EFFICIENT BIASED ESTIMATION
AND
APPLICATIONS
TO
LINEAR MODELS

A thesis presented in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
in Statistics
at Massey University

Terry Moore
1981

Abstract

Efficient Biased Estimation and Applications to Linear Models

In recent years biased estimators have received a great deal of attention because they can often produce more accurate estimates in multiparameter problems. One sense in which biased estimators are often more accurate is that the mean square error is smaller.

In this work several parametric families of estimators are examined and good values of the parameters are sought by approximate analytical arguments. These parametric values are then tested by computing and plotting graphs of the mean square error. In this way the risks of various estimators may be seen and it is possible to discard some estimators which have large risk.

The risk functions are computed by numerical integration - a method faster and more accurate than the usual simulation studies. The advantage of this is that it is possible to evaluate a greater number of estimators; however, the method only copes with spherically symmetric estimators.

The relationship of biased estimation to the use of prior information is made clear. This leads to discussion of partially spherically symmetric estimators and the fact that, although not uniformly better than spherically symmetric ones, they are usually better in a practical sense.

It is shown how the theoretical results may be applied to the linear model. The linear model is discussed in the very general case in which it is not of full rank and there are linear restrictions on the parameter. A kind of weak prior knowledge which is often assumed for such a model makes the partially symmetric estimators attractive.

Distributions of spherically symmetric estimators are briefly discussed.

Acknowledgements

I should like to thank Dr. Richard J. Brook and Prof. Brian Hayman of the department of Mathematics and Statistics, Dr. Brook for supervision of this project and Prof. Hayman for acting as second supervisor as well as giving more substantial help during Dr. Brook's sabbatical leave. I should also like to thank several members of the department with whom I had profitable discussions which helped clarify my thinking. These include Dr. Howard Edwards, Dr. mike Carter, Dr. Rob Pringle, Mr. John Reynolds, Mr. Adrian Swift and Dr. Peter Thomson. My thanks also go to Mrs. Gail Tyson for typing a small but significant part of the manuscript and to the long suffering Dr. Ted Drawneek of the computer unit who had to considerably modify his plotter package to cope with my requirements.

Preface

In recent years it has become apparent that biased estimators often give estimates which are more accurate than unbiased estimators. One way of measuring the accuracy of an estimator is by means of its mean square error. Stein was the first to show that the usual unbiased estimator for the mean of a multivariate normal distribution is inadmissible in the sense that there are estimators with smaller mean square error. In fact the mean square error of the James-Stein estimator which shrinks the usual estimates towards the origin, is often very much smaller. Of course, an estimator which is not unbiased is biased. This seems to be a bad property of an estimator - but so, it would seem, is the property of being inadmissible. In fact both words are technical terms and should not be thought of as having their everyday meanings. There are a variety of ways of measuring the bias of an estimator and a variety of ways of measuring its mean deviation from the true value. The properties of the estimator depend critically on how these things are measured.

In chapter 1 we review some of the properties of estimators and suggest Bayes and empirical Bayes estimators as tools for finding estimators with good properties with respect to repeated sampling. Some of the ways of doing so are surveyed and the results suggest the form which good estimators might take. These estimators shrink the usual estimates towards the origin as does the James-Stein estimator. There is little in this chapter which is new.

Chapter 2 leans heavily on the work of Stein and in particular we prove a result which Stein only proves asymptotically.

It is well known that a linear model can be transformed into the canonical form for which Stein proved his results. We show how to apply the James-Stein estimator directly to the general linear model whether or not of full rank and with or without linear restrictions imposed upon it. We then prove the result alluded to above which shows that separate shrinkages in several linear subspaces of the parameter space are generally better than one over-all shrinkage. The result also gives a bound on the loss of mean square error which may be incurred by such separate shrinkages. Graphs of the difference in risk for shrinkages in two subspaces and the risk for a single subspace shrinkage are plotted in three dimensions together with a

contour map showing the region of improvement.

Chapter 3 is closely related to work of Lindley and Smith and to work of Tiao and Zellner. The results are again given for the non-full rank model with linear restrictions. This generalisation poses difficulties when stages of prior information are incorporated in a natural order. It is in this part of chapter 3 that the novelty lies.

Another approach to estimation, Theil's so called "minimum mean square error estimation", is the topic of chapter 4. This criterion does not lead to an estimator as the statistic calculated depends upon unknown parameters. How this statistic itself can be estimated, and the properties of the estimators thus obtained, are discussed. Some distributional properties of quadratic forms and their ratios are derived in a discussion of consistent estimation. The resulting estimators belong to a parametric family of estimators. The various approaches to estimation of the shrinkage factor suggest possible parameter values which are then tested by numerical computation of the risk function. Graphs of these are plotted and displayed at the end of chapter 5. This material is mostly the creation of the author.

Chapter 5 discusses iterative improvement of the estimators of chapter 4. Although this was originally discussed by Hemmerle, we consider several different and novel approaches and compute and plot the risk functions of the resulting estimators. Graphs of the risk functions of these estimators are plotted together with the graphs of the estimators of chapter 4.

The theoretical computation of the risk functions for shrunken estimators was postponed until chapter 6 so that it could first be seen for what class of estimators this should be done. A wide selection of different formulae for the risk are given with the proofs arranged in a systematic manner. If only a few of the formulae are required then the proofs can be simplified by ignoring certain previous results used for computing other forms of the risk. If this is done then more elegant proofs than those used given are obtained. Some generalisations to non-spherically symmetric estimators are given and these are new. These expressions lead to an easier proof of a minimaxity condition than that given by Strawderman in a generalisation of a theorem of Baranchik, and a

non-minimaxity theorem of Efron and Morris is generalised to the non-spherically symmetric case.

In chapter 7 some risk estimate domination results of Efron and Morris are generalised by using an unbiased estimator for the risk in the manner of Efron and Morris. This generalisation is not completely successful but some results are obtained.

The distributions of James-Stein and other shrunken estimators have never been given. Possibly of more use is the distribution of the Studentised version. In chapter 8 this is shown to be a transformation of a multivariate t-distribution. Some of the results in chapter 4 on ratios of quadratic forms will lead, with tedious computations, to moments of the James-Stein estimator but this was not done as the Studentised version is of more value.

We have not given a complete bibliography of work in the general area covered by this work, nor have we referred to every paper in the more precise areas in this thesis. The works cited are directly related to the development of this work.

In order to make this work as self contained as possible we have given some standard results along with their proofs and have appended some general mathematical formulae which have been used heavily.

Equations and theorems have been numbered consecutively within each section and are referred to in that section by their numbers. When referenced outside their own section their numbers are prefixed by the chapter and section number. Diagrams are numbered consecutively throughout the whole thesis.

Contents

Chapter 1
Point Estimation

1.1 Criteria for Choosing Estimators

In this work we shall justify the choice of estimators by their sampling theory properties. However, one particular sampling theory property, that of unbiasedness, shall be of no interest to us. One reason for this is that estimators with a small amount of bias are often vastly better in terms of mean square error than unbiased estimators. We shall use the mean square error of an estimator as a criterion, the smaller the mean square error the better, since this penalises very strongly estimators which tend, on average, to be far from correct. The mean square error is the risk function corresponding to a quadratic loss function. Although the loss function is often an arbitrary choice, quadratic loss is usually fairly tractable (especially when the sampling distribution is normal) and behaves in a reasonable manner in that, the greater the difference between an estimate and the true value, the greater the loss. It has been argued that a loss function should be bounded, but in the case of a sampling distribution which is normal, quadratic loss and any bounded loss functions which approximate it near the true parameter value do not give very different results.

A growing number of statisticians, but still a minority, prefer to use Bayesian methods. Given a prior distribution, $p(\theta)$, for the parameter θ (which may be proper or improper), and the likelihood function $\ell(\theta|X)$, the posterior distribution, $f(\theta|X) \propto p(\theta)\, \ell(\theta|X)$ may be calculated. This posterior distribution should be proper for a reasonable point estimator for θ to be computable from it. Possible choices for point estimator are the mean, median or mode of the posterior distribution. Usually the mean is chosen as this often leads to admissible estimators when the loss function is quadratic, that is, no estimator has uniformly smaller risk. In the next section we shall state the usual definitions and prove this well known theorem since it justifies the methods we shall use henceforward.

1.1.1 Loss Functions and Risk Functions

Let $X_1, X_2, \ldots, X_n$ be a sample from a distribution with parameter θ (the parameter, the observations, or both may be scalars or vectors). Let the likelihood function be $\ell(\theta|X_1,\ldots,X_n)$ and let $\hat{\theta}(X_1,\ldots,X_n)$ be an estimator for θ. We denote the loss function by $\mathcal{L}(\hat{\theta}(X_1,\ldots,X_n),\theta)$.

Given a prior distribution $p(\theta)$ for θ the *Bayes Posterior Risk*, $r_{\hat{\theta}}(X_1,\ldots,X_n)$ is given by

$$r_{\hat{\theta}}(X_1,\ldots,X_n) = E[\int(\hat{\theta}(X_1,\ldots,X_n),\theta)|X_1,\ldots,X_n]$$

where the expectation is taken with respect to the posterior distribution of θ.

A sampling theorist, having no prior distribution, cannot compute this. Instead he may compute the *risk function*, $R_{\hat{\theta}}(\theta)$, given by

$$R_{\hat{\theta}}(\theta) = E[\int(\hat{\theta}(X_1,\ldots,X_n), \theta)|\theta]$$

where the expectation is taken with respect to the probability distribution for $X_1,\ldots,X_n$ given θ.

Taking the expectation over both the sample space $\$(X)$ and the parameter space $\$(\theta)$ gives the *Bayes risk*

$$\bar{R}_{\hat{\theta}} = E[E[\int(\hat{\theta}(X_1,\ldots,X_n),\theta)|\theta]].$$

It must be noted that some or all of these quantities may not exist, although the existence of the Bayes risk implies the existence of the other two. This follows from Fubini's theorem which allows us to replace a double integral by repeated single intervals, in either order, if the former exists. That is

$$\begin{aligned}\bar{R}_{\hat{\theta}} &= E[R_{\hat{\theta}}(\theta)] \\ &= E[E[\int(\hat{\theta}(X_1,\ldots,X_n), \theta)|\theta]] \\ &= E[\ell(\hat{\theta}(X_1,\ldots,X_n),\theta)] \\ &= E[E[\ell(\hat{\theta}(X_1,\ldots,X_n),\theta)|X_1,\ldots,X_n]] \\ &= E[r_{\hat{\theta}}(X_1,\ldots,X_n)].\end{aligned}$$

It is when the loss function is unbounded or the prior is improper that the Bayes risk may fail to exist.

1.1.2 Admissibility

To a sampling theorist, the risk function, being independent of any prior distribution, can be used for the comparison of estimators. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators and if $R_{\hat{\theta}}(\theta) \leq R_{\hat{\theta}_2}(\theta)\ \forall\ \theta \in \$\{\theta\}$ (where $\$\{\theta\}$ is the parameter space) then θ_1 is said to dominate $\hat{\theta}_2$. The risk function therefore defines a partial ordering of the set of estimators for θ. It is not, unfortunately, a total ordering because there are pairs of estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for which $\$\{\theta\}$ may be partitioned into subsets S_1,S_2,S_3 in such a way that $S_1 \neq \emptyset$ $S_3 \neq \emptyset$ and $R_{\hat{\theta}_1}(\theta) < R_{\hat{\theta}_2}(\theta)\ \forall\ \theta \in S_1$, $R_{\hat{\theta}_2}(\theta) < R_{\hat{\theta}_1}(\theta)\ \forall\ \theta \in S_3$

and $R_{\hat{\theta}_1}(\theta) = R_{\hat{\theta}}(\theta)\ \forall\ \theta \in S_2$. This means that one estimator is not always superior to the other. However, a minimal element in the partial ordering is superior to every other element (estimator) to which it is comparable. Such an estimator is said to be *admissible*. An estimator, $\hat{\theta}_1$, is admissible, then, if there is no other estimator $\hat{\theta}_2$ for which $R_{\hat{\theta}_2}(\theta) \leq R_{\hat{\theta}_1}(\theta)\ \forall\ \theta \in \$\{\theta\}$.

Since $\bar{R}_{\hat{\theta}} = E[r_{\hat{\theta}}(X_1,\ldots,X_n)]$ (if it exists) it follows that an admissible estimator has minimal Bayes risk for any prior distribution for which the Bayes risk exists. Similarly, minimising the Bayes posterior risk gives minimum Bayes risk and hence gives rise to an admissible estimator. *It is to be noted that this only applies if there is an estimator for which the Bayes risk exists.*

The estimator which minimises the Bayes posterior risk is called the *Bayes estimator*. If the Bayes risk for the Bayes estimator does not exist then this estimator is not necessarily admissible, but it may be. In the case of a p-variate normal distribution, $X \sim N_p(\theta,\Sigma)$ with Σ known and with uniform prior distribution for θ and loss function $\mathcal{L}(\hat{\theta},\theta) = (\hat{\theta}-\theta)^T(\hat{\theta}-\theta)$ the Bayes estimator is the minimum variance unbiased estimator (also the least squares estimator and maximum likelihood estimator). This estimator is admissible if $p = 1$ or $p = 2$ but not admissible if $p \geq 3$ (as was first shown by Stein (1955)).

It is clear that, even for a non-Bayesian, a powerful tool for finding admissible estimators is to assume a prior distribution and find the corresponding Bayes estimator. In many cases it can be shown, Fergusson (1967), that an admissible estimator must be a Bayes estimator or a generalised Bayes estimator (i.e. a Bayes estimator based on an improper prior distribution). In this case the importance of Bayes estimators to a non-Bayesian is obvious.

The converse problem is finding whether an estimator is a Bayes estimator and, if so, finding the prior distribution for which it is, has been discussed by Strawderman (1971). Strawderman and Cohen (1971) given conditions under which an improper Bayes estimator is admissible or inadmissible for the case of the multivariate normal distribution with known variance, while Brown (1966) gives classes of prior distributions which lead to admissible estimators.

In this brief summary precise details have not been given. Fergusson (1967) gives more precise proofs of the connection between admissible estimators and Bayes estimators.

1.2 Estimators for the Mean of a Multivariate Normal Distribution under Quadratic Loss

Suppose, for simplicity, that a random variable, X, has a multivariate normal distribution $X \sim N_p(\mu,\sigma^2 I)$ with σ^2 known. (We use the symbol $N_p(\mu,V)$ for a p-variate normal distribution with mean μ and dispersion matrix V). On the basis of a sample $X_1,\ldots,X_n$ we wish to estimate μ under the quadratic loss function

$$\mathcal{L}(\hat{\mu},\mu) = (\hat{\mu}-\mu)^T(\hat{\mu}-\mu)/\sigma^2.$$

The minimum variance unbiased estimator is $\hat{\mu}(X_1,\ldots,X_n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and this minimises the risk among the class of unbiased estimators whatever the value of μ. This is also the maximum likelihood estimator for μ.

If we do not wish to restrict ourselves to unbiased estimators then we cannot uniformly minimise the risk but we can search for admissible estimators and these are found amongst the Bayes estimators. We shall therefore choose a prior distribution for μ. Now with bounded loss function and proper prior we are assured of an admissible estimator. However, our loss function is not bounded and we do not wish to restrict ourselves to using a proper prior when we have little prior knowledge as this might weight our estimates unfairly towards the prior mean.

The most obvious prior distribution for μ is, perhaps, the uniform prior. This leads to the posterior probability distribution for μ being proportional to the likelihood function. Since the likelihood function is

$$\begin{aligned}\ell(\mu;X_1,\ldots,X_n) &= \exp\{-\tfrac{1}{2}\sigma^{-2}\sum_{i=1}^{n}\|(X_i-\mu)\|^2\} \\ &= \exp\{-\tfrac{1}{2}\sigma^{-2}\sum_{i=1}^{n}\|X_i-\bar{X}\|^2\}\exp\{-\frac{n}{2\sigma^2}\|X-\mu\|^2\}\end{aligned}$$

where $\|a\| = (a^Ta)^{\frac{1}{2}}$, we have the result that this is symmetric about $\hat{\mu} = \bar{X}$ so that the maximum likelihood estimator coincides with the posterior mean. This estimator turns out to be admissible in one or two dimensions but inadmissible in three or more dimensions (Stein(1955))

If we transform our parameters to polar coordinates the uniform prior distribution becomes a spherically symmetric distribution with density of $r = (\mu^T\mu)^{\frac{1}{2}} = \|\mu\|$ given by $p(r) = r^{p-1}$. This is a non-uniform density and puts a large weighting on large values of $\|\mu\|$. We might try a prior distribution without this feature. For example,

the priors, $p(r) = r^a$, $a < p-1$, overcome this defect to a greater or lesser extent depending on the value of a. For any value of a this density for r gives rise to the density for μ, $p(\mu) = (\mu^T\mu)^{\frac{1}{2}(a-p+1)} = (\mu^T\mu)^t$. If t is negative then $a < p-1$ and we avoid weighting large values of μ too heavily.

In order to find the Bayes estimator we shall show that, for a quadratic loss function, the mean of the posterior distribution minimises the Bayes posterior risk.

We have

$$\begin{aligned}
&\frac{\partial}{\partial\hat{\mu}} E[(\hat{\mu}(X)-\mu)^T(\hat{\mu}(X)-\mu)|X] \\
&= E[\frac{\partial}{\partial\hat{\mu}}(\hat{\mu}(X)-\mu)^T(\hat{\mu}(X)-\mu)|X] \\
&= E[2(\hat{\mu}(X)-\mu)|X] \\
&= 2\{\hat{\mu}(X) - E[\mu|X]\}.
\end{aligned}$$

This is zero if and only if $\hat{\mu}(X) = E[\mu|X]$ and this value clearly gives a minimum.

In the normal distribution case we have

$$\hat{\mu}(X_1,\ldots,X_n) = \frac{\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\mu(\mu^T\mu)^t \exp\{-\frac{n}{2\sigma^2}\|\bar{X}-\mu\|^2\}d\mu}{\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}(\mu^T\mu)^t \exp\{-\frac{n}{2\sigma^2}\|\bar{X}-\mu\|^2\}d\mu}.$$

In order that the integral in the denominator should converge we must have $t > -\frac{p}{2}$ (ie $a > -1$). With this restriction on t we may calculate $\hat{\mu}$ as follows. Let $\nu \sim N_p(\bar{X},\frac{\sigma^2}{n} I)$ then, if the density of ν is $p(\nu|\bar{X})$, we have

$$\hat{\mu}(X_1,\ldots,X_n) = \frac{E[\nu(\nu^T\nu)^t]}{E[(\nu^T\nu)^t]}$$

and $$\frac{\partial}{\partial\bar{X}} p(\nu|\bar{X}) = \frac{n}{\sigma^2}(\nu-\bar{X})\,p(\nu|\bar{X}).$$

Therefore, $$E[\nu(\nu^T\nu)^t] = \bar{X}E[(\nu^T\nu)^t] + \frac{\sigma^2}{n}\frac{\partial}{\partial\bar{X}} E[(\nu^T\nu)^t]$$

so that $$\hat{\mu}(X_1,\ldots,X_n) = \bar{X} + \frac{\sigma^2}{n}\frac{\frac{\partial}{\partial\bar{X}} E[\|\nu\|^{2t}]}{E[\|\nu\|^{2t}]}.$$

If $\nu \sim N_p(\bar{X}, \frac{\sigma^2}{n} I)$ then $\|\nu\| \sim \frac{\sigma^2}{n}\chi(p,z)$

where $z = \frac{n}{2\sigma^2}\bar{X}^T\bar{X}$ (and $\therefore$ $\frac{\partial z}{\partial\bar{X}} = \frac{n}{\sigma^2}\bar{X}$).

Using the properties of hypergeometric functions and moments of χ^2 distributions given in appendix 2 we obtain

$$E[\|\nu\|^{2t}] = \left(\frac{2\sigma^2}{n}\right)^t \left(\frac{p}{2}\right)_t e^{-z}\ {}_1F_1\left(\frac{p}{2}+t;\ \frac{p}{2};\ z\right)$$

and $$\frac{\partial}{\partial z} E[\|\nu\|^{2t}] = \left(\frac{2\sigma^2}{n}\right)^t \left(\frac{p}{2}\right)_t e^{-z}\left[\left(1+\frac{2t}{p}\right){}_1F_1\left(\frac{p}{2}+t+1;\frac{p}{2}+1;z\right) - {}_1F_1\left(\frac{p}{2}+t;\frac{p}{2};z\right)\right]$$

giving us the result that

$$\hat{\mu}(X_1,\ldots,X_n) = \frac{\frac{1}{2}p+t}{\frac{1}{2}p}\ \frac{{}_1F_1(\frac{1}{2}p+t+1;\frac{1}{2}p+1;z)}{{}_1F_1(\frac{1}{2}p+t;\frac{1}{2}p;z)}$$

Since ${}_1F_1(a;a;z) = e^z$, the special case $t = 0$ gives $\hat{\mu} = \bar{X}$ which is the well known special case of a uniform prior for μ.

If t is a positive integer or if $\frac{1}{2}p + t$ is a negative integer then this expression gives a rational function of z. However, for the former case we do not expect the estimator to perform well, while for the latter case the estimator is not a Bayes estimator since the integral does not converge. For other values of t we do not expect to have a rational function of z.

Now $\hat{\mu}$ is a scalar multiple of $\bar{X}$. We shall show that for $-\frac{1}{2}p < t < 0$ the multiplying factor lies between zero and one.

From the asymptotic expansion for the confluent hypergeometric function ${}_1F_1(a;b;z) \sim \frac{\Gamma(b)}{\Gamma(a)} e^z z^{a-b}\ {}_2F_0(1-a,b-a;;1/z)$ we obtain

$$\frac{\frac{1}{2}p+t}{\frac{1}{2}p}\ \frac{{}_1F_1(\frac{1}{2}p+t+1;\frac{1}{2}p+1;z)}{{}_1F_1(\frac{1}{2}p+t;\frac{1}{2}p;z)} \sim \frac{{}_2F_0(-\frac{1}{2}p-t,-t;;1/z)}{{}_2F_0(1-\frac{1}{2}p-t,-t;;1/z)}$$

$$\sim \frac{z+t(\frac{1}{2}p+t)}{z+t(\frac{1}{2}p+t-1)} \sim 1 \text{ as } z \to \infty.$$

Thus as $z \to \infty$ the multiplying factor tends to 1. Also $0 < \frac{\frac{1}{2}p+t}{\frac{1}{2}p} < 1$ so that when $z = 0$ the multiplying factor lies between zero and one. We now complete the proof by showing that the multiplying factor is an increasing function of z. This is a special case of the following lemma given in Lehman (1959).

Lemma 1 If for $i = 0,1,2,3,\ldots$ $a_i > 0$ and $b_i > 0$ and if the series $\sum_{i=0}^{\infty} a_i z^i$ converges to $A(z)$ and the series $\sum_{i=0}^{\infty} b_i z^i$ converges to $B(z)$ then for $z > 0$ $f(z) = \frac{A(z)}{B(z)}$ is an increasing function of z if $\frac{a_i}{b_i}$ is an increasing function of i.

Proof For f(z) to be an increasing function of z it is necessary that $f'(z) > 0$. Now $f'(z) = [A'(z)B(z) - A(z)B'(z)]/[B(z)]^2$ and so we require that $A'(z)B(z) - A(z)B'(z) > 0$.

Now
$$A'(z)\,B(z) - A(z)\,B'(z) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[i\,a_i\,b_j z^{i+j-1} - i\,a_j b_i\,z^{i+j-1}]$$
$$= \sum_{n=0}^{\infty}\sum_{k=0}^{n} k(a_k b_{n-k} - a_{n-k}b_k)z^{n-1}$$
$$= \sum_{n=1}^{\infty} c_n\, z^{n-1}$$

where
$$c_n = \sum_{0\le k<\frac{1}{2}n} k(a_k b_{n-k} - a_{n-k}b_k) + \sum_{\frac{1}{2}n<k\le n} k(a_k b_{n-k} - a_{n-k}b_k)$$
$$= \sum_{0\le k<\frac{1}{2}n} k(a_k b_{n-k} - a_{n-k}b_k) + \sum_{0\le k<\frac{1}{2}n} (n-k)(a_{n-k}b_k - a_k b_{n-k})$$
$$= \sum_{0\le k<\frac{1}{2}n} (n-2k)(a_{n-k}b_k - a_k b_{n-k})$$

Now for $0 \le k < \frac{1}{2}n$ $n - 2k > 0$ and $n-k > k$

thus if $a_s b_k > a_k b_s$ for $s > k$ then $c_n > 0$.

However, $a_s b_k > a_k b_s$ is equivalent to $\dfrac{a_s}{b_s} > \dfrac{a_k}{b_k}$ since $b_i > 0$ for all i and the result follows.

In our application of this lemma

$$a_i = \frac{(\frac{1}{2}p+t)_{i+1}}{(\frac{1}{2}p)_{i+1}}\,\frac{1}{i!}, \qquad b_i = \frac{(\frac{1}{2}p+t)_i}{(\frac{1}{2}p)_i}\,\frac{1}{i!}$$

so we have
$$\frac{a_i}{b_i} = \frac{\frac{1}{2}p + t + i}{\frac{1}{2}p + i} = 1 + \frac{t}{\frac{1}{2}p + i}$$

which is an increasing function of i if and only if $t < 0$. We have thus shown that the multiplying factor lies between zero and one if $-\frac{1}{2}p < t < 0$. This means that $\hat{\mu}$ is a shrinkage of $\bar{X}$.

The estimator $\hat{\mu}(X_1,\ldots,X_n)$ given above may be computed easily for small values of z, or, using the asymptotic expansion given previously, for large z. For intermediate values of z the aid of a computer may be required. It seems desirable to find a more easily calculated approximation to the shrinkage factor. Let h(z) be the shrinkage factor. Now as $z \to 0$ and as $z \to \infty$, h(z) is asymptotically equal to a bilinear function. We shall approximate h(z) by such a function. We first prove the following lemma.

Lemma 2 If $f(z) = \dfrac{\sum_{i=0}^{\infty} a_i z^i}{\sum_{i=0}^{\infty} b_i s^i}$ and $g(z) = \dfrac{a_0+a_1 z}{b_0+b_1 z}$

then as $z \to 0$ $f(z) \sim g(z)$ and $f'(z) \sim g'(z)$.

Proof Clearly $f(z) \sim g(z)$.

Now $$f'(z) = \frac{\sum_{i=1}^{\infty}\sum_{j=0}^{\infty} i(a_i b_j - a_j b_i) z^{i+j-1}}{(\sum_{j=0}^{\infty} b_j z^j)^2} \sim \frac{a_1 b_0 - a_0 b_1}{(b_0+b_1 z)^2} \text{ as } z \to 0,$$

and $$g'(z) = \frac{a_1 b_0 - a_0 b_1}{(b_0 + b_1 z)^2}.$$

Thus $f'(z) \sim g'(z)$.

Applying this result to $h(z)$ shows that the first two terms of the Taylor series in the numerator and denominator approximates $h(z)$ in value and derivative at $z = 0$.

A similar lemma for large values of z is as follows.

Lemma 3 If $f(z) \sim \dfrac{\sum_{i=0}^{\infty} a_i z^{-i}}{\sum_{i=0}^{\infty} b_i z^{-i}}$ and $g(z) = \dfrac{a_0}{b_0}$

where the series are asymptotic expansions as $z \to \infty$ then $f(z) \sim g(z)$ and $f'(z) \sim g'(z)$ as $z \to \infty$.

Proof Clearly $f(z) \sim g(z)$.

Now in a similar manner to the previous lemma

$$f'(z) \sim \frac{a_0 b_1 - a_1 b_0}{b_2+b_1 z+b_0 z^2} \to 0 \text{ as } z \to \infty.$$

Since $g'(z) = 0$ the result is proved.

This shows that for large z we may approximate $h(z)$ in value and first derivative using just the constant terms. The next lemma gives a bilinear function which approximates another bilinear function for small z in value and first derivative and approaches a given constant for large z.

Lemma 4 Let $f(z) = \dfrac{a+bz}{c+dz}$ and $g(z) = \dfrac{\alpha+\beta z}{\gamma+\delta z}$. Then $f(z) \sim g(z)$ and $f'(z) \sim g'(z)$ as $z \to 0$ and $g(z) \to k$ as $z \to \infty$ if $\alpha = \dfrac{a}{c}\gamma$, $\delta = \dfrac{1}{c}\dfrac{bc-ad}{kc-a}\gamma$, $\beta = \dfrac{k}{c}\dfrac{bc-ad}{kc-a}\gamma$.

Proof We must have $\beta = k\delta$ to satisfy the condition for large z and $\alpha = \dfrac{a}{c}\gamma$ so that $f(0) = g(0)$. In order that $f'(0) = g'(0)$ we must

have $\frac{bc-ad}{c^2} = \frac{\beta\gamma-\alpha\delta}{\gamma^2}$. Thus $k\gamma\delta - \frac{a}{c}\gamma\delta = \frac{bc-ad}{c^2}\gamma^2$ and $\gamma = \frac{1}{c}\frac{bc-ad}{kc-a}\gamma$.

