ASYMPTOTIC CONSISTENCY OF THE JAMES-STEIN SHRINKAGE ESTIMATOR

BY

MUNGO ALEX SAMUEL

A dissertation submitted to The University of Zambia in partial fulfilment of the requirements of the degree of Master of Science(MSc) in Statistics

THE UNIVERSITY OF ZAMBIA
LUSAKA

2019

ABSTRACT

This study establishes the asymptotic consistency of the James-Stein shrinkage estimator (JSSE) $\hat{\boldsymbol{\beta}}_n^*$ of some parameter $\boldsymbol{\theta}$, which is obtained by shrinking a maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_n$. Our shrinking strategy involves creating a subspace to shrink to by partitioning the parameter space Ω into two components. From this partition, our interest is in the whole parameter space Ω and one of the partitioned components which we refer to as the sub-parameter space Ω_o. Due to this partition, we have another maximum likelihood estimator for the parameter in the sub-parameter space Ω_o and we call it the restricted maximum likelihood estimator (RMLE) $\tilde{\boldsymbol{\theta}}_n^o$ which is the shrinkage target. Therefore in this framework we consider three estimators, the JSSE $\hat{\boldsymbol{\beta}}_n^*$, RMLE $\tilde{\boldsymbol{\theta}}_n^o$ and MLE $\hat{\boldsymbol{\theta}}_n$. We use Hansen's approach to derive the asymptotic distribution of the James-Stein shrinkage estimator (JSSE). With regularity conditions for the MLE considered, we obtain the asymptotic distribution as a multivariate normal distribution with some shrinkage effect values. We use the Taylor's theorem and limit theorems on this distribution to show that the James-Stein shrinkage estimator is asymptotically consistent as long as the initial (MLE) estimator is consistent. The asymptotic distributional bias (ADB) is evaluated for each of the three estimators. Results show that the JSSE $\hat{\boldsymbol{\beta}}_n^*$ and RMLE $\tilde{\boldsymbol{\theta}}_n^o$ are asymptotically biased while the unrestricted MLE $\hat{\boldsymbol{\theta}}_n$ is asymptotically unbiased. Furthermore we show that the JSSE $\hat{\boldsymbol{\beta}}_n^*$ is also asymptotically efficient. Lastly, simulation plots are done in R for the mean squared error (MSE) for sample size values of 30, 2000, 8000, 50000 and 100000 using the R multivariate model to compare the unbiased estimator (MLE) and the James-Stein shrinkage estimator in order to show lower MSE of the latter. Results also show that the James-Stein shrinkage estimator converges faster compared to the MLE. We conclude from the study that the James-Stein shrinkage estimator (JSSE) obtained by shrinking a maximum likelihood estimator (MLE) is asymptotically consistent and efficient.

Keywords: Convergence, Efficiency, Maximum likelihood estimator, Mean squared error, Shrinkage.

DECLARATION

The work described in this Master of Science in Statistics dissertation was carried out under the supervision of Dr V. M. Nawa, Department of Mathematics and Statistics, University of Zambia, Lusaka.

I, the undersigned, hereby declare that the Master of Science in Statistics dissertation represents my original work and has not otherwise been submitted in any form for any degree or diploma to this or any other University. Any other work done by others has been acknowledged and referenced accordingly.

Signed:

.....
MUNGO ALEX SAMUEL (Student)

Date: .....

COPYRIGHT

@2019

APPROVAL

This dissertation of Mungo Alex Samuel has been approved as fulfilling the requirements or partial fulfilment of the requirements for the award of a Master of Science in Statistics by the University of Zambia.

Examiner 1 .....

Signature: .....

Date: .....

Examiner 2 .....

Signature: .....

Date: .....

Examiner 3 .....

Signature: .....

Date: .....

Chairperson Board of Examiners .....

Signature: .....

Date: .....

Supervisor .....

Signature: .....

Date: .....

DEDICATION

I dedicate this work to my parents Zacharia Beek Mungo and Addie Luwe Mungo, my brother Ackim and to my sisters. This is what your patience generated.

ACKNOWLEDGEMENTS

First of all, I thank the almighty God for this precious opportunity and for granting me good health throughout this study.

Secondly, I thank my supervisor Dr V. M. Nawa for the wonderful guidance, tireless support and patience he had shown during my entire study. Am also grateful to Dr A. M. Ngwengwe for equipping me with mathematical knowledge and skills, and the Head of Department Dr I. D. Tembo for his full support to solve the challenges I faced during the whole period of my study.

Thirdly, I thank my family and my friends for the tireless encouragement throughout my study which provided me extra energy to successfully complete my dissertation. I am grateful to my friend Chota Monday for providing me a suitable environment during my stay at campus in Dag 1.

Finally, I thank the Department of Mathematics and Statistics at the University of Zambia for according me a chance to study, the financial support and wonderful guidance I received throughout my study.

NOTATION AND CONVENTIONS

X	The vector of observed variables
n	Sample size value
$\bar{X}$	The sample mean
S^2	Sample variance
$\mathbb{R}^p$	The set of elements of p-dimensional
$\boldsymbol{\theta}$	Parameter of interest
$\boldsymbol{\theta}_o$	Assumed true value
σ^2	Population variance
μ	Population mean
Ω	Parameter space with elements in $\mathbb{R}^p$
Ω_0	Sub-parameter space contained in the parameter space Ω
$\hat{\boldsymbol{\theta}}_n$	Maximum likelihood estimator (`MLE`) for the parameter space Ω
$\tilde{\boldsymbol{\theta}}_n^o$	Restricted maximum likelihood estimator (`RMLE`) for the sub-parameter space Ω_o
$\hat{\boldsymbol{\beta}}_n^*$	The James-Stein shrinkage estimator obtained by shrinking the `MLE`
$\mathbf{A}^\top$	The transpose of the matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	The inverse of the matrix $\mathbf{A}$
$g'(\boldsymbol{\theta})$	The derivative of the function g of $\boldsymbol{\theta}$
$f_{\boldsymbol{\theta}}(X)$	The density function of the observations on X indexed with parameter $\boldsymbol{\theta}$
$\mathbf{L}(\boldsymbol{\theta}, X)$	The likelihood function
$\ell(\boldsymbol{\theta}, X)$	The log likelihood function
$\mathbf{I}(\boldsymbol{\theta}, X)$	The information function
$\mathbf{J}(\boldsymbol{\theta}, X)$	The fisher information
$\mathbf{S}(\boldsymbol{\theta}, X)$	The score function
$R(\boldsymbol{\theta}, X)$	The risk of X with respect to the parameter $\boldsymbol{\theta}$
$\mathbf{b}_{\boldsymbol{\theta}}(\mathbf{T})$	The bias of the estimator $\mathbf{T}$ with respect to the parameter $\boldsymbol{\theta}$
$\mathbf{V}$	Variance covariance matrix
Z	The Standard Normal distribution
χ_p^2	The Chi-square distribution with p degrees of freedom
$N_p(\boldsymbol{\theta}, \mathbf{V})$	The p-multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and variance $\mathbf{V}$
$\Pr(X)$	The probability of X
$A \subset B$	A is contained in B
$x \in B$	x is an element of B
$X_n \longrightarrow_p C$	X_n converges in probability to C
$X_n \longrightarrow_d X$	X_n converges in distribution to X
$X \sim Z$	X follows the standard normal distribution Z
$\mathbb{E}_{\boldsymbol{\theta}}[X]$	The expectation of X with respect to the parameter $\boldsymbol{\theta}$
$\lvert\mathbf{A}\rvert$	The norm of the matrix $\mathbf{A}$
$\lim_{n \longrightarrow \infty} X_n$	The limit of X_n as n approaches ∞
$\max_{\boldsymbol{\theta} \in \Omega}$	The maximum of $\boldsymbol{\theta}$ in the parameter space Ω
$\sum_{i=1}^{n} X_i$	The summation of the X_i's from $i = 1$ up to n
$\sup_{\boldsymbol{\theta} \in \Omega}$	The supreme of $\boldsymbol{\theta}$ in the parameter space Ω
$\therefore$	Therefore.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

The focus of this dissertation is to investigate the asymptotic behaviour of the James-Stein shrinkage estimator (JSSE) $\hat{\boldsymbol{\beta}}_n^*$ obtained by shrinking a maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_n$, more especially its consistency as $n \longrightarrow \infty$ when we have observed variables which follow a p-multivariate normal distribution $N_p(\boldsymbol{\theta}, \mathbf{V})$. We analyse the asymptotic consistency of this shrinkage estimator from its asymptotic convergence to some normal distribution which results from the asymptotic normality of the maximum likelihood estimator (MLE). This is presented in two main parts, the analysis of the consistency of the James-Stein shrinkage estimator and the rate at which it converges. Though the shrinkage estimator is biased it is important in estimation because there is a realisation that efficiency dominates all these other properties. In the report [16], Efron shows how bias dominates unbiasedness in estimation.

1.1. Background

The idea of shrinking an estimator came in 1956 when Stein [51] established that we can reduce the mean squared error (MSE) of an estimator if we give up a little on bias of a given estimator. This means that the efficiency of the new estimator is desirable in a way it estimates the "true value". Therefore, this works well when the number of parameters is more than two ($p \geq 3$) called the "James-Stein classic condition." With this condition, when a MLE is shrunk, it does well in terms of the MSE compared to the initial estimator (MLE). The new estimator obtained is closer to the "true value" than the initial estimator and it is called the **James-Stein shrinkage estimator**. Hansen [28] shows that a very simple shrinkage modification can achieve substantially smaller asymptotic risk (weighted mean squared error) making the conventional MLE inefficient. The magnitude of the improvement depends on the distance between the "true" parameter value and a parametric restriction. If the distance is small then the reduction in risk can be quite substantial. Even when the distance is moderately large the reduction in risk can be significant. With all these modifications and restrictions, we ask ourselves if this desirable shrinkage estimator maintains its properties as the sample size n increases without bound.

Therefore, in this study we concentrate on the asymptotic behaviour of the James-Stein shrinkage estimator specifically its asymptotic consistency. If we have a sequence of estimators, then we check if the shrinkage estimator is consistent for the estimates. We proceed by investigating the asymptotic distribution of the James-Stein shrinkage estimator through the whole parameter space Ω and the restricted parameter space Ω_o. To allow continuity, we do this within the neighbourhood $n^{-\frac{1}{2}}$ supported by a constant h. We then review Hansen's approach in [27] and [28]

for deriving the asymptotic distribution of the James-Stein shrinkage estimator. Furthermore, we study the convergence of the shrinkage estimator as the shrinkage distance diverge.

1.2. Problem Statement

A lot of work has been done on the development and properties of the James-Stein shrinkage estimator. However, not much work has been done on the asymptotic behaviour of this type of a shrinkage estimator, especially on the asymptotic consistency when we have a multivariate normal distribution. Therefore, this study will investigate the consistency of the James-Stein shrinkage estimator in order to establish the asymptotic behaviour of the estimator. In the study we review Hansen's approach of finding the asymptotic distribution and then use the asymptotic distribution to study the asymptotic behaviour, thus the asymptotic consistency of the James-Stein shrinkage estimator.

1.3. Aim of the Study

To study the asymptotic behaviour of the James-Stein shrinkage estimator, more especially the asymptotic consistency using concepts generated from the development of the asymptotic distribution of this estimator.

1.4. Research Objectives

1. Construct the James-Stein shrinkage estimator using the `MLE` which follows a p-multivariate normal $N_p(\boldsymbol{\theta}, \mathbf{V})$.

2. Review the approach used by Hansen [28] to find the asymptotic distribution of the generalised James-Stein shrinkage estimator.

3. Check the asymptotic consistency of the constructed James-Stein shrinkage estimator using the asymptotic distribution by employing the limit and large sample theorems.

4. Check if the James-Stein shrinkage estimator is asymptotically efficient.

1.5. Research Questions

1. Are we able to reduce the mean squared error (`MSE`) of the `MLE` $\hat{\boldsymbol{\theta}}_n$ to have a James-Stein shrinkage estimator?

2. Is a sequence of James-Stein shrinkage estimators consistent?

3. Are we able to review Hansen's [28] approach of finding the asymptotic distribution of the James-Stein shrinkage estimator in order to use it?

4. Is the James-Stein shrinkage estimator asymptotically efficient?

5. What development does shrinking bring on a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$?

1.6. Significance of the Study

This study is important because the theory of shrinkage techniques plays an important role in developing efficient statistical estimators which play a key role in statistical inference. This technique is also important in regularising ill-posed inference problems by minimising the mean squared error (MSE). When error is minimised in an estimator, it implies that the precision of the estimator is improved and this further improves statistical decision making. Error reduction remains vital in estimation since error management is always a factor in statistical inference. Therefore, this study will increase literature on our knowledge and understanding of the asymptotic distribution and asymptotic behaviour of the James-Stein shrinkage estimator especially its consistency which ensures efficiency in statistical inference. This will further enable us understand the stability of this shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ when the sample size value n is large.

1.7. Organisation of the Dissertation

The work is organised as follows; Chapter 2 presents literature on some of the major contributions on shrinkage estimation. Chapter 3 is the methodology and also consists of preliminaries. In this chapter we first present the methodology used in our study and then introduce basic concepts of shrinking an estimator and how to achieve a good shrinkage estimator. Necessary definitions and results useful for our study are also presented. Chapter 4 presents an analysis on the parametric structure and the asymptotic distribution, here we formulate a statistical model and then come up with the form of the James-Stein shrinkage estimator which is used in the main results. In the last part of Chapter 4 we derive the asymptotic distribution of the James-Stein shrinkage estimator using Hansen's approach in [28] and then analyse it further for easy use in the results. Chapter 5 presents the main results. The first section show the consistency of the maximum likelihood estimator (MLE). In the second section we show theoretically that the shrinkage estimator is asymptotically consistent by using the asymptotic distribution presented in the last part of Chapter 4. The asymptotic distributional bias for the three estimators in play are also evaluated, and then we show that the shrinkage estimator is asymptotically efficient. The forth section of Chapter 5 analyses the rate of convergence of the James-Stein shrinkage estimator by linking up asymptotic consistency and asymptotic convergence using concepts of that used on the maximum likelihood estimator (MLE). The last part of Chapter 5 shows simulation plots produced using the statistical package R. The graphs show simulation plots which compare the James-Stein shrinkage estimator (JSSE) and maximum likelihood estimator (MLE) with their respective mean squared error (MSE) values. These simulation plots are produced using R version 3.1.2 by considering different sample size values. In Chapter 6 we give a discussion on the results of the study. The last chapter presents a conclusion and recommendations for the whole study. The appendix and bibliography are presented at the end.

Results, definitions, lemmas, corollaries and theorems are numbered successively. By Lemma 4.2.3 we mean Lemma 3 of Section 2 of Chapter 4. The symbol □ is used to indicate that the proof has ended.

Chapter 2

LITERATURE REVIEW

Fisher [18] proposed a formal method of finding a maximum likelihood estimator and its asymptotic distribution. Later on, he showed that under smoothness conditions, the maximum likelihood estimator is the most efficient with the lowest asymptotic squared mean error among all asymptotically normally distributed estimators. This theory by Fisher is vital for this study because all the results we get on the James-Stein shrinkage estimator totally depend on the maximum likelihood estimator (MLE).

2.1. Preview on Shrinkage Estimation

The literature on shrinkage estimation is enormous, so we only mention a few of the most relevant contributions. Stein [51] set the genesis of the development of minimax shrinkage estimators of the multivariate normal mean under quadratic loss function and suggested that they were more efficient than the maximum likelihood estimator by observing that a p-multivariate MLE is inadmissible when the dimension exceeds two. Working with his student James, they developed an estimator they called the James-Stein shrinkage estimator which has reduced mean squared error though biased. In [35], James and Stein used shrinking techniques to reduce the performance risk of the ordinary least squares (OLS) estimator in a multivariate linear regression model. In this study they constructed a James-Stein shrinkage estimator from the OLS estimator and showed that the shrinkage estimator had lower risk compared to the original estimator. Baranchik [3] showed that the positive part James-Stein shrinkage estimator has lower risk compared to a James-Stein shrinkage estimator which covers both the negative and positive parts. Berger in [5] gave a discussion on selecting a minimax estimator of a multivariate normal mean by considering different forms of the James-Stein type estimators. Stein [52] used shrinking techniques to estimate the mean of the multivariate normal distribution. In [9] Carter and Ullah constructed the sampling distribution and F-ratios of James-Stein shrinkage estimators obtained from the OLS estimator. In the report [16], Efron justifies the James-Stein shrinkage estimator by analysing its bias against the unbiasedness of the MLE. The MLE has a lot of good properties but since the James-Stein shrinkage estimator dominates it in terms of the MSE, he concludes that it is worthy to study the JSSE. Due to this fact, the JSSE is preferred to the MLE in practical applications. In [47] and [48] Oman developed estimators that shrink p-multivariate MLEs towards linear subspaces.

George [22] proposed minimax multiple shrinkage estimators that allow multiple specifications for selection of a set of targets which shrink a given estimator. In [20], Geyer looked at the asymptotics of constrained M-estimators which also fall in the class of shrinkage estimators. Stein-type shrinkage is related to model

averaging, in fact, in linear regression with two nested models, the Mallows model averaging (MMA) estimator of Hansen [26] is precisely a Stein-type shrinkage estimator. Hansen [27], derived the asymptotic distribution of a generalised James-Stein shrinkage estimator given a maximum likelihood estimator. In his study he uses a positive part James-Stein shrinkage estimator to develop this asymptotic distribution. Hansen in the study [30] showed that we can shrink an estimator towards any subspace either linear or non linear and showed that a positive James-Stein shrinkage estimator is efficient and has lower risk than the MLE in the study [28].

In all these studies on shrinkage estimators, little attention has been paid to the asymptotic behaviour of the James-Stein shrinkage estimator especially the asymptotic consistency and efficiency of the estimator. Asymptotic consistency is investigated in relationship with the increase, decrease and divergence of shrinking distance with close analysis on the effect of shrinking the MLE. Therefore, in this framework we study the asymptotic behaviour of the James-Stein shrinkage estimator, check if it is consistent for $\boldsymbol{\theta}$ when we have a p-multivariate normal distribution $N_p(\boldsymbol{\theta}, \mathbf{V})$, and then analyse the asymptotic efficiency and rate of convergence. This gives us a clear picture on the asymptotic properties of the MLE which are preserved when we shrink it. The shrinkage estimator considered is constructed using Hansen's approach in [27] for easy derivation of the asymptotic distribution.

2.2. Asymptotic Distributional Risk (ADR)

We now discuss the literature on asymptotic distributional risk of the maximum likelihood estimator and the James-Stein shrinkage estimator. We exclude the restricted maximum likelihood estimator from the discussion on asymptotic risk because we are more interested in comparing the risk of the MLE $\hat{\boldsymbol{\theta}}_n$ and the JSSE $\hat{\boldsymbol{\beta}}_n^*$ though in some cases we still refer to it. Our focus is also on evaluating the improvement shrinking brings on an estimator, hence we consider literature on the comparison of the initial estimator and the estimator we obtain after shrinking, not the shrinkage target. We therefore consider different results in some selected studies on the asymptotic risk of the James-Stein shrinkage estimator.

We begin by considering the study in [1]. Ahmed et. al in [1] and Hossain et. al [33], begin the study of asymptotic risk by assuming that for the estimator $\mathbf{T}_n$ of $\boldsymbol{\theta}$, the asymptotic distribution function of $\mathbf{T}_n$ under $\{\mathbf{P}_n\}$ exists and is given by

$$\mathbf{F}(X) = \lim_{n \longrightarrow \infty} \Pr\left(\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \leq X | \mathbf{P}_n\right) \tag{2.1}$$

where $\mathbf{F}(X)$ is non degenerate. Then they define the asymptotic distributional risk of $\mathbf{T}_n$ as

$$\mathbf{ADR}(\mathbf{T}_n) = tr\left(\mathbf{Q} \int_{R_P} \int XX^\top d\mathbf{F}(X)\right) = tr(\mathbf{QV}) \tag{2.2}$$

where $\mathbf{V}$ is the dispersion matrix for the distribution $\mathbf{F}(X)$. Using the expression (2.2) with some other assumptions and conditions on the matrices $\mathbf{Q}$ and $\mathbf{V}$ they evaluate the asymptotic risk estimates. From the estimated values they conclude that as the sample size value n grows large $R\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_n^o\right) \leq R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n^*\right) \leq R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n\right)$. From this study we see that in terms of risk the restricted maximum likelihood

estimator has lower risk compared to the other two estimators. The James-Stein shrinkage estimator has also lower risk compared to the maximum likelihood estimator. But as $n \longrightarrow \infty$ all the three estimators will have the same estimated risk.

Analysing asymptotic risk further, we consider Section 2 of Green and Strawderman [23] where they compared the risk and bias of the James-Stein shrinkage estimator by plotting risk against bias. The graph in this study demonstrates that for smaller values of bias the risk is small and that bias increases with risk. At some point the change in bias causes only a small change in risk and as bias becomes large the risk becomes constant. Therefore from this study we conclude that as $n \longrightarrow \infty$ the risk for $\hat{\boldsymbol{\theta}}_n$ is the same or slightly less than that of the `MLE` $\hat{\boldsymbol{\beta}}_n^*$.

Another study we consider is Hannes and Pötscher [25]. In Section 2.1 of [25], Hannes and Pötscher showed that the risk of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is smaller than that of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ which follows from the proof in the studies [51], [35] and [3]. They noted that all these results show that improvement in risk of $\hat{\boldsymbol{\beta}}_n^*$ over $\hat{\boldsymbol{\theta}}_n$ is substantial when the parameter $\boldsymbol{\theta}$ is close to zero. This is the main reason why we shrink towards zero. Furthermore it is shown in the same study [25] that if $n^{\frac{1}{2}}\boldsymbol{\theta}$ is small, then $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n^*)$ is close to $R(\mathbf{0}, \hat{\boldsymbol{\beta}}_n^*) = 2$, which is substantially smaller than $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) = p$. For large $n^{\frac{1}{2}}\boldsymbol{\theta}$ (when $n \longrightarrow \infty$) the risk of $\hat{\boldsymbol{\beta}}_n^*$ is close to the risk of $\hat{\boldsymbol{\theta}}_n$. Actually, $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n^*)$ converges to p for every $\boldsymbol{\theta} \neq \mathbf{0}$ as $n \longrightarrow \infty$. Casella and Hwang in [10] discuss the risk of James-Stein estimators by constructing bounds on the risk ratio $R(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n^*)/R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n)$. They found that the positive part and the ordinary James-Stein estimators have the same bound, and that there is a difference though minimal in the asymptotic risk of the `JSSE` and `MLE`.

These studies show that to evaluate the estimate of the asymptotic risk of an estimator say $\mathbf{T}_n$ is difficult. Hannes and Pötscher in [25] state that the risk of the James-Stein shrinkage estimator is difficult to estimate because it cannot be estimated uniformly consistently over balls around the origin. Hansen in [28] and [29] states that the risk of an estimator $\mathbf{T}_n$ in general is difficult to evaluate, and may not even be finite unless $\mathbf{T}_n$ has sufficient finite moments.

With these difficulties in mind, Hansen evaluates a meaningful estimate of the asymptotic risk. Considering an estimator $\mathbf{T}_n$, according to Hansen in [28] we study the `ADR` in terms of the quadratic risk loss as $n \longrightarrow \infty$. The risk of an estimator $\mathtt{T}_n$ is its expected loss $\mathbb{E}_{\boldsymbol{\theta}}\left[\ell\left(\mathbf{T}_n - \boldsymbol{\theta}\right)\right]$. In general this expectation is not specific in its value. Therefore according to [28] to obtain a useful approximation and ensure existence, Hansen [28] uses a trimmed loss and takes limits as the sample size value $n \longrightarrow \infty$. Therefore Hansen in [28] concludes that the asymptotic trimmed risk of any estimator $\mathbf{T}_n$ satisfying a quadratic trimmed risk function defined by Hansen can be calculated. From these calculations, Hansen obtained that $R(h, \hat{\boldsymbol{\beta}}_n^*) < R(h, \hat{\boldsymbol{\theta}}_n)$ where h is the localising constant of shrinkage.

The inequality $R(h, \hat{\boldsymbol{\beta}}_n^*) < R(h, \hat{\boldsymbol{\theta}}_n)$ in [28] shows that under trimmed risk the James- Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ has lower risk compared to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$. Therefore in general, different studies considered show that even if it is not possible to evaluate the actual asymptotic risk estimate values, we conclude that the asymptotic risk for the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is less than or equal to that of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$. Hence we see that

even at an asymptotic value ($n \longrightarrow \infty$) the properties of the shrinkage estimator (JSSE) and the maximum likelihood estimator (MLE) are relatively the same. The James- Stein shrinkage estimator maintains its bias and lower risk compared to the initial estimator (MLE) we shrink. This is also discussed by Yuzo in [56].

2.3. Rate of Convergence of the Maximum Likelihood Estimator

We discuss the rate of convergence of the maximum likelihood estimator according to the analysis by Halonen in [24] and then use these concepts to analyse the rate of convergence for the shrinkage estimator (JSSE) in the forth section of Chapter 5. In [14], Cramér gave a proof of the convergence of the MLE which is very similar to Doob's [15] proof, with almost identical conditions. Cramér's proof was derived under regularity conditions of the MLE which are the same as those considered in Chapter 4 under Assumptions 4.3.1 and 4.3.2. With the likelihood equation of the function $f_\theta(X_i)$ denoted as

$$\mathbf{L}(\theta) = \prod_{i=1}^{n} f_\theta(X_i), \tag{2.3}$$

using Taylor's theorem Cramér showed that at the MLE $\hat{\theta}$,

$$\frac{1}{n}\frac{\partial}{\partial\theta}\left(\ln \mathbf{L}(\hat{\theta})\right) = B_0 + B_1(\hat{\theta} - \theta_o) + \frac{1}{2}\zeta B_2(\hat{\theta} - \theta_o)^2 = 0 \tag{2.4}$$

where θ_o is the assumed true value and $\zeta \in (-1, 1)$ depends on θ and n, where

$$B_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f_{\theta_o}(x_i)$$

and $B_0 \longrightarrow_p 0$ as $n \longrightarrow \infty$. Equation (2.3) is equated to zero in order to find the maximum when we differentiate.

$$B_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \ln f_{\theta_o}(x_i)$$

and $B_1 \longrightarrow_p -t^2 = \mathbf{I}(\theta_o)$ as $n \longrightarrow \infty$. Then B_2 is given by

$$B_2 = \frac{1}{n}\sum_{i=1}^{n} H(x_i) \longrightarrow_p \mathbb{E}_\theta\left[H(x)\right] < M \quad as \quad n \longrightarrow \infty$$

where $H(x)$ is the function of x and M is the bound for the bound of the expectation of $H(x)$. Cramér further showed that rearranging (2.3) gives

$$t\sqrt{n}(\hat{\theta} - \theta_o) = \frac{\frac{1}{t\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} f_{\theta_o}(x_i)}{-\frac{B_1}{t^2} - \zeta(\hat{\theta} - \theta_o)\frac{B_2}{2t^2}} = \xi^* \tag{2.5}$$

$\xi^* \longrightarrow_d Z \sim N(0, 1)$ as $n \longrightarrow \infty$. Since the dominator as $n \longrightarrow \infty$ becomes

$$-\frac{B_1}{t^2} - \zeta(\hat{\theta} - \theta_o)\frac{B_2}{2t^2} \longrightarrow_p -\frac{-t^2}{t^2} - 0 = 1$$

and the numerator as $n \longrightarrow \infty$ gives

$$\frac{1}{t\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} f_{\theta_o}(x_i) \longrightarrow_d N(0, 1)$$

by the central limit theorem (CLT), then from (2.5) Cramér obtained

$$t\sqrt{n}(\hat{\theta} - \theta_o) \longrightarrow_p N(0,1) \tag{2.6}$$

as $n \longrightarrow \infty$.

Lehmann in [43] used a similar argument differing by the use of a $\theta_n^* \in (\theta_o, \hat{\theta}_n)$ instead of the ζ used above to make the equality exact with identical conditions to establish the same result. Thus it is clear that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ converges in distribution to a standard normal distribution at the $\sqrt{n}$ rate. This result motivates the idea of analysing the rate of convergence of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ in order to examine whether shrinking affects the rate of convergence. We will use the arguments in [44] and [50] to determine the rate of convergence of the JSSE in Chapter 5. We further analyse the asymptotic normality of the MLE using the concept of differentiability in quadratic mean (DQM), the concept we also apply when analysing the convergence rate of the shrinkage estimator.

2.3.1 Differentiability in Quadratic Mean of the MLE

We discuss the concept of differentiability in quadratic mean (DQM) according to Halonen in the study [24]. This leads us to the concept of local area normality (LAN) which we apply on the shrinkage estimator (JSSE) to determine the order of convergence and find the rate of convergence. Le Cam in [39], discusses the assumptions needed to prove the asymptotic normality of the maximum likelihood estimates. In [41], Le Cam further established the important concept of local asymptotic normality, which implies that the log likelihood ratio for the sequence of the distribution of a maximum likelihood estimator (MLE) can be approximated by a normal distribution as the sample size value increases. Now in the study [39], Le Cam showed that for certain probability functions, the traditional conditions (assumptions) can be replaced by a simpler assumption of differentiability in quadratic mean (DQM), which implies differentiability in norm of the square root of the probability density as an element of an L^2 space. Thus it links to the concept of quadratic approximation property for the log-likelihoods known as local asymptotic normality (LAN). Halonen in [24] states the formal definition of local area normality for a distribution with one parameter and pdf $f_\theta(x)$ as follows;
if

$$\theta_n = \theta_o + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{2.7}$$

then $f_\theta(x)$ is locally asymptotically normal if

$$\ln\left(\frac{\prod_{i=1}^{n} f_{\theta_n}(x_i)}{\prod_{i=1}^{n} f_{\theta_o}(x_i)}\right) = (\theta_n - \theta_o)\sqrt{n}\sqrt{\mathbf{I}(\theta_o)}Z - \frac{n}{2}(\theta_n - \theta_o)^2\mathbf{I}(\theta_o) + O_p(1) \tag{2.8}$$

were $Z \sim N(0,1)$ and $O_p(x)$ is the probability order of x. The equation (2.7) also implies that$(\theta_n - \theta_o) = O_p\left(\frac{1}{\sqrt{n}}\right)$. If the log-likelihood function is assumed to be differentiable, Halonen in [24] showed that (2.8) becomes

$$\frac{\partial}{\partial\theta_n}\ln\left(\frac{\prod_{i=1}^{n} f_{\theta_n}(x_i)}{\prod_{i=1}^{n} f_{\theta_o}(x_i)}\right) = \sqrt{n}\sqrt{\mathbf{I}(\theta_o)}Z - n(\theta_n - \theta_o)\mathbf{I}(\theta_o) + O_p(1) = 0$$

and thus we have

$$\sqrt{\mathbf{I}(\theta_o)}\sqrt{n}(\theta_n - \theta_o) + O_p(1) = Z.$$

Le Cam in the study [41] used the concept of differentiability in quadratic mean (DQM) defined for a univariate pdf as follows;
if $f_\theta(x)$ is differentiable in quadratic mean in one dimension, where $\xi_\theta(x) = \sqrt{f_\theta(x)}$ then

$$\xi_\theta(x) = \xi_{\theta_0}(x) + (\theta - \theta_o)\Delta\theta_o(x) + r_\theta(x) \tag{2.9}$$

where

$$\|r_\theta(x)\| := \sqrt{\int_{\mathbb{R}} r_\theta(x)^2 dx} = O_p(\theta - \theta_o) \quad as \quad n \longrightarrow \theta_o.$$

Furthermore, Le Cam establishes that the DQM condition implies local asymptotic normality, which implies the rate of convergence is of the order $\frac{1}{\sqrt{n}}$ if the condition of differentiability in quadratic mean is met. According to Halonen in [24], the reason why the DQM leads to $O_p\left(\frac{1}{\sqrt{n}}\right)$ convergence without the requirements of the second derivative is that the square root of a probability density function (pdf) is an element of L^2 space with norm of 1.

Now considering a sequence of the maximum likelihood estimator (MLE), $\hat{\theta}_n$. Halonen in [24] used the Taylor's theorem and the central limit theorem (CLT) to show that this sequence satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_o) \longrightarrow_d N\left(0, \mathbf{I}(\theta_o)^{-1}\right)$$

as $n \longrightarrow \infty$. The asymptotic distribution itself is not helpful since we have to evaluate the information matrix at the value of the parameter. However, we can consistently estimate the asymptotic variance of the MLE by evaluating the information matrix at the MLE, i.e

$$\sqrt{n}(\hat{\theta_n} - \theta_o) \longrightarrow_d N\left(0, \mathbf{I}(\hat{\theta}_n)^{-1}\right) \quad as \quad n \longrightarrow \infty.$$

Therefore we have seen that the differentiability in quadratic mean (DQM) of the maximum likelihood estimator (MLE) implies local asymptotic normality (LAN) of the MLE. Hence it is of the rate of convergence of order $\frac{1}{\sqrt{n}}$ and converges at the rate $\sqrt{n}$. We use this same concept to determine the order and rate of convergence of the James-Stein shrinkage estimator in Chapter 5.

