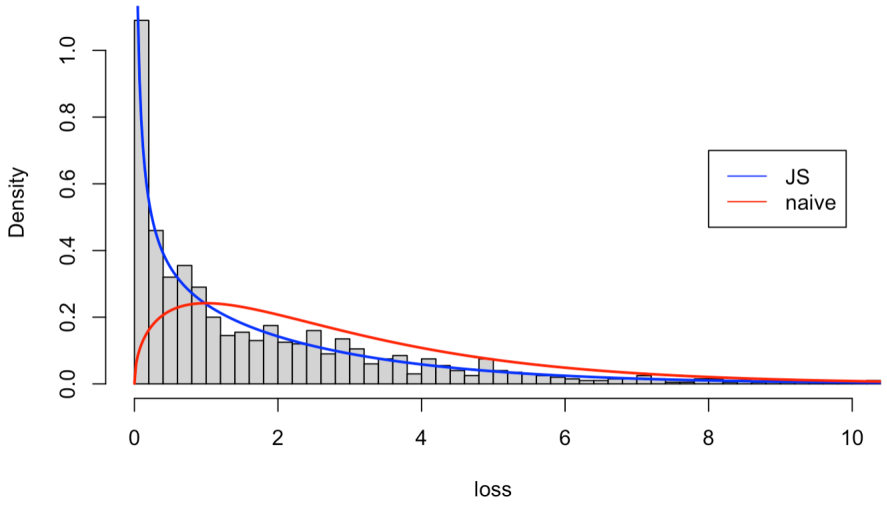


Figure 4: Comparison between the JS estimator and naive estimator for $p = 5$ and $\theta = 0_V$

The histogram above is obtained by sampling from the multivariate standard normal distribution, with dimension $p = 5$ and calculating the loss incurred by the JS estimator to estimate $\theta = 0_V$. By considering the shape of the distributions, it is quite apparent that the JS estimator is preferable as it is distributed around values closer to 0 as compared to the naive estimator.

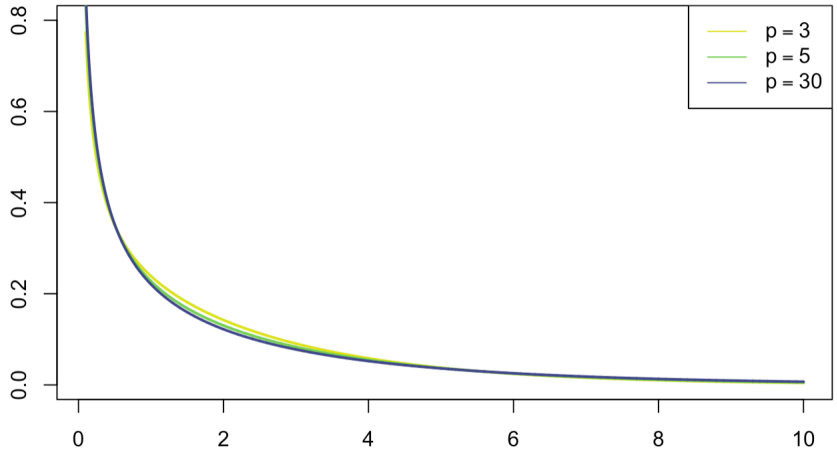


Figure 5: Distribution of L as p varies

Moreover, we can observe that the distribution of L remains relatively similar as p varies. This complements the idea that the risk remains the same for all $p \geq 3$. This provides strong justification for why the JS estimator is preferable compared to the naive estimator in the case where $\theta = 0_V$. This is especially apparent as the number of means being estimated, p, becomes large. Intuitively, since $\theta = 0_V$ and we are 'shrinking towards' the origin, it makes sense that in this case, the JS estimator performs much better than the naive estimator when comparing their losses.

3.4 The General Case - Distribution of the Difference in Loss

In the general case where $\theta \in \mathbb{R}^p$, the loss function is given by:

$$L = \left|\left|\left(1 - \frac{p-2}{||X||^2}\right) X - \theta\right|\right|^2 \tag{3}$$

Rather than analysing the distribution of the loss function of the JS estimator as in the previous section, we will directly consider the distribution of the difference of the JS and naive loss functions, D. The following approach was suggested by Professor Young, our project supervisor. By expanding and taking the difference of equations (3) and (2), we obtain:

$$D = \mathcal{L}(\delta_{JS}, \theta) - \mathcal{L}(\delta_0, \theta) = \frac{2a\theta^T X + a^2}{||X||^2} - 2a$$

where $a = p - 2$. From this, we would like to find the cdf of D, F_D. Consider when $t \neq -2a$. By definition,

$$\begin{aligned}
F_D(t) &= P(D < t) \\
&= P\left(\frac{2a\theta^T X + a^2}{||X||^2} - 2a < t\right) \\
&= P\left(c||X||^2 - 2a\theta^T X - a^2 > 0\right) \qquad (c = 2a + t) \\
&= P\left(||X||^2 - \frac{2a\theta^T X}{c} - \frac{a^2}{c} \lesseqgtr 0\right) \\
&= P\left(\sum_{n=1}^{p}\left(X_i - \frac{a\theta}{c}\right)^2 - \frac{a^2\theta^T\theta}{c^2} - \frac{a^2}{c} \lesseqgtr 0\right) \\
&= P\left(Z^T Z \lesseqgtr \frac{a^2}{c} + \frac{a^2\theta^T\theta}{c^2}\right)
\end{aligned}$$

where $Z = (Z_1, ..., Z_p)^T, Z_i \sim N\left(\left(1 - \frac{a}{c}\right)\theta_i, 1\right)$. Then, $Y = Z^T Z \sim \chi^2_p(\delta)$ where $\delta = \left(1 - \frac{a}{c}\right)^2 \theta^T\theta$ is the non-centrality parameter. We note that the cdf of Y, $F_Y(y)$, is given by $1 - Q_{p/2}\left(\sqrt{\delta}, \sqrt{y}\right)$, where $Q_M(a, b)$ is the Marcum Q-function.

Now, when $t = -2a$, we get

$$\begin{aligned}
F_D(-2a) &= P(D < -2a) \\
&= P\left(\theta^T X < -\frac{a}{2}\right) \\
&= F_{\tilde{Y}}\left(-\frac{a}{2}\right)
\end{aligned}$$

where $\tilde{Y} = \theta^T X \sim N(\theta^T\theta, \theta^T\theta)$

Finally,

$$F_D(t) = \begin{cases} 1 - Q_{p/2}\left(\sqrt{\delta}, \sqrt{\frac{a^2}{2a+t} + \frac{a^2\theta^T\theta}{(2a+t)^2}}\right) & -2a - \theta^T\theta \leq t < -2a \\ \Phi\left(\frac{-a-2\theta^T\theta}{2\sqrt{\theta^T\theta}}\right) & t = -2a \\ Q_{p/2}\left(\sqrt{\delta}, \sqrt{\frac{a^2}{2a+t} + \frac{a^2\theta^T\theta}{(2a+t)^2}}\right) & t > -2a \end{cases} \tag{4}$$

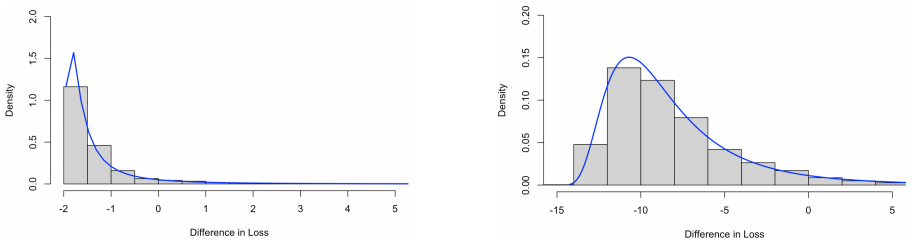
where $\delta = \left(1 - \frac{a}{c}\right)^2 \theta^T\theta$ and Φ is the cdf of the standard normal distribution.

A few interesting observations can be made about the distribution. Firstly, we note that for a given value of t, the non-central χ^2 distribution we depend

on changes as the non-centrality parameter is dependent on t. As a result, it becomes very difficult to derive a closed form for the pdf of D.

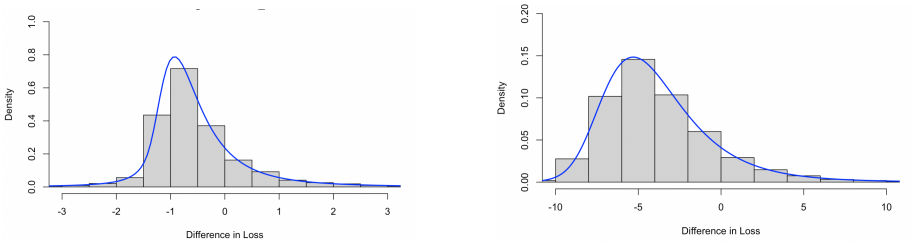
Apart from that, the distribution depends on θ only through $\theta^T\theta$ and not the individual means, θ_i.