We may now apply these lemmas to the function h(z). In this case

$k = 1,\ a = \frac{\frac{1}{2}p+t}{\frac{1}{2}p},\ b = \frac{(\frac{1}{2}p+t)_2}{(\frac{1}{2}p)_2},\ c = 1,\ d = \frac{\frac{1}{2}p+t}{\frac{1}{2}p}.$

We have $\alpha = \frac{\frac{1}{2}p+t}{\frac{1}{2}p}\gamma$, $\beta = \delta = \left[\frac{(\frac{1}{2}p+t)_2}{(\frac{1}{2}p)_2} - \left(\frac{\frac{1}{2}p+t}{\frac{1}{2}p}\right)^2 \gamma \Big/ \left(1 - \frac{\frac{1}{2}p+t}{\frac{1}{2}p}\right)\right]$

$$= -\frac{\frac{1}{2}p+t}{t}\left[\frac{\frac{1}{2}p+t+1}{\frac{1}{2}p+1} - \frac{\frac{1}{2}p+t}{\frac{1}{2}p}\right]\gamma$$

$$= (\tfrac{1}{2}p+t)\,\gamma \,/\, [\tfrac{1}{2}p(\tfrac{1}{2}p+1)]$$

Taking $\gamma = (\frac{1}{2}p)_2$ we obtain as our bilinear approximation to h(z),

$$h(z) = \frac{(\frac{1}{2}p+t)(\frac{1}{2}p+1)+(\frac{1}{2}p+t)z}{\frac{1}{2}p(\frac{1}{2}p+1)+(\frac{1}{2}p+t)z} = (\tfrac{1}{2}p+t)\,\frac{\frac{1}{2}p+1+z}{\frac{1}{2}p(\frac{1}{2}p+1)+(\frac{1}{2}p+t)z}$$

It is clear that we could use higher order approximations to h(z) but this bilinear approximation will be quite good; also it is doubtful whether much improvement can be gained by going beyond biquadratic approximations.

1.2.1 Other Prior Distributions

The reason that the maximum likelihood estimator tends to overestimate the length of μ is that it is based on a uniform prior. The surface of a sphere of radius r is proportional to r^{p-1} which means that a uniform prior weights large values of r highly. We overcome this by using a prior distribution which puts smaller weighting on large values of r in compensation for this. Stein (1962) calls this effect "the surface-volume effect".

If $p(\mu)$ is the probability density for μ then on taking polar coordinates $(\|\mu\|,\theta)$ we may write $\delta(\theta) = \frac{1}{\|\mu\|}\mu$ so that $\|\delta\| = 1$. The Jacobian of the transformation will be

$$J = \left\| \frac{\partial\mu}{\partial\|\mu\|} \vdots \frac{\partial\mu}{\partial\theta} \right\| = \left\| \delta \vdots \|\mu\| \frac{\partial\delta}{\partial\theta} \right\|$$

$$= \|\mu\|^{p-1} \left\| \delta \vdots \frac{\partial\delta}{\partial\theta} \right\|.$$

We thus have the probability density for $(\|\mu\|,\theta)$

$$p(\|\mu\|,\theta) = p(\mu)\|\mu\|^{p-1}\left\| \delta \vdots \frac{\partial\delta}{\partial\theta} \right\|.$$

This proves the transformation law given previously.

Another way to overcome the surface volume effect is to use a proper prior distribution. In particular we could use a normally distributed prior. We shall consider a family of prior distributions

which includes the previous family and the normal family as special cases. Consider the prior

$$p(\mu|\tau) \propto (\mu^T\mu)^t \exp\{-\tfrac{1}{2}\tau^{-2} (\mu-\alpha)^T(\mu-\alpha)\}$$

When $t = 0$ we obtain the normal prior while, in the limit as $\tau \to \infty$ we obtain the prior of the last section. The posterior distribution is given by

$$p(\mu) \propto \exp\{-\frac{n}{2\sigma^2} \|\mu-\bar{X}\|^2\} (\mu^T\mu)^t \exp\{-\frac{1}{2\tau^2} \|\mu-\alpha\|^2\}$$

$$= (\mu^T\mu)^t \exp\{-\tfrac{1}{2}[\frac{n}{\sigma^2}(\mu^T\mu-2\bar{X}^T\mu+\bar{X}^T\bar{X})+\frac{1}{\tau^2}(\mu^T\mu-2\alpha^T\mu+\alpha^T\alpha)]\}$$

$$\propto (\mu^T\mu)^t \exp\{-\tfrac{1}{2}[(\frac{n}{\sigma^2}+\frac{1}{\tau^2})\mu^T\mu - 2(\frac{n}{\sigma^2}\bar{X}^T+\frac{1}{\tau^2}\alpha^T)\mu]\}$$

$$\propto (\mu^T\mu)^t \exp\{-\tfrac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2}) \|\mu - \frac{\frac{n}{\sigma^2}\bar{X}+\frac{1}{\tau^2}\alpha}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\|^2\}.$$

This is of the same form as the posterior which corresponds to the prior $(\mu^T\mu)^t$. The coefficient $\frac{n}{\sigma^2}$ has been replaced by $\frac{n}{\sigma^2} + \frac{1}{\tau^2}$ and the vector $\bar{X}$ has been replaced by $\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} (\frac{n}{\sigma^2}\bar{X}+\frac{1}{\tau^2}\alpha)$. Thus the mean of the posterior distribution is also of the same form as before.

Usually no value of τ^2 is known but it is possible to estimate τ^2 from the data (the so-called empirical Bayes estimators). We shall illustrate this in the next section for the case $t = 0$. Alternatively we may choose a prior distribution for τ^2 (a so-called two stage Bayesian method).

Consider a prior distribution for τ^2, $p(\tau^2) \propto \tau^{-2c}$. The prior for μ obtained by integrating with respext to τ^2 is

$$p(\mu) \propto \int_0^\infty (\mu^T\mu)^t \exp\{-\tfrac{1}{2}\tau^{-2}\|\mu-\mu\|^2\} \tau^{-2c} d\tau^2$$

$$= (\mu^T\mu)^t \frac{2^{c-1}\Gamma(c-1)}{\|\mu-\alpha\|^{2(c-1)}}$$

$$\propto \frac{\|\mu-\alpha\|^{2(c-1)}}{\|\mu\|^{2t}}$$

In the case in which $\alpha = 0$ this reduces to $p(\mu) \propto \|\mu\|^{2(c-1-t)}$ which is of the form previously considered.

1.2.2 Empirical Bayes Estimators

Instead of choosing a prior for τ^2 and integrating it is possible to estimate τ^2. This may be done because the variation in the usual estimates of the μ_i consists of two components under the random effects model considered. These components are the variation of the μ_i about their mean and the random variation of the $X_{i.}$. The within samples sum of squares estimates σ^2 while the between samples sum of squares estimates $\frac{\sigma^2}{n} + \tau^2$.

Writing X_{ij} for the jth component of the ith sample vector the random effects model gives

$$\hat{\tau}^2 = \frac{1}{n}\left[\frac{1}{p-1}\sum_{i=1}^{p} n(\bar{X}_{i.} - \bar{X}_{..})^2 - \frac{1}{(p-1)(n-1)}\sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.})^2\right]$$

$$\hat{\sigma}^2 = \frac{1}{(p-1)(n-1)}\sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.})^2 .$$

This random effects model assumes unknown variance and that each component of α is the same and can be estimated by $\bar{X}_{..}$. If we modify for the case $\alpha = 0$ and known variance we obtain

$$\hat{\tau}^2 = \frac{1}{p}\sum_{i=1}^{p}\bar{X}_{i.}^2 - \frac{\hat{\sigma}^2}{n} .$$

Using this model the posterior distribution for μ is

$$\mu \sim N_p\left(\frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\,\bar{X},\ \frac{\tau^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2} I\right)$$

and the posterior mean is $\frac{n\tau^2}{n\tau^2+\sigma^2}\,\bar{X} = \left(1 - \frac{\sigma^2}{n\tau^2+\sigma^2}\right)\bar{X}$.

Substituting the above estimator for τ^2 gives

$$\tilde{\mu} = \left(1 - \frac{p\sigma^2}{n\sum_{i=1}^{p}\bar{X}_{i.}^2}\right)\bar{X} = \left(1 - \frac{p}{n}\frac{\sigma^2}{\bar{X}^T\bar{X}}\right)\bar{X}.$$

Using our previous notation $z = \frac{n}{2\sigma^2}\,\bar{X}^T\bar{X}$ we have

$$\tilde{\mu} = \left(1 - \frac{p}{2z}\right)\bar{X} = \frac{z - \frac{1}{2}p}{z}\,\bar{X}.$$

This again is a shrinkage of $\bar{X}$ where the shrinkage factor is a bilinear function of z. It is not the same bilinear function which we derived previously.

When the denominator of a bilinear shrinkage is proportional to z it turns out to be easier to calculate the risk than for a general bilinear shrinkage. James and Stein (1960) gave a modified form of

the above estimator in both the case of known variance and the case of unknown variance. For the latter case we may substitute $\hat{\sigma}^2$ for σ^2. Since the empirical Bayes estimator is only an approximation to the full Bayesian estimator it seems reasonable to check whether a slight modification will give smaller risk. They found that, for the estimator $\breve{\mu}_1 = \frac{z-a}{z}\bar{X}$ the risk function which we shall calculate in chapter 6 is given by

$$E\left[\left\|\left(1-\frac{a}{z}\right)\bar{X}-\mu\right\|^2\right] = p - 4a\left\{(p-2) - \frac{n+2}{n}a\right\} E\left[\frac{1}{p-2+2K}\right]$$

where K has a Poisson distribution with parameter $\frac{\mu^T\mu}{2\sigma^2}$ if the variance is unknown, while if it is known then the risk is

$$E\left\|\left(1-\frac{a}{z}\right)\bar{X}-\mu\right\|^2 = p - 4a\left\{(p-2) - a\right\} E\left[\frac{1}{p-2+2K}\right]$$

with the same Poisson distribution for K.

The estimator is uniformly better than $\hat{\mu}$ if $a = \frac{1}{2}p - 1$ in the case of known variance or if $a = \frac{\frac{1}{2}n}{\frac{1}{2}n+1}(\frac{1}{2}p-1)$ in the case of unknown variance. In fact for values of a between zero and twice this value the estimator is minimax which means that it has uniformly smaller risk than the maximum likelihood estimator.

1.2.3 Admissibility

For the James-Stein estimator given above the shrinkage factor will be negative if $z < a$. Intuitively this would seem to be a bad thing. The reason the estimator performs well on average is that there is only a small chance that $z < a$. If we define $a_+ = a$ if $a \geq 0$ and $a_+ = 0$ $a \leq 0$ then the estimator $\breve{\mu}_1^+ = \left(\frac{z-a}{z}\right)_+^+ \bar{X}$ seems likely to provide an improvement over the James-Stein estimator. In fact this is so and this proves that the James-Stein estimator is not admissible. The estimator $\breve{\mu}_1^+$ is the truncated James-Stein estimator and is not admissible either.

Efron and Morris (1972) have compared the above estimators with the full Bayes estimators. They quote the loss of efficiency due to estimating τ^2 and show that it is small. Thus the James-Stein estimator is almost admissible in a sense because the proper prior we have considered gives rise to an admissible estimator.

Our previous estimator based on the prior $p(\mu) \propto (\mu^T\mu)^t$ is a possible candidate for admissibility since the only admissible estimators are Bayes estimators. Since the Bayes risk does not exist we cannot guarantee admissibility by this result. A paper of

Brown (1971) shows which prior distributions lead to admissible estimators. The proof shows under what conditions a sequence of admissible estimators based on proper priors converges to an admissible estimator. Strawderman and Cohen (1971) derive from this the following simple criterion for the case of known variance. An improper Bayes estimator is admissible if it is a proper shrinkage (i.e. shrinkage factor < 1) of the maximum likelihood estimator - the result applying to estimation of the normal mean under quadratic loss when the variance is known. This means that our Bayes estimator previously derived is admissible but not our bilinear approximations to it.

Some of the above forms of prior knowledge, as well as many other related methods which lead to Stein-like estimators have been summarised in a review paper, Zellner and Vandaele (1972).

1.2.4 Unknown Variance

We have already seen how unknown variance may be dealt with empirically. The full Bayesian procedure of choosing a prior distribution for σ^2 leads to rather intractable integrals. Suppose we use a prior distribution $p(\sigma^2) \propto (\sigma^2)^b \exp(-\frac{e}{\sigma^2})$. We may integrate out the σ^2 fairly easily and we are left with an intractable integral for μ. Alternatively, if we integrate out the μ we are left with an intractable integral for σ^2. These may be solved numerically or by asymptotic expansions but the solution does not give easily computed estimators. An alternative is to use the mode of the posterior distribution for μ and this is more easily computed. We shall discuss these methods in more detail in chapter 3.

1.2.5 Linear Models

We have described the case of a sample from a multivariate normal distribution. This is a special case of a linear model since we may write $X_i = \mu + \varepsilon_i$ where $\varepsilon_i \sim N_p(0,\sigma^2 I)$. We may then write

$$\begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix} = \begin{bmatrix} I \\ \cdot \\ \cdot \\ \cdot \\ I \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \quad \text{where} \begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \sim N_{np}(0,\sigma^2 I)$$

For known variance $\bar{X}$ is a sufficient statistic for μ while for unknown variance

$S = \sum_{i=1}^{n} (X_i - \bar{X})^T (X_i - \bar{X})$ and $\bar{X}$ are jointly sufficient for μ.

Now $\bar{X} \sim N_p(\mu, \frac{\sigma^2}{n} I)$ and $S \sim \chi^2_{n-1}$ independently of $\bar{X}$. We may therefore study the problem of a single observed X with distribution $X \sim N_p(\mu, \sigma^2 I)$ and, if σ^2 is unknown, an independent variate S with distribution $S \sim \chi^2_n$. The linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$ takes this form if we take our sufficient statistics to be $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}^2 = Y^T (I - X(X^T X)^{-1} X^T) Y$. This is the so-called canonical form of the model. In chapter 2 we shall apply the methods of this chapter to the linear model in this way as well as working with the model directly.

1.2.6 Criticism

Estimators obtained in the last few sections have the property of shrinking each component estimate towards a common value. As we have discussed the problem the shrinkage is towards zero, but it is easy to modify the methods to give a shrinkage towards any value or towards the overall mean. Since the shrinkage factor depends on all the data the estimate of one particular coordinate is affected by data concerning other coordinates. In the case we have considered the distributions of the coordinates are independent and unless we have prior knowledge that the coordinates are close together it seems unreasonable to use estimators which have this property.

The reason for the reduction in risk is that if the coordinates are in fact equal in mean then pooling the data is more efficient and protects us against any data for one coordinate being an outlier. If the coordinates are close we are still protected against outlying data. If all the means really are far apart, little harm will be done since with high probability the shrinkage will be close to unity. The main danger is that a minority of coordinates may be atypical of the rest. Apart from this possibility the shrunken estimators can at best greatly improve our estimation and at worst do only a little harm. Unfortunately the possibility of a small number of atypical components cannot be ignored. This is a criticism of ensemble loss functions: the James-Stein estimator does what is required of it - namely gives smaller risk than the usual estimator. In the alarming case above the majority of components have their components of risk reduced slightly at the expense of the minority which could have

unacceptably large components of the risk. A slight modification due to Efron and Morris (1979) seems to give us the best of both worlds. At small expense to the ensemble risk a rule which limits the amount of shrinkage allowed can reduce the individual component risks to nearly the same value as for the unshrunken estimator.

Critics of shrunken estimators argue that the methods suggest that we should join separate models together into one so that estimates in one case improve those for the others. The absurdity of doing so when the other problems are irrelevant to the problem in hand is self evident. In order to add weight to this criticism, many absurd suggestions of this sort have been made, for example, that baseball batting averages (or even random normal numbers) should be used to improve prediction of the effectiveness of a drug. Barnard asks: "Why should not all our estimation problems be combined into one grand melée?" In fact, if we do combine them, then our estimators will hardly differ from the maximum likelihood estimator and the risk will be smaller by a negligible amount.

It is only in problems in which most of the components have means which are close together that we obtain a useful reduction in the risk and only where we believe this to be likely should we use this method. In order to protect ourselves against one component risk being increased we should use a limited shrinkage rule. After all, a patient is interested in *his* diagnosis and the risk to *him* and has less concern for the risk to other patients who happen to have been examined on the same day.

Later we shall see that a prior distribution which keeps some component estimators independent of some of the others, in other words we are not combining the estimation problems, can give an even greater reduction in ensemble risk than the crude shrunken estimators.

Chapter 2

Modified James-Stein Estimators Applied to Linear Models

2.1 Introduction

In this chapter we shall discuss the canonical form for the linear model, which was set up in chapter 1, and extend the James-Stein estimator so that it does not necessarily shrink the maximum likelihood estimator towards the origin. In the discussion to Stein(1962), Lindley suggested shrinking towards the common mean of the coordinates while Stein(1966) suggested shrinking some components towards one value and others towads different values. We shall consider estimators which shrink the usual estimator towards several orthogonal hyperplanes thus generalising both of these suggestions. We note, however, that Stein(1955) clearly had this in mind for applications of his ideas.

Having developed these estimators, we shall show how they may be applied to the linear model. Both the full rank model and the non-full rank model will be considered as well as restricted linear models.

2.2 Shrinkage of the Maximum Likelihood Estimator Towards a Hyperplane

Suppose $X \sim N_p(\mu, \sigma^2 I)$ and $S \sim \frac{1}{n}\sigma^2\chi^2_n$ (if σ^2 is known then we put $n = \infty$ and $S = \sigma^2$). We shall estimate μ under the loss function $L(\hat{\mu}, \mu, \sigma^2) = \|\hat{\mu} - \mu\|^2/\sigma^2$. We wish to shrink the maximum likelihood estimator, X, towards the hyperplane $H\mu^* = h$ where rank $H = r$.

We may write X as the sum of three terms as follows

$$X = Gh + (I - GH)X + G(HX - h)$$

where $HGH = H$ and $GH = (GH)^T$.

Then Gh is the projection of the origin onto the solution space of $H\mu^* = h$ (i.e. the hyperplane towards which we are shrinking), $(I - GH)X$ is the projection of X onto the null space of H (i.e. the parallel hyperplane $H\mu^* = 0$) and $G(HX - h)$ is the projection of $X - Gh$ onto the column space of H^T (i.e. the orthogonal complement of the null space of H).

The idea is to shrink the component of $X - Gh$ orthogonal to the null space of H by an amount dependent on X and S without changing the component of X in the null space of H. This suggests using the estimator

$$\breve{\mu} = Y + \left(1 - \frac{cS}{\|Z\|^2}\right)Z \quad \text{where} \quad Y = Gh + (I - GH)X \quad \text{and} \quad Z = G(HX - h).$$

Let $E[Y] = \eta = Gh + (I - GH)\mu$, $E[Z] = \zeta = G(H\mu - h)$ so that, since

$X = Y + Z$, $\mu = \eta + \zeta$. Figure 1 shows the relationships among the variables and parameters defined above.

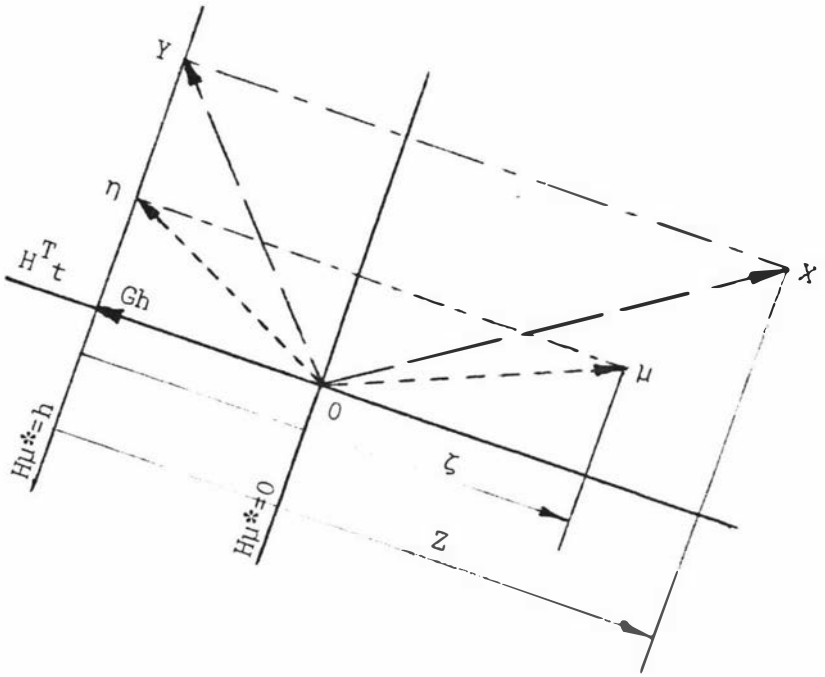

Figure 1 Variables Defined in Relation to the Maximum Likelihood Estimator

We shall show that the vectors $(I - GH)X$ and $G(HX - h)$ are indeed orthogonal (in the geometrical sense, i.e. with the inner product $\langle a , b \rangle_A = a^T A b$ where A is symmetric and positive definite, a and b are orthogonal if and only if $\langle a , b \rangle_A = 0$; We take $A = I$ in this case). We wish to show that $(I - GH)X^T G(HX - h) = 0$. Now, if $h = Hu$ then the left hand side is

$$\begin{aligned} X^T(I - GH)^T GH(X - u) &= X^T(I - GH)GH(X - u) \\ &= X^T(GH - GHGH)(X - u) \\ &= 0 \end{aligned}$$

which is the required result.

We shall also show that Y and Z are orthogonal in the sense of being statistically independent. We have

$$\begin{aligned} E[(Y - \eta)(Z - \zeta)^T] &= E[\{(I - GH)(X - \mu)\}\{GH(X - \mu)\}^T] \\ &= (I - GH)\,\mathrm{var}\,X\,(GH)^T \\ &= \sigma^2(I - GH)(GH)^T \\ &= \sigma^2(I - GH)(GH) \\ &= 0 . \end{aligned}$$

Finally we wish to find the distributions of the norms of the random variables defined above. We have $\| X \|^2 \sim \sigma^2 \chi_p^2(\tfrac{1}{2} \| \mu \|^2)$

since $\|X\|^2$ is the sum of squares of independent normal variates.
Also $\|Z\|^2 = Z^TZ = [GH(X-u)]^T[GH(X-u)]$
$= (X-u)^T(GH)^TGH(X-u)$
$= (X-u)^TGH(X-u)$

and since GH is idempotent $\|Z\|^2 \sim \sigma^2\chi^2_r(\tfrac{1}{2}\|\zeta\|^2)$ where r = rank H = rank GH . Similarly,

$$\|Y-\eta\|^2 = \{(I-GH)(X-\mu)\}^T\{(I-GH)(X-\mu)\}$$
$$= (X-\mu)^T(I-GH)^T(I-GH)(X-\mu)$$
$$= (X-\mu)^T(I-GH)(X-\mu)$$

and since $I-GH$ is idempotent of rank $p-r$, $\|Y-\eta\|^2 \sim \sigma^2\chi^2_{p-r}$. Also, since $(I-GH)GH = 0$, or since Y and Z are independent, $\|Y-\eta\|^2$ and $\|Z\|^2$ are independent.

In this chapter the only use we make of the risk function is as a motivation for using the estimator $\breve{\mu}$ and its generalisations. The proof of the formula for the risk function of the ordinary James-Stein estimator will therefore not be given here but will be delayed until chapter 6 where it will be given as a special case of the risk of a more general class of estimators. Here we shall quote the result. The risk function for the estimator $\breve{\mu}_1 = \left(1 - \frac{cS}{\|X\|^2}\right)X$ given in Stein(1966) is

$$E\left[\left\|\left(1 - \frac{cS}{\|X\|^2}\right)X - \mu\right\|^2\right] = p - c\{2(p-2) - \frac{n+2}{n}c\}E\left[\frac{1}{p-2+2K}\right]$$

where K has a Poisson distribution with parameter $\tfrac{1}{2}\sigma^{-2}\|\mu\|^2$. This is the result already given in chapter 1. We now give a slight generalisation - if X has a singular normal distribution then there is a matrix, L, such that LX has a normal distribution of the same rank and with $LL^T = I$, therefore replacing X by LX does not change the risk function - the risk is the same in the singular case.

Now, for the estimator $\breve{\mu}$ we have the risk function

$$E[\|Y + \phi(\|Z\|^2, S)Z - \mu\|^2]$$
$$= E[\|Y - \eta + \phi(\|Z\|^2, S)Z - \zeta\|^2]$$
$$= E[\|Y-\eta\|^2] + 2E[Y-\eta]^TE[\phi(\|Z\|^2, S)Z - \zeta] + E[\|\phi(\|Z\|^2, S)Z - \zeta\|^2]$$
$$= E[\|Y-\eta\|^2] + E[\|\phi(\|Z\|^2, S)Z - \zeta\|^2]$$

where $\phi(\|Z\|^2, S) = \left(1 - \frac{cS}{\|Z\|^2}\right)Z$ (the result is also true for general ϕ). The risk function is thus

$$R(\breve{\mu}, \mu, \sigma^2) = p - r + \left\{r - c\left[2(r-2) + \frac{n+2}{n}c\right]E\left[\frac{1}{r-2+2K}\right]\right\}$$

where K has a Poisson distribution with parameter $\frac{1}{2}\sigma^{-2} \| \zeta \|^2 = \frac{1}{2}\sigma^{-2} \| GH\mu - Gh \|^2 = \frac{1}{2}\sigma^{-2} \| G(H\mu - h) \|^2$. Thus we have

$$R(\tilde{\mu}, \mu, \sigma^2) = p - c\left\{2(r-2) - \frac{n+2}{n} c\right\} E\left[\frac{1}{r-2+2K}\right].$$

The minimum value of this is less than p (the risk for the maximum likelihood estimator) and occurs when $c = n\frac{r-2}{n+2}$ so long as $r \geq 3$, i.e. $p \geq s + 3$ where $s = p - r$.

The estimator just developed has the property that for small values of F the signs of the components of X are reversed. Intuitively this would seem to be a bad property, and, as we shall see in chapter 6, it is possible to obtain a slight uniform reduction in risk by using the positive part shrinkage in which negative values of the shrinkage factor are replaced by zero. By doing so, the saving in risk near $\mu = 0$ is quite marked. Using this shrinkage we find that there is no longer a uniformly best value of c, but that for $c \leq 2n\frac{r-2}{n+2}$ the estimator remains minimax. James and Stein suggested keeping to the value of c which is optimal for the ordinary James-Stein estimator, while Efron and Morris(1973a,1976) suggested the value $c = \min(2n\frac{r-2}{n+2}, n\frac{r-0.66}{n-0.66})$ which, when the second value is the minimum, gives an approximate 50% F value if, in fact, $\mu = 0$. The second value is the minimum when p and n take the values given in table 1. The resulting estimator is therefore a smoothed version of a preliminary test estimator. Its risk, while not uniformly less than that of the ordinary James-Stein estimator, can do much better and is never much worse.