Chapter 3

METHODOLOGY

This chapter presents two sections, the methodology used and preliminaries. The first section gives a guide on how the study was investigated. The second section presents results and definitions important for the study.

3.1. Methodology Overview

A James-Stein shrinkage estimator shrunk towards a target is considered in this study. The shrinkage target is created through the construction of a restricted sub-parameter space. This is done by partitioning the main parameter space by introducing a mapping which selects the mean estimates close to zero. Therefore we will have two maximum likelihood estimators (MLEs), the first one maximises $\boldsymbol{\theta}$ in the entire parameter space Ω and we denote it by $\hat{\boldsymbol{\theta}}_n$. The second one, which is also our shrinkage target maximises $\boldsymbol{\theta}$ in the restricted sub-parameter space Ω_o, which is the subset of the entire parameter space and we denote it by $\tilde{\boldsymbol{\theta}}_n^o$, which shrinks the initial estimator $\hat{\boldsymbol{\theta}}_n$ towards zero. This means that our shrinking strategy is designed to achieve an estimator which is more efficient than the initial estimator. This is well explained in Chapter 4 where details on the parametric structure (statistical model) are given.

We derive the three asymptotic distributions for all the three estimators in play while taking care of our area of interest which is the asymptotic behaviour of the James-Stein shrinkage estimator. The asymptotic distribution of the restricted maximum likelihood estimator (RMLE) is obtained by linking the distribution of the unrestricted maximum likelihood estimator (MLE) and the shrinkage estimator as a consequence of the shrinking target.

We proceed by first defining a shrinkage estimator as proposed by James and Stein [35]. Then we highlight the development of the James-Stein shrinkage estimator with reference to the mean squared error (MSE). A number of limit and large sample theorems are stated which in the process give a good guidance to review Hansen's approach of finding the asymptotic distribution. This asymptotic distribution is used as a tool together with the limit and probability theorems to check for asymptotic consistency of the James-Stein shrinkage estimator. Then the asymptotic distributional bias for the three estimators are evaluated. We utilise the distributional bias value of the James-Stein shrinkage estimator to examine the asymptotic efficiency of the estimator. We further analyse the rate of convergence and then simulate some plots in R to compare the MSE of the initial estimator (MLE) to that of the shrinkage estimator (JSSE) using different sample size values.

3.2. Preliminaries

Here we present results which are useful for the study. Definitions, propositions, lemmas and theorems on estimators are stated, where necessary background information about a concept is also provided. This part is segmented into four sections. The first section presents definitions and results on estimators in general. The second section presents results and definitions on shrinkage estimators. The third section is on results and definitions on maximum likelihood estimators (MLE) and in the last section we consider definitions and some results on probability and convergence.

3.2.1 General Results on Estimators

We begin by giving the following definitions in order to distinguish between an estimate and an estimator in this discussion. We also need these definitions in order to define the mean squared error (MSE).

Definition 3.2.1 *A **statistic** $T(X)$, is a function of sample data which does not depend on the unknown parameter θ.*

We consider an example on a statistic before defining an estimate.

Example 3.2.2 *The following are some examples of a statistic;*

1. *Sample Mean*
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
2. *Sample Variance*
$$S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Definition 3.2.3 *An **estimate** is an observed value of a random variable called an **estimator**.*

Note 3.2.4 *Although the statistic $T(X)$ is not a function of θ, its distribution can depend on the parameter θ.*

Definition 3.2.5 *An **estimator** is a statistic used for estimating an unknown parameter. i.e an estimator is a random variable. e.g. $\bar{X}$.*

Therefore, an estimate is an observed value of a random variable (estimator). e.g. $\bar{x}$. We now define the error loss and the risk function, and introduce the concept of bias in estimators by stating the following definitions.

Definition 3.2.6 *Let $X = \{X_1, ..., X_n\}$ be a random sample. A statistic $T(X)$ is an unbiased estimator of θ if*

$$\mathbb{E}_\theta\left[T(X)\right] = \theta$$

for all $\theta \in \Omega$, where Ω is the parameter space.

Definition 3.2.7 *Given an estimator $T(X)$ for θ such that $\theta \in \Omega$. A **loss fuction** or distance between the estimator and the true value denoted by $\ell(\theta, T(X))$ is given by either*

$$\ell(\theta, T(X)) = (T(X) - \theta)^2$$

*called the **squared error** loss function, or*

$$\ell(\theta, T(X)) = \mid T(X) - \theta \mid$$

*called the **absolute value error** loss function.*

Definition 3.2.8 *When the loss function is averaged over all possible values of the data we get the **risk function** given by*

$$R(\theta, T(X)) = \mathbb{E}_\theta \left[\ell(\theta, T(X))\right].$$

Definition 3.2.9 *Given an estimator $T(X)$ of θ, the bias for $T(X)$ over θ is given by*

$$\boldsymbol{b}_\theta(T(X)) = \mathbb{E}_\theta \left[T(X)\right] - \theta$$

*where θ is the parameter. When $\boldsymbol{b}_\theta(T(X)) = 0$, then $T(X)$ is called an **unbiased estimator** of θ and when $\boldsymbol{b}_\theta(T(X)) \neq 0$, $T(X)$ is called a **biased estimator** of θ.*

The following lemma stated below is important for the study. In all the calculations of the mean squared error (`MSE`) in the study we use this lemma.

Lemma 3.2.10 *([11], p.330)*

Let $T(X)$ be an estimator of θ. Then, the mean squared error (`MSE`) of $T(X)$ is give by

$$\boldsymbol{MSE}_\theta(T(X)) = \boldsymbol{V}_\theta(T(X)) + [\boldsymbol{b}_\theta(T(X))]^2$$

where the bias is as defined in Definition 3.2.9 and $\boldsymbol{V}_\theta(T(X))$ is the variance of the estimator $T(X)$.

From Lemma 3.2.10, we note that if an estimator is unbiased then the mean squared error is equal to its variance otherwise we add the square of the bias.

We now consider the results on Stein's Lemma and identities which make it easer to evaluate expectations of means of a normal distribution, concepts we use in this study.

Lemma 3.2.11 *([52], Lemma 1)*

Let x be a standard normal $N(0,1)$ real random variable and let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be an indefinite integral of the lebesgue measurable function g', essentially the derivative of g. Suppose also that $\mathbb{E} \mid g'(x) \mid < \infty$. Then

$$\mathbb{E}\left[g'(x)\right] = \mathbb{E}\left[xg(x)\right]. \tag{3.1}$$

Proof

Let $\phi(x)$ denote the standard normal density with derivative $\phi'(x) = -x\phi(x)$. Then

$$
\begin{aligned}
\mathbb{E}\left[g'(x)\right] &= \int_{-\infty}^{\infty} g'(x)\phi(x)dx \\
&= \int_{0}^{\infty} g'(x)\left\{\int_{x}^{\infty} z\phi(z)dz\right\}dx - \int_{-\infty}^{0} g'(x)\left\{\int_{-\infty}^{x} z\phi(z)dz\right\}dx \\
&= \int_{0}^{\infty} z\phi(z)\left\{\int_{0}^{z} g'(x)dx\right\}dz - \int_{-\infty}^{0} z\phi(z)\left\{\int_{z}^{0} g'(x)dx\right\}dz \quad \textit{by Fubini's Theorem([42])} \\
&= \left(\int_{0}^{\infty} + \int_{-\infty}^{0}\right)\left[z\phi(z)\left\{g(z) - g(0)\right\}\right]dz \qquad \textit{by additive of lebesgue integretion} \\
&= \int_{-\infty}^{\infty} zg(z)\phi(z)dz \\
&= \mathbb{E}\left[xg(x)\right],
\end{aligned}
$$

hence we have (3.1). □

Lemma 3.2.12 ***(Stein's Identity)****([54], p.1444)*
Let $X \sim N(\mu, \sigma^2)$ and let $g(x)$ be such that $\mathbb{E}\left[g'(x)\right] < \infty$. Then

$$\mathbb{E}\left[g(X)(X-\mu)\right] = \sigma^2\mathbb{E}\left[g'(X)\right]$$

where μ and σ^2 are the mean and variance respectively.

Lemma 3.2.13 (***Stein's Identity; Multivariate Case***)
Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and let $g(\boldsymbol{x}) : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ such that $\mathbb{E}\left[dg_i/dx_i\right] < \infty$. Then

$$\mathbb{E}\left[(\boldsymbol{X}-\boldsymbol{\mu}).g(\boldsymbol{X})\right] = \sigma^2\mathbb{E}\left[\Delta.g(\boldsymbol{X})\right]$$

where $\Delta.g(\boldsymbol{X}) = \sum_{i=1}^{p} \frac{dg_i(\boldsymbol{x})}{dx_i}$.

The above results on Stein's Lemma and identities are also explained in [11] and [42] under Stein's estimation. In our study we also consider the generalised form presented as Lemma 3.2.13.

In the next section we present work on shrinkage estimators using the analysis from the studies [36], [22] and [49] as background concepts to use when constructing the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ in Chapter 4.

3.2.2 Shrinkage Estimators

Shrinkage estimators are dependent on properties of the initial estimator we intend to shrink. Need arises to shrink an estimator so that we improve ill-posed estimators to have a new estimator which effectively estimates the true value of the parameter. The shrinking distance or the target to which to shrink determine a good effective shrinking technique. Therefore, given any unbiased estimator $\hat{\theta}$, we have a new estimator

$$\hat{\theta}^* = \frac{1}{1+\lambda}\hat{\theta} \tag{3.2}$$

such that $\lambda > 0$ which is clearly biased. If λ is zero then $\hat{\theta}^*$ is the same with $\hat{\theta}$ and if λ is very large then $\hat{\theta}^*$ shrinks to zero. If $\lambda < 0$ then we take λ as 0.

Note 3.2.14 *To achieve desirable results of our shrinkage estimator in (3.2), a constraint is introduced on the values of λ. Thus if $\lambda < 0$ then $\lambda = 0$. So when shrinking we just shrink for values of $\lambda \geq 0$.*

Hence it is the fraction $\frac{1}{1+\lambda}$ which brings in the difference in terms of the mean squared error (`MSE`) of the two estimators as explained in [19]. We pay more attention on how the `MSE` is affected when we apply shrinking techniques with the aim of having the `MSE` reduced when it is compared to the `MSE` of the initial estimator. Therefore, calculating the `MSE` of $\hat{\theta}^*$ from (3.2) for $\lambda \geq 0$ and $T = \frac{1}{1+\lambda}\hat{\theta}$ we have

$$\begin{aligned}
\mathbf{MSE}_\theta(T) &= \mathbb{E}_\theta\left[T-\theta\right]^2 \\
&= \mathbb{E}_\theta\left[\sum_{k=1}^{P}\left(\frac{1}{1+\lambda}\hat{\theta}_k - \theta_k\right)^2\right] \\
&= \mathbb{E}_\theta\left[\sum_{k=1}^{P}\left(\frac{1}{1+\lambda}\hat{\theta}_k - \mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) + \mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \theta_k\right)^2\right] \\
&= \mathbb{E}_\theta\left\{\sum_{k=1}^{P}\left[\left(\frac{1}{1+\lambda}\hat{\theta}_k - \mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right)\right) + \left(\mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \theta_k\right)\right]^2\right\} \\
&= \sum_{k=1}^{P}\mathbb{E}_\theta\left[\frac{1}{1+\lambda}\hat{\theta}_k - \mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right)\right]^2 + \sum_{k=1}^{P}\left[\mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \theta_k\right]^2 + \\
&\sum_{k=1}^{P} 2\left(\mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right)\right)\left(\mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \theta_k\right) \\
&= \sum_{k=1}^{P}\mathbb{E}_\theta\left[\frac{1}{1+\lambda}\hat{\theta}_k - \mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right)\right]^2 + \sum_{k=1}^{P}\left[\mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \theta_k\right]^2.
\end{aligned}$$

Continuing by using the definition of variance and bias of an estimator we have

$$\mathbf{MSE}_\theta(T) = \sum_{k=1}^{P}\mathbf{V}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) + \sum_{k=1}^{P}\left[\mathbf{b}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right)\right]^2. \tag{3.3}$$

Considering the first term of (3.3) and proceeding by linearity of variance we have

$$\sum_{k=1}^{P}\mathbf{V}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) = \left(\frac{1}{1+\lambda}\right)^2\sum_{k=1}^{P}\mathbf{V}_\theta(\hat{\theta}_k) = \left(\frac{1}{1+\lambda}\right)^2\sum_{k=1}^{P}\sigma^2 = \left(\frac{1}{1+\lambda}\right)^2 P\sigma^2. \tag{3.4}$$

We now look at the second term of (3.3) which is the bias term. We proceed by using Definition 3.2.9 to have

$$\sum_{k=1}^{P}\left[\mathbf{b}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right)\right]^2 = \sum_{k=1}^{P}\left[\mathbb{E}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) - \theta_k\right]^2 = \sum_{k=1}^{P}\left[\left(\frac{1}{1+\lambda}\right)\mathbb{E}_\theta(\hat{\theta}_k) - \theta_k\right]^2, \tag{3.5}$$

using the identity

$$\frac{1}{1+t} = 1 - t + \frac{t^2}{1+t}$$

(3.5) becomes

$$\begin{aligned}\sum_{k=1}^{P}\mathbf{b}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) &= \sum_{k=1}^{P}\left[\left(1-\lambda+\frac{\lambda^2}{1+\lambda}\right)\mathbb{E}_\theta(\hat{\theta}_k)-\theta_k\right]^2 \\ &= \sum_{k=1}^{P}\left[\left(1+\frac{\lambda^2}{1+\lambda}\right)\mathbb{E}_\theta(\hat{\theta}_k)-\theta_k\right]^2 \quad \textit{by Note 3.2.14, } -\lambda<0 \textit{ implies } \lambda=0 \\ &= \sum_{k=1}^{P}\theta_k^2\left[\left(1+\frac{\lambda^2}{1+\lambda}\right)^2-2\left(1+\frac{\lambda^2}{1+\lambda}\right)+1\right] \\ &= \left(\frac{\lambda^2}{1+\lambda}\right)^2\sum_{k=1}^{P}\theta_k^2.\end{aligned}$$

Now θ_k is an unknown mean, therefore we use its estimate $\hat{\theta}_k$ so that

$$\sum_{k=1}^{P}\mathbf{b}_\theta\left(\frac{1}{1+\lambda}\hat{\theta}_k\right) = \left(\frac{\lambda^2}{1+\lambda}\right)^2\sum_{k=1}^{P}\hat{\theta}_k^2. \tag{3.6}$$

Hence using (3.3) and combining (3.4) and (3.6) we have

$$\widehat{\mathbf{MSE}}_\theta(T) = p\sigma^2\left(\frac{1}{1+\lambda}\right)^2+\left(\frac{\lambda^2}{1+\lambda}\right)^2\sum_{k=1}^{p}\hat{\theta}_k^2 \tag{3.7}$$

where the first term is the variance component and it is largest when λ is zero and the second term is the squared bias which grows when λ grows. With these considerations, the variance decomposition suggests that we can do better in terms of `MSE` than the initial estimator $\hat{\theta}$ if we select a good value of λ. The optimal for this value is

$$\lambda = \frac{p\sigma^2}{\sum_{k=1}^{p}\hat{\theta}_k^2}$$

suggested by James and Stein [35]. Therefore, a James-Stein type shrinkage estimator of the form

$$\hat{\theta}_k^* = \left(1-\frac{(p-2)\sigma^2}{\sum_{k=1}^{p}\hat{\theta}_k^2}\right)\hat{\theta}_k \tag{3.8}$$

was estabilised in [35]. In the same study James and Stein showed that the James-Stein shrinkage estimator of the form (3.8) has the `MSE`

$$\sum_{k=1}^{p}\mathbf{E}_\theta\left(\hat{\theta}_k^*-\theta_k\right)^2 \leq 2\sigma^2 \tag{3.9}$$

compared to that of the `MLE`

$$\sum_{k=1}^{p}\mathbf{V}_\theta\left(\hat{\theta}_k\right) = p\sigma^2 \tag{3.10}$$

with bias $\mathbf{b}_\theta(\hat{\theta}_k)=0$. Analysing (3.9) and (3.10) further, we note that for values of $p<2$, the `MLE` does well in terms of `MSE`. When $p=2$ both estimators have the

same MSE. When $p \geq 3$ the James-Stein shrinkage estimator does better in terms of MSE and this is the classical condition for James-Stein shrinkage estimators. Therefore, we can have another form of the James-Stein shrinkage estimator from the same type of (3.8) by replacing the term $\sum_{k=1}^{p} \hat{\theta}_k^2 \sim \chi_p^2$ by $\hat{\boldsymbol{\theta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}_n \sim \chi_p^2$ when we have p-random vectors and the covariance matrix $\mathbf{V}$. Thus we have

$$\hat{\boldsymbol{\theta}}_n^* = \left(1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}_n}\right) \hat{\boldsymbol{\theta}}_n \tag{3.11}$$

for $p \geq 3$ when the shrinkage target is zero. To have a generalised form we set a new shrinkage target other than zero according to [27]. Thus for any shrinkage target $\tilde{\boldsymbol{\theta}}_n^o$ we have a generalised James-Stein shrinkage estimator

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n - \left(\frac{p-2}{n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right)^\top \mathbf{V}^{-1} \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right)}\right) \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right) \tag{3.12}$$

where $p \geq 3$ and $\mathbf{V}$ is the covariance matrix of the initial estimator. The shrinkage target determines how good the shrinking is.

Baranchik in the study [3] showed that a positive part James-Stein shrinkage estimator of the form

$$\hat{\boldsymbol{\theta}}_n^* = \max\left(0, 1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}_n}\right) \hat{\boldsymbol{\theta}}_n \tag{3.13}$$

has lower risk compared to the James-Stein shrinkage estimator in (3.11) which get both negative and positive values. Hansen in [27] presents the form (3.13) as

$$\hat{\boldsymbol{\theta}}_n^* = \left(1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}_n}\right)_+ \hat{\boldsymbol{\theta}}_n \tag{3.14}$$

where the notation $(a)_+$ trims a to be always positive. Hence from (3.12), the positive part generalised James-Stein shrinkage estimator is given by

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n - \left(\frac{p-2}{n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right)^\top \mathbf{V}^{-1} \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right)}\right)_+ \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right) \tag{3.15}$$

where $p \geq 3$. Therefore in this study we use the forms (3.14) and (3.15) to get our results.

We consider the following definitions on shrinkage estimators. We start by stating a general definition on a shrinkage estimator and then narrow it to James and Stein's shrinkage type estimator in [35].

Definition 3.2.15 *A shrinkage estimator is an estimator that, either explicitly or implicitly, incorporates the effects of shrinkage. In loose terms this means that a naive or raw estimate is improved by combining it with other information. The term relates to the notion that the improved estimate is made closer to the value supplied by the other information than the raw estimation.*

Definition 3.2.16 *For $p \geq 3$, let $\hat{\boldsymbol{\theta}}_n$ be an estimator of normally distributed random p-vector with unknown mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{V}$, i.e. $\hat{\boldsymbol{\theta}}_n \sim N_p(\boldsymbol{\theta}, \boldsymbol{V})$. Given an unbiased estimator $\hat{\boldsymbol{\theta}}_n$, a **positive part James-Stein shrinkage** estimator is given by*

$$\hat{\boldsymbol{\theta}}_n^* = \left(1 - \frac{b}{a + \parallel \hat{\boldsymbol{\theta}}_n \parallel^2}\right)_+ \hat{\boldsymbol{\theta}}_n$$

with $a, b > 0$ where $(.)_+$ means $(.) \geq 0$. We denote the James-Stein shrinkage estimator obtained from an estimator $\hat{\boldsymbol{\theta}}_n$ by $\hat{\boldsymbol{\theta}}_n^*$.

Note 3.2.17 *If we let $b = p - 2$ and set $\left(a + \parallel \hat{\boldsymbol{\theta}}_n \parallel^2\right) = \left(\hat{\boldsymbol{\theta}}_n^\top \boldsymbol{V}^{-1} \hat{\boldsymbol{\theta}}_n\right)$, then we have (3.14) which when we generalise gives (3.15).*

Lehman *et al* [42] establishes the following result from [51], [35] and [17]. We use the result to show that the James-Stein shrinkage estimator has lower risk compared to the `MLE`.

Lemma 3.2.18 *([42], Theorem 5.1, p.355)*

Let $\hat{\boldsymbol{\theta}}_n \sim N_p(\boldsymbol{\theta}, \mathbf{I})$ such that $p \geq 3$ and let the estimator $\hat{\boldsymbol{\theta}}_n^*$ of $\boldsymbol{\theta}$ be given by

$$\hat{\boldsymbol{\theta}}_n^* = \left(1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \mathbf{I}^{-1} \hat{\boldsymbol{\theta}}_n}\right)_+ \hat{\boldsymbol{\theta}}_n$$

where $\mathbf{I}$ is a pxp identity matrix. Then the risk function of $\hat{\boldsymbol{\theta}}_n^*$ is

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n^*) = 1 - \frac{(p-2)^2}{p} \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{1}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\right]$$

where the expectation is taken with respect to the parameter $\boldsymbol{\theta}$.

Theorem 3.2.19 *([51], p.199)*

For $p \geq 3$, let X be a normally distributed random p-vector with unknown mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{I}$, the identity matrix. In addition to the usual estimator $\hat{\boldsymbol{\theta}}_n$ consider the James-Stein shrinkage estimator $\hat{\boldsymbol{\theta}}_n^*$ given by

$$\hat{\boldsymbol{\theta}}_n^* = \left(1 - \frac{b}{a + \parallel \hat{\boldsymbol{\theta}}_n \parallel^2}\right)_+ \hat{\boldsymbol{\theta}}_n$$

with $a, b > 0$. For sufficiently small b and large a, $\hat{\boldsymbol{\theta}}_n^*$ is strictly better than $\hat{\boldsymbol{\theta}}_n$, in fact,

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}\right]^2 < p = \mathbb{E}_{\boldsymbol{\theta}} \left[\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right]^2$$

for all $\boldsymbol{\theta}$.

In the next proposition, we prove that the James-Stein shrinkage estimator has lower `MSE` compared to the `MLE` which is important for the study.

Proposition 3.2.20

Let $\hat{\boldsymbol{\theta}}_n \sim N_p(\boldsymbol{\theta}, \mathbf{I})$ be a maximum likelihood estimator of $\boldsymbol{\theta} \in \Omega$ where Ω is a parameter space with elements in $\mathbb{R}^p$. For $p \geq 3$, the James-Stein shrinkage estimator $\hat{\boldsymbol{\theta}}_n^*$ obtained by shrinking $\hat{\boldsymbol{\theta}}_n$ given by

$$\hat{\boldsymbol{\theta}}_n^* = \left(1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \mathbf{I} \hat{\boldsymbol{\theta}}_n}\right)_+ \hat{\boldsymbol{\theta}}_n$$

has lower risk loss (`MSE`) compared to the risk loss (`MSE`) of the initial `MLE` $\hat{\boldsymbol{\theta}}_n$.

Proof

Let $\hat{\theta}_n \sim N_p(\boldsymbol{\theta}, \mathbf{I})$*, given the James-Stein shrinkage estimator*

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_n^* &= \left(1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \boldsymbol{V}^{-1}\hat{\boldsymbol{\theta}}_n}\right)_+ \hat{\boldsymbol{\theta}}_n \\
&= \left(1 - \frac{p-2}{\hat{\boldsymbol{\theta}}_n^\top \mathbf{I}\hat{\boldsymbol{\theta}}_n}\right)_+ \hat{\boldsymbol{\theta}}_n \qquad\qquad \textit{since} \quad \boldsymbol{V}^{-1} = \mathbf{I} \\
&= \hat{\boldsymbol{\theta}}_n - \frac{(p-2)}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\hat{\boldsymbol{\theta}}_n \\
\hat{\boldsymbol{\theta}}_n^* &= \hat{\boldsymbol{\theta}}_n - g(\hat{\boldsymbol{\theta}}_n)
\end{aligned}$$

where the function g *is given by* $g(\hat{\boldsymbol{\theta}}_n) = \frac{(p-2)}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\hat{\boldsymbol{\theta}}_n$ *and* $\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n \sim \chi^2_p$ *for* $p \geq 3$*. Then*

$$\begin{aligned}
\boldsymbol{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n^*) &= \mathbb{E}_{\boldsymbol{\theta}}\left[\parallel \hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta} \parallel^2\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\parallel \hat{\boldsymbol{\theta}}_n - g(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{\theta} \parallel^2\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left\{\left[\parallel (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}_n) \parallel^2\right]\right\} \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\parallel (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \parallel^2\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\parallel g(\hat{\boldsymbol{\theta}}_n) \parallel^2\right] - 2\mathbb{E}_{\boldsymbol{\theta}}\left[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}).g(\hat{\boldsymbol{\theta}}_n)\right]
\end{aligned}$$

where all expectations are taken with respect to the distribution of $\hat{\boldsymbol{\theta}}_n$ *given* $\boldsymbol{\theta}$*. The first expectation*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\parallel \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \parallel^2\right] = \boldsymbol{V}(\hat{\boldsymbol{\theta}}_n) = \sigma^2 p = p \tag{3.16}$$

since $\sigma^2 = 1$*. The second expectation will be*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[g \parallel \hat{\boldsymbol{\theta}}_n \parallel^2\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[(p-2)^2 \frac{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)^2}\right] = (p-2)^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right] \tag{3.17}$$

using $\parallel \boldsymbol{t} \parallel = \boldsymbol{t}^\top \boldsymbol{t}$*. To simplify the third expectation we use Stein's identity in Lemma 3.2.13. When we apply it on* $g(\hat{\boldsymbol{\theta}}_n) = \frac{(p-2)}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\hat{\boldsymbol{\theta}}_n$ *we have*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}).g(\hat{\boldsymbol{\theta}}_n)\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{(p-2)^2}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right] = (p-2)^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right]. \tag{3.18}$$

Therefore from (3.16), (3.17) and (3.18) we have

$$\begin{aligned}
\boldsymbol{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n^*) &= p + (p-2)^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right] - 2(p-2)^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right] \\
&= p - (p-2)^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right] \\
&= p - \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{(p-2)^2}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}\right].
\end{aligned}$$

Now the term $(p-2) > 0$ *and* $\frac{1}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)}$ *is an inverse moment of a non central* χ^2 *distribution which depend on* $\boldsymbol{\theta}$ *since* $\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n \sim \chi^2_p$. *Since* $\frac{1}{\boldsymbol{\theta}^\top \boldsymbol{\theta}} > 0$ *for all* $\boldsymbol{\theta} \in \mathbb{R}^p$, *then*

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{(p-2)^2}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)} \right] > 0$$

for all $\boldsymbol{\theta}$. *Hence*

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\| \hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta} \|^2 \right] = p - \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{(p-2)^2}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)} \right] < p = \mathbb{E}_{\boldsymbol{\theta}} \left[\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \|^2 \right]$$

for all $\boldsymbol{\theta} \in \mathbb{R}^p$. *This means that the risk loss (`MSE`) for the James-Stein shrinkage estimator* $\hat{\boldsymbol{\theta}}_n^*$ *is lower than that of the `MLE`* $\hat{\boldsymbol{\theta}}_n$. □

3.2.3 Fundamentals of Maximum Likelihood Estimators

In this section we state results and definitions on maximum likelihood estimators useful for our discussion. Some proofs for some results are in the appendix. We will need the following definitions on maximum likelihood estimators.

Definition 3.2.21 *Let* $X_1, ..., X_n$ *be a random sample from the model* $\{f_\theta(X) : \theta \in \Omega\}$, *then*

1. *the* ***likelihood function*** *of* $X_1, ..., X_n$ *is given by*

$$\boldsymbol{L}(\theta) = \boldsymbol{L}(\theta, X) = \prod_{i=1}^{n} f_\theta(X_i).$$

2. *the* ***log likelihood function*** *of* $X_1, ..., X_n$ *is given by*

$$\boldsymbol{\ell}(\theta) = \boldsymbol{\ell}(\theta, X) = \log \boldsymbol{L}(\theta, X)$$

where $\boldsymbol{L}(\theta, X)$ *is as defined in 1.*

Definition 3.2.22 *Let* $X_1, ..., X_n$ *be a random sample from the model* $\{f_\theta(X) : \theta \in \Omega\}$, *then*

1. *the function*

$$\boldsymbol{S}(\theta) = \boldsymbol{S}(\theta, X) = \frac{\partial}{\partial \theta} \boldsymbol{\ell}(\theta, X)$$

is called the ***score function*** *of* $X_1, ..., X_n$, *where* $\boldsymbol{\ell}(\theta, X)$ *is the log likelihood function.*

2. *the function*

$$\boldsymbol{I}(\theta) = \boldsymbol{I}(\theta, X) = -\frac{\partial}{\partial \theta} \boldsymbol{S}(\theta, X) = -\frac{\partial^2}{\partial \theta^2} \boldsymbol{\ell}(\theta, X)$$

is called the ***information function*** *of* $X_1, ..., X_n$, *where* $\boldsymbol{\ell}(\theta, X)$ *and* $\boldsymbol{S}(\theta, X)$ *are the log likelihood and score function respectively.*

3. *the function*

$$\boldsymbol{J}(\theta) = \boldsymbol{J}(\theta, X) = \mathbb{E}_\theta \left[\boldsymbol{I}(\theta, X) \right]$$

is called the ***expected*** *or* ***fisher information function****, where* $\boldsymbol{I}(\theta, X)$ *is as defined in 2 above.*

Definition 3.2.23 *Let $X_1, ..., X_n$ be a random sample from the model $\{f_\theta(x) : \theta \in \Omega\}$ such that $\Omega \in \mathbb{R}^2$, then a value $\hat{\theta} = \hat{\theta}(x)$ which maximizes the likelihood function $\boldsymbol{L}(\theta, x)$ is called a maximum likelihood estimate of θ. Therefore, $\hat{\theta} = \hat{\theta}(X)$ is called a* ***maximum likelihood estimator*** *(MLE) of θ.*

Note 3.2.24 *In many cases, $\hat{\theta}$ is found by solving the score equation $\boldsymbol{S}(\theta, x) = 0$ and checking that $\boldsymbol{I}(\hat{\theta}, x) > 0$.*

Definition 3.2.25 *Consider a parametric model in which the joint distribution of $X = (X_1, ..., X_n)$ has a density $f_{\boldsymbol{\theta}}(X)$ with respect to a parameter space $\Omega \in \mathbb{R}^p$. Then the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is a solution to the maximization problem*

$$\max_{\boldsymbol{\theta} \in \Omega} f_{\boldsymbol{\theta}}(x).$$

Note 3.2.26 *Note that the solution to an optimization problem is invariant to a strictly monotone increasing transformation of the objective function. The MLE can be obtained as a solution to the following problem*

$$\max_{\boldsymbol{\theta} \in \Omega} \log \boldsymbol{L}(\boldsymbol{\theta}, x) = \max_{\boldsymbol{\theta} \in \Omega} \boldsymbol{\ell}(\boldsymbol{\theta}, x).$$

The following proposition assures us of having only one MLE and restricted maximum likelihood estimator (RMLE) for the whole parameter space Ω and subparameter space Ω_o respectively in the set up of the parameter structure in the analysis (Chapter 4).

Proposition 3.2.27 ***(Sufficient condition for Uniqueness of MLE)***

If the parameter space Ω is convex and the likelihood function $\boldsymbol{L}(\theta, x)$ is strictly concave in θ, then the MLE is unique when it exists. If the observations on X are identically independent distributed (iid) with density $f_\theta(x_i)$ for each observation, then we can write the likelihood function as

$$\boldsymbol{L}(\theta, x) = \prod_{i=1}^{n} f_\theta(x_i) \Longrightarrow \boldsymbol{\ell}(\theta, x) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

by using the definition of a likelihood function and then taking logarithms.

In this study we have a mapping for shrinkage, so we use the theorem below to obtain the plug-in estimators which we use to obtain the results.

Theorem 3.2.28 ***(Invariance Property)****([11], Theorem 7.2.10)*

If $\hat{\theta}$ is the maximum likelihood estimator of θ and $\tau(\theta)$ is a function of θ, then $\tau(\hat{\theta})$ is the maximum likelihood estimator of $\tau(\theta)$.