However, we can still visualise the pdf by fixing θ and p. For all $t \geq -2a - \theta^T\theta$, We first obtain $F_D(t)$ from their respective distributions (as given in (4)). This allows us to map out the cdf of D. We then use finite difference methods (forward difference in this case) to numerically obtain the derivative of F_D. We can then plot the vector of derivatives we have obtained to produce the pdf plots.



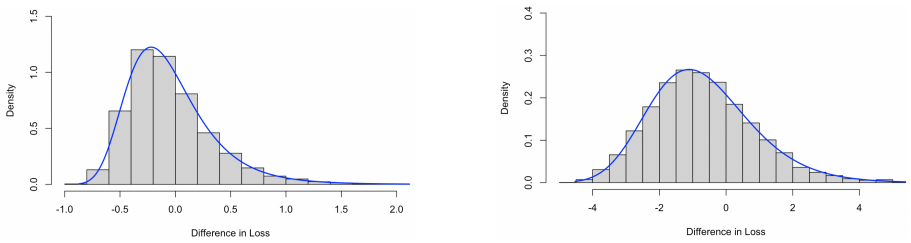
(a) $p = 3$ (b) $p = 10$

Figure 6: Distribution of the difference in loss when $\theta = 0_V$



(a) $p = 3$ (b) $p = 10$

Figure 7: Distribution of the difference in loss when $\theta_i = 1$



(a) $p = 3$ (b) $p = 10$

Figure 8: Distribution of the difference in loss when $\theta_i = 3$

We first consider when $\theta_i = k$ for all $i \in \{1, ..., p\}$ where k is a constant. We note that the general shape of the graph is consistent with our expectations where it is positively skewed and the distribution is concentrated in the negative region of the 'difference in loss' axis. This suggests that the JS estimator is preferable. We can observe that as the magnitude of k increases, the distribution becomes less positively skewed and more symmetric instead. Furthermore, the pdf shifts closer to 0.

On the other hand, when p increases, we see that the distributions shift towards more negative values of 'difference in loss'. As expected, this shift is especially pronounced when $\theta = 0$ as we know that the loss incurred by the JS estimator remains similar as p increases while the expected loss incurred by the naive estimator scales proportionately to p.

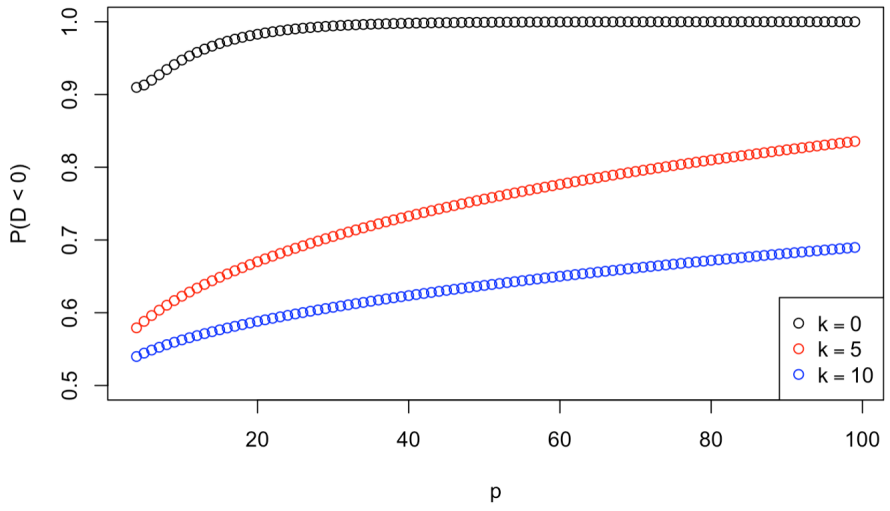


Figure 9: $P(\mathcal{L}(\delta_{JS}, \theta) < \mathcal{L}(\delta_0, \theta))$ as p varies for different values of $\theta_i = k$ for all $i \in \{1, ..., p\}$

Since we are mainly interested in $P(\mathcal{L}(\delta_{JS}, \theta) < \mathcal{L}(\delta_0, \theta))$ or $P(D < 0)$, we

can plot this probability against p. In the above graph, we plot $P(D < 0)$ for 3 different values of k. We note that the curves are monotone increasing and concave. Although all curves follow a similar trend, we see that $P(D < 0)$ is especially large (between 0.9 and 1) for $k = 0$. This follows our intuition: that the JS estimator performs particularly well when we shrink towards the true mean, zero.

We note that $P = P(\mathcal{L}(\delta_{JS}, \theta) < \mathcal{L}(\delta_0, \theta)) > 0.5$ for all p and as p increases, P also increases. However, for a fixed $p \geq 3$ and $k > 0$, P is less than the case where $\theta = 0_V$. Generally, for a fixed $p \geq 3$, $P > 0.5$ for all k and as the magnitude of k increases, P decreases towards 0.5.

We can conclude that it is still preferable to use the JS estimator for any given θ although the difference is not as noticeable.

Although we have only considered the case where θ_i is the same for all $i \in \{1, ..., p\}$, we recall the important fact that the distribution of D is only dependent on θ through the quantity $\theta^T\theta$. Hence, for any mean vector $\theta = (\theta_1, ..., \theta_p)^T \in \mathbb{R}^p$, its distribution will be the same as when the mean vector is $\theta' = (\tilde{\theta}, ..., \tilde{\theta}) \in \mathbb{R}^p$ where $\tilde{\theta} = \sqrt{\frac{\theta_1^2 + ... + \theta_p^2}{p}}$. Hence, the analysis in the general case is similar to our previous discussion.

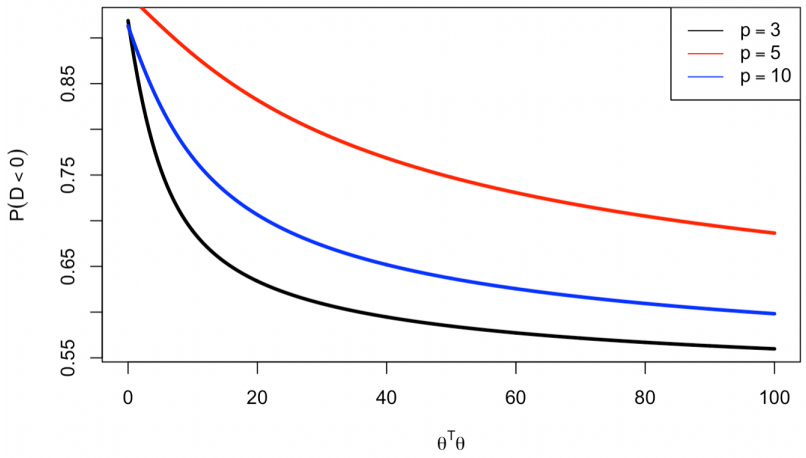


Figure 10: $P(\mathcal{L}(\delta_{JS}, \theta) < \mathcal{L}(\delta_0, \theta))$ as $\theta^T\theta$ varies for different values of p

We see that, as expected, when $\theta^T\theta$ increases, $P(D < 0)$ is monotone decreasing and convex. However, $P(D < 0)$ is always greater than 0.5 when $p \geq 3$. This suggests that the loss incurred by the JS estimator is similar smaller than the naive estimator.

3.5 Conclusion

In this section, we have compared the JS estimator against the naive estimator by analysing the distribution of their loss functions. We have shown that the JS estimator is preferable as compared to the naive estimator. This is especially true in the case where $\theta = 0_V$ where we have observed that the distribution of $\mathcal{L}(\delta_{JS}, 0)$ remains similar as p increases whereas $\mathcal{L}(\delta_0, 0)$ is distributed further away from 0 with increasing p (as it follows a χ^2_p distribution). This suggests that the loss incurred by the JS estimator does not change much whilst the loss incurred by the naive estimator scales linearly with p as p increases.

By considering the distribution of the difference between the loss function of the JS estimator and the naive estimator, D, we were able to analyse the case where $\theta \in \mathbb{R}$ is a general mean vector of length p. In this, we observed a few key results.

Firstly, the distribution of D is positively skewed, with its skewness decreasing as $\theta^T\theta$ increases.

Secondly, we find that as p increases, the distribution shifts to the left. This indicates that the difference in loss between the JS estimator and naive estimator increases as p increases (the loss of the JS estimator is much smaller as p increases). This was further verified by specifically looking at $P(D < 0)$.