Table 1 Values of $r \geq 3$ and $n \geq 1$ for which $\frac{r-0.66}{n-0.66} \geq 2\frac{r-2}{n+2}$

r	3	4	4 - 5	4 - 6 *	4 - 7	4 - 13	≥ 3
n	≥ 1	4 - 14	4 - 7	4 - 6 *	4 - 5	4	1 - 3

* *equality if $r = 6$ and $n = 6$*

Using the program described in chapter 5, the risk functions (as functions of $\lambda = \frac{1}{2}\sigma^{-2} \| \mu \|^2$) of various estimators of the James-Stein type and their positive part versions were computed. Plots of the difference in risk between each estimator and the ordinary James-Stein estimator, and of the difference in risk between the same estimator and the Efron and Morris estimator are shown in figures 2 to 4. In these graphs the curves have been labelled according to the point where they cut the risk axis. The abbreviation JSα refers to

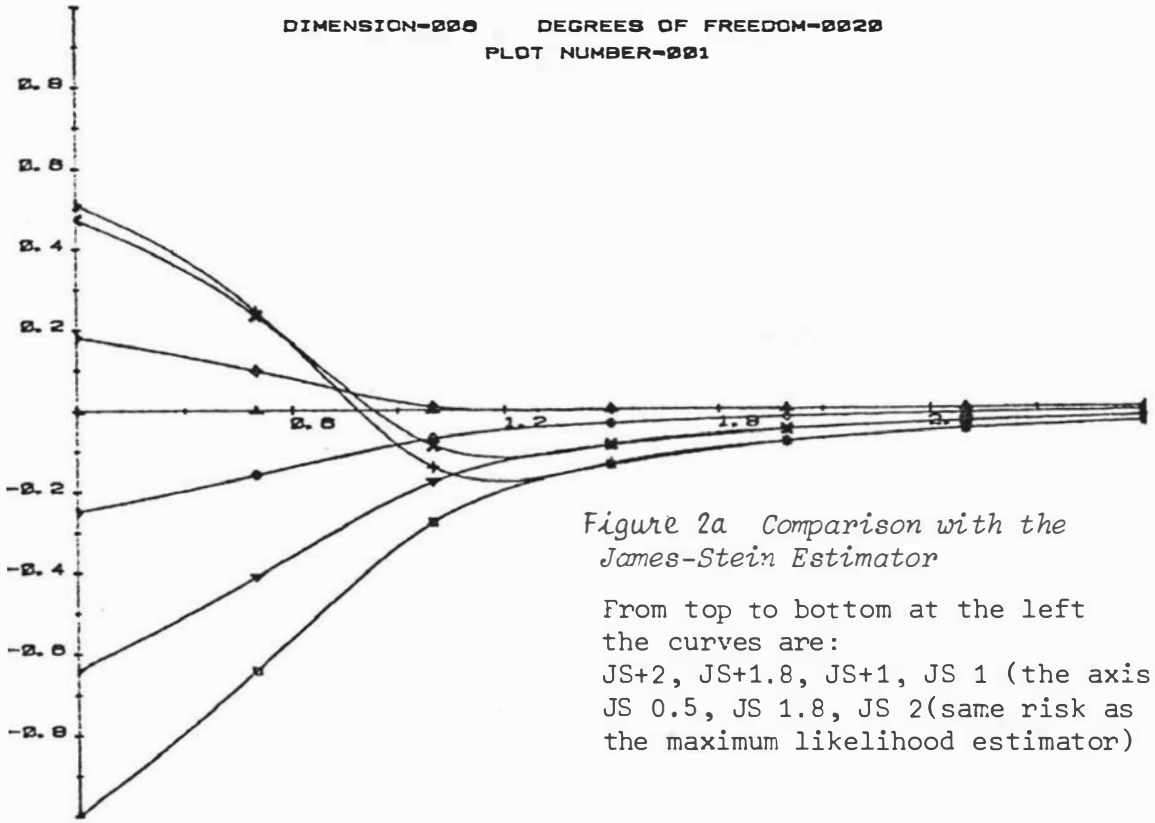

Figure 2a Comparison with the James-Stein Estimator

From top to bottom at the left the curves are:
JS+2, JS+1.8, JS+1, JS 1 (the axis)
JS 0.5, JS 1.8, JS 2(same risk as the maximum likelihood estimator)

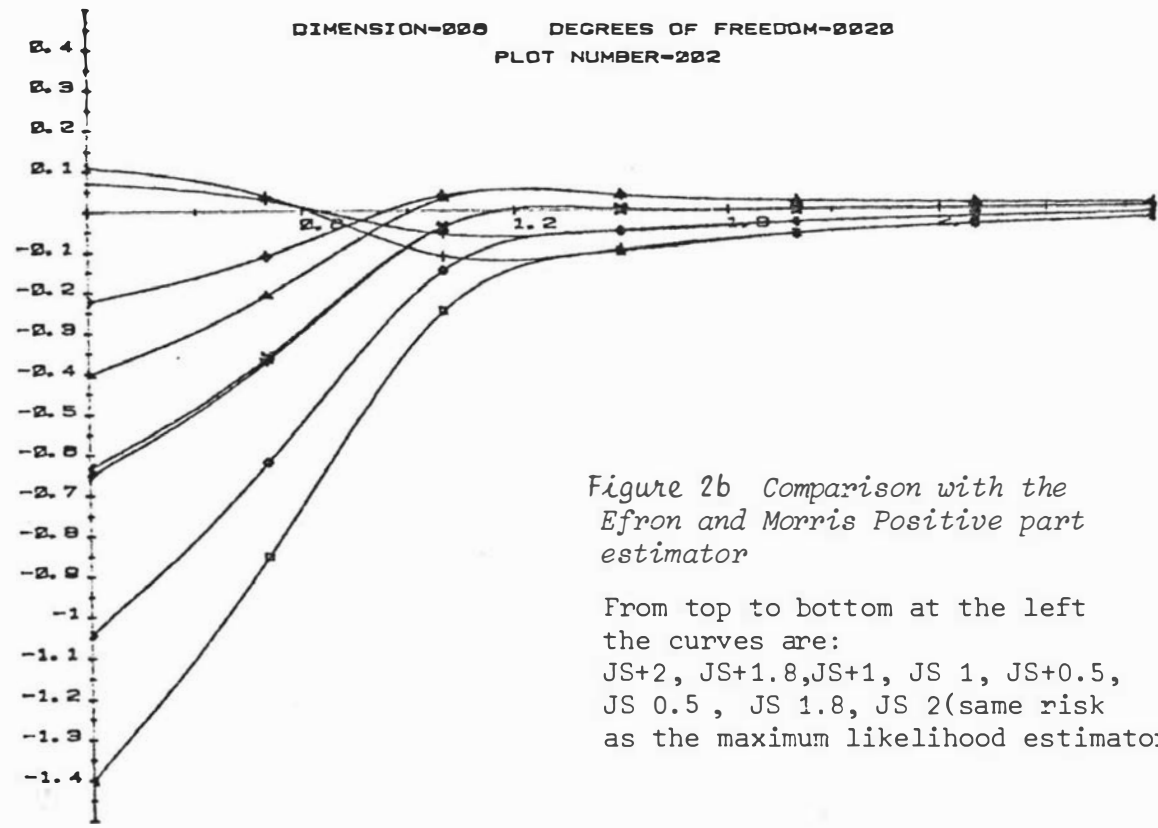

Figure 2b Comparison with the Efron and Morris Positive part estimator

From top to bottom at the left the curves are:
JS+2, JS+1.8,JS+1, JS 1, JS+0.5, JS 0.5 , JS 1.8, JS 2(same risk as the maximum likelihood estimator

Figure 2 Comparison of the Risk Functions of Stein-like Estimators

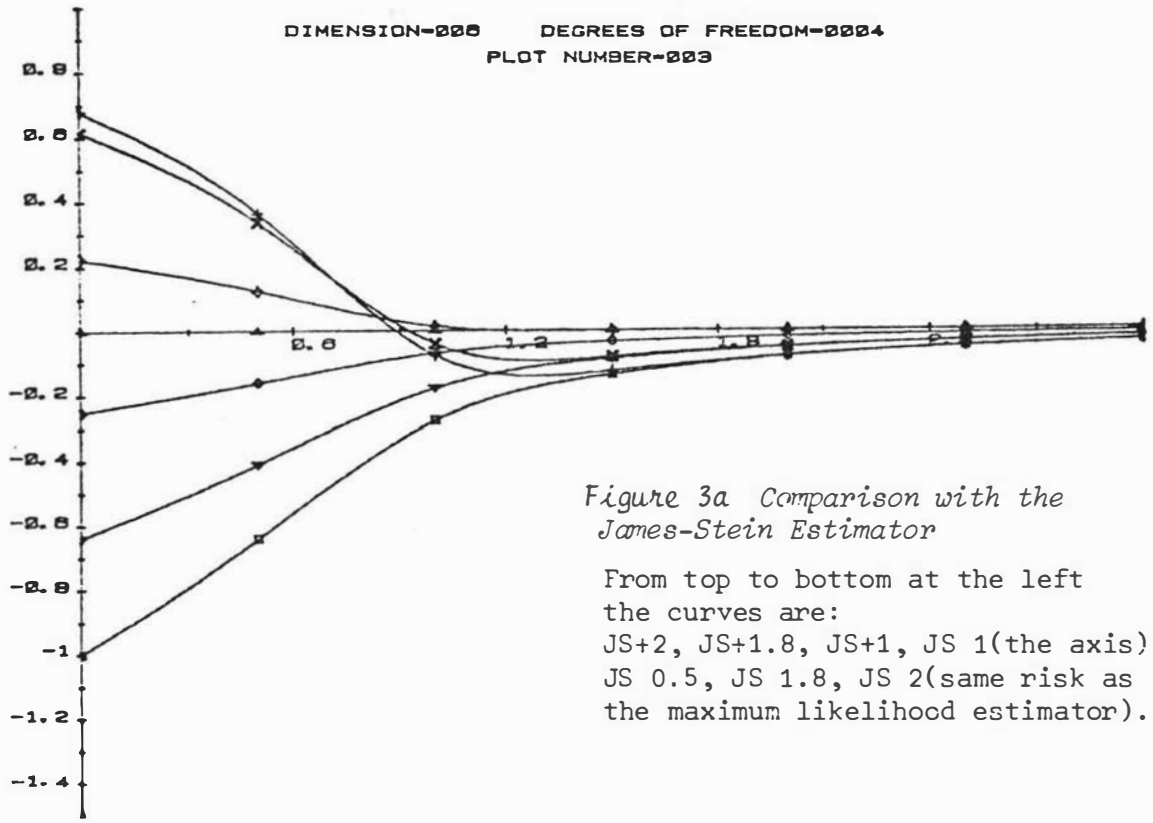

Figure 3a Comparison with the James-Stein Estimator

From top to bottom at the left the curves are:
JS+2, JS+1.8, JS+1, JS 1(the axis) JS 0.5, JS 1.8, JS 2(same risk as the maximum likelihood estimator).

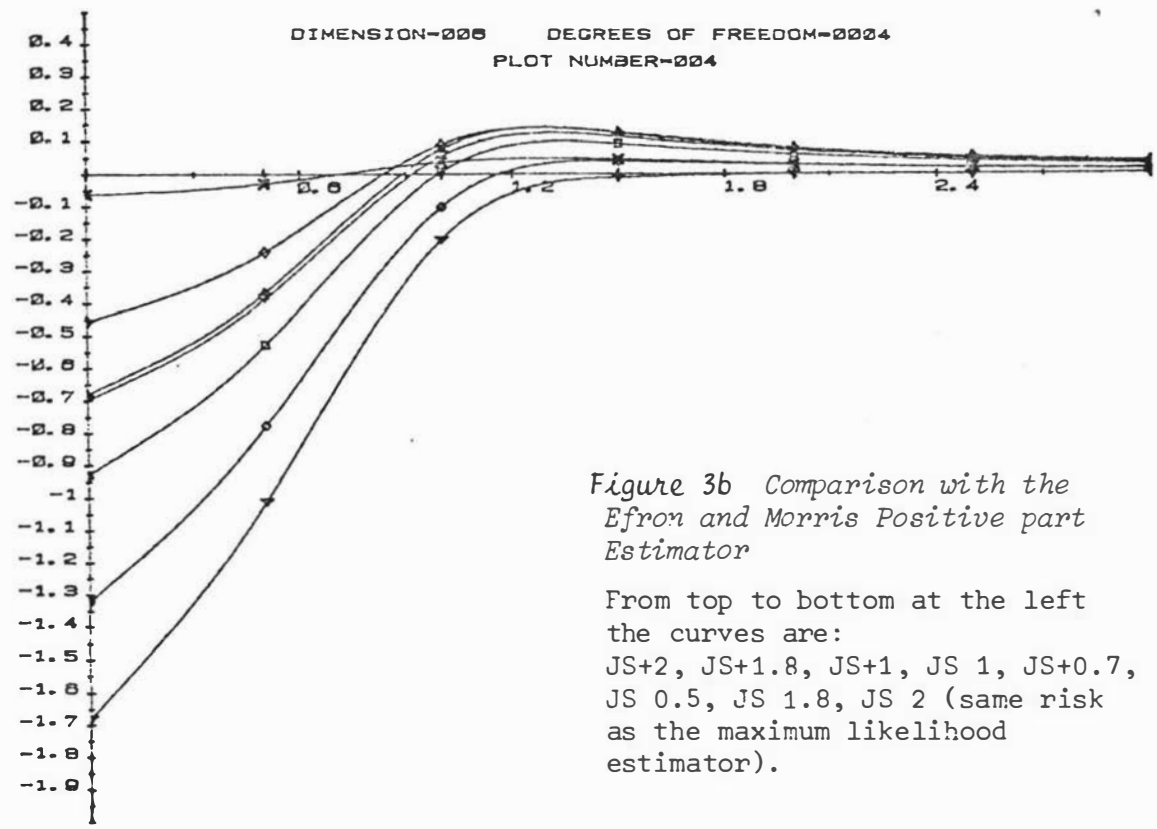

Figure 3b Comparison with the Efron and Morris Positive part Estimator

From top to bottom at the left the curves are:
JS+2, JS+1.8, JS+1, JS 1, JS+0.7, JS 0.5, JS 1.8, JS 2 (same risk as the maximum likelihood estimator).

Figure 3 Comparison of the Risk Functions of Stein-like Estimators

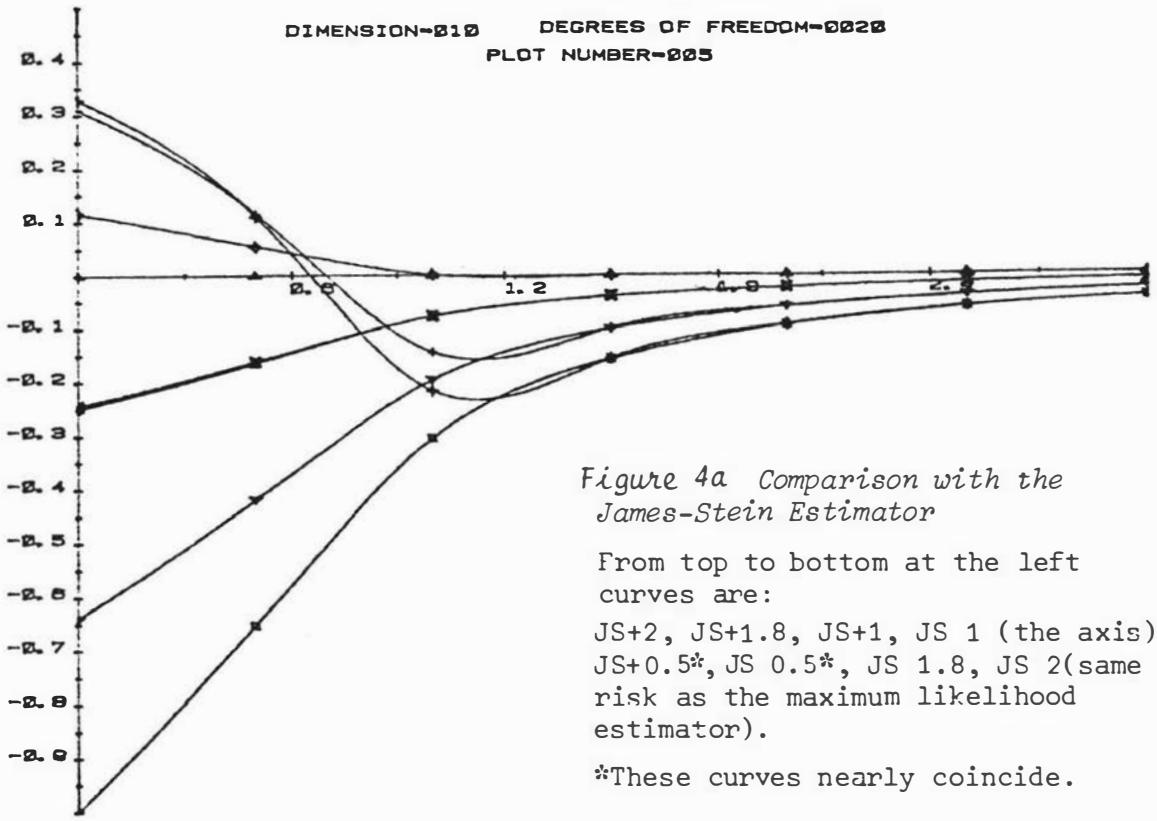

Figure 4a Comparison with the James-Stein Estimator

From top to bottom at the left curves are:
JS+2, JS+1.8, JS+1, JS 1 (the axis) JS+0.5*, JS 0.5*, JS 1.8, JS 2(same risk as the maximum likelihood estimator).

*These curves nearly coincide.

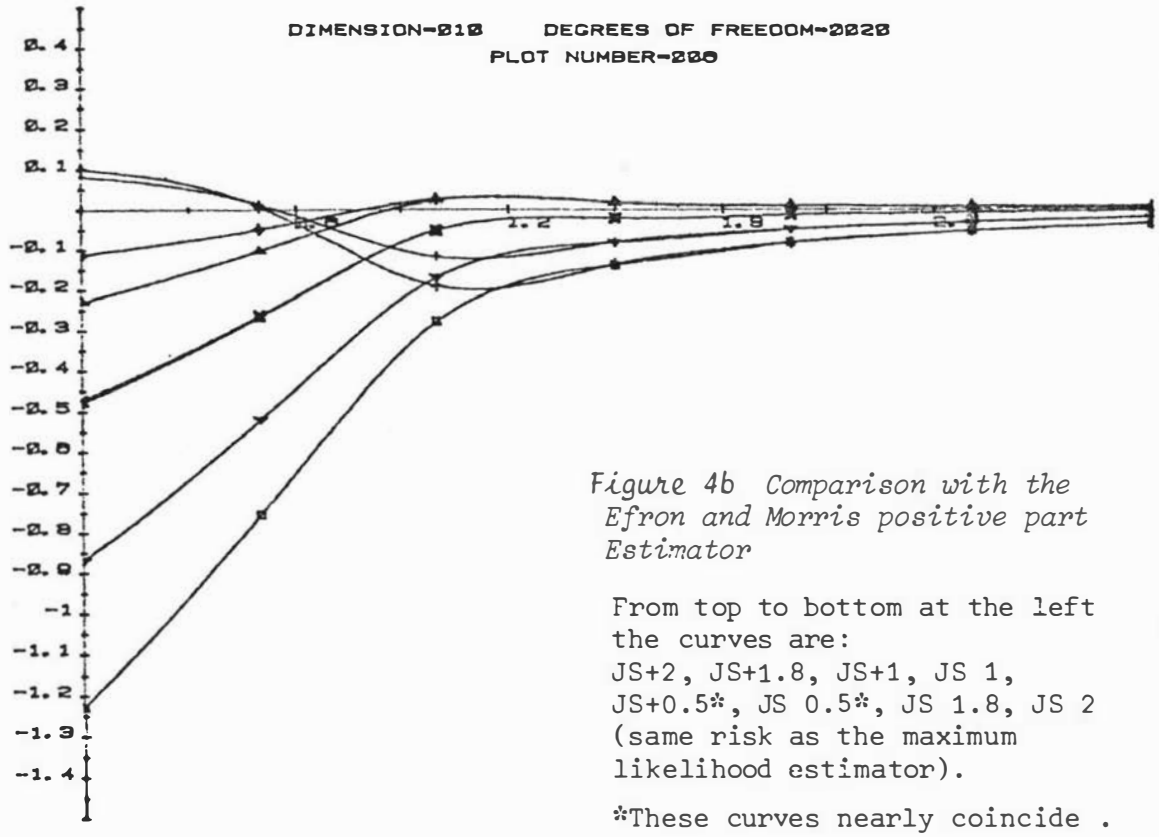

Figure 4b Comparison with the Efron and Morris positive part Estimator

From top to bottom at the left the curves are:
JS+2, JS+1.8, JS+1, JS 1, JS+0.5*, JS 0.5*, JS 1.8, JS 2 (same risk as the maximum likelihood estimator).

*These curves nearly coincide .

Figure 4 Comparison of the Risk Functions of Stein-like Estimators

$\delta(X, S) = \left(1 - \frac{cS}{\|X\|^2}\right) X$ of which the James-Stein estimator is a special case. For the positive part version of this estimator we have written $JS+\alpha$. The value c is the value recommended by James and Stein so for the James-Stein estimator we put $\alpha = 1$. Although these result are well known, the author has not seen the graphs plotted elsewhere.

Since the curves were smoothed using cubic splines, the graphs tend to have too steep a gradient at $\mu = 0$, the theoretical gradient being zero. The fit elsewhere is good - that it is not perfect is shown by the crossing of graphs for the positive part and corresponding non-positive part estimators: mid way between the points of intersection the difference is only about 0.01.

2.2.1 Special Cases

If we put $h = 0$ and take H to have full column rank then the null space of H and the hyperplane $H\mu^* = h$ reduce to a single point (have dimension zero). In this case the estimator $\breve{\mu}$ is the James-Stein estimator and the shrinkage is towards the origin. Stein suggested choosing the origin at the best prior estimate for each coordinate. Alternatively, without essentially changing the estimator, we may choose h such that Gh is the best prior estimate of the mean. In particular, taking $H = I_p$ and $G = I_p$, h will be the prior estimate for μ. Here $r = p$ and the hypothesis is $\mu_1 = \mu_2 = \ldots = \mu_p = \mu_0$.

If we do not wish to make p prior estimates then we may follow the suggestion of Lindley and let the data choose the origin towards which all coordinates are to be shrunk. We do this by putting $h = 0$ and

$$H = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{bmatrix}$$

in which case our prior hypothesis is $\mu_1 = \mu_2 = \ldots = \mu_p = \mu_0$, μ_0 unknown, and the estimator shrinks each coordinate towards the common mean. In other words X is shrunk towads the line $\mu^* = [1,1,\ldots,1]^T\alpha$. In this case $r = p - 1$ and we have a reduction in risk if $p \geq 4$. If $p = 3$ then the estimator gives no reduction in risk over the maximum likelihood estimator (in fact it is the maximum likelihood estimator) while the ordinary James-Stein estimator does. However, when $p \geq 4$ neither this estimator nor the James-Stein estimator is uniformly the

more efficient, The James-Stein estimator being more efficient if μ_0 is close to the true mean and less efficient if it is not.

Taking $h \neq 0$ gives a hypothesis of the form $\mu_i = \alpha_i + \mu_0$, with μ_0 unknown and the α_i known, towards which we shrink the maximum likelihood estimator. Since H has full row rank every generalised inverse is a right inverse satisfying $GH = (GH)^T$ which means that it is the unique Penrose inverse. The Penrose inverse is

$$G = \frac{1}{p}\begin{bmatrix} p-1 & p-2 & \cdots & 2 & 1 \\ -1 & p-2 & \cdots & 2 & 1 \\ -1 & -2 & \cdots & 2 & 1 \\ \vdots & \vdots & & & \vdots \\ -1 & -2 & \cdots & -(p-2) & 1 \\ -1 & -2 & \cdots & -(p-2) & -(p-1) \end{bmatrix}$$

and taking $h = [\alpha_1-\alpha_2, \alpha_2-\alpha_3, \ldots, \alpha_{p-1}-\alpha_p]^T$ ensures that our estimator shrinks the maximum likelihood estimator towards the required line.

Alternatively, we may obtain the same estimator with

$$H = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \qquad G = \frac{1}{p}\begin{bmatrix} p-1 & -1 & -1 & \cdots & -1 \\ -1 & p-1 & -1 & \cdots & -1 \\ -1 & -1 & p-1 & \cdots & -1 \\ \vdots & \vdots & \vdots & & \vdots \\ -1 & -1 & -1 & \cdots & p-1 \\ -1 & -1 & -1 & \cdots & -1 \end{bmatrix}$$

and $h = [\alpha_1-\alpha_p, \alpha_2-\alpha_p, \ldots, \alpha_{p-1}-\alpha_p]^T$

or with

$$H = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & -2 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & -3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & 1 & -p \end{bmatrix} \qquad G = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{12} & \cdots & \frac{1}{p(p+1)} \\ -\frac{1}{2} & \frac{1}{6} & \frac{1}{12} & \cdots & \frac{1}{p(p+1)} \\ 0 & -\frac{1}{3} & \frac{1}{12} & \cdots & \frac{1}{p(p+1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{p}{p(p+1)} \end{bmatrix}$$

and $h = [\alpha_1-\alpha_2, \alpha_1+\alpha_2-2\alpha_3, \ldots, \alpha_1+\alpha_2+\ldots+\alpha_{p-1}-(p-1)\alpha_p]^T$.

Whichever form we take for H, $GH = I - \frac{1}{p} 1 1^T$ and

$$Gh = \alpha - \bar{\alpha} = [\alpha_1, \ldots, \alpha_p]^T - \frac{1}{p}\sum_{i=1}^{p} \alpha_i \, 1 .$$

If H splits into several mutually orthogonal submatrices $H = [H_1^T, H_2^T, \ldots, H_t^T]^T$ with $H_i H_j = 0$ for $i \neq j$ then the shrinkage can be divided into components orthogonal to each of the null spaces $H_i\mu^* = 0$. Figure 5 represents the hyperplane in which Z lies, the axes representing the null spaces of H_1 and H_2. (In this diagram the hyperplane $H\mu^* = h$ is represented as a single point).

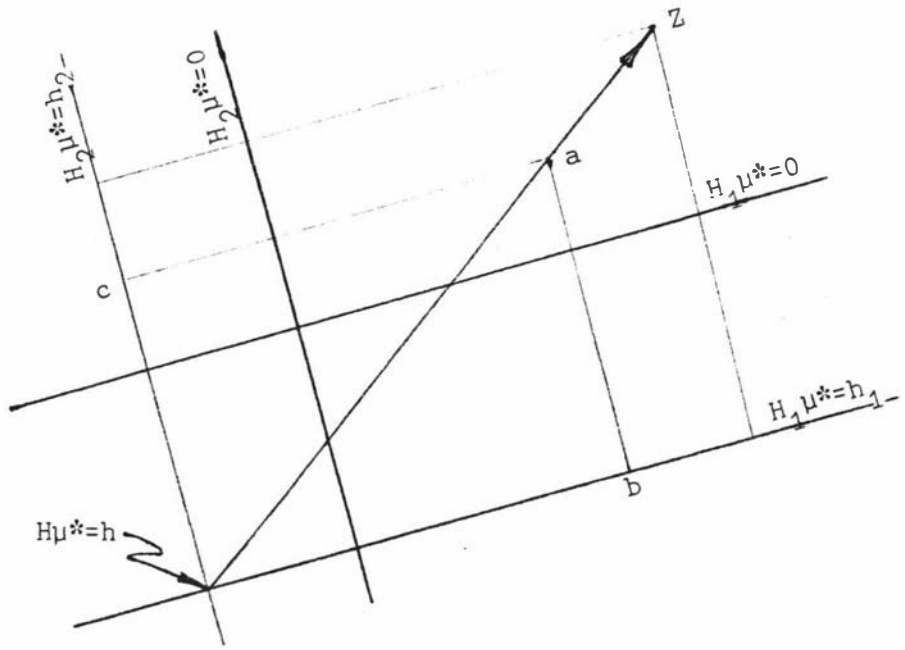

Figure 5 Shrinkage of the Maximum Likelihood Estimator towards a Hyperplane

The points in the diagram labelled a, b and c are respectively the vector towards which Z shrinks and the components of this vector in each of the hyperplanes $H_1\mu^* = h_1$ and $H_2\mu^* = h_2$.

2.2.2 Generalised Shrunken Estimators

The last special case considered suggests a possible generalisation. The shrinkage towards each of the hyperplanes $H_i\mu^* = h_i$ does not need to involve the same shrinkage factor in each case. The amount of shrinkage in the case considered is proportional to the weight of evidence in the sample for the hypothesis $H\mu = h$ (this weight of evidence being measured by $\frac{n}{pF} = \frac{S}{\|Z\|^2}$). In the case $t = 2$, if Z is close to the hyperplane $H_1\mu^* = h_1$ but far from the hyperplane $H_2\mu^* = 0$ then the hypothesis $H\mu^* = h$ is likely to have little evidence in its favour and the shrinkage factor will be small. On the other hand the data provides evidence that $H_1\mu^* = h_1$ and we might prefer a larger shrinkage factor towards this hyperplane and a smaller

shrinkage factor towards the hyperplane $H_2\mu^* = h_2$ (which has even greater evidence in its favour than has the hypothesis $H\mu^* = h$).

In the more general case in which $t \geq 2$ we shall split Z into orthogonal components

$$Z = Z_1 + Z_2 + \ldots + Z_t$$

where $Z_i = G_i(H_iX - h_i)$, $H_iH_j^T = 0$ for $i \neq j$, $H_iG_iH_i = H_i$, $G_iH_i = (G_iH_i)^T$, $H^T = [H_1^T,\ldots,H_t^T]$, $h^T = [h_1^T,\ldots,h_t^T]$ and $G = [G_1,\ldots,G_t]$.

We first show that $HGH = H$ and $GH = (GH)^T$. We have

$$GH = G_1H_1 + \ldots + G_tH_t$$

and each component on the right is symmetric.

Now $$HGH = \begin{bmatrix} H_1G_1H_1 + \sum_{i\neq 1} H_1G_iH_i \\ \vdots \\ H_tG_tH_t + \sum_{i\neq t} H_tG_iH_i \end{bmatrix}$$

but $H_jG_iH_i = H_j(G_iH_i)^T = H_jH_i^TG_i = 0$

and $H_iG_iH_i = H_i$.

Thus $HGH = H$.

We now show that the Z_i are mutually orthogonal. We have

$$\begin{aligned} Z_i^TZ_j &= (H_iX - h_i)^TG_i^TG_j(H_jX - h_j) \\ &= (X - u_i)^TH_i^TG_i^TG_jH_j(X - u_j) \end{aligned}$$

where $h_i = H_iu_i$ for each i.

Since each G_iH_i is symmetric

$$\begin{aligned} Z_i^TZ_j &= (X - u_i)^TG_iH_iH_j^TG_j^T(X - u_j) \\ &= 0. \end{aligned}$$

We may thus split X as

$$X = Gh + (I - GH)X + \sum_{i=1}^{t} G_i(H_iX - h_i)$$

where the terms after the summation sign are mutually orthogonal. We shall show that they are also orthogonal to $(I - GH)X$. We have

$$X^T(I - GH)^TG_i(H_iX - h_i) = X^T(I - GH)G_iH_i(X - u_i).$$

Now $$GHG_iH_i = \sum_{j=1}^{t} G_jH_jG_iH_i = \sum_{j=1}^{t} G_jH_jH_i^TG_i^T$$

$$= G_iH_iH_i^TG_i^T = G_iH_iG_iH_i = G_iH_i.$$

Thus $(I - GH)G_iH_i = 0$ and $\{(I - GH)X\}^T G_i(H_iX - h_i) = 0$.

Applying the shrinkage to each component of Z gives the estimator

$$= Y + \sum_{i=1}^{t} \left(1 - \frac{c_i S}{\|Z_i\|^2}\right) Z_i \qquad \text{where} \qquad Y = Gh + (I - GH)X .$$

As before we shall let $E[Y] = \eta$, $E[Z] = \zeta$ and $\operatorname{rank} H = r$. Also, putting $E[Z_i] = \zeta_i$ and $\operatorname{rank} H_i = r_i$ we have $r = \sum_{i=1}^{t} r_i$.

We show that Y and the Z_i are all independent. Firstly,

$$\begin{aligned}
\operatorname{cov}(Z_i, Z_j) &= \operatorname{cov}(G_iH_iX, G_jH_jX) = G_iH_i \operatorname{var} X \, (G_jH_j)^T \\
&= \sigma^2 G_iH_iH_j^T G_j^T \\
&= 0 \quad \text{if and only if} \quad H_iH_j^T = 0 \quad \text{for } i \neq j . \quad \text{Secondly,} \\
\operatorname{cov}(Y, Z_i) &= \operatorname{cov}\{(I - GH)X, G_iH_iX\} = (I - GH) \operatorname{var} X \, (G_iH_i)^T \\
&= \sigma^2 (I - GH)(G_iH_i)^T \\
&= 0 \quad \text{if and only if} \quad H_iH_j^T = 0 \quad \text{for } i \neq j
\end{aligned}$$

As shown previously, $\|Y - \eta\|^2 \sim \sigma^2 \chi^2_{p-r}$ and $\|Z_i\|^2 \sim \sigma^2 \chi^2_{r_i}(\tfrac{1}{2}\|\zeta_i\|^2)$ the distributions being independent.

