Shrinkage Estimators

Richard Samworth

St. John's College

I declare that this essay is work done as part of the Part III Examination. It is the result of my own work, and except where stated otherwise, includes nothing which was performed in collaboration. No part of this essay has been submitted for a degree or any such qualification.

Signed ....

Home address:
Cherry Tree House,
Askham Bryan,
York.
YO23 3QS

Shrinkage Estimators

1 Introduction

In his celebrated 1956 paper, Stein [28] considered the problem of estimating the mean vector θ of a p-variate normal distribution with identity covariance under quadratic loss, based on one observation. In other words, with $X|\theta \sim N_p(\theta, I)$, the problem is to estimate θ under the loss function

$$L(\theta, \delta) = ||\delta - \theta||^2 = \sum_{i=1}^{p} (\delta_i - \theta_i)^2.$$

The admissibility of the estimator $\delta_0(X) = X$ for $p = 1$ had already been proved by, e.g., Blyth [6], and Stein's aim had been to demonstrate admissibility for $p \geq 2$. Instead he found that X was admissible for $p = 2$ and inadmissible for $p \geq 3$. In particular, Stein showed that 'shrinkage' estimators of the form

$$\delta_{a,b}(X) = \left(1 - \frac{a}{b + ||X||^2}\right) X$$

dominate X for $a > 0$ sufficiently small and b sufficiently large. Five years later, James and Stein [25] observed that $0 < a < 2(p-2)$ and $b = 0$ suffice, i.e. that estimators of the form

$$\delta_a(X) = \left(1 - \frac{a}{||X||^2}\right) X$$

dominate X for $0 < a < 2(p-2)$. Furthemore, they were able to show that δ_{p-2} had the uniformly smallest risk of any estimator in this class.

From a mathematical viewpoint, one of the disturbing features of this result concerns the notions of invariance and equivariance. The problem of estimating θ on the basis of X does not depend on a choice of origin; formally, it is location invariant, in the sense that the loss function L satisfies

$$L(\theta + a, \delta + a) = L(\theta, \delta) \text{ for all } a \in \mathbb{R}^p.$$

One might therefore expect a location equivariant estimator, that is, one satisfying

$$\delta(X + a) = \delta(X) + a \text{ for all } a \in \mathbb{R}^p,$$

to be preferable to one depending on a choice of origin. It is clear, however, that while X is equivariant, the James-Stein estimators δ_a are not.

Since Stein's paper was published, much work has been done to try to generalise the result to the estimation of the mean vector θ of a p-variate location parameter family. Through many small steps, much progress has been made, and these developments are outlined in Section 4. Before this, in Section 2, we give an empirical Bayes argument to indicate that certain shrinkage estimators might be expected to dominate X in greater generality than the multivariate normal case. In Section 3, we give a detailed treatment of the problem mentioned above, where X has a p-variate normal distribution. Apart from the obvious importance of this case in its own right, both the theorems and techniques of proof are indicative of the more general results. In addition, we are able to exhibit admissible estimators for this case. Section 5 discusses the paper of Evans and Stark [21], which contains the most powerful result to date, and which uses the idea of representing the distribution of X as that of a stopped Brownian motion.

2 An Empirical Bayes Argument

The following empirical Bayes argument is well-known (see Efron and Morris [19], Lehmann [26, pp. 299–302], Brandwein and Strawderman [12]), and leads to the optimal James-Stein estimator of Section 2.

Let $X|\theta \sim N_p(\theta, I)$ with $p \geq 2$, and suppose the prior distribution on θ is $\theta \sim N_p(0, bI)$, where b is unknown. Suppose we wish to estimate θ under quadratic loss. The posterior density of θ given X is easily calculated as follows:

$$\begin{aligned}\pi(\theta|X) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{p}(X_i - \theta_i)^2 - \frac{1}{2b}\sum_{i=1}^{p}\theta_i^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(1+\frac{1}{b}\right)\sum_{i=1}^{p}\left(\theta_i - \frac{X_i}{1+\frac{1}{b}}\right)^2\right).\end{aligned}$$

In other words, $\theta|X \sim N_p\left(\frac{bX}{b+1}, \frac{b}{b+1}\right)$.

The Bayes estimate $\delta_b(X)$ of θ is given by the posterior mean, i.e.

$$\delta_b(X) = \left(1 - \frac{1}{b+1}\right)X.$$

Of course, δ_b is not a valid estimator of θ as b is unknown, but we can estimate it from the data. One way to do this (for another, see Efron and Morris [19] or Lehmann [26, pp. 299–302]) is to first calculate the marginal distribution of X. This can either be done directly using the formula (with obvious notation)

$$f_X(x) = \int_{\mathbb{R}^p} f(x;\theta)\pi(\theta)d\theta,$$

or by noting that $(X - \theta)|\theta \sim N_p(0, I)$, independent of θ, so that $X - \theta$ and θ are independent. Thus the marginal distribution of $X = X - \theta + \theta$ is p-variate normal with mean 0 and covariance matrix $(b+1)I$.

If X were scalar, one observation would give no information about the variance b. However, when X is a p-vector, we can use one observation to gain a meaningful estimate of b, based on the fact that the components of X are independently normally distributed with mean 0 and variance $b+1$. This means that $\frac{\|X\|^2}{b+1} \sim \chi_p^2$, so $\mathbb{E}\left(\frac{p-2}{\|X\|^2}\right) = \frac{1}{b+1}$, since $\mathbb{E}(\frac{1}{\chi_p^2}) = \frac{1}{p-2}$. Thus we might reasonably estimate $\frac{1}{b+1}$ by $\frac{p-2}{\|X\|^2}$.

This procedure yields the estimator $\left(1 - \frac{p-2}{\|X\|^2}\right)X$, which is precisely the James-Stein estimator δ_{p-2} of Section 1.

Remark: Observing that $\mathbb{E}\left(\frac{p-2}{\|X\|^2}\right) = \frac{1}{b+1}$ and using $\frac{p-2}{\|X\|^2}$ as an estimate of $\frac{1}{b+1}$ is particularly convenient, in that it leads to the optimal James-Stein estimator. However, it is also slightly contrived, and one might equally well note that $\mathbb{E}\left(\frac{\|X\|^2}{p}\right) = b+1$, whereupon estimating $b+1$ by $\frac{\|X\|^2}{p}$ yields the estimator δ_p. This estimator also dominates X for $p \geq 5$, and is close to the optimal estimator if p is large.

Remark: It is reasonably clear that the assumption of normality for the prior distribution of θ should not be crucial if p is large, since the effect of the data on the posterior distribution will 'swamp' that of the prior distribution.

3 The Normal Case

3.1 Location Parameter Family Preliminaries

Definition 3.1. *Suppose $f(\mathbf{x})$ is a p-variate density function. The family of densities $f(\mathbf{x} - \theta)$, indexed by $\theta \in \mathbb{R}^p$ is called a location parameter family, and θ is called the location parameter.*

Remark: Note that whenever $\mathbb{E}_\theta(X) < \infty$, we may, by shifting the indexing parameter θ if necessary, assume without loss of generality that $\mathbb{E}_0(X) = 0$.

Remark: If $\mathbb{P}_\theta$ is the probability measure corresponding to the density $f(\mathbf{x}-\theta)$, then the distribution of X under $\mathbb{P}_\theta$ is the same as the distribution of $X + \theta$ under $\mathbb{P}_0$. In particular, $\mathbb{E}_\theta(X) = \theta$.

Recall that the general problem is to estimate the mean vector θ of a p-variate location parameter family under the quadratic loss function

$$L(\theta, \delta) = ||\delta - \theta||^2 = \sum_{i=1}^{p} (\delta_i - \theta_i)^2. \tag{1}$$

When $X|\theta \sim N_p(\theta, I)$, the risk function of the estimator $\delta_0(X) = X$ is given by

$$R(\theta, \delta_0) = \mathbb{E}_\theta ||X - \theta||^2 = \sum_{i=1}^{p} \mathrm{Var}(X_i) = p,$$

a constant, independent of θ. In fact, this result holds more generally:

Lemma 3.2. *Suppose the density of X is from a location parameter family. Then the risk of any equivariant estimator $\delta(X)$ is constant.*

Proof. We have

$$R(\theta, \delta) = \mathbb{E}_\theta ||\delta(X) - \theta||^2 = \mathbb{E}_0 ||\delta(X + \theta) - \theta||^2 = \mathbb{E}_0 ||\delta(X)||^2,$$

a constant, independent of θ. □

Thus if we can find a location equivariant estimator with smallest (constant) risk, it makes sense to speak of a best equivariant estimator.

When X is an observation from a location parameter family with mean vector θ, the estimator $\delta_0(X) = X$ is the best equivariant estimator (see Berger [4, pp. 247–248] for a proof). Furthermore, it follows from Kiefer [24] that X is minimax. Brown [14] extended Stein's original result, showing that the best equivariant estimator of a location parameter is inadmissible for quadratic loss (1) when $p \geq 3$.

It is clear that if we can find an estimator $\delta(X)$ which dominates X, then it too will be minimax, so in particular any admissible estimator is minimax. Conversely, any minimax estimator must be at least as good as X, in that

$$R(\theta, \delta) \geq R(\theta, X) \text{ for all } \theta \in \mathbb{R}^p.$$

Therefore, if it is more convenient, we might say that a class of estimators is minimax rather than that it dominates X, although typically it is only at the endpoints of a range of minimax estimators that we might have equality of the risk function for all values of θ.

3.2 James-Stein Estimators

As previously noted, Stein [28] proved that when $X|\theta \sim N_p(\theta, I)$, estimators of the form

$$\delta_{a,b}(X) = \left(1 - \frac{a}{b + ||X||^2}\right) X$$

dominate X for $a > 0$ sufficiently small and b sufficiently large. His proof involved an information inequality argument. James and Stein [25] improved the result, giving an explicit class of dominating estimators

$$\delta_a(X) = \left(1 - \frac{a}{||X||^2}\right) X$$

fo $0 < a < 2(p-2)$. Their new proof used the fact that a non-central chi-squared random variable with p degrees of freedom and non-centrality parameter $||\theta||^2$ (written $\chi^2_p(||\theta||^2)$) has the same distribution as a random variable W, where W is obtained by taking $K \sim \text{Poi}\left(\frac{||\theta||^2}{2}\right)$, and letting $W|K \sim \chi^2_{p+2K}$. Both this proof and the earlier one are relatively long and are omitted here.

However in 1981, Stein [30] published his 'unbiased estimation of risk' technique, which had been known to him since 1974, and which simplifies the proof of the James-Stein result substantially. It depends upon a simple integration by parts identity, which has become known as Stein's Lemma.