Finally, by fixing p and looking at $P(D < 0)$, we find that as $\theta^T\theta$ increases, $P(D < 0)$ monotonically decreases toward 0.5. This suggests that when $\theta^T\theta$ is large, their losses are similar, with the JS estimator performing slightly 'better' (lower loss incurred) than the naive estimator.

4 Improving δ^+_{JS}

4.1 Motivation

As we have shown previously, the James-Stein estimator $\delta_{JS} = (1 - \frac{p-2}{||X||^2})X$ is dominated by it's positive part $\delta^+_{JS} = (1 - \frac{p-2}{||X||^2})\mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}X$, where $\mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}$, is our indicator function which is equal to 1 when the sub-scripted condition is satisfied, and 0, otherwise.

Ideally we would find an **admissible** estimator δ^Q which **dominates** δ^+_{JS} - however such an estimator is extremely difficult to find *explicitly*, thus we must weigh up; admissibility or dominance?

Before we proceed, there exist many **admissible** estimators; notably we would use minimax and Bayes estimators to produce these easily. (See *Maruyama's* paper for how this works.) [4]

Admissibility does not guarantee dominance over every other estimator, it guarantees non-dominance *from* every other estimator - this does not necessarily provide an explicit improvement over δ^+_{JS}.

Generally, we seek to examine the class of shrinkage estimators (δ_{JS} is itself a shrinkage estimator), a specific set varies the shrinkage factor:

$$\delta_\phi = (1 - \frac{\phi(||X||^2)}{||X||^2})X$$

These estimators "shrink" their values towards an arbitrary point - hence their name.

Notably, *Kubokawa* (1991) and *Maruyama* (1996) [5] introduced $\delta^K(X)$ and $\delta^M_\alpha(X)$, which dominate δ_{JS} (δ^K turns out to be admissible!). It is useful to compare how these "famous" estimators perform against δ^+_{JS}. The estimators presented in figure 11 are:

- δ_{JS}
- δ^+_{JS}
- $\delta^K = (1 - \frac{\phi_K(||X||^2)}{||X||^2})X$
- $\delta^M_\alpha = (1 - \frac{\phi_\alpha(||X||^2)}{||X||^2})X$ denoted Mα for $\alpha = 2, 10$

Note: for explicit formulae for ϕ_K & ϕ_α see *Maruyama's* paper referenced [5].

In analyzing figure 11, this particular graph is for when our distribution is $N_p(\theta, \sigma^2 I_p)$ where σ^2 is unknown and we use a sample variance approximation instead (see [5] for details on how the estimators change). Whilst this is not exactly the same setup we have originally, the comparison of risks is useful and can be used to infer information about the risk of our setup in question. Here the non-centrality parameter is $\frac{||\theta||}{\sigma}$ (which directly dictates the behaviour of $||\theta||^2$), but we can fix $\sigma = 1$ to get closer to our case.

At the origin of the figure, where $||\theta||$ is small we see only δ^M_{10} comes close to matching the performance of δ^+_{JS}. Whereas δ^M_2 has a higher risk by 0.3 and δ^K has a higher risk by 0.7 - these are the largest differences between the estimators and δ^+_{JS} in the whole figure.

However, once we stop "over-shrinking" and increase $||\theta||$ we see reduced performance from δ^M_{10} and improved (and more consistent) performance from δ^M_2. For δ_K we see it provides the best performance in terms of risk at these larger values of $||\theta||$. However we are unable to "beat" δ^+_{JS} when our mean vector is small.

Crucially, even though δ^K is admissible, it is difficult to decide whether one chooses to use it over δ^+_{JS} as we can see it does not perform well at small values of $||\theta||^2$!

Thus it is in our interests in seeking an improvement over δ^+_{JS}, we seek dominance over admissibility i.e. we will attempt to find an estimator δ which **dominates** δ^+_{JS}.

It must be noted that most proofs of inadmissibility of δ^+_{JS} have been constructive. Hence, gaining an explicit representation of this estimator is rather powerful.

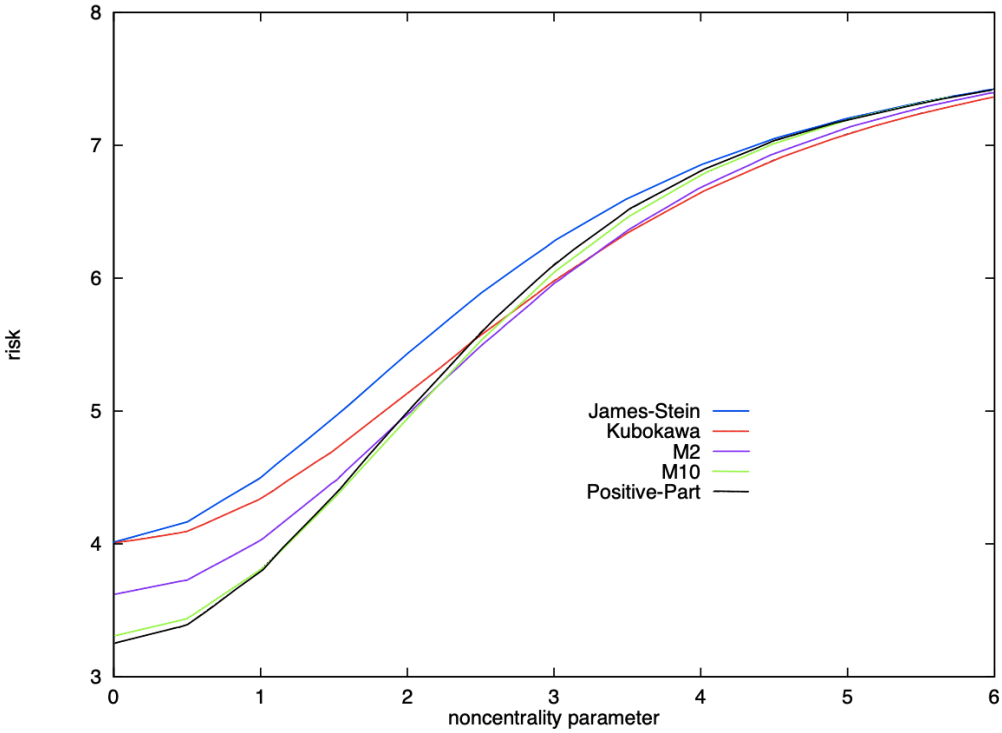


Figure 11: Comparison of the risks of δ_{JS}, δ^+_{JS}, δ^K, δ^M_α with $\alpha = 2, 10$. p = 8

4.2 Stein's Lemma - reworked

In order to give an explicit representation of this estimator that dominates δ^+_{JS}, we must rework Stein's Lemma.

This lemma can be thought of as a more generalized and specialized version of *Stein's Lemma* (1973), which was proven previously. For the sake of saving space I will present a slightly simplified version over the one presented in *Shao* and *Strawderman's* paper. [6]

Let $X \sim N_p(\theta, I_p)$; let $H(.)$ be a continuous function on $[a, b]$; and let $H'(.)$ have at most a finite number of discontinuities 0 lea $= a_0 < a_1 < ... < a_{k+1} = b$. If as we approach each discontinuity from the left and right our gradient remains bounded; i.e. for $i = 0, ..., k+1$, both $H'(a_i^+)$ and $H'(a_i^-)$ are finite, then:

$$E[(X-\theta)^T X H(||X||^2) I_{\{a<||X||^2<b\}}] = \\ E[pH(||X||^2) + 2||X||^2 H'(||X||^2)] I_{\{a<||X||^2<b\}} + \digamma(a) - \digamma(b)$$

For the proof see reference [6].

The motivation for introducing this lemma is to simplify the expectation of an inner product. For our purposes the values $\digamma(a), \digamma(b)$ will vanish (or are negligible).

4.3 Shrinkage Estimators

As mentioned above, the class of Shrinkage Estimators is useful in improving on δ_{JS}^+. *Shao and Strawderman's* paper was the first to present an estimator which strictly dominates δ_{JS}^+, which provides shrinkage on a *specific* region of X (as opposed to shrinkage everywhere for the estimators in figure 11):

$$\delta(a, g, X) = \delta_+^{JS}(X) - \frac{ag(||X||^2)}{||X||^2} X \mathbb{I}_{\{p-2 \le ||X||^2 \le p\}}$$

This estimator provides an improvement upon δ_{JS}^+ solely on the set $\{p - 2 \le ||X||^2 \le p\}$. We use this idea and provide a simpler estimator which provides improvement on $\{X = (X_1, X_2, ..., X_p) : \frac{p-2}{||X||^2} \le 1\}$; a region that is unaffected by δ_{JS}^+. The main idea is to add a term of the form:

$$\alpha(\frac{1}{||X||^2})^m \mathbb{I}_{\frac{p-2}{||X||^2} \le 1} X$$

We have $\alpha = \alpha(p)$ and $m \in \mathbb{Z}$.