We define the following concepts on regularity conditions for the maximum likelihood estimator.

Definition 3.2.29 *Let $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + n^{-\frac{1}{2}}\boldsymbol{h}$ be a sequence of estimates where $\boldsymbol{\theta}_o$ is the "true value" and $\boldsymbol{h}$ be a fixed parameter which provides a neighbourhood for shrinking $(n^{-\frac{1}{2}})$ for the sample size value n. An estimator T_n is called* ***regular*** *if*

$$\sqrt{n}\,(T_n - \boldsymbol{\theta}_n) \longrightarrow_d \boldsymbol{\psi}$$

for some random variable $\boldsymbol{\psi}$ which does not depend on $\boldsymbol{h}$.

Note 3.2.30 *A regular sequence of estimators is one whose asymptotic distribution remains the same in shrinking neighbourhoods of the true parameter value.*

Definition 3.2.31 *Consider a family of probability density functions* $\{f_\theta(x) : \theta \in \Omega\}$. *Let* $A = \{x : f_\theta(x) > 0\}$. *Then*

$$\int_A f_\theta(x)dx = 1$$

and

$$\int_A \frac{\partial}{\partial\theta} f_\theta(x)dx = \frac{\partial}{\partial\theta}\int_A f_\theta(x)dx. \tag{3.19}$$

Models that permit the interchange of integral and the derivative like in (3.19) and the calculation of the fisher information are called ***regular models****.*

Regularity and smoothness are combined in the three lemmas below. They show results on regularity assumptions of the `MLE` which are referred to in this study when deriving the asymptotic distribution.

Lemma 3.2.32

Let $X = (X_1, ...X_n)$ *be a random sample from a model* $\{f_\theta(x) : \theta \in \Omega\}$ *which satisfies regularity conditions in Definition 3.2.31 such that the density* $f_\theta(x)$ *is continuous. Then for* $A = \{x : f_\theta(x) > 0\}$ *we have*

$$\frac{\partial}{\partial\theta}\int_A f_\theta(x)dx = \int_A \frac{\partial f_\theta(x)}{\partial\theta}dx = 0$$

and

$$\frac{\partial^2}{\partial\theta\partial\theta^\top}\int_A f_\theta(x)dx = \int_A \frac{\partial^2 f_\theta(x)}{\partial\theta\partial\theta^\top}dx = 0.$$

Lemma 3.2.33 *([45], Theorem 2.3.3)*

Let $X = \{X_1, ...X_n\}$ *be a random sample from a model* $\{f_{\theta_o}(x) : \theta \in \Omega\}$ *which is regular as defined above. Let* $A = \{x : f_{\theta_o}(x) > 0\}$, *then*

$$\mathbb{E}_{\theta_o}\left[\frac{\partial \log f_{\theta_o}(x)}{\partial\theta}\right] = 0.$$

Lemma 3.2.34 *([11], Lemma 7.3.11)*

Let $X = \{X_1, ...X_n\}$ *be a random sample from a model* $\{f_{\theta_o}(x) : \theta \in \Omega\}$ *which satisfies the regularity conditions above in Definition 3.2.31 , then*

$$\mathbb{E}_{\theta_o}\left[\frac{\partial \log f_{\theta_o}(X)}{\partial\theta}\frac{\partial \log f_{\theta_o}(X)}{\partial\theta^\top}\right] = \mathbb{E}_{\theta_o}\left[-\frac{\partial^2 \log f_{\theta_o}(X)}{\partial\theta\partial\theta^\top}\right].$$

The sequences considered in the next chapters uses the Taylor's theorem and the mean value theorem to expand and determine their convergence. We will expand using the Taylor's theorem about a middle point chosen using the mean value theorem (`MVT`) between an estimator and the assumed true parameter value $\boldsymbol{\theta}_o$. We therefore state the two theorems without proving them.

Theorem 3.2.35 ***(Taylor's Theorem)***

Suppose f is a real-valued function of x that can be differentiated $n+1$ times in an interval $\boldsymbol{\mathcal{I}}$ containing x_o with the $(n+1)^{th}$ derivative integrable on $\boldsymbol{\mathcal{I}}$. If $x_o+a \in \boldsymbol{\mathcal{I}}$ then we have

$$f(x_o+a) = f(x_o) + \sum_{k=1}^{n} \frac{f^k(x_o)a^k}{k!} + \int_{x_o}^{x_o+a} f^{(n+1)}(t)\frac{(x_o+a-t)^n}{n!}dt$$

where $x = x_o + a$ and the expansion is about a.

Theorem 3.2.36 ***(Mean Value Theorem (MVT))***

Let $f : [\boldsymbol{a}, \boldsymbol{b}] \longrightarrow \mathbb{R}$ be a continuous function on the closed interval $[\boldsymbol{a}, \boldsymbol{b}]$, and differentiable on the open interval $(\boldsymbol{a}, \boldsymbol{b})$, where $\boldsymbol{a} < \boldsymbol{b}$. Then there exists some $\boldsymbol{c} \in (\boldsymbol{a}, \boldsymbol{b})$ such that

$$f'(\boldsymbol{c}) = \frac{f(\boldsymbol{b}) - f(\boldsymbol{a})}{\boldsymbol{b} - \boldsymbol{a}}$$

In the next section we discuss the concepts of probability and convergence as discussed in [14]. This will guide us when analysing the asymptotic convergence of the shrinkage estimator and establishing the main results for the study.

3.2.4 Some Probability Results

In this section we consider definitions and results on probability and convergence which are useful to get our results in the study. We begin by defining convergence in probability.

Definition 3.2.37 ***(Convergence in Probability)*** *A sequence of random variables $X_1, ..., X_n$ converges in probability to a constant C if for each $\varepsilon > 0$*

$$\lim_{n\to\infty} \Pr\left(|X_n - C| > \varepsilon\right) = 0$$

or

$$\lim_{n\to\infty} \Pr\left(|X_n - C| < \varepsilon\right) = 1.$$

We write $X_n \longrightarrow_p C$.

The following definitions are according to Chernoff [12]. We use them to determine the probability order, analyse the quadratic mean and the rate of convergence of the shrinkage estimator in Chapter 5.

Definition 3.2.38 *X_n converges in probability to 0 with increasing n if for every $\varepsilon > 0$ $\lim \Pr\left(\mid X_n \mid \leq \varepsilon\right) = 1$. We write $\Pr \lim_{n\longrightarrow\infty} X_n = 0$ to mean X_n converges in probability to 0 with increasing n. Furthermore*

$$\Pr \lim_{n\longrightarrow\infty} X_n = X \; if \; \Pr \lim_{n\longrightarrow\infty} (X_n - X) = 0.$$

The definition above is the same as Definition 3.2.37 when the constant $C = 0$.

Definition 3.2.39 *X_n is of probability order $O_p\left(f(n)\right)$ if*

$$\Pr \lim_{n\longrightarrow\infty} X_n/f(n) = 0.$$

We write $X_n = O_P\left(f(n)\right)$ to mean X_n is of probability order $O_p\left(f(n)\right)$.

We apply the following results in the analysis of the results in Chapter 5 when showing the consistence of the RMLE $\tilde{\boldsymbol{\theta}}_n^o$.

Theorem 3.2.40 ***(Chebyshev's Inequality)***

Suppose X is a random variable with $\mathbb{E}(X) = \mu$ and $\boldsymbol{V}(X) = \sigma^2 < \infty$ then for each $k > 0$, then

$$\Pr(|X - \mu| > k) < \frac{\sigma^2}{k^2}.$$

Theorem 3.2.41 ***(Weak Law of Large Numbers (WLLN))***

Suppose $X_1, ...X_n$ is a random sample from a distribution with $\mathbb{E}(X_i) = \mu$ and $\boldsymbol{V}(X_i) = \sigma^2 < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow_p \mu.$$

We obtain the following result directly from the weak law of large numbers (WLLN) on the convergence of the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \longrightarrow_p \sigma^2.$$

We use Theorem 3.2.41 to show the convergence of the James-Stein shrinkage estimator to some asymptotic normal distribution. To establish convergence in distribution for the three estimators $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}_n^o$ and $\hat{\boldsymbol{\theta}}_n^*$ we need the following definition.

Definition 3.2.42 ***(Convergence in Distribution)*** *Let X_n be a sequence of random variables. We say X_n converges in distribution to a random variable X if*

$$\lim_{n\to\infty} \Pr(X_n \leq x) = \Pr(X \leq x)$$

for all values of x at which the right hand side $F(x) = \Pr(X \leq x)$ is continuous. We denote it by

$$X_n \longrightarrow_d X.$$

The definition above requires a sequence of random variables to analyse convergence in distribution. To evaluate asymptotic values, we consider the concept of convergence in distribution and probability. To investigate the asymptotic consistency of the James-Stein shrinkage estimator, we combine these two concepts. Therefore the following results show the combination of these concepts.

Theorem 3.2.43 ***(Central Limit Theorem (CLT))***

Suppose $X_1, ..., X_n$ is a random sample from a distribution with $\mathbb{E}(X_i) = \mu$ and $\boldsymbol{V}(X_i) = \sigma^2 < 0$. Then

$$Y_n = \frac{\sum_{i=1}^{n} X_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \longrightarrow_d Z \sim N(0,1).$$

Theorem 3.2.44 ***(Limit Theorems)***

1. *If $X_n \longrightarrow_p \boldsymbol{a}$ where $\boldsymbol{a}$ is a constant and g is a real-valued function which is continuous at $\boldsymbol{a}$, then*

$$g(X_n) \longrightarrow_p g(\boldsymbol{a}).$$

2. *If* $X_n \longrightarrow_d X$ *and* g *is a real-valued then*

$$g(X_n) \longrightarrow_d g(X).$$

3. *If* $X_n \longrightarrow_p X$ *then* $X_n \longrightarrow_d X$.

As a consequence of Theorem 3.2.44 above we have the following corollary on convergence of the sample variance.

Corollary 3.2.45

Suppose $X_1, ..., X_n$ *is a random sample from a distribution with* $\mathbb{E}(X_i) = \mu$ *and* $\boldsymbol{V}(X_i) < \infty$. *For the sample variance* S^2 *we have*

$$S \longrightarrow_p \sigma, \frac{S}{\sigma} \longrightarrow_p 1 \quad and \quad \frac{S^2}{\sigma^2} \longrightarrow_p 1.$$

Theorem 3.2.46 ***(Slutsky's Theorem)***

Let X_n *and* Y_n *be a sequence of random variables. If* $X_n \longrightarrow_p \boldsymbol{a}$ *and* $Y_n \longrightarrow_d Y$, *then*

1. $X_n + Y_n \longrightarrow_d \boldsymbol{a} + Y$.
2. $X_n Y_n \longrightarrow_d \boldsymbol{a}Y$.
3. $\frac{Y_n}{X_n} \longrightarrow_d \frac{Y}{\boldsymbol{a}}$ *provided* $\boldsymbol{a} \neq 0$.

Theorem 3.2.47

Suppose $\boldsymbol{a}$ *and* $\boldsymbol{b} > 0$ *are constants and* $n^{\boldsymbol{b}}(X_n - \boldsymbol{a}) \longrightarrow_d X$. *Let* g *be a real valued function that is differentiable and whose derivative* g' *is continuous at* $\boldsymbol{a}$. *Then*

$$n^{\boldsymbol{b}}\left[g(X_n) - g(\boldsymbol{a})\right] \longrightarrow_d g'(\boldsymbol{a})X.$$

We now define the concept of consistency of an estimator.

Definition 3.2.48 *Consider a sequence of estimators* T_n *where the subscript* n *indicate that the estimators has been obtained from the data* $(X_1, ..., X_n)$ *with sample size* n. *Then the sequence is said to be a* ***consistent sequence*** *of estimators of* $\tau(\theta)$ *if*

$$T_n \longrightarrow_p \tau(\theta)$$

for all $\theta \in \Omega$.

Note 3.2.49 *To solve certain kinds of sample-size problems, it is helpful to take Definition 3.2.48 in the context where we involve an* ε *and* δ. *That is,* $\hat{\boldsymbol{\theta}}_n$ *is consistent for* $\boldsymbol{\theta}$ *if for all* $\varepsilon > 0$ *and* $\delta > 0$, *there exist an* $n(\varepsilon, \delta)$ *such that*

$$\lim_{n \longrightarrow \infty} \Pr(\mid \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \mid < \varepsilon) > 1 - \delta \qquad for \quad n > n(\varepsilon, \delta).$$

Definition 3.2.50 *A sequence of estimators* T_n *is* $\boldsymbol{\sqrt{n} - consistent}$ *for* $\boldsymbol{\theta}$ *if* $\sqrt{n}\,(T_n - \boldsymbol{\theta})$ *is bounded in probability, that is, if*

$$T_n - \boldsymbol{\theta} = O_P\left(\frac{1}{\sqrt{n}}\right)$$

where $O_p(a)$ *is the probability order and* n *is the sample size value.*

When the term $\sqrt{n}(T_n - \boldsymbol{\theta})$ in Definition 3.2.50 above becomes $k\sqrt{n}(T_n - \boldsymbol{\theta})$ due to bias effect, and if it is bounded in probability such that

$$T_n - \boldsymbol{\theta} = O_P\left(\frac{1}{k\sqrt{n}}\right)$$

then T_n is said to be $\boldsymbol{k\sqrt{n} - consistent}$.

Theorem 3.2.51

Suppose $X_1, ..., X_n$ is a random sample from a regular statistical model $\{f_\theta(x) : \theta \in \Omega\}$. Let $\boldsymbol{S}_1(\theta, X) = \frac{\partial}{\partial\theta} log f_\theta(x)$ be a score function for a sample of size one. Then with probability tending to one as $n \longrightarrow \infty$, the likelihood equation

$$\sum_{i=1}^{n} \boldsymbol{S}_1(\theta, X_i) = 0$$

has a root $\hat{\theta}_n$ such that $\hat{\theta}_n$ converges in probability to θ_o, the true value of the parameter, as $n \longrightarrow \infty$. i.e the `MLE` is consistent ($\hat{\theta}_n \longrightarrow_p \theta_o$).

Proposition 3.2.52

Let g be a continuously differentiable function of $\boldsymbol{\theta} \in \mathbb{R}^p$. If a sequence of `MLEs` $\hat{\boldsymbol{\theta}}_n$, satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \longrightarrow_d N\left(\boldsymbol{0}, \boldsymbol{I}(\theta_o)^{-1}\right)$$

then as $n \longrightarrow \infty$

1. *$g(\hat{\boldsymbol{\theta}}_n)$ converges in probability almost everywhere to $g(\boldsymbol{\theta}_o)$.*
2. *$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_o)\right) \longrightarrow_d N\left(\boldsymbol{0}, g'(\boldsymbol{\theta}_o)\mathbf{I}(\boldsymbol{\theta}_o)^{-1}g(\boldsymbol{\theta}_o)\right)$.*

Proposition 3.2.52 above holds by direct mapping. This result is used in Chapter 4 to analyse the asymptotic distribution, and in Chapter 5 to show the consistency of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}^*_n$. We use it to link the estimators $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}^o_n$ and $\hat{\boldsymbol{\theta}}^*_n$ to the plug-in estimators $\hat{\boldsymbol{\beta}}_n$, $\tilde{\boldsymbol{\beta}}^o_n$ and $\hat{\boldsymbol{\beta}}^*_n$ respectively. We also consider the following theorem for the same fact.

Theorem 3.2.53 *([45], Theorem 2.7.4)*

Let $X_1, ..., X_n$ be a random sample from a regular statistical model $\{f_\theta(x) : \theta \in \Omega\}$. Suppose that θ is a consistent root of the likelihood equation. Let

$$\boldsymbol{J}_1(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right]$$

be the fisher information for a sample of size one. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_o) \longrightarrow_d N\left(0, \frac{1}{\boldsymbol{J}_1(\theta_o)}\right)$$

where θ_o is the true value of the parameter. This result may also be written as

$$\sqrt{n\boldsymbol{J}_1(\theta_o)}(\hat{\theta}_n - \theta_o) = \sqrt{\boldsymbol{J}(\theta_o)}(\hat{\theta}_n - \theta_o) \longrightarrow_d Z \sim N(0, 1).$$

Note 3.2.54 *Theorem 3.2.53 asserts;*

1. *the* `MLE` *is asymptotically unbiased.*
2. *the asymptotic variance of* `MLE` *approaches the Cramér Rao Lower Bound* (`CRLB`).

In this study we need to evaluate the asymptotic distributional bias for the three estimators in play. We therefore state the following definition.

Definition 3.2.55 *Given an estimator T_n of the parameter vector $\boldsymbol{\theta}$ where n is the sample size value. The asymptotic distributional bias denoted by* `ADB` *is given by*

$$\boldsymbol{ADB}(T_n) = \lim_{n \longrightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}} \left[\sqrt{n}(T_n - \boldsymbol{\theta}) \right]$$

where $T_n = \{T_n, n = 1, 2, ...\}$ and $\boldsymbol{\theta} = \{\theta_n, n = 1, 2, ...\}$.

We use the following definition to establish the asymptotic efficiency of the James-Stein shrinkage estimator as part of the results in the last section of Chapter 5.

Definition 3.2.56 *Let $X_1, ..., X_n$ be independent and identically distributed (*`iid`*) according to a probability density $f_\theta(X)$ satisfying suitable regularity conditions. Suppose that $\boldsymbol{T}_n = \boldsymbol{T}_n(X_1, ..., X_n)$ is asymptotically normal say that*

$$\sqrt{n}\,(\boldsymbol{T}_n - \boldsymbol{\theta}) \longrightarrow_d N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\theta})) \qquad \boldsymbol{V}(\boldsymbol{\theta}) \geq 0$$

for the positive definite matrix $\boldsymbol{V}(\boldsymbol{\theta})$, where $\boldsymbol{T}_n$ is estimating $\boldsymbol{\theta}$. Then a sequence of estimators $\{\boldsymbol{T}_n\} = \{\boldsymbol{T}_n(X_1, ..., X_n)\}$ satisfying

$$\lim_{n \longrightarrow \infty} [n\,\boldsymbol{V}(\boldsymbol{T}_n)] = [\boldsymbol{J}(\boldsymbol{\theta})]^{-1} \tag{3.20}$$

for the fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is said to be ***asymptotically efficient****. Particularly (3.20) becomes*

$$\boldsymbol{V}(\boldsymbol{\theta}) = [\boldsymbol{J}(\boldsymbol{\theta})]^{-1}.$$

The next lemma is required to prove Theorem 4.3.2 which shows the convergence in distribution of all the three estimators considered in the study.

Lemma 3.2.57 *([28], Lemma 1)*

Let $\boldsymbol{T} = \{\boldsymbol{T}_n : n = 1, 2, ...\}$ denote a sequence of estimators and $\boldsymbol{\theta}_n$ denote a sequence of parameter estimates. For any estimator $\boldsymbol{T}_n$ satisfying

$$\sqrt{n}\,(\boldsymbol{T}_n - \boldsymbol{\theta}_n) \longrightarrow_d \psi, \tag{3.21}$$

for some random variable ψ, if $\boldsymbol{\theta}_n \longrightarrow \boldsymbol{\theta}_o$ as $n \longrightarrow \infty$, and for any loss function satisfying regularity conditions then the asymptotic risk R of the estimator $\boldsymbol{T}$ is given by

$$R(\boldsymbol{h}, \boldsymbol{T}) = \mathbb{E}\left(\psi^\top \boldsymbol{W} \psi\right)$$

where $\boldsymbol{h}$ is a non-centrality parameter and $\boldsymbol{W}$ is a weight matrix (equal to the variance matrix) for the random variable ψ.

Proof

Let $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + n^{-\frac{1}{2}}\boldsymbol{h}$ *where* n *is the sample size value,* $\boldsymbol{\theta}_n$ *is a sequence of parameters estimates and* $\boldsymbol{\theta}$ *is a parameter. By the regularity conditions of the loss function* ℓ *we have that* $\ell(\boldsymbol{\theta}_n, \boldsymbol{\theta}_n) = 0$*. This imply that* $\ell(\boldsymbol{\theta}_n, \boldsymbol{\theta})$ *is minimised and differentiable at* $\boldsymbol{\theta} = \boldsymbol{\theta}_n$*, so the first-order expansion condition is*

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}_n, \boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} = 0.$$

Then by a second-order Taylor series expansion of $\ell(\boldsymbol{\theta}_n, \boldsymbol{T}_n)$ *about* $\boldsymbol{\theta}_n$ *gives*

$$n\ell(\boldsymbol{\theta}_n, \boldsymbol{T}_n) = n\ell(\boldsymbol{\theta}_n, \boldsymbol{\theta}_n) + n\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}_n, \boldsymbol{\theta}_n)^\top (\boldsymbol{T}_n - \boldsymbol{\theta}_n) + n(\boldsymbol{T}_n - \boldsymbol{\theta}_n)^\top \boldsymbol{W}(\boldsymbol{\theta}_n^*)(\boldsymbol{T}_n - \boldsymbol{\theta}_n).$$

Since the first and second terms on the right hand side are zero, we have

$$n\ell(\boldsymbol{\theta}_n, \boldsymbol{T}_n) = n(\boldsymbol{T}_n - \boldsymbol{\theta}_n)^\top \boldsymbol{W}(\boldsymbol{\theta}_n^*)(\boldsymbol{T}_n - \boldsymbol{\theta}_n) \tag{3.22}$$

for some $\boldsymbol{\theta}_n^*$ *on the line segment joining* $\boldsymbol{T}_n$ *and* $\boldsymbol{\theta}_n$*. The expression (3.21) and* $\boldsymbol{\theta}_n \longrightarrow \boldsymbol{\theta}_o$ *imply* $\boldsymbol{T}_n \longrightarrow_p \boldsymbol{\theta}_o$ *and hence* $\boldsymbol{\theta}_n^* \longrightarrow_p \boldsymbol{\theta}_o$*. By the continuity of the loss function* ℓ *from the regularity conditions, it follows that* $\boldsymbol{W}(\boldsymbol{\theta}_n^*) \longrightarrow_p \boldsymbol{W}(\boldsymbol{\theta}_o) = \boldsymbol{W}$*. Combining this with (3.21) we find*

$$n\ell(\boldsymbol{\theta}_n, \boldsymbol{T}_n) \longrightarrow_d \psi^\top \boldsymbol{W}\psi.$$

As shown by Lemma 6.1.14 of Lehman and Casella in [42] the above convergence can be used to establish that

$$R(\boldsymbol{h}, \boldsymbol{T}) = \lim_{\zeta \longrightarrow \infty} \lim_{n \longrightarrow \infty} \inf \mathbb{E}_{\boldsymbol{\theta}} \min\left[n\boldsymbol{\ell}(\boldsymbol{\theta}_n, \boldsymbol{T}_n), \zeta\right]$$

$$R(\boldsymbol{h}, \boldsymbol{T}) = \mathbb{E}\left(\psi^\top \boldsymbol{W}\psi\right)$$

for some scale constant ζ*.* □

In Chapter 4 we partition the parameter space by introducing a restriction on it to set up a subspace to shrink to. This is in order to have a shrinkage target. To do that, we use the concepts in the following theorem.

Theorem 3.2.58 *([45], Theorem 4.4.8)*

Suppose $X_1, X_2, ..., X_n$ *is a random sample from a regular statistical model* $\{f_{\boldsymbol{\theta}}(X) : \boldsymbol{\theta} \in \Omega\}$ *with* Ω *an open set in p-dimensional Euclidean space. Consider a subset of* Ω *defined by* $\Omega_o = \{\boldsymbol{\theta}(\eta) : \eta \in$ *open subset of the k-dimensional Euclidean space* $\}$ *where* $k < p$*. Then the likelihood ratio statistic defined by*

$$\Lambda_n(X) = \frac{\sup_{\boldsymbol{\theta}\in\Omega} \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i)}{\sup_{\boldsymbol{\theta}\in\Omega_o} \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i)} = \frac{\sup_{\boldsymbol{\theta}\in\Omega} \boldsymbol{L}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Omega_o} \boldsymbol{L}(\boldsymbol{\theta})}$$

is such that under the restriction Ω_o

$$2\log\Lambda_n(X) \underset{d}{\longrightarrow} \omega \sim \chi^2_{(p-k)}$$

for $k < p$*.*

Note 3.2.59 *The number of degrees of freedom is the difference between the number of parameters that need to be estimated in the general model and the number of parameters to be estimated under the restrictions imposed on* Ω_o.

Lastly, we state the following definition which provides a mathematical definition for domination of an estimator by another estimator.

Definition 3.2.60 *If* $\boldsymbol{T}_1$ *and* $\boldsymbol{T}_2$ *are estimators for the parameter* $\boldsymbol{\theta}$*, then* $\boldsymbol{T}_1$ *is said to dominate* $\boldsymbol{T}_2$ *if*

1. *its mean squared error (*`MSE`*) is smaller for at least some value of* $\boldsymbol{\theta}$.
2. *the mean squared error (*`MSE`*) of* $\boldsymbol{T}_1$ *does not exceed that of* $\boldsymbol{T}_2$ *for any value of* $\boldsymbol{\theta}$.

Formally, $\boldsymbol{T}_1$ *dominates* $\boldsymbol{T}_2$ *if*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mid \boldsymbol{T}_1 - \boldsymbol{\theta} \mid^2\right] \leq \mathbb{E}_{\boldsymbol{\theta}}\left[\mid \boldsymbol{T}_2 - \boldsymbol{\theta} \mid^2\right]$$

holds for all $\boldsymbol{\theta}$*, with strict inequality holding somewhere.*

Chapter 4

SHRINKAGE ANALYSIS AND THE ASYMPTOTIC DISTRIBUTION

In this chapter we set up a statistical model that is used in the study. The first section presents work on the set up of the parameters and how we shrink the MLE towards a sub-parameter space. The second section discusses work on estimation and the structures for the estimators. In the third section we present the generalised James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ we obtain after shrinking the MLE according to the estimators in Section 4.2. Then in the last section we discuss the asymptotic distributions for the three estimators in play.

4.1. Parametric Structure

Our interest is to improve upon the MLE $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ by shrinking it towards the restricted estimator $\tilde{\boldsymbol{\theta}}_n^o$ which is well defined in Section 4.2.2. We note that the new estimator (shrinkage estimator) we obtain is also p-dimensional. We use the squared error loss to measure estimation efficiency of the shrinkage estimator. We begin by looking at the parametric structure before we discuss the concept of estimation.

Suppose that we observe a random array of vectors $\tilde{\boldsymbol{X}}_n = \{X_{1n}, ..., X_{nn}\}$ of independent and identically distributed (iid) random variables from a density $f_{\boldsymbol{\theta}}(X)$ indexed by a parameter $\boldsymbol{\theta} \in \Omega$ where Ω is a parameter space with its elements in $\mathbb{R}^p$ such that $X \sim N_p(\boldsymbol{\theta}, \mathbf{V})$. Let Ω_o be a sub-parameter space partitioned from the whole parameter space Ω by a parametric restriction

$$\Omega_o = \{\boldsymbol{\theta} \in \Omega : \mathbf{e}(\boldsymbol{\theta}) = \mathbf{0}\} \tag{4.1}$$

where $\mathbf{e}(\boldsymbol{\theta}) : \mathbb{R}^p \longrightarrow \mathbb{R}^m$. Let $\mathbf{E}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{e}(\boldsymbol{\theta})^\top$ be a shrinkage matrix. The sub-parameter space Ω_o provides a simple model of interest towards which to shrink. If $m = p$ then will have a singleton zero vector $\Omega_o = \{\mathbf{0}\}$ and if $m < p$, we have a sub-model of particular interest. In the restriction, $\boldsymbol{\theta}$ is simply the kernel of $\mathbb{R}^p$. Oman in [48] showed that this sub-parameter space can be linear or non linear. In this case we have a linear sub-parameter space and we are shrinking towards zero. We view $\boldsymbol{\theta}$ partitioned into two sub-parameter spaces as $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ with elements of the form

$$\left(\theta_1^{(1)}, ..., \theta_m^{(1)}, \theta_{m+1}^{(2)}, ..., \theta_{m+q}^{(2)}\right)$$

were $p = q + m$ is the dimension of $\boldsymbol{\theta}$. This means that $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^m$ and $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^q$ respectively. Therefore if $m = p$, we have a zero vector $(0, ..., 0)$ of p-dimensional and we consider this as a special case. Otherwise we consider the vector elements up to m-dimension. Thus we have created a sub-parameter space of dimension m,

hence $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)} \in \mathbb{R}^m$. Therefore the maximum likelihood estmator which maximises the parameter in this sub-parameter space is our shrinkage target and we call it the restricted maximum likelihood estimator (RMLE) denoted by $\tilde{\boldsymbol{\theta}}_n^o$. This is the same as what Amirdjanova and Woodroofe call target estimator in [2]. The other sub-parameter space $\boldsymbol{\theta}^{(2)}$ has the elements of the form $(\theta_{m+1}^{(2)}, ..., \theta_{m+q}^{(2)})$ with q-dimension. We regard the parameter $\boldsymbol{\theta}^{(2)}$ as "nuisance" because it does not play any role in creating a shrinkage target but it is important for the estimates. This sub-parameter space is similar to the one in the last example of Section 2 of Oman in [48] and Hansen in the study [28]. Though Hansen in [28] considers both the two partitioned sub-parameter spaces.

Therefore, considering that Ω_o is a linear subspace, we can write

$$\mathbf{e}(\boldsymbol{\theta}) = \mathbf{E}^\top \boldsymbol{\theta} - \mathbf{a} \tag{4.2}$$

where $\mathbf{E}$ is a $p \times m$ matrix and $\mathbf{a}$ is an $m \times 1$ vector. From these dimensions of the matrices $\mathbf{E}$ and $\mathbf{a}$, and knowing that the parameter $\boldsymbol{\theta}$ is a $p \times 1$ matrix, we obtain $\mathbf{e}(\boldsymbol{\theta})$ as an $m \times 1$ matrix. This implies that $\mathbf{e}(\boldsymbol{\theta})$ is maintained as sub-parameter space of m-dimension since it is an $m \times 1$ matrix. One common set up of $\mathbf{e}(\boldsymbol{\theta})$ we use is to shrink the elements of $\boldsymbol{\theta}$ to a common value, in which case we would set

$$\mathbf{E} = \begin{bmatrix} -1 & 0 & . & . & 0 \\ 1 & 0 & . & . & 0 \\ 0 & -1 & . & . & 0 \\ 0 & 1 & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & -1 \\ 0 & 0 & . & . & 1 \end{bmatrix}_{p \times m} \tag{4.3}$$

and $\mathbf{a} = \mathbf{0}$.This becomes appropriate since $\boldsymbol{\theta}$ values are disaggregate coefficients and we obtain a useful approximation by shrinking these coefficients towards a common value. We present the following examples which illustrate different choices we can impose on the matrices $\mathbf{E}$ and $\mathbf{a}$ to have a same linear sub-parameter space which allow us to shrink towards a zero vector.

Example 4.1.1

Suppose $p = 8$ and $m = 4$. Thus from (4.2) and (4.3) we have

$$\mathbf{E} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \end{bmatrix} \qquad \textit{and} \qquad \boldsymbol{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Taking the transpose of $\mathbf{E}$ *and multiplying with* $\boldsymbol{\theta}$ *using matrix product we have*

$$\mathbf{E}^\top\boldsymbol{\theta} = \begin{bmatrix}\theta_2-\theta_1\\ \theta_4-\theta_3\\ \theta_6-\theta_5\\ \theta_8-\theta_7\end{bmatrix} \quad \textit{implying} \quad \boldsymbol{e}(\boldsymbol{\theta}) = \mathbf{E}^\top\boldsymbol{\theta} - \boldsymbol{a} = \begin{bmatrix}\theta_2-\theta_1-0\\ \theta_4-\theta_3-0\\ \theta_6-\theta_5-0\\ \theta_8-\theta_7-0\end{bmatrix} = \begin{bmatrix}\theta_2-\theta_1\\ \theta_4-\theta_3\\ \theta_6-\theta_5\\ \theta_8-\theta_7\end{bmatrix} = \begin{bmatrix}\theta_1^{(1)}\\ \theta_2^{(1)}\\ \theta_3^{(1)}\\ \theta_4^{(1)}\end{bmatrix}.$$

Thus we have a linear sub-parameter from $\boldsymbol{\theta} \in \mathbb{R}^8$ *to* $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^4$*. In this case the "nuisance" parameters are in* $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^4$*.*

Example 4.1.2

Suppose we have the value of p *and the matrix* $\boldsymbol{\theta}$ *set as in Example 4.1.1. Let*

$$\mathbf{E} = \begin{bmatrix}1&0&0&0&0\\ 0&0&0&0&0\\ 0&1&0&0&0\\ 0&0&0&0&0\\ 0&0&1&0&0\\ 0&0&0&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\end{bmatrix} \quad \textit{and} \quad \boldsymbol{a} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}.$$

Then for $m = 5$ *and computing* $\boldsymbol{e}(\boldsymbol{\theta})$ *we have*

$$\boldsymbol{e}(\boldsymbol{\theta}) = \mathbf{E}^\top\boldsymbol{\theta} - \boldsymbol{a} = \begin{bmatrix}\theta_1-1\\ \theta_3-1\\ \theta_5-1\\ \theta_7-1\\ \theta_8-1\end{bmatrix} = \begin{bmatrix}\theta_1^{(1)}\\ \theta_2^{(1)}\\ \theta_3^{(1)}\\ \theta_4^{(1)}\\ \theta_5^{(1)}\end{bmatrix}.$$

Therefore we have a linear sub-parametric space $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^5$*. The other sub-parameter space remaining is* $\boldsymbol{\theta}^{(2)} = (\theta_3, \theta_4, \theta_6)$ *which we set as* $\left(\theta_1^{(2)}, \theta_2^{(2)}, \theta^{(2)}\right)$ *and refer to as "nuisance" parameters. So* $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^3$ *with* $\boldsymbol{\theta} \in \mathbb{R}^8$ *and* $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right)$*.*

In Example 4.1.1 if we let $\theta_j, j = 1, 2, ..., p$ have the same value, then we have $\mathbf{e}(\boldsymbol{\theta}) = \mathbf{0}$. If we let $\theta_j = 1, j = 1, 2, ..., p$ in Example 4.1.2 will have the same sub-parameter space (zeros) also. Therefore considering both conditions, it implies that in both examples will be shrinking towards zero. In this study we consider the case of Example 4.1.1.