Lemma 3.3 (Stein's Lemma). *Let X have a univariate normal distribution with mean θ and variance 1, and let g be a real-valued differentiable function satisfying $\mathbb{E}_\theta |g'(X)| < \infty$. Then*

$$\mathbb{E}_\theta \left(g(X)(X - \theta)\right) = \mathbb{E}_\theta (g'(X)).$$

Proof. We have

$$\begin{aligned}
\mathbb{E}_\theta \left(g(X)(X-\theta)\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x)(x-\theta) e^{-\frac{1}{2}(x-\theta)^2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) \frac{d}{dx}\left(-e^{-\frac{1}{2}(x-\theta)^2)}\right) dx \\
&= \left[-\frac{1}{\sqrt{2\pi}} g(x) e^{-\frac{1}{2}(x-\theta)^2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x) e^{-\frac{1}{2}(x-\theta)^2}\, dx \\
&= \mathbb{E}_\theta (g'(X))
\end{aligned}$$

since the condition on g' is sufficient to ensure the integrated term is zero. □

Remark: This lemma can be used to simplify calculations of higher order moments. For instance, if $X|\theta \sim N(\theta, I)$, then

$$\mathbb{E}_\theta(X^3) = \mathbb{E}_\theta\left(X^2(X-\theta+\theta)\right) = 2\mathbb{E}_\theta(X) + \theta\mathbb{E}_\theta(X^2) = 2\theta + \theta(\theta^2+1) = 3\theta + \theta^3.$$

We use this lemma to obtain an unbiased estimate of the risk of $\delta_a(X)$, and hence the required result.

Theorem 3.4. *The estimator $\delta_a(X)$ dominates X for $0 < a < 2(p-2)$ when $p \geq 3$. Furthermore, the estimator*

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

dominates every other estimator in this class.

Proof.

$$\begin{aligned}
R(\theta, \delta_a(X)) &= \mathbb{E}_\theta \left\| \left(1 - \frac{a}{||X||^2}\right) X - \theta \right\|^2 \\
&= \mathbb{E}_\theta ||X-\theta||^2 + a^2 \mathbb{E}_\theta \left(\frac{1}{||X||^2}\right) - 2a\mathbb{E}_\theta \left(\frac{X^T(X-\theta)}{||X||^2}\right) \\
&= p + a^2\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right) - 2a\sum_{i=1}^p \mathbb{E}_{\theta_i}\left(\frac{X_i(X_i-\theta_i)}{\sum_{j=1}^p X_j^2}\right)
\end{aligned}$$

Applying Stein's lemma to each component X_i, we obtain

$$\begin{aligned}
R(\theta, \delta_a(X)) &= p + a^2\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right) - 2a\sum_{i=1}^p \mathbb{E}_{\theta_i}\left(\frac{d}{dX_i}\left(\frac{X_i}{\sum_{j=1}^p X_j^2}\right)\right) \\
&= p + a^2\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right) - 2a\sum_{i=1}^p \mathbb{E}_{\theta_i}\left(\frac{\sum_{j=1}^p X_j^2 - 2X_i^2}{(\sum_{j=1}^p X_j^2)^2}\right) \\
&= p + a^2\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right) - 2a\mathbb{E}_\theta\left(\frac{p||X||^2 - 2||X||^2}{||X||^4}\right) \\
&= p + (a^2 - 2a(p-2))\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right).
\end{aligned}$$

We note that the quadratic $a^2 - 2a(p-2)$ is negative for $0 < a < 2(p-2)$, and attains its minimum at $a = p-2$. This proves the theorem. □

It is both interesting and instructive to compare the risk functions of δ_{p-2} and δ_0, and this means calculating $\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right)$ explicitly. Observe that $||X||^2|\theta \sim \chi_p^2(||\theta||^2)$, and we have already noted that we can rewrite this distribution in terms of a random variable K, where $K \sim \mathrm{Poi}\left(\frac{||\theta||^2}{2}\right)$, and $||X||^2|K \sim \chi^2_{p+2K}$. Thus

$$\begin{aligned}
\mathbb{E}_\theta\left(\frac{1}{||X||^2}\right) &= \mathbb{E}_\theta\left(\mathbb{E}\left(\frac{1}{||X||^2}\,\middle|\, K\right)\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{p+2K-2}\right) \\
&= e^{-\frac{||\theta||^2}{2}}\sum_{k=0}^{\infty}\frac{1}{p+2k-2}\frac{||\theta||^{2k}}{2^k k!}.
\end{aligned}$$

Hence the risk of δ_{p-2} is given by

$$R(\theta,\delta_{p-2}(X)) = p-(p-2)^2 e^{-\frac{||\theta||^2}{2}}\sum_{k=0}^{\infty}\frac{1}{p+2k-2}\frac{||\theta||^{2k}}{2^k k!}.$$

Although we cannot write this sum in a closed form, by approximating it with a large but finite number of terms, I have obtained an accurate representation of the graph $R(\theta,\delta_{p-2}(X))$ against $||\theta||$ (see Figure 1). Note particularly that at the origin, the risk of δ_{p-2} is 2 for all values of $p \geq 3$.

Observe that δ_{p-2} is still inadmissible: it is dominated by the 'positive-part' James-Stein estimator

$$\delta^+_{p-2}(X) = \left(\max\left(0, 1-\frac{p-2}{||X||^2}\right)\right)X$$

due to the fact that the shrinkage factor becomes negative for small $||X||^2$. If $\mathbb{P}_\theta(||X||^2 < p-2) \equiv q(||\theta||)$, say, then a very similar calculation to the one above gives

$$R(\theta,\delta^+_{p-2}(X)) = q(||\theta||)||\theta||^2 + (1-q(||\theta||))\left(p-(p-2)^2 e^{-\frac{||\theta||^2}{2}}\sum_{k=0}^{\infty}\frac{1}{p+2k-2}\frac{||\theta||^{2k}}{2^k k!}\right)$$

This facilitates a numerical approximation to the risk function of δ^+_{p-2}, which I have also included in Figure 1.

Remarkably, even δ^+_{p-2} is inadmissible; we return to the question of admissibility in Section 3.4.

3.3 Choice of Attractor

The estimators δ_a in Section 3.2 shrink X towards the origin, and this is where it is observed (see Figure 1) that the improvement in risk over the usual estimator is greatest. If $\theta_0 \in \mathbb{R}^p$, estimators of the form

$$\delta_{a,\theta_0}(X) = \theta_0 + \left(1-\frac{a}{||X-\theta_0||^2}\right)||X-\theta_0||$$

shrink X towards θ_0 and it is clear that δ_{a,θ_0} dominates X provided $0 < a < 2(p-2)$, with the greatest improvement in risk being at θ_0. We are therefore interested in the problem of choosing the best 'attractor' θ_0. The Bayesian answer, of course, is that we should select θ_0 such that θ is close to

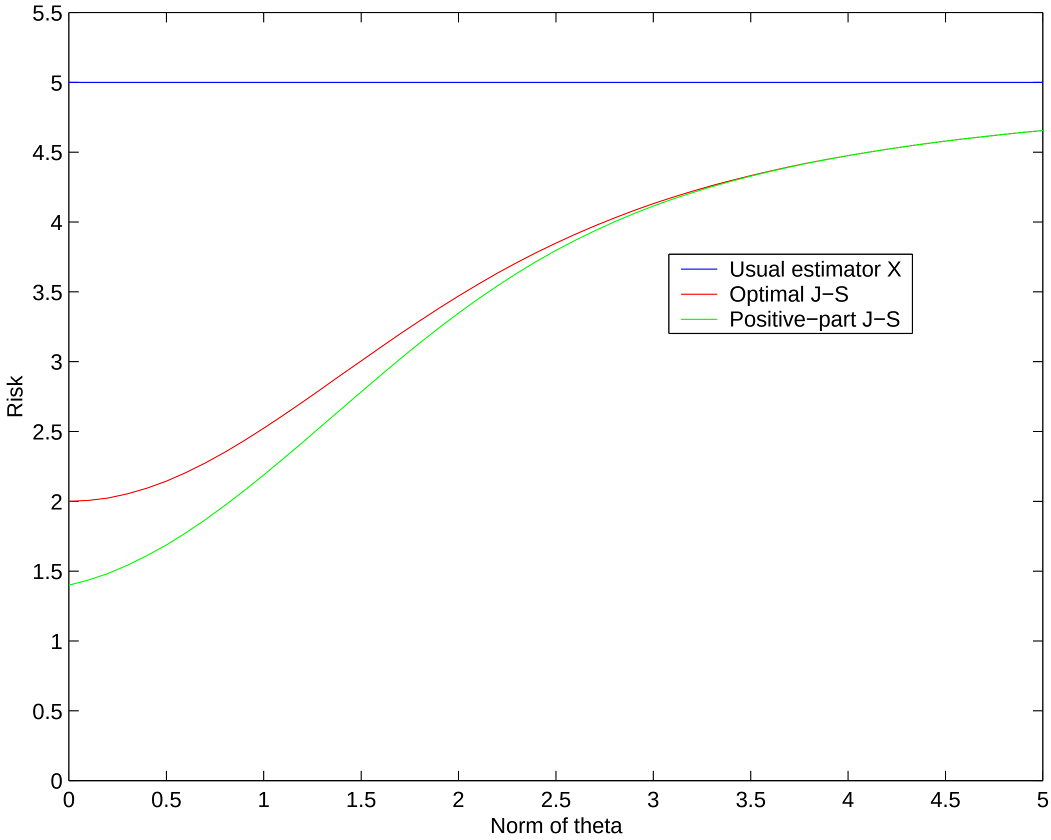

Figure 1: Risks of X, the optimal James-Stein estimator δ_{p-2} and its positive-part counterpart δ_{p-2}^{+} for $p = 5$

θ_0 with high prior probability. In fact, shrinkage estimation is only of practical benefit when there is some reason to believe that θ is likely to lie in a fairly small region. In the absence of such prior information, however, we may be able to select a resonable value of θ_0 from the data.

For instance, in the important special case where our prior belief is that the components of θ are equal, we can carry out a similar analysis to that of Section 2. Recall then that we had $X|\theta \sim N_p(\theta, I)$, but now suppose firstly that $p \geq 4$, and secondly that the prior distribution on θ is $\theta \sim N_p(\xi, bI)$. The marginal distribution of X is now $N_p(\xi, (b+1)I)$. This suggests using $\overline{X}\mathbf{1} = \left(\frac{1}{p}\sum_{i=1}^{p} X_i\right)\mathbf{1}$ as an estimator of ξ (where $\mathbf{1}$ is a p-vector of ones). Further, since

$$\sum_{i=1}^{p}(X_i - \overline{X})^2 \sim (b+1)\chi^2_{p-1},$$

we might, by the same reasoning as in Section 2, estimate $\frac{1}{b+1}$ by

$$\frac{p-3}{\sum_{i=1}^{p}(X_i - \overline{X})^2}.$$

In this way, we obtain the estimator

$$\overline{X}\mathbf{1} + \left(1 - \frac{p-3}{\sum_{i=1}^{p}(X_i - \overline{X})^2}\right)(X - \bar{X}\mathbf{1}). \tag{2}$$

This estimator used was first proposed in 1962 by Lindley in a discussion of a paper by Stein [29]. An excellent informal example and discussion of its utility in a practical situation is given in Efron and Morris [20].

As is to be expected, we pay a price in terms of risk for our estimate of θ_0; the minimum risk for the estimator in Equation (2) is 3 for all values of $p \geq 4$, and is attained on the set $\{\theta \in \mathbb{R}^p : \theta_1 = \theta_2 = \ldots = \theta_p\}$.