We now calculate the risk function for $\check{\mu}$. Noting that $E[Y - \eta] = 0$, $E[Z_i] = \zeta_i$ and that $Z_i^T Z_j = 0$ we have for $\phi_i(\|Z_i\|^2, S) = 1 - \dfrac{c_i S}{\|Z_i\|^2}$

$$\begin{aligned}
R(\check{\mu}, \mu, \sigma^2) &= E\Big[\Big\|Y + \sum_{i=1}^{t} \phi_i(\|Z_i\|^2, S) Z_i - \mu\Big\|^2\Big] \\
&= E\Big[\Big\|Y - \eta + \sum_{i=1}^{t} \{\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\}\Big\|^2\Big] \\
&= E\big[\|Y - \eta\|^2\big] + E\Big[\Big\| \sum_{i=1}^{t} \{\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\}\Big\|^2\Big] \\
&\quad + 2E[Y - \eta]^T E\Big[\sum_{i=1}^{t} \{\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\}\Big] \\
&= E\big[\|Y - \eta\|^2\big] + E\Big[\Big\| \sum_{i=1}^{t} \{\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\}\Big\|^2\Big] \\
&= E\big[\|Y - \eta\|^2\big] + \sum_{i=1}^{t} E\big[\|\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\|^2\big] \\
&\quad + \sum_{i \neq j} E[E[\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i \mid S]^T E[\phi_j(\|Z_j\|^2, S) Z_j - \zeta_j \mid S] \\
&= E\big[\|Y - \eta\|^2\big] + \sum_{i=1}^{t} E\big[\|\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\|^2\big] \\
&\quad + \sum_{i \neq j} E[E[\{\phi_i(\|Z_i\|^2, S) - 1\} Z_i \mid S]^T E[\{\phi_j(\|Z_j\|^2, S) - 1\} Z_j \mid S] \\
&= E\big[\|Y - \eta\|^2\big] + \sum_{i=1}^{t} E\big[\|\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\|^2\big] \\
&\quad + \sum_{i \neq j} E[\{\phi_i(\|Z_i\|^2, S) - 1\}\{\phi_j(\|Z_j\|^2, S) - 1\} Z_i^T Z_j] \\
&= E\big[\|Y - \eta\|^2\big] + \sum_{i=1}^{t} E\big[\|\phi_i(\|Z_i\|^2, S) Z_i - \zeta_i\|^2\big].
\end{aligned}$$

(This is also true for general ϕ_i).

Thus $R(\tilde{\mu},\mu,\sigma^2) = p - r + \sum_{i=1}^{t} \left\{ r_i - c_i\left[2(r_i-2) - \frac{n+2}{n} c_i\right] E\left[\frac{1}{r_i-2+2K_i}\right]\right\}$

where the K_i have Poisson distributions with parameter $\frac{1}{2}\|\zeta_i\|^2$.

Each term in the sum is minimised when $c_i = \frac{n}{n+2}(r_i - 2)$. With these values for the c_i the risk function becomes

$$R(\tilde{\mu},\mu,\sigma^2) = p - \frac{n}{n+2} \sum_{i=1}^{t} (r_i - 2)^2 E\left[\frac{1}{r_i-2+2K_i}\right].$$

For general values of the c_i the risk simplifies to

$$R(\tilde{\mu},\mu,\sigma^2) = p - \sum_{i=1}^{t} c_i\left\{2(r_i - 2) - \frac{n+2}{n} c_i\right\} E\left[\frac{1}{r_i-2+2K_i}\right].$$

Also, the orthogonality of the Z_i imply the orthogonality of the ζ_i and $\|\zeta\|^2 = \|\sum_{i=1}^{t} \zeta_i\|^2 = \sum_{i=1}^{t} \|\zeta_i\|^2 + \sum_{i\neq j} \zeta_i^T \zeta_j$. Therefore $\|\zeta\|^2 = \sum_{i=1}^{t} \|\zeta_i\|^2$.

2.2.3 Comparison of James-Stein Estimators

It might be imagined that, since shrunken estimators are not uniformly better than the maximum likelihood estimator when $p \leq 2$, but can be when $p \geq 3$, the larger the value of p the greater the gain in efficiency. Unfortunately it is difficult to find a meaningful basis of comparison between different models. However, the greatest possible reduction in risk occurs when $\|\zeta\| = 0$ and using the ordinary James-Stein estimator the risk is

$$R(\tilde{\mu},\mu,\sigma^2) = p - \frac{n}{n+2}(p-2) = \frac{2p}{n+2} + \frac{2n}{n+2}$$

$$= 2\,\frac{p+n}{n+2}.$$

Although this increases with p the reduction in risk over that for the maximum likelihood estimator is $\frac{n(p-2)}{n+2}$ and this also increases with p. A fairer basis of comparison is the risk relative to the risk for the maximum likelihood estimator and this is $2\,\frac{1+n/p}{n+2}$ which decreases with p.

The above argument seems to suggest that we should make the dimension of the hyperplane towards which we are shrinking as small as possible, but since we are comparing different estimators for the same model the argument does not apply. In fact, if μ lies on the hyperplane and we use the generalised James-Stein estimator then

$$R(\tilde{\mu},\mu,\sigma^2) = p - \sum_{i=1}^{t} (r_i - 2)\frac{n}{n+2}$$
$$= p - \frac{nr}{n+2} + \frac{2tn}{n+2}$$

and if $t = 1$ then we would like r to be as large as possible. Alternatively, for fixed r we would like t as small as possible. This argument favours the ordinary James-Stein estimator *if* μ *lies on the hyperplane towards which we are shrinking*. To gain maximum advantage from this estimator we have to make a good prior guess for μ. If our guess is poor then the estimator and its risk differ very little from the maximum likelihood estimator and the risk for that estimator.

On the other hand, taking $t > 1$ may make a saving in risk if our guess of some coordinates is good even though others have been guessed badly. This is despite the fact that the potential saving is not as great. Stein(1955) argued this way, an argument which seems sensible on intuitive grounds for any estimator which uses prior information. When the form of the estimator is given we no longer have to rely on intuitive arguments but may analyse the situation more precisely. This we shall proceed to do.

We shall write the risk for the generalised James-Stein estimator in a form which makes comparison with the ordinary James-Stein estimator easier. We have

$$R(\tilde{\mu},\mu,\sigma^2) = p - \frac{n}{n+2}\sum_{i=1}^{t}(r_i - 2)E\left[\frac{r_i - 2}{r_i - 2 + 2K_i}\right]$$
$$= p - \frac{n}{n+2}\sum_{i=1}^{t}(r_i - 2)E\left[1 - \frac{2K_i}{r_i - 2 + 2K_i}\right]$$
$$= p - \frac{n(r-2t)}{n+2} + \frac{n}{n+2}\sum_{i=1}^{t}(r_i - 2)E\left[\frac{2K_i}{r_i - 2 + 2K_i}\right]$$
$$= \frac{n}{n+2}(p-r) + \frac{2p}{n+2} + \frac{n}{n+2}2t + \frac{n}{n+2}E\left[\sum_{i=1}^{t}\frac{(r_i-2)\times 2K_i}{(r_i-2)+2K_i}\right].$$

Now $\frac{xy}{x+y}$ is a concave function of x and y when x and y are positive and therefore

$$\sum_{i=1}^{t}\frac{(r_i-2)\times 2K_i}{(r_i-2)+2K_i} \le \frac{(r-2t)\times 2K}{(r-2t)+2K} \le \frac{(r-2)\times 2K}{(r-2)+2K} \quad \text{where} \quad K = \sum_{i=1}^{t} K_i,$$

the latter inequality following from the fact that $\frac{xy}{x+y}$ is increasing in x. The equalities hold in the trivial case $t = 1$ and the first equality also holds, for $t \neq 1$, if and only if (i), for each i,

$K_i = 0$ or (ii) for each i, $r_i = 2$. This follows from the fact that $\frac{xy}{x+y}$ is strictly concave. When $t \neq 1$ the second equality only holds if, for each i, $K_i = 0$.

Now each K_i has a Poisson distribution with parameter $\frac{1}{2}\|\zeta_i\|^2$ and we may take them to be independent in which case K has a Poisson distribution with parameter $\frac{1}{2}\|\zeta\|^2$. Therefore

$$R(\breve{\mu},\mu,\sigma^2) \leq \frac{n}{n+2}\,2t + \frac{n}{n+2}\,(p-r) + \frac{n}{n+2}\,E\left[\frac{(r-2)\times 2K}{(r-2)+2K}\right]$$
$$= \frac{n}{n+2}\,(2t-2) + R_1(\breve{\mu}_1,\mu,\sigma^2)$$

where $R_1(\breve{\mu}_1,\mu,\sigma^2)$ is the risk function for the estimator with $t = 1$. Equality holds if $\|\zeta\|^2 = 0$.

This result is a more precise form of a result of Stein(1966) which he describes as a crude approximation valid when r and n are large (Stein took $r = p$). In fact it is not essential for his argument to take n to be large. Stein's approximation replaces $r_i - 2$ by r_i and $E\left[\frac{r_i \times 2K_i}{r_i + 2K_i}\right]$ by $\frac{r_i \times \|\zeta\|^2}{r_i + \|\zeta_i\|^2}$ then uses the same concavity argument to show that $\sum_{i=1}^{t} \frac{r_i \times \|\zeta_i\|^2}{r_i + \|\zeta_i\|^2} \leq \frac{r\,\|\zeta\|^2}{r + \|\zeta\|^2}$. Doing this ignores the term $\frac{n}{n+2}\,(2t-2)$ which is small compared with $\frac{n}{n+2}\,r$ when each r_i is large. Also the approximations to the expectations are good for large r_i only if $\|\zeta_i\|^2$ is small compred with r_i or large compared with r_i, in other words the approximation is good when $\frac{1}{r_i}\|\zeta_i\|^2$ is small or large, but not necessarily for intermediate values. This, therefore does not answer the question as to which estimator is the better at values of the ζ_i neither close to nor far from the origin. Stein's result does not appear to be very useful.

Our result, on the other hand, shows that

$$R(\breve{\mu},\mu,\sigma^2) - R_1(\breve{\mu}_1,\mu,\sigma^2) \leq \frac{n}{n+2}\,(2t-2)$$

which gives an upper bound on the amount by which $\breve{\mu}$ is worse than $\breve{\mu}_1$; but does not prove that it can ever be better.

Suppose that the average of the r_i is k, i.e. $k = \frac{r}{t}$. The maximum saving in risk over the maximum likelihood estimator is $\frac{n}{n+2}\,(r-2t) = \frac{n}{n+2}\,(k-2)t$. The ratio of this saving to the saving (for

the same value of r) when $t = 1$ is $\frac{(k-2)t}{r-2} = \frac{(k-2)t}{kt}\frac{r}{r-2} = \frac{k-2}{k}\frac{r}{r-2}$. The amount of reduction in savings as a proportion of the single savings is $\frac{(r-2)-(k-2)t}{r-2} = \frac{2(t-1)}{r-2} = \frac{2}{k}\frac{kt-k}{r-2} = \frac{2}{k}\frac{r-k}{r-2}$. This reduction in savings is less than a proportion $\frac{2}{k}$ of the single shrinkage savings. Thus, for $k = 3$, we lose less than $2/3$ of the possible savings, less than $1/2$ if $k = 4$, and, for $k = 6$, less than $1/3$.

We shall now show that there is a potential gain in efficiency. Suppose that a set $\mathfrak{A}$ of the i are such that ζ_i for $i \in \mathfrak{A}$ are close enough to the origin for $\frac{(r_i-2)\times\|\zeta_i\|^2}{(r_i-2)+\|\zeta_i\|^2}$ to be a good approximation to $E\left[\frac{(r_i-2)\times 2K_i}{(r_i-2)+2K_i}\right]$ and that for $i \notin \mathfrak{A}$, ζ_i is far enough from the origin for the same approximation to be good. (By A2.5.1, this is the first term of an asymptotic expansion far from the origin which also gives accurate results close to the origin). Now $\sum_{i=1}^{t}\|\zeta_i\|^2 = \|\zeta\|^2$ will be large in this case. Suppose it is large enough for the same approximation to be accurate. Let the number of elements of $\mathfrak{A}$ be u. We have

$$R(\check{\mu},\mu,\sigma^2) \doteqdot \frac{n}{n+2}(p-r) + \frac{2p}{n+2} + \frac{n}{n+2}2t + \frac{n}{n+2}\sum_{i=1}^{t}\frac{(r_i-2)\times\|\zeta_i\|^2}{(r_i-2)+\|\zeta_i\|^2}.$$

Using the fact that for $i \in \mathfrak{A}$, $\|\zeta_i\|^2 \doteqdot 0$ and for $i \notin \mathfrak{A}$, $\|\zeta_i\|^2$ is large gives

$$R(\check{\mu},\mu,\sigma^2) \doteqdot \frac{n}{n+2}(p-r) + \frac{2p}{n+2} + \frac{n}{n+2}2t + \frac{n}{n+2}\sum_{i\notin\mathfrak{A}}(r_i-2)$$

and if $u \neq t$ then $\|\zeta\|^2$ is large and

$$R_1(\check{\mu}_1,\mu,\sigma^2) \doteqdot \frac{n}{n+2}(p-r) + \frac{2p}{n+2} + \frac{n}{n+2}\times 2 + \frac{n}{n+2}(r-2) = p.$$

Thus

$$R(\check{\mu},\mu,\sigma^2) - R_1(\check{\mu}_1,\mu,\sigma^2) \doteqdot -\frac{n}{n+2}\sum_{i\in\mathfrak{A}}(r_i-2) \quad \text{for } u \neq t.$$

Since this is negative when each $r_i > 2$ this represents a gain in efficiency of at least $\frac{n}{n+2}$ for $u \neq t$, but a loss of efficiency of $\frac{n}{n+2}(2t-2)$ if $u = t$.

Note that corresponding to each of the terms in which $\|\zeta_i\|^2$ is large, the component of the risk is the same (to a high degree of accuracy) as that for the estimator which does no shrinking towards the corresponding hyperplane. Therefore we obtain approximately the

same reduction in risk by not shrinking those components. This shows why a Lindley type shrinkage can often be better than a Stein type shrinkage, for, if we split one of the hyperplanes into orthogonal hyperplanes, we can, as just shown, decrease the risk when one of the components of ζ_i is large and the other is small.

We shall present graphs of the risk function for the James-Stein estimator and three dimensional graphs of the difference in risk between the double and single shrinkage rules. The risk is scaled so that the maximum reduction over the maximum likelihood estimator is unity.

For the single shrinkage, the reduction in risk is given by the formula

$$\Delta R \propto E\left[\frac{r-2}{r-2+2K}\right] = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{r-2}{r-2+2k} = e^{-\lambda}\,{}_1F_1(\tfrac{1}{2}r-1;\tfrac{1}{2}r;\lambda)\,.$$

Using Euler's theorem in appendix 1 for the confluent hypergeometric function the latter expression is

$${}_1F_1(1;\tfrac{1}{2}r-1;-\lambda) = \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\frac{1}{(\tfrac{1}{2}r-1)_k}\,.$$

This formula is slightly more efficient than the other because it avoids evaluating $e^{-\lambda}$. It also gives an error estimate since the error is less than the first neglected term in the expansion. Both series require a large number of terms for large λ and this results in an exponent overflow condition in finite precision arithmetic. The second formula has the further disadvatage of causing severe loss of significant digits when there are far fewer terms than the number needed to cause exponent overflow. Accordingly we used the latter series only for $\lambda \le 20$: for larger values of λ we could use the asyptotic expansion (see appendix 1)

$${}_1F_1(1;\tfrac{1}{2}r-1;-\lambda) \sim \frac{\Gamma(\tfrac{1}{2}r-1)}{\Gamma(\tfrac{1}{2}r-2)}\,\lambda^{-1}\,{}_2F_0(3-\tfrac{1}{2}r,1;;1/\lambda)$$

$$= \sum_{k=0}\frac{[\tfrac{1}{2}r-3]_k}{(\tfrac{1}{2}r-1)(-\lambda)^{k+1}}$$

which terminates if r is an even integer greater than 4, but, since this can require the calculation of many terms if λ is smaller than $\frac{1}{2}r$, we preferred the asymptotic expansion A2.5.1 which generalises Stein(1966),

$$E\left[\frac{\alpha}{\alpha+K}\right] \sim \frac{\alpha}{\alpha+\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n\mu_n}{(\alpha+\lambda)^n}$$

where μ_n is the n th central moment of the Poisson distribution and

$\alpha = \frac{1}{2}r-1$. This was taken as far as the term corresponding to $n = 5$.

The reduction in risk was plotted for values of $r = 3, 4, 5, 8$ and ∞ - the latter value being interpreted as a limiting value - and for $\phi = \sqrt{\lambda/r}$ from zero to five. The closer the curve is to the origin, the smaller is the value of r (called "degrees of freedom" in the heading to the graph). The graph is shown in figure 6.

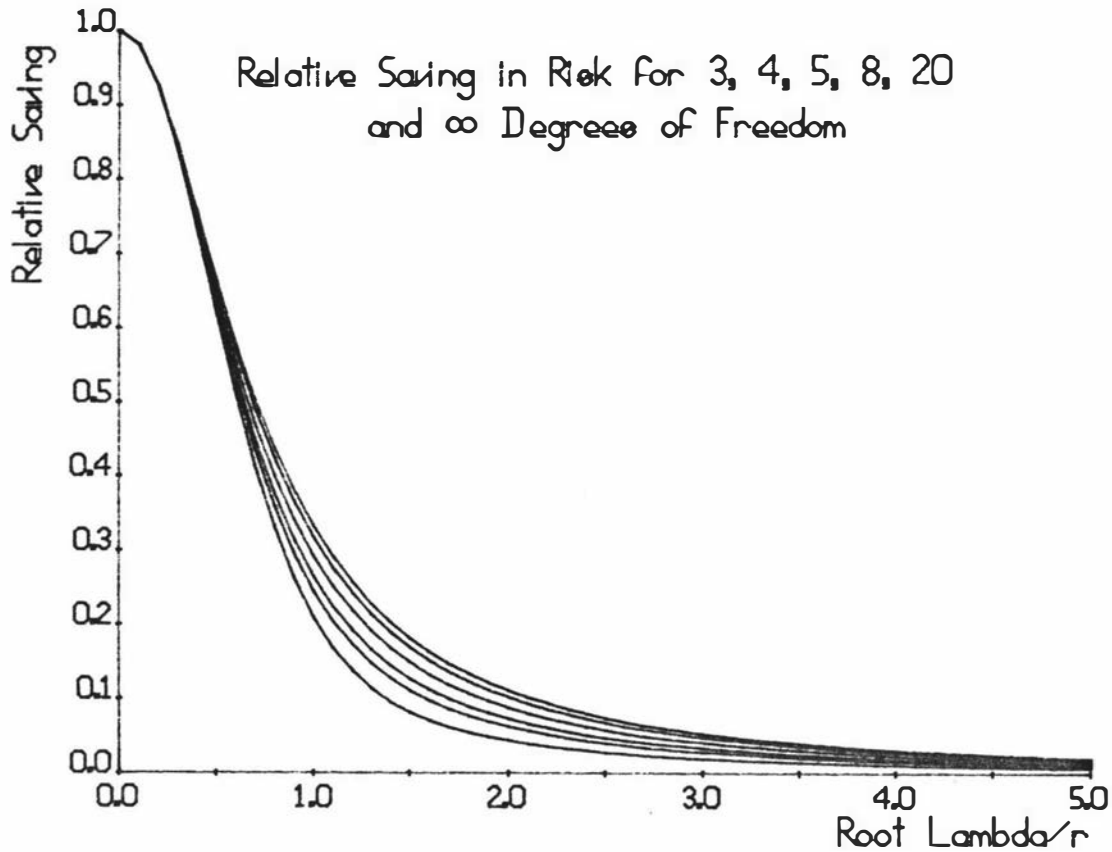

Figure 6 Saving in Risk for James Stein Estimators in Comparison with the Maximum Likelihood Estimator

For the difference in risk between the double and single shrinkages, $\phi_1 = \sqrt{\lambda_1/r_1}$ and $\phi_2 = \sqrt{\lambda_2/r_2}$ were each evaluated from zero to five. In order to show the region of improvement, a contour map of the surface was plotted with the zero contour clearly indicated. Contour spacings were such that either 10 or 20 contour lines were drawn between zero and the highest point.

By putting $\Delta R_i = 0$ if $r_i = 1$ or $r_i = 2$, we can consider the case of the Lindley type shrinkage. It is clear that in general the graph has two "wings" near the axes whose height increase with increasing r_i. If $r_1 = 1$ then the corresponding wing disappears. The central region over which the single shrinkage is best, although large, only has

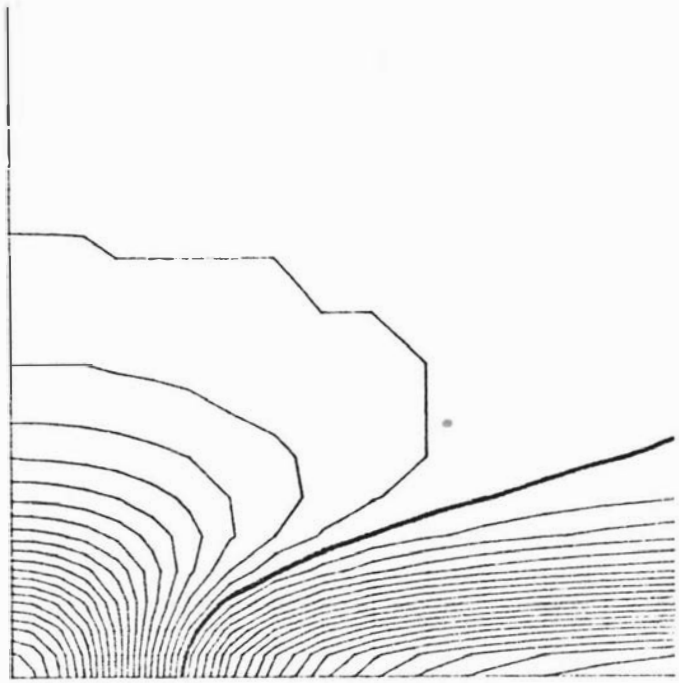

Figure 7a *Contour Map*

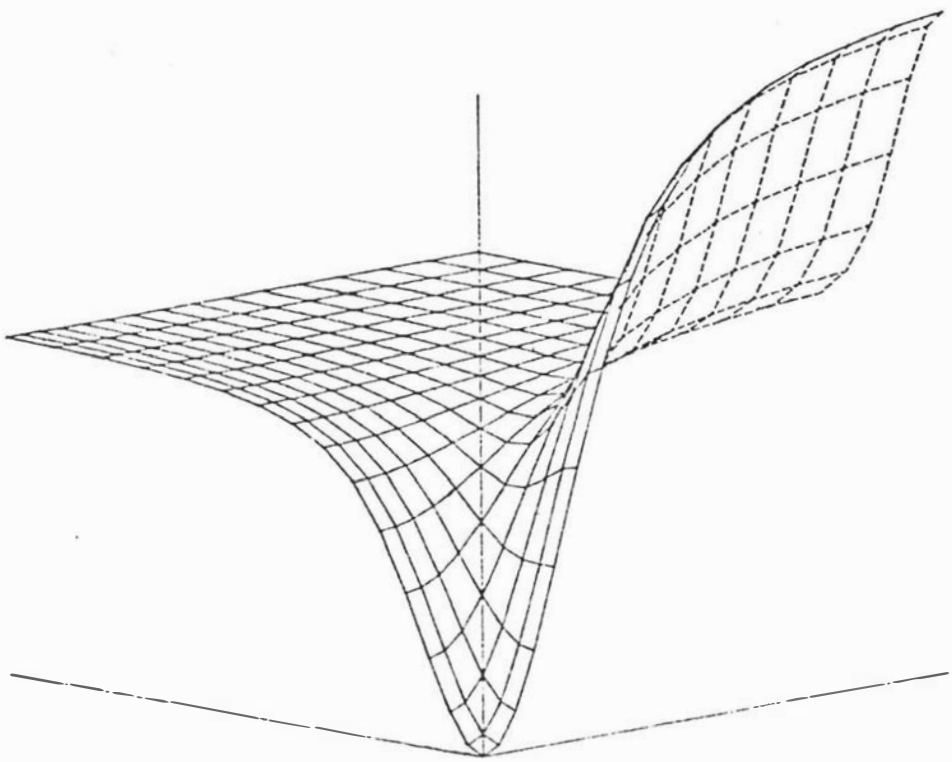

Figure 7b *Pictorial View*

Figure 7 *Difference in Risk Between Separate Shrinkage and Combined Shrinkage Estimators (dimensions 1 and 3)*

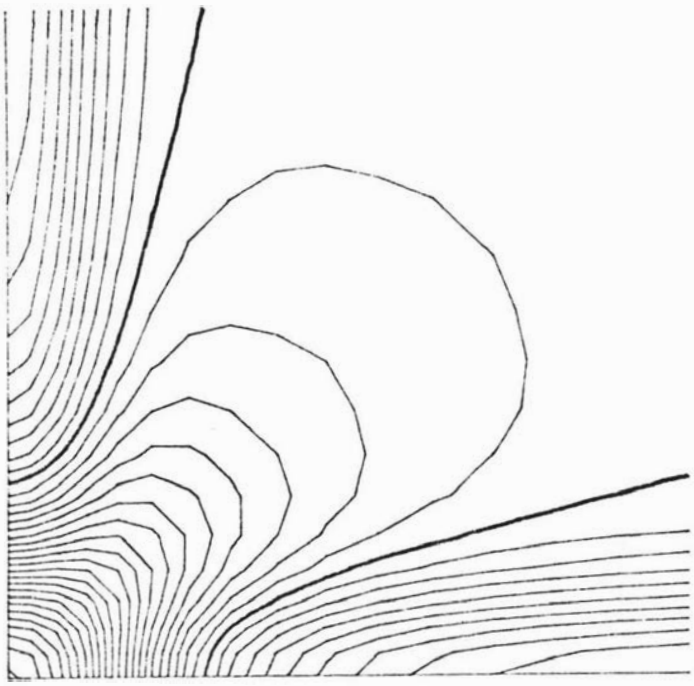

Figure 8a *Contour Map*

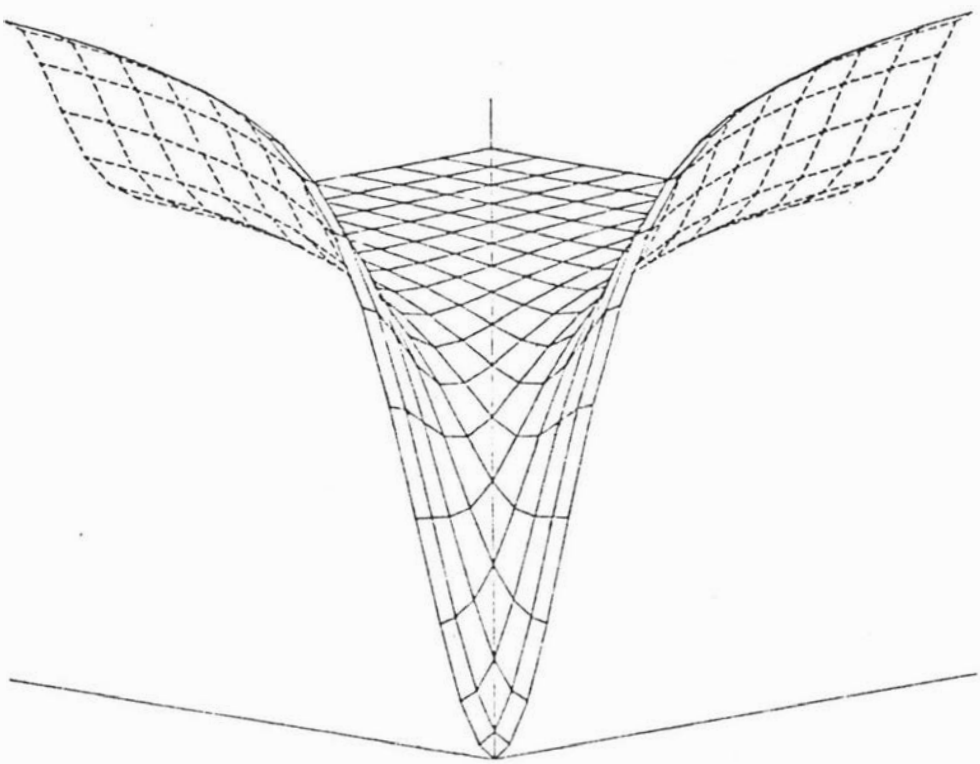

Figure 8b *Pictorial View*

Figure 9 *Difference in Risk Between Separate Shrinkage and Combined Shrinkage Estimators (Dimensions 3 and 3)*