N.B. Through recursive addition of the above term it is possible to gain a class of shrinkage estimators with continuous risk improvement (see [7] for full details), however we only present one such estimator here.

4.4 Improved estimator

$$\delta_b^{(2)} = \delta_{JS}^+ + b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X$$

Where b is a positive constant that may depend on p. [7]

Analysing the function at face-value we get the same performance as δ_{JS}^+ for $\{X = (X_1, ..., X_p) \mid \frac{p-2}{||X||^2} > 1\}$ (i.e. $\delta_b^{(2)} = 0$ on this region), so our recursive term only activates on our region of X where we do not "over-shrink".

4.4.1 Risk Calculation

Calculating the risk (under $L(\delta, \theta) = ||\delta - \theta||^2$) [7]

$$R(\delta_b^{(2)}, \theta) = E[||\delta_b^{(2)} - \theta||^2] = E[||\delta_{JS}^+ + b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X - \theta||^2] \quad (5)$$

Then using the identity $||a+b||^2 =< a+b, a+b >= ||a||^2 + ||b||^2 + 2 < a, b >$ and linearity of expectation on (5) we get

$$E[||(\delta_{JS}^+ - \theta) + b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X||^2] =$$

$$E[||\delta_{JS}^+ - \theta||^2] + E[||b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X||^2] + 2E[< (\delta_{JS}^+ - \theta), b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X >] \textbf{ (i)}$$

Simplifying $E[||b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X||^2]$ as follows:

$$E[||b(\frac{1}{||X||^2})^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X||^2] = E[\frac{b^2}{||X||^8} ||X||^2 \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}] = b^2 E[\frac{1}{||X||^6} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]$$

Furthermore upon expansion of δ_{JS}^+ in the inner product we get:

$$< ((X - \theta) - (\frac{p-2}{||X||^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X)), \frac{b}{(||X||^2)^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X >$$

$$=< (X - \theta), \frac{b}{(||X||^2)^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X > - < \frac{p-2}{||X||^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X, \frac{b}{(||X||^2)^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X >$$

$$= \underbrace{< (X - \theta), \frac{b}{(||X||^2)^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1} X >}_{\textbf{(ii)}} - b(p-2) \frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}$$

On taking expectation of the expression **(ii)**, we require the reworked Stein's Lemma (section 4.2), taking $H(||X||^2) = \frac{1}{(||X||^2)^2}$ gives us:

$$b(p-4) E[\frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]$$

Now substituting back into **(i)** and replacing $R(\delta^+_{JS}, \theta) = E[||\delta^+_{JS} - \theta||^2]$ as follows

$$= R(\delta^+_{JS}, \theta) + b^2 E[\frac{1}{||X||^6} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}] + 2b(p-4) E[\frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}] - 2b(p-2) E[\frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]$$

$$= \boxed{R(\delta^+_{JS}, \theta) + \underbrace{b^2 E[\frac{1}{||X||^6} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]}_{\textbf{(iii)}} \underbrace{- 4bE[\frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]}_{\textbf{(iv)}}}$$

We now have an expression $R(\delta^{(2)}_b, \theta)$ in terms of $R(\delta^+, \theta)$ allowing us to provide a sufficient condition on b to guarantee dominance.

4.4.2 Proof of dominance & finding optimal b

We claim:

A sufficient condition for dominance of $\delta^{(2)}_b$ over δ^+_{JS} is:

$$0 \leq b \leq 4(p-2)$$

Proof. Examining expression **(iii)**:

$$b^2 E[\frac{1}{||X||^6} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}] = b^2 E[\frac{1}{||X||^4} \frac{1}{||X||^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]$$

Now noting our indicator function requires the condition $\frac{p-2}{||X||^2} \leq 1 \iff \frac{1}{||X||^2} \leq \frac{1}{p-2}$ to be satisfied for terms **(iii) and (iv)** to be non-zero, so we apply the inequality above to our expression above.

$$b^2 E[\frac{1}{||X||^4} \frac{1}{||X||^2} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}] \leq \frac{1}{p-2} E[\frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]$$

Now replacing in the full expression of the risk:

$$R(\delta^{(2)}_b, \theta) \leq R(\delta^+_{JS}, \theta) + b(\frac{b}{p-2} - 4) \underbrace{E[\frac{1}{||X||^4} \mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]}_{\textbf{(v)}}$$

Noting that **(v)** is non-negative gives us the sufficient condition for dominance that is:

$$\boxed{\frac{b}{p-2} - 4 \leq 0 \iff 0 \leq b \leq 4(p-2)}$$

It can be easily verified that the value of b which minimises $b(\frac{b}{p-2} - 4)$ is $\hat{b} = 2(p-2)$. □

We have thus successfully proved dominance of $\delta_b^{(2)}$ over δ_{JS}^+ and in turn the inadmissibility of δ_{JS}^+!

4.4.3 Analysis

For analysing the improved estimator we reference the graphs provided in the paper referenced. Plotting the risk ratios ($\frac{R(\delta_{JS}^+,\theta)}{R(X,\theta)}$ vs $\frac{R(\delta_b^{(2)},\theta)}{R(X,\theta)}$) can provide a more standardised and "nice" graph than by plotting the risks by themselves. To check how our risk ratios vary, we change $\lambda = ||\theta||^2$, which is the non-centrality parameter of the distribution of $X_1^2 + ... + X_p^2$.

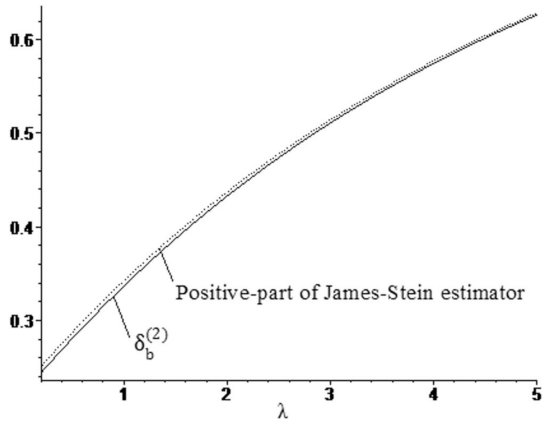


Figure 12: Risk ratio of $\frac{R(\delta_{JS}^+,\theta)}{R(X,\theta)}$ & $\frac{R(\delta_b^{(2)},\theta)}{R(X,\theta)}$ for $p = 6$ [7]

We can see for smaller values of λ we get a better estimation. In the above graph $1 \leq \lambda \leq 6$ so one may think we do not get $\frac{p-2}{||X||^2} > 1$. However in comparison to the mean values - the variance $\sigma_i = 1$ is large, so $||X||^2$ is larger than expected, which in turn forces the condition $\frac{p-2}{||X||^2} \leq 1$ to be satisfied; activating our improvement.

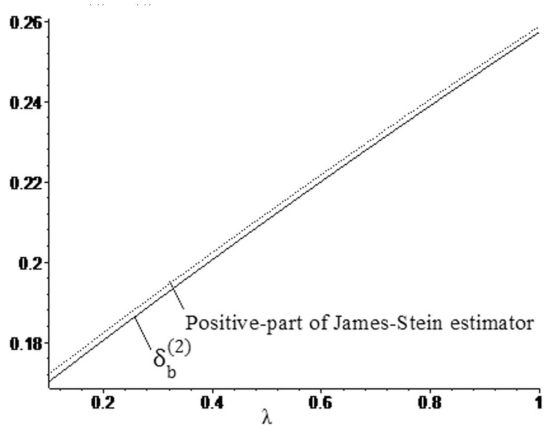


Figure 13: Risk ratio of $\frac{R(\delta^+_{JS},\theta)}{R(X,\theta)}$ & $\frac{R(\delta^{(2)}_b,\theta)}{R(X,\theta)}$ for $p = 8$ [7]

Figure 13 makes it more clear that at our smaller values of $||\theta||^2$ we get more effective behaviour from the improved estimator.

4.4.4 Simulations

Running R simulations (see Appendix A for the code) to compute the average loss (estimate for the risk) for the estimators X, δ^+_{JS} and $\delta^{(2)}_b$ respectively at larger samples.