3.4 Finding Admissible Estimators

We return to the problem of finding admissible estimators when $X|\theta \sim N_p(\theta, I)$ and $p \geq 3$. Recall from Section 3.2 that both the optimal James-Stein estimator δ_{p-2} and its positive-part counterpart δ^+_{p-2} were inadmissible. The following theorem, whose proof is straightforward and deferred to the Appendix (Theorem 6.1), shows that the search for an admissible estimator reduces to finding a prior distribution with an analytically tractable Bayes estimator. Note that a proper prior measure is one with finite total mass (so by scaling, we may assume its total mass to be 1).

Theorem 3.5. *Suppose the distribution of X given θ is such that the risk functions $R(\theta, \delta)$ under quadratic loss are continuous in θ for all decision rules δ. Suppose further that the proper prior measure Π gives positive probability to any subset of $\mathbb{R}^p$ of positive Lebesgue measure. Then a Bayes rule with respect to Π is admissible.*

Remark: When $X|\theta \sim N_p(\theta, I)$, the risk functions are continuous in θ for all decision rules δ (for a proof, see Ferguson [22, pp. 139–140]).

Before we can find any Bayes estimators, we will need the following extension to the James-Stein class of minimax estimators, due to Baranchik [1] and Strawderman [31].

Lemma 3.6. *Estimators of the form*

$$\delta_r(X) = \left(1 - \frac{r(\|X\|^2)}{\|X\|^2}\right) X$$

are minimax for $0 \le r(\|X\|^2) \le 2(p-2)$, provided $r(\cdot)$ is increasing.

Proof. For any spherically symmetric estimator, i.e. one of the form $h(\|X\|^2)X$, the difference in risk between itself and X is given by

$$\mathbb{E}_\theta \|h(\|X\|^2)X - \theta\|^2 - \mathbb{E}_\theta \|X - \theta\|^2 = \mathbb{E}_\theta \left(\|X\|^2 h^2(\|X\|^2)\right) - 2\theta^T \mathbb{E}_\theta \left(h(\|X\|^2)X\right) + \|\theta\|^2 - p.$$

Consider the first term on the right-hand side. Using the same repesentation of a non-central chi-squared random variable as was mentioned in Section 3.2, we have

$$\mathbb{E}_\theta \left(\|X\|^2 h^2(\|X\|^2)\right) = e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\|\theta\|^2}{2}\right)^k}{k!} \mathbb{E}_\theta \left(\chi^2_{p+2k} h^2(\chi^2_{p+2k})\right). \tag{3}$$

To compute $2\theta^T \mathbb{E}_\theta \left(h(\|X\|^2)X\right)$, we make an orthogonal transformation such that $\theta = (\|\theta\|, 0, 0, \dots, 0)$. Note that because of spherical symmetry, this does not affect the expectations. A simple calculation (see the Appendix, Sublemma 6.2), shows that

$$2\theta^T \mathbb{E}_{\|\theta\|} \left(h(\|X\|^2)X\right) = 4e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \left(\frac{\|\theta\|^2}{2}\right)^k \frac{k}{k!} \mathbb{E}\big(h(\chi^2_{p+2k})\big), \tag{4}$$

where $\mathbb{E}_{\|\theta\|}$ denotes the expected value when $\theta = (\|\theta\|, 0, 0, \dots, 0)$. Finally, observe that

$$\|\theta\|^2 = \|\theta\|^2 e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\|\theta\|^2}{2}\right)^k}{k!} = e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \frac{2\left(\frac{\|\theta\|^2}{2}\right)^{k+1}}{k!} = e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \frac{2k\left(\frac{\|\theta\|^2}{2}\right)^k}{k!}, \tag{5}$$

since the $k = 0$ term in the final expression does not contribute to the sum.

Putting Equations (3),(4) and (5) together, we end up with

$$R\big(\theta, h(\|X\|^2)X\big) - R(\theta, X) = e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\|\theta\|^2}{2}\right)^k}{k!} \Big(\mathbb{E}\big(\chi^2_{p+2k} h^2(\chi^2_{p+2k})\big) - 4k\mathbb{E}\big(h(\chi^2_{p+2k})\big) - p + 2k\Big)$$

To show the above difference is negative, it suffices to show that for every integer k,

$$\mathbb{E}\big(\chi^2_{p+2k} h^2(\chi^2_{p+2k})\big) - 4k\mathbb{E}\big(h(\chi^2_{p+2k})\big) - p + 2k \le 0.$$

We are interested in estimators of the form

$$h(\|X\|^2) = 1 - \frac{r(\|X\|^2)}{\|X\|^2}$$

and with the stated assumptions on $r(\cdot)$, we can compute as follows:

$$\begin{aligned}
& \mathbb{E}\big(\chi^2_{p+2k} h^2(\chi^2_{p+2k})\big) - 4k\mathbb{E}\big(h(\chi^2_{p+2k})\big) - p + 2k \\
= \; & \mathbb{E}\left\{ \chi^2_{p+2k}\left(1 - \frac{r(\chi^2_{p+2k})}{\chi^2_{p+2k}}\right)^2 - 4k\left(1 - \frac{r(\chi^2_{p+2k})}{\chi^2_{p+2k}}\right) - p + 2k \right\} \\
= \; & \mathbb{E}\left\{ \frac{r^2(\chi^2_{p+2k})}{\chi^2_{p+2k}} - 2r(\chi^2_{p+2k}) + 4k\frac{r(\chi^2_{p+2k})}{(\chi^2_{p+2k})} \right\} \text{ using } \mathbb{E}(\chi^2_{p+2k}) = p + 2k \\
= \; & \mathbb{E}\left\{ r(\chi^2_{p+2k})\left(\frac{r(\chi^2_{p+2k})}{\chi^2_{p+2k}} + \frac{4k}{\chi^2_{p+2k}} - 2\right) \right\} \\
\leq \; & \mathbb{E}\left\{ r(\chi^2_{p+2k})\left(\frac{2(p-2)+4k}{\chi^2_{p+2k}} - 2\right) \right\} \text{ using } r(\cdot) \leq 2(p-2) \\
= \; & \text{Cov}\left(r(\chi^2_{p+2k}), \frac{2p-4+4k}{\chi^2_{p+2k}} - 2 \right) + \mathbb{E}\big(r(\chi^2_{p+2k})\big)\mathbb{E}\left(\frac{2p-4+4k}{\chi^2_{p+2k}} - 2\right) \\
\leq \; & 0,
\end{aligned}$$

since the covariance between an increasing function and a decreasing function is non-positive, and $\mathbb{E}\left(\frac{2p-4+4k}{\chi^2_{p+2k}} - 2\right) = 0$. This completes the proof of the lemma. □

Remark: This lemma shows that the class of minimax estimators is considerably richer for $p \geq 3$ than it is for $p = 1$ or $p = 2$, when X is the unique minimax estimator.

The challenge is now to find prior distributions on θ such that the resulting Bayes estimators satisfy the conditions of Lemma 3.6. Baranchik [1] claimed (as a result of private communication with Charles Stein) to have found at least one admissible estimator using such a technique, but we prefer to follow the approach of Strawderman [31].

We consider two-stage priors for θ, where at the first stage $\theta|\lambda \sim N_p(0, \frac{1-\lambda}{\lambda})$, and at the second stage, $\lambda \sim (1-b)\lambda^{-b}$ on $0 < \lambda \leq 1$. Note that the prior distribution on λ is proper if and only if $-\infty < b < 1$. The posterior distribution of θ given X and λ is $N_p((1-\lambda)X, 1-\lambda)$, and the Bayes estimator of θ is given by

$$\begin{aligned}
\mathbb{E}(\theta|X) &= \mathbb{E}\big(\mathbb{E}(\theta|X,\lambda) \bigm| X\big) \\
&= \mathbb{E}\big((1-\lambda)X \bigm| X\big) \\
&= \big(1 - \mathbb{E}(\lambda|X)\big)X. \qquad (6)
\end{aligned}$$

We can therefore prove the following theorem:

Theorem 3.7. *For $p \geq 5$, the Bayes estimator given in Equation (6) is minimax, provided $\frac{1}{2}(6-p) \leq b < 1$.*

Remark: The condition that $p \geq 5$ is necessary and sufficient to ensure that the marginal prior distribution on θ is proper.

Proof. By computing the joint posterior distribution of θ and λ given X, integrating out θ and then taking the expected value of the marginal posterior distribution of λ given X, we obtain

$$\mathbb{E}(\lambda|X) = \frac{1}{||X||^2}\left[p+2-2b-\frac{2}{\int_0^1 \lambda^{\frac{p}{2}-b}\exp(\frac{1-\lambda}{2}||X||^2)d\lambda}\right] \tag{7}$$

(For the details of this calculation, see the Appendix, Lemma 6.3). Defining $r(||X||^2)$ to be the term inside the square brackets on the right-hand side of Equation (7), we see that the Bayes estimator of Equation (6) is precisely of the form studied in Lemma 3.6. It therefore remains to show that $r(\cdot)$ is increasing and $0 \leq r(\cdot) \leq 2(p-2)$. The first requirement follows since $\int_0^1 \lambda^{\frac{p}{2}-b}\exp(\frac{1-\lambda}{2}||X||^2)d\lambda$ is increasing in $||X||^2$; to show the second, we note first that

$$r(0) = p+2-2b-\frac{2}{\int_0^1 \lambda^{\frac{p}{2}-b}d\lambda} = p+2-2b-2(\frac{p}{2}-b+1) = 0.$$

Thus the conditions of Lemma 3.6 will be satisfied provided $p+2-2b \leq 2(p-2)$, or equivalently

$$\frac{1}{2}(6-p) \leq b.$$

It is clear that for $p \geq 5$ we can choose b in the required range. This completes the proof of the theorem. □

The cases $p = 3$ and $p = 4$ remain. We will need the notion of a generalised Bayes estimator:

Definition 3.8. *If the random variable X has density $f(x;\theta)$ and Π is an improper prior measure, then a generalised Bayes estimator $\delta(X)$ with respect to the loss function (1) is one which minimises*

$$\int_{\mathbb{R}^p} ||\delta-\theta||^2 f(x;\theta)\, d\Pi(\theta).$$

Remark: In the case where $X|\theta \sim N_p(\theta, I)$, it is clear that the generalised Bayes estimator is given by

$$\delta(X) = \frac{\int_{\mathbb{R}^p} \theta e^{-||X-\theta||^2}\, d\Pi(\theta)}{\int_{\mathbb{R}^p} e^{-||X-\theta||^2}\, d\Pi(\theta)}$$

In other words, we simply treat the improper prior distribution Π as if it were proper, and compute the posterior mean as usual to find the generalised Bayes estimator.

Corollary 3.9. *For $p \geq 3$, the estimator (6) is generalised Bayes and minimax, provided $\frac{1}{2}(6-p) \leq b < \frac{1}{2}(p+2)$.*

Proof. The generalised Bayes estimator (6) exists provided the integral in Equation (7) exists, for which we require $\frac{p}{2} - b > -1$. This gives the right-hand inequality. The fact that (6) is minimax follows in the same way as in Theorem 3.7, and the range of minimax values for b is non-empty for $p \geq 3$. □

Unfortunately, generalised Bayes estimators need not be admissible, and the verification of admissibility can be difficult. However, in the multivariate spherically symmetric case, Brandwein and Cohen [17] gave a simple condition which ensures that certain generalised Bayes estimators are admissible. We state this condition without proof below.