In other cases, Ω_o may be a non-linear sub-parameter space. For example one of the restrictions could be $\mathbf{e}(\boldsymbol{\theta}) = \boldsymbol{\theta}^{(1)}\boldsymbol{\theta}^{(2)} - 1$ which would shrink the coefficients towards $\boldsymbol{\theta}^{(1)}\boldsymbol{\theta}^{(2)} = 1$. In general, shrinking towards a non-linear sub-parameter space may be useful when a statistical model or hypothesis implies a set of non-linear restrictions on the coefficients. But for this study we consider only shrinking towards a linear sub-parameter space.

4.1.1 Parameter of Interest and Shrinkage Dimension

The goal of shrinking the initial estimator $\hat{\boldsymbol{\theta}}_n$ towards a linear sub-parameter space is to obtain a James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ which estimates $\boldsymbol{\theta}$ the parameter of interest of p-dimensional. In this study the whole parameter $\boldsymbol{\theta}$ is of interest

apart from when we are obtaining the shrinkage target $\tilde{\boldsymbol{\theta}}_n^o$ then we consider the m-dimensional parameter space. Let $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\theta})$ for some differentiable function $\boldsymbol{g} : \mathbb{R}^m \longrightarrow \mathbb{R}^p$. Let $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\theta})^\top$ be a matrix harmonising shrinkage dimension m and the dimension of interest p. The function $\boldsymbol{g}$ increases the dimension of the estimator by increasing the corresponding entries of the estimator if it is less than p. We introduce this function to ensure that the dimension of the estimators match with the parameter we are estimating which is of dimension p. We also note that the entries in an estimator are not uniquely corresponded to each entry of the vector of estimates because the random variables are independent and identically distributed (`iid`). To be specific, the restricted maximum likelihood estimator $\tilde{\boldsymbol{\theta}}_n^o$ estimating $\boldsymbol{\theta} \in \Omega_o$ has dimension m which is less than p. Therefore to harmonise the dimension, the function $\boldsymbol{g}$ increases the dimension from m to p through a plug-in estimator $\tilde{\boldsymbol{\beta}}_n^o = \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_n^o)$. If $m = p$ the function $\mathbf{g}$ maintains the dimension since it is the same. This means that the matrix $\mathbf{G} = \mathbf{G}(\tilde{\boldsymbol{\theta}}_n^o)$ is of dimension $m \times p$ which changes to $p \times m$ when we get the transpose $\mathbf{G}^\top$ and of dimension $p \times p$ when $\mathbf{G} = \mathbf{I}_p$. The structure of the matrix $\mathbf{G}$ is that the element entries are ones in the main diagonal and the rest are zeros, and if it is not a square matrix the last column will have ones from the last row of square to the last entry which is dependant on the function $\boldsymbol{g}$.

Therefore in this context we see that we have introduced $\boldsymbol{\beta}$ which is the parameter in a shrinkage sense. Since we are considering a general estimation case, when $\boldsymbol{\beta} = \boldsymbol{\theta}$ the entire parameter space is of interest apart from when we are choosing the shrinkage target. Hence, in this case $\mathbf{G} = \mathbf{I}_p$. The shrinkage direction $\mathbf{e}(\boldsymbol{\theta})$ may, but does not necessarily need to, relate to the parameter of interest which is determined by the map $\boldsymbol{g}(\boldsymbol{\theta})$.

4.1.2 Loss and Risk Function

Here we give some background concept on the loss and risk functions so that we understand the link between risk and loss in terms of the mean squared error (`MSE`). Also since in the study we quantify the shrinkage estimator by lower mean squared error, it is therefore important to study these two functions (loss and risk function). The loss function in a point estimation problem like in this case for this study reflects the fact that if the estimator $\mathbf{T}$ is close to the parameter $\boldsymbol{\theta}$, then the decision made from $\mathbf{T}$ is reasonable and little loss is incurred. Therefore the estimator is regarded to have lower mean squared error and effective. If the estimator $\mathbf{T}$ is far from $\boldsymbol{\theta}$, then the loss incurred is large. In this case the estimator $\mathbf{T}$ is said to have high mean squared error and less effective. Therefore, the loss function is a non-negative function that generally increases as the distance between the estimator $\mathbf{T}$ and the parameter $\boldsymbol{\theta}$ increases.

With all parameters set as described in the previous sections, we consider the loss function $\ell(\boldsymbol{\theta}, \mathbf{T})$ which is the same as that considered by Berger in [4] and require that it should have a second derivative to make it easier to expand up to the second order of the Taylor's theorem in our proofs as considered by Hansen in [28]. Secondly, as in [28] we require that the loss function $\ell(\boldsymbol{\theta}, \mathbf{T})$ is uniquely minimised at $\mathbf{T} = \boldsymbol{\theta}$. This is required in the main results when analysing the asymptotic distribution and asymptotic consistency of the James-Stein shrinkage estimator (`JSSE`).

From Definition 3.2.7, we have one particular loss function given by

$$\ell(\theta, \mathbf{T}) = [\mathbf{T} - \theta]^2 \tag{4.4}$$

called the squared error loss function whose corresponding risk called the mean squared error (MSE) is given by

$$R(\theta, \mathbf{T}) = \mathbf{MSE}_\theta(\mathbf{T}) = \mathbb{E}_\theta [\mathbf{T} - \theta]^2 \tag{4.5}$$

where the expectation is taken with respect to the parameter θ. Another loss function is

$$\ell(\boldsymbol{\theta}, \mathbf{T}) = \mid \mathbf{T} - \boldsymbol{\theta} \mid \tag{4.6}$$

which we referred to as the absolute error loss function in Definition 3.2.7. Its corresponding risk, called the mean absolute error, is given by

$$R(\boldsymbol{\theta}, \mathbf{T}) = \mathbb{E}_{\boldsymbol{\theta}} \mid \mathbf{T} - \boldsymbol{\theta} \mid \tag{4.7}$$

where the loss function $\ell(\boldsymbol{\theta}, \mathrm{T})$ is averaged with respect to $\boldsymbol{\theta}$. Now, in this study we only consider the form (4.4) with its corresponding risk (MSE) (4.5). Since we are working with a multivariate case, the loss function (4.4) is expressed as

$$\ell(\boldsymbol{\theta}, \mathbf{T}) = (\mathbf{T} - \boldsymbol{\theta})^\top \mathbf{W} (\mathbf{T} - \boldsymbol{\theta}) \tag{4.8}$$

for some weight matrix $\mathbf{W} > 0$. The equation for the mean squared error (MSE) in (4.5) is the same as the one expressed explicitly in Lemma 3.2.10 which is the expression we use for the simulation plots in Chapter 5 to calculate the mean squared error values for both the MLE and James-Stein shrinkage estimator (JSSE). Considering the form (4.8) further, in this study the weight matrix $\mathbf{W}$ will be equal to the covariance matrix. Thus we have the case that either $\mathbf{W} = \mathbf{I}_p$ the identity matrix or $\mathbf{W} = \mathbf{V}^{-1}$ the inverse of the covariance matrix. We will prefer using $\mathbf{W} = \mathbf{V}^{-1}$ in most of the analysis because according to Hansen in [28] this choice renders the loss function invariant to rotations of the coefficient of the parameter vector $\boldsymbol{\theta}$. Hence depending on a statistical problem involved, the squared loss (4.8) assumes different forms. For example, in the following statistical problem given by Hansen in [28], the loss function is averaged using the integrated distance.

Example 4.1.3 *([28], Example 1)*

In the (possibly non-linear) regression model $y_i = g(x_i, \beta) + e_i$ with $e_i \sim f(e, \eta)$ for some parametric density $f(e, \eta)$, an estimate of the conditional mean function takes the form $g_{\boldsymbol{T}}(X)$. A common measure of accuracy is the integrated distance, which is

$$\ell(\boldsymbol{\beta}, \boldsymbol{T}) = \int \left(g_{\boldsymbol{T}}(X) - g_{\boldsymbol{\beta}}(X)\right) w(X) dX$$

for some smooth distance measure $d(u) \geq 0$. In general, $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ is a non quadratic yet smooth function of $\boldsymbol{T}$. If the regression function is linear in the parameters, $g_{\boldsymbol{\beta}}(X) = X^\top \boldsymbol{\beta}$, and $d(u) = u^2$ is quadratic, then the integrated squared error equals

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{T}) &= \int \left(X^\top \boldsymbol{T} - X^\top \boldsymbol{\beta}\right)^2 w(X) dX \\ &= (\boldsymbol{\beta} - \boldsymbol{T})^\top \boldsymbol{W} (\boldsymbol{\beta} - \boldsymbol{T}) \end{aligned}$$

with $\boldsymbol{W} = \int XX^\top w(X) dX = \mathbb{E}_{\boldsymbol{\beta}} \left[X_i X_i^\top w(X_i)\right]$, and thus takes the form (4.8) with a specific weight matrix.

Therefore, the consideration of the whole parameter space as the parameter of interest in this study makes the theory of shrinkage simplified than Hansen's consideration in the study [28], where the parameter of interest is just part of the whole parameter space. Hence with this consideration we use a simplified setting by letting $\mathbf{W} = \mathbf{V}^{-1}$ in the loss function (4.8). Under this setting many of the formulae simplify, and therefore we have $\mathbf{G} = \mathbf{I}_p$ and $\mathbf{W} = \mathbf{V}^{-1}$ which Hansen in [28] called the canonical case. In practical applications it may be a little restrictive to consider the whole parameter vector as the parameter of interest but we do that as in this case we focus on the general estimation problem and our main interest is on the asymptotic behaviour of the shrinkage estimator we obtain.

4.2. Estimation

We now present the estimation process. We state the functions, estimates and estimators of interest according to the parametric set up. We begin by stating the likelihood function and develop it up to the maximum likelihood estimators using Definitions 3.2.21 to 3.2.25 in Chapter 3.

The likelihood function for the sample is

$$\mathbf{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i) \tag{4.9}$$

and the log likelihood function is given by

$$\boldsymbol{\ell}_n(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i)$$

which becomes

$$\boldsymbol{\ell}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i) \tag{4.10}$$

when we take the log of the product, hence the score function is

$$\mathbf{S}_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i) \right). \tag{4.11}$$

We consider two standard estimators of $\boldsymbol{\theta}$. The unrestricted maximum likelihood estimator (`MLE`) which maximises (4.10) over $\boldsymbol{\theta} \in \Omega$

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Omega} \boldsymbol{\ell}_n(\boldsymbol{\theta})$$

and the restricted `MLE` which maximizes (4.10) over $\boldsymbol{\theta} \in \Omega_o$

$$\tilde{\boldsymbol{\theta}}_n^o = \arg\max_{\boldsymbol{\theta} \in \Omega_o} \boldsymbol{\ell}_n(\boldsymbol{\theta})$$

The information matrix is given by

$$\mathbf{I}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i) \tag{4.12}$$

and the fisher information is given by

$$\mathbf{J}_n(\boldsymbol{\theta}) = \mathbb{E}_\theta \left[\mathbf{I}_n(\boldsymbol{\theta})\right]$$

$$\mathbf{J}_n(\boldsymbol{\theta}) = \mathbb{E}_\theta \left[\mathbf{S}_n(X_i, \boldsymbol{\theta})\mathbf{S}_n(X_i, \boldsymbol{\theta})^\top\right]$$

where $\mathbf{S}(x, \theta) = \frac{\partial}{\partial\theta} \log f_\theta(x)$. Therefore proceeding we get

$$\mathbf{J} = \mathbf{J}_n(\boldsymbol{\theta}) = \mathbb{E}_\theta \left[\sum_{i=1}^{n} -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i)\right]$$

and using scaled log likelihood function $\boldsymbol{\ell}_n = \frac{1}{n}\sum_{i=1}^{n} \log f_{\hat{\boldsymbol{\theta}}_n}(X_i)$ we have

$$\widehat{\mathbf{J}}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\hat{\boldsymbol{\theta}}_n}(X_i) \tag{4.13}$$

a consistent estimator for $\mathbf{J}_n(\boldsymbol{\theta})$. Now the variance can be found using

$$\mathbf{V}_n = (\mathbf{J}_n(\boldsymbol{\theta}))^{-1}$$

and it gives

$$\mathbf{V} = \mathbf{V}_n(\boldsymbol{\theta}) = \left(\mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i)\right]\right)^{-1} \tag{4.14}$$

which is the covariance matrix and its estimate is given by

$$\widehat{\mathbf{V}}_n = \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\hat{\boldsymbol{\theta}}_n}(X_i)\right)^{-1} \tag{4.15}$$

which is the one we use in the analysis of the set up of estimators.

4.2.1 Unrestricted Estimator

We consider a maximum likelihood estimator maximising $\boldsymbol{\theta}$ in the whole parameter space Ω. We refer to this estimator as the unrestricted estimator. Therefore, the standard estimator of $\boldsymbol{\theta}$ is the unrestricted maximum likelihood estimator (`MLE`). From (4.21), the log likelihood function is

$$\boldsymbol{\ell}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i).$$

The `MLE` maximises this likelihood function over $\boldsymbol{\theta} \in \Omega$,

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Omega} \boldsymbol{\ell}_n(\boldsymbol{\theta}).$$

We assume that the maximum is unique so that $\hat{\boldsymbol{\theta}}_n$ is well defined. Thus we rely on the property stated in Proposition 3.2.27 in the preliminaries. Let $\widehat{\mathbf{V}}_n$ denote any consistent estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_n$ as given in (4.15), such as

$$\widehat{\mathbf{V}}_n = \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\hat{\boldsymbol{\theta}}_n}(X_i)\right)^{-1}.$$

The `MLE` for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_n = \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$ given as a function of $\hat{\boldsymbol{\theta}}_n$ with the function $\boldsymbol{g}$ allowing the parameter dimension go up to p. Let $\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = \widehat{\mathbf{G}}_n^\top \widehat{\mathbf{V}}_n \widehat{\mathbf{G}}_n$ denote the standard estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}_n$, where $\widehat{\mathbf{G}}_n = \mathbf{G}(\hat{\boldsymbol{\theta}}_n)$. We use this set up in the construction of the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$.

4.2.2 Restricted Estimator

In this section we investigate the set up of the restricted maximum likelihood estimator $\tilde{\boldsymbol{\theta}}_n^o$. This is the shrinking target, meaning that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ is pulled towards $\tilde{\boldsymbol{\theta}}_n^o$ when we shrink it. To achieve the objective of shrinking towards the sub-parameter space Ω_o, we define the restricted maximum likelihood estimator $\tilde{\boldsymbol{\theta}}_n^o$ which satisfies the condition $\tilde{\boldsymbol{\theta}}_n^o \in \Omega_o$. To have this condition satisfied, Hansen [28] considers three possibilities defined below.

1. Restricted maximum likelihood estimator (`RMLE`)

$$\tilde{\boldsymbol{\theta}}_n^R = \arg\max_{\boldsymbol{\theta}\in\Omega_o} \boldsymbol{\ell}_n(\boldsymbol{\theta}) \tag{4.16}$$

2. Efficient minimum distance (`EMD`)

$$\tilde{\boldsymbol{\theta}}_n^E = \arg\max_{\boldsymbol{\theta}\in\Omega_o} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top \widehat{\mathbf{V}}_n^{-1} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \tag{4.17}$$

3. Projection

$$\tilde{\boldsymbol{\theta}}_n^P = \hat{\boldsymbol{\theta}}_n - \widehat{\mathbf{V}}_n \widehat{\mathbf{E}}_n (\widehat{\mathbf{E}}_n^\top \widehat{\mathbf{V}}_n \widehat{\mathbf{E}}_n)^{-1} \mathbf{e}(\hat{\boldsymbol{\theta}}_n) \tag{4.18}$$

   for $\widehat{\mathbf{E}}_n = \mathbf{E}(\hat{\boldsymbol{\theta}}_n)$.

Hansen analyses these three possibilities as all satisfying the condition $\tilde{\boldsymbol{\theta}}_n^R$, $\tilde{\boldsymbol{\theta}}_n^E$, $\tilde{\boldsymbol{\theta}}_n^P \in \Omega_o$, for linear restrictions (4.1). Since we are only considering a linear sub-parameter space Ω_o, any of the three restricted estimators will represent the restricted estimator. He notes that the three estimators are asymptotically equivalent under the assumptions in the next section and thus we generally write the restricted estimator as $\tilde{\boldsymbol{\theta}}_n^o$ without drawing a distinction between (4.16), (4.17) and (4.18). It would also be possible to consider other distance or projection estimators as in (4.17) or (4.18) but with the weight matrix $\widehat{\mathbf{V}}^{-1}$ replaced with another choice. Other choices, however, would lead to asymptotically inefficient estimators, so we confine our attention to these three estimators defined by Hansen [28]. Given the estimator $\tilde{\boldsymbol{\theta}}_n^o$, we note that it is of dimension m since it is in Ω_o. Therefore to increase the dimension to p, we have a plug-in restricted estimator for $\boldsymbol{\beta}$ given by $\tilde{\boldsymbol{\beta}}_n^o = \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_n^o)$.

Analysing further the general form of the restricted estimator $\tilde{\boldsymbol{\theta}}_n^o$, assuming that for some symmetric matrix $\mathbf{A}$ such that $\text{rank}(\mathbf{AE}) = p$, then we have

$$\tilde{\boldsymbol{\theta}}_n^o = \hat{\boldsymbol{\theta}}_n - \mathbf{AE}(\mathbf{E}^\top \mathbf{AE})^{-1} \mathbf{e}(\hat{\boldsymbol{\theta}}_n) + O_p(n^{-\frac{1}{2}})$$

where

$$O_p(n^{-\frac{1}{2}}) = \lim_{n\to\infty} \frac{1}{\sqrt{n}}$$

.

as defined in [12]. This asserts that the covariance matrix estimate for the unrestricted estimator $\hat{\boldsymbol{\theta}}_n$ is consistent, which shows that the restricted estimator can be asymptotically constructed from the unrestricted estimator. The assumption that $\mathbf{AE}$ is of full rank is required so that $\tilde{\boldsymbol{\theta}}_n^o$ is well defined. It allows the matrix

$\mathbf{A}$ to be deficient rank (not of full rank), but in this case $\mathbf{E}$ cannot lie in the null space of $\mathbf{A}$. A simple restricted estimator arising from this assertion is

$$\tilde{\boldsymbol{\theta}}_n^o = \hat{\boldsymbol{\theta}}_n - \widehat{\mathbf{A}}\widehat{\mathbf{E}}(\widehat{\mathbf{E}}^\top \widehat{\mathbf{A}}\widehat{\mathbf{E}})^{-1}\mathbf{e}(\hat{\boldsymbol{\theta}}_n) \tag{4.19}$$

where $\widehat{\mathbf{E}} = \mathbf{E}(\hat{\boldsymbol{\theta}}_n)$ and $\widehat{\mathbf{A}}_n$ is a consistent estimator of $\mathbf{A}$. In particular we have,

$$\tilde{\boldsymbol{\theta}}_n^o = \hat{\boldsymbol{\theta}}_n - \widehat{\mathbf{V}}\widehat{\mathbf{E}}(\widehat{\mathbf{E}}^\top \widehat{\mathbf{V}}\widehat{\mathbf{E}})^{-1}\mathbf{e}(\hat{\boldsymbol{\theta}}_n)$$

which is the same as the projection (4.18) and this is asymptotically efficient under $\mathbf{e}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. When $\mathbf{e}(\boldsymbol{\theta})$ is linear the restricted estimator will typically take the first form of (4.18), and in this case $\mathbf{e}(\tilde{\boldsymbol{\theta}}_n^o) = \mathbf{0}$ so $\tilde{\boldsymbol{\theta}}_n^o \in \Omega_o$. When $\mathbf{e}(\boldsymbol{\theta})$ is non-linear then a typical restricted estimator will not satisfy (4.18) exactly.

4.3. Generalised James-Stein Shrinkage Estimator

The purpose of this section is to present a generalised James-Stein shrinkage estimator and use it to find the asymptotic distribution which is required for our results in Chapter 5. We use concepts from the studies [7], [8], [13], [21] and [37].

4.3.1 Shrinkage Estimator

Suppose the `MLE` $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta} \in \Omega$ with elements in $\mathbb{R}^p$ is distributed as $N_p(\boldsymbol{\theta}, \mathbf{V})$ and $\tilde{\boldsymbol{\theta}}_n^o$ is the restricted `MLE`. To state the generalised James-Stein shrinkage estimator we use the estimators $\hat{\boldsymbol{\beta}}_n = \mathbf{g}(\hat{\boldsymbol{\theta}}_n)$ and $\tilde{\boldsymbol{\beta}}_n^o = \mathbf{g}(\tilde{\boldsymbol{\theta}}_n^o)$ as functions of $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n^o$ respectively for easy link up of the asymptotic distributions and shrinking strategy (direction). Using the form (3.15) in Chapter 3, the generalised positive part James-Stein shrinkage estimator is given by

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n - \left(\frac{p-2}{n(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)^\top \mathbf{V}^{-1}(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)} \right)_+ (\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o) \tag{4.20}$$

where $(x)_+ = x1(x \geq 0)$ is a positive trimming function which keeps the value in the brackets always positive or zero. Now, let $D_n = n(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)^\top \mathbf{V}^{-1}(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)$, then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n^* &= \hat{\boldsymbol{\beta}}_n - \left(\frac{p-2}{D_n} \right)_+ (\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o) \\
&= \hat{\boldsymbol{\beta}}_n - \left(\frac{p-2}{D_n} \right)_+ \hat{\boldsymbol{\beta}}_n + \left(\frac{p-2}{D_n} \right)_+ \tilde{\boldsymbol{\beta}}_n^o \\
&= \left(1 - \frac{p-2}{D_n} \right)_+ \hat{\boldsymbol{\beta}}_n + \left(\frac{p-2}{D_n} \right)_+ \tilde{\boldsymbol{\beta}}_n^o \\
&= \left(1 - \frac{p-2}{D_n} \right)_+ \hat{\boldsymbol{\beta}}_n + \left(1 - \left(1 - \frac{p-2}{D_n} \right)_+ \right) \tilde{\boldsymbol{\beta}}_n^o
\end{aligned}$$

$$\therefore \hat{\boldsymbol{\beta}}_n^* = \left(1 - \frac{p-2}{D_n} \right)_+ \hat{\boldsymbol{\beta}}_n + \left(1 - \left(1 - \frac{p-2}{D_n} \right)_+ \right) \tilde{\boldsymbol{\beta}}_n^o. \tag{4.21}$$

Furthermore, let $\widehat{\tau}_n = p - 2$ and

$$\widehat{w} = \left(1 - \frac{p-2}{D_n}\right)_+$$

which becomes

$$\widehat{w} = \left(1 - \frac{\widehat{\tau}_n}{D_n}\right)_+ \tag{4.22}$$

where $(x)_+$ is as defined earlier on. Then (4.21) becomes

$$\hat{\boldsymbol{\beta}}_n^* = \widehat{w}\hat{\boldsymbol{\beta}}_n + (1 - \widehat{w})\tilde{\boldsymbol{\beta}}_n^o \tag{4.23}$$

expressed in terms of $\widehat{w}$, $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}_n^o$ as a weighted average. Therefore, according to Hansen in [28], the proposed generalised James-Stein shrinkage estimator of $\boldsymbol{\beta}$ is a weighted average of the `MLE` and the restricted estimator where the weight $\widehat{w}$ takes the form (4.22) and "+" keeps the value of $\widehat{w}$ to be always non-negative.

We analyse the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ of the form (4.23) as a weighted average of the `MLE` on the amount of shrinkage using the weight $\widehat{w}$. When the weight $\widehat{w} = 0$ the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^* = \tilde{\boldsymbol{\beta}}_n^o$ the restricted estimator, and when $\widehat{w} = 1$, we have $\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n$ the unrestricted estimator (`MLE`). The condition $\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n$ $(p = 2)$ implies that there is no shrinkage and $\hat{\boldsymbol{\beta}}_n^* = \tilde{\boldsymbol{\beta}}_n^o$ implies full shrinkage. Therefore it is worthy to analyse $\widehat{w}$ further because it determines the amount of shrinkage imposed on the unrestricted estimator $\hat{\boldsymbol{\beta}}$. From (4.22), it is clear that the fraction $\frac{\widehat{\tau}_n}{D_n}$ determines the value of $\widehat{w}$ for $p - 2$ and

$$D_n = n\ell(\hat{\boldsymbol{\beta}}_n, \tilde{\boldsymbol{\beta}}_n^o) \tag{4.24}$$

where n is the sample size value and p is the dimension of the parameter. There are two possible conditions which can make $\widehat{w} = 0$, thus when either the fraction $\frac{\widehat{\tau}_n}{D_n} = 1$ or $\frac{\widehat{\tau}_n}{D_n} > 1$. This means that when $\widehat{\tau}_n > D_n$ we obtain a negative value of $\widehat{w}$, and when we apply the positive trimming function "+" which is well explained in [3], we get $\widehat{w} = 0$. Therefore in this case we have full shrinkage up to the shrinking target $\tilde{\boldsymbol{\beta}}_n^o$. When $D_n > \widehat{\tau}_n$ which implies $\frac{\widehat{\tau}_n}{D_n} < 1$, shrinkage takes place and both $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}_n^o$ are involved and the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ will be just an average of the two estimators. Hence the fraction $\frac{\widehat{\tau}_n}{D_n}$ depends on the dimension of the parameter $\boldsymbol{\theta}$ since $\widehat{\tau}_n = p - 2$ and also the scaled distance between the `MLE` $\hat{\boldsymbol{\beta}}_n$ and `RMLE` $\tilde{\boldsymbol{\beta}}_n^o$ since D_n is given by

$$D_n = n(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)^\top \mathbf{V}^{-1}(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o) \tag{4.25}$$

a scaled loss function with sample size value n and $\mathbf{W} = \mathbf{V}^{-1}$ as the default choice. Hansen in [28] states that the full shrinkage estimator is the classic James-Stein shrinkage estimator and the partial shrinkage estimator with linear $\mathbf{e}(\boldsymbol{\theta})$ is similar to Oman's [47] shrinkage estimator. He further stresses that the partial shrinkage estimator is also a special case of Hansen's [26] Mallows Model Averaging (`MMA`) estimator.

4.3.2 Asymptotic Distribution

In this section we analyse the asymptotic distribution of the James-Stein shrinkage estimator (`JSSE`) $\hat{\boldsymbol{\beta}}_n^*$. We borrow Hansen's [28] approach and methods to show

that the James-Stein shrinkage estimator asymptotically converges to some normal distribution determined by the asymptotic distribution of the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_n$ and the restricted maximum likelihood estimator (RMLE) $\tilde{\boldsymbol{\beta}}_n^o$. This is the asymptotic distribution we use in Chapters 5 to show the results of the study. In [28], Hansen uses the local asymptotic normality approach of Le Cam [40] and van der Vaart [55] and analyses the convergence in distribution by considering parameter sequences of the form

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + n^{-\frac{1}{2}} h \tag{4.26}$$

where $\boldsymbol{\theta}_n \in \Omega_o$ and $h \in \mathbb{R}^P$. In the sequence $\boldsymbol{\theta}_n$ in (4.26), $\boldsymbol{\theta}_o$ is the true value and the centering value with h as a localising parameter. Since we have $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\theta})$ as the parameter of interest, then we have $\boldsymbol{\beta}_n = \boldsymbol{g}(\boldsymbol{\theta}_n)$ and $\boldsymbol{\beta}_o = \boldsymbol{g}(\boldsymbol{\theta}_o)$. According to Hansen [28], the centering value $\boldsymbol{\theta}_o$ and the sequence $\boldsymbol{\theta}_n$ are localised to the restricted parameter space Ω_o while $n^{-\frac{1}{2}}$ for a fixed h provides a neighbourhood for $\boldsymbol{\theta}_o$. The neighbourhood shrinks (reduces) as $n \longrightarrow \infty$ to have $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$. Also we note that $\lim_{n \longrightarrow \infty} \boldsymbol{\theta}_n = \boldsymbol{\theta}$ the parameter. Therefore the analysis in asymptotic distribution for the estimators is investigated along the sequences $\boldsymbol{\theta}_n$.

We now state a standard set of regularity conditions sufficient for asymptotic normality of the conventional estimators.

Assumption 4.3.1 *([28],Assumption 1)*

1. *The observations X_i are independent and identically distributed (iid) drawn from the density $f_{\boldsymbol{\theta}_n}(X)$, where $\boldsymbol{\theta}_n$ satisfies (4.26), $\boldsymbol{\theta}_o$ is in the interior of Ω_o, and Ω is compact.*
2. *If $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\top$ then $f_{\boldsymbol{\theta}}(X) \neq f_{\boldsymbol{\theta}^\top}(X)$.*
3. *$\log f_{\boldsymbol{\theta}}(X)$ is continuous at each $\boldsymbol{\theta} \in \Omega$ with probability one.*
4. $\mathbb{E}_{\theta_o}\left[\sup_{\boldsymbol{\theta}\in\Omega} |\log f_{\boldsymbol{\theta}}(X_i)|\right] < \infty$.
5. $\mathbb{E}_{\theta_o}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}_o}(X_i)\right]$ *exists and is non-singular.*
6. *For some neighbourhood $\wp$ of $\boldsymbol{\theta}_o$,*
   (a) *$f_{\boldsymbol{\theta}}(X)$ is twice continuously differentiable,*
   (b) $\int \sup_{\boldsymbol{\theta}\in\wp} \parallel \frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(X) \parallel dX < \infty$,
   (c) $\int \sup_{\boldsymbol{\theta}\in\wp} \parallel \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f_{\boldsymbol{\theta}}(X) \parallel dX < \infty$,
   (d) $\mathbb{E}_{\theta_o} \sup_{\boldsymbol{\theta}\in\wp} \parallel \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i) \parallel < \infty$.
7. *$\boldsymbol{E}(\boldsymbol{\theta})$ is continuous in some neighbourhood of $\boldsymbol{\theta}_o$, and rank $(\boldsymbol{E}(\boldsymbol{\theta}_o)) = m$.*
8. *$\boldsymbol{G}(\boldsymbol{\theta})$ is continuous in some neighbourhood of $\boldsymbol{\theta}_o$.*

We highlight each assumption for better understanding. Part 1-6 of Assumption 4.3.1 are the conditions listed in Theorem 3.3 of Newey and McFadden [46] for the asymptotic normality of the MLE $\hat{\boldsymbol{\theta}}_n$. These assumptions establishes the regularity conditions for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ to converge asymptotically to a normal distribution. Part 1 of Assumption 4.3.1 specifies that the observations

are independent and identically distributed, the parameter satisfies the local sequences (4.26) and the centering $\boldsymbol{\theta}_o$ value is in the interior of Ω_o so that Taylor's expansion methods can be employed. Part 2 of Assumption 4.3.1 establishes identification of the maximum likelihood estimator in the process of solving the score function. Part 3 and 4 of Assumption 4.3.1 are used to establish consistency of $\hat{\boldsymbol{\theta}}_n$ which is an important property for the results on consistency of the James-Stein shrinkage estimator. Part 5 of Assumption 4.3.1 requires the Fisher information to exist and to be non-singular. Part 6 of Assumption 4.3.1 are regularity conditions to establish asymptotic normality. Part 7 and 8 of Assumption 4.3.1 allow asymptotic normality to extend to the estimators $\hat{\boldsymbol{\beta}}_n$, $\tilde{\boldsymbol{\beta}}_n^o$, and $\hat{\boldsymbol{\beta}}_n^*$ and these are the main conditions for the continuous mapping theorem which we use when obtaining the asymptotic distribution of the James-Stein shrinkage estimator.