Our simulation runs as follows with $p = 100$:

1. For $i = 2, ..., 20$ we compute a random mean vector θ_i (100 entries), with the property that each entry $-i \leq \theta_{ij} \leq i$ $(j = 1, ..., 100)$ to provide a loose bound on $||\theta_i||^2$.
2. To get an estimate of the *risk* of each estimator we compute the average loss. We run 500 trials, in each trial we take a sample, $X_1, ..., X_{100}$.
3. For each sample we compute $\delta_{\text{naive}}, \delta^+_{JS}$ & $\delta^{(2)}_b$ and the loss (under $||.||^2$).
4. After repeating this 500 times we average out the losses for each estimator & plot the average loss for a given θ_i.

Note: θ_{ij} denotes the jth entry of θ_i.

Plotting for $\delta_{\text{naive}} = X$ against δ^+_{JS}:

From both figures 14 & 15 we can see significant improvement for smaller i (lower $||\theta^2||$) for both estimators.

We would like to inspect the *amount* of improvement in using $\delta^{(2)}_b$ in practice over δ^+_{JS}. Plotting the average loss difference $\textbf{avg. loss}_{\delta^+_{JS}} - \textbf{avg. loss}_{\delta^{(2)}_b}$ in

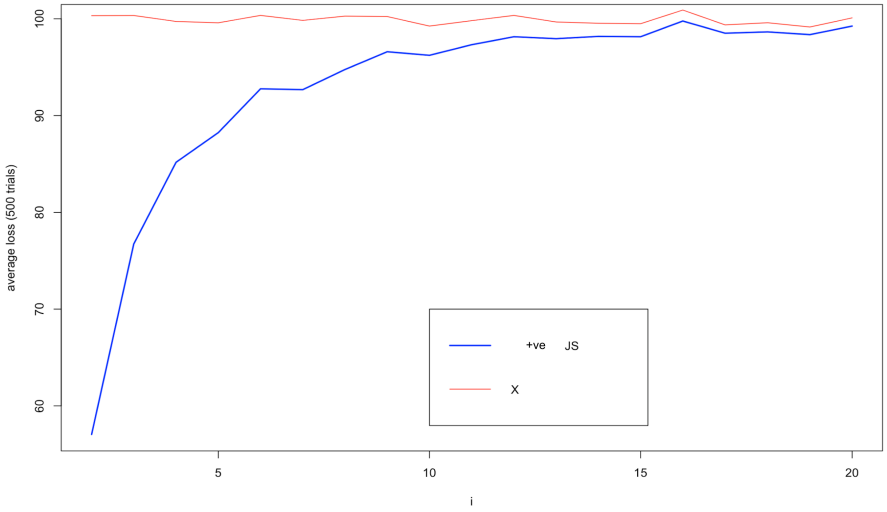


Figure 14: Comparison of average loss for X & δ_{JS}^{+} for $|\theta_{ij}| \leq i$, $j = 1, ..., 100$

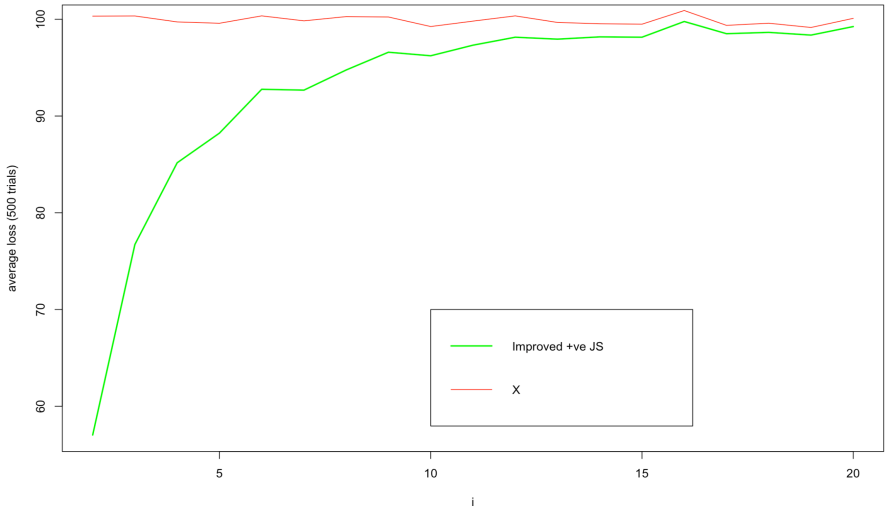


Figure 15: Comparison of average loss for X & $\delta_{b}^{(2)}$ for $|\theta_{ij}| \leq i$, $j = 1, ..., 100$

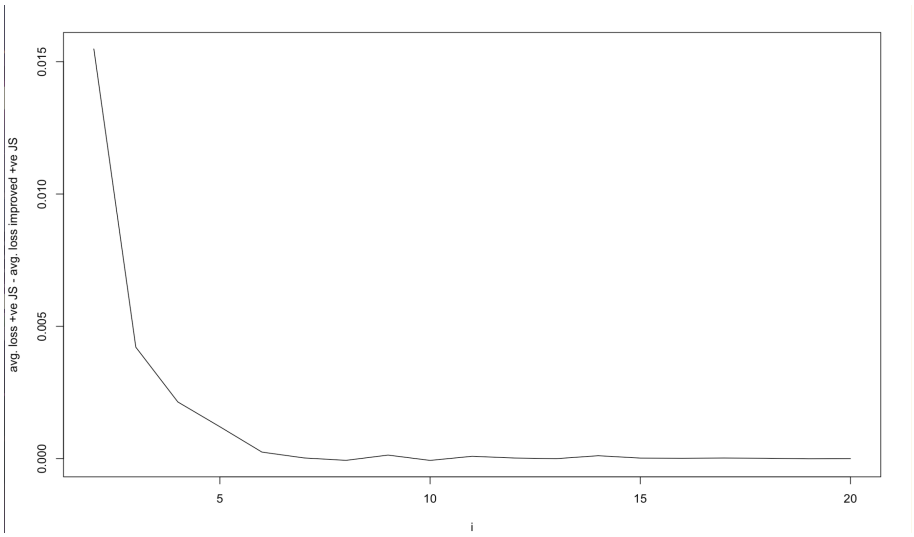


Figure 16: Plot of $\text{avg.loss}_{\delta_{JS}^+} - \text{avg.loss}_{\delta_b^{(2)}}$ for $|\theta_{ij}| \leq i$, $j = 1, ..., 100$

figure 16.

In figure 16 we obtain the most significant improvement in using $\delta_2^{(b)}$ at smaller $||\theta||^2$ - corroborating the evidence seen in figures 12 & 13.

4.5 Conclusion & Performance Explanation

We have shown that it is possible to gain an estimator which dominates δ_{JS}^+ - perhaps in practice it is not such a significant improvement to dismiss the James-Stein estimator from practical use. However, certainly when estimating the mean of multivariate normal models for which we know $||\theta||^2$ small it is significantly advantageous to use $\delta_2^{(b)}$.

The reasoning for why we get better performance for smaller $||\theta||^2$ in $\delta_2^{(b)}$ is due to expression **(v)**. Once the sufficient condition on b specified in section 4.4.2 is satisfied we are guaranteed non-positivity of $E[\frac{1}{||X||^4}\mathbb{I}_{\frac{p-2}{||X||^2} \leq 1}]$. For smaller $||\theta||^2$ it is clear that $\frac{1}{||X||^4}$ would be larger - hence we get a greater risk/average loss difference between the two estimators at these smaller values of $||\theta||^2$.

5 Norms

The loss function for the James Stein estimator was originally defined as

$$L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||^2 \tag{6}$$

which is the norm on the Euclidean space, referred to as L2 in the following graphs, under this loss function the James Stein estimator makes the naive estimator inadmissible, but how does the behaviour of the two estimators change if the loss function is redefined as

$$L(\theta, \hat{\theta}) = \sum_{i=1}^{p} |\theta_i - \hat{\theta}_i| \tag{7}$$

which is the absolute error norm, referred to as L1 in the following graphs.

We want to see if using the absolute error norm over the Euclidean norm will lead to the naive being a better estimator over the James Stein estimator, i.e. does redefining the norm change the overall results.

5.1 Zero θ

To begin we assume that $X_1, ..., X_p \sim$ N(0,1), and we want to investigate how L2 and L1 behave when varying p, whilst also comparing the naive estimator to the James Stein. We use that

$$E[X_i] = 0 \tag{8}$$

as the James Stein estimator works best when θ is close to zero.