Theorem 3.10. *Suppose $\delta(X)$ is a bounded risk generalised Bayes estimator of the form $\delta(X) = h(\|X\|^2)X$. If there exists an M such that $h(y) \leq 1 - \frac{p-2}{y}$ for all $y > M$, then $\delta(X)$ is admissible.*

Remark: Since the generalised Bayes estimators (6) are minimax, their risk is bounded above by p.

The work in the remainder of this subsection, although almost certainly known to Strawderman and others, has not been published to the author's knowledge. In order to find admissible generalised Bayes estimators for $p = 3$ and $p = 4$ of the given form, it remains to find a b in the range $\frac{1}{2}(6-p) \leq b < \frac{1}{2}(p+2)$ such that $r(y) \geq p-2$, where

$$r(y) = p + 2 - 2b - \frac{2}{\int_0^1 \lambda^{\frac{p}{2}-b} \exp\left(\frac{1-\lambda}{2}y\right) \, d\lambda}.$$

Equivalently, we must find b in the given range such that

$$4 - 2b - \frac{2}{\int_0^1 \lambda^{\frac{p}{2}-b} \exp\left(\frac{1-\lambda}{2}y\right) \, d\lambda} \geq 0 \tag{8}$$

We note that, by monotone convergence, for any fixed $b \in [\frac{1}{2}(6-p), \frac{1}{2}(p+2))$,

$$\int_0^1 \lambda^{\frac{p}{2}-b} e^{\frac{1-\lambda}{2}y} \, d\lambda \to \infty \text{ as } y \to \infty,$$

so any $b < 2$ will satisfy Equation (8).

By this discussion and Corallary 3.9, we have therefore proved the following theorem:

Theorem 3.11. *For $p = 3$, if $b \in [\frac{3}{2}, 2)$ then the generalised Bayes estimator (6) is admissible. For $p = 4$, if $b \in [1, 2)$, then (6) is admissible.*

4 Extensions to other distributions

We now give a discussion of the developments that have been made towards generalising Stein's ideas for the multivariate normal case to other location parameter family distributions. We concentrate on results which have shown new families of distributions to be amenable to shrinkage estimation, rather than those which have extended the range of parameter values within such a family. For such extensions, see, e.g. James and Stein [25], Berger[3] and Bock[8]. Also omitted in this discussion is the work on finding estimators which dominate X for other loss functions, such as the general quadratic loss

$$L(\theta, \delta) = (\delta - \theta)^T D(\delta - \theta) \text{ where } D \text{ is a given positive definite matrix,}$$

and non-decreasing, concave functions of quadratic loss. See e.g., Bhattacharya [5], Bock [7] and Brandwein and Strawderman[11, 13], for results of this nature. Instead, we merely note that in general it is found that shrinkage estimators are relatively robust to such changes in loss function.

As was observed in Section 2, Stein [29] pointed out that the normality assumption should not be essential to his argument. In fact, Brown [14] proved that the best equivariant estimator of a location parameter was inadmissible for almost arbitrary loss, including the case of quadratic loss (1), where X is the best equivariant estimator. In the light of this, Strawderman [32] considered 'variance mixtures' of multivariate normal distributions, that is, density functions of the form

$$f(||x-\theta||) = \int_{\mathbb{R}^p} \frac{1}{(2\pi\sigma^2)^{p/2}} e^{-\frac{1}{2\sigma^2}||x-\theta||^2}\, dG(\sigma), \tag{9}$$

where $G(\cdot)$ is a known cumulative distribution function (cdf) on $(0,\infty)$. The motivation for considering such distributions arises from the fact that a random variable X with such a density is, given σ, distributed as a p-variate normal with mean θ and covariance matrix $\sigma^2 I$, while the unconditioned distribution of σ is $G(\cdot)$. Two important special cases are:

i) The p-variate normal distribution $N_p(\theta, \sigma^2 I)$, which has $G(\cdot)$ degenerate at $\sigma = 1$

ii) The type I p-variate t-distribution, which has $\frac{1}{\sigma^2} \sim \frac{\chi^2_m}{m}$. This is the distribution of T, where $T_j = \frac{X_j}{S}$ for $j = 1, 2, \ldots, p$, with $X|\theta \sim N_p(\theta, I)$, and $S^2 \sim \frac{\chi^2_m}{m}$.

Strawderman proved the following theorem:

Theorem 4.1. *Let X be a single observation from the density (9). Then, for $p \geq 3$, estimators of the form*

$$\delta_r(X) = \left(1 - \frac{r(||X||^2)}{||X||^2}\right) X \tag{10}$$

are minimax provided:

i) $0 \leq r(\cdot) \leq \frac{2}{\mathbb{E}_0(||X||^{-2})}$

ii) $r(||X||^2)$ *is increasing in* $||X||^2$

iii) $\frac{r(||X||^2)}{||X||^2}$ *is decreasing in* $||X||^2$

iv) $\mathbb{E}_0(||X||^2) < \infty$

v) $\mathbb{E}_0(||X||^{-2}) < \infty$.

Strawderman's proof is very similar to Lemma 3.6 for the p-variate normal case and is omitted.

Remark: Conditions i) and ii) are the same as in Lemma 3.6; condition v) is relatively weak, and will be satisfied if the density f is bounded in a neighbourhood of the origin.

Berger [3] gives a slight extension to Strawderman's result. Special cases for which he finds classes of minimax estimators are

i) A double exponential density of the form

$$f(||x-\theta||) = \frac{e^{-||x-\theta||}}{a_p \Gamma(p)}$$

(where a_p is the surface area of the unit sphere); and

ii) The Cauchy-like density

$$f(\|x-\theta\|) = \frac{2\Gamma(a)}{(1+\|x-\theta\|^2)^a} a_p \frac{\Gamma(\frac{p}{2})}{\Gamma(a-\frac{p}{2})}.$$

He shows these densities are of the form given in Equation (9) by proving that if f is completely monotonic on $(0,\infty)$ (i.e. $(-1)^n \frac{d^n}{ds^n} f(s) \geq 0$ for all $n \in \mathbb{N}_0$), then it is of the required form.

The next big steps forward were made by Brandwein and Strawderman [10] and Brandwein [9]. By means of lengthy explicit calculations, they together found minimax estimators of similar form to those of Strawderman [32] for spherically symmetric unimodal (s.s.u.) distributions, and then Brandwein extended the work to all spherically symmetric (s.s.) distributions. Both the conditions and conclusions of the main theorem are very similar in the two cases, so we omit the statement and proof of the earlier result. However, the essence of the proof was first to find minimax estimators for p-dimensional spherical uniform distributions $X|\theta \sim U\{\|X-\theta\| \leq R\}$. They were then able to characterise s.s.u. distributions as 'R-mixtures' of spherical uniforms. That is, they showed that the density of an s.s.u. distribution about θ with respect to Lebesgue measure is of the form

$$f(\|x-\theta\|) = \int_0^\infty \frac{c}{R^p} \mathbb{I}_S(x,R)\, dF(R)$$

where $S = \{x : \|x-\theta\| \leq R\}$, $F(\cdot)$ is a cdf on $(0,\infty)$, and c is a positive constant. In this way, they could prove that their class of minimax estimators was valid for all s.s.u. distributions.

Brandwein made use of the similarly intuitive result proved in, e.g., Dempster [18, pp.271–2], that if X has a s.s. distribution about θ, then the conditional distribution of X given that $\|X-\theta\| = R$ is uniform over the set $\{x : \|x-\theta\| = R\}$. Thus, having found minimax estimators for the case where $X|\theta \sim \mathrm{U}\{\|X-\theta\| = R\}$, she was able to extend her result to all s.s. distributions. Her result can be stated as follows:

Theorem 4.2. *Let X have a p-dimensional spherically symmetric distribution about θ. Then, for $p \geq 4$, estimators δ_r of Equation (10) are minimax provided:*

i) $0 \leq r(\cdot) \leq 2\frac{p-2}{p}\frac{1}{\mathbb{E}_0(\|X\|^{-2})}$

ii) $r(\|X\|^2)$ *is increasing in* $\|X\|^2$

iii) $\frac{r(\|X\|^2}{\|X\|^2}$ *is decreasing in* $\|X\|^2$

iv) $\mathbb{E}_0(\|X\|^{-2}) < \infty$.

Remark: The earlier s.s.u. result was shown to hold with marginally improved upper bounds on $r(\cdot)$. Brandwein and Strawderman were able to replace the fraction $\frac{p-2}{p}$ by $\frac{p}{p+2}$ for $p \geq 4$, and also give a result for $p = 3$, with the same fraction replaced by 0.375.

Remark: This theorem was proved in a shorter and more elegant manner as a special case of a more general class of dominating estimators $\delta(X) = X + ag(X)$, given by Brandwein and Strawderman [13]. They applied the divergence theorem to the cross-product term in the expression for the risk of such an estimator (calculated conditionally on $\|X-\theta\| = R$) to generalise Stein's lemma

(Lemma 3.3). However, the hypotheses on g are slightly technical and are not discussed here. Cellier and Fourdrinier [16] were able to somewhat weaken these hypotheses.

There are two appealing features of Theorem 4.2. Firstly, it demonstrates the applicability of shrinkage estimators to all s.s. distributions generated by a one-dimensional cdf for $F(\cdot)$ for R on $(0,\infty)$ provided that

$$\infty > \mathbb{E}_0(\|X\|^{-2}) = \mathbb{E}_F\left(\mathbb{E}_0(\|X\|^{-2}|R)\right) = \mathbb{E}_F(R^{-2}).$$

Secondly, the theorem shows that shrinkage estimators are robust with respect to distributional assumptions. The upper bound $2\frac{p-2}{p}\frac{1}{\mathbb{E}_0(\|X\|^{-2})}$ of condition i) in Theorem 4.2 is the best which holds uniformly for all s.s. distributions. Although better upper bounds are possible for specific distributions, the improvement is usually not significant. Note in particular that if $r(\cdot)$ is constant, then for any s.s. distribution, the risk of δ_r at 0 is smaller than that of X if and only if

$$0 < \mathbb{E}_0\|X\|^2 - \mathbb{E}_0\left\|\left(1-\frac{r}{\|X\|^2}\right)X\right\|^2 = \mathbb{E}_0\left(-2r+\frac{r^2}{\|X\|^2}\right).$$

In other words, we must have $r < \frac{2}{\mathbb{E}_0(\|X\|^{-2})}$, which is only larger than the upper bound of condition i) by a factor of $\frac{p}{p-2}$.

Finally, we can now follow Brandwein [9] and give a discussion of the practically important situation of multiple observations. We start with a simple lemma characterising spherically symmetric estimators, the proof of which is the author's own.