We therefore specify regularity conditions for the loss function $\ell(\boldsymbol{\beta}, \mathbf{T})$ used in Theorem 4.3.2.

Assumption 4.3.2

The loss function $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ satisfies

1. $\ell(\boldsymbol{\beta}, \boldsymbol{T}) \geq 0$
2. $\ell(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$
3. $\boldsymbol{W}(\boldsymbol{\beta}) = \frac{1}{2}\frac{\partial^2}{\partial \boldsymbol{T} \partial \boldsymbol{T}^\top}\ell(\boldsymbol{\beta}, \boldsymbol{T}) \mid_{\boldsymbol{T}=\boldsymbol{\beta}}$ *is continuous in a neighbourhood of* $\boldsymbol{\beta}_o$.

Part 1 and 2 of Assumption 4.3.2 are conventional properties of the loss function which are as discussed in Chapter 1.1 of Lehmann and Casella [42]. Thus the non negative property of the loss function. Part 3 of Assumption 4.3.2 is stronger, requiring the loss function to be smooth (locally quadratic) and is as stated in [28]. Under squared loss (4.8), $\mathbf{W}(\boldsymbol{\beta}) = \mathbf{W}$. In general, we define $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta}_o)$ where $\boldsymbol{\beta}_o$ is the assumed true parameter value in the shrinkage sense.

We state the following lemma on the asymptotic distribution of the maximum likelihood estimator (`MLE`). The lemma is used in the proof for the asymptotic distribution of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$.

Lemma 4.3.1

Suppose that the estimator $\hat{\boldsymbol{\theta}}_n$ satisfies the conditions under the regularity Assumptions 4.3.1. Then along the sequences 4.26 we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \longrightarrow_d Z \sim N_p(\boldsymbol{0}, \boldsymbol{V}) \tag{4.27}$$

where $\hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator, n is the sample size value and $\boldsymbol{\theta}$ is the parameter of dimension p.

Proof

Let $\boldsymbol{\theta}_o$ be equal to the "True" parameter value. Let $S_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta})$, $S_n'(\boldsymbol{\theta}) = \frac{\partial S}{\partial \boldsymbol{\theta}}$, $S_n''(\boldsymbol{\theta}) = \frac{\partial S_n'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Then from the Mean Value Theorem (`MVT`) there exist $\bar{\boldsymbol{\theta}} \in (\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_o)$ such that

$$S_n''(\bar{\boldsymbol{\theta}}) = \frac{S_n'(\hat{\boldsymbol{\theta}}_n) - S_n'(\boldsymbol{\theta}_o)}{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o},$$

and making $S_n'(\hat{\boldsymbol{\theta}}_n)$ *the subject of the formula we have*

$$S_n'(\hat{\boldsymbol{\theta}}_n) = S_n'(\boldsymbol{\theta}_o) + S_n''(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o).$$

But by definition $S_n'(\hat{\boldsymbol{\theta}}_n) = 0$*, therefore*

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) = -S_n''(\bar{\boldsymbol{\theta}}_n)^{-1} S_n'(\boldsymbol{\theta}_o)$$

which becomes

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) = -\left(\frac{1}{n} S_n''(\bar{\boldsymbol{\theta}}_n)\right)^{-1} \frac{1}{\sqrt{n}} S_n'(\boldsymbol{\theta}_o). \tag{4.28}$$

when we introduce $\sqrt{n}$ *both sides. Let*

$$A = -\left(\frac{1}{n} S_n''(\bar{\boldsymbol{\theta}}_n)\right)^{-1} \tag{4.29}$$

in equation (4.28). We have

$$plim \frac{1}{n} S_n''(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \quad \forall \quad \boldsymbol{\theta} \notin g(\boldsymbol{\theta}_o) \neq 0 \quad from \ \ Assumption \ 4.3.1.$$

Using part 3 and 4 of Assumption 4.3.1, if

$$plim \ \hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_o$$

then

$$\bar{\boldsymbol{\theta}}_n \in (\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_o) \Rightarrow \bar{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}_o$$

for some assumed true value $\boldsymbol{\theta}_o$*. Therefore, from (4.29)*

$$plim \ A = -\left(g(\boldsymbol{\theta}_o)\right)^{-1} = \boldsymbol{Q}$$

a positive definite matrix. From (4.28), $\frac{1}{n} S_n'(\boldsymbol{\theta}_o)$ *is a sample average such as* $\frac{1}{n} S_n'(\boldsymbol{\theta}_o) = \frac{1}{n} \sum \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$*. Therefore by the central limit theorem (*`CLT`*) we have*

$$\sqrt{n}\left(\frac{1}{n} S_n'(\boldsymbol{\theta}_o) - \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n} S_n'(\boldsymbol{\theta}_o)\right]\right) \longrightarrow N_p(\boldsymbol{0}, \boldsymbol{V}) \tag{4.30}$$

where $\boldsymbol{V} = Var\left(\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$*. Now*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n} S_n'(\boldsymbol{\theta}_o)\right] &= \frac{1}{n} \sum \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial \log \boldsymbol{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \\ &= \frac{1}{n} \sum 0 \\ &= 0 \end{aligned}$$

since $\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial \log \boldsymbol{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = 0$ *from Lemma 3.2.33. Thus (4.30) becomes*

$$\sqrt{n}\left(\frac{1}{n} S_n'(\boldsymbol{\theta}_o)\right) = \frac{1}{\sqrt{n}} S_n'(\boldsymbol{\theta}_o) \longrightarrow N_p(\boldsymbol{0}, \boldsymbol{V}). \tag{4.31}$$

Hence from (4.28) and (4.31) we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) \longrightarrow N_p(\boldsymbol{0}, \boldsymbol{Q}\boldsymbol{V}\boldsymbol{Q})$$

for a positive definite matrix $\boldsymbol{Q}$*. Using Lemma 3.2.34 we have*

$$\boldsymbol{V}\left[\frac{\partial \log f_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}}\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right],$$

therefore

$$plim\ \frac{1}{n}S''(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[S''(\boldsymbol{\theta}_o)\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right]$$

$$plim\ A = \boldsymbol{Q} = -\mathbb{E}_{\boldsymbol{\theta}}\left[S''(\boldsymbol{\theta}_o)\right]^{-1},$$

but

$$-\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}_o\partial \boldsymbol{\theta}_o}\right] = \boldsymbol{J}$$

the fisher information matrix, and

$$\boldsymbol{J} = \boldsymbol{V}^{-1}.$$

Since $\boldsymbol{V}$ *is the covariance matrix from a multivariate standard normal then*

$$\boldsymbol{V}^{-1} = \boldsymbol{V}.$$

Also we have

$$-\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}_o\partial \boldsymbol{\theta}_o}\right] = \boldsymbol{V},$$

therefore

$$\boldsymbol{Q} = \boldsymbol{V}^{-1}.$$

Thus

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) \longrightarrow_d N_p(\boldsymbol{0}, \boldsymbol{Q}\boldsymbol{V}\boldsymbol{Q}) = N_p(\boldsymbol{0}, \boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1})$$
$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) \longrightarrow_d N_p(\boldsymbol{0}, \boldsymbol{V}^{-1}).$$
$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) \longrightarrow_d Z \sim N_p(\boldsymbol{0}, \boldsymbol{V})$$

which is the result (4.27). Also see Theorem 3.3 of Newey and McFadden [46]. □

The above lemma establishes the asymptotic normality of the maximum likelihood estimator (`MLE`) which satisfies the regularity conditions stated in Assumption 4.3.1. This result is very important because it gives us an asymptotic property of the estimator we are shrinking. Therefore this helps us to analyse the asymptotic properties of the shrinkage estimator (`JSSE`) we obtain.

We now present the asymptotic distributions of the plug-in maximum likelihood estimator $\hat{\boldsymbol{\beta}}_n$, restricted maximum likelihood estimator $\tilde{\boldsymbol{\beta}}_n^o$ and James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$. Before we state the theorem we begin by defining the matrix

$$\mathbf{B} = \mathbf{E}(\mathbf{E}^\top\mathbf{V}\mathbf{E})^{-1}\mathbf{E}^\top\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}^\top\mathbf{V}\mathbf{E}(\mathbf{E}^\top\mathbf{V}\mathbf{E})^{-1}\mathbf{E}^\top \tag{4.32}$$

which has the same dimension as that of matrix $\mathbf{V}$, were $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}_n)$, $\mathbf{E} = \mathbf{E}(\boldsymbol{\theta}_o)$, and

$$\mathbf{V} = \left(\mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\log f_{\boldsymbol{\theta}_o}(X_i)\right]\right)^{-1},$$

are as defined in the parametric structure Section 4.1 and estimation Section 4.2 respectively. We state the following theorem which shows the asymptotic distribution of the estimators $\hat{\boldsymbol{\beta}}_n$, $\tilde{\boldsymbol{\beta}}_n^o$ and $\hat{\boldsymbol{\beta}}_n^*$.

Theorem 4.3.2 *([28], Theorem 1)*

Suppose that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ satisfies the conditions under Assumptions 4.3.1 and 4.3.2, along the sequences (4.26) for $\boldsymbol{\beta}_n = \boldsymbol{\theta}_n$, $\tilde{\boldsymbol{\beta}}_n^o = \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_n^o)$ and $\hat{\boldsymbol{\beta}}_n = \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right) \longrightarrow_d \mathbf{G}^\top Z \sim \mathbf{G}^\top N_p(\boldsymbol{0}, \boldsymbol{V}) \tag{4.33}$$

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_n^o - \boldsymbol{\beta}_n\right) \longrightarrow_d \mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z+h)\right) \tag{4.34}$$

$$D_n = n\ell\left(\hat{\boldsymbol{\beta}}_n, \tilde{\boldsymbol{\beta}}_n^o\right) \longrightarrow_d \xi = (Z+h)^\top \boldsymbol{B}(Z+h) \tag{4.35}$$

$$\widehat{w} \longrightarrow_d w(Z) = \left(1 - \frac{\tau}{\xi}\right)_+ \tag{4.36}$$

where $\widehat{\tau}_n \longrightarrow \tau > 0$. Then the asymptotic distribution of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n\right) \longrightarrow_d w(Z)\mathbf{G}^\top Z + (1 - w(Z))\,\mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z+h)\right) \tag{4.37}$$

where the form the shrinkage estimator is the same as in (4.20). Furthermore, equations (4.33) to (4.37) hold jointly.

Proof

This proof is adopted from Theorem 1 of Hansen [28]. From Lemma 4.3.1 we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \longrightarrow_d Z \sim N_p\left(\boldsymbol{0}, \boldsymbol{V}\right) \tag{4.38}$$

and from standard arguments in the derivation in Section 9.1 of Newey and McFadden [46], we have also

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n^o - \boldsymbol{\theta}_n\right) \longrightarrow_d Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z+h). \tag{4.39}$$

From (4.38) and (4.39) and under part 8 of Assumption 4.3.1, we can apply the delta method to find that for some $\boldsymbol{\theta}_n^* \longrightarrow_p \boldsymbol{\theta}_o$,

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right) &= \sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_n)\right) \\ &= \boldsymbol{G}(\boldsymbol{\theta}_n^*)^\top \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \longrightarrow_d \mathbf{G}^\top Z\end{aligned}$$

and similarly

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_n^o - \boldsymbol{\beta}_n\right) = \boldsymbol{G}(\boldsymbol{\theta}_n^*)^\top \sqrt{n}(\tilde{\boldsymbol{\theta}}_n^o - \boldsymbol{\theta}_n) \longrightarrow_d \mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z+h)\right),$$

implying that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o\right) \longrightarrow_d \mathbf{G}^\top \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z+h). \tag{4.40}$$

Thus we have established (4.33) and (4.34). Now taking a second order Taylor expansion around $\hat{\boldsymbol{\beta}}_n$ *we have*

$$n\ell\left(\hat{\boldsymbol{\beta}}_n, \tilde{\boldsymbol{\beta}}_n^o\right) = n\ell\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\beta}}_n\right) + n\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}) \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n}^{\top} \left(\tilde{\boldsymbol{\beta}}_n^o - \hat{\boldsymbol{\beta}}_n\right) + n\left(\tilde{\boldsymbol{\beta}}_n^o - \hat{\boldsymbol{\beta}}_n\right)^{\top} \boldsymbol{W}(\boldsymbol{\beta}_n^*)\left(\tilde{\boldsymbol{\beta}}_n^o - \hat{\boldsymbol{\beta}}_n\right)$$

where $\boldsymbol{\beta}_n^*$ *lies on a line segment joining* $\tilde{\boldsymbol{\beta}}_n^o$ *and* $\hat{\boldsymbol{\beta}}_n$*. By part 1 of Assumption 4.3.2,* $\ell\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\beta}}_n\right) = 0$*. By part 3 of Assumption 4.3.2, the fact that* $\ell\left(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}\right)$ *is minimised at* $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_n$*, and the differentiability implied by part 3 of Assumption 4.3.2, then* $\frac{\partial}{\partial \boldsymbol{\beta}}\ell\left(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}\right) \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n} = 0$*. Under part 3 of Assumption 4.3.2 and the consistency of* $\tilde{\boldsymbol{\beta}}_n^o$ *and* $\hat{\boldsymbol{\beta}}_n$*, it follows that* $\boldsymbol{W}(\boldsymbol{\beta}_n^*) \longrightarrow_p \boldsymbol{W}$*. Combined with (4.28) we have*

$$\begin{aligned} n\ell\left(\hat{\boldsymbol{\beta}}_n, \tilde{\boldsymbol{\beta}}_n^o\right) &= n\ell\left(\tilde{\boldsymbol{\beta}}_n^o - \hat{\boldsymbol{\beta}}_n\right)^{\top} \boldsymbol{W}(\boldsymbol{\beta}_n^*)\left(\tilde{\boldsymbol{\beta}}_n^o - \hat{\boldsymbol{\beta}}_n\right) \\ &\longrightarrow_d (Z+h)^{\top}\boldsymbol{E}(\mathbf{E}^{\top}\boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^{\top}\boldsymbol{V}\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}^{\top}\boldsymbol{V}\boldsymbol{E}(\mathbf{E}^{\top}\boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^{\top}(Z+h) \\ &= (Z+h)^{\top}\boldsymbol{B}(Z+h) \\ &= \xi \end{aligned}$$

which is (4.35). We show (4.36) and (4.37) using the continuous mapping theorem (combination of Proposition 3.2.52 and Theorem 3.2.53). Thus from equations (4.23) and (4.22) we have

$$\widehat{w} \longrightarrow_d w(Z) = \left(1 - \frac{\tau}{\xi}\right)$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n\right) \longrightarrow_d w(Z)\mathbf{G}^{\top}Z + (1-w(Z))\mathbf{G}^{\top}\left(Z - \boldsymbol{V}\boldsymbol{E}(\mathbf{E}^{\top}\boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^{\top}(Z+h)\right)$$

respectively. Hence the proof. □

We state the following result we obtain direct from Theorem 4.3.2.

Corollary 4.3.3

With all the estimators and parameters set as in Theorem 4.3.2, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o\right) \longrightarrow_d \mathbf{G}^{\top}\boldsymbol{V}\boldsymbol{E}(\mathbf{E}^{\top}\boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^{\top}(Z+h)$$

where the matrices $\boldsymbol{V}$*,* $\boldsymbol{E}$ *and* $\boldsymbol{G}$ *are as defined before and* $h \in \mathbb{R}^p$*. The estimators* $\hat{\boldsymbol{\beta}}_n$ *and* $\tilde{\boldsymbol{\beta}}_n^o$ *are the plug-in unrestricted and restricted maximum likelihood estimators respectively, and* $Z \sim N_p(\boldsymbol{\theta}, \boldsymbol{V})$*.*

Theorem 4.3.2 gives expressions for the joint asymptotic distribution of the maximum likelihood estimator (`MLE`), restricted maximum likelihood estimator (`RMLE`), and James-Stein shrinkage estimator (`JSSE`) as transformations of the normal standard distribution Z and the non-centrality parameter h. The asymptotic distribution of $\hat{\boldsymbol{\beta}}_n^*$ is written as a random weighted average of the asymptotic distributions of $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}_n^o$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}_n^*$ is obtained for parameter sequences $\boldsymbol{\theta}_n$ tending towards a point $\boldsymbol{\theta}_o$ in the restricted parameter space Ω_o. The case of fixed $\boldsymbol{\theta} \notin \Omega_o$ can be obtained by letting h diverge towards infinity, in

which case $\xi \longrightarrow_p \infty$, $w(Z) \longrightarrow_p 1$ a result which is shown in Lemma 5.1.1, and the distribution on the right-hand-side of (4.33) tends towards $\mathbf{G}^\top Z \sim N_p(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$ where $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{G}^\top \mathbf{V}\mathbf{G}$. Equation (4.35) provides the asymptotic distribution of the distance-type statistic D_n. The limit random variable ξ controls the weight $w(Z)$ and thus the degree of shrinkage, so it is worth investigating further. Hansen in [28] showed that its expected value is

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}(\xi) &= \mathbb{E}\left[(Z+h)^\top \mathbf{B}(Z+h)\right] \\ &= h^\top \mathbf{B}h + \mathbb{E}_{\boldsymbol{\theta}}\left[tr(\mathbf{B}ZZ^\top)\right] \\ &= h^\top \mathbf{B}h + tr(\mathbf{B}\mathbf{V}) \end{aligned} \tag{4.41}$$

and in the canonical case when $\mathbf{G} = \mathrm{I}_p$ and $\mathbf{W} = \mathbf{V}^{-1}$ he showed that (4.41) simplifies to

$$\mathbb{E}\left(\xi\right) = h'\mathbf{B}h + p \tag{4.42}$$

since $tr(\mathbf{B}\mathbf{V}) = tr(\mathbf{V}^{-1}\mathbf{V}) = tr(\mathbf{I}_p) = p$. Furthermore, Hansen [28] explains that in this case, $\xi \sim \chi^2_p(h^\top \mathbf{B}h)$, a non-central chi-square random variable with non-centrality parameter $h^\top \mathbf{B}h$ and degrees of freedom p, and that the scalar $h^\top \mathbf{B}h$ captures how the divergence of $\boldsymbol{\theta}$ from the restricted region Ω_o affects the distribution of ξ. This will be useful to consider when analysing the asymptotic distributional bias value for the shrinkage estimator $\hat{\boldsymbol{\beta}}^*_n$.

The result in Corollary 4.3.3 is important to consider because it shows the link in the asymptotic distribution of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_n$ and restricted maximum likelihood estimator $\tilde{\boldsymbol{\beta}}^o_n$. The distribution represents the asymptotic distribution the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_n$ converges to when it is estimating the restricted maximum likelihood estimator $\tilde{\boldsymbol{\beta}}^o_n$. Hence the disparity between the two maximum likelihood estimators in distribution helps in determining how good the shrinking target is. Also since the asymptotic distribution is normally distributed, it is helpful in the analysis of the consistency of the restricted maximum likelihood estimator $\tilde{\boldsymbol{\beta}}^o_n$. Therefore everything depend on the sub-parameter space Ω_o used containing the shrinkage target. It is therefore good to have an idea of the region the true parameter value may lie according to the distribution one is looking at for efficient setting of the shrinkage target.

Chapter 5

ASYMPTOTIC BEVIOUR OF THE JAMES-STEIN SHRINKAGE ESTIMATOR

This chapter focuses on the asymptotic behaviour of the James-Stein shrinkage estimator (JSSE). It presents the key findings of our study. We investigate the consistency of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ by first considering the unrestricted MLE $\hat{\boldsymbol{\theta}}_n$. It is therefore important to first show that is consistent for $\boldsymbol{\theta}$ as $n \longrightarrow \infty$ which totally depends on its asymptotic distribution. Parts 3 and 4 of Assumption 4.3.1 are used to show the consistency of the MLE $\hat{\boldsymbol{\theta}}_n$. Then the consistency of the shrinkage estimator $\hat{\boldsymbol{\beta}}_n$ is established from the consistency of $\hat{\boldsymbol{\theta}}_n$. Thus in the first section of this chapter we show that the MLE $\hat{\boldsymbol{\theta}}_n$ and JSSE are asymptotically consistent. In the second section we investigate the asymptotic distributional bias. In the third section we show that the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is asymptotically efficient and then analyse the rate of convergence in the section which follows. In the last section we present simulation plots produced in R. We begin by considering asymptotic consistency and we use general concepts from the studies [31], [38] and [57].

5.1. Asymptotic Consistency

Asymptotic consistency is a fundamental property in estimators. This property requires that an estimator converges to the "true value" as the sample size value $n \longrightarrow \infty$. Consistency, just like other asymptotic properties concerns a sequence of estimators although we normally refer them to as "one" because we always generate the sequence from one type of the estimator. So it is a sequence of an estimator we refer to when we say the "consistency of an estimator". It is therefore worthy to investigate whether the James-Stein shrinkage estimator obtained from a maximum likelihood estimator satisfies this property. It should be mentioned that this is the main focus of this study. Therefore , we first show that the MLE $\hat{\boldsymbol{\theta}}_n$ is consistent as $n \longrightarrow \infty$.

5.1.1 Consistency of the Maximum Likelihood Estimator

To show that $\hat{\boldsymbol{\theta}}_n$ is asymptotically consistent for $\boldsymbol{\theta}$, we show that

$$\hat{\boldsymbol{\theta}}_n \longrightarrow_p \boldsymbol{\theta}.$$

Thus

$$\lim_{n\to\infty} \Pr\left(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| < \varepsilon\right) = 1 \tag{5.1}$$

for an $\varepsilon > 0$ using Definition 3.2.48. We take it that the Assumption 4.3.1 hold for the regularity condition of the MLE $\hat{\boldsymbol{\theta}}_n$. From (4.27) of Lemma 4.3.1 in Chapter

4, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \longrightarrow_d Z \sim N_p(\mathbf{0}, \mathbf{V}).$$

We consider the case where the parameter dimension is one ($p = 1$) for easy mathematical manipulation and then use Theorem 3.2.53 and Proposition 3.2.52 to generalise it to the case when we have p-dimensional parameter vector. Therefore, we proceed by using the methods in [11]. When we consider a univariate case from the distribution of the array $\tilde{\boldsymbol{X}}$ in Section 4.1 of Chapter 4, we have $X_1, X_2, ..., X_n$ being independent and identically distributed (`iid`) such that $X_i \sim N(0, 1)$. The maximum likelihood estimator for θ is

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We recall that $\bar{X}_n \sim N\left(0, \frac{1}{n}\right)$ since $\sigma^2 = 1$. By using (5.1) we have

$$\begin{aligned}
\Pr\left(\mid \hat{\theta}_n - \theta \mid < \varepsilon\right) &= \Pr\left(\mid \bar{X}_n - \theta \mid < \varepsilon\right) \\
&= \Pr\left(-\varepsilon < \bar{X}_n - \theta < \varepsilon\right) \\
&= \Pr\left(\frac{-\varepsilon}{\sqrt{\frac{1}{n}}} < \frac{\bar{X}_n - \theta}{\sqrt{\frac{1}{n}}} < \frac{\varepsilon}{\sqrt{\frac{1}{n}}}\right) \\
&= \Pr\left(-\varepsilon\sqrt{n} < Z < \varepsilon\sqrt{n}\right) \qquad \textit{since} \quad \frac{\bar{X}_n - \theta}{\sqrt{\frac{1}{n}}} \sim Z = N(0, 1).
\end{aligned}$$

Therefore

$$\lim_{n \longrightarrow \infty} \Pr\left(-\varepsilon\sqrt{n} < Z < \varepsilon\sqrt{n}\right) = \Pr\left(-\infty < Z < \infty\right) = 1.$$

Thus $\hat{\theta}_n$ is asymptotically consistent for θ. The implication of having p-dimensional vector is that $\hat{\boldsymbol{\theta}}_n$ will be a p-dimensional vector of estimators of these parameters. Since each estimator in this vector of estimators is consistent then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ for p-dimensional parameter vector as $n \longrightarrow \infty$.

Since we have shown that the initial estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically consistent, we proceed to check whether the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ we obtain when we shrink $\hat{\boldsymbol{\theta}}_n$ is consistent.

5.1.2 Consistency of the James-Stein Shrinkage Estimator

We now examine whether the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is consistent for $\boldsymbol{\theta}_n$. We analyse this consistency in two ways depending on how large the value of h is for the sequence $\boldsymbol{\theta}_n$. To do this, we begin by stating two lemmas before stating the main theorem which provide part of the results for the study. The following lemma is important for the results in Theorem 5.1.3. It provides concepts on how the distance type statistic D_n and the estimate $\widehat{w}$ converge in probability.

Lemma 5.1.1

Suppose we have the shrinkage value estimate $\widehat{w}$ as in Theorem 4.3.2 such that

$$\widehat{w} \longrightarrow_d w(Z) = \left(1 - \frac{\tau}{\xi}\right)_+ \tag{5.2}$$

where $\xi = (Z+h)^\top \boldsymbol{B}(Z+h) \sim \chi^2_p\left(h^\top \boldsymbol{B}h\right)$ a non central Chi-square distribution with non centrality parameter $h^\top \boldsymbol{B}h$, $\tau = p-2$, $p \geq 3$ and matrix $\boldsymbol{B}$ is as defined in (4.32). With the sequence $\boldsymbol{\theta}_n$ of estimates, if $h \longrightarrow \infty$ then

$$w(Z) \longrightarrow_p 1 \tag{5.3}$$

and if h is fixed then

$$w(Z) \longrightarrow_p 0 \quad otherwise \quad w(Z) \longrightarrow_p \boldsymbol{r} \tag{5.4}$$

where $\boldsymbol{r}$ is a constant such that $0 < \boldsymbol{r} < 1$ and $(a)_+$ in (5.2) is a positive trimming function which keeps what is in the brackets greater than or equal to zero.

Proof

We begin by considering the first case when h diverges. Suppose that $h \longrightarrow \infty$ then

$$(Z+h) \longrightarrow_p \infty \quad as \quad n \longrightarrow \infty. \tag{5.5}$$

Therefore from (5.5) we have

$$(Z+h)^\top \boldsymbol{B}(Z+h) = \xi \longrightarrow_p \infty \tag{5.6}$$

as $h \longrightarrow \infty$ and $n \longrightarrow \infty$. Now considering that

$$\widehat{w} \longrightarrow_d w(Z) = \left(1 - \frac{\tau}{\xi}\right)_+$$

as $n \longrightarrow \infty$. Thus using (5.6) we have

$$w(Z) \longrightarrow_p \left(1 - \frac{\tau}{\infty}\right)_+ \quad as \quad n \longrightarrow \infty$$

and gives

$$w(Z) \longrightarrow_p (1-0)_+ \quad as \quad n \longrightarrow \infty$$

which then

$$w(Z) \longrightarrow_p 1 \tag{5.7}$$

as $n \longrightarrow \infty$. Hence we have established (5.3).

Secondly suppose that h is fixed. Then we have the sequence

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + n^{-\frac{1}{2}} h$$

which becomes $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$ as $n \longrightarrow \infty$ implying that $\hat{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}_o$ as $n \longrightarrow \infty$. Suppose

$$\xi = \chi^2_p\left(h^\top \boldsymbol{B}h\right) \longrightarrow_p \boldsymbol{D} \quad as \quad n \longrightarrow \infty \tag{5.8}$$

where $\boldsymbol{D}$ *is a constant,* h *is fixed and* $\boldsymbol{B}$ *is not affected by an increase in the sample size value* n*, then*

$$w(Z) \longrightarrow_d \left(1 - \frac{\tau}{\boldsymbol{D}}\right)_+$$

where $p \geq 3$*. If* $\frac{\tau}{\boldsymbol{D}} = 1$ *as* $n \longrightarrow \infty$*, then*

$$w(Z) \longrightarrow_p (1-1)_+$$
$$w(Z) \longrightarrow_p 0,$$

if $\frac{1}{\boldsymbol{D}} > 1$ *then the value inside the brackets will be negative and by definition of "+" we end with zero. This will vary as* p *changes but still considering* $\xi \sim \chi^2_p(h^\top \boldsymbol{B}h)$ *the probability of* ξ *depends on the degrees of freedom* p *and will vary according to the chi-square distribution, implying that the ratio* $\frac{\tau}{\boldsymbol{D}} = \boldsymbol{M} \geq 1$ *as* $n \longrightarrow \infty$*. Therefore we have*

$$w(Z) \longrightarrow_p \left(1 - \frac{\tau}{\boldsymbol{D}}\right)_+ \quad as \quad n \longrightarrow \infty$$

which gives

$$w(Z) \longrightarrow_p (1 - \boldsymbol{M})_+ \quad as \quad n \longrightarrow \infty$$

for a constant $\boldsymbol{M} > 1$*. Proceeding in the same way we have*

$$w(Z) \longrightarrow_p (1 - \boldsymbol{M})_+ \quad as \quad n \longrightarrow \infty, \quad for \quad 1 - \boldsymbol{M} < 0,$$
$$w(Z) \longrightarrow_p 0 \quad as \quad n \longrightarrow \infty$$

by definition of $(x)_+$*. Thus*

$$w(Z) \longrightarrow_p 0 \quad as \quad n \longrightarrow \infty. \tag{5.9}$$

Otherwise, if the ratio $\frac{\tau}{\boldsymbol{D}}$ *is such that* $0 < \frac{\tau}{\boldsymbol{D}} < 1$ *as* $n \longrightarrow \infty$*, we have*

$$w(Z) \longrightarrow_p \boldsymbol{r}$$

where $\boldsymbol{r} \in (0, 1)$*.* □

Lemma 5.1.1 establishes that $w(Z)$ will converge to a constant in $[0, 1]$ depending on the value of p which will determine the ratio $\frac{p-2}{\xi}$ as $n \longrightarrow \infty$. This means that $w(Z)$ is asymptotically uniformly distributed over $[0, 1]$.

We state and prove the following lemma needed to prove Theorem 5.1.3.