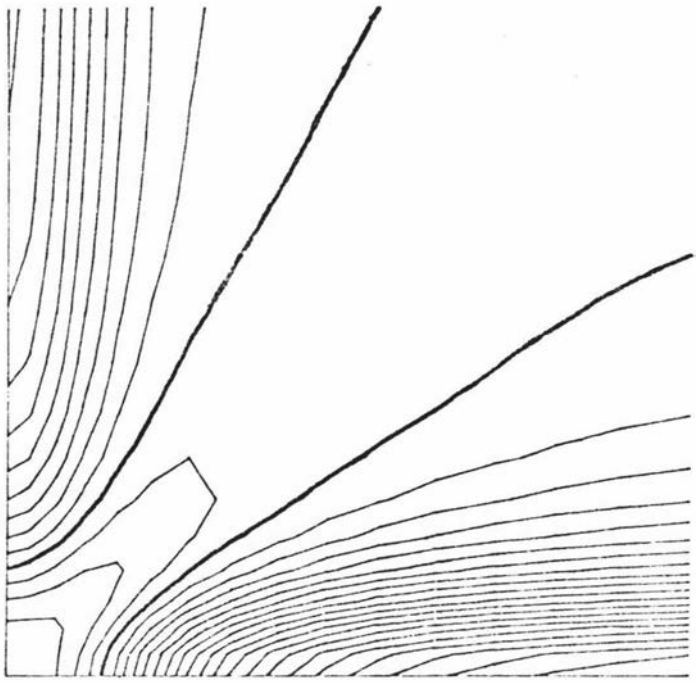

Figure 9a *Contour Map*

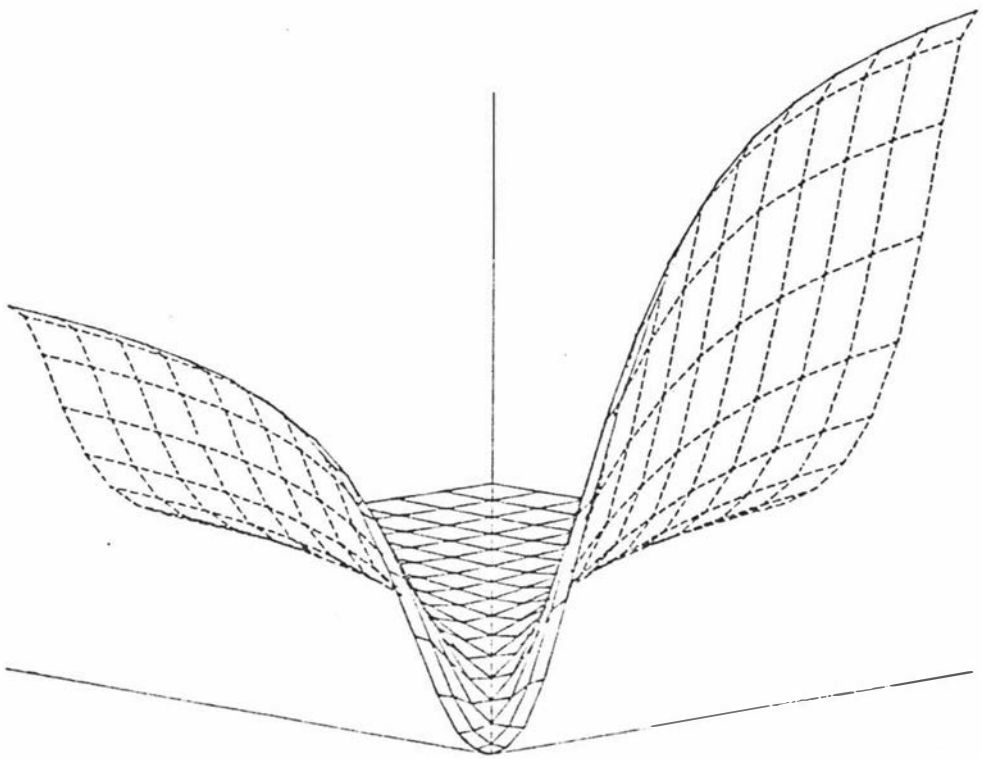

Figure 9b *Pictorial View*

Figure 9 *Difference in Risk Between Separate Shrinkage and Combined Shrinkage Estimators (Dimensions 3 and 5)*

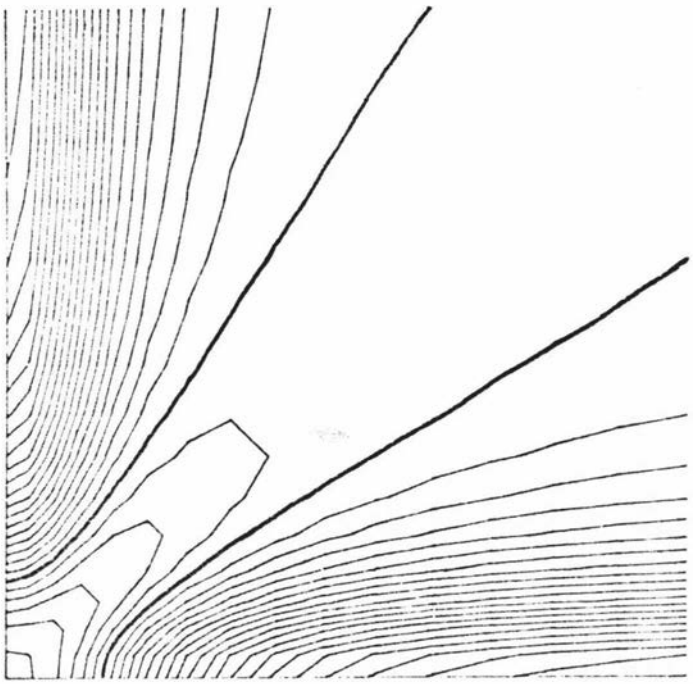

Figure 10a *Contour Map*

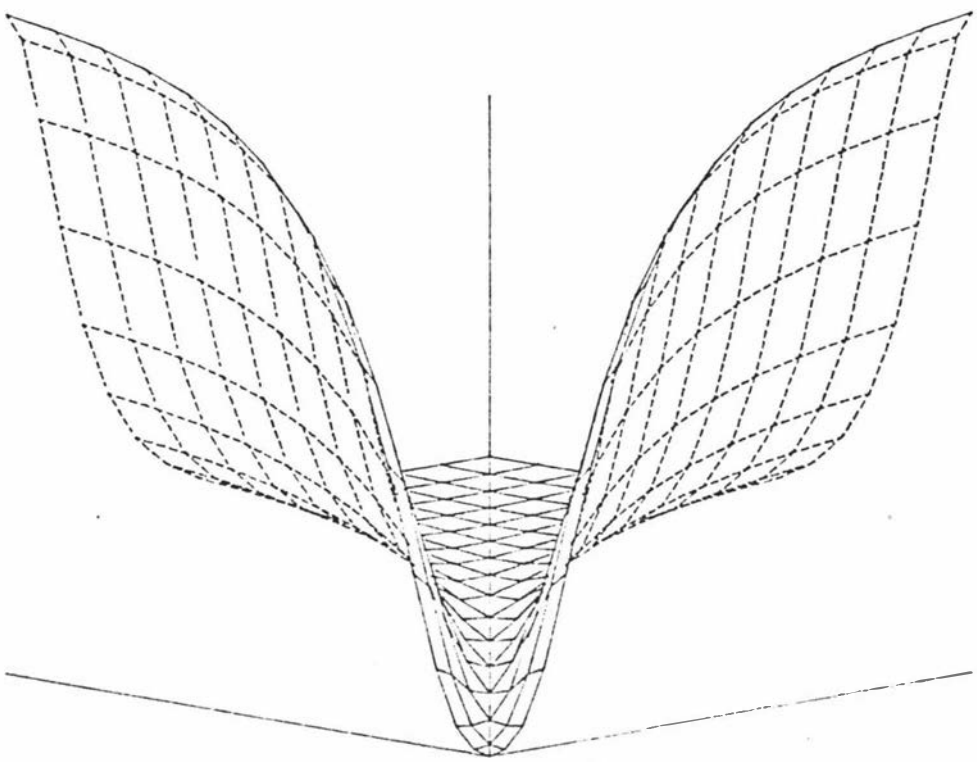

Figure 10b *Pictorial View*

Figure 10 *Difference in Risk Between Separate Shrinkage and Combined Shrinkage Estimators (Dimensions 10 and 10)*

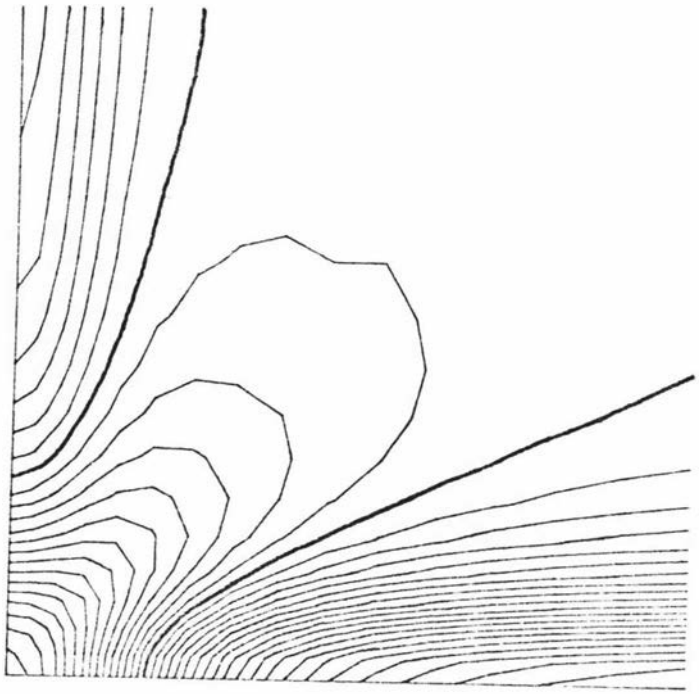

Figure 11a *Contour Map*

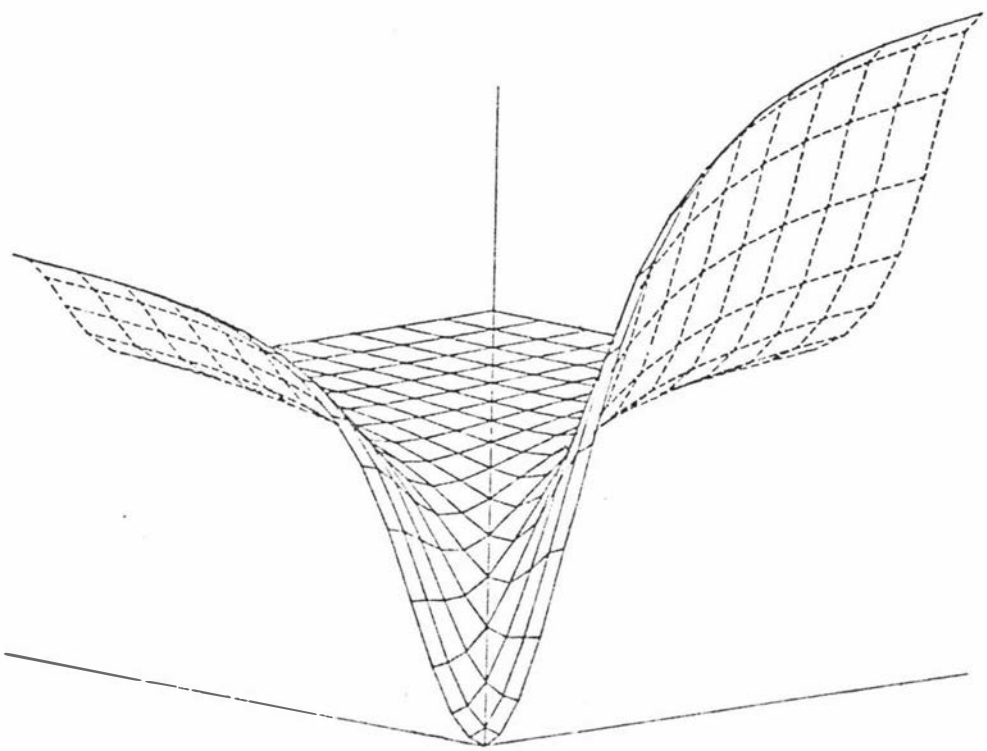

Figure 11b *Pictorial View*

Figure 12 *Difference in Risk Between Separate Shrinkage and Combined Shrinkage Estimators (Dimensions 7 and 13)*

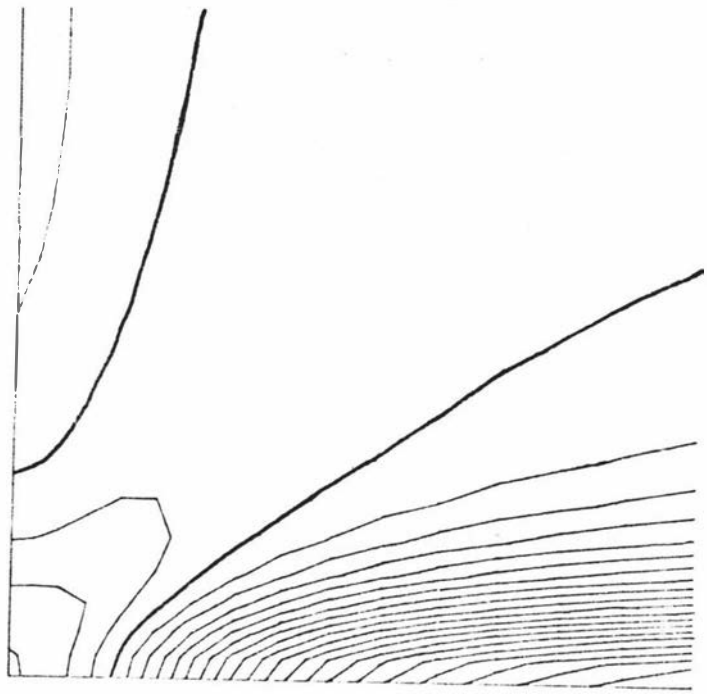

Figure 12a *Contour Map*

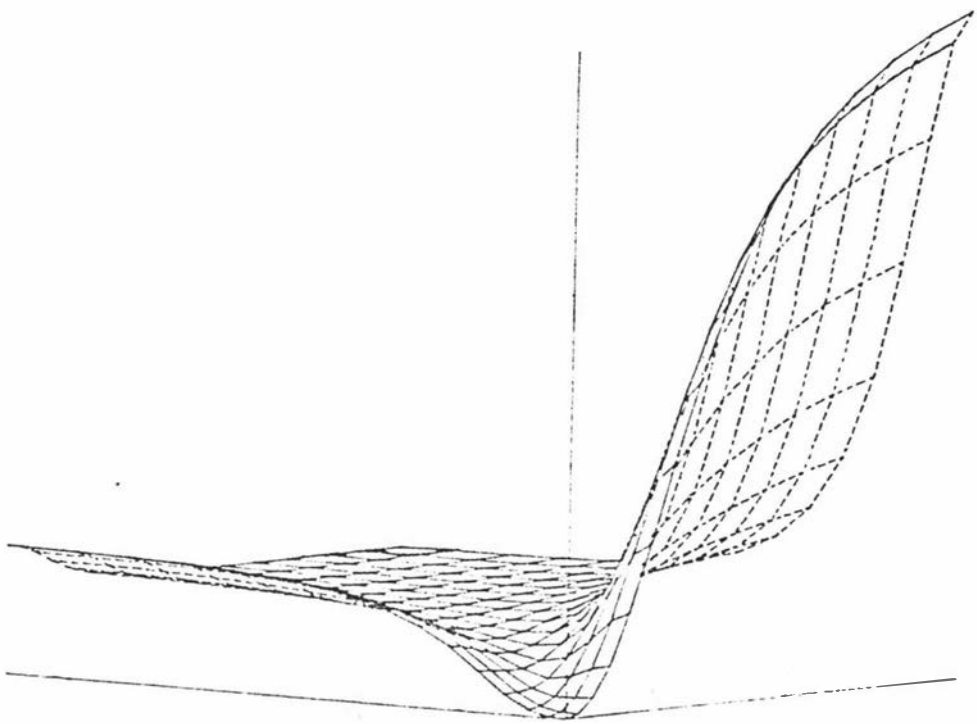

Figure 12b *Pictorial View*

Figure 12 *Difference in Risk Between Separate Shrinkage and Combined Shrinkage Estimators (Dimensions 3 and 17)*

a small region of large imrovement near the origin. The central nearly plane area is close to the plane $\Delta R = 0$. These graphs, plotted for various values of r_1 and r_2, are shown in figures 7 to 12.

2.2.4 Generalised James-Stein Estimators in Practice

We have shown that the lower the dimension of the hyperplane towards which we shrink the usual estimator, the greater the potential reduction in risk over that for the maximum likelihood estimator. However, we have also shown that Lindley type shrinkages towards higher dimensional hyperplanes and separate shrinkages within certain orthogonal hyperplanes can give smaller risk in practice. Which estimator is better is determined by how good is our prior guess of the mean vector. If we can guess some components more accurately than others, or using a Lindley type contraction, guess that some components are close to one another (without guessing their mean), then we will do better with a componentwise shrinkage. *This shows why we should not combine totally unrelated problems together - the risk is likely to be greater if we do.*

The best procedure seems to be that we choose subspaces in which the components are liable to be similar in value to one another and shrink within those spaces, either to a fixed point if we have good prior knowledge, or towards a common mean if our prior knowledge is not so good. Furthermore, we can avoid one component of the risk becoming large by using an Efron and Morris type limited shrinkage rule. This inflates the ensemble risk by only a small amount while protecting a few components from being too greatly affected by the majority. We note, however, that by using small enough groups of components this protection is not so important.

2.3 General Variance Matrix

Suppose that in our model X has the distribution $X \sim N_p(\mu,\sigma^2 V)$. It is possible to transform back into canonical form by a transformation matrix L. We have $LX \sim N_p(L\mu,\sigma^2 LVL^T)$ in which case we want the loss function to become

$$\mathcal{L}(L\hat{\mu},L\mu,\sigma^2) = \| L\hat{\mu} - L\mu \|^2 / \sigma^2 .$$

In order that LX have variance matrix $\sigma^2 I$, $LVL^T = I$ which implies that $V = L^{-1}L^{T\,-1} = (L^TL)^{-1}$. Being symmetric and of full rank, V^{-1} can always be factorised as $V^{-1} = L^TL$. Now

$$\mathcal{L}(L\hat{\mu},L\mu,\sigma^2) = (\hat{\mu} - \mu)^T L^T L(\hat{\mu} - \mu)/\sigma^2$$

$$= (\hat{\mu} - \mu)^T V^{-1}(\hat{\mu} - \mu)/\sigma^2$$
$$= \| \hat{\mu} - \mu \|^2_{V^{-1}}/\sigma^2$$

where $\|x\|^2_A = x^T A x$. We take this to be the loss function for $\hat{\mu}$ as an estimator for μ. In other words we shall consider the model $X \sim N_p(\mu, \sigma^2 V)$, $S \sim \frac{\sigma^2}{n} \chi^2_n$ independently of X and we use the loss function $\mathcal{L}(\hat{\mu}, \mu, \sigma^2) = \| \hat{\mu} - \mu \|^2_{V^{-1}}/\sigma^2$. Although we could repeat the above theory without transforming to canonical form it is not necessary to do so as we may transform the result derived previously.

The hypothesis $H\mu = h$ may be written $HL^{-1}L\mu = h$. We shall apply the previous results to LX. We require a generalised inverse of HL^{-1} satisfying the symmetry conditions of section 2.2. Now if G is a generalised inverse of H then LG is a generalised inverse of HL^{-1} since $HL^{-1}LGHL^{-1} = HGHL^{-1} = HL^{-1}$. Also $LGHL^{-1}$ is symmetric if and only if $LGHL^{-1} = L^{-1^T} H^T G^T L^T$ and since $V^{-1} = L^T L$ this is equivalent to $GHV = (GHV)^T$. If H is to be split as $H = [H_1^T, \ldots, H_t^T]^T$ then we similarly take G_i to be a generalised inverse of H_i with $G_i H_i V$ symmetric and take $G = [G_1, \ldots, G_t]$ in which case GHV will be symmetric and G will be a generalised inverse of H.

We may now write

$$X = Gh + (I - GH)X + \sum_{i=1}^{t} G_i (H_i X - h_i)$$

and premultiplying by L we obtain

$$LX = LGh + (I - LGHL^{-1})LX + \sum_{i=1}^{t} LG_i(H_i L^{-1} LX - h_i).$$

Putting

$$Y = Gh + (I - GH)X, \quad Z_i = G_i(H_i X - h_i) \quad \text{and} \quad Z = \sum_{i=1}^{t} Z_i$$

we observe that LY, LZ and the LZ_i satisfy the orthogonality conditions derived previously if $(H_i L^{-1})(H_j L^{-1})^T = 0$ for $i \neq j$. The latter condition is equivalent to $H_i L^{-1} L^{-1^T} H_j^T = 0$ which is equivalent to $H_i V H_j^T = 0$ for $i \neq j$. Thus when the matrices H_i satisfy the orthogonality condition $H_i V H_j^T$ for $i \neq j$, the decomposition $X = Y + \sum_{i=1}^{t} Z_i$ divides X into orthogonal components with respect to the inner product $\langle a, b \rangle_{V^{-1}} = a^T V^{-1} b$ and LX into orthogonal components $LX = LY + \sum_{i=1}^{t} LZ_i$ with respect to the inner

product $\langle a,b\rangle = a^T b$. They will also be statistically independent.

We may now calculate the risk for the estimator

$$\breve{\mu} = Y + \sum_{i=1}^{t} \left(1 - \frac{c_i S}{\|Z_i\|^2_{V^{-1}}}\right) Z_i$$

directly using the distributions

$$\|Y - \eta\|^2_{V^{-1}} \sim \sigma^2 \chi^2_{p-r} \quad \text{and} \quad \|Z_i\|^2_{V^{-1}} \sim \sigma^2 \chi^2_{r_i}(\tfrac{1}{2}\|\zeta_i\|^2_{V^{-1}})$$

or alternatively by using the special case alredy considered. In either case we have

$$E\left[\|Y + \sum_{i=1}^{t} \phi_i(\|Z_i\|^2_{V^{-1}}, S)Z_i - \mu\|^2_{V^{-1}}\right]$$

$$= E\left[\|LY + \sum_{i=1}^{t} \phi_i(\|LZ_i\|^2, S)LZ_i - L\mu\|^2\right]$$

$$= p - \sum_{i=1}^{t} c_i\{2(r_i-2) - \frac{n+2}{n} c_i\}E\left[\frac{1}{r_i-2+2K_i}\right]$$

where $\phi_i(\|Z_i\|^2_{V^{-1}}, S) = 1 - \dfrac{c_i S}{\|Z_i\|^2_{V^{-1}}}$ and K_i has a Poisson distribution with parameter $\tfrac{1}{2}\|Z_i\|^2_{V^{-1}}$. This formula is almost identical with the special case derived previously. It relies on the use of the loss function $L(\hat{\mu},\mu,\sigma^2) = \|\hat{\mu} - \mu\|^2_{V^{-1}}/\sigma^2$. Other quadratic loss functions will give more complicated formulae for the risk.

2.4 Generalised James-Stein Estimators for the Parameters of a Linear Model

In this section we shall change the notation of the previous section and shall use the symbol X to denote a fixed matrix. We wish to estimate β in the linear model $Y = X\beta + e$, $e \sim N_n(0,\sigma^2 V)$. We shall take X to be a matrix of full column rank; Y is the vector of observations.

Now the usual estimators for β and σ^2 are

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-p}(Y - X\hat{\beta})^T V^{-1}(Y - X\hat{\beta})$$

which are unbiased. These statistics are independent and jointly sufficient for β, however there is no single sufficient statistic for β so that no estimator for β may be based on a sufficient statistic. We may base estimators on both of the above statistics (or just on $\hat{\beta}$) but there is no need to use the original observed vector Y.

Accordingly we shall apply the results of section 2.3 to the random variables $\hat{\beta}$ and $\hat{\sigma}^2$. Now $\hat{\beta} \sim N_p(\beta,\sigma^2(X^TV^{-1}X)^{-1})$ and $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$. We may thus replace X in section 2.3 by $\hat{\beta}$, S by $\hat{\sigma}^2$ and n by $n-p = \nu$. If we choose a set of hypotheses concerning β, $H_i\beta = h_i$ with $H_i(X^TV^{-1}X)^{-1}H_j^T = 0$ for $i \neq j$ then we are ready to apply the results to our model giving the estimator

$$\check{\beta} = Gh + (I - GH)\hat{\beta} + \sum_{i=1}^{t}\left(1 - \frac{c_i\hat{\sigma}^2}{\|G_i(H_i\hat{\beta} - h_i)\|^2_{X^TV^{-1}X}}\right)G_i(H_i\hat{\beta} - h_i)$$

where $H_iG_iH_i = H_i$, $G_iH_i(X^TV^{-1}X)^{-1}$ is symmetric, $h = [h_1^T,\ldots,h_t^T]^T$, $G = [G_1,\ldots,G_t]$ and $H = [H_1^T,\ldots,H_t^T]^T$. By the result of section 2.3 this estimator has the risk function

$$R(\check{\beta},\beta,\sigma^2) = p + \sum_{i=1}^{t} c_i\{2(r_i-2) - \frac{\nu+2}{\nu}c_i\}E\left[\frac{1}{r_i-2+2K_i}\right]$$

where K_i has a Poisson distribution with parameter $\frac{1}{2\sigma^2}\|G_i(H_i\beta - h_i)\|^2_{X^TV^{-1}X}$ and $r_i = \operatorname{rank} H_i$. This has risk uniformly less than that for the usual estimator, $\hat{\beta}$, when each c_i is such that $0 < c_i < \frac{2(r_i-2)\nu}{\nu+2}$ and has a minimum when $c_i = \frac{\nu}{\nu+2}(r_i-2)$ in which case the risk is

$$R(\check{\beta},\beta,\sigma^2) = p - \frac{\nu}{\nu+2}\sum_{i=1}^{t} E\left[\frac{(r_i-2)^2}{r_i-2+2K_i}\right].$$

2.4.1 Geometrical Construction of James-Stein Estimators in Linear Models

The space in which Y lies can be decomposed as the direct sum $T = T_0 \oplus T^*$ where T_0 is the error space and T^* is the row space of X^T. The vector Y can be written

$$Y = X\hat{\beta} + \hat{e} = X(X^TV^{-1}X)^{-1}X^TV^{-1}Y + (I - X(X^TV^{-1}X)^{-1}X^TV^{-1})Y$$

with $\hat{e} \in T_0$ and $X\hat{\beta} \in T^*$, $X\hat{\beta}$ being orthogonal to $\hat{e}$ with respect to the inner product $\langle a,b\rangle = a^TV^{-1}b$. The usual estimator for $X\beta$ is the component of Y in the space T_0. This gives $\hat{\sigma}^2 = \nu^{-1}\|\hat{e}\|^2_{V^{-1}}$. Using the notation $P_{\mathcal{U}}Y$ for the projection of the vector Y onto the subspace $\mathcal{U}$ we have $\hat{e} = P_{T_0}Y$ and $X\hat{\beta} = P_{T^*}Y$. Since X has full column rank, the equation $X\beta = \mu$ has a unique solution if it is consistent and it follows that we only need to estimate $\mu = X\beta$.

The hypothesis $H\beta = h$ may be written in terms of μ since if X

has full column rank the equation $H = H^+X$ has a solution $H^+ = HX^-$ where X^- is any generalised inverse of X; so that $H\beta = h$ may be written as $HX^-\mu = h$. We may therefore estimate μ with an estimator which shrinks the maximum likelihood estimator towards each of the hyperplanes $H_iX^-\mu = h_i$ and then, on multiplying by X^- produce the corresponding estimator for β.

The hypotheses $H_i\beta = h_i$ (or $H_iX^-\mu = h_i$) provide a decomposition of T^* into a constant vector XGh and a set of orthogonal subspaces $T_0^*, \ldots, T_t^*$. As in section 2.2 (putting β in place of the vector X) β may be written as

$$\beta = Gh + (I - GH)\beta + \sum_{i=1}^{t} G_i(H_i\beta - h_i)$$

$$\text{so } X\beta = XGh + X(I - GH)\beta + \sum_{i=1}^{t} XG_i(H_i\beta - h_i)$$

where $(I - GH)^TX^TV^{-1}XG_iH_i = 0$ and $H_i^TG_i^TX^TV^{-1}XG_jH_j = 0$ for $i \neq j$ which shows that $T_0^*, T_1^*, \ldots, T_t^*$ are indeed orthogonal with respect to the inner product $\langle .,. \rangle_{X^TV^{-1}X}$.

The estimator previously written explicitly can now be written in terms of the projections of Y onto the spaces $T_0, T_0^*, T_1^*, \ldots, T_t^*$.

We have $$\breve{\mu} = P_{T^+}0 + P_{T_0^*}Y + \sum_{i=1}^{t}\left(1 - \frac{c_i \| P_{T_0^*}Y\|^2_{V^{-1}}}{\nu \|P_{T_i^*}Y\|^2_{V^{-1}}}\right) P_{T_i^*}Y$$

where T^+ is the hyperplane $HX^-\mu^* = h$. This gives

$$\breve{\beta} = X^-P_{T^+}0 + X^-P_{T^*}Y + \sum_{i=1}^{t}\left(1 - \frac{c_i \| P_{T_0^*}Y\|^2_{V^{-1}}}{\nu \|X^-P_{T_i^*}Y\|^2_{X^TV^{-1}X}}\right) X^-P_{T_i^*}Y$$

which is the estimator given previously.

This geometrical view of the estimator $\breve{\beta}$ avoids any explicit mention of computational procedures and, since $\breve{\mu}$ is unique even when X does not have full column rank, allows for generalisation to non-full rank models.

2.5 Linear Models of Less than Full Rank

We shall now consider the linear model $Y = X\beta + e$, $e \sim N_n(0,\sigma^2V)$ where X does not have full column rank. Although β is not estimable, certain linear combinations of its elements are. If $k = \text{rank } X$ then there are k linearly independent estimable functions. It is well known that a non-full rank model can be transformed into a full rank one by reparametrising in terms of a set of estimable functions. Such

a procedure is mentioned, for example, in Pringle and Rayner (1971). We shall expand a little on this.