Lemma 4.3. *An estimator $\delta(X)$ is spherically symmetric, i.e. is of the form $\delta(X) = h(\|X\|^2)X$ for some real-valued function h, if and only if it satisfies*

$$\delta(XP) = \delta(X)P \textit{ for all } p\times p \textit{ orthogonal matrices } P.$$

Proof. If δ is spherically symmetric, then for any $p \times p$ orthogonal matrix P,

$$\delta(XP) = h(\|XP\|^2)XP = h(\|X\|^2)XP = \delta(X)P.$$

Conversely, if the condition holds, then let $\delta(X) = h(X)X$, for some $p\times p$ matrix $h(X)$, with entries h_{ij}, say. We have

$$h(X)XP = \delta(X)P = \delta(XP) = h(XP)XP \text{ for all orthogonal } P.$$

Thus $h(X) = h(XP)$ for all orthogonal P.

Suppose $\|X\| = \|Y\|$, so that there exists an orthogonal matrix Q such that $Y = XQ$. Then

$$h(X) = h(XQ) = h(Y).$$

Thus $h \equiv h(\|X\|^2)$.

Finally, we must show that h is a scalar multiple of the identity matrix, so that in fact $\delta(X) = h(\|X\|^2)X$, for some real-valued function h. To this end, let P be the orthogonal matrix given by

$$P = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -1 & 0 & 0 & \dots & 0 \end{pmatrix},$$

so that

$$\begin{aligned} P^T h P &= \begin{pmatrix} 0 & 0 & \dots & 0 & -1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} h_{11} & h_{12} & h_{13} & \dots & h_{1p} \\ h_{21} & h_{22} & h_{23} & \dots & h_{2p} \\ h_{31} & h_{32} & h_{33} & \dots & h_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{p1} & h_{p2} & h_{p3} & \dots & h_{pp} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -1 & 0 & 0 & \dots & 0 \end{pmatrix} \\ &= \begin{pmatrix} h_{pp} & -h_{p1} & -h_{p2} & \dots & -h_{p,p-1} \\ -h_{1p} & h_{11} & h_{12} & \dots & h_{1,p-1} \\ -h_{2p} & h_{21} & h_{22} & \dots & h_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -h_{p-1,p} & h_{p-1,1} & h_{p-1,2} & \dots & h_{p-1,p-1} \end{pmatrix}. \end{aligned}$$

So if $P^T hP = h$, then

$$\begin{aligned} h_{pp} &= h_{11} = h_{22} = \dots = h_{p-1,p-1} \\ -h_{p1} &= h_{12} = h_{23} = \dots = h_{p-1,p} = h_{p1} \\ -h_{p2} &= h_{13} = h_{24} = \dots = h_{p-1,1} = h_{p2} \\ &\vdots \\ -h_{p,p-1} &= h_{1p} = h_{21} = \dots = h_{p,p-1}. \end{aligned}$$

The top string of equalities shows that the diagonal entries of h are equal, while the the lower lines show that every off-diagonal entry is zero. Hence h is a scalar multiple of the identity matrix, as required. □

Similarly, a random variable X whose cdf is absolutely continuous with respect to Lebesgue measure is spherically symmetric about θ, i.e. its density is a function of $\|x - \theta\|$, if and only if $(X - \theta)P$ has the same distribution as $(X - \theta)$ for every $p \times p$ orthogonal matrix P.

Now suppose that we have n observations $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ from a s.s. distribution about θ. The best equivariant estimator with respect to quadratic loss (1) in such a situation is Pitman's estimator, given by

$$\delta(X^{(1)}, X^{(2)}, \dots, X^{(n)}) = X^{(1)} - \mathbb{E}_0(X^{(1)} | X^{(2)} - X^{(1)}, X^{(3)} - X^{(1)}, \dots, X^{(n)} - X^{(1)}).$$

Often, this is analytically intractable, and the sample mean $\overline{X}$, or maximum likelihood estimator (MLE) may be preferred. All these estimators are spherically symmetric location equivariant estimators (provided for the MLE case, it is unique) . That is, the estimators satisfy

$$\delta(X^{(1)}P + a, X^{(2)}P + a, \dots, X^{(n)}P + a) = \delta(X^{(1)}, X^{(2)}, \dots, X^{(n)})P + a,$$

for all $p \times p$ orthogonal matrices P, and $a \in \mathbb{R}^p$. We can now prove the following theorem:

Theorem 4.4. *Suppose $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ are i.i.d. from a p-dimensional spherically symmetric distribution about θ, and that $\delta(X^{(1)}, X^{(2)}, \dots, X^{(n)})$ is a spherically symmetric location equivariant estimator. Then $\delta(X^{(1)}, X^{(2)}, \dots, X^{(n)})$ also has a spherically symmetric distribution about θ.*

Proof. For any (measurable) set S, we have

$$\begin{aligned}
\mathbb{P}\Big((\delta(X^{(1)}, X^{(2)}, \dots, X^{(n)}) - \theta)P \in S\Big) &= \mathbb{P}\Big(\delta((X^{(1)} - \theta)P, (X^{(2)} - \theta)P, \dots, (X^{(n)} - \theta)P) \in S\Big) \\
&\quad \text{because } \delta \text{ is spherically symmetric} \\
&= \mathbb{P}\Big(\delta((X^{(1)} - \theta), (X^{(2)} - \theta), \dots, (X^{(n)} - \theta)) \in S\Big) \\
&\quad \text{since } (X_i - \theta)P \text{ has the same distribution as } (X_i - \theta) \\
&= \mathbb{P}\Big((\delta(X^{(1)}, X^{(2)}, \dots, X^{(n)}) - \theta) \in S\Big) \\
&\quad \text{by location equivariance of } \delta.
\end{aligned}$$

Hence $\delta(X^{(1)}, X^{(2)}, \dots, X^{(n)})$ has a spherically symmetric distribution about θ. □

This means that, provided $p \geq 4$, we can apply Brandwein's result in Theorem 4.2 directly to any one of the common estimators mentioned above to obtain new ones which dominate them in terms of risk.

5 The paper of Evans and Stark

All the generalisations of Stein's proof of the inadmissibility of the best equivariant estimator in the normal case have thus far been confined to spherically symmetric distributions. Evans and Stark [21] have recently proved that inadmissibility for $p \geq 3$ holds for a far wider class of distributions, namely those location parameter families which have finite first and second moments, plus one further condition which is discussed below. Their proof is rather different in flavour from the results discussed in Section 4, relying more on probability theory and properties of Brownian motion. We do not delve too deeply into the probabilistic foundations here (see, for example, Rost [27] for relevant results), but as motivation do discuss a mathematical connection between admissibility and Brownian motion noticed by Brown [15].

The fact that the best equivariant estimator is admissible for $p \leq 2$ and inadmissible for $p \geq 3$ is reminiscent of a dimension-dependent property of Brownian motion, namely that it is (neighbourhood) recurrent for $p \leq 2$ and transient for $p \geq 3$. Brown establishes, that to each estimator δ, there corresponds a diffusion on p-dimensional space, and that the estimator is admissible if and only if the corresponding diffusion is recurrent. Furthermore, he shows that the diffusion related to the best equivariant estimator is (essentially) Brownian motion.

The theorem which Evans and Stark prove is divided into three parts. In the first part, they show that, subject to the conditions mentioned above, estimators of the form

$$\delta_a^{ES}(X) = \left(1 - \frac{a}{1 + \|X\|^2}\right) X \tag{11}$$

dominate X for $a > 0$ sufficiently small. Although powerful, this part alone gives no indication of how small a must be to lead to an improved estimator. So in the second and third parts, Evans and Stark apply the first part of the theorem to give a range of values of a for which estimators of the form (11) dominate X in two special cases: firstly, where the distribution of X is contained in some ball around θ, and secondly where it does not intersect such a ball.

The theorem and proof as stated here differ from the original in two major respects. Firstly, we state and prove the theorem for a location parameter family rather than a single random variable in order to be consistent with what has gone before; and secondly, we fill in many of the details omitted by Evans and Stark in their rather condensed style, in the hope that this will aid clarity.

Theorem 5.1. *Let X be an observation from a location parameter family indexed by $\theta \in \mathbb{R}^p \, (p \geq 3)$ such that*

i) X is not almost surely equal to θ

ii) $\mathbb{E}_\theta(X) = \theta$ for all $\theta \in \mathbb{R}^p$

iii) $\mathbb{E}_\theta \|X - \theta\|^2 < \infty$ for all $\theta \in \mathbb{R}^p$

iv) $\mathbb{E}_\theta \|X\|^{2-p} \leq \|\theta\|^{2-p}$ for all $\theta \in \mathbb{R}^p$.

Then

1) Estimators of the form

$$\delta_a^{ES}(X) = \left(1 - \frac{a}{1 + \|X\|^2}\right) X$$

dominate X for $a > 0$ sufficiently small.

2) If the support of the distribution of X given θ is contained in the ball $\{x \in \mathbb{R}^p : \|x - \theta\| \leq A\}$, then any

$$a \in \left(0, 2\frac{p-2}{p}\left(\frac{\alpha^*}{2+\alpha^*}\right)^6 \mathbb{E}_\theta \|X - \theta\|^2\right)$$

suffices for $\delta_a^{ES}(X)$ to dominate X, where α^* is the unique positive root of

$$(p-2)\alpha^6 A^2 \left(1 + (2+\alpha)^2 A^2\right)^2 - p(2+\alpha)^4 = 0.$$

3) If the support of the distribution of X given θ does not intersect the ball $\{x \in \mathbb{R}^p : \|x - \theta\| \leq A\}$, then any

$$a \in \left(0, 2\frac{p-2}{p}A^2\right)$$

suffices for $\delta_a^{ES}(X)$ to dominate X.

Remark: Condition iv) may appear surprising, but as Evans and Stark state, Rost [27] demonstrates in his solution of Skorokhod's problem for transient Markov processes, this is precisely the condition required on X for there to exist a stopping time T such that the distribution of B_T is that of $X - \theta$. Here, $B \equiv (B_t)_{t \geq 0}$ is a standard Brownian motion starting at $0 \in \mathbb{R}^p$, and B_T is the stopped process, defined by $B_T(\omega) = B_{T(\omega)}(\omega)$. Note that since we may assume $\mathbb{E}_0(X) = 0$, we must have $B_T < \infty$ a.s., and hence $T < \infty$ a.s..