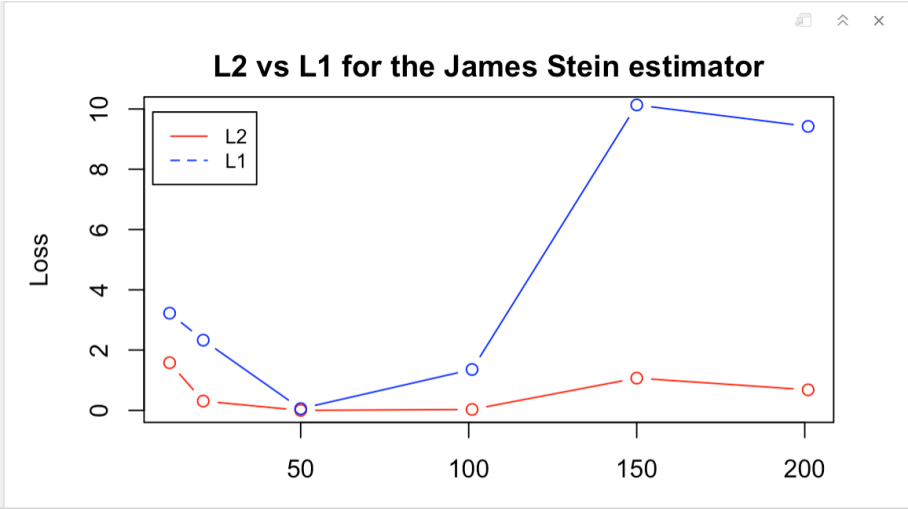


Figure 17

The graph in figure 17 shows us that for the James Stein estimator we always get a better result for the Euclidean norm compared to the absolute error norm, however there is little difference when p is small but as p increases L2 is the much better estimator for the James Stein estimator.

The graph in figure 18 gives us a contrary result compared to the James Stein estimator, as for the naive estimator using the absolute error norm gives a lower loss value for p, so for expectation close to zero the two estimators prefer different norms.

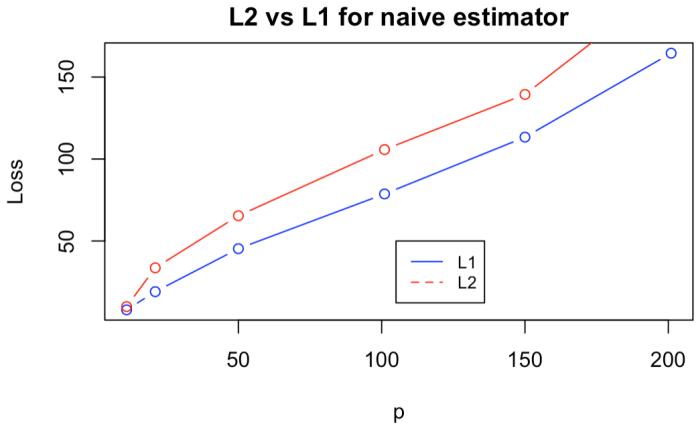


Figure 18

The figure 19 graph of differences between the losses of the James Stein and naive estimator for the different norms clearly shows that regardless of the norm chosen to define the loss function, for a lower θ the James Stein estimator is a better estimator.

5.2 Non zero θ

In practice, the expectations of the normal distributions might not always be zero or near zero, (e.g. IQ of a group of individuals) or a binomial approximation might not have an expectation near zero, so the assumption of N(0,1) may not be extremely accurate in practice. By varying the expectation of the p normal distributions between 10 and 100, we will be able to see the performance of these estimators in a more realistic situation.

The graph in figure 20 below shows that the absolute error norm provides a better loss value than the Euclidean norm for the James Stein estimator when θ is higher.

Figure 21 shows that for the naive estimator the absolute error norm provides a loss lower than L2, so for higher expectation both estimators give a smaller loss value under L1.

The low absolute value of the differences between the James Stein estimator and naive estimator for L2 and L1 in figure 22 shows that for a higher expectation, there is little difference in the estimators regardless of the norm used.

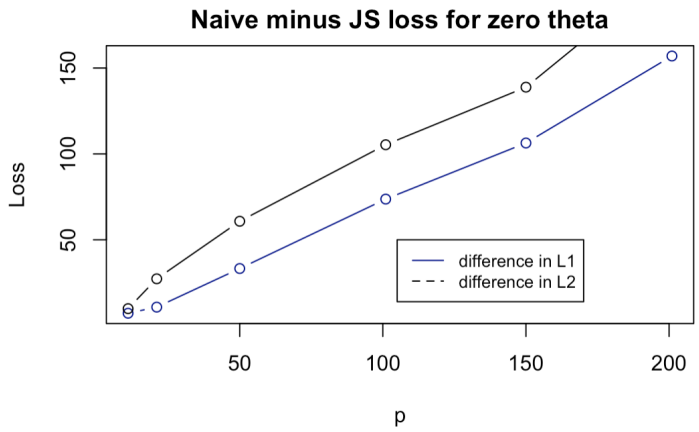


Figure 19

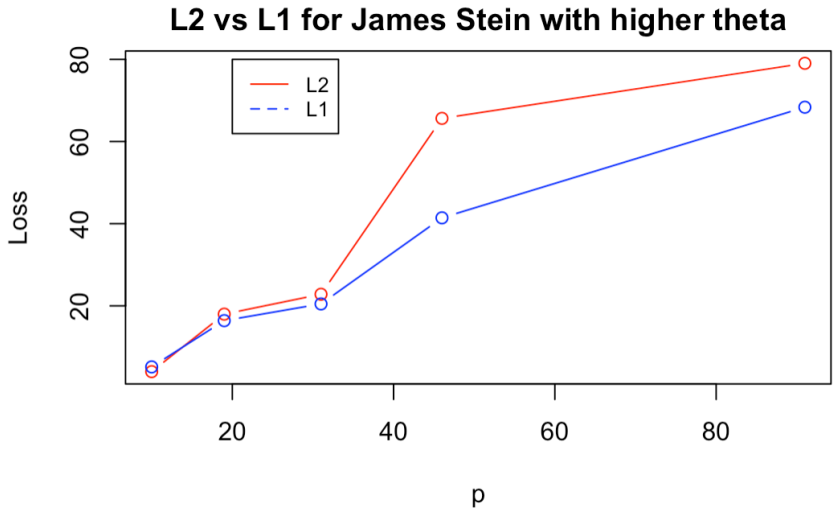


Figure 20

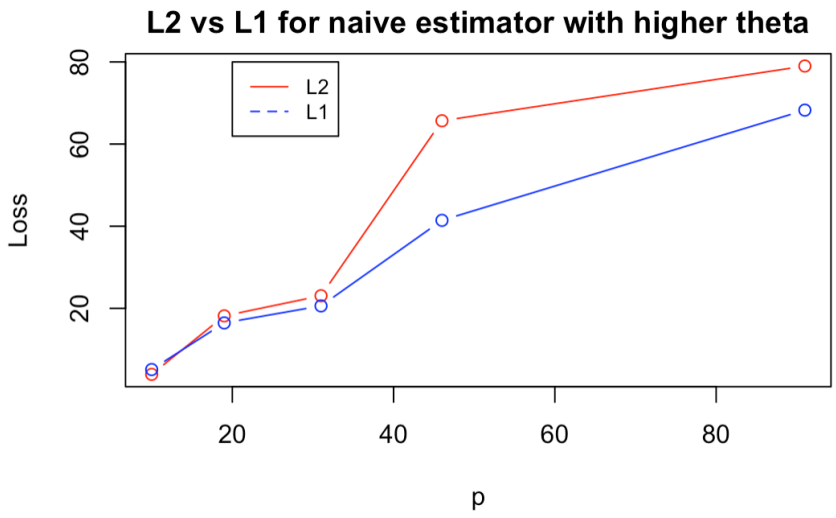


Figure 21

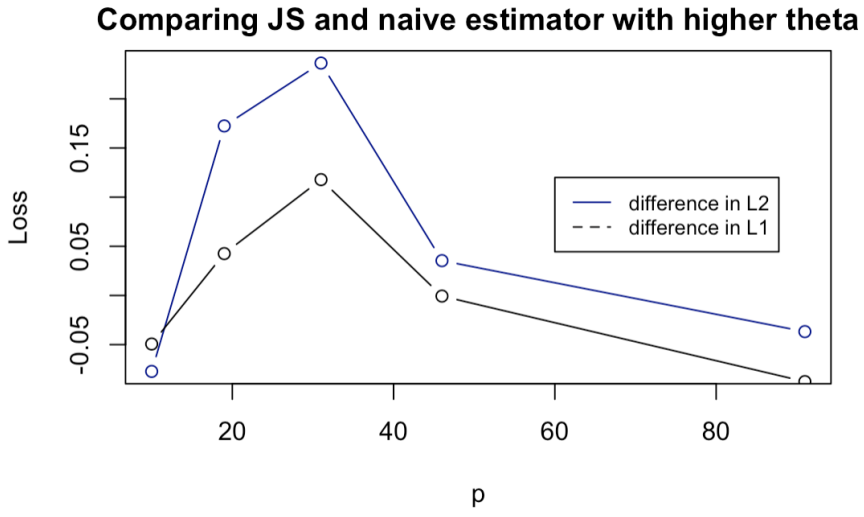


Figure 22

5.2.1 Conclusion

The only major difference in accuracy of the estimators was the loss incurred via the James Stein estimator for zero θ. Under this assumption, the Euclidean norm provided a much lower loss than the absolute error norm, in all other scenarios the two norms behaved similarly. But the graph of differences shows that changing the loss function does not really affect which estimator is more favourable - when θ is zero, the James Stein is favourable and for higher θ, both the James Stein and naive estimator provide similar results.