Let $\Lambda\beta$ be a vector of k linearly independent estimable functions. This is equivalent to the condition $\Lambda = TX$ where T and Λ have full row rank k. We wish to reparametrise the model in terms of $\beta^* = \Lambda\beta$. We shall at first discuss the transformed model in general whether or not it has full rank. Let the transformed model be $Y = X^*\beta^* + e$ and therefore $X^*\beta^* = X^*\Lambda\beta = X\beta$ for all β. This implies $X^*\Lambda = X$. In order that both models have the same rank we require that X and X^* have the same column space (we have just shown that the column space of X is a subspace of that of X^*) which implies that $X^* = X\Lambda^*$ for some matrix Λ^*.

If $\lambda^T\beta$ is estimable in the original model then

$$\lambda^T\beta = \alpha^T X\beta = \alpha^T X^*\beta^* = \lambda^{*T}\beta^* \quad \text{where} \quad \lambda^{*T} = \alpha^T X^*$$

so that estimable functions in one model correspond to estimable functions in the other. Also if β^o is a solution to the normal equations $X^TV^{-1}X\beta = X^TV^{-1}Y$ then $\Lambda\beta^o$ is a solution to the normal equations $X^{*T}V^{-1}X^*\beta^* = X^{*T}V^{-1}Y$ and conversely if β^{*o} is a solution to the latter equations then $\Lambda^*\beta^{*o}$ is a solution to the former. This follows since $X^TV^{-1}X\beta = X^TV^{-1}Y$ implies $\Lambda^{*T}X^TV^{-1}X\beta = \Lambda^{*T}X^TV^{-1}Y$ or $X^{*T}V^{-1}X^*\Lambda\beta = X^{*T}V^{-1}Y$ and conversely $X^{*T}V^{-1}X^*\beta^* = X^{*T}V^{-1}Y$ implies $\Lambda^TX^{*T}V^{-1}X^*\beta = \Lambda^TX^{*T}V^{-1}Y$ or $X^TV^{-1}X\Lambda^*\beta^* = X^TV^{-1}Y$.

If $H\beta = h$ is a testable hypothesis then

$$H = UX = UX^*\Lambda = H^*\Lambda \quad \text{where} \quad H^* = UX^*$$

and the hypothesis may be written as $H^*\beta^* = h$.

In general we wish to preserve the property that if $\widetilde{\lambda^T\beta}$ is an estimator for $\lambda^T\beta$ and $\widetilde{\beta}$ is a solution to the equation $\lambda^T\widetilde{\beta} = \widetilde{\lambda^T\beta}$ then the corresponding estimator for $\lambda^{*T}\beta^* = \lambda^T\beta$ in the tansformed model is $\lambda^{*T}\widetilde{\beta^*}$ where $\widetilde{\beta^*} = \Lambda\widetilde{\beta}$. If this is so then the loss function

$$\begin{aligned}
L^*(\widetilde{\beta^*},\beta^*,\sigma^2) &= \sigma^{-2}\,\|\widetilde{\beta^*} - \beta^*\|^2_{X^{*T}V^{-1}X^*} \\
&= \sigma^{-2}(\widetilde{\beta^*} - \beta^*)^TX^{*T}V^{-1}X^*(\widetilde{\beta^*} - \beta^*) \\
&= \sigma^{-2}(\widetilde{\beta} - \beta)^T\Lambda^TX^{*T}V^{-1}X^*\Lambda(\widetilde{\beta} - \beta) \\
&= \sigma^{-2}(\widetilde{\beta} - \beta)^TX^TV^{-1}X(\widetilde{\beta} - \beta) \\
&= \sigma^{-2}\,\|\widetilde{\beta} - \beta\|^2_{X^TV^{-1}X} \\
&= L(\widetilde{\beta},\beta,\sigma^2)
\end{aligned}$$

so that one loss function transforms to the other. In the case of the

full rank model this is the loss function already considered.

Now suppose that $H^T = [H_1^T, H_2^T, \ldots, H_t^T]$, $h^T = [h_1^T, h_2^T, \ldots, h_t^T]$ and $H_i = A_i X$ for $i = 1, 2, 3, \ldots, t$. In the transformed model the hypotheses which correspond to the hypotheses $H_i\beta = h_i$ will be $H_i^* \beta^* = h_i$ where $H_i^* = A_i X^*$. The conditions for independence of the $H_i\beta^o - h_i$ and of the $G_i H_i \beta^o$ and $(I - GH)\beta^o$ are

$$\operatorname{cov}(H_i\beta^o, H_j\beta^o) = H_i(X^TV^{-1}X)^-X^TV^{-1}X(X^TV^{-1}X)^-H_j^T = 0 \quad \text{and}$$

$$\operatorname{cov}(XG_iH_i\beta^o, X(I - GH)\beta^o) = X(I - GH)(X^TV^{-1}X)^-X^TV^{-1}X(X^TV^{-1}X)^-H_i^TG_i^TX^T = 0;$$

β^o being a least squares solution to the normal equations, G_i being a generalised inverse of H_i for which $X^TV^{-1}XG_iH_i$ is symmetric and $G = [G_1, G_2, \ldots, G_t]$. The matrix G_i gives the projection of β^o onto the plane $H_i\beta^o = h_i$ the projection being $P(\beta^o) = \beta^o - G_i(H_i\beta^o - h_i)$. If the former condition holds for $i \neq j$ then the latter becomes

$$X(I - G_iH_i)(X^TV^{-1}X)^-X^TV^{-1}X(X^TV^{-1}X)^-H_i^TG_i^TX^T = 0.$$

Since the hypotheses are testable we may use the properties of generalised inverses of $X^TV^{-1}X$ to reduce these conditions to

$$H_i(X^TV^{-1}X)^-H_j^T = 0 \quad \text{and} \quad X(I - G_iH_i)(X^TV^{-1})^-H_i^TG_i^TX^T = 0.$$

Before showing that these conditions are equivalent to the corresponding conditions for the transformed model and that these conditions are independent of the generalised inverse used we shall give some useful formulae and collect together those already given. These will be stated as a lemma.

Lemma 1 If (i) $X^*\Lambda = X$ (ii) $X\Lambda^* = X^*$
(iii) $AX = H$ and (iv) $AX^* = H^*$

then

1. $H^*\Lambda = H$ $\quad H\Lambda^* = H^*$
2. $\Lambda^*(X^{*T}V^{-1}X^*)^-\Lambda^{*T}$ is a generalised inverse of $X^TV^{-1}X$
3. $X(X^TV^{-1}X)^-X^TV^{-1}X = X$
4. $H(X^TV^{-1}X)^-X^TV^{-1}X = H$
5. If $G^* = \Lambda G$ then $XG = X^*G^*$ and $HG = H^*G^*$
6. G is a generalised inverse of H is equivalent to $G^* = \Lambda G$ is a generalised inverse of H^*
7. $X^TV^{-1}XGH$ is symmetric is equivalent to $XGH(X^TV^{-1}X)^-X^T$ is symmetric
8. If $G^* = \Lambda G$ then $X^{*T}V^{-1}X^*G^*H^*$ is symmetric is equivalent to

$X^T V^{-1} XGH$ is symmetric

9. Similar formulae are true if each starred variable is interchanged with each corresponding unstarred variable.

Proof 1. $H^*\Lambda = AX^*\Lambda = AX = H \qquad H\Lambda^* = AX\Lambda^* = AX^* = H^*$

2. $X^T V^{-1} X\Lambda^*(X^{*T} V^{-1} X^*)^- \Lambda^{*T} X^T V^{-1} X$
$= \Lambda^T X^{*T} V^{-1} X^* (X^{*T} V^{-1} X^*)^- X^{*T} V^{-1} X^* \Lambda$
$= \Lambda^T X^{*T} V^{-1} X^* \Lambda$
$= X^T V^{-1} X .$

3. $X^T V^{-1} X (X^T V^{-1} X)^- X^T V^{-1} X = X^T V^{-1} X ,$
but V^{-1} can be written $V^{-1} = P^T P$ where P has full rank. Therefore $PX(X^T P^T PX)^- X^T P^T PX = PX$
so that $X(X^T V^{-1} X)^- X^T V^{-1} X = X .$

4. Since $H = AX ,$ 3. implies
$H = H(X^T V^{-1} X)^- X^T V^{-1} X .$

5. $XG = X^*\Lambda G = X^* G^* \qquad HG = H^*\Lambda G = H^* G^*$

6. $HGH = H \Leftrightarrow H^* G^* H = H \Rightarrow H^* G^* H\Lambda^* = H\Lambda^* \Leftrightarrow H^* G^* H^* = H^*$
$H^* G^* H^* = H^* \Leftrightarrow HGH^* = H^* \Rightarrow HGH^*\Lambda = H^*\Lambda \Leftrightarrow HGH = H .$

7. $X^T V^{-1} XGH$ is symmetric implies that if $(X^T V^{-1} X)^-$ is symmetric then $(X^T V^{-1} X)^- X^T V^{-1} XGH (X^T V^{-1} X)^-$ is symmetric. This implies that $X(X^T V^{-1} X)^- X^T V^{-1} XGH (X^T V^{-1} X)^- X^T$ is symmetric i.e. $XGH(X^T V^{-1} X)^- X^T$ is symmetric. Since this is invariant to the generalised inverse used (since $H(X^T V^{-1} X)^- X^T = AX(X^T V^{-1} X)^- X^T$) this holds in general. Conversely, $XGH(X^T V^{-1} X)^- X^T$ is symmetric implies $X^T V^{-1} XGH(X^T V^{-1} X)^- X^T V^{-1} X$ is symmetric, i.e. $X^T V^{-1} XGH$ is symmetric.

8. $X^{*T} V^{-1} X^* G^* H^* = \Lambda^{*T}\{X^T V^{-1} XGH\}\Lambda^*$ and
$X^T V^{-1} XGH = \Lambda^T\{X^{*T} V^{-1} X^* G^* H^*\}\Lambda$
so that the symmetry of the parts in brackets implies the symmetry of the left hand sides.

9. By the symmetry of the conditions the result follows.

We now state a theorem which shows the equivalence of the orthogonality conditions in the transformed model with those in the original model and gives a condition for $X(I - GH)\beta^o$ to be orthogonal to $XG_i H_i \beta^o$.

Theorem 1 1. $H_i(X^T V^{-1} X)^- H_j^T = H_i^*(X^{*T} V^{-1} X^*)^- H_j^{*T}$ and is invariant to the generalised inverse used.

2. If $X^T V^{-1} XG_i H_i$ is symmetric then $X(I - G_i H_i)(X^T V^{-1} X)^- H_i^T G_i^T X^T = 0.$

Proof 1. Since $H_i = A_iX$ and $H_j = A_jX$ invariance follows.

Now $H_i^*(X^{*T}V^{-1}X^*)^-H_j^{*T} = H_i\Lambda^*(X^{*T}V^{-1}X^*)^-\Lambda^{*T}H_j^T$

$= H_i(X^TV^{-1}X)^-H_j^T$

by lemma 1 (part 2).

2. $X(I - G_iH_i)(X^TV^{-1}X)^-H_i^TG_i^TX^T$

$= (X - XG_iA_iX)(X^TV^{-1}X)^-H_i^TG_i^TX^T$

$= XG_iH_i(X^TV^{-1}X)^-X^T - XG_iA_iXG_iH_i(X^TV^{-1}X)^-X^T$

$= X(G_iH_i - G_iH_iG_iH_i)(X^TV^{-1}X)^-X^T$

$= 0.$

We now transform to the full rank model and apply the James-Stein technique to the estimator $\hat{\beta}^*$ for β^*. The estimator is

$$\overset{\frown}{\beta}^* = G^*h + (I - G^*H^*)\hat{\beta}^* + \sum_{i=1}^{t}\left\{1 - \frac{c_i\hat{\sigma}^2}{\|G_i^*(H_i^*\hat{\beta}^* - h_i)\|^2_{X^{*T}V^{-1}X^*}}\right\} \times G_i^*(H_i^*\hat{\beta}^* - h_i)$$

with $0 < c_i < 2\frac{n-k}{n-k+2}(r_i-2)$ where $r_i = \operatorname{rank} H_i$. The optimum value of c_i is $c_i = \frac{n-k}{n-k+2}(r_i-2)$.

Now writing $\overset{\frown}{\Lambda\beta}$ for $\overset{\frown}{\beta}^*$ we have

$$\overset{\frown}{\Lambda\beta} = \Lambda Gh + \Lambda(I - GH)\beta^\circ + \Lambda\sum_{i=1}^{t}\left\{1 - \frac{c_i\hat{\sigma}^2}{\|G_i(H_i\beta^\circ - h_i)\|^2_{X^TV^{-1}X}}\right\} \times G_i(H_i\beta^\circ - h_i).$$

One solution for $\Lambda\overset{\frown}{\beta^\circ} = \overset{\frown}{\Lambda\beta^\circ}$ is

$$\overset{\frown}{\beta^\circ} = Gh + (I - GH)\beta^\circ + \sum_{i=1}^{t}\left\{1 - \frac{c_i\hat{\sigma}^2}{\|G_i(H_i\beta^\circ - h_i)\|^2_{X^TV^{-1}X}}\right\}G_i(H_i\beta^\circ - h_i),$$

and this generates estimators for all the estimable functions.

For completeness we restate the conditions on G_i and H_i. We must have $H_i(X^TV^{-1}X)^-H_j^T = 0$ for $i \neq j$, $H_iG_iH_i = H_i$ and $X^TV^{-1}XG_iH_i$ is symmetric (which is equivalent to the symmetry of $XG_iH_i(X^TV^{-1}X)^-X^T$).

2.6 Restricted Models of Less than Full Rank

We wish to estimate the estimable functions of β in the linear model $Y = X\beta + e$, $e \sim N_n(0,\sigma^2V)$ under the restrictions $R\beta = r$. If the restrictions are of the form of a hypothesis which is testable in the unrestricted model then they provide a genuine resriction on the

model. Otherwise they restrict the parameter space and remove some of the arbitrariness due to the model's not being of full rank. We shall consider the general case in which some of the restrictions may be of one type and the rest of the other.

Suppose $\lambda^T\beta$ is an estimable function and μ is an arbitrary vector in the same dimension as r. We may write

$$\lambda^T\beta = \lambda^T\beta + \mu^T(R\beta - r) = (\lambda + R^T\mu)^T\beta - \mu^T r$$

so an estimable function may be written in many ways in the form $\nu^T\beta + \alpha$. In the unrestricted model if $\hat{f}$ is an estimator for $f(\beta)$ then it is natural to take $\hat{f} + c$ as an estimator for $f(\beta) + c$ so it is only necessary to consider homogeneous linear functions; however in the restricted case we have shown that it is useful to consider non-homogeneous linear functions of the form $\lambda^T\beta + \alpha$.

Definition A function $\lambda^T\beta + \alpha$ is estimable if and only if, for some c and some vector t (where t and c are not necessarily unique), we have $\lambda^T\beta + \alpha = t^T E[Y] + c$, the condition holding for all for which $R\beta = r$.

In the unrestricted model this corresponds to the usual condition since $\lambda^T\beta + \alpha = t^T E[Y] + c = t^T X\beta + c$ for all β implies that $\lambda^T = t^T X$ and $\alpha = c$. We shall find a condition for estimability in the restricted model. The restriction $R\beta = r$ is equivalent to the condition $\beta = m + M\xi$ where $\operatorname{rank} M = \dim\beta - \operatorname{rank} R$, $Rm = r$ and $RM = 0$ with ξ being arbitrary. One possible value of M is $I - R^-R$ and for m is R^-r where R^- is a generalised inverse of R. Another choice is $(I - R^-R)Q$ where Q is such that $(I - R^-R)Q$ consists of a maximal set of linearly idependent columns of $(I - R^-R)$. In any case M satisfies the equation $M = (I - R^-R)M$ since $RM = 0$. Now $\lambda^T\beta + \alpha = t^T X\beta + c$ for all β with $R\beta = r$ is equivalent to $\lambda^T(m + M\xi) + \alpha = t^T X(m + M\xi) + c$ for all ξ and this is equivalent to $\lambda^T m + \alpha = t^T Xm + c$ and $\lambda^T M = t^T XM$. Solving the latter gives $\lambda^T - t^T X = \rho^T K$ where ρ is arbitrary and $KM = 0$ with $\operatorname{rank} K = \dim\beta - \operatorname{rank} M$. Since R satisfies this condition we may take $K = R$ giving $\lambda^T = t^T X + \rho^T R$. This gives $\rho^T Rm + \alpha = c$ or $\alpha = c - \rho^T R$. The condition for estimability is therefore $\lambda^T = t^T X + \rho^T R$ and this might have been guessed since $t^T X\beta$ is estimable in the unrestricted model and $\rho^T R\beta$ is known in the restriced model.

We now wish to state conditions under which testable hypotheses $H_i\beta = h_i$ are orthogonal in the restricted model (they do not need to be

orthogonal in the unrestricted model) and conditions on the G_i such that $G_i(H_i\beta - h_i)$ is the deviation from the intersection of the plane $H_i\beta = h_i$ with the plane $R\beta = r$.

Transforming the model to an unrestricted model gives $Y - Xm = XM\xi + e$. Since $M = (I - R^-R)M$ the column space of M is a subspace of that of $I - R^-R$ and since $\operatorname{rank} M = \operatorname{rank}(I - R^-R)$ the column spaces are equal and $I - R^-R = MK$ for some matrix K. Thus

$$AM = 0 \Rightarrow AMK = 0 \Rightarrow A(I - R^-R) = 0 \Rightarrow A(I - R^-R)M = 0 \Rightarrow AM = 0$$

so that $AM = 0$ is equivalent to $A(I - R^-R) = 0$. Let $X^* = XM$ and $H_i^* = H_iM$. In the transformed model $Y - Xm = X^*\xi + e$ we know conditions for the testable hypotheses $H_i^*\xi = h_i$ to be orthogonal and for $(I - G_i^*H_i^*)\xi$ to be the projection of ξ onto the hyperplane $H_i^*\xi = 0$. These conditions are

$$H_i^*(X^{*T}V^{-1}X^*)^-H_j^* = 0, \quad H_i^*G_i^*H_i^* = H_i^*$$

and the symmetry of

$$X^{*T}V^{-1}X^*G_i^*H_i^* \quad \text{and} \quad X^*G_i^*H_i^*(X^{*T}V^{-1}X^*)^-X^{*T}$$

(the latter symmetry conditions being equivalent). We wish to express these conditions in terms of the original model.

If we put $G_i = MG_i^*$ we obtain

$$H_iG_i = H_iMG_i^* = H_i^*G_i^*, \quad XG_i = XMG_i^* = X^*G_i^* \quad \text{and}$$

$$H_iG_iH_iM = H_i^*G_i^*H_i^* = H_i^* = H_iM \quad \text{which is equivalent to}$$

$H_iG_iH_i(I - R^-R) = H_i(I - R^-R)$. Conversely, if G_i satisfies the latter condition and is in the column space of M then $G_i = MG_i^*$ for some G_i^* and $H_iMG_i^*H_iM = H_iM$ which is equivalent to $H_i^*G_i^*H_i^* = H_i^*$. Also $X^{*T}V^{-1}X^*G_i^*H_i^* = M^TX^TV^{-1}XG_iH_iM$ so that the symmetry of $X^{*T}V^{-1}X^*G_i^*H_i^*$ is equivalent to the symmetry of $M^TX^TV^{-1}XG_iH_iM$ - a condition which does not depend on M. Also

$$X^*G_i^*H_i^*(X^{*T}V^{-1}X^*)^-X^{*T} = XG_iH_iM(M^TX^TV^{-1}XM)^-M^TX^T$$

so that one is symmetric if and only if the other is. Since the symmetry conditions in the starred variables are equivalent this is also independent of M. We shall show this in another way.

Let $XFX^TV^{-1}XM = XM$ (i.e. $M^TX^TV^{-1}XFX^TV^{-1}XM = M^TX^TV^{-1}XM$) and let F have all its rows and columns in the row and column spaces of M. Therefore $F = M\Phi M^T$ for some Φ. This implies that Φ is a generalised inverse of $M^TX^TV^{-1}XM$. We show that XFX^T is invariant to

the choice of F. Let F_1 and F_2 be two matrices satisfying the conditions with $F_i = M\Phi_i M^T$.

Now $XMF_1X^TV^{-1}XM = XM = XF_2X^TV^{-1}XM$

which implies $XM\Phi_1M^TX^TV^{-1}XM = XM\Phi_2M^TX^TV^{-1}XM$

so that $XM\Phi_1M^TX^T = XM\Phi_2M^TX^T$

which gives $XF_1X^T = XF_2X^T$.

Now one possible value of F is $F = M(M^TX^TV^{-1}XM)^-M^T$ so we may write $XG_iH_iM(M^TX^TV^{-1}XM)^-M^TX^T$ as $XG_iA_iXFX^T$ which is invariant to the choice of F.

The matrix F is useful in the solution of the normal equations in the restricted model since, in the transformed model,

$\xi^\circ = \Phi M^TX^TV^{-1}(Y - Xm)$ where Φ is a generalised inverse of $M^TX^TV^{-1}XM$. Putting $\beta^\circ = m + M\xi^\circ$, $R\beta^\circ = Rm + RM\xi^\circ = r$ and

$$\begin{aligned}\beta^\circ &= m + M\Phi M^TX^TV^{-1}(Y - Xm)\\ &= m + FX^TV^{-1}(Y - Xm)\\ &= FX^TV^{-1}Y + (I - FX^TV^{-1}X)m .\end{aligned}$$

Naturally this is not invariant to F unless β is estimable. If $\Lambda\beta$ is estimable then $\Lambda = AX + BR$ and

$$\Lambda\beta^\circ = AXFX^TV^{-1}Y + A(X - XFX^TV^{-1}X)m + Br$$

which is invariant to F. It is also invariant to m, different values of m giving the same solution. Finally,

$$\begin{aligned}E[\Lambda\beta^\circ] &= AXFX^TV^{-1}X\beta + A(X - XFX^TV^{-1}X)m + Br\\ &= AXFX^TV^{-1}X(m + M\xi) + AXm + Br - AXFX^TV^{-1}Xm\\ &= AXFX^TV^{-1}XM\xi + AXm + Br\\ &= AXM\xi + AXm + Br\\ &= AX\beta + BR\beta\\ &= \Lambda\beta\end{aligned}$$

so that $\Lambda\beta^\circ$ is an unbiased estimator for $\Lambda\beta$. The invariance of the solution to m may be proved as follows. If m_1 and m_2 are two particular solutions to $R\beta = r$ then $m_1 - m_2 = M\theta$ for some θ. The difference between estimators $\Lambda\beta_1^\circ$ and $\Lambda\beta_2^\circ$, where these correspond to the values m_1 and m_2 respectively, is

$$\begin{aligned}\Lambda\beta_1^\circ - \Lambda\beta_2^\circ &= A(X - XFX^TV^{-1}X)M\theta\\ &= A(XM - XFX^TV^{-1}XM)\theta\\ &= 0 .\end{aligned}$$

The variance of β° is given by

$$\begin{aligned}\text{var } \beta^{\circ} &= FX^{T}V^{-1}(\text{var } Y)V^{-1}X^{T}F^{T} \\ &= \sigma^{2}FX^{T}V^{-1}VV^{-1}XF^{T} \\ &= \sigma^{2}FX^{T}V^{-1}XF^{T}\end{aligned}$$

and $\text{var } \Lambda\beta^{\circ} = \sigma^{2}\Lambda FX^{T}V^{-1}XF^{T}\Lambda^{T} = \sigma^{2}\Lambda FX^{T}V^{-1}XF\Lambda$

since $\Lambda FX = \Lambda M\Phi M^{T}X = AXM\Phi M^{T}X = AXFX^{T}$

is invariant to F and as F^{T} satisfies the conditions on F we may replace F^{T} by F (in fact we can use a symmetric F).

Also $\Lambda FX^{T}V^{-1}XF\Lambda = AXFX^{T}V^{-1}XM\Phi M^{T}\Lambda$

$= AXM\Phi M^{T}\Lambda$

$= \Lambda F\Lambda^{T}$.

Therefore, $\text{var } \Lambda\beta^{\circ} = \sigma^{2}\Lambda F\Lambda^{T}$.

We now summarise the conditions on the H_i and G_i which ensure the orthogonality of $H_i\beta^{\circ} - h_i$ and $H_j\beta^{\circ} - h_j$ and of $X(I - GH)\beta^{\circ}$ and $XG_iH_i\beta^{\circ}$. The conditions are

$H_iM(M^{T}X^{T}V^{-1}XM)^{-}M^{T}H_j = 0$ (i.e. $H_iFH_j = 0$)

and $M^{T}X^{T}V^{-1}XG_iH_iM$ is symmetric for any matrix M for which rank M + rank R = dim β and $M = (I - R^{-}R)M$ (e.g. M may be taken to be $I - R^{-}R$).

Now we shall write down the generalised James-Stein value which corresponds to β°. It is

$$\breve{\beta}^{\circ} = Gh + (I - GH)\beta^{\circ} + \sum_{i=1}^{t}\left\{1 - \frac{c_i\hat{\sigma}^{2}}{\|G_i(H_i\beta^{\circ} - h_i)\|^{2}_{X^{T}V^{-1}X}}\right\}G_i(H_i\beta^{\circ} - h_i)$$

since $\|G_i^{*}\{H_i^{*}\xi^{\circ} - (h_i - H_im)\}\|^{2}_{X^{*T}V^{-1}X^{*}} = \|G_i(H_i\beta^{\circ} - h_i)\|^{2}_{X^{T}V^{-1}X}$.

This follows from the following:

1. $H_i\beta = h_i \Leftrightarrow h_i(m + M\xi) = h_i$

$\Leftrightarrow H_iM\xi = h_i - H_im$

$\Leftrightarrow H_i^{*}\xi = h_i - H_im$

2. $X^{*}G_i^{*}\{H_i^{*}\xi^{\circ} - (h_i - H_im)\} = XG_i(H_iM\xi^{\circ} - h_i + H_im)$

$= XG_i(H_i\beta^{\circ} - h_i)$.

2.7 Discussion

The estimators which we have developed in this chapter are based on the idea of shrinking the maximum likelihood estimator orthogonally towards several hyperplanes.

We showed that these ideas can be applied directly to the non-

full rank linear model and to restricted linear models and expresssed the results in a form which involved only the original parameters. (Although we did use reparametrisation as a tool for proving these results, the parametrisation was chosen arbitrarily). The main value of this is that, no matter how the model is reparametrised, the result is shown to be the same (or more precisely, invariant to the transformation used).

From a practical point of view, however, the generalised inverses or near generalised inverses needed in the calculation are most easily calculated by carrying out the transformation (at least in part). Also it is unlikely that orthogonal hyperplanes will arise in practice except in simpler models which are nearer to the canonical form. For both these reasons it seems likely that the estimators will best be found by reparametrising the model.

A more important consideration is the gain in efficiency that the James-Stein approach to estimation affords. Some criticism has been given in the literature based on the mistaken belief that the more estimation problems which are combined together the greater should be the efficiency of the James-Stein estimator. We have shown that combining problems together only to the extent that we use a combined estimator for the error variance, but otherwise keeping them separate, is likely to produce a smaller risk than by combining all the problems together with a single shrinkage factor.

Apart from the high probability of mis-specifying the variance matrix when unrelated problems are combined (it is clearly dangerous to assume all variances to be equal) the chance of identifying a suitable choice of origin on the basis of vague prior knowledge is poor. By separating the problems, a good choice of origin for one component problem will lower the risk even when the others are poorly chosen. This increases the chance of reducing the risk and makes the method attractive. It might be worth trying to prove that we also may gain in efficiency by estimating the error variance separately in each component problem and thus keeping the problems entirely separate. It seems very likely that this is so.

A further possibility is to let the data decide to what extent the problems are to be combined. This approach has been discussed by Efron and Morris(1973b). They introduce a data dependent shrinkage factor which, under favourable circumstances, gives equal shrinkage

to all components while, under unfavourable cicumstances, leaves the problems separate. In general the result will be in between these two extremes.

In fact, this suggestion of Efron and Morris can often be fitted into the geralised James-Stein framework when there are at least three hyperplanes. This can be done by introducing a fourth hyperplane. For example, if the hypotheses were

$\beta_{ij} = \alpha_j$ for $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, t$

where

$\beta = [\beta_{00} \ \beta_{11} \ \beta_{12} \cdots \beta_{1n_1} \ \beta_{21} \ \beta_{22} \cdots \beta_{2n_2} \ \ldots \ \beta_{tn_t}]$

then a further prior hypothesis could be $\alpha_j = 0$ for $j = 1, 2, \ldots, t$. The support for the latter hypothesis would determine the extent to which the combined shrinkage towards $\beta = 0$ occurs.

Chapter 3

Bayesian Estimation in the Linear Model

3.1 Introduction

In this chapter we shall show how certain prior distributions for the parameters of a linear model give rise to estimators of a similar form to those of the previous chapter - being a shrinkage of the maximum likelihood estimator towards each of a set of hyperplanes. This is a generalisation of the estimators in Chapter 1 which were developed for the estimation of the mean of a multivariate normal distribution.

This shows a similarity between a Bayesian approach and the James-Stein approach, but note that the justification - as opposed to the motivation - of the latter is entirely sampling theoretic.

The major point of this chapter is to show how Bayesian methods may be applied to non-full rank models.

As we pointed out in Chapter 1, in order to obtain estimators with good sampling properties we must use estimators which are (at least approximately) derived from prior distributions. Although necessary this condition is not sufficient since estimators derived from an improper prior, although sometimes admissible, can often be improved upon quite considerably. However, this observation justifies the use of Bayesian methods even in the absence of strong prior knowledge (which naturally justifies their use). It is rare that no prior knowledge exists, but weak prior knowledge is, by definition, not precise enough to be incorporated into a prior distribution. Even so, whether prior knowledge is weak or non-existent, we must try to postulate a prior distribution if we want our estimators to be good in a sampling theory sense.