Proof. 1) We have

$$\begin{aligned} R(\theta, X) - R\big(\theta, \delta_a^{ES}(X)\big) &= \mathbb{E}_\theta \|X-\theta\|^2 - \mathbb{E}_\theta \left\| \left(1 - \frac{a}{1+\|X\|^2}\right) X - \theta \right\|^2 \\ &= 2a\mathbb{E}_\theta \left(\frac{X^T(X-\theta)}{1+\|X\|^2}\right) - a^2 \mathbb{E}_\theta \left(\frac{\|X\|^2}{(1+\|X\|^2)^2}\right). \end{aligned}$$

Since $a > 0$, the result will follow provided

$$a\mathbb{E}_\theta \left(\frac{\|X\|^2}{(1+\|X\|^2)^2}\right) < 2\mathbb{E}_\theta \left(\frac{X^T(X-\theta)}{1+\|X\|^2}\right) \text{ for all } a \in \mathbb{R}^p.$$

Now, by condition iv) and the remark following it, let T be a stopping time such that the distribution of B_T is the same as that of $X - \theta$. The condition on a becomes

$$a\mathbb{E}\left(\frac{\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2}\right) < 2\mathbb{E}\left(\frac{B_T.(B_T+\theta)}{1+\|B_T+\theta\|^2}\right) \text{ for all } a \in \mathbb{R}^p. \tag{12}$$

Our first aim is find a lower bound for the right-hand side of (12) using a generalisation of Stein's lemma (Lemma 3.3). We would like to use Girsanov's formula, but in order to do this directly we would need both T and B_T to be bounded. We therefore let $S_n = \inf\{s \geq 0 : \|B_s\| = n\}$, and define $T_n = T \wedge S_n \wedge n$, in the hope that our result for T_n will still hold when we let n tend to infinity. Girsanov's formula applied to T_n states that, for any $y \in \mathbb{R}^p$, $\epsilon \in \mathbb{R}$ and any bounded (measurable) function $F : \mathbb{R}^p \to \mathbb{R}$, we have

$$\mathbb{E}\big(\exp(\epsilon y.B_{T_n} - \frac{1}{2}\epsilon^2\|y\|^2 T_n)F(B_{T_n})\big) = \mathbb{E}\big(F(B_{T_n} + \epsilon T_n y)\big).$$

If F, in addition to being bounded, is continuous, with bounded and continuous first partial derivatives, we can differentiate both sides of this equation with respect to ϵ, inside the expectation, and evaluate the derivatives at $\epsilon = 0$. This gives

$$\mathbb{E}\big(y.B_{T_n} F(B_{T_n})\big) = \mathbb{E}\big(T_n y.\nabla F(B_{T_n})\big). \tag{13}$$

This is the generalisation of Stein's lemma which we were seeking. Now let

$$F(B_{T_n}) = \frac{(B_{T_n})_i + \theta_i}{1+\|B_{T_n}+\theta\|^2},$$

and let y be the ith coordinate vector, where $1 \leq i \leq p$. Note that $|F(B_{T_n})| \leq 1$, and F is continuous. Furthermore, we have

$$\frac{\partial F}{\partial (B_{T_n})_i} = \frac{1+\|B_{T_n}+\theta\|^2 - 2\big((B_{T_n})_i+\theta_i\big)^2}{(1+\|B_{T_n}+\theta\|^2)^2} \text{ so that } \left|\frac{\partial F}{\partial (B_{T_n})_i}\right| \leq 1,$$

and

$$\frac{\partial F}{\partial (B_{T_n})_j} = \frac{-2\big((B_{T_n})_i+\theta_i\big)\big((B_{T_n})_j+\theta_j\big)}{(1+\|B_{T_n}+\theta\|^2)^2} \text{ for } j \neq i, \text{ so that } \left|\frac{\partial F}{\partial (B_{T_n})_j}\right| \leq 2.$$

As all these partial derivatives are also continuous, we may apply Equation (13) to obtain

$$\mathbb{E}\left(\frac{(B_{T_n})_i\big((B_{T_n})_i+\theta_i\big)}{1+\|B_{T_n}+\theta\|^2}\right) = \mathbb{E}\left(T_n \frac{1+\|B_{T_n}+\theta\|^2 - 2\big((B_{T_n})_i+\theta_i\big)^2}{(1+\|B_{T_n}+\theta\|^2)^2}\right).$$

Summing these equations over $i = 1, 2, \dots, p$ gives

$$\mathbb{E}\left(\frac{B_{T_n}.(B_{T_n}+\theta)}{1+||B_{T_n}+\theta||^2}\right) = \mathbb{E}\left(T_n\frac{p+(p-2)||B_{T_n}+\theta||^2}{(1+||B_{T_n}+\theta||^2)^2}\right).$$

Now, applying the bounded convergence theorem (which is valid since the sequence of random variables inside the expectation is uniformly bounded above by the constant 1), we have

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{B_{T_n}.(B_{T_n}+\theta)}{1+||B_{T_n}+\theta||^2}\right) = \mathbb{E}\left(\frac{B_T.(B_T+\theta)}{1+||B_T+\theta||^2}\right).$$

We are therefore able to use Fatou's lemma to obtain

$$\begin{aligned}\mathbb{E}\left(\frac{B_T.(B_T+\theta)}{1+||B_T+\theta||^2}\right) &= \liminf_{n\to\infty}\mathbb{E}\left(\frac{B_{T_n}.(B_{T_n}+\theta)}{1+||B_{T_n}+\theta||^2}\right)\\ &= \liminf_{n\to\infty}\mathbb{E}\left(T_n\frac{p+(p-2)||B_{T_n}+\theta||^2}{(1+||B_{T_n}+\theta||^2)^2}\right)\\ &\geq \mathbb{E}\left(T\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right).\end{aligned}$$

We now have a lower bound for the right-hand side of (12), and the proof of part 1) will be complete if we can find an $a > 0$ such that

$$a\mathbb{E}\left(\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right) < 2\mathbb{E}\left(T\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right).$$

To prove the existence of such an a, it suffices to show that:

I) $\mathbb{E}\left(\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)$ is bounded above on compact sets in $\mathbb{R}^p$

II) $\mathbb{E}\left(T\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)$ is bounded below, away from zero, on compact sets in $\mathbb{R}^p$

III) $\liminf_{||\theta||\to\infty}\left\{\frac{||\theta||^2 2\mathbb{E}\left(T\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)}{||\theta||^2\mathbb{E}\left(\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)}\right\} > 0.$

We prove I, II and III in order. Let $\theta \in \mathbb{R}^p$, and take a sequence $(\theta_n)_{n\geq 1} \in \mathbb{R}^p$ such that $\theta_n \to \theta$ as $n \to \infty$. Let

$$Y_n = \frac{||B_T+\theta_n||^2}{(1+||B_T+\theta_n||^2)^2}$$

and

$$Y = \frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}.$$

Then $|Y_n| \leq 1$ for each integer n, and $Y_n \to Y$ almost surely, as $n \to \infty$. So by the bounded convergence theorem, $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$ as $n \to \infty$. In other words, the map

$$\theta \mapsto \mathbb{E}\left(\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)$$

is continuous. This proves I.

Suppose again that we have a sequence $(\theta_n)_{n\geq 1}$ such that $\theta_n \to \theta$ as $n \to \infty$. By Fatou's lemma,

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}\left(T\,\frac{p+(p-2)||B_T+\theta_n||^2}{(1+||B_T+\theta_n||^2)^2}\right) &\geq \mathbb{E}\left\{\liminf_{n\to\infty}\left(T\,\frac{p+(p-2)||B_T+\theta_n||^2}{(1+||B_T+\theta_n||^2)^2}\right)\right\} \\ &= \mathbb{E}\left(T\,\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right).\end{aligned}$$

Thus the map

$$\theta \mapsto \mathbb{E}\left(T\,\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)$$

is lower semi-continuous. Moreover, this function is strictly positive at each point. For,

$$\mathbb{E}\left(T\,\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right) \geq (p-2)\mathbb{E}\left(\frac{T}{(1+||B_T+\theta||^2)}\right),$$

and since B_T is not almost surely 0 (by condition i)), it follows that T is not almost surely 0.

Hence $\mathbb{E}\left(T\,\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)$ is bounded below, away from zero, on compact sets in $\mathbb{R}^p$. This proves II.

We show III by demonstrating that the lim inf of the numerator is strictly positive, and that the denominator tends to a finite positive limit.

Applying Fatou's lemma to the numerator gives

$$\begin{aligned}\liminf_{||\theta||\to\infty} ||\theta||^2 2\mathbb{E}\left(T\,\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right) &\geq 2\mathbb{E}\left(\liminf_{||\theta||\to\infty} ||\theta||^2 T\,\frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right) \\ &\geq 2(p-2)\mathbb{E}(T) \\ &> 0.\end{aligned}$$

Applying the bounded convergence theorem to the denominator gives

$$\lim_{||\theta||\to\infty} ||\theta||^2 \mathbb{E}\left(\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right) = \mathbb{E}\left(\lim_{||\theta||\to\infty} ||\theta||^2\,\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right) = 1.$$

This proves III and hence completes the proof of part 1) of the theorem.

2) When the distribution of X given θ is contained in the ball $\{x \in \mathbb{R}^p : ||x-\theta|| \leq A\}$, we find an explicit range of values for a for which the result holds by first finding a lower bound for

$$\frac{\inf_{||x-\theta||\leq A}\left(\frac{p+(p-2)||x+\theta||^2}{(1+||x+\theta||^2)^2}\right)}{\sup_{||x-\theta||\leq A}\left(\frac{||x+\theta||^2}{(1+||x+\theta||^2)^2}\right)}$$

which holds for all $\theta \in \mathbb{R}^p$. We do this by choosing an arbitrary $\alpha > 0$, and considering separately the cases $||\theta|| \geq (1+\alpha)A$ and $||\theta|| < (1+\alpha)A$. We can then choose α to maximise this range of values for a. Next we find a lower bound for $\mathbb{E}(T)$, and then put these together to give the required result.

Thus, fix $\alpha > 0$. For $||\theta|| \geq (1+\alpha)A$ and $||x-\theta|| \leq A$, we have

$$||x|| \leq ||x-\theta|| + ||\theta|| \leq A + ||\theta|| \leq \left(\frac{1}{1+\alpha}+1\right)||\theta|| = \frac{2+\alpha}{1+\alpha}||\theta||.$$

Similarly,

$$||x|| \geq ||\theta|| - ||\theta - x|| \geq ||\theta|| - A \geq \left(1-\frac{1}{1+\alpha}\right)||\theta|| = \frac{\alpha}{1+\alpha}||\theta||.$$

Thus, for $||\theta|| \geq (1+\alpha)A$, we have

$$\begin{aligned}
\inf_{||x-\theta||\leq A}\left(\frac{p+(p-2)||x||^2}{(1+||x||^2)^2}\right) &\geq \frac{(p-2)\left(\tfrac{\alpha}{1+\alpha}\right)^2||\theta||^2}{\left(1+\left(\tfrac{2+\alpha}{1+\alpha}\right)^2||\theta||^2\right)^2} \\
&= (p-2)\left(\frac{\alpha}{2+\alpha}\right)^6 \frac{\left(\tfrac{2+\alpha}{1+\alpha}\right)^2||\theta||^2}{\left(\left(\tfrac{\alpha}{2+\alpha}\right)^2+\left(\tfrac{\alpha}{1+\alpha}\right)^2||\theta||^2\right)^2} \\
&\geq (p-2)\left(\frac{\alpha}{2+\alpha}\right)^6 \frac{\left(\tfrac{2+\alpha}{1+\alpha}\right)^2||\theta||^2}{\left(1+\left(\tfrac{\alpha}{1+\alpha}\right)^2||\theta||^2\right)^2} \\
&\geq (p-2)\left(\frac{\alpha}{2+\alpha}\right)^6 \sup_{||x-\theta||\leq A}\frac{||x||^2}{(1+||x||^2)^2}. \qquad (14)
\end{aligned}$$

Now for $||\theta|| < (1+\alpha)A$ and $||x-\theta|| \leq A$, we have

$$0 \leq ||x|| \leq ||x-\theta|| + ||\theta|| \leq A + ||\theta|| \leq (2+\alpha)||\theta||.$$