6 James-Stein Estimator Experiment

6.1 James-Stein Estimator

Our findings so far has indicated that the J-S estimator δ_{JS} dominates the MLE δ_0, meaning the risk of the J-S Estimator is less than or equal to the risk of the MLE estimator where the risk of $\hat{\theta}$ is given by:

$$R(\hat{\theta}, \theta) = E[L(\hat{\theta}, \theta)] \tag{9}$$

Choosing J-S estimator to shrink towards the sample mean, it gives us an estimator [8]:

$$\delta_{JS}(X) = \bar{x} + [1 - \frac{p-2}{||x_i - \bar{x}||^2}](x_i - \bar{x}) \tag{10}$$

Let c be the shrinkage factor:

$$c = 1 - \frac{p-2}{||x_i - \bar{x}||^2} \tag{11}$$

Where p is the number of unknown means and $\bar{x}$ is the grand average.

6.2 Football Example

Using the J-S estimator to estimator goal to shot ratios of particular football players in a season, and comparing it to their long term average, in theory should be a improvement to the MLE estimator. If we consider 10 professional football players, and use their 20/21 league season statistics [9] to estimate their long term probability of scoring from a shot, we can see whether the J-S estimator is an improvement to the MLE estimator. If we consider the outcome of a shot, there are two outcomes, success or failure (goal or not respectively), therefore we can model each players goal attempts to a binomial distribution:

$$x_i \sim Binomial(n, \theta_i) \tag{12}$$

Where n represents the number of shot attempts in 20/21 league season, and θ_i represents the long term probability of scoring with a shot attempt.

Name:	Shots 20/21:	Goals 20/21:	Average 20/21	Long Term Shots	Long Term Goals	Long Term Average
Lionel Messi	196	30	0.1531	1335	230	0.1723
Cristiano Ronaldo	168	29	0.1726	1345	215	0.1599
Lewnadowski	137	41	0.2993	689	156	0.2264
Mo Salah	126	22	0.1746	621	110	0.1771
Kylian Mbappe	104	27	0.2596	316	78	0.2468
Harry Kane	137	23	0.1679	884	160	0.181
Lukaku	96	24	0.25	669	110	0.1644
Benzema	123	22	0.1789	650	119	0.1831
Ibrahimovic	81	15	0.1852	425	73	0.1718
Luis Suarez	104	21	0.2019	748	167	0.2233

Figure 23: Data for Goals Shots

Looking at figure 23, The data represents the shots taken, and goals scored from the 20/21 season, and long term shots and goals scored. Due to gaps in long term data, the majority of the players data is based on the 14/15 season to 20/21 season, with the exception of Lewandowski, Ibrahimovic, which data started in 16/17 and Mbappe, which data started in 18/19. Reasons for this are due to players being in different leagues, meaning some information are not recorded in those leagues, and others being much younger than the other football players. Looking through the data of each player, we can see that the difference between the 1 season average and long term average for some is quite a considerable amount, which would come down to peak performances of a football player coming in certain years of their life. These are factors which need to be considered as they lead to the differences in the sample means and long term means.

Since our n for each Binomial distribution is very large, we can take a normal approximation to get:

$$x_i \sim Normal(\theta_i, \sigma_0^2) \tag{13}$$

where σ_0^2 is the binomial variance:

$$\sigma_0^2 = \frac{\bar{\theta}(1-\bar{\theta})}{n_i} \tag{14}$$

Where $\bar{\theta}$ is the average of all players long term averages.

$$y_i = \frac{x_i}{\sigma_i^2} \tag{15}$$

Using the transformation in (15) [10], and applying it to the J-S Estimator (10), leads to an estimator of the form [8]:

$$\delta_{JS} = \bar{x} + [1 - \frac{(k-3)\sigma_0^2}{||x_i - \bar{x}||^2}](x_i - \bar{x}) \tag{16}$$

With Shrinkage value c:

$$c = 1 - \frac{(p-3)\sigma_0^2}{||x_i - \bar{x}||^2} \tag{17}$$

Using the data from Figure 23, $\bar{x} = 0.20431$, which represents the grand average. As the Binomial Variance would be different for every player, due to the difference in the value n, we will take the mean of all variances, $\sigma_0^2 = 0.001291501$. This value is obtained by using equation (14) for each player, then finding the mean of all, where we have used $\bar{\theta} = 0.19061$, the average of all long term averages (the true mean). With $p = 10$, and $||x_i - \bar{x}||^2$ being the sum of all the differences between the sample mean and the individual goal average of the players squared, we get $||x_i - \bar{x}||^2 = 0.021020689$.
Applying all these values to find the value of c (17), gives c = 0.569923431. Now, we can apply these values to find the J-S Estimator for each footballer and compare it to the 20/21 season.

6.3 Results

Names:	20/21 Average (x_i)	J-S Estimator (JS_i)	Long Term Average:	|Long Term - JS_i|^2	|Long Term - x_i|^2
Lionel Messi	0.1531	0.1751	0.1723	0.0000080631969922	0.00036953986515307
Cristiano Ronaldo	0.1726	0.1862	0.1599	0.0006962435239025	0.00016301535078545
Robert Lewandowski	0.2993	0.2584	0.2264	0.0010260446914246	0.00530784791454047
Mo Salah	0.1746	0.1874	0.1771	0.0001049378861174	0.00000640333303539
Kylian Mbappe	0.2596	0.2358	0.2468	0.0001213164905205	0.00016332690672201
Harry Kane	0.1679	0.1836	0.1810	0.0000065721104046	0.00017193145237022
Romelu Lukaku	0.2500	0.2303	0.1644	0.0043461435137743	0.00732316376985483
Karim Benzema	0.1789	0.1898	0.1831	0.0000455803560576	0.00001776735850776
Zlatan Ibrahimovic	0.1852	0.1934	0.1718	0.0004688981997707	0.00018010926471775
Luis Suarez	0.2019	0.2029	0.2233	0.0004131278833810	0.00045535100742653

Figure 24: J-S Estimator

Comparing J-S Estimator and 20/21 average for each football player, we can see that $|\theta - \delta_{JS}|^2$ is smaller than $|\theta - x_i|^2$ for 6 out of the 10 football players, (θ representing the true value for each player) and slightly worse off for the 4 remaining players, this can be seen from columns 5 and 6 in figure 24, and the bars in figure 25. Reasons for the δ_{JS} estimator being an improvement for some players and not others can come down to all the factors which affect a players performance in a particular season, the main one being players peak in certain ages meaning they are likely to perform better. For example, two players in particular which the δ_{JS} estimator was less accurate than the MLE 20/21 average was Cristiano Ronaldo and Zlatan Ibrahimovic. Being two of the oldest players on the list, they are likely to be out of their prime. Further more, for Benzema & Mo Salah, the other two players which the δ_{JS} estimator was not an improvement, the differences in the two estimators were not significant, as both season average and J-S was relatively close to their true value.