A useful technique for embodying weak prior knowledge is to use a two stage prior: the first stage is a prior distribution involving unknown parameters (frequently a proper prior) and the second stage is a prior distribution for the parameters of the first stage prior. This technique can be taken to several more stages if desired. A similar method can be used even with strong prior knowledge based on past data since there must have been a time before that data was collected. A weak prior distribution for the parameter of interest, when combined with the past data, gives rise to a stronger posterior which is then used as a prior for the current data set. Clearly this

method can also be applied to several stages - each data set giving rise to a stronger prior to be applied to the next data set.

The former approach was used by Lindley and Smith (1972) for estimating the parameter β in the linear model $Y = X\beta + e$, $e \sim N(0,\Sigma)$. The parameter vector β was given a normally distributed prior $\beta \sim N(\mu\mathbf{1},\Sigma_1)$ and μ was given a uniform distribution. This guaranteed an exchangeable distribution for the components of β. Assuming that Σ and Σ_1 are known they found the posterior distribution for β and its mean. In the case where Σ and Σ_1 were each known up to a multiplicative constant ($\Sigma = \sigma^2 V$, $\Sigma_1 = \sigma_1^2 V_1$) they found the mode of the posterior distribution and showed that the parameters σ^2 and σ_1^2 could be estimated thus giving an empirical Bayes estimator for β. Assuming inverse χ^2 distributions for σ^2 and σ_1^2 Lindley and Smith were also able to find the posterior distribution for β, but in this case the calculation involves difficult numerical integration.

The latter approach was used by Tiao and Zellner (1964). Using the same model they considered past data $Y_1 = X_1 \beta + e_1$, $e_1 \sim N(0,\Sigma_1)$ with a uniform prior for $(\beta, \log \sigma, \log \sigma_1)$ where $\Sigma = \sigma^2 V$ and $\sigma_1^2 = \sigma^2 V_1$ with V and V_1 known (=I in their paper). Unless the variances are estimated from the data, the unknown variance case gives rise to similar difficult integrals to those obtained by Lindley and Smith. Tiao and Zellner, generalising a technique of Fisher, give an asymptotic expansion for the integral. As Fienberg points out in the discussion to the Lindley and Smith paper, this integral can also be simplified using the results of Dickey (1968).

In the next section we shall show that the two approaches are special cases of the same general scheme and that this scheme leads to the estimators of chapter 2.

3.2 Posterior Distribution of the Parameter Vector

In order to reconcile the above approaches we show that the ultimate prior after all previous stages, in the case of known variance matrix, consists of a normal distribution of the components of the parameter vector in the direction of some hyperplane and a uniform distribution perpendicular to it. There is a slight complication when previous observations are in the form of non-full rank models, a case which might arise when these observations are of some components of the parameter vector only. In this case the prior

distribution for non-estimable functions must be ignored if it is improper (a proper prior causes no difficulty). The reason for ignoring the prior distribution for non-estimable functions is that an improper prior will give rise to a posterior distribution which is also improper. The mean of such a posterior will be undefined and this reinforces our view that the functions are not estimable. Note that, although it is not usually stated explicitly that non-estimable components are ignored, the above procedure is in fact the usual one. Any model can have extra irrelevant parameters added to it, but if they are irrelevant then it would be better if they were ignored. This is what this method does - estimates of the other parameters should be the same whether the irrelevant parameters are included in or excluded from the model. This can be done by factoring the joint density into a factor involving the estimable function and a factor involving the non-estimable functions. The latter is ignored (or, in the case of a proper prior, it integrates to unity). Two other approaches achieve the same effect: Box and Tiao (1973) use locally uniform priors which are proper but approximate a uniform distribution over the region of interest; Lindley (1965) suggests that a uniform distribution be regarded as a family of conditional distributions - uniform on each bounded region - so that for any bounded region we may condition on it to achieve a proper density. Lindley's approach is similar to that of Jeffreys (1961) who remarks that an integral over an infinite range is defined as the limit of a family of integrals over a finite range as the range tends to infinity (however Jeffreys fails to point out that for improper priors, in contrast to the case for proper priors, the integrand is repeatedly renormalised as the range increases).

We shall show the result of using a uniform prior for ϕ on both the marginal density for θ and on the posterior density for ϕ given θ when we consider the linear model $\theta = A\phi + \varepsilon$, $\varepsilon \sim N_n(0,\Sigma)$. We do not assume that A has full column rank.

We first write the model in terms of just the estimable functions then we find the marginal and posterior distributions and transform back to the original coordinates.

Let A be an $n\times p$ matrix of rank r, let Λ be an $r\times p$ matrix such that the components of $\Lambda\phi$ form a complete set of estimable functions (ie there exist matrices T and B for which $\Lambda = TA$ and $A = B\Lambda$) and let M be a $(p-r)\times p$ matrix complementary to Λ. The matrix $\begin{bmatrix}\Lambda\\ M\end{bmatrix}$ is

of full rank. Let $[\Lambda^- \ M^-]$ be its inverse. We have $\Lambda\Lambda^- = I$, $MM^- = I$, $\Lambda M^- = 0$, $M\Lambda^- = 0$ and $\Lambda^-\Lambda + M^- M = I$. Also $AM^- = B\Lambda M^- = 0$ and $A\Lambda^- = B\Lambda\Lambda^- = B$. Putting $\psi_1 = \Lambda\phi$, $\psi_2 = M\phi$ and $\psi = \begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix}$ we obtain $\phi = \Lambda^-\psi_1 + M^-\psi_2$. Now $A\phi = A\Lambda^-\psi_1 + AM^-\psi_2 = B\psi_1$ so that, as might be expected, the parameters ψ_2 are irrelevant. Thus the model $\theta = A\phi + \varepsilon$ may be written as $\theta = [B \ 0]\psi + \varepsilon$ or as $\theta = B\psi_1 + \varepsilon$.

$$\text{Also } A^T\Sigma^{-1}A = \begin{bmatrix}\Lambda\\ M\end{bmatrix}^T [\Lambda^- \ M^-]^T A^T\Sigma^{-1}A [\Lambda^- \ M^-] \begin{bmatrix}\Lambda\\ M\end{bmatrix}$$

$$= \begin{bmatrix}\Lambda\\ M\end{bmatrix}^T \begin{bmatrix}B^T\\ 0\end{bmatrix} \Sigma^{-1} [B \ \ 0] \begin{bmatrix}\Lambda\\ M\end{bmatrix}$$

$$= \begin{bmatrix}\Lambda\\ M\end{bmatrix}^T \begin{bmatrix}B^T\Sigma^{-1}B & 0\\ 0 & 0\end{bmatrix} \begin{bmatrix}\Lambda\\ M\end{bmatrix}$$

$$\text{and } \phi^T A^T \Sigma^{-1} A\phi = \psi_1^T B^T \Sigma^{-1} B \psi_1$$

$$= \psi^T \begin{bmatrix}B^T \Sigma^{-1} B & 0\\ 0 & 0\end{bmatrix} \psi.$$

A similar transformation to this is to be found in Zellner (1971), but Zellner rewrites the density function instead of rewriting the model. The following paragraph shows the two approaches to be equivalent.

We now show that, if we transform $A^T\Sigma^{-1}A$ by a congruence transformation of the form $Q^T A^T \Sigma^{-1} A Q = \begin{bmatrix}C & 0\\ 0 & 0\end{bmatrix}$ where $Q = [Q_1 \ Q_2]$ and $Q_1^T A^T \Sigma^{-1} A Q_1 = C$, then the elements of $Q_1^-\phi$, where $[Q_1 \ Q_2]^{-1} = \begin{bmatrix}Q_1^-\\ Q_2^-\end{bmatrix}$, are a complete set of estimable functions.

We have $Q_2^T A^T \Sigma^{-1} A Q_2 = 0$ so that $AQ_2 = 0$. Also

$$Q_1 Q_1^- + Q_2 Q_2^- = I$$

so that

$$A = AQ_1Q_1^- + AQ_2Q_2^- = AQ_1Q_1^- .$$

Therefore rank Q_1 = rank Q_1^- = rank A = rank C and the column space of A is contained in the column space of Q_1. However, from the equality of the ranks, the column spaces are identical and hence there is a matrix T such that $Q_1^- = TA$. This establishes the result.

$$\text{Now } p(\theta|\phi) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\theta-A\phi)^T \Sigma^{-1}(\theta-A\phi)\}$$

and using a uniform prior $p(\phi) = c^p$ for ϕ we have

$$p(\theta,\phi) = \frac{c^p}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\theta-A\phi)^T\Sigma^{-1}(\theta-A\phi)\} .$$

Let ϕ° be a solution to the normal equations

$$A^T\Sigma^{-1}A\phi^\circ = A^T\Sigma^{-1}\theta$$

so that $$A^T\Sigma^{-1}(\theta - A\phi^\circ) = 0 .$$

We then have

$$\begin{aligned}(\theta - A\phi)^T\Sigma^{-1}(\theta - A\phi) &= \{(\theta - A\phi^\circ) - A(\phi - \phi^\circ)\}^T\Sigma^{-1}\{(\theta - A\phi^\circ) - A(\phi - \phi^\circ)\} \\ &= (\theta - A\phi^\circ)^T\Sigma^{-1}(\theta - A\phi^\circ) + (\phi - \phi^\circ)A^T\Sigma^{-1}A(\phi - \phi^\circ)\end{aligned}$$

and

$$\begin{aligned}(\theta - A\phi^\circ)^T\Sigma^{-1}(\theta - A\phi^\circ) &= \theta^T\Sigma^{-1}(\theta - A\phi^\circ) \\ &= \theta^T\Sigma^{-1}\{I - A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta \\ &= \theta^T\{\Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta .\end{aligned}$$

This gives

$$p(\theta,\phi) = \frac{c^p}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}\theta^T\{\Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta] \times \exp\{-\tfrac{1}{2}(\phi-\phi^\circ)A^T\Sigma^{-1}A(\phi-\phi^\circ)\} .$$

Also, putting $\psi^\circ = \begin{bmatrix}\Lambda \\ M\end{bmatrix}\phi^\circ$, $\psi_1^\circ = \Lambda\phi^\circ$ and $\psi_2^\circ = M\phi^\circ$, we have

$$p(\theta,\psi) = \frac{c^p}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}\theta^T\{\Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta] \times \exp\{-\tfrac{1}{2}(\psi_1-\psi_1^\circ)^TB^T\Sigma^{-1}B(\psi_1-\psi_1^\circ)\} .$$

In order to find the posterior density for ψ and the marginal distribution for θ we integrate with respect to ψ_1 and ignore the irrelevant parameters ψ_2 as we have already explained in section 3.1. Alternatively consider the distribution of ψ_2 as the limiting form of a family of proper distributions: we may take the limit after integrating. We therefore obtain

$$p(\theta) = \frac{c^r}{(2\pi)^{\frac{1}{2}(n-r)}|\Sigma|^{\frac{1}{2}}|B^T\Sigma^{-1}B|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}\theta^T\{\Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta]$$

and

$$p(\psi|\theta) = \frac{c^{p-r}}{(2\pi)^{\frac{1}{2}r}|B^T\Sigma^{-1}B|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\psi_1-\psi_1^\circ)^TB^T\Sigma^{-1}B(\psi_1-\psi_1^\circ)\}$$

or in terms of ϕ,

$$p(\phi|\theta) = \frac{c^{p-r}}{(2\pi)^{\frac{1}{2}r}|B^T\Sigma^{-1}B|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\phi-\phi^\circ)^TA^T\Sigma^{-1}A(\phi-\phi^\circ)\} .$$

Notice that both $p(\theta)$ and $p(\phi|\theta)$ have the form of a mulivariate normal distribution with singular precision matrix (Lindley's

terminology) and that this can be transformed to the product of a proper multivariate normal distribution and a uniform distribution (indeed, $p(\psi|\theta)$ is of that form). To see this for $p(\theta)$ note that there is an orthogonal matrix P such that

$$P^T\{\Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}P = \begin{bmatrix} I_{n-r} & 0 \\ 0 & 0 \end{bmatrix}.$$

We now show that Lindley and Smith's model is equivalent to one of the form used by Tiao and Zellner. Putting

$$C = I - A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}$$

we obtain

$$\begin{aligned} C^T\Sigma^{-1}C &= C^T\Sigma^{-1}\Sigma\Sigma^{-1}C \\ &= \Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1} \end{aligned}$$

so that the marginal distribution for θ is of the same form as that obtained from the linear model $\theta = C\phi + \varepsilon$, $\varepsilon \sim N(0,\Sigma)$ by taking the posterior distribution of ϕ given $\theta = 0$.

In the Lindley-Smith approach the first stage prior is uniform and gives rise to the marginal distribution which we have just found. We shall show that, using this marginal distribution as a prior for another normal distribution, gives a similar posterior. In the Tiao-Zellner approach we show that the posterior at the next stage has a similar form.

Let $\theta = A\phi + \varepsilon$, $\varepsilon \sim N_n(0,\Sigma)$

and let $p(\phi) = c\exp\{-\tfrac{1}{2}(\phi-\alpha)^TQ(\phi-\alpha)\}$

where Q does not necessarily have full rank. Now

$$p(\theta,\phi) = \frac{c}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}}\exp\{-\tfrac{1}{2}(\phi-\alpha)^TQ(\phi-\alpha) + (\theta-A\phi)^T\Sigma^{-1}(\theta-A\phi)\}.$$

In order to complete the square for the quadratic expression in ϕ,

$$\begin{aligned} &(\phi-\alpha)^TQ(\phi-\alpha) + (\theta-A\phi)^T\Sigma^{-1}(\theta-A\phi) \\ &\quad = (\phi-\alpha)^TQ(\phi-\alpha) + (\phi-\phi^o)^TA^T\Sigma^{-1}A(\phi-\phi^o) + (\theta-A\phi^o)^T\Sigma^{-1}(\theta-A\phi^o) \end{aligned}$$

where $\phi^o = (A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\theta$, we need the following lemma.

Lemma 1 If A and B are two matrices with B and $A - B$ positive or negative semi-definite and if A and B are symmetric then, for any generalised inverse, A^-, of A, $AA^-B = B$ (and hence $AA^-(A - B) = A - B$) and BA^-B is invariant to the choice of A^-.

Proof The result is essentially that in example 3.7 of Rao(1973).

As the proof is not given there we shall prove the result here. We first show that the null space of A is contained in the null space of B. This follows since

$$\begin{aligned} Ax = 0 &\Rightarrow x^TAx = 0 \\ &\Rightarrow x^T(A-B)x + x^TBx = 0 \\ &\Rightarrow x^TBx = 0 \\ &\Rightarrow Bx = 0 . \end{aligned}$$

Since the column space of a matrix is the orthogonal complement of the null space, this shows that the column space of B is contained in the column space of A. Thus there is a matrix T such that $B = AT$. Therefore $AA^-B = AA^-AT = AT = B$. Also $BA^-B = B^TA^-B = T^TA^TA^-AT = T^TAA^-AT = T^TAT = T^TA^TT = BT$ and this does not depend on A^-.

We can now state and prove the completion of the square result to which we alluded above. We shall state it as a lemma.

Lemma 2 If G and H are positive semi-definite then

$$\begin{aligned} &(x-a)^TG(x-a) + (x-b)^TH(x-b) \\ &\quad = \{x-(G+H)^-(Ga+Hb)\}^T(G+H)\{x-(G+H)^-(Ga+Hb)\} \\ &\qquad + (b-a)^T\{G-G(G+H)^-G\}(b-a) . \end{aligned}$$

Proof By lemma 1

(1) $(G+H)(G+H)^-G = G$ and $(G+H)(G+H)^-H = H$

from which we deduce that

(2) $G - G(G+H)^-G = H(G+H)^-G$ and $H - H(G+H)^-H = G(G+H)^-H$.

Now
$$\begin{aligned} &(x-a)^TG(x-a) + (x-b)^TH(x-b) \\ &\quad = x^T(G+H)x - 2x^T(Ga+Hb) + a^TGa + b^THb \\ &\quad = \{x-(G+H)^-(Ga+Hb)\}^T(G+H)\{x-(G+H)^-(Ga+Hb)\} \\ &\qquad + a^TGa + b^THb - (Ga+Hb)^T(G+H)^{-T}(G+H)(G+H)^-(Ga+Hb) \end{aligned}$$

from (1). Simplifying the constant term we obtain

$$\begin{aligned} &a^TGa + b^THb - (Ga+Hb)^T(G+H)^-(Ga+Hb) \\ &\quad = a^T\{G-G(G+H)^-G\}a + b^T\{H-H(G+H)^-H\}b - 2A^TG(G+H)^-Hb \end{aligned}$$

using (1). Although this formula has an elegant symmetry, for practical applications it is more convenient to use (2) to write the constant term in one of the forms

$$a^T\{G-G(G+H)^-G\}a + b^T\{G-G(G+H)^-G\}b - 2a^T\{G-G(G+H)^-G\}b$$

or $a^T\{H - H(G+H)^-H\}a + b^T\{H - H(G+H)^-H\}b - 2a^T\{H - H(G+H)^-H\}b$.

These may also be written

$$(a-b)^T\{G - G(G+H)^-G\}(a-b) , \quad (a-b)^T\{H - H(G+H)^-H\}(a-b) ,$$
$$(a-b)^TH(G+H)^-G(a-b) \text{ or as } (a-b)^TG(G+H)^-H(a-b) .$$

We now apply this lemma to the quadratic form in ϕ and obtain

$$\begin{aligned}(\phi-\alpha)Q(\phi-\alpha) &+ (\theta-A\phi)^T\Sigma^{-1}(\theta-A\phi)\\ &= \{\phi - (Q+A^T\Sigma^{-1}A)^-(Q\alpha + A^T\Sigma^{-1}A\phi^o)\}^T\\ &\qquad \times (Q + A^T\Sigma^{-1}A)\{\phi - (Q+A^T\Sigma^{-1}A)^-(Q\alpha + A^T\Sigma^{-1}A\phi^o)\\ &\quad + (\theta-A\phi^o)^T\Sigma^{-1}(\theta-A\phi^o)\\ &\quad + (\phi^o-\alpha)^T\{A^T\Sigma^{-1}A - A^T\Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}A\}(\phi^o-\alpha) .\end{aligned}$$

Substituting for ϕ^o and simplifying the last two terms we obtain

$$\begin{aligned}&\theta^T\{\Sigma^{-1} - \Sigma^{-1}A(A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta\\ &\quad + (\theta-A\alpha)^T\Sigma^{-1}A\{(A^T\Sigma^{-1}A)^- - (Q+A^T\Sigma^{-1}A)^-\}A^T\Sigma^{-1}(\theta-A\alpha)\\ &= \theta^T\{\Sigma^{-1} - \Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta\\ &\qquad - 2\theta^T\Sigma^{-1}A\{(A^T\Sigma^{-1}A)^- - (Q+A^T\Sigma^{-1}A)^-\}A^T\Sigma^{-1}A\alpha\\ &\qquad + \alpha^TA^T\Sigma^{-1}A\{(A^T\Sigma^{-1}A)^- - (Q+A^T\Sigma^{-1}A)^-\}A^T\Sigma^{-1}A\alpha\\ &= \theta^T\{\Sigma^{-1} - \Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}\theta\\ &\qquad - 2\theta^T\{\Sigma^{-1} - \Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}A\alpha\\ &\qquad + \alpha^TA^T\{\Sigma^{-1} - \Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}A\alpha\\ &= (\theta-A\alpha)^T\{\Sigma^{-1} - \Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}(\theta-A\alpha) .\end{aligned}$$

Alternatively,

$$(\theta-A\phi^o)^T\Sigma^{-1}(\theta-A\phi^o) = (\theta-A\alpha)\Sigma^{-1}(\theta-A\alpha) - (\phi^o-\alpha)^TA^T\Sigma^{-1}A(\phi^o-\alpha)$$

so the quadratic form becomes

$$(\theta-A\alpha)^T\Sigma^{-1}(\theta-A\alpha) - (\phi^o-\alpha)^TA^T\Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}A(\phi^o-\alpha)$$

which again yields the result.

The joint density of θ and ϕ can now be written as

$$p(\theta,\phi) = \frac{c}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}\|\phi - (Q+A^T\Sigma^{-1}A)^-(Q\alpha + A^T\Sigma^{-1}A\phi^o)\|^2_{Q+A^T\Sigma^{-1}A}\}$$
$$\times \exp\{-\tfrac{1}{2}\|\theta - A\alpha\|^2_{\Sigma^{-1}-\Sigma^{-1}A(Q+A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}}\} .$$

We now wish to integrate out the estimable functions of ϕ and ignore the non-estimable functions. In this context, however, we have not yet defined the term "estimable function" adequately. We do so now. Let B and Σ_1 be matrices such that $Q = B^T\Sigma_1^{-1}B$. A set of functions will be said to be estimable if its members are the elements of a vector of the form $(TB + UA)\phi$. These functions have as a basis those functions

which are estimable in the original model together with those functions which become estimable because of the prior distribution. Integrating over the estimable functions of ϕ and ignoring the others gives the marginal density for θ given by

$$p(\theta) \propto \exp[-\tfrac{1}{2}(\theta-A\alpha)^T\{\Sigma^{-1} - \Sigma^{-1}A(Q + A^T\Sigma^{-1}A)^-A^T\Sigma^{-1}\}(\theta-A\alpha)] .$$

Dividing this into the joint density of θ and ϕ gives the posterior density for ϕ,

$$p(\phi|\theta) \propto \exp[-\tfrac{1}{2}\,\|\phi - (Q + A^T\Sigma^{-1}A)^-(Q\alpha + A^T\Sigma^{-1}A\phi^{\circ})\|^2_{Q+A^T\Sigma^{-1}A}\ ]$$

The constant in each case is found by integrating over the estimable functions only and ignoring the others. Note that the case $Q = 0$ is the case of the uniform prior density already considered: in this case the terms containing α are zero.

We can now apply these results to the linear model. With prior knowledge as specified by Tiao and Zellner we have the model

$$Y = X\beta + \varepsilon , \qquad \varepsilon \sim N(0,\Sigma)$$

and prior observations

$$Y_i = X_i\beta + \varepsilon_i , \qquad \varepsilon_i \sim N(0,\Sigma_i) \qquad i = 1, 2, \ldots , t .$$

We shall write $Y_{t+1} = Y$, $X_{t+1} = X$ and $\Sigma_{t+1} = \Sigma$.

If we suppose we have a prior density for β of the form

$$p(\beta) \propto \exp\{-\tfrac{1}{2}(\beta-\alpha)^TQ(\beta-\alpha)\} ,$$

the case $Q = 0$ giving a uniform prior, then we may apply the last result repeatedly to obtain

$$p(\beta|Y_1,\ldots,Y_t,Y) \propto \exp[-\tfrac{1}{2}\,\|\beta - (Q + \sum_{i=1}^{t+1} X_i^T\Sigma_i^{-1}X_i)^-(Q\alpha + \sum_{i=1}^{t+1} X_i^T\Sigma_i^{-1}X_i\beta^{\circ}_{(i)}\|^2_{Q+\sum_{i=1}^{t+1} X_i\Sigma_i^{-1}X_i}\ ]$$

where $\beta^{\circ}_{(i)} = (X_i^T\Sigma_i^{-1}X_i)^-X_i^T\Sigma_i^{-1}Y_i$.

The form of this expression makes it clear that the same result is obtained by grouping some of the sets of prior observations together (or together with the current data set) or by taking the observations in a different order. This is true in general when the prior is proper. In this case it is now established for certain improper prior distributions. Tiao and Zellner did not do this but, instead, applied the uniform prior density directly to the full data set of current and prior observations.

The posterior mean for β (if the density is proper) is

$$\beta^{o} = (Q + \sum_{i=1}^{t+1} X_i^T \Sigma_i^{-1} X_i)^{-}(Q\alpha + \sum_{i=1}^{t+1} X_i^T \Sigma_i^{-1} Y_i)$$

where the generalised inverse is, in fact, an inverse. When the posterior density is improper this determines the estimable functions uniquely but not the non-estimable functions. In the case of an improper density, the integral is not defined uniquely, but can be given any value by choosing a suitable definition of the integral. However, as noted above, the estimable functions are determined uniquely.

A principle value for the integral $\int_{-\infty}^{\infty} f(x)\,dx$ is defined to be $P_g \int_{-\infty}^{\infty} f(x)\,dx = \lim_{N\to\infty} \int_{-N}^{g(N)} f(x)\,dx$ where $g(.)$ is a monotone increasing function of N for which $\lim_{N\to\infty} g(N) = \infty$. Principle values may converge to any limit, diverge to $+\infty$ or $-\infty$ or exhibit still more aberrant behaviour. This may be considered an advantage since the lack of uniqueness indicates that it is wrong to try to estimate the non-estimable functions.

Note that if $Q = 0$ then this is the result given by the usual sampling theory estimator in Theil(1971) obtained by combining previous and current observations. It is only applicable when the variance matrices are all known.

Consider now the same linear model but with prior knowledge a generalisation of that specified by Lindley and Smith. We have

$$\begin{aligned} Y &= X\beta + \varepsilon\,, & \varepsilon &\sim N(0,\Sigma) \\ \beta &= A_t\theta_t + \varepsilon_t\,, & \varepsilon_t &\sim N(0,\Sigma_t) \\ \theta_{i+1} &= A_i\theta_i + \varepsilon_i\,, & \varepsilon_i &\sim N(0,\Sigma_i) \quad \text{for } i = 0, 1, 2, \ldots, t-1 \end{aligned}$$

and $p(\theta_0) \propto \exp\{-\tfrac{1}{2}(\theta_0-\alpha)^T Q(\theta_0-\alpha)\}$.

The case of a uniform prior distribution at the final stage, given by $Q = 0$, is the case considered by Lindley and Smith.

Lindley and Smith found the posterior distribution of β given Y with prior knowledge given by the first stage. This depends on θ_t. By using the second stage prior for θ_t they eliminate this parameter by integration and obtain a result depending on θ_{t-1}. Repeating this process finally leads to a distribution depending on α (presumed to be known). To avoid dificulties with improper prior distributions, they take Q to be of full rank and calculate the limiting posterior

distribution for β as Q tends to zero. We, on the other hand, meet the challenge head on. Our approach also works in the reverse order using each stage of the prior to determine the prior knowledge at the next stage. Unfortunately the formula for the prior knowledge at each successive stage does not take quite such a simple form as with the approach of Tiao and Zellner. In practice, however, only two stages will usually be needed and rarely will it be necessary to go beyond three stages.

We now give the marginal distribution for θ_s after the last s stages, i.e. the prior at the $t-s+1$ stage. We have

$$p(\theta_s) \propto \exp\{-\tfrac{1}{2}(\theta_s - A_{s-1}A_{s-2}\ldots A_1\alpha)^T Q_{(s)}(\theta_s - A_{s-1}A_{s-2}\ldots A_1\alpha)$$

where $Q_{(s)}$ is defined inductively by $Q_{(0)} = Q$ and

$$Q_{(i+1)} = \Sigma_{i+1}^{-1} - \Sigma_{i+1}^{-1}A_{i+1}(Q_{(i)} + A_{i+1}^T\Sigma_{i+1}^{-1}A_{i+1})^- A_{i+1}^T\Sigma_{i+1}^{-1}.$$

Using the notation $\theta_{t+1} = \beta$, $p(\theta_t)$ is the ultimate prior distribution for β. We may now deduce that the posterior distribution for β is given by

$$p(\beta|\alpha) \propto \exp\{-\tfrac{1}{2}\|Y - (Q_{(t)}+X^T\Sigma^{-1}X)^-(Q_{(t)}A_{t-1}\ldots A_1\alpha + X^T\Sigma^{-1}X\beta^\circ\|^2_{Q_{(t)}+X^T\Sigma^{-1}X}\}.$$

This is the same as the result given by Lindley and Smith since if Q is non-singular then the marginal and conditional distributions of β may be found by integrating in any order; also the limit of our result as $Q \to 0$ is the same as that given by Lindley and Smith. In general we could allow Q to tend to any singular limit (since if Q_0 is singular then $\forall\ \delta > 0$, $Q = Q_0 + \delta I$ is positive definite and therefore non-singular, and we can choose Q as close as we like to Q_0) so, for the more general case, it does not matter in which order the integrations are carried out. Owing to the complicated form of the marginal distributions a direct proof that the order of integration is unimportant is difficult.

It is worth noting that a combination of the approaches of Lindley and Smith and of Tiao and Zellner may be used. We may replace the prior distribution for β used by the latter authors by a multistage prior of the form used by the former authors. A further extension, which may be applied to either approach, or to the joint approach, is to replace the prior distribution

$$p(\beta) \propto \exp\{-\tfrac{1}{2}(\beta-\alpha)^T Q(\beta-\alpha)\}$$

by the prior

$$p(\beta) \propto (\beta^T\beta)^{-c} \exp\{-\tfrac{1}{2}(\beta-\alpha)^T Q(\beta-\alpha)\}.$$

This will still be an exchangeable prior if $\alpha = 0$ and $Q = I$ but exchangeability is lost when prior observations are incorporated. This is to be expected since the prior observations will not usually give the same information about each coordinate of the parameter vector. If the model is of full rank then the effect on the posterior distribution for β will merely be to multiply it by $(\beta^T\beta)^{-c}$.