Thus, for $||\theta|| < (1+\alpha)A$,

$$\begin{aligned}
\inf_{||x-\theta||\leq A}\left(\frac{p+(p-2)||x||^2}{(1+||x||^2)^2}\right) &\geq \frac{p}{\left(1+(2+\alpha)^2A^2\right)^2} \\
&\geq \frac{p}{(2+\alpha)^2A^2\left(1+(2+\alpha)^2A^2\right)^2}(2+\alpha)^2A^2 \\
&\geq \frac{p}{(2+\alpha)^2A^2\left(1+(2+\alpha)^2A^2\right)^2}\sup_{||x-\theta||\leq A}\frac{||x||^2}{(1+||x||^2)^2}. \qquad (15)
\end{aligned}$$

Putting Equations (14) and (15) together, we see that, for all $\theta \in \mathbb{R}^p$,

$$\inf_{||x-\theta||\leq A}\left(\frac{p+(p-2)||x||^2}{(1+||x||^2)^2}\right) \geq \left((p-2)\left(\frac{\alpha}{2+\alpha}\right)^6 \wedge \frac{p}{(2+\alpha)^2A^2\left(1+(2+\alpha)^2A^2\right)^2}\right)\sup_{||x-\theta||\leq A}\frac{||x||^2}{(1+||x||^2)^2}.$$

Note that since the function

$$f_1(\alpha) = (p-2)\left(\frac{\alpha}{2+\alpha}\right)^6$$

increases monotonically from 0 to 1 as α ranges from 0 to ∞, and the function

$$f_2(\alpha) = \frac{p}{(2+\alpha)^2 A^2 \left(1 + (2+\alpha)^2 A^2\right)^2}$$

decreases monotonically from $\frac{p}{4A^2(1+4A^2)^2}$ to 0 as α ranges from 0 to ∞, the maximum value of the function

$$\alpha \mapsto \left((p-2)\left(\frac{\alpha}{2+\alpha}\right)^6 \wedge \frac{p}{(2+\alpha)^2 A^2 \left(1 + (2+\alpha)^2 A^2\right)^2}\right)$$

occurs at $\alpha = \alpha^*$, say, where α^* is the unique positive solution to the equation $f_1(\alpha) = f_2(\alpha)$. This is equivalent to saying α^* solves

$$(p-2)\alpha^6 A^2 \left(1 + (2+\alpha)^2 A^2\right)^2 - p(2+\alpha^4) = 0.$$

The corresponding value at the maximum is $(p-2)\left(\frac{\alpha^*}{2+\alpha^*}\right)^6$.

To compute a lower bound for $\mathbb{E}(T)$, we use the fact that $(||B_t||^2 - tp)_{t \geq 0}$ is a martingale. Recall that $T_n = T \wedge S_n \wedge n$, where $S_n = \inf\{t \geq 0 : ||B_t|| = n\}$. Hence

$$\begin{aligned}
\mathbb{E}(T) &= \lim_{n\to\infty} \mathbb{E}(T_n) \text{ by monotone convergence} \\
&= \frac{1}{p} \lim_{n\to\infty} \mathbb{E}(||B_{T_n}||^2) \text{ by the optional stopping theorem} \\
&= \frac{1}{p} \liminf_{n\to\infty} \mathbb{E}(||B_{T_n}||^2) \\
&\geq \frac{1}{p} \mathbb{E}(||B_T||^2) \text{ by Fatou's lemma} \\
&= \frac{1}{p} \mathbb{E}_\theta(||X - \theta||^2).
\end{aligned}$$

Thus,

$$\frac{2\mathbb{E}\left(T \frac{p+(p-2)||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)}{\mathbb{E}\left(\frac{||B_T+\theta||^2}{(1+||B_T+\theta||^2)^2}\right)} \geq 2\mathbb{E}(T) \frac{\inf_{||x-\theta|| \leq A}\left(\frac{p+(p-2)||x+\theta||^2}{(1+||x+\theta||^2)^2}\right)}{\sup_{||x-\theta|| \leq A}\left(\frac{||x+\theta||^2}{(1+||x+\theta||^2)^2}\right)} \geq 2\frac{p-2}{p}\left(\frac{\alpha^*}{2+\alpha^*}\right)^6 \mathbb{E}_\theta(||X - \theta||^2).$$

This completes the proof of part 2) of the theorem.

3) For this case, Evans and Stark exploit a result from a paper by Fitzsimmons [23]. According to this work, they state that not only may we assume that there exists a stopping time T such that the

distribution of B_T is the same as that of $X - \theta$, but also that this stopping time is of a particular form. This will enable us to obtain an explicit expression for $\mathbb{E}(T|B_T)$ (which is in fact a constant), and hence when we compute

$$\mathbb{E}\left(T \frac{p + (p-2)\|B_T + \theta\|^2}{(1 + \|B_T + \theta\|^2)^2} \,\middle|\, B_T \right),$$

this constant can be taken outside the expectation. In order to describe the form of the stopping time, we make the following definitions:

Definition 5.2. *A finely open set is one which a Brownian motion takes a strictly positive time to exit, almost surely, when started at any point in the set. In other words,*

$$\mathbb{P}^x(\tau_A > 0) = 1 \textit{ for all } x \in A,$$

where $\tau_A = \inf\{t \geq 0 : B_t \notin A\}$ *is the first exit time from* A.

Definition 5.3. *A finely closed set is the complement of a finely open set.*

We may suppose that the stopping time $T \equiv T_{C(U)}$, where $\{C(u) : 0 \leq u \leq 1\}$ is a decreasing family of finely closed sets such that $C(u) \subset \{x \in \mathbb{R}^p : \|x - \theta\| \geq A\}$. Here, $U \sim U[0,1]$, independently of B, and $T_{C(u)}$ is the first entry time of $C(u)$, i.e. $T_{C(u)} = \inf\{t \geq 0 : B_t \in C(u)\}$. The important point is that each $C(u)$ does not intersect the ball of radius A around θ.

Now, let $S_{A,n} = S_A \wedge n \equiv (\inf\{t \geq 0 : \|B_t\| = A\}) \wedge n$.

Then

$$\begin{aligned}
\mathbb{E}(T|B_T) &\geq \mathbb{E}(S_A|B_T) \text{ since } B \text{ hits level } A \text{ before any } \{C(u) : 0 \leq u \leq 1\} \\
&= \mathbb{E}(S_A) \text{ by the Strong Markov property and rotational invariance of } B \\
&= \mathbb{E}(\lim_{n\to\infty} S_{A,n}) \\
&= \lim_{n\to\infty} \mathbb{E}(S_{A,n}) \text{ by bounded convergence} \\
&= \frac{A^2}{p} \text{ by the optional stopping theorem applied to the martingale } \{\|B_t\|^2 - tp\}_{t\geq 0}.
\end{aligned}$$

We can now compute the following:

$$\begin{aligned}
\frac{2\mathbb{E}\left(T \frac{p+(p-2)\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2}\right)}{\mathbb{E}\left(\frac{\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2}\right)} &= \frac{2\mathbb{E}\left\{\mathbb{E}\left(T \frac{p+(p-2)\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2} \,\middle|\, B_T\right)\right\}}{\mathbb{E}\left(\frac{\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2}\right)} \\
&\geq 2\frac{A^2}{p} \frac{\mathbb{E}\left(\frac{p+(p-2)\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2}\right)}{\mathbb{E}\left(\frac{\|B_T+\theta\|^2}{(1+\|B_T+\theta\|^2)^2}\right)} \\
&\geq 2\frac{p-2}{p}A^2.
\end{aligned}$$

This completes the proof of the third and final part of the theorem. □

Remark: It is natural to try to ascertain the restrictiveness of condtion iv) of the theorem, which stated that the distribution of X given θ must satisfy

$$\mathbb{E}_\theta ||X||^{2-p} \leq ||\theta||^{2-p}.$$

Bass [2] (see equation II.3.6) shows, in his discussion of potential theory, that if the distribution of X is spherically symmetric about θ, and σ_r is normalised surface measure on the sphere of radius r in $\mathbb{R}^p$, then

$$\mathbb{E}_\theta ||X||^{2-p} = \int ||x+\theta||^{2-p}\, \sigma_r(dx) = r^{2-p} \wedge ||\theta||^{2-p} \leq ||\theta||^{2-p}.$$

Thus the theorem can be applied to all spherically symmetric distributions. In general, Evans and Stark state that it follows from Fitzsimmons [23] that condition iv) of the theorem holds if and only if $X-\theta$ has the exit distribution of a standard Brownian motion from a finely open domain containing θ, or is a mixture of such distributions. As an extreme example to demonstrate the applicability of the theorem, Evans and Stark point out that this class contains singular distributions supported on fractal sets of non-integral dimension. Note, more importantly and intuitively, that spherically symmetric distributions arise only when the domain is a ball centred at θ.

In conclusion, we have seen that shrinkage estimators dominate X in great generality, and can be of considerable practical importance in certain situations. To the author's knowledge, the problem of finding admissible estimators for any distribution other than a multivariate normal distribution remains an open question.

6 Appendix

Theorem 6.1. *Suppose the distribution of X given θ is such that the risk functions $R(\theta,\delta)$ under quadratic loss are continuous in θ for all decision rules δ. Suppose further that the proper prior measure Π gives positive probability to any subset of $\mathbb{R}^p$ of positive Lebesgue measure. Then a Bayes rule with respect to Π is admissible.*

Proof. Let δ be a Bayes rule with respect to Π, and suppose for a contradiction that δ' dominates δ, so that

$$\begin{aligned} R(\theta,\delta') &\leq R(\theta,\delta) \text{ for all } \theta \in \mathbb{R}^p \\ R(\theta_0,\delta') &< R(\theta_0,\delta) \text{ for some } \theta_0 \in \mathbb{R}^p. \end{aligned}$$

Let $\epsilon = R(\theta_0,\delta) - R(\theta_0,\delta')$. By continuity of the risk functions, there exists $\eta > 0$ such that

$$R(\theta,\delta) - R(\theta,\delta') > \frac{\epsilon}{2} \text{ for all } ||\theta-\theta_0|| < \eta.$$

Thus, denoting the Bayes risk of δ with respect to Π by $r(\Pi,\delta)$, and writing
$A = \{\theta \in \mathbb{R}^p : ||\theta - \theta_0|| < \eta\}$, we have

$$\begin{aligned} r(\Pi,\delta) - r(\Pi,\delta') &= \int_{\mathbb{R}^p} \left(R(\theta,\delta) - R(\theta,\delta')\right) d\Pi(\theta) \\ &> \frac{\epsilon}{2}\Pi(A) \\ &> 0. \end{aligned}$$

This contradicts the fact that δ is a Bayes rule with respect to Π. Hence δ is admissible. □

Sublemma 6.2. *Suppose $X|\theta \sim N_p(\theta, I)$, and $h(\|X\|^2)X$ is a spherically symmetric estimator of θ. Then*

$$2\theta^T \mathbb{E}_{\|\theta\|}\left(h(\|X\|^2)X\right) = 4\sum_{k=0}^{\infty} e^{-\frac{\|\theta\|^2}{2}} \left(\frac{\|\theta\|^2}{2}\right)^k \frac{k}{k!}\mathbb{E}\left(h\left(\chi^2_{p+2k}\right)\right),$$

where $\mathbb{E}_{\|\theta\|}$ denotes the expected value when $\theta = (\|\theta\|, 0, 0, \ldots, 0)$.