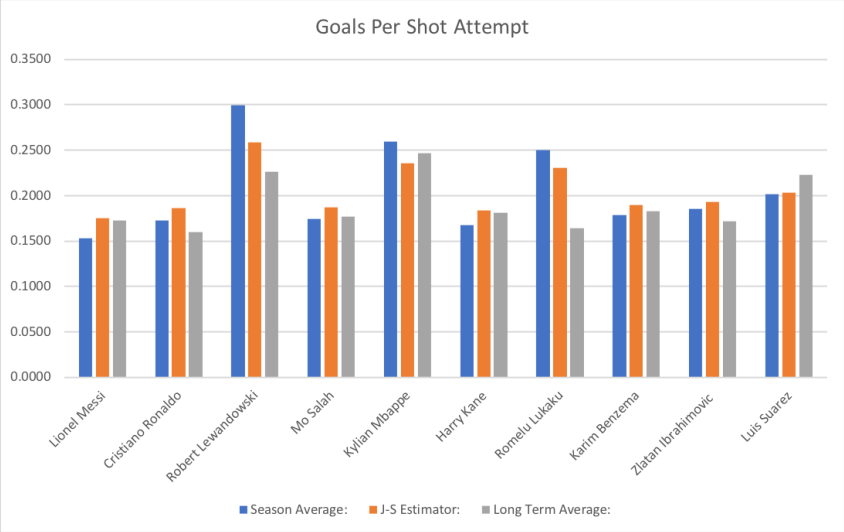


Figure 25: Bar Chart Comparing J-S, MLE, True Value

6.4 Conclusion

In order to see if the J-S estimator is an improved estimator to the MLE, we can look at their Euclidean norms, $||\delta_{JS} - \theta||^2$ and $||x_i - \theta||^2$, with the smaller value being on average closer to the true value. To do this, we will look at:

$$Error_{js} = ||\delta_{JS} - \theta||^2 \tag{18}$$

$$Error_{sa} = ||\delta_{JS} - \theta||^2 \tag{19}$$

Where equation (18) represents the error for the δ_{JS} estimator and (19) represents the error for the 20/21 Season Average. Using the square of Columns 5 and 6 from Figure 22, and substituting the values into (18) and (19), we get:

$$Error_{js} = 0.007236928 \tag{20}$$

$$Error_{sa} = 0.014158456 \tag{21}$$

From (20) and (21), we can see that the δ_{JS} estimator is an improvement as an estimator for the true value of the long term mean θ, over the MLE for the 20/21 season. This therefore concludes that the δ_{JS} Estimator is a better estimator for a footballers goal to shot ratio in a league season.

References

[1] G A Young and RL Smith. *Essentials of Statistical Inference*, volume 16. Cambridge University Press, 2005.

[2] Marc S Paolella. *Intermediate probability: A computational approach*. John Wiley & Sons, 2007.

[3] Theodore Wilbur Anderson et al. Introduction to multivariate statistical analysis. 1958.

[4] Yuzo Maruyama. Minimax admissible estimation of a multivariate normal mean and improvement upon the James-Stein estimator. *Doctor Thesis, Faculty of Economics, University of Tokyo*, 2000.

[5] Yuzo Maruyama. Improving on the James-Stein estimator. *Statistics & Risk Modeling*, 17(2):137–140, 1999.

[6] Peter Yi-Shi Shao and William E Strawderman. Improving on the James-Stein positive-part estimator. *The Annals of Statistics*, pages 1517–1538, 1994.

[7] Abdenour Hamdaoui. On shrinkage estimators improving the positive part of James-Stein estimator. *Demonstratio Mathematica*, 54(1):462–473, 2021.

[8] B Efron. Morris c. *Stein's paradox in statistics. Scientiijc American*, 236:119–127, 1977.

[9] Infogol - live football scores, stats, fixtures, results and tips. `https://www.infogol.net/en`.

[10] Isaac Matejin. James-Stein estimation in baseball. 2018.

Appendix A Code For Figures

Listing 1: Example code for figure 1-2

```
library(MASS)
library(caTools)
library(latex2exp)

b <- function(w, p, t){w^(p/2 - 2)*exp(t/2 * w)}

data_vec = c()
for(t in 0:50){
  t_val = t
  p_val = 3
  z <- function(t, p) {sapply(t, function (t_i)
  integrate(b, lower = 0, upper = 1,
  p=p_val, t = t_val)$value)}
  y = p_val - ((p_val-2)^2/(2*exp(t/2))) * z(1)
  data_vec = c(data_vec, y)
}

mean_vec = c()
for(t in seq(0,5,0.5)){
  jsloss_vec = c()
  for(i in 1:20000){
    p = 3
    mu = rep(t, p)
    x = mvrnorm(n=p, mu = mu, Sigma = diag(p))[1,1:p]
    js_coeff = 1 - (p-2)/(sum(x * x))
    js_diff = js_coeff * x - mu
    loss = norm(js_diff, type = "2")^2
    jsloss_vec = c(jsloss_vec, loss)
  }
  mean_vec = c(mean_vec, mean(jsloss_vec))
}
xfit = 3 * seq(0,5,0.5) * seq(0,5,0.5)
plot(0:50, data_vec, lwd = 2,
type = 'l', col = 'blue',
xlab = TeX(r'($||\theta||^2$)'),
ylab = "Risk", ylim = c(2, 3))
points(xfit, mean_vec)
legend('bottomright',
legend=c("Simulated", "Theoretical"),
pch = c(1, NaN), lty = c(NaN, 1),
col = c('black', 'blue') )
```

Listing 2: Example code for figure 6-8

```
data_vec = c()
mu = rep(3, p)
p = 10
diff_vec = c()
for(i in 1:5000){
  x = mvrnorm(n=p, mu = mu, Sigma = diag(p))[1,1:p]
  js_coeff = 1 - (p-2)/(sum(x * x))
  js_diff = js_coeff * x - mu
  jsloss = norm(js_diff, type = "2")^2
  nloss = norm(x - mu, type = "2")^2
  diff_vec = c(diff_vec, jsloss - nloss)

len = 1000
h = hist(diff_vec, breaks = 100)
xfit = seq(min(diff_vec), max(diff_vec), length = len)
a = p - 2
c = 2*a + xfit
xmod = a^2/(c) + a^2*(sum(mu * mu))/(c^2)
ncp = (1 - a/c)^2 * sum(mu * mu)
yfit = c()
for(i in 1:len){
  if (c[i] > 0) {
  val = 1 - pchisq(xmod[i], df = p, ncp = ncp[i])
  } else {
  val = pchisq(xmod[i], df = p, ncp = ncp[i])
  }
  yfit = c(yfit, val)
}

hist(diff_vec, prob = TRUE, breaks = 17, xlab = "Difference in Loss",
xlim = c(-5, 5), ylim = c(0, 0.4))
lines(xfit[1:len-1], fd_y, lwd=2, col='blue')
```

Note that many figures in the section 'Distribution of Loss Function' are created by code similar to Listing 2.

Listing 3: Base code for figures 14-16

```
library(MASS)
p = 100 #number of random variables
improv_est <- function(x,b){
  sub_coeff = 1- (p-2)/(sum(x * x))
  coeff = max(0, sub_coeff)
  if (sub_coeff >= 0){
    return((coeff + b*(1/sum(x * x))^2)*x)
  }
```

```
  else{
    return(0)
  }
}
sim_run <-function(mu){
  jsloss_vec = c()
  nloss_vec = c()
  imploss_vec = c()
  b = 2*(p-2)
  for(i in 1:500){

    x = mvrnorm(n=p, mu = mu, Sigma = diag(p))[1,1:p]
    js_coeff = max(0, 1 - (p-2)/(sum(x * x)))
    js_diff = js_coeff * x - mu
    loss = norm(js_diff, type = "2")^2
    jsloss_vec = c(jsloss_vec, loss)
    nloss_vec = c(nloss_vec, norm(x - mu, type = "2")^2)
    imp_diff = improv_est(x, b) - mu
    imp_loss = norm(imp_diff, type = "2")^2
    imploss_vec = c(imploss_vec, imp_loss)
  }
  mnloss = mean(nloss_vec)
  mjsloss = mean(jsloss_vec)
  mimploss = mean(imploss_vec)
  return(c(mnloss, mjsloss, mimploss))
}
meanvec = c()
meanjsvec = c()
meanimpvec = c()
for(i in 2:20){
  mu = runif(n = p, min = -i, max = i)
  #generate a random mean vector uniformly
  fullVec = sim_run(mu)
  meanvec = c(meanvec, fullVec[1])
  meanjsvec = c(meanjsvec, fullVec[2])
  meanimpvec = c(meanimpvec, fullVec[3])
}
xrange = c(2:20)
```

Listing 4: Plotting figures 14-16

```
plot(xrange, meanjsvec, type="l",
col="blue",xlab="i", ylab="average loss (500 trials)", lwd=2)
lines(xrange, meanvec, type="l", col="red")
legend(10,70 ,c("+ve JS", "X"), lwd=c(2,1),
col=c("blue", "red"), y.intersp = 1.5)
```

```
plot(xrange, meanimpvec, type="l",
col="green",xlab="i", ylab="average loss (500 trials)", lwd=2)
lines(xrange, meanvec, type="l", col="red")
legend(10,70 ,c("Improved +ve JS", "X"), lwd=c(2,1),
col=c("green", "red"), y.intersp = 1.5)

diff_vec = meanjsvec - meanimpvec
plot(xrange, diff_vec, type="l",
xlab="i", ylab="avg. loss +ve JS - avg. loss improved +ve JS")
```