Lemma 5.1.2

Let $\hat{\boldsymbol{\beta}}_n$ *and* $\tilde{\boldsymbol{\beta}}^o_n$ *be the plug-in maximum likelihood estimator and restricted maximum likelihood estimator respectively. From (4.40) of Theorem 4.3.2 we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}^o_n\right) \longrightarrow_d \mathbf{G}^\top \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h) = \eta(Z + h)$$

where $\eta = \mathbf{G}^\top \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top$*. Let* $\boldsymbol{\mu} = \eta h$ *be the mean of* $\eta(Z + h)$ *where* $\eta(Z + h) \sim N_p\left(\eta h, \eta^\top \boldsymbol{V}\eta\right)$*. Then by the Law of Large Numbers (LLN) and the Chebyshev's inequality, if* $\hat{\boldsymbol{\theta}}_n$ *converges in probability to the "true value"* $\boldsymbol{\theta}_o$ *then also* $\tilde{\boldsymbol{\theta}}^o_n \longrightarrow_p \boldsymbol{\theta}_o$ *as the sample size value* $n \longrightarrow \infty$*. Thus for any* $\varepsilon > 0$*,*

$$\lim_{n \longrightarrow \infty} \Pr\left(\mid \tilde{\boldsymbol{\theta}}^o_n - \boldsymbol{\theta}_o \mid > \varepsilon\right) = 0$$

where h *is fixed as in (4.26). Implying* $\tilde{\boldsymbol{\theta}}^o_n$ *is also asymptotically consistent for* $\boldsymbol{\theta}$*.*

Proof

We consider a case for $p = 1$ *(univariate) and generalise to p-multivariate normal. Suppose that* h *is fixed, then the sequence* $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + n^{-\frac{1}{2}}h$ *becomes* $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$ *as* $n \longrightarrow \infty$*. From Section 4.1 we have the array* $\tilde{\boldsymbol{X}}$ *which contains vectors* $X_1, X_2, ..., X_n$ *which are independent and identically distributed (*`iid`*). Assume that* $Var(X_i) = \sigma^2$ *for all* $i < \infty$*. Then we have*

$$Var(\hat{\theta}_n) = \frac{\eta_1^2\sigma^2}{n} \quad and \quad \mathbb{E}_\theta(\hat{\theta}_n) = \mu = \eta_1 h_1 \tag{5.10}$$

as $n \longrightarrow \infty$*. By the Weak Law of Large Numbers (*`WLLN`*) and Sluksky's theorem proceeding from (5.10) we have*

$$\hat{\theta}_n = \bar{X}_n \longrightarrow_p \mu = \eta_1 h_1 = \mathbb{E}_\theta(\hat{\theta}_n) \tag{5.11}$$

as $n \longrightarrow \infty$ *where* $\eta_1 h_1$ *is an element of* ηh *and*

$$\frac{S^2}{n} \longrightarrow_p Var(\bar{X}_n) = Var(\hat{\theta}_n) = \frac{\eta_1^2\sigma^2}{n} = \frac{\eta_1^2}{n} \tag{5.12}$$

as $n \longrightarrow \infty$ *for* σ^2 *where* $\frac{\eta_1^2}{n}$ *is one of the diagonal elements of the matrix* $(\eta^\top \boldsymbol{V}\eta)$*. But from the regularity conditions of maximum likelihood estimators, the* `RMLE` $\tilde{\theta}_n^o$ *is consistent for the mean* μ*, thus we have*

$$\tilde{\theta}_n^o \longrightarrow_p \mu = \eta_1 h_1 = \theta_o \tag{5.13}$$

as $n \longrightarrow \infty$*. Now using Chebyshev's inequality on* $\hat{\theta}_n$ *for all* $\varepsilon > 0$ *we have*

$$\begin{aligned}
\Pr\left(\mid \hat{\theta}_n - \mu \mid > \varepsilon\right) &\leq \frac{\sigma^2}{n\varepsilon^2} \\
\Pr\left(\mid \tilde{\theta}_n^o - \theta \mid > \varepsilon\right) &\leq \frac{\eta_1^2}{n\varepsilon^2} \qquad\qquad as \;\; n \longrightarrow \infty \;\; by \;\; using \; (5.13) \\
&= \frac{\eta_1^2}{n\varepsilon^2} \longrightarrow 0
\end{aligned}$$

as $n \longrightarrow \infty$*. Thus* $\tilde{\theta}_n^o$ *is consistent for* θ*. Consequently,* $\tilde{\boldsymbol{\theta}}_n^o$ *is asymptotically consistent for* $\boldsymbol{\theta}$*.* □

Lemma 5.1.2 establishes that the maximum likelihood estimator (`MLE`) $\hat{\boldsymbol{\theta}}_n$ is asymptotically consistent for the restricted maximum likelihood estimator (`RMLE`) $\tilde{\boldsymbol{\theta}}_n^o$. As $n \longrightarrow \infty$ within some neighbourhood $\tilde{\boldsymbol{\theta}}_n^o \longrightarrow_p \boldsymbol{\theta}_o$, the true value. Similarly we expect that $\hat{\boldsymbol{\theta}}_n \longrightarrow_p \boldsymbol{\theta}_o$ within some neighbourhood. Thus making the two maximum likelihood estimators to converge in probability to some "true value" $\boldsymbol{\theta}_o$ as $n \longrightarrow \infty$. From the continuous mapping theorem and using similar arguments above, Lemma 5.1.2 also implies that the restricted maximum likelihood estimator $\tilde{\boldsymbol{\theta}}_n^o$ is asymptotically consistent for $\boldsymbol{\theta}$. This is important in the analysis of the asymptotic consistency of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ when we have a case where the the maximum likelihood estimator we are shrinking is equal to the shrinkage target. We now present the main result of this study in the following theorem.

Theorem 5.1.3

*Let $\boldsymbol{\theta} \in \Omega$, where Ω is a parameter space with elements in $\mathbb{R}^p$. Suppose we have a James-Stein shrinkage estimator (**JSSE**) $\hat{\beta}_n^*$ which is obtained by shrinking the maximum likelihood estimator (**MLE**) $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta} \in \Omega$ where our shrinkage target $\tilde{\boldsymbol{\theta}}_n^o$ is the restricted maximum likelihood estimator (**RMLE**) of $\boldsymbol{\theta} \in \Omega_o$ a partitioned sub-parameter space from Ω by the restriction (4.1). Then the **JSSE** is given by*

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n - \left(\frac{p-2}{n(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)^\top \boldsymbol{V}^{-1} (\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)} \right)_+ (\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)$$

where $\hat{\boldsymbol{\beta}}_n = \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$, $\tilde{\boldsymbol{\beta}}_n^o = \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_n^o)$, $p \geq 3$ and $(x)_+$ is a positive trimming function. If $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ as $n \longrightarrow \infty$ then the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is also consistent for $\boldsymbol{\theta}$, where we have the sequence $\boldsymbol{\theta}_n$ defined as in (4.26).

Proof

Let $\boldsymbol{\beta}_n = \boldsymbol{\theta}_n$ where $\boldsymbol{\theta}_n$ is as defined in (4.26). To show that $\hat{\boldsymbol{\beta}}_n^*$ is consistent for $\boldsymbol{\theta}$ as $n \longrightarrow \infty$ we consider the value of h which determines the neighbourhood of the sequence $\boldsymbol{\theta}_n$, when it diverges to infinity and when it is just fixed. Suppose that h diverges, to evaluate

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n\right) \longrightarrow_d w(Z)\mathbf{G}^\top Z + (1 - w(Z))\,\mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top (Z + h)\right) \tag{5.14}$$

as $n \longrightarrow \infty$, we first consider $w(Z)$ from (4.36). By Lemma 5.1.1 we have

$$w(Z) \longrightarrow_p 1 \quad as \quad n \longrightarrow \infty. \tag{5.15}$$

Hence from (5.15) and substituting $w(Z)$ by 1 in (5.14) we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d 1\mathbf{G}^\top Z + (1-1)\,\mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top (Z + h)\right)$$

and yields

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d \mathbf{G}^\top Z \sim \mathbf{G}^\top N_p(\boldsymbol{0}, \boldsymbol{V})$$

which gives

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d N_p(\boldsymbol{0}, \boldsymbol{V}_\beta) \tag{5.16}$$

as $n \longrightarrow \infty$ where $\boldsymbol{V}_\beta = \mathbf{G}^\top \boldsymbol{VG}$. Thus we have

$$\hat{\boldsymbol{\beta}}_n^* \longrightarrow_p \boldsymbol{\beta}_n$$

if

$$\lim_{n\to\infty} \Pr\left(|\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n| > \varepsilon\right) = 0$$

for any $\varepsilon > 0$. Hence $\hat{\boldsymbol{\beta}}_n^*$ is consistent for $\boldsymbol{\beta}_n = \boldsymbol{\theta}_n$ and $\lim_{n\longrightarrow\infty} \boldsymbol{\theta}_n = \boldsymbol{\theta}$.

Secondly, suppose h is fixed as a constant value, then the sequence

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + n^{-\frac{1}{2}}h$$

becomes $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$ *as* $n \longrightarrow \infty$. *From this equality we have* $\hat{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}_o$ *as* $n \longrightarrow \infty$ *and two conditions arise. The first one is that the sequence* $\boldsymbol{\theta_n}$ *will be within the restricted parameter space* Ω_o *with* $\boldsymbol{\theta}_o \in \Omega_o$. *From the restriction (4.1) of* Ω_o *this will mean that the shrinkage target is exactly at the true value and our consideration will be just on one parameter space. Therefore, we have* $\hat{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\theta}}_n^o$, *but from (4.39)*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n^o - \boldsymbol{\theta}_n) \longrightarrow_d \zeta = Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)$$

which will be the same as the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ *since we only consider the sub-parameter space* Ω_o, *and the shrinkage value* $\boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top$ *affects it. Thus*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \longrightarrow_d \zeta = Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h) \tag{5.17}$$

as $n \longrightarrow \infty$ *because we are estimating* $\boldsymbol{\theta} \in \Omega_o$ *and from Proposition 3.2.27 there will be no difference between the* `MLE` *and* `RMLE`. *Due to this equality of the two maximum likelihood estimators, from (4.23) and substituting (5.17) in (5.14) we have*

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d w(Z)\mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)\right) + (1 - w(Z)) \\ \mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)\right)\end{aligned}$$

as $n \longrightarrow \infty$, *which becomes*

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d \left[w(Z)\mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)\right)\right] - \\ \left[w(Z)\mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)\right)\right] + \\ \mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)\right)\end{aligned}$$

as $n \longrightarrow \infty$, *and then simplifies to*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d \mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z + h)\right) \tag{5.18}$$

as $n \longrightarrow \infty$, *which is the same as the asymptotic distribution of* $\tilde{\boldsymbol{\beta}}_n^o = \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_n^o)$. *Therefore using Lemma 5.1.2 and (5.18), the consistency of the James-Stein shrinkage estimator* $\hat{\boldsymbol{\beta}}_n^*$ *follows from the consistency of* $\tilde{\boldsymbol{\theta}}_n^o$.

Lastly, we consider the case when we have two well defined parameter spaces, Ω *and* Ω_o. *Then we have that* $\hat{\boldsymbol{\theta}}_n \neq \tilde{\boldsymbol{\theta}}_n^o$. *Analysing (5.14) further, we consider the value* $\boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top$ *which is not affected by the sample size value n but it is a value affected by an increase or decrease in the number of parameters p. Since*

$$Z \sim N_p(\boldsymbol{0}, \boldsymbol{V})$$

then

$$Z + h \sim N_p(h, \boldsymbol{V}) \tag{5.19}$$

by linearity property of the normal distribution. Also implying

$$\eta(Z + h) \sim N_p(\eta h, \eta^\top \boldsymbol{V}\eta) \tag{5.20}$$

for some matrix η *of dimension* $p \times p$. *From (5.4) of Lemma 5.1.1 we have*

$$w(Z) \longrightarrow_p 0 \quad if \quad \frac{\tau}{D_n} \geq 1 \tag{5.21}$$

as $n \longrightarrow \infty$. *Therefore from (5.18) we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d w(Z)\mathbf{G}^\top Z + (1 - w(Z))\mathbf{G}^\top (Z - \eta(Z + h))$$

for some shrinkage value effect matrix $\eta = \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top$. *Evaluating this asymptotic distribution as* $n \longrightarrow \infty$ *we have*

$$\begin{aligned} w(Z)\mathbf{G}^\top Z + (1 - w(Z))\mathbf{G}^\top (Z - \eta(Z + h)) &\longrightarrow_d (0)\mathbf{G}^\top Z + (1 - 0)\mathbf{G}^\top (Z - \eta(Z + h)) \\ &\longrightarrow_d \mathbf{G}^\top (Z - \eta(Z + h)) \end{aligned}$$

as $n \longrightarrow \infty$. *Thus*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n) \longrightarrow_d \mathbf{G}^\top \left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top (Z + h)\right) \quad as \quad n \longrightarrow \infty.$$

Hence using Lemma 5.1.2, the consistency of $\hat{\boldsymbol{\beta}}_n^*$ *follows from the consistency of* $\tilde{\boldsymbol{\beta}}_n^o = \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_n^o)$ *which is consistent since* $\tilde{\boldsymbol{\theta}}_n^o$ *is consistent. Similarly if* $w(Z) \longrightarrow \boldsymbol{r} \in (0, 1)$, *the consistency of the James-Stein shrinkage estimator* $\hat{\boldsymbol{\beta}}_n^*$ *follow from the consistency of the restricted maximum likelihood estimator and also the fact that* $\hat{\boldsymbol{\theta}}_n$ *is consistent for* $\boldsymbol{\theta}$. *Thus the shrinkage estimator* $\hat{\boldsymbol{\beta}}_n^*$ *is asymptotically consistent for* $\boldsymbol{\theta}$. □

We consider the following corollary obtained direct from the results discussed in the previous sections concerning the consistency of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ which is obtained by shrinking a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ towards a shrinkage target $\tilde{\boldsymbol{\theta}}_n^o$.

Corollary 5.1.4

Let $\boldsymbol{\theta} \in \Omega$, *where* Ω *is a parameter space with elements in* $\mathbb{R}^p$. *Suppose we have a James-Stein shrinkage estimator* $\hat{\boldsymbol{\beta}}_n^*$ *which is obtained by shrinking the maximum likelihood estimator* $\hat{\boldsymbol{\theta}}_n$ *of* $\boldsymbol{\theta} \in \Omega$ *where the shrinking target* $\tilde{\boldsymbol{\theta}}_n^o$ *is the restricted maximum likelihood estimator of* $\boldsymbol{\theta} \in \Omega_o$ *a sub-parameter space of* Ω. *If* $\hat{\boldsymbol{\beta}}_n^*$ *and* $\hat{\boldsymbol{\theta}}_n$ *are asymptotically consistent for* $\boldsymbol{\theta}$ *then the restricted maximum likelihood estimator* $\tilde{\boldsymbol{\theta}}_n^o$ *is also asymptotically consistent for* $\boldsymbol{\theta}$.

We now examine the asymptotic bias behaviour of the three estimators, the maximum likelihood estimator, the restricted maximum likelihood estimator and the James-Stein shrinkage estimator, by evaluating their asymptotic distributional bias.

5.2. Asymptotic Distributional Bias (ADB)

In this section we derive expressions for the asymptotic distributional bias (ADB) for the estimators $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}_n^o$ and $\hat{\boldsymbol{\beta}}_n^*$. The objective is to estimate the unknown parameter vector $\boldsymbol{\theta}$ by some estimator T_n when performance is evaluated by squared error loss and to establish the asymptotic distributional bias of the James-Stein shrinkage estimator for easy analysis of the asymptotic efficiency.

We study the ADB for the three estimators by analysing the asymptotic bias values. We therefore present the expression for the asymptotic distributional bias of the estimators. The ADB of an estimator T_n from Definition 3.2.55 is given by

$$\mathbf{ADB}(T_n) = \lim_{n \longrightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}} \left[\sqrt{n}(T_n - \boldsymbol{\theta})\right]. \tag{5.22}$$

We present the asymptotic distributional bias of the three estimators $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}_n^o$ and $\hat{\boldsymbol{\beta}}_n^*$ in the theorem below.

Theorem 5.2.1

Suppose that Assumptions 4.3.1 and 4.3.2 stated earlier in Chapter 4 hold. Then under $\{P_n\}$ a sequence of parameter dimension with the sample size value n and $p \geq 3$, the ADBs of the estimators $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}^o_n$ and $\hat{\boldsymbol{\beta}^*}_n$ are respectively

1. $\boldsymbol{ADB}(\hat{\boldsymbol{\theta}}_n) = 0$
2. $\boldsymbol{ADB}(\tilde{\boldsymbol{\theta}}^o_n) = -\boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h$
3. $\boldsymbol{ADB}(\hat{\boldsymbol{\beta}}^*_n) = -\vartheta \mathbf{G}^\top \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h$

where $\vartheta = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{p-2}{\xi}\right]$ for $p \geq 3$.

Proof

1.

$$\begin{aligned}
\boldsymbol{ADB}(\hat{\boldsymbol{\theta}}_n) &= \lim_{n\longrightarrow\infty} \mathbb{E}_{\boldsymbol{\theta}}\left(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\right) \\
&= \lim_{n\longrightarrow\infty} 0 \quad since \quad \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \longrightarrow_d Z \sim N_p(\boldsymbol{0}, \boldsymbol{V}) \quad as \quad n \longrightarrow \infty \\
&= 0 \\
\therefore \boldsymbol{ADB}(\hat{\boldsymbol{\theta}}_n) &= 0.
\end{aligned} \tag{5.23}$$

2.

$$\begin{aligned}
\boldsymbol{ADB}(\tilde{\boldsymbol{\theta}}^o_n) &= \lim_{n\longrightarrow\infty} \mathbb{E}_{\boldsymbol{\theta}}\left(\sqrt{n}(\tilde{\boldsymbol{\theta}}^o_n - \boldsymbol{\theta})\right) \\
&= \lim_{n\longrightarrow\infty} \left(-\boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h\right) \\
&= -\boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h \\
\therefore \boldsymbol{ADB}(\tilde{\boldsymbol{\theta}}^o_n) &= -\boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h
\end{aligned} \tag{5.24}$$

from (4.39) of Theorem 4.3.2.

3. $\boldsymbol{ADB}(\hat{\boldsymbol{\beta}}^*_n) = \lim_{n\longrightarrow\infty} \mathbb{E}_{\boldsymbol{\theta}}\left[\sqrt{n}(\hat{\boldsymbol{\beta}}^*_n - \boldsymbol{\beta}_n)\right]$. *From (4.37) of Theorem 4.3.2 we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*_n - \boldsymbol{\beta}_n) \longrightarrow_d w(Z)\mathbf{G}^\top Z + (1-w(Z))\mathbf{G}^\top\left(Z - \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top(Z+h)\right)$$

where

$$w(Z) = \left(1 - \frac{p-2}{\xi}\right)_+.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}\left[w(Z)\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[1 - \left(\frac{p-2}{\xi}\right)\right] \\
&= 1 - \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{p-2}{\xi}\right] \\
&= 1 - \vartheta
\end{aligned}$$

where $\vartheta = \mathbb{E}\left[\frac{p-2}{\xi}\right]$, $p \geq 3$ *and* $\mathbb{E}_{\boldsymbol{\theta}}(Z) = 0$ *as* $n \longrightarrow \infty$ *since* $Z \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. *Then*

$$\begin{aligned}
\boldsymbol{ADB}(\hat{\boldsymbol{\beta}}_n^*) &= \lim_{n\longrightarrow\infty}\left[(1-\vartheta)\,0 + (1-1+\vartheta)\left(-\mathbf{G}^\top\left(\boldsymbol{VE}(\mathbf{E}^\top\boldsymbol{VE})^{-1}\mathbf{E}^\top h\right)\right)\right] \\
&= \lim_{n\longrightarrow\infty}\left[-\vartheta\mathbf{G}^\top\left(\boldsymbol{VE}(\mathbf{E}^\top\boldsymbol{VE})^{-1}\mathbf{E}^\top h\right)\right] \\
&= -\vartheta\mathbf{G}^\top\left(\boldsymbol{VE}(\mathbf{E}^\top\boldsymbol{VE})^{-1}\mathbf{E}^\top h\right) \\
\therefore \boldsymbol{ADB}(\hat{\boldsymbol{\beta}}_n^*) &= -\vartheta\mathbf{G}^\top\left(\boldsymbol{VE}(\mathbf{E}^\top\boldsymbol{VE})^{-1}\mathbf{E}^\top h\right)
\end{aligned} \tag{5.25}$$

where $\vartheta = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{p-2}{\xi}\right]$ *for* $p \geq 3$. □

Corollary 5.2.2

When the fixed constant $h = \boldsymbol{0}$, *the asymptotic distributional bias values of the three estimators are zero. Therefore* $h \neq \boldsymbol{0}$.

Remark 5.2.3 *This condition actually arises from the shrinking neighbourhood. So if* $h = \boldsymbol{0}$, *then the implication is that there is no shrinkage. Therefore, all the three estimators will be the same. Thus always h is not equal to zero. Furthermore, it is clear that the shrinking process brings some bias to an estimator as much as it improves the* `MSE`.

Expressions (5.23), (5.24) and (5.25) above show the asymptotic distributional bias values for the `MLE`, `RMLE` and `JSSE` respectively.

In the next section we investigate the asymptotic efficiency of the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$. We utilise the result obtained on the asymptotic bias of the James-Stein shrinkage estimator and analyse its efficiency as the sample size value n approaches infinity. This analysis will provide a clear picture on the estimation measure of the shrinkage estimator to the assumed true value as the sample size value n grows without bound.

5.3. Asymptotic Efficiency

Analysing further the asymptotic properties of the shrinkage estimator, we investigate the asymptotic efficiency of the estimator. In the previous section we have shown that the shrinkage estimator is asymptotically biased and from the literature review we have found that it has a smaller asymptotic distributional risk. Therefore, it is prompting to check if it is also asymptotically efficient. We proceed by using the Cramér-Rao Bound (`CRB`) although this is commonly used for unbiased estimators. We begin by discussing briefly the bound concerning biased estimators before using it.

5.3.1 Bound on the Variance of Biased Estimators

Consider an estimator $T_n = T_n(\boldsymbol{\theta})$ of dimension p with bias $\mathbf{b}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[T_n] - \boldsymbol{\theta}$, and let $\boldsymbol{\Phi}(\boldsymbol{\theta}) = \mathbf{b}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \boldsymbol{\theta}$. By the Cramér-Rao bound, any j^{th} component of an unbiased estimator of dimension p whose expectation is $\boldsymbol{\Phi}_j(\boldsymbol{\theta})$ has variance greater than or equal to $\frac{(\boldsymbol{\Phi}_j'(\boldsymbol{\theta}))^2}{\mathbf{J}_{jj}(\boldsymbol{\theta})}$ for all $j = 1, 2, ..., p$. Thus, any estimator T_n whose bias is given by a function $\mathbf{b}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ satisfies

$$\mathbf{V}_{jj}(T_n) \geq \frac{\left[1 + \mathbf{b}'_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})\right]^2}{\mathbf{J}_{jj}(\boldsymbol{\theta})} \tag{5.26}$$

for all $j = 1, 2, ..., p$ where $\mathbf{J}_{jj}(\boldsymbol{\theta})$ is the j^{th} component of the fisher information matrix $\mathbf{J}(\boldsymbol{\theta})$. The unbiased version of the bound is a special case of this result, with $\mathbf{b}(\boldsymbol{\theta}) = \mathbf{0}$. But from equation (5.26) we find that the **mean squared error** of a biased estimator is bounded by

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(T_{n_j} - \boldsymbol{\theta}_j)^2\right] \geq \frac{\left[1 + \mathbf{b}'_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})\right]^2}{\mathbf{J}_{jj}(\boldsymbol{\theta})} + \mathbf{b}_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})^2 \tag{5.27}$$

using the standard decomposition of the `MSE` for all $j = 1, 2, ..., p$. Using these concepts we proceed to check if the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}^*_n$ is asymptotically efficient. We use the concepts in [34] and [6] to apply the Cramér-Rao bound (`CRB`).

5.3.2 Asymptotic Efficiency of the James-Stein Estimator

From (5.27) and (5.25) we are sure that we have a bound for the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}^*_n$. We state the following theorem as part of our results.

Theorem 5.3.1

Let $\hat{\boldsymbol{\beta}}^*_n$ be a James-Stein shrinkage estimator obtained by shrinking a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ where the two estimators are as defined in Section 4.2. Given the asymptotic bias $\boldsymbol{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n)$ of the `JSSE` $\hat{\boldsymbol{\beta}}^*_n$, the Cramér-Rao bound for $\hat{\boldsymbol{\beta}}^*_{n_j}$ is given by

$$\boldsymbol{CRB} = \frac{\left[1 + \boldsymbol{b}'_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\beta}}^*_n)\right]^2}{\boldsymbol{J}_{jj}(\boldsymbol{\theta})} \quad for \;\; all \;\; j = 1, 2, ..., p \tag{5.28}$$

where $\boldsymbol{J}(\boldsymbol{\theta})$ is the fisher information and $\boldsymbol{b}'_{\boldsymbol{\theta}_j}$ is the derivative of the j^{th} element of the bias vector. Then

$$\frac{\boldsymbol{CRB}}{\boldsymbol{V}_{jj}(\hat{\boldsymbol{\beta}}^*_n)} = 1 \quad as \;\; n \longrightarrow \infty,$$

*and thus the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}^*_n$ is **asymptotically efficient** for all $j = 1, 2, ..., p$.*

Proof

We analyse the asymptotic efficient by evaluating the Cramér Rao bound as $n \longrightarrow \infty$. Consider the bias of the estimator $\hat{\boldsymbol{\beta}}^*_n$ from (5.25) and let

$$\boldsymbol{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n) = -\mathbf{G}^\top \vartheta \, \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h \tag{5.29}$$

where $\vartheta = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{p-2}{\xi}\right]$ for $p \geq 3$. The expectation $\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{p-2}{\xi}\right]$ of the fraction $\frac{p-2}{\xi}$ which follows a distribution determined by the distribution $\xi \sim \chi^2_p\left(h^\top \boldsymbol{B}h\right)$ has a value (constant) free of the parameter $\boldsymbol{\theta}$. Therefore we regard it as a scalar. Let $\alpha = -\vartheta$ then (5.29) becomes

$$\boldsymbol{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n) = \alpha \mathbf{G}^\top \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h \tag{5.30}$$

and $\boldsymbol{b}'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n)$ will be

$$\begin{aligned} \boldsymbol{b}'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n) &= \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n) \\ &= \alpha \frac{\partial}{\partial \boldsymbol{\theta}} \left(\mathbf{G}^\top \boldsymbol{VE}(\mathbf{E}^\top \boldsymbol{VE})^{-1}\mathbf{E}^\top h\right) \end{aligned} . \tag{5.31}$$

a matrix of dimension $p \times 1$. *Using (5.26) and then combining (5.29) and (5.31) we obtain*

$$\frac{\left[1 + \boldsymbol{b}'_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\beta}}^*_n)\right]^2}{\boldsymbol{J}_{jj}(\boldsymbol{\theta})} = \frac{\left[1 + \alpha \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{G}^\top \boldsymbol{V}\boldsymbol{E}(\mathbf{E}^\top \boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^\top h\right)\right]^2}{\boldsymbol{J}_{jj}(\boldsymbol{\theta})} \quad for\ all \quad j = 1, 2, ..., p \tag{5.32}$$

where $\frac{\partial}{\partial \boldsymbol{\theta}_j}$ *is the partial derivative of the* j^{th} *element,* $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta}) = \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i)\right)^{-1}$, $\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{\theta}_o) = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{e}(\boldsymbol{\theta})^\top$ *and*

$$\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{n} -\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i)\right].$$

Proceeding analysing the bound, we begin by considering the terms involved. Thus

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{E} = \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \boldsymbol{e}(\boldsymbol{\theta})^\top$$

remains the same as $n \longrightarrow \infty$. *We have*

$$\boldsymbol{V} = \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \log f_{\boldsymbol{\theta}}(X_i)\right)^{-1} \longrightarrow \boldsymbol{\Sigma}_p \tag{5.33}$$

as $n \longrightarrow \infty$ *where the elements of* $\boldsymbol{\Sigma}_p$ *are zeros apart from the diagonal elements* $\boldsymbol{V}_{jj}(\boldsymbol{\theta})$ *which are ones for* $j = 1, 2, ..., p$ *since the observations are* ***iid*** *and follow a p-multivariate standard normal distribution. Thus from (5.33) we have*

$$\frac{\partial}{\partial \boldsymbol{\theta}_j} \boldsymbol{V}_{jj}(\boldsymbol{\theta}) \longrightarrow 0 \quad and \quad \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{V} \longrightarrow \boldsymbol{0} \tag{5.34}$$

as $n \longrightarrow \infty$ *for* $j = 1, 2, ..., p$. *This implies that*

$$\frac{\partial}{\partial \boldsymbol{\theta}_j} \mathbf{G}^\top \boldsymbol{V}\boldsymbol{E}(\mathbf{E}^\top \boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^\top h \longrightarrow 0 \quad and \quad \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}^\top \boldsymbol{V}\boldsymbol{E}(\mathbf{E}^\top \boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^\top h \longrightarrow \boldsymbol{0} \tag{5.35}$$

for $j = 1, 2, ..., p$ *as* $n \longrightarrow \infty$. *Then from (5.35) we have*

$$\boldsymbol{b}'_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\beta}}^*_n) \longrightarrow 0 \quad and \quad \boldsymbol{b}'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*_n) \longrightarrow \boldsymbol{0} \tag{5.36}$$

for $j = 1, 2, ..., p$ *as* $n \longrightarrow \infty$. *Therefore from (5.35) and (5.36), then using (5.32) we have*

$$\begin{aligned} \frac{\left[1 + \boldsymbol{b}'_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\beta}}^*_n)\right]^2}{\boldsymbol{J}_{jj}(\boldsymbol{\theta})} &= \frac{\left[1 + \alpha \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{G}^\top \boldsymbol{V}\boldsymbol{E}(\mathbf{E}^\top \boldsymbol{V}\boldsymbol{E})^{-1}\mathbf{E}^\top h\right)\right]^2}{\boldsymbol{J}_{jj}(\boldsymbol{\theta})} \\ &= \frac{[1+0]^2}{\boldsymbol{J}_{jj}(\boldsymbol{\theta})} \\ &= [\boldsymbol{J}_{jj}(\boldsymbol{\theta})]^{-1} \end{aligned} \tag{5.37}$$

as $n \longrightarrow \infty$ *for* $j = 1, 2, ..., p$. *Since for* $j = 1, 2, ..., p$ *we have* $\boldsymbol{V}_{jj}(\boldsymbol{\theta}) = \boldsymbol{J}_{jj}(\boldsymbol{\theta})^{-1}$ *then* $\boldsymbol{V} = \boldsymbol{J}^{-1}$ *as* $n \longrightarrow \infty$. *Hence from (5.37) we have the variance for the James-Stein shrinkage estimator* $\hat{\boldsymbol{\beta}}^*_{n_j}$ $\boldsymbol{V}_{jj}(\hat{\boldsymbol{\beta}}^*_{n_j})$ *converges to the* ***CRB*** *as* $n \longrightarrow \infty$ *for all* $j = 1, 2, ..., p$. *This means that*

$$\frac{\boldsymbol{CRB}}{\boldsymbol{V}_{jj}(\hat{\boldsymbol{\beta}}^*_n)} \longrightarrow \frac{\boldsymbol{V}_{jj}(\hat{\boldsymbol{\beta}}^*_n)}{\boldsymbol{V}_{jj}(\hat{\boldsymbol{\beta}}^*_n)} = 1 \quad as \quad n \longrightarrow \infty$$

for all $j = 1, 2, ..., p$. *Thus the James-Stein shrinkage estimator* ***JSSE*** $\hat{\boldsymbol{\beta}}^*_n$ *is asymptotically efficient.* □

Theorem 5.3.1 above establishes the asymptotic efficiency of the JSSE $\hat{\boldsymbol{\beta}}_n^*$ by showing that using the Cramér-Rao bound (CRB) we have a bound for the MSE of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$.

5.4. Rate of Convergence of the James-Stein Estimator

We now investigate the rate of convergence of the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ (JSSE) by using concepts applied on the MLE discussed in Section 2.3 of the literature review and in Hoeffding [32]. To proceed we consider the shrinkage estimator of the form in (3.15) using plug-in maximum likelihood estimators $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}_n^o$. Let $\hat{\boldsymbol{\beta}}_n^*$ be a James-Stein shrinkage estimator obtained when we shrink the MLE $\hat{\boldsymbol{\beta}}_n = \mathbf{g}(\hat{\boldsymbol{\theta}}_n)$ defined earlier before as

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n - \left(\frac{p-2}{n(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)^\top \mathbf{V}^{-1} (\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)} \right)_+ (\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o)$$

for $p \geq 3$ and $\tilde{\boldsymbol{\beta}}_n^o = \mathbf{g}(\tilde{\boldsymbol{\theta}}_n^o)$. We proceed to find the rate of convergence of this estimator by using its relationship with the MLE. Since the shrinkage target value may have no effect on the convergence rate, for easier transformation of our sequence $\boldsymbol{\theta}_n$ we set $\tilde{\boldsymbol{\theta}}_n^o = \mathbf{0}$ implying $\tilde{\boldsymbol{\beta}}_n^o = \mathbf{0}$. Thus we have

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n - \left(\frac{p-2}{n\left(\hat{\boldsymbol{\beta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\beta}}_n\right)} \right)_+ \hat{\boldsymbol{\beta}}_n$$

which becomes

$$\hat{\boldsymbol{\beta}}_n^* = \left(1 - \frac{p-2}{\hat{\boldsymbol{\beta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\beta}}_n} \right)_+ \hat{\boldsymbol{\beta}}_n$$

when we factor out $\hat{\boldsymbol{\beta}}_n$ and drop out the n in the denominator to have a form with a lower MSE according to the James-Stein shrinkage strategy in [35]. Let

$$k = \left(1 - \frac{p-2}{\hat{\boldsymbol{\beta}}_n^\top \mathbf{V}^{-1} \hat{\boldsymbol{\beta}}_n} \right)_+, \tag{5.38}$$

then

$$\hat{\boldsymbol{\beta}}_n^* = k \hat{\boldsymbol{\beta}}_n. \tag{5.39}$$

Now consider the sequence

$$\hat{\boldsymbol{\beta}}_{n_j} = \boldsymbol{\beta}_{o_j} + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{5.40}$$

for $j = 1, 2, ..., p$ where $\boldsymbol{\beta}_{o_j}$ is the "true" j^{th} parameter value. From the equality in(5.39) we have

$$\hat{\boldsymbol{\beta}}_n = \frac{1}{k} \hat{\boldsymbol{\beta}}_n^* \tag{5.41}$$

for the shrinkage value k. Therefore substituting the right hand side of (5.41) in (5.40) the sequence $\hat{\boldsymbol{\beta}}_{n_j}$ becomes

$$\frac{1}{k} \hat{\boldsymbol{\beta}}_{n_j}^* = \boldsymbol{\beta}_{o_j} + O_p\left(\frac{1}{\sqrt{n}}\right),$$

hence we have the sequence

$$\hat{\boldsymbol{\beta}}^*_{n_j} = k\boldsymbol{\beta}_{o_j} + O_p\left(\frac{1}{\sqrt{n}}\right)k \tag{5.42}$$

which is in terms of the shrinkage estimator with the shrinking effect value k such that $0 < k \leq 1$. Analysing this sequence further shows that it satisfies the smoothness regularity conditions for the `MLE`, therefore we can use Proposition 3.2.52.