Proof.

$$\begin{aligned}
& 2\theta^T \mathbb{E}_{\|\theta\|}\left(h(\|X\|^2)X\right) \\
= \; & \frac{2\|\theta\|}{(2\pi)^{\frac{p}{2}}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h\left(\sum_{i=1}^{p} x_i^2\right) x_1 \exp\left(-\frac{1}{2}(x_1 - \|\theta\|)^2 - \frac{1}{2}\sum_{i=2}^{p} x_i^2\right) dx_1\, dx_2 \ldots dx_p \\
= \; & \frac{2\|\theta\|}{(2\pi)^{\frac{p}{2}}} e^{-\frac{\|\theta\|^2}{2}} \frac{d}{d\|\theta\|} \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h\left(\sum_{i=1}^{p} x_i^2\right) \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{p} x_i^2 - 2\|\theta\| x_1\right)\right) dx_1\, dx_2 \ldots dx_p \right\} \\
= \; & \frac{2\|\theta\|}{(2\pi)^{\frac{p}{2}}} e^{-\frac{\|\theta\|^2}{2}} \frac{d}{d\|\theta\|} \left\{ e^{\frac{\|\theta\|^2}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h\left(\sum_{i=1}^{p} x_i^2\right) \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{p} x_i^2 - 2\|\theta\| x_1 + \|\theta\|^2\right)\right) dx_1\, dx_2 \ldots dx_p \right\} \\
= \; & 2\|\theta\| e^{-\frac{\|\theta\|^2}{2}} \frac{d}{d\|\theta\|} \left\{ e^{\frac{\|\theta\|^2}{2}} \mathbb{E}\big(h(\chi^2_p(\|\theta\|^2))\big) \right\} \\
= \; & 2\|\theta\| e^{-\frac{\|\theta\|^2}{2}} \frac{d}{d\|\theta\|} \left\{ e^{\frac{\|\theta\|^2}{2}} \mathbb{E}\big(h(\chi^2_{p+2K})\big) \right\} \text{ where } K \sim \text{Poi}\left(\frac{\|\theta\|^2}{2}\right) \\
= \; & 2\|\theta\| e^{-\frac{\|\theta\|^2}{2}} \frac{d}{d\|\theta\|} \left\{ \sum_{k=0}^{\infty} \frac{\left(\frac{\|\theta\|^2}{2}\right)^k \mathbb{E}\big(h(\chi^2_{p+2k})\big)}{k!} \right\} \\
= \; & 2\|\theta\| e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \|\theta\| \frac{k}{k!} \left(\frac{\|\theta\|^2}{2}\right)^{k-1} \mathbb{E}\big(h(\chi^2_{p+2k})\big) \\
= \; & 4e^{-\frac{\|\theta\|^2}{2}} \sum_{k=0}^{\infty} \left(\frac{\|\theta\|^2}{2}\right)^k \frac{k}{k!} \mathbb{E}\left(h\left(\chi^2_{p+2k}\right)\right).
\end{aligned}$$

□

Lemma 6.3. *Suppose $X|\theta \sim N_p(\theta, I)$ and we have a two-stage prior for θ. At the first stage $\theta|\lambda \sim N_p\left(0, \frac{1-\lambda}{\lambda}\right)$, and at the second stage, $\lambda \sim (1-b)\lambda^{-b}$ on $0 < \lambda \leq 1$, where $-\infty < b < 1$. Then*

$$\mathbb{E}(\lambda|X) = \frac{1}{\|X\|^2}\left[p + 2 - 2b - \frac{2}{\int_0^1 \lambda^{\frac{p}{2}-b} \exp(\frac{1-\lambda}{2}\|X\|^2) d\lambda}\right]$$

Proof. The joint posterior distribution of θ and λ is given by

$$\pi(\theta, \lambda|X) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{p}(X_i - \theta_i)^2\right) \frac{\lambda^{\frac{p}{2}-b}}{(1-\lambda)^{\frac{p}{2}}} \exp\left(-\frac{\lambda}{2(1-\lambda)}\sum_{i=1}^{p}\theta_i^2\right).$$

Hence the marginal posterior of λ is obtained by integrating out θ:

$$\begin{aligned}\pi(\lambda|X) &\propto \int_{\mathbb{R}^p} \frac{\lambda^{\frac{p}{2}-b}}{(1-\lambda)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\left(1+\frac{\lambda}{1-\lambda}\right)\sum_{i=1}^{p}\left(\theta_i - \frac{X_i}{1+\frac{\lambda}{1-\lambda}}\right)^2 - \frac{1}{2}\sum_{i=1}^{p} X_i^2 + \frac{1}{2}\sum_{i=1}^{p}\frac{X_i^2}{1+\frac{\lambda}{1-\lambda}}\right) d\theta \\ &\propto \lambda^{\frac{p}{2}-b} e^{-\frac{\lambda}{2}\|X\|^2}.\end{aligned}$$

Thus the marginal posterior mean can be calculated as follows:

$$\begin{aligned}\mathbb{E}(\lambda|X) &= \frac{\int_0^1 \lambda^{\frac{p}{2}-b+1} e^{-\frac{\lambda}{2}\|X\|^2}\, d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-b} e^{-\frac{\lambda}{2}\|X\|^2}\, d\lambda} \\ &= \frac{\left[-\frac{2}{\|X\|^2}\lambda^{\frac{p}{2}-b+1} e^{-\frac{\lambda}{2}\|X\|^2}\right]_0^1 + \frac{2}{\|X\|^2}(\frac{p}{2}-b+1)\int_0^1 \lambda^{\frac{p}{2}-b} e^{-\frac{\lambda}{2}\|X\|^2}\, d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-b} e^{-\frac{\lambda}{2}\|X\|^2}\, d\lambda} \\ &= \frac{1}{\|X\|^2}\left[p+2-2b-\frac{2}{\int_0^1 \lambda^{\frac{p}{2}-b}\exp(\frac{1-\lambda}{2}\|X\|^2) d\lambda}\right],\end{aligned}$$

as required. □

References

[1] Baranchik, A. (1964) *Multiple regression and estimation of the mean of a multivariate normal distribution*, Technical Report 51, Dept. Statistics, Stanford Univ..

[2] Bass, R. (1995) *Probabilistic techniques in analysis*, Springer, New York.

[3] Berger, J. (1975) *Minimax estimation of location vectors for a wide class of densities*, Ann. Statist., **3**, 1318–1328.

[4] Berger, J. (1980) *Statistical decision theory*, Springer-Verlag, New York.

[5] Bhattacharya, P. *Estimating the mean of a multivariate normal population with general quadratic loss function*, Ann. Math. Statist., **37,** 1819–1824.

[6] Blyth, C. (1951), *On minimax statistical decision procedures and their admissibility*, Ann. Math. Statist., **22**, 22–42.

[7] Bock, M. *Minimax estimators of the mean of a multivariate normal distribution*, Ann. Statist., **3,** 209–218.

[8] Bock, M. (1985), *Minimax estimators that shift towards a hypersphere for location vectors of spherically symmertric distributions*, J. Multivariate Anal., **17**, 127–147.

[9] Brandwein, A. (1979) *Minimax estimation of the mean of spherically symmetric distributions under general quadratic loss*, J. Multivariate Anal., **9**, 579–588.

[10] Brandwein, A. and Strawderman, W. (1978) *Minimax estimation of location parameters for spherically symmetric unimodal distributions*, Ann. Statist., **6**, 377–416.

[11] Brandwein, A. and Strawderman, W. (1980) *Minimax estimation of location parameters for spherically symmetric distributions with concave loss*, Ann. Statist., **8,** 279–284.

[12] Brandwein, A. and Strawderman, W. (1990), *Stein estimation: the spherically symmetric case*, Statist. Sci., **5**, 356–369.

[13] Brandwein, A. and Strawderman, W. (1991) *Generalizations of James-Stein estimators under spherical symmetry*, Ann. Statist., **19**, No.3, 1639–1650.

[14] Brown, L. (1966) *On the admissibility of invariant estimators of one or more location parameters*, Ann. Math. Statist., **37**, 1087–1135.

[15] Brown, L. (1971) *Admissible esimators, recurrent diffusions and insoluble boundary value problems*, Ann. Math. Statist., **42,** No. 3, 855–903.

[16] Cellier, D. and Fourdrinier, D. (1995) *Shrinkage estimators under spherical symmetry for the general linear model*, J. Multivariate Anal., **52**, 338–351.

[17] Cohen, A. and Strawderman, W. (1971) *Admissibility of estimators of the mean of a multivariate normal distribution with quadratic loss*, Ann. Math. Statist., **42**, 270–296.

[18] Dempster, A. (1969) *Elements of continuous multivariate analysis*, Addison-Wesley, Reading, Mass..

[19] Efron, B. and Morris, C. (1973), *Stein's estimation rule and its competitors - an empirical Bayes approach*, J. Amer. Statist. Assoc., **68,** 117–130.

[20] Efron, B. and Morris, C. (1977), *Stein's Paradox in Statistics*, Scientific American, May issue, 119–127.

[21] Evans, S. and Stark, P. (1996), *Shrinkage estimators, Skorohod's problem and stochastic integration by parts*, Ann. Statist., **24**, No. 2, 809–815.

[22] Ferguson, T. (1967) *Mathematical statistics: a decision theoretic approach*, Academic Press, New York.

[23] Fitzsimmons, P. (1991) *Skorokhod embedding by randomised hitting times*, Seminar on Stochastic Processes, 1990, 183–192. Birkhauser, Boston.

[24] Kiefer, J. (1957) *Invariance, minimax sequential estimation and continuous time processes*, Ann. Math. Statist., **28**, 573–601.

[25] James, W. and Stein, C. (1961), *Estimation with quadratic loss* Proc. 4th Berkeley Symposium, **1**, 361–379. Univ. of California Press.

[26] Lehmann, E. (1983), *Theory of point estimation*, Wiley, New York.

[27] Rost, H. (1971), *The stopping distributions of a Markov process*, Invent. Math., **14**, 1–16.

[28] Stein, C. (1956), *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proc. 3rd Berkeley Symposium, **1**, 197–206. Univ. of California Press.

[29] Stein, C. (1962) *Confidence sets for the mean of a multivariate normal distribution (with discussion)*, J. Roy. Statist. Soc. Ser. B, **24**, 265–296.

[30] Stein, C. (1981) *Estimation of the mean of a multivariate normal distribution*, Ann. Statist., **9**, 1135–1151.

[31] Strawderman, W. (1971) *Proper Bayes minimax estimators of the multivariate normal mean*, Ann. Math. Statist., **42**, 385–388.

[32] Strawderman, W. (1974) *Minimax estimation of location parameters for certain spherically symmetric distributions*, J. Multivariate Anal., **4**, 255–264.