Let $\boldsymbol{\beta}^*_{o_j} = k\boldsymbol{\beta}_{o_j}$ be the true value in the shrinkage sense which is obtained when we scale the true value $\boldsymbol{\beta}_{o_j}$ with the shrinkage factor k. Then the sequence (5.42) becomes

$$\hat{\boldsymbol{\beta}}^*_{n_j} = \boldsymbol{\beta}^*_{o_j} + O_p\left(\frac{1}{\sqrt{n}}\right)k \tag{5.43}$$

implying that

$$\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right) = O_p\left(\frac{1}{\sqrt{n}}\right)k \tag{5.44}$$

for all $j = 1, 2, ..., p$. This means that $\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right)$ is still within the neighbourhood of $\frac{1}{\sqrt{n}}$ since $0 < k \leq 1$. Therefore, using the second order Taylor's theorem we have

$$\ln \frac{\prod_{i=1}^{n} f_{\hat{\boldsymbol{\beta}}^*_{n_j}(x_i)}}{\prod_{i=1}^{n} f_{\boldsymbol{\beta}^*_{o_j}(x_i)}} = \left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right)\sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)}Z_j - \frac{n}{2}\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right)^2 \mathbf{I}_{jj}(\boldsymbol{\beta}^*_o) + O_p(1) \tag{5.45}$$

for $j = 1, 2, ..., p$. Since

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mathbf{L}(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}$$

for the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_n$, then also

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mathbf{L}(\hat{\boldsymbol{\beta}}^*_n) = \mathbf{0}$$

implying that

$$\frac{\partial}{\partial \boldsymbol{\beta}_j} \ln \mathbf{L}(\hat{\boldsymbol{\beta}}^*_{n_j}) = 0 \tag{5.46}$$

for all $j = 1, 2, ..., p$. Assuming that the log-likelihood function is differentiable, from (5.45) and (5.46) we have

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}^*_{n_j}} \ln \left(\frac{\prod_{i=1}^{n} f_{\hat{\boldsymbol{\beta}}^*_{n_j}(x_i)}}{\prod_{i=1}^{n} f_{\boldsymbol{\beta}^*_{o_j}(x_i)}}\right) = \left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right)^{1-1}\sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)}Z_j - \frac{2n}{2}(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j})^{2-1}\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o) + O_p(1) \tag{5.47}$$

and simplifying (5.47) becomes

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}^*_{n_j}} \ln \left(\frac{\prod_{i=1}^{n} f_{\hat{\boldsymbol{\beta}}^*_{n_j}(x_i)}}{\prod_{i=1}^{n} f_{\boldsymbol{\beta}^*_{o_j}(x_i)}} \right) = \sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)} Z_j - n\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right)\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o) + O_p(1) = 0$$

implying

$$\sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)} Z_j - n\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right)\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o) + O_p(1) = 0. \tag{5.48}$$

Rearranging (5.48) we have

$$n\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right) + O_p(1) = \sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)} Z_j \tag{5.49}$$

for $j = 1, 2, ..., p$ where $Z_j \sim (0, \mathbf{V}_{\boldsymbol{\beta}_j})$ where $\mathbf{V}_{\boldsymbol{\beta}_j}$ is the variance of the j^{th} element of the covariance matrix $\mathbf{V}_{\boldsymbol{\beta}}$ of the distribution $= \mathbf{G}^\top N_p(\mathbf{0}, \mathbf{V})$ presented in Theorem 4.3.2 in Chapter 4 and thus Z_j is the standard normal distribution for the j^{th} element of $\hat{\boldsymbol{\beta}}_n$. Now dividing the left and right hand side of (5.49) by $\sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)}$ we obtain

$$\sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)}\left(\hat{\boldsymbol{\beta}}^*_{n_j} - \boldsymbol{\beta}^*_{o_j}\right) + O_p(1) = Z_j \tag{5.50}$$

where Z_j is the distribution of the j^{th} element of $\hat{\boldsymbol{\beta}}_n$ and $Z \sim N_p(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$. Using sequence (5.42), equation (5.50) becomes

$$k\sqrt{n}\sqrt{\mathbf{I}_{jj}(\boldsymbol{\beta}^*_o)}\left(\hat{\boldsymbol{\beta}}_{n_j} - \boldsymbol{\beta}_{o_j}\right) + O_p(1) = Z_j \tag{5.51}$$

for some $\boldsymbol{\beta}^*_{o_j} \longrightarrow \boldsymbol{\beta}_{o_j}$ and $j = 1, 2, ..., p$ where $Z \sim N_p(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$ and $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{G}^\top \mathbf{V}\mathbf{G}$. The distribution Z_j is a normal distribution for each j^{th} element of the plug-in estimator $\hat{\boldsymbol{\beta}}_n$ as described before in the analysis above.

Thus equation (5.51) establishes the condition which implies local asymptotic normality (`LAN`) and differentiability in quadratic mean (`DQM`) for the estimator $\hat{\boldsymbol{\beta}}^*_n$ which implies that the rate of convergence is of order $\frac{1}{k\sqrt{n}}$ and rate $k\sqrt{n}$. This can also be achieved if we use the fact that the risk bound of the James-Stein shrinkage estimator is bounded by that of the `MLE` and the latter converges at the rate $\sqrt{n}$. Hence the James-Stein shrinkage estimator is $k\sqrt{n}$-consistent.

5.5. Simulation Plots

In this section we compare the behaviour of the mean squared error (`MSE`) of the maximum likelihood estimator (`MLE`) $\hat{\boldsymbol{\theta}}_n$ and the James-Stein shrinkage estimator (`JSSE`) $\hat{\boldsymbol{\beta}}^*_n$ as the sample size value n increases. We use R to simulate plots of the `MSE` for different sample size values of n using the R package library (**MASS**). The data is generated using a 3×3 correlation matrix ρ to get the covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.3 & 0.1 \\ 0.3 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{bmatrix}$$

which is symmetric and the variance in the major diagonal is 1 representing a standard normal variance. Thus we take the case when $p = 3$ and this meets the James-Stein classical condition of $p \geq 3$. Since $X \sim N_3(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, we have

$$\hat{\boldsymbol{\theta}}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_{in} \quad for \quad p = 3 \tag{5.52}$$

making the MLE $\hat{\boldsymbol{\theta}}_n$ a 3×1 matrix which implies that the dimension for the shrinkage estimator is also 3×1. Now from Lemma 3.2.10 and knowing that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ is unbiased and the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is biased, the following expressions where used to calculate the mean squared error MSE of the two estimators. Using (5.52), for the MLE we have

$$\mathbf{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n) = \mathbf{V}(\hat{\boldsymbol{\theta}}_n) = \mathbf{V}(\bar{X}_n) \tag{5.53}$$

for $p = 3$ and $\mathbf{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = 0$, and for the James-Stein shrinkage estimator

$$\mathbf{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_n^*) = \mathbf{V}(\hat{\boldsymbol{\beta}}_n^*) + \left[\mathbf{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_n^*)\right]^2 \tag{5.54}$$

which using (5.39) becomes

$$\mathbf{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_n^*) = \mathbf{V}(k\hat{\boldsymbol{\theta}}_n) + \left[\mathbf{b}_{\theta}(k\hat{\boldsymbol{\theta}}_n)\right]^2 \tag{5.55}$$

where k is a shrinkage value which shrinks the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ to a James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ for $p \leq 3$ and it is equivalent to the one given in equation (5.38). The bias $\mathbf{b}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_n^*)$ is as defined in Definition 3.2.9. Thus we obtain the mean squared error of the shrinkage estimator in (5.55) by using (5.53) which is obtained from Lemma 3.2.10. The shrinkage value k is evaluated using the expression

$$k = \left(1 - \frac{1}{\bar{X}_n^{\top}\mathbf{V}(\bar{X}_n)^{-1}\bar{X}_n}\right) \tag{5.56}$$

for $p = 3$. The commands for all expressions and plots produced in R are provided in the appendix.

We proceed by first considering a sample size value of $n = 30$ to analyse the nature of the plots and how the trends for the mean squared error change as n increases. The plots for this sample size value are plotted on the same graph for easy comparison of the behaviour of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ and the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$. The other sample size values considered are 2000, 8000, 50000 and 100000 so that we examine the trends of the graphs as $n \longrightarrow \infty$ since we are interested in the asymptotic behaviour. Plots for the sample size value of 50000 are presented separately for the two estimators, each estimator has its own MSE graph. This is to show the mean squared error value concentration for the two estimators, $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\beta}}_n^*$ respectively.

We now present the following plots which were simulated using the sample size value of 30 and 2000.

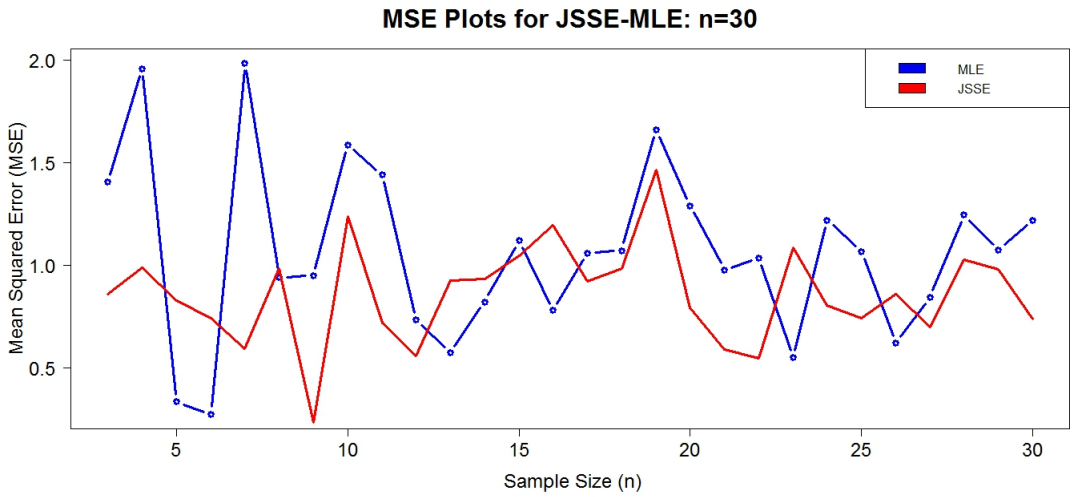

Figure 5.5.1: `MSE` plots for the `MLE` and `JSSE` for $n = 30$

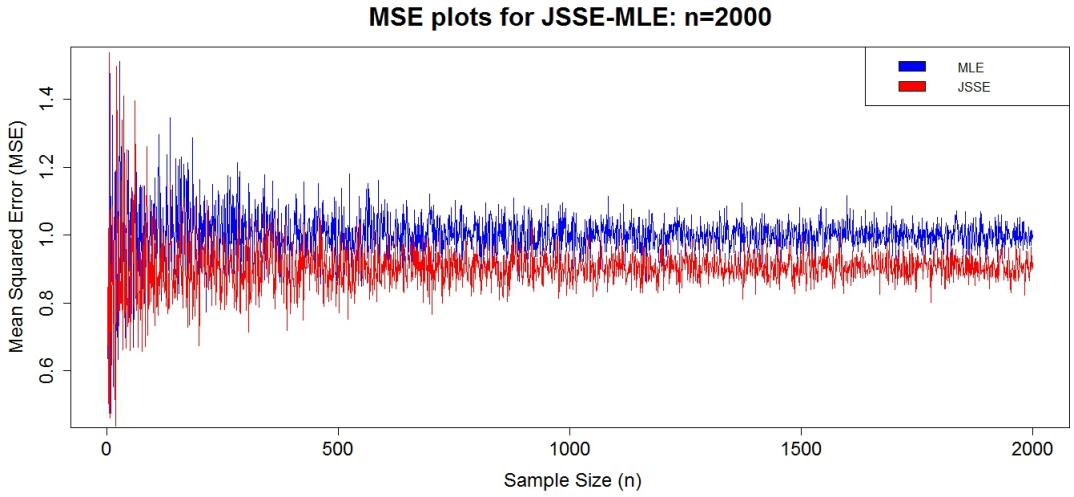

Figure 5.5.2: `MSE` plots for the `MLE` and `JSSE` for $n = 2000$

The graphs in Figure 5.5.1 show the `MSE` plots for the `MLE` and `JSSE` obtained by simulating the mean squared error (`MSE`) for the sample size value of 30 and those in Figure 5.5.2 are the `MSE` plots for both estimators simulated using n =2000. As the sample size approaches the value n =30, the line plot for the `JSSE` shows a lower mean squared error value. The trend is the same in Figure 5.5.2 as we increase the sample size value n from 30 to 2000. The graph for the `JSSE` shows a lower `MSE` compared to that of the `MLE` and the graphical picture trend for the two estimators is almost the same in the way they converge.

To have a clear picture of the trend, we consider large sample size values which produce well scaled plots as n increases. We therefore consider simulation plots obtained using the sample size values of 8000, 50000 and 100000. We present the following `MSE` plots for the `JSSE` and `MLE` for $n = 8000$ and 50000. For $n = 500000$ the `MSE` graphs for each estimator (`MLE` and `JSSE`) are plotted separately on single graphs.

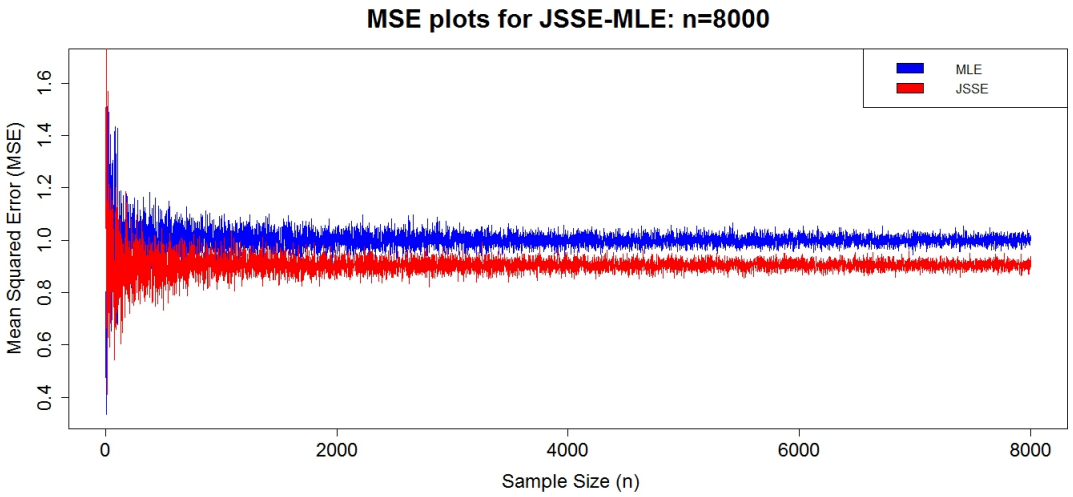

Figure 5.5.3 : `MSE` plots for the `MLE` and `JSSE` for $n = 8000$

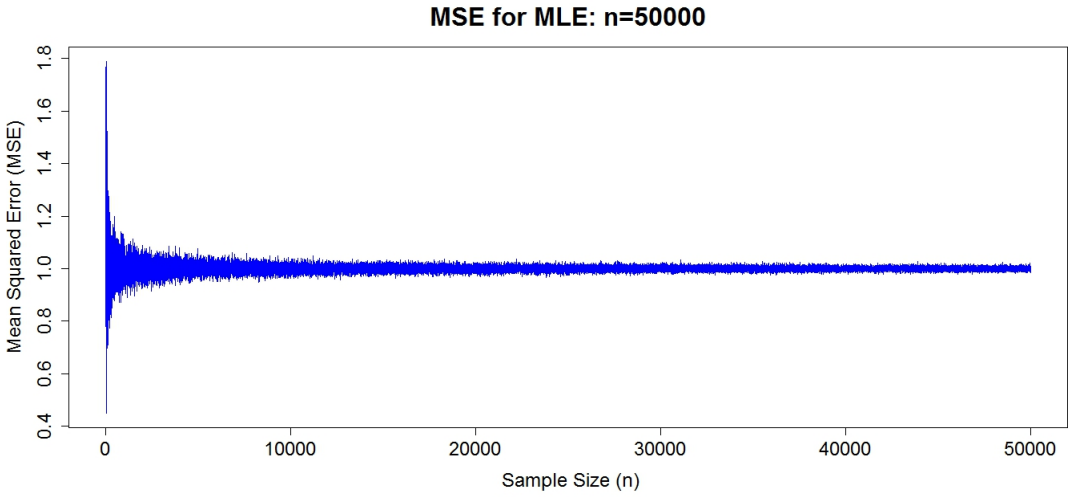

Figure 5.5.4: `MSE` plot for the `MLE` for $n = 50000$

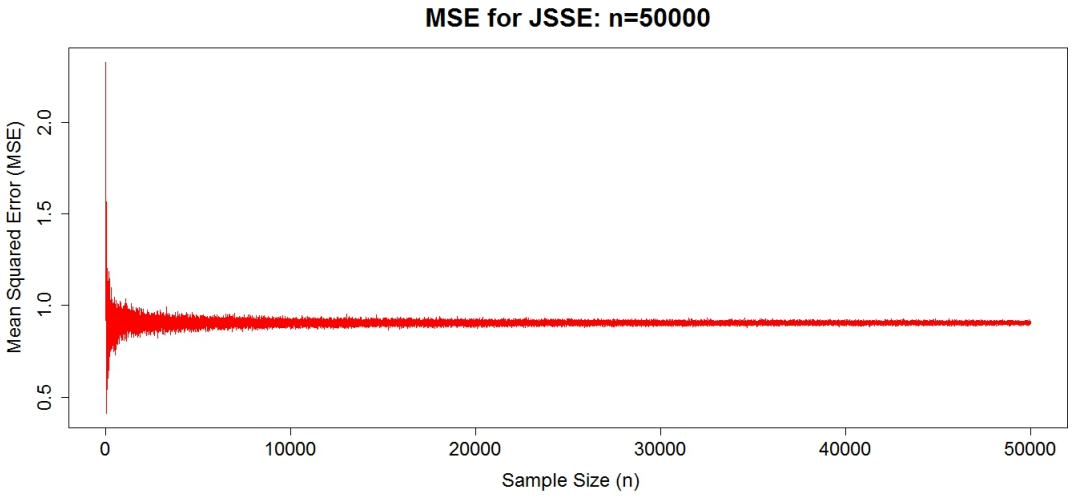

Figure 5.5.5: `MSE` plot for the `JSSE` for $n = 50000$

Figures 5.5.3, 5.5.4 and 5.5.5 above show the `MSE` graphs for the maximum likelihood estimator (`MLE`) and James-Stein shrinkage estimator (`JSSE`) obtained by simulating the mean squared error (`MSE`) using sample size values of 8000 and 50000. They show the graphical trends of the `MSE` for the maximum likelihood

estimator $\hat{\boldsymbol{\theta}}_n$ and James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ when the sample size value n is increased to 8000 and then to 50000.

Figure 5.5.4 shows that the MSE value for the James-Stein shrinkage estimator is concentrated on 0.9 while Figure 5.5.5 shows that the MSE for the maximum likelihood estimator is concentrated on 1.0. The graph for the James-Stein shrinkage estimator is a little bit thinner than the graph for the maximum likelihood estimator for the sample size value of 50000. For close comparison, we consider the following MSE plots for the James-Stein shrinkage estimator and the maximum likelihood estimator for a sample size value of 100000 plotted on the same graph.

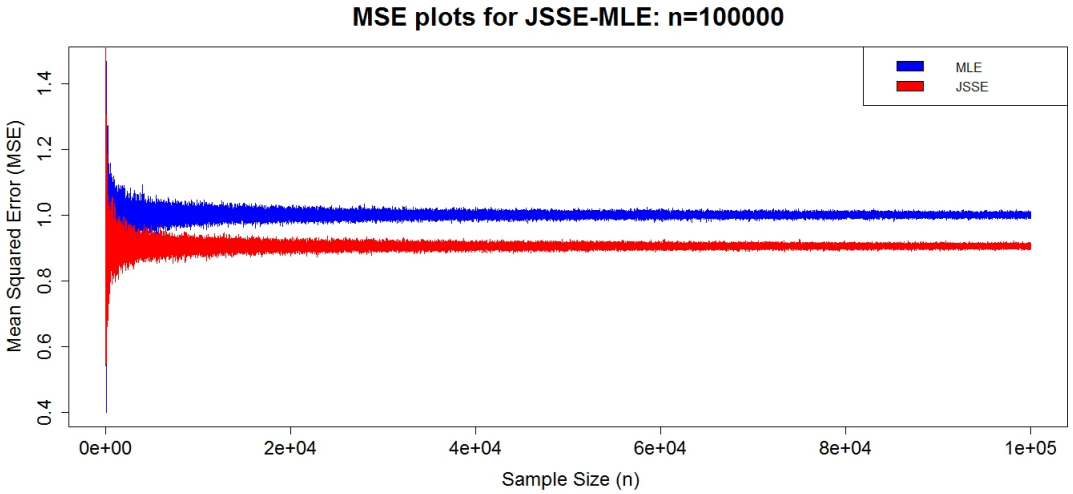

Figure 5.5.6: MSE plots for the MLE and JSSE for $n = 100000$

Figure 5.5.6 shows the MSE graphs for the James-Stein shrinkage estimator (JSSE) and maximum likelihood estimator (MLE) obtained by simulating the mean squared error (MSE) using a sample size value of 100000. The graphs show the error trends which are obtained when we increased the sample size value to 100000.

Collectively the scaled plots show that there is some reduction in the mean squared error for the James-Stein shrinkage estimator compared to that of the initial estimator (MLE). The trend in mean squared error for both the maximum likelihood estimator and the James-Stein shrinkage estimator shows that as the sample size value n increases, the MSE values converge to some value. The plots suggest that the James-Stein shrinkage estimator converges to a lower MSE value compared to the maximum likelihood estimator.

In Section 5.1 we have shown that the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ obtained by shrinking a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically consistent. This consistency follows from the consistency of the MLE $\hat{\boldsymbol{\theta}}_n$ and RMLE $\tilde{\boldsymbol{\theta}}_n^o$ (shrinkage target). Section 5.2 has established that the James-Stein shrinkage estimator is asymptotically biased, a property it posses even with small sample size values. We have also observed that the James-Stein shrinkage estimator has lower asymptotic risk compared to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$. The third section has established that the MSE of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ achieves the CRB and therefore it is asymptotically efficient. Section 5.4 has examined the rate of convergence of the James-Stein shrinkage estimator and in the last section we have presented MSE plots for the JSSE and MLE.

Chapter 6

DISCUSSION OF FINDINGS

In this chapter we discuss the key findings of our study. In the first section we examine the consistency of the James-Stein shrinkage estimator established in Theorem 5.1.3. In the second section we discuss the asymptotic values obtained in Theorem 5.2.1. The third section provides a discussion on the asymptotic efficiency of the JSSE established in Theorem 5.3.1 and the last section discusses the findings obtained from the simulation plots presented in the last section of Chapter 5.

6.1. Consistency of the James-Stein Shrinkage Estimator

Our interest to check whether the shrinkage estimator is asymptotically consistent for $\boldsymbol{\theta}$ arises from the fact that it converges to some normal distribution. In Section 4.2 of [30], Hansen showed that the James-Stein shrinkage estimator uniformly dominates the ordinary least squares (OLS) estimator in terms of MSE and that the risk bound of the JSSE is less than that of the latter. Since the properties of the OLS and MLE are the same, we expect that the bounds are the same when we have a MLE. Thus convergence of the MLE implies convergence of the shrinkage estimator. Therefore, this makes consideration of the consistency of the James-Stein shrinkage estimator which is obtained by shrinking the MLE an interesting area of attention. According to our shrinking strategy, the consistency of the shrinkage estimator is shown by considering different cases which arise from the definition of the sequence $\boldsymbol{\theta}_n$.

First we considered the case when $\boldsymbol{\theta}_n$ has a neighbourhood which is not restricted. This happens when we let h diverge ($h \longrightarrow \infty$). When $h \longrightarrow \infty$, our neighbourhood of consideration becomes the whole parameter space and $\xi \longrightarrow_p \infty$ and $\widehat{w} \longrightarrow_p 1$. Hence there is no distinction on how the parameters in Ω and Ω_o are asymptotically distributed. As a result the asymptotic distribution of the James-Stein shrinkage estimator will be the same as for the MLE as $n \longrightarrow \infty$. Thus the two estimators are asymptotically distributed the same under this condition. Hence the consistency of $\hat{\boldsymbol{\beta}}_n^*$ follows from the consistency of $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$.

The second case is when we have h as a fixed value. Here the two parameter spaces are well defined and distinctive in terms of where the parameters of interest are located. When $n \longrightarrow \infty$ then $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$ (the "true value") because $n^{-\frac{1}{2}}h \longrightarrow 0$, implying that it is possible to be just within the restricted parameter space when we take the asymptotic values. Therefore, if we are within the restricted parameter space Ω_0 then there is no difference between the maximum likelihood estimators, the MLE and RMLE. Thus the MLE is distributed the same as the RMLE as $n \longrightarrow \infty$. As a consequence of having the two maximum likelihood estimators distributed the same, the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ ends up to be asymptotically

distributed the same as $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n^o$, hence $\hat{\boldsymbol{\beta}}_n^*$ is consistent for $\boldsymbol{\theta}$ which follows from the consistency of $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n^o$. Furthermore, we have $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n^o\right) \longrightarrow_p 0$ as $n \longrightarrow \infty$. Although its under invariant estimators, Stone in Proposition 2.2 of [53] showed that the result holds as long as both estimators converge asymptotically to the same normal distribution. Therefore in this case the `MLE` and `RMLE` satisfy this condition without necessarily being invariant estimators.

Lastly we have the sequence $\boldsymbol{\theta}_n$ not only within the restricted parameter space, but in both Ω and Ω_0. We simplify the asymptotic distribution by using η to represent the scalar asymptotic values. Since we get the asymptotic distribution which is normally distributed. We apply the law of large numbers (`LLN`) as $n \longrightarrow \infty$, therefore $\hat{\boldsymbol{\beta}}_n^*$ converges to $\boldsymbol{\theta}$ for $\varepsilon > 0$ since $\Pr\left(|\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n| > \varepsilon\right) = 0$ and $\boldsymbol{\beta}_n = \boldsymbol{\theta}_n$ where $\lim_{n \longrightarrow \infty} \boldsymbol{\theta}_n = \boldsymbol{\theta}$. Hence generally $\hat{\boldsymbol{\beta}}_n^*$ is asymptotically consistent for $\boldsymbol{\theta}$.

6.2. Asymptotic Distributional Bias Values

Expressions (5.23), (5.24) and (5.25) in Section 5.2 of Chapter 5 show the asymptotic bias values for the `MLE`, `RMLE` and `JSSE` respectively. The `MLE` $\hat{\boldsymbol{\theta}}_n$ is asymptotically distributionally unbiased while the restricted `MLE` $\tilde{\boldsymbol{\theta}}_n^o$ and the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ are asymptotically distributionally biased. When the parameter dimension is $p = 2$, the James-Stein shrinkage estimator is asymptotically unbiased. Though when this is the case there is no reduction in the risk loss and `MSE`. This further implies that there is no shrinkage ($\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\theta}}_n$). Also, the asymptotic distributional bias of the James-Stein shrinkage estimator depends on the value of p compared to the asymptotic distributional bias of the restricted maximum likelihood estimator which is constant. Thus both shrinking and partitioning (creation of a sub-parameter space) bring some bias to an estimator even at an asymptotic value. This is seen from (5.24) and (5.25) that an estimator $\tilde{\boldsymbol{\theta}}_n^o$ obtained due to creation (partitioning) of a sub-parameter space is asymptotically biased and $\hat{\boldsymbol{\beta}}_n^*$ an estimator obtained due to shrinkage is also asymptotically biased.

6.3. Asymptotic Efficiency of the Shrinkage Estimator

Theorem 5.3.1 has established the asymptotic efficiency of the `JSSE` $\hat{\boldsymbol{\beta}}_n^*$ by showing that using the Cramér-Rao bound (`CRB`) we have a bound for the `MSE` of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$. The bound utilises the bias we obtain in (5.26) of Section 5.2. This bias converges to a constant when we evaluate it as $n \longrightarrow \infty$. We note that the only matrix which can have expressions involving the parameter $\boldsymbol{\theta}$ is $\mathbf{V}$. Therefore we show that $\mathbf{V}$ converges to a constant free of the parameter as $n \longrightarrow \infty$. Hence we have the whole shrinkage expression as a constant, and when we differentiate the constant with respect to the parameter it yields zero. Consequently the partial derivative of the bias with respect to the parameter $\boldsymbol{\theta}$ gives zero. Therefore evaluating the bound as $n \longrightarrow \infty$, we see that it converges to the variance, hence the result. The property of asymptotic efficiency is very important because it expresses the stability of the estimator when we have large sample size values.

We have seen that on the aspect of bias, the James-Stein shrinkage estimator maintains its bias in small and large sample size values. On the other hand it does

well in the risk loss even when we have very large sample size values. In terms of efficiency, the James-Stein shrinkage estimator is asymptotically efficient. This further shows that the shrinkage estimator is consistent as we consider different sample size values even when the sample size value is large. Therefore the James-Stein shrinkage estimator exhibits good asymptotic properties which are desirable even though it is asymptotically biased. This means that the properties of the estimator are not distorted by an increase in the sample size value n. Thus making it a preference when we want to come up with a good estimator through shrinking techniques because the new estimator we obtain when we shrink is stable even when the sample size value increases without bound.

6.4. MSE Comparisons from the Simulation Plots

Graphical presentations comparing the two estimators (`MLE` and `JSSE`) are presented as simulation plots in the last section of Chapter 5. The simulation plots are produced using the statistical package R. This section analysed the effect of the sample size value n on the mean squared error (`MSE`) for both estimators. The single plots for both the `MLE` and `JSSE` involving large sample size values show some important features of the asymptotic behaviour of these estimators. The linear trend of the `MSE` of the maximum likelihood estimator is concentrated on 1.0. The James-Stein shrinkage estimator also shows a linear trend just like the `MLE` although the concentration of the `MSE` is on 0.9 as the sample size value is increased to 50000. In each case the graphs for both estimators (`JSSE` and `MLE`) were plotted using the same sample size values for all the simulations, but the linear trends of the `MLE` and the `JSSE` differ in smoothness. The `MSE` graph for the James-Stein shrinkage estimator becomes thinner a little bit faster than the `MSE` graph for the maximum likelihood estimator. As the sample size value increases without bound the graphs will become very thin and converge to some value. Hence the `MSE` for the `JSSE` will converge first compared to the `MSE` for the `MLE`. The rate at which the graph for the shrinkage estimator drops to become smooth is a little bit faster compared to the graph for the `MLE` as the sample size value increases. Therefore, we conclude that the James-Stein shrinkage estimator converges at a faster rate compared to the maximum likelihood estimator though the difference is minimal.

The other result we obtain from the graphs is about consistency of the estimators. Since the initial estimator (`MLE`) is asymptotically consistent then so is the shrinkage estimator (`JSSE`) since the graphical trend of the two estimators is the same. Both graphs converge to some value in a neighbourhood created for any epsilon, implying that there will be no error in estimating the true parameter value at the point of convergence. Hence from the graphs we conclude that the James-Stein shrinkage estimator is asymptotically consistent and efficient. Since we are able to achieve a reduction in `MSE`, we also conclude that the James-Stein shrinking strategy used in the study is effective.

The James-Stein shrinkage estimator is more effective than the maximum likelihood estimator as shown in the `MSE` simulation plots that it has a lower mean squared error value compared to that of the `MLE`. Since error is always there in estimation then we justify shrinking (minimising error) as a very important technique for yielding effective and efficient estimators.

Chapter 7

CONCLUSION AND RECOMMENDATIONS

This chapter provides a conclusion and recommendations of the results for the whole study. We state and analyse the results obtained in line with the objectives of the study and compare them to other relevant results for some of the studies considered in the introduction in Chapter 1. In the last paragraph we make recommendations for future research.

The study considered a James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ constructed using James and Stein's shrinking techniques [35]. To have a generalised form of the JSSE, we partitioned the main parameter space into two partitions to obtain a sub-parameter space Ω_o which provided the shrinkage target $\tilde{\boldsymbol{\theta}}_n^o$ (RMLE). We considered a linear sub-parameter space Ω_o which brought some simplifications on the structures of the matrices $\mathbf{G}$, $\mathbf{V}$ and $\mathbf{E}$ determining the shrinkage value $\mathbf{G}^\top\mathbf{V}\mathbf{E}(\mathbf{E}^\top\mathbf{V}\mathbf{E})^{-1}\mathbf{E}^\top h$ for $h \in \mathbb{R}^p$. With the shrinkage target in place, we expressed the JSSE explicitly as a weighted average in terms of the weight (shrinkage factor) $\widehat{w}$, RMLE $\tilde{\boldsymbol{\beta}}_n^o$ and MLE $\hat{\boldsymbol{\beta}}_n$, therefore providing a good form to determine its asymptotic distribution. Hence achieving the first objective of the study of constructing a James-Stein shrinkage estimator from the MLE which follows a p-multivariate normal distribution $N_p(\boldsymbol{\theta}, \mathbf{V})$.

To understand very well the asymptotic behaviour of the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$, the study reviewed Hansen's [28] approach of finding the asymptotic distribution of the shrinkage estimator. Using the asymptotic normality of the MLE $\hat{\boldsymbol{\beta}}_n$, we showed that the restricted maximum likelihood estimator $\hat{\boldsymbol{\beta}}_n^o$ asymptotically converges to some normal distribution affected by some value due to shrinkage and creation of the sub-parameter space Ω_o. Since the shrinkage estimator was expressed as a weighted average of the two maximum likelihood estimators (MLE and RMLE), we also showed that the weight $\widehat{w}$ asymptotically converges to some distribution denoted by $w(Z)$ determined by the Chi-square distribution $\xi \sim \chi_p^2(h^\top\mathbf{B}h)$. This was done in order to have the asymptotic distributions for all the components which determine the shrinkage estimator. This approach is also similar to the approach Carter and Ullah [9] used to derive the sampling distribution of the shrinkage estimator obtained by shrinking the OLS estimator. Now with the asymptotic distributions of $\hat{\boldsymbol{\beta}}_n$, $\tilde{\boldsymbol{\beta}}_n^o$ and $\widehat{w}$ determined, in Theorem 4.3.2 we showed that the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ obtained from the MLE $\boldsymbol{\theta}$ converges asymptotically to some normal distribution with some shrinkage effect value. This shows that when we shrink a MLE which is asymptotically normally distributed, the new shrinkage estimator we obtain asymptotically converges to some normal distribution determined by the asymptotic distributions of the shrinkage factor, shrinkage target and initial estimator (MLE). Though Hansen [28] states how the distribution $w(Z)$ asymptotically converges in probability as ξ diverges,

in this study we went further to show all possibilities on how the distribution of the shrinkage factor converges in probability as $n \longrightarrow \infty$. Depending on the value ξ asymptotically converges to in probability, we showed that $w(Z) \longrightarrow_p \mathbf{r} \in [0, 1]$. This result is important for the consistency of the JSSE because it shows that even at an asymptotic value ($n \longrightarrow \infty$), the shrinkage factor $\widehat{w}$ regulates shrinkage and the value to which $\hat{\boldsymbol{\beta}}_n^*$ asymptotically converges to in probability. Therefore achieving the second objective for our study of reviewing Hansen's [28] approach of finding the asymptotic distribution of the JSSE.

Since shrinking is dependent on properties of the initial estimator as stated in Section 3.2.2, it was important to determine whether the RMLE $\tilde{\boldsymbol{\beta}}_n^o$ is consistent for $\boldsymbol{\theta}$ when we have the asymptotic distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n^o\right)$ as $n \longrightarrow \infty$. This is because naturally from the regularity conditions, the consistency of the MLE $\hat{\boldsymbol{\beta}}_n$ extends to the RMLE $\tilde{\boldsymbol{\beta}}_n^o$. But because of the difference in the set up of the parameter in every case arising from the two parameter spaces Ω and Ω_o, Lemma 5.1.2 proves to be vital. In Lemma 5.1.2 we showed that the RMLE $\tilde{\boldsymbol{\beta}}_n^o$ is consistent for $\boldsymbol{\theta}$ when both parameter spaces (Ω and Ω_o) are well defined. It was shown in Theorem 5.1.3 that from the consistency of the MLE $\hat{\boldsymbol{\beta}}_n$ and RMLE $\tilde{\boldsymbol{\beta}}_n^o$, the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is asymptotically consistent for $\boldsymbol{\theta}$. Thus achieving the objective for checking the asymptotic consistency of the James-Stein shrinkage estimator. This is an important asymptotic property of the estimator $\hat{\boldsymbol{\beta}}_n^*$ because it shows the stability of the estimator when we have large values for the sample size value n. Actually in practical applications as discussed by Efron [17], we normally consider large sample size values for effective estimation. Hence making the JSSE good for practical applications.

To check whether the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is biased when the sample size value n is large, we evaluated the asymptotic distributional bias (ADB) using the asymptotic distribution obtained in the study. The ADBs for the three estimators in play were evaluated and the results in Theorem 5.2.1 show that the MLE $\hat{\boldsymbol{\theta}}_n$ is asymptotically unbiased while the RMLE $\tilde{\boldsymbol{\theta}}_n^o$ and JSSE $\hat{\boldsymbol{\beta}}_n^*$ are asymptotically biased. The definition and method used to evaluate these ADBs is similar to the one Ahmed *et. al* [1] used. The results Ahmed *et. al* [1] found are the same as what we obtained concerning biasedness of the estimators, and just differ in the bias values of the RMLE $\tilde{\boldsymbol{\theta}}_n^o$ and JSSE $\hat{\boldsymbol{\beta}}_n^*$ due to the difference in the way the restricted estimator and shrinkage estimator were constructed. In both studies it is shown that both the restricted estimator and shrinkage estimator are asymptotically biased. This means that both shrinking and creation of a sub-parameter space (partitioning) to shrink to bring bias which is preserved even when the sample size value n is large. In practical applications, bias becomes a factor when it affects the distribution of the estimator, therefore in this case it might not be a big concern for the JSSE since it is normally distributed and has a lower MSE value compared to the initial estimator (MLE). As discussed by Efron [17], the effectiveness (reduced mean squared error) of the JSSE dominates biasedness.

After showing that the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ is asymptotically biased, we used the bias value to show that the variance for the shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ achieves the Cramér Rao-bound (CRB). Due to complications in evaluating the bound using matrices, we evaluated the bound component wise. The results showed that the j^{th} component (for $j = 1, 2, ..., p$) of the variance matrix of the JSSE asymptotically achieves the Cramér Rao-bound, hence making the whole p-

dimensional estimator $\hat{\boldsymbol{\beta}}_n^*$ to be efficient as $n \longrightarrow \infty$. This again shows the stability and effectiveness of the James-Stein shrinkage estimator as the sample size value n becomes large. From its consistency and efficiency, we conclude that the `JSSE` converges in probability to the "true value" as $n \longrightarrow \infty$.

The convergence rate of the James-Stein shrinkage estimator was analysed using the concept of differentiability in quadratic mean (`DQM`) and local asymptotic normality (`LAN`) as used by Halonen in [24] to analyse the convergence rate of the `MLE`. We found that the James-stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ converges at $k\sqrt{n}$ rate while the initial estimator (`MLE`) converges at $\sqrt{n}$ rate. To further analyse the rate of convergence we produced simulation plots in R. The simulation plots showed that the concentration of the `MSE` value for the `JSSE` is at 0.9 while that of the `MLE` is at 1.0. The `MSE` plots show that the mean squared error for both estimators (`MLE` and `JSSE`) stabilise as $n \longrightarrow \infty$ and converges to some value. We conclude from the simulation plots that the James-Stein shrinkage estimator obtained by shrinking the `MLE` converges faster and has lower `MSE` compared to the `MLE`. Hence achieving the last objective of checking whether the James-Stein is asymptotically efficient.

This study established asymptotic consistency and efficiency by only analysing the increase in the sample size value n with a constant (non-varying) p the number of parameters. The future perspective of this study is to analyse the effect of increasing the number of parameters p relative to an increase in the sample size value n on the mean squared error (`MSE`) of the James-Stein shrinkage estimator. The other perspective is to study the shrinkage factor, specifically the denominator $D_n \sim \chi_p^2$ and examine different forms of possible distributions D_n can follow to reduce the asymptotic mean squared error and convergence rate and then compare the two estimators (the `JSSE` and the improved `JSSE`) using the same method used in the study. One might also want to check if the James-Stein shrinkage estimator $\hat{\boldsymbol{\beta}}_n^*$ converges in distribution to the restricted maximum estimator $\tilde{\boldsymbol{\theta}}_n^o$.

APPENDIX

A. Selected Proofs

We provide the following proofs for some results used in the study.

Lemma 3.2.33 (Proof)

Assume the density $f_\theta(x)$ is continuous and that we are integrating over the whole parameter space.

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\frac{\partial \log f_{\theta_0}(x)}{\partial\theta}\right] &= \int_A \frac{\partial \log f_{\theta_0}(x)}{\partial\theta} f_{\theta_0}(x)dx \\
&= \int_A \frac{1}{f_{\theta_0}} \frac{\partial f_{\theta_0}(x)}{\partial\theta} f_{\theta_0} dx \quad \textit{by definition of the derivative of a} \ \log \\
&= \int_A \frac{\partial f_{\theta_0}(x)}{\partial\theta} dx \\
&= \frac{\partial}{\partial\theta} \int_A f_{\theta_0}(x)dx \\
&= \frac{\partial}{\partial\theta} 1 \quad \textit{since by definition} \ \int_A f_{\theta_0}(x) = 1 \\
&= 0 \quad \textit{when we differentiate with respect to } \theta.
\end{aligned}$$

Lemma 3.2.34 (Proof)

Assume $f_\theta(x)$ is continuous and we integrate over A such that $A = \{x : f_\theta(x) > 0\}$.

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\frac{\partial^2 \log f_{\theta_0}(x)}{\partial\theta\partial\theta'}\right] &= \int_A \frac{\partial^2 \log f_{\theta_0}(x)}{\partial\theta\partial\theta'} f_{\theta_0}(x)dx \\
&= \int_A \frac{\partial}{\partial\theta}\left(\frac{\partial \log f_{\theta_0}(x)}{\partial\theta'}\right) f_{\theta_0}(x)dx \\
&= \int_A \frac{\partial}{\partial\theta}\left(\frac{1}{f_{\theta_0}(x)} \frac{\partial f_{\theta_0}(x)}{\partial\theta'}\right) f_{\theta_0}(x)dx \\
&= \int_A \left[-\frac{1}{(f_{\theta_0}(x))^2} \frac{\partial f_{\theta_0}(x)}{\partial\theta} \frac{\partial f_{\theta_0}(x)}{\partial\theta'} + \frac{1}{f_{\theta_0}(x)} \frac{\partial^2 f_{\theta_0}(x)}{\partial\theta\partial\theta'}\right] f_{\theta_0}(x)dx \\
&= -\int_A \left[\frac{1}{f_{\theta_0}(x)} \frac{\partial f_{\theta_0}(x)}{\partial\theta}\right]\left[\frac{1}{f_{\theta_0}(x)} \frac{\partial f_{\theta_0}(x)}{\partial\theta'}\right] f_{\theta_0}(x)dx + \int \frac{\partial^2 f_{\theta_0}(x)}{\partial\theta\partial\theta'} dx
\end{aligned}$$

$$= -\int_A \frac{\partial \log f_{\theta_0}(x)}{\partial \theta} \frac{\partial \log f_{\theta_0}(x)}{\partial \theta'} f_{\theta_0}(x) dx \quad since \int_A \frac{\partial^2 f_{\theta_0}(x)}{\partial \theta \partial \theta'} dx = 0$$

$$= -\mathbb{E}_{\theta_0}\left[\frac{\partial \log f_{\theta_0}(x)}{\partial \theta} \frac{\partial \log f_{\theta_0}(x)}{\partial \theta'}\right].$$

Lemma 3.2.18 (Proof)

From the study [52] Stein showed that the risk for a positive James-Stein shrinkage estimator $\hat{\boldsymbol{\theta}}_n^*$ is given by

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n^*) = 1 + \frac{1}{p}\mathbb{E}_{\boldsymbol{\theta}}\left[|g(\hat{\boldsymbol{\theta}}_n)|^2\right] - \frac{2}{p}\sum_{i=1}^{p}\mathbb{E}_{\theta}\frac{\partial}{\partial \hat{\theta}_i} g_i(\hat{\theta}) \tag{7.1}$$

where $g_i(\hat{\theta}) = (p-2)^2 \frac{\hat{\theta}_i}{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}{}^\top \hat{\boldsymbol{\theta}}_n}$ and $|g(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})|^2 = \frac{(p-2)^2}{\hat{\boldsymbol{\theta}}_n}$. Differentiating we have

$$\frac{\partial}{\partial \theta_i} g_i(\hat{\boldsymbol{\theta}}_n) = \frac{(p-2)}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)^2}\left[\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n - 2\hat{\theta}_i^2\right]$$

and hence

$$\begin{aligned}
\sum_{i=1}^{p}\frac{\partial}{\partial \theta_i} g_i(\hat{\boldsymbol{\theta}}_n) &= \frac{(p-2)}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)^2}\sum_{i=1}^{p}\left[\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n - 2\hat{\theta}_i^2\right] \\
&= \frac{(p-2)}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)^2}\left[p(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n) - 2\sum_{i=1}^{p}\hat{\theta}_i^2\right] \\
&= \frac{(p-2)}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)^2}\left[p(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n) - 2(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)\right] \quad since \quad \sum_{i=1}^{p}\hat{\theta}_i^2 = \hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n \\
&= \frac{(p-2)}{(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)^2}(\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n)(p-2) \\
&= \frac{(p-2)^2}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}.
\end{aligned}$$

Therefore, substituting $\sum_{i=1}^{p}\frac{\partial}{\partial \theta_i} g_i(\hat{\boldsymbol{\theta}}_n) = \frac{(p-2)^2}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}$ in (7.1) we get

$$\begin{aligned}
R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n^*) &= 1 + \frac{1}{p}\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{(p-2)^2}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\right] - \frac{2}{p}\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{(p-2)^2}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\right] \\
&= \frac{(p-2)^2}{p}\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{\hat{\boldsymbol{\theta}}_n^\top \hat{\boldsymbol{\theta}}_n}\right].
\end{aligned}$$

Theorem 3.2.28 (Proof)

Let $\hat{\alpha}$ denote the value that maximises $\mathbf{L}^*(\alpha/X)$. We must show that $\mathbf{L}^*(\hat{\alpha}/X) = \mathbf{L}^*(\tau(\hat{\theta})/X)$. Now, as stated in the theorem that the maximum of $\mathbf{L}$ and $\mathbf{L}^*$ coincide, so we have

$$\begin{aligned}
\mathbf{L}^*(\hat{\alpha}/X) &= \sup_{\alpha}\sup_{\{\theta: r(\theta)=\alpha\}}\mathbf{L}(\theta/X) \qquad by\ definition\ of\ \mathbf{L}^* \\
&= \sup_{\theta}\mathbf{L}(\theta/X) \qquad (7.2) \\
&= \mathbf{L}(\hat{\theta}/X) \quad by\ definition\ of\ \hat{\theta}
\end{aligned}$$

where the second equality follows because the iterated maximization is equal to the unconditional maximization over θ, which is attained at the maximum $\hat{\theta}$. Furthermore

$$\begin{aligned} \mathbf{L}(\hat{\theta}/X) &= \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} \mathbf{L}(\theta/X) \quad \textit{since } \hat{\theta} \textit{ is the } \texttt{MLE} \\ &= \mathbf{L}^*(\tau(\hat{\theta})/X) \quad \textit{by definition of } \mathbf{L}^*. \end{aligned} \tag{7.3}$$

Hence combining the equalities in (7.2) and (7.3) we have that $\mathbf{L}^*(\hat{\alpha}/X) = \mathbf{L}^*(\tau(\theta)/X)$ and that $\tau(\hat{\theta})$ is the maximum likelihood estimator MLE of $\tau(\theta)$.

Corollary 3.2.45 (Proof)

From the weak law of large numbers (WLLN) we have

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \longrightarrow_p \sigma^2. \tag{7.4}$$

Taking square roots on both sides of (7.4) we get

$$S \longrightarrow_p \sigma. \tag{7.5}$$

Thus using Slutsky and limit theorems on probability and distribution convergence, it follows from (7.5) that

$$\frac{S}{\sigma} \longrightarrow_p \frac{\sigma}{\sigma} = 1 \tag{7.6}$$

and $\frac{S^2}{\sigma^2} \longrightarrow_p 1$ follows direct from (7.6).

Proposition 3.2.52 (Proof)

The first part is a straight application of Slutsky's theorem. It therefore hold directly from Slutsky's result. For the second part, we do a Taylor expansion of $g(\hat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}_o$ to obtain

$$g(\hat{\boldsymbol{\theta}}_n) = g(\boldsymbol{\theta}_o) + \frac{\partial g(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o\right) \tag{7.7}$$

for some $\boldsymbol{\theta}^*$ such that $\boldsymbol{\theta}^* \longrightarrow \boldsymbol{\theta}_o$ as $n \longrightarrow \infty$. Hence rearranging (7.7) and introducing $\sqrt{n}$ both sides we have

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_o)\right) = \frac{\partial g(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o\right).$$

Now, as $n \longrightarrow \infty$, we have

$$\frac{\partial g(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \longrightarrow \frac{\partial g(\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} = g'(\boldsymbol{\theta}_o) \quad \textit{almost surely}$$

and since

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o\right) \longrightarrow_d N\left(0, \mathbf{I}(\boldsymbol{\theta}_o)^{-1}\right),$$

then

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_o)\right) \longrightarrow_d N\left(0, g'(\boldsymbol{\theta}_o)\mathbf{I}(\boldsymbol{\theta_o})^{-1}g(\boldsymbol{\theta}_o)\right) \quad \textit{hence the result.}$$

B. Commands of Graphs in R

The R statistical package used for these commands is R version 3.1.2 (2014-10-31).

Function Programme

Start

```
ms <- function(n) {
   rho <- cbind(c(1, .3, .1), c(.3,  1, .2), c(.1, .2, 1))
   library(MASS)
   x <-mvrnorm (n, mu=1:3,   Sigma=rho)
   E<- matrix(c(sum(x[, 1])/n, sum(x[, 2])/n, sum(x[, 3])/n), 3, 1)
   V <- var(E)
  return(V)
   }
```

Note 1

- *ms represent the* `MSE` *function for the* `MLE` *and n is the sample size.*
- *We use $p = 3$ as the number of parameters that is why the rho is a 3×3 matrix.*
- *E is the maximum likelihood estimator (*`MLE`*) $\hat{\boldsymbol{\theta}}_n$ since for the normal distribution it is $\bar{X}$.*
- *V is the variance and the* `MSE` *since the bias for the* `MLE` *is zero.*

```
me <- function(n){
   rho <- cbind(c(1, .3, .1), c(.3, 1, .2), c(.1, .2, 1))
   library(MASS)
   x <-mvrnorm (n, mu=1:3,  Sigma=rho)
   E<- matrix(c(sum(x[, 1])/n,  sum(x[, 2])/n, sum(x[, 3])/n), 3,  1)
   F <- var(x)
   G <- t(E)%*%F
   H <- G%*%E
   I <- 1/H
   J <- (1-I)
   D <- drop(J) # as a scalar
   K <- D*E
   bias2 <- (mean(K)-mean(E))^2
   L <- var(K)
   M <-(L+bias2)
   return(M)
   }
```

Note 2

- *me is the* `MSE` *function for the* `JSSE` *and n is the sample size.*
- *F is the variance of X.*

- *H is the D_n statistic value.*
- *D is a scalar of shrinkage.*
- *K is the James-Stein shrinkage estimator (`JSSE`).*
- *bias2 is the squared bias of the `JSSE`.*
- *L is the variance of the `JSSE`.*
- *M is the Mean Squared Error (`MSE`) of the `JSSE`.*
- *ME in the "for loop" below represent the `MSE` of the `JSSE`.*

For Loop

```
samplesize <- 3:30, 2000, 8000, 100000
          MSE <- c()
          ME <- c()
          for (i in samplesize){
          MSE <- c(MSE, ms(i))
          ME <- c(ME, me(i))
          }
```

End

The commands above were used to produce the graphs which have both the graphs for the `MLE` and `JSSE`. The commands for the single graphs are presented below.

MLE (Maximum Likelihood Estimator)

```
ms <- function(n) {
   rho <- cbind(c(1, .3, .1), c(.3, 1, .2), c(.1, .2, 1))
library(MASS)
   x <-mvrnorm (n, mu=1:3, Sigma=rho)
   E<- matrix(c(sum(x[, 1])/n,  sum(x[, 2])/n, sum(x[, 3])/n), 3, 1)
   V <- var(E)
   return(V)
   }
samplesize <- 3:50000
          MSE <- c()
          for (i in samplesize){
          MSE <- c(MSE, ms(i))
          }
plot(samplesize, MSE,
     main="MSE for MLE: n=50000",
     xlab="Sample Size (n)",
     ylab="Mean Squared Error  (MSE)",
     col="blue",
     type="l", prob=TRUE,
     cex.lab=1.5, cex.axis=1.5,   cex.main=2.0, cex.sub=1.5)
```

JSSE (James-Stein Shrinkage Estimator)

```
me <- function(n){
rho <- cbind(c(1, .3, .1), c(.3, 1, .2), c(.1, .2, 1))
library(MASS)
x <-mvrnorm (n, mu=1:3, Sigma=rho)
E<- matrix(c(sum(x[, 1])/n, sum(x[, 2])/n, sum(x[, 3])/n), 3, 1)
F <- var(x)
G <- t(E)%*%F
H <- G%*%E
I <- 1/H
J <- (1-I)
D <- drop(J) # as a scalar
K <- D*E
bias2 <- (mean(K)-mean(E))^2
L <- var(K)
M <- L+bias2
return(M)
}
samplesize <- 3:50000
ME <- c()
for (i in samplesize){
ME <- c(ME, me(i))
}
plot(samplesize, ME,
main="MSE for JSSE: n=50000",
xlab="Sample Size (n)",
ylab="Mean Squared Error (MSE)",
col="red",
type="l", prob=TRUE,
cex.lab=1.5, cex.axis=1.5, cex.main=2.0, cex.sub=1.5)
```

Plotting

For $n = 30$

```
samplesize <- 3:30
MSE <- c()
ME <- c()
for (i in samplesize){
MSE <- c(MSE, ms(i))
ME <- c(ME, me(i))
}
plot(samplesize, MSE,
main="MSE Plots for JSSE-MLE: n=30",
xlab=" Sample Size (n)",
ylab= "Mean Squared Error (MSE)",
col="blue",
type="b",lwd=3.5, prob=TRUE,
cex.lab=1.5, cex.axis=1.5, cex.main=2.0, cex.sub=1.5, las=1)
lines(samplesize, ME,
col="red",lwd=3)
legend ("topright",
```

```
c( "MLE", "JSSE"),
fill=c("blue", "red")
)
# For n = 2000, 8000 and 100000 #
samplesize <- 3:2000, 8000, 100000,
MSE <- c()
ME <- c()
for (i in samplesize){
MSE <- c(MSE, ms(i))
ME <- c(ME, me(i))
}
plot(samplesize, MSE,
main="MSE plots for JSSE-MLE: n=2000",
xlab=" Sample Size (n)",
ylab= "Mean Squared Error (MSE)",
col="blue",
type="l", prob=TRUE,
cex.lab=1.5, cex.axis=1.5, cex.main=2.0, cex.sub=1.5)
lines(samplesize, ME,
col="red")
legend ("topright",
c( "MLE", "JSSE"),
fill=c("blue", "red")
)
```

REFERENCES

[1] Ahmed, S. E., Doksum, S. and You, J. (2007). "Shrinkage, Pretest and Absolute Penalty Estimators in Partially Linear Models." *Australian and New Zealand Journal of Statistics*, **49**(4), 435-454.

[2] Amirdjanova, A. and Woodroofe, M. (2004). "Shrinkage Estimation for Convex Polyhedral Cones." *Statistics and Probability Letters*, **70**, 87-94.

[3] Baranchik, A. J. (1964). "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution." *Technical Report No. 51,* Department of Statistics, Stanford University.

[4] Berger, J. O. (1976). "Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss." *Journal of Multivariate Analysis*, **6**, 256-264.

[5] Berger, L. O. (1982). "Selecting a Minimax Estimator of a Multivariate Normal Mean." *The Annals of Statistics*, **10**, 81-92.

[6] Bobrovsky, B. Z., Mayer-Wolf, E. and Zakai, M. (1987). "Some Classes of Global Cramér-Rao Bounds." *The Annals of Statistics,* **15**(4), 1421-1438.

[7] Bock, M. E. (1988). "Shrinkage estimators: Pseudo-Bayes rules for Normal mean vectors." *In statistical decision theory and related topics*, **1**, 281-297.

[8] Carsten, H. B. and Michael, J. D. (2006). "A Shrinkage Estimator for Spectral Densities." *Biometrics*. **93**(1), 179-195.

[9] Carter, R. L. and Ullah, A. (1984). "The Sampling Distribution of Shrinkage Estimators and their F-Ratios in Regression Model." *Journal of Econometrics*, **25**, 109-122.

[10] Casella, G. and Hwang, J. T. (1982). "Limit expressions for the risk of James-Stein estimators." *Canadian Journal of Statistics*, **10**, 305-309.

[11] Casella, G. and Berger, R. L. (2002). Statistical Inference. *Second edition, New York*: Thomson Learning Inc.

[12] Chernoff, H. (1956). "Large-sample theory: Parametric case." *The Annals of Mathematical Statistics* , **27**(1), 1-22.

[13] Constrantirios, G. (1996). " Partial Least Squares Algorithm Yields Shrinkage Estimators." *The Annals of Statistics*, **24**(2), 816-824.

[14] Cramẽr, H. (1946). Mathematical methods of Statistics. *New York:* Princeton University Press.

[15] Doob, J. L. (1934). "Probability and Statistics." *American Mathematical Society,* **36**(4), 759-775.

[16] Efron, B. (1975). "Biased versus unbiased estimation." *Advances in mathematics, New York:* Academic press.

[17] Efron, B. and Morris, C. (1975). "Data Analysis using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association*, **70**(350), 311-319.

[18] Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistical." *philosophical Transactions of the Royal Society*, **222**, 311-319.

[19] Fisher, J. T. and Xiaoqian, S. (2011). "Improved Stein-type Shrinkage Estimators for the high-dimensional Multivariate Normal Covariance matrix." *Computational Statistics and Data Analysis*, **55**, 1909-1918.

[20] Geyer, C. J. (1994). "On the Asymptotics of Constrained M-estimation." *Annals of Statistics,* **22**, 1993-2010.

[21] George, E. I (1986). "Combining Minimum Shrinkage Estimators." *Journal of the American,* **81**(394), 437-445.

[22] George, E. I. (1986). "Minimax multiple shrinkage estimation." *The Annals of Statistics*, **14**(1), 188-205.

[23] Green, E. J. and Strawderman, W. E. (1991). "A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators." *Journal of the America Statistical Association,* **86**(416), 1001-1006.

[24] Halonen, B. (2015). Rate of Convergence of Maximum Likelihood Estimators under Relaxed Smoothness Conditions on the Likelihood Function. *Master's Report,* Michigan Technological University.

[25] Hannes, L. and Pötsher, B. M. (2006). "Performance Limits for Estimates of Risk or Distribution of Shrinkage-Type Estimators and some General lower Risk- bound results." *Econometric Theory*, **22**(1), 69-97.

[26] Hansen, E. B. (2007). "Least Squares Model Averaging." *Econometrica*, **75**, 1175-1189.

[27] Hansen, E. B. (2008). "Generalised Shrinkage Estimators." *www.ssc.wisc.edu/∼ bhansen.* United States.

[28] Hansen, E. B. (2016). "Efficient Shrinkage in Parametric Models." *Journal of Econometrics*, **190**(2016), 115-132.

[29] Hansen, E. B. (2016). "The Risk of James-Stein and Lasso Shrinkage." *Econometric Reviews*, **35**(8-10), 1456-1470.

[30] Hansen, E. B. (2015). "Shrinkage Efficiency Bounds." *www.ssc.wisc.edu/∼bhansen.* United States.

[31] Hansen, E. B. (2014). "Asymptotic moments of Autogressive estimators with a near unit root and minimax." *Advances in Econom trics*, **33**, 3-21.

[32] Hoeffding, W. (1948). "A Class of Statistics with Asymptotically Normal Distribution." *Annals Mathematical Statistics,* **19**(3), 293-325.

[33] Hossain, S., Kjell, A. D. and Ahmed, E. S. (2009). "Positive shrinkage, Improved pretest and absolute penalty estimators in Partially linear Models." *Linear Algebra and its Application,* **430**, 2749-2761.

[34] Hodges, J. L. Jr. and Lehmann, E. L. (1951). "Some applications of the Cramér-Rao Inequality." *Proceedings of second Berkeley Symposium on mathematical statistics and probability*, University of California Press, 13-22.

[35] James, W. and Stein, C. (1961). "Estimation with Quadratic Loss." *Berkeley Symposium on Mathematical Statistics and Probability*, Stanford University.

[36] Ki, F. and Tsui, K. (1990). "Multiple Shrinkage Estimators of means in Exponential Families." *The Canadian Journal of Statistics*, **18**(1), 31-46.

[37] Knight, k. (2008). "Shrinkage Estimator for Nearly Singular Designs." *Econometrics Theory,* **2**, 323-337.

[38] Knight, K. and Wenjiang, F. (2000). "Asymptotics for Lasso-type Estimators." *Annals of Statistics,* **28**(5), 1356-1378.

[39] Le Cam, L. (1970). " Assumptions used to Prove Asymptotic Normality of Maximum Likelihood Estimates." *The Annals of Mathematical Statistics,* **3**(41), 802-828.

[40] Le Cam, L. (1970). "Limits of experiments." *Proceedings of the sixth Berkeley symposium on Mathematical Statistics and Probability,* University of California press, **1**,245-261.

[41] Le Cam, L. (1986). Asymptotic Methods in Statistical Decision Theory. *New York:* Springer-Verlang.

[42] Lehmann, E. L., Casella, E., George, I. (1998). Theory of Point Estimation. *Second edition, New York:* Springer-Verlang.

[43] Lehmann, E. L., Casella, E., George, I. (2004). Elements of Large Sample Theory. *New York:* Springer-Verlang.

[44] Mann, H. B. and Wald, A. (1943). "On stochastic limit and order relationships." *Annals of mathematical statistics,* **14**, 217-226.

[45] McLeish, D. L. and Struthers, C. A. (2013). "Estimation and Hypothesis", Supplementary Lecture notes. *University of Waterloo,* Ontario, Canada.

[46] Newey, W. K. and Mcfadden, D. L. (1994). "Large sample estimation and hypothesis testing." *Handbook of Econometrics*, **4**, 2111-2245.

[47] Oman, S. D. (1982a). "Contracting towards subspaces when estimating the mean of a multivariate normal distribution." *Journal of Multivariate Analysis*, **12**, 270-290.

[48] Oman, S. D. (1982b). "Shrinking towards subspaces in multiple linear regression." *Technometrics*, **(**24), 307-311.

[49] Rising, J. K. and Wyner, A. J. (2012). "Partial Kelly Portfolios and Shrinkage Estimators." *IEEE International Symposium on Information Theory Proceedings*, 1618-1622.

[50] Robinson, P. M. (1988). "Root-N-Consistent Semiparametric Regression." *Econometrica*, **56**(4), 931-954.

[51] Stein, C. (1956). "Inadmissibility of the usual Estimator for the Mean of a Multivariate Normal Distribution." *Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 197-206.

[52] Stein, C. (1981). "Estimation of the Mean of a Multivariate Normal Distribution." *The Annals of Statistics*, **9**(6), 1135-1151.

[53] Stone, C. J. (1974). "Asymptotic properties of Estimators of a location parameter." *Annals of Statistics,* **2**(6), 1127-1137.

[54] Sudheesh, K. K. (2009). "On Stein's Identity and its Applications." *Statistics and Probability Letters,* **79**(12), 1444-1449.

[55] Van der Vaart, A. W. (1998). Asymptotic Statistics. *London:* Cambridge University Press.

[56] Yuzo, M. (2007). "Some notes on improving upon the James-Stein Estimator." *The University of Tokyo,* Tokyo.

[57] Zhipen, L. (2010). "Adaptive GMM shrinkage Estimation with Consistent Moment selection." *Department of Economics*. UC Los Angels.