3.3 Estimation Under Prior Linear Hypotheses

The foregoing theory works just as well if prior data

$$Y_i = X_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0,\Sigma_i)$$

is replaced by prior beliefs

$$h_i = H_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0,\Sigma_i).$$

In this case the prior knowledge will be weaker and this will imply that each Σ_i is large. The ε_i reflect the fact that our belief that $h_i = H_i\beta$ is uncertain and we may be in error. We shall impose the restriction that our prior beliefs are mutually consistent, that is the equation $H\beta = h$ where $H^T = [H_1^T, H_2^T, \ldots, H_t^T]$ and $h = [h_1^T, h_2^T, \ldots, h_t^T]^T$ has a solution for β. We shall let the rank of H_i be r_i and the rank of H be r, the matrices H_i being independent so that $r = \Sigma_{i=1}^t r_i$.

Choosing as a prior distribution for β

$$p(\beta) \propto \exp\{-\tfrac{1}{2}(\beta-\alpha)^T Q(\beta-\alpha)\}$$

(where this might perhaps be the result of multistage prior information) the posterior distribution will be

$$p(\beta|Y,h_1,\ldots,h_t)$$
$$\propto \exp\{-\tfrac{1}{2}\|\beta - (Q + \sum_{i=1}^t H_i^T\Sigma_i^{-1}H_i + X^T\Sigma^{-1}X)^-(Q\alpha + \sum_{i=1}^t H_i^T\Sigma_i^{-1}h_i + X^T\Sigma^{-1}Y\|^2\}$$

where the norm is $\|.\| = \|.\|_{Q + \sum_{i=1}^t H_i^T\Sigma_i^{-1}H_i + X^T\Sigma^{-1}X}$.

This gives rise to estimators $\Lambda\beta^\circ$ for functions $\Lambda\beta$ which are estimable in the original model, where

$$\beta^\circ = (Q + \sum_{i=1}^t H_i^T\Sigma_i^{-1}H_i + X^T\Sigma^{-1}X)^-(Q\alpha + \sum_{i=1}^t H_i^T\Sigma^{-1}h_i + X^T\Sigma^{-1}Y).$$

In order to see how the analysis at the end of chapter 1 may have a parallel in the regression model we shall use a prior distribution

$$p(\beta) \propto \{\beta^T(Q + \textstyle\sum_{i=1}^t H_i^T\Sigma_i^{-1}H_i + X^T\Sigma^{-1}X)\beta\}^{-c}\exp\{-\tfrac{1}{2}(\beta-\alpha)^TQ(\beta-\alpha)\}.$$

The matrix of the quadratic form in β has been chosen so that the prior probability only depends on the estimable functions relative to the posterior distribution of β. Prior knowledge of this form may not be entirely realistic; however, if it produces admissible estimators and does not introduce too much non-sample information, then it may be worthwhile. The posterior distribution in this case will be

$$p(\beta|Y,h_1,\ldots,h_t) \propto (\beta^T W\beta)^{-c} \exp\{-\tfrac{1}{2}||\beta - W^-(Q\alpha + \sum_{i=1}^{t} H_i^T \Sigma_i^{-1} h_i + X^T\Sigma^{-1}Y\,||_W^2\}$$

where $W = Q + \sum_{i=1}^{t} H_i^T \Sigma_i^{-1} H_i + X^T\Sigma^{-1}X$.

If W is non-singular then the expectation of β exists, otherwise it does not. Using the same convention of ignoring the non-estimable functions that we have used before, we may calculate $E[\Lambda\beta]$ where the elements of $\Lambda\beta$ are estimable functions relative to the prior distribution. Defining β^o as in the equation above and defining $z = \tfrac{1}{2}n\beta^{oT}W\beta^o$ we may use a result of chapter 1 to obtain

$$E[\Lambda\beta] = \Lambda \frac{\tfrac{1}{2}p+c}{\tfrac{1}{2}p} \frac{{}_1F_1(\tfrac{1}{2}p+c+1;\tfrac{1}{2}p+1;z)}{{}_1F_1(\tfrac{1}{2}p+c;\tfrac{1}{2}p;z)} \beta^o .$$

Notice that this is a scalar shrinkage of $\Lambda\beta^o$ which, in the case when $\alpha = 0$, $h_i = 0$ for $i = 1,\ldots,t$, is a matrix shrinkage of the maximum likelihood estimator for $\Lambda\beta$. It is doubtful whether this double shrinkage is realistic unless each is mild (i.e. c is small and $r = \sum_{i=1}^{t} \operatorname{rank} H_i$ is not too large).

3.4 The case of Unknown Variance

When some or all of the variance matrices $\Sigma,\Sigma_1,\ldots,\Sigma_t$ are unknown we must either estimate them to produce empirical Bayes estimators or assume prior distributions for them and integrate them out of the model. There is clearly not enough information from which to estimate the variance matrices completely so we shall suppose that $\Sigma = \sigma^2 V$ and $\Sigma_i = \sigma_i^2 V_i$ for $i = 1,2,\ldots,t$. This assumption seems not unreasonable for the full Bayes estimators as well as for empirical Bayes estimators.

3.4.1 The Empirical Bayes Case

We might base our estimators for the variances on the maximum likelihood estimators for $\Lambda\beta$ given a set of prior estimates or guesses of the variances. This scheme could be used iteratively, the new weights providing a new estimator for $\Lambda\beta$ from which to compute new variance estimates. We shall not investigate the convergence of this scheme. An alternative method is to use the theory of MINQUE estimators. However,

these can lead to negative estimates as can the iterative scheme outlined above.

We shall only consider the case in which the prior knowledge is vague and the prior hypotheses are all testable in the original model. It is also convenient to suppose the prior hypotheses to be mutually orthogonal and that the Q of section 3.3 is zero. Since a non-zero Q can be considered to be the result of an earlier stage prior hypothesis, the latter assumption is not very restrictive. With the assumption of vague prior knowledge an estimate of $\Lambda\beta$ ignoring this knowledge is a good prior estimate upon which to base estimates of σ^2 and σ_i^2 for $i = 1, 2, \ldots, t$. These estimates will then be used to provide a better estimator for $\Lambda\beta$. It will not be necessary to proceed to a second iteration.

Since $\Sigma, \Sigma_1, \ldots, \Sigma_t$ are unknown we cannot calculate β°. Suppose we use $W, W_1, \ldots, W_t$ as estimates of $\Sigma^{-1}, \Sigma_1^{-1}, \ldots, \Sigma_t^{-1}$. We may then calculate an approximation, $\tilde{\beta}$, to β°. Let W_* be the block diagonal matrix $W_* = \mathrm{diag}(W_1, W_2, \ldots, W_t)$ and let $\Sigma_* = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_t)$. With $H^T = [H_1^T, \ldots, H_t^T]$, $h^T = [h_1^T, \ldots, h_t^T]$ and $S = X^T W X + H^T W_* H$ the vector $\tilde{\beta} = S^-(X^T W Y + H^T W_* h)$ estimates β when the variances are estimated by fixed matrices W^{-1}, W_*^{-1}. Since

$$Y - X\tilde{\beta} = (I - XS^-X^TW)Y - XS^-H^TW_*h$$

and

$$h - H\tilde{\beta} = -HS^-X^TWY + (I - HS^-H^TW_*)h$$

and since $\mathrm{var}\, Y = \Sigma$, $\mathrm{var}\, h = \Sigma_*$ and $\mathrm{cov}(Y,h) = 0$, we have

$$\mathrm{var}(Y - X\tilde{\beta}) = (I - XS^-X^TW)\Sigma(I - WXS^-X^T) + XS^-H^TW_*\Sigma_*W_*HS^-X^T$$

$$\mathrm{var}(h - H\tilde{\beta}) = HS^-X^TW\Sigma WXS^-H^T + (I - HS^-H^TW_*)\Sigma_*(I - W_*HS^-H^T)$$

and

$$\mathrm{cov}(Y - X\tilde{\beta}, h - H\tilde{\beta}) = -(I - XS^-X^TW)\Sigma WXS^-H^T - XS^-H^TW_*\Sigma_*(I - W_*HS^-H^T).$$

If, in fact, $W = \Sigma^{-1}$ and $W_* = \Sigma_*^{-1}$ then these reduce to

$$\mathrm{var}(Y - X\tilde{\beta}) = \Sigma - XS^-X^T$$

$$\mathrm{var}(h - H\tilde{\beta}) = \Sigma_* - HS^-H^T$$

and

$$\mathrm{cov}(Y - X\tilde{\beta}, h - H\tilde{\beta}) = -XS^-H^T.$$

Since we suppose the variance matrices to be known up to a multiplicative constant we shall suppose that $W = a\Sigma^{-1}$. Ignoring prior knowledge is equivalent to taking $W_* = 0$ in which case we obtain

$$\mathrm{var}(Y - X\tilde{\beta}) = \Sigma - aXS^-X^T = \Sigma - X(X^T\Sigma^{-1}X)^-X^T$$

$$\text{var}(h - H\tilde{\beta}) = \Sigma_* + aHS^{-}H^T = \Sigma_* + H(X^T\Sigma^{-1}X)^{-}H^T$$

and

$$\text{cov}(Y-X\tilde{\beta}, h-H\tilde{\beta}) = 0 .$$

Now taking $\Sigma = \sigma^2 V$ gives

$$E[(Y-X\tilde{\beta})^T V^{-1}(Y-X\tilde{\beta})] = \text{tr}\{I - X(X^TV^{-1}X)^{-}X^TV^{-1}\} = (n-p)\sigma^2$$

and

$$E[(h-H\tilde{\beta})^T\Sigma_*^{-1}(h-H\tilde{\beta})] = \text{tr}\{I + \sigma^2 H(X^TV^{-1}X)^{-}H^T\Sigma_*^{-1}\} .$$

The above calculations help us to calculate estimators for $\sigma^2, \sigma_1^2, \ldots, \sigma_t^2$ which will be substituted into the expressions for $\tilde{\beta}$ to provide an estimator for β.

We now wish to use the orthogonality of the hypotheses and the fact that $\Sigma_i = \sigma_i^2 V_i$ to calculate $E[(h_i - H_i\tilde{\beta})(h_j - H_j\tilde{\beta})^T]$. Clearly when $i \neq j$ this gives $H_i(X^T\Sigma^{-1}X)^{-}H_j^T$ which is zero by the orthogonality of the H_i. If $i = j$ then we obtain the value

$$\Sigma_i + \sigma^2 H_i(X^TV^{-1}X)^{-}H_i^T = \sigma_i^2 V_i + \sigma^2 H_i(X^TV^{-1}X)^{-}H_i^T$$

from which we see that

$$E[(h_i-H_i\tilde{\beta})^T V_i^{-1}(h_i-H_i\tilde{\beta})] = n_i\sigma_i^2 + \sigma^2\text{tr}\{(X^TV^{-1}X)^{-}H_i^T V_i^{-1}H_i\}$$

where n_i is the number of rows of H_i.

We may now give unbiased estimators for σ^2 and the σ_i^2. They are

$$\tilde{\sigma}^2 = \frac{1}{n-p}(Y - X\tilde{\beta})^T V^{-1}(Y - X\tilde{\beta})$$

and

$$\tilde{\sigma}_i^2 = \frac{1}{n_i}[(h_i-H_i\tilde{\beta})^T V_i^{-1}(h_i-H_i\tilde{\beta}) - \tilde{\sigma}^2\text{tr}\{(X^TV^{-1}X)^{-}H_i^T V_i^{-1}H_i\}]$$

If any of the $\tilde{\sigma}_i^2$ turn out to be negative then the corresponding variance σ_i^2 is likely to be small and an exact restriction is probably required. Alternatively a different prior distribution might give rise to positive estimates.

Using these estimates for the variance our empirical Bayes estimator for β will be given by

$$\beta^* = \{\frac{1}{\tilde{\sigma}^2}X^TV^{-1}X + \sum_{i=1}^{t}\frac{1}{\tilde{\sigma}_i^2}H_i^T V_i^{-1}H_i\}^{-}\{\frac{1}{\tilde{\sigma}^2}X^TV^{-1}X + \sum_{i=1}^{t}\frac{1}{\tilde{\sigma}_i^2}H_i^T V_i^{-1}h_i\} .$$

In order to deal with the case of exact restrictions we shall find the limiting value of this expression as some of the variances tend to zero. This is a generalisation of a method of Brook and Wallace(1973).

Suppose we let all the variances tend to zero except for those in the set $\{\sigma_i^2 : i \notin I\}$ where $I \subset \{1, 2, \ldots, t\}$. Thus for $i \notin I$ the restrictions are not exact. If $i \notin I$ then $H_i\beta$ is estimable and the columns of H_i^T are in the column space of X^T and of

$\frac{1}{\sigma^2} X^T V^{-1} X + \sum_{i \in I} \frac{1}{\sigma_i^2} H_i^T V_i^{-1} H_i$. We now use the formula for the inverse of the sum of matrices and let $\sigma_i^2 \to 0$ for $i \in I$. Let

$$A = \frac{1}{\sigma^2} X^T V^{-1} X + \sum_{i \in I} \frac{1}{\sigma_i^2} H_i^T V_i^{-1} H_i$$

$$a = \frac{1}{\sigma^2} X^T V^{-1} Y + \sum_{i \in I} \frac{1}{\sigma_i^2} H_i^T V_i^{-1} h_i$$

$B = \text{diag}\, \frac{1}{\tilde{\sigma}_i^2} V_i^{-1}$ for $i \notin I$, $H_-^T = [H_{i_1}^T, H_{i_2}^T, \ldots, H_{i_u}^T]$ and $h_-^T = [h_{i_1}^T, h_{i_2}^T, \ldots, h_{i_u}^T]$ where $\{i_1, i_2, \ldots, i_u\} = \{1,2,\ldots,t\} \setminus I$.

With these definitions we have

$$\beta^* = \{A^- - A^- H_-^T (B^{-1} + H_- A^- H_-^T)^{-1} H_- A^-\}(a + H_-^T B h_-).$$

Now if α is the largest eigenvalue of B^{-1} and if $B^{-1} = \alpha C$ then B^{-1} and α tend to zero together and (assuming each H_i has full row rank)

$$\beta^* \to \{A^- - A^- H_-^T (H_- A^- H_-^T)^{-1} H_- A^-\} a + \lim_{\alpha \to 0} \left\{ \frac{1}{\alpha} A^- H_-^T C h_- - \frac{1}{\alpha} A^- H_-^T (\alpha C^{-1} + H_- A^- H_-^T)^{-1} H_- A^- H_-^T C h_- \right\}.$$

The second term is equal to

$$\lim_{\alpha \to 0} \frac{1}{\alpha} A^- H_-^T (\alpha C^{-1} + H_- A^- H_-^T)^{-1} (\alpha C^{-1} + H_- A^- H_-^T - H_- A^- H_-^T) C h_-$$

$$= \lim_{\alpha \to 0} A^- H_-^T (\alpha C^{-1} + H_- A^- H_-^T)^{-1} h$$

$$= A^- H_-^T (H_- A^- H_-^T)^{-1} h .$$

Therefore $\beta^* \to A^- a - A^- H_-^T (H_- A^- H_-^T)^{-1} (H_- A^- a - h_-)$.

(We note that $A^- a$ is the estimator which ignores the exact restrictions).

The rule given resembles a positive part estimator in that negative estimates of a variance component imply shrinking the estimator for β onto the hyperplane corresponding to that component instead of over-shrinking as would otherwise be the case.

3.4.2 The Bayes Case

Instead of estimating the variance components we may choose a prior distribution for them and integrate them out of the model. Unfortunately this requires the use of numerical methods. We shall assume inverse gamma distributions for the variance components since this distribution is the conjugate distribution for this problem.

We shall suppose that

$$p(\sigma^2|\alpha,a^2) = (\tfrac{1}{2}\alpha)^{\alpha}\frac{a^{2\alpha}}{\Gamma(\alpha)}\frac{1}{(\sigma^2)^{\alpha+1}}\exp\left\{-\frac{1}{2}\frac{\alpha a^2}{\sigma^2}\right\}$$

and $$p(\sigma_i^2|\alpha_i,a_i^2) = (\tfrac{1}{2}\alpha_i)^{\alpha_i}\frac{(a_i)^{2\alpha_i}}{\Gamma(\alpha_i)}\frac{1}{(\sigma_i^2)^{\alpha_i+1}}\exp\left\{-\frac{1}{2}\frac{\alpha_i a_i^2}{\sigma_i^2}\right\}$$

all independently of one another and of β.

The special cases of these given by $\alpha \to 0$, $\alpha_i \to 0$ (in which case the normalising constants tend to zero and the distributions become improper) are the invariant priors given by Jeffreys.

Using the general result of section 3.3 that

$$p(\beta|Y,h_1,\ldots,h_t) \propto \text{const}(\sigma^2,\{\sigma_i^2\})\exp\{-\tfrac{1}{2}\|\beta-\beta^{\circ}\|^2_{Q+\sum_{i=1}^{t}H_i^T\Sigma_i^{-1}H_i+X^T\Sigma^{-1}X}\}$$

where

$$\beta^{\circ} = (Q+\sum_{i=1}^{t}H_i^T\Sigma_i^{-1}H_i+X^T\Sigma^{-1}X)^{-}(Q\alpha+\sum_{i=1}^{t}H_i^T\Sigma_i^{-1}h_i+X^T\Sigma^{-1}Y)$$

the joint density of β, σ^2 and the σ_i^2 is

$$p(\beta,\sigma^2,\{\sigma_i^2\}|Y,h_1,\ldots,h_t,\alpha,a^2,\{\alpha_i,a_i^2\})$$
$$\propto \frac{1}{(\sigma^2)^{\alpha+1}\prod_{i=1}^{t}(\sigma_i^2)^{\alpha_i+1}}\exp\left\{-\tfrac{1}{2}\|\beta-\beta^{\circ}\|^2_S-\frac{1}{2}\frac{\alpha a^2}{\sigma^2}-\frac{1}{2}\sum_{i=1}^{t}\frac{\alpha_i a_i^2}{\sigma_i^2}\right\}$$

where $$S = Q+\sum_{i=1}^{t}H_i^T\Sigma_i^{-1}H_i+X^T\Sigma^{-1}X$$

giving the marginal density for β,

$$p(\beta|Y,\{h_i\},\alpha,a^2,\{\alpha_i,a_i^2\})$$
$$\propto \int_0^{\infty}\cdots\int_0^{\infty}\exp\left\{-\tfrac{1}{2}\|\beta-\beta^{\circ}\|^2_S-\frac{\alpha a^2}{2\sigma^2}-\frac{1}{2}\sum_{i=1}^{t}\frac{\alpha_i a_i^2}{\sigma_i^2}\right\}\frac{d\sigma^2 d\sigma_1^2\ldots d\sigma_t^2}{(\sigma^2)^{\alpha+1}\prod_{i=1}^{t}(\sigma_i^2)^{\alpha_i+1}}.$$

Since the expression for β° is not a very simple function of $\sigma^2,\sigma_1^2,\ldots,\sigma_t^2$, this integral is intractable. In fact

$$\|\beta-\beta^{\circ}\|^2_S = \beta^T S\beta - 2\beta^T S\beta^{\circ}+\beta^{\circ T}S\beta^{\circ}$$
$$= \beta^T S\beta - 2\beta^T(Q\alpha+\sum_{i=1}^{t}H_i^T\Sigma_i^{-1}h_i+X^T\Sigma^{-1}Y)$$
$$+\beta^{\circ T}(Q\alpha+\sum_{i=1}^{t}H_i^T\Sigma_i^{-1}h_i+X^T\Sigma^{-1}Y).$$

It is the last term which gives rise to the intractibility of the integral since it contains S^-.

If we integrate in a different order then we can make the integral a little more tractable. The joint distribution of $Y,h_1,\ldots,h_t,\beta,\sigma^2$

and $\sigma_1,\ldots,\sigma_t$ is given by

$$p(Y,h_1,\ldots,h_t,\beta,\sigma^2,\sigma_1^2,\ldots,\sigma_t^2|\alpha,\alpha_1,\ldots,\alpha_t,a^2,a_1^2,\ldots,a_t^2)$$

$$\propto \frac{1}{(\sigma^2)^{\frac{1}{2}n+\alpha+1}\prod_{i=1}^{t}(\sigma_i^2)^{\frac{1}{2}n_i+\alpha_i+1}} \exp\left\{-\tfrac{1}{2}\left[\|Y-X\beta\|^2_{\Sigma^{-1}} + \sum_{i=1}^{t}\|h_i-X_i\beta\|^2_{\Sigma_i^{-1}}\right]\right\}$$

$$\times \exp\left\{-\tfrac{1}{2}\left[\frac{\alpha a^2}{\sigma^2} + \sum_{i=1}^{t}\frac{\alpha_i a_i^2}{\sigma_i^2}\right\}\right] p(\beta)$$

where $p(\beta) \propto \exp\{-\tfrac{1}{2}(\beta-\alpha)^T Q(\beta-\alpha)\}$.

Integrating with respect to the variances we obtain the marginal distribution,

$$p(Y,h_1,\ldots,h_t,\beta|\alpha_1,\ldots,\alpha_t,a_1^2,\ldots,a_t^2,\alpha,a^2)$$

$$\propto p(\beta)\,\Gamma(\alpha)\prod_{i=1}^{t}\Gamma(\alpha_i)\left(\|Y-X\beta\|^2_{V^{-1}}+\alpha a^2\right)^{-(\alpha+\frac{1}{2}n)}\prod_{i=1}^{t}\left(\|h_i-X_i\beta\|^2_{V_i^{-1}}+\alpha_i a_i^2\right)^{-(\alpha_i+\frac{1}{2}n_i)}$$

$$= f \quad \text{(say)}.$$

Usually Q will be taken to be zero in which case $p(\beta) = 1$. In this case this density is a product of multivariate t distributions. If Q is taken to be non-zero then it is likely to be known only up to an unknown scale factor in which case this could be integrated in the same way as the σ_i^2 to give another multivariate t factor.

The marginal distribution of β is not easy to find, if found, however, then this leads immediately to the conditional density

$$p(\beta|Y,h_1,\ldots,h_t,\alpha,a^2,\alpha_1,\ldots,\alpha_t,a_1^2,\ldots,a_t^2) \propto \frac{f}{\int\ldots\int f\,dy\,dh_1\,dh_2\ldots dh_t}.$$

Note that the integral is a multiple integral of dimension $n+n_1+n_2+\ldots+n_t$ and the expected value of β is the ratio of two such integrals.

As explained earlier, this may be calculated in one of two ways. The results of Dickey(1968) reduce the integral to one of dimension t, which for small t may be found numerically. Alternatively we can use the method of Tiao and Zellner(1964) which uses an asymptotic expansion for the integral. Either way the work involved is laborious and a simpler method is desirable. One approximate method is to use the mode of the posterior distribution given by $\frac{\partial f}{\partial \beta} = 0$. Taking $Q = 0$ we have

$$\frac{1}{f}\frac{\partial f}{\partial \beta} = \frac{\partial \log f}{\partial \beta} = 2(\alpha+\tfrac{1}{2}n)\{\|Y-X\beta\|^2_{V^{-1}} + \alpha a^2\}^{-1} X^T V^{-1}(Y-X\beta)$$

$$+ \sum_{i=1}^{t} 2(\alpha_i+\tfrac{1}{2}n_i)\{\|h_i-X_i\beta\|^2_{V_i^{-1}} + \alpha_i a_i^2\}^{-1} H_i^T V_i^{-1}(h_i-X_i\beta)$$

and this gives a solution satisfying

$$\left[\frac{\alpha+\frac{1}{2}n}{\|Y-X\hat{\beta}\|^2_{V^{-1}}+\alpha a^2}X^TV^{-1}X+\sum_{i=1}^{t}\frac{\alpha_i+\frac{1}{2}n_i}{\|h_i-X_i\hat{\beta}\|^2_{V_i^{-1}}+\alpha_i a_i^2}H_i^TV_i^{-1}H_i\right]\hat{\beta}$$

$$=\frac{\alpha+\frac{1}{2}n}{\|Y-X\hat{\beta}\|^2_{V^{-1}}+\alpha a^2}X^TV^{-1}Y+\sum_{i=1}^{t}\frac{\alpha_i+\frac{1}{2}n_i}{\|h_i-X_i\hat{\beta}\|^2_{V_i^{-1}}+\alpha_i a_i^2}H_i^TV_i^{-1}h_i\ .$$

This would need to be solved iteratively. The similarity to the eampirical Bayes solution should be noted. The difference lies in the variance estimates of the form

$$\frac{1}{2\alpha+n}\left\{\|Y-X\hat{\beta}\|^2_{V^{-1}}+\alpha a^2\right\}\quad\text{and}\quad\frac{1}{2\alpha_i+n_i}\left\{\|h_i-H_i\hat{\beta}\|^2_{V_i^{-1}}+\alpha_i a_i^2\right\}$$

which, for small α, would seem likely to have positive bias.

3.5 Comparison with Generalised James-Stein Estimators

When the respective variance matrices are known the Bayes estimators are of the form of a constant plus a matrix shrinkage of the least squares estimator. The estimator may also be regarded as a weighted average of the least squares solution and the exact values satisfying the hypotheses. Thus the Bayes estimator, like the generalised James-Stein estimators, shrink the usual estimators towards the prior hyperplanes. This also applies in the unknown variance case, however, the estimator is then extremely complicated in contrast to the James-Stein estimators. On grounds of tractability it seems better to use the latter, especially since they are "almost admissible".

Chapter 4

Minimum Mean Square Error Estimation

4.1 Introduction

In this chapter we shall consider the linear model

$$Y = X\beta + e, \quad E[e] = 0, \quad \text{var } e = \Sigma \quad \text{where} \quad \Sigma = \sigma^2 V$$

and V is unknown. We shall either suppose σ^2 to be unknown or shall take $\sigma^2 = 1$. (There is no need to first study the canonical form of the model as it is just as easy to deal with the model directly).

We shall compare estimators by using either the mean square error matrix or the weighted sum of squared errors. Suppose that $\beta^*(.)$ is an estimator for β and β^* is the corresponding estimate, then the former is given by

$$M(\beta^*(.), \beta) = E[(\beta^* - \beta)(\beta^* - \beta)^T]$$

and, for a weighting matrix W, the latter is given by

$$M_W(\beta^*(.), \beta) = E[(\beta^* - \beta)^T W(\beta^* - \beta)].$$

Another way of comparing estimators is by using the mean square error of prediction. Prediction may be performed at the points given by the matrix X, or at a set of future points. We shall only consider the former. The mean square error of prediction is given by

$$\begin{aligned} M_P(X\beta^*(.), X\beta) &= E[(X\beta^* - X\beta)(X\beta^* - X\beta)^T] \\ &= X\, M(\beta^*(.), \beta)\, X^T. \end{aligned}$$

If X has full column rank then this equality establishes a one-to-one correspondence between $M_P(X\beta^*(.), X\beta)$ and $M(\beta^*(.), \beta)$ and it is immaterial whether we measure the mean square error in the p-dimensional β-space or in the n-dimensional Y-space. For the same reason, this also applies to prediction of $X_1\beta$ if X_1 is a matrix of full column rank.

4.2 Comparison of Estimators

We may say that one estimator is better than another if its mean square error is smaller than the mean square error of the other. We shall make this precise in the following definitions.

Definition 4.1 Given two estimators for β, $\beta_1^*(.)$ and $\beta_2^*(.)$, we say that $\beta_1^*(.)$ is at least as good as $\beta_2^*(.)$ and write $\beta_1^*(.) \leq \beta_2^*(.)$ if and only if, for all p-vectors λ, $\lambda^T\beta_1^*(.)$ has scalar mean square error less than that for $\lambda^T\beta_2^*(.)$ as estimators for $\lambda^T\beta$.

Definition 4.2 Given two estimators for β, $\beta_1^*(.)$ and $\beta_2^*(.)$, we say that $\beta_1^*(.)$ is at least as good as $\beta_2^*(.)$ with respect to a symmetric positive definite matrix Q if and only if

$M_Q(\beta_1^*(.), \beta) \leq M_Q(\beta_2^*(.), \beta)$. We denote this by $\beta_1^*(.) \leq_Q \beta_2^*(.)$.

After defining some inequalities for matrices we may rewrite definition 4.1 in terms of $M(.,.)$.

Definition 4.3 Given two $p \times p$ matrices A and B, A is less than or equal to B (written $A \leq B$) if and only if $B - A$ is non-negative definite.

Definition 4.4 Given two $p \times p$ matrices A and B, A is less than or equal to B with respect to a positive definite matrix Q (written $A \leq_Q B$) if and only if $\text{tr } AQ \leq \text{tr } BQ$.

Now $\lambda^T\beta_1^*(.) \leq \lambda^T\beta_2^*(.)$ if and only if

$$E[(\lambda^T\beta_1^*(.) - \lambda^T\beta)^2] \leq E[(\lambda^T\beta_2^*(.) - \lambda^T\lambda)]$$

and this is equivalent to the statement that

$$\lambda^T M(\beta_1^*(.), \beta)\lambda \leq \lambda^T M(\beta_2^*(.), \beta)\lambda .$$

Thus an equivalent form of definition 4.1 is $\beta_1^*(.) \leq \beta_2^*(.)$ if and only if $M(\beta_1^*(.), \beta) \leq M(\beta_2^*(.), \beta)$.

Also, for a random vector t, $E[t^TQt] = \text{tr } E[tt^T]Q$. Thus, putting $\beta_1^* - \beta$ and $\beta_2^* - \beta$ in turn for t we see that $\beta_1^*(.) \leq_Q \beta_2^*(.)$ if and only if $M(\beta_1^*(.), \beta) \leq_Q M(\beta_2^*(.), \beta)$.

We shall now show that $\leq$ is a stronger partial ordering than $\leq_Q$. For all symmetric positive definite matrices Q, $A \leq_Q B$ implies $\text{tr } AQ \leq \text{tr } PQ$ implies $A \leq_Q B$. It follows by applying the result to $M(\beta_i^*(.), \beta)$ that $\beta_1^*(.) \leq \beta_2^*(.) \Rightarrow \beta_1^*(.) \leq_Q \beta_2^*(.)$.

4.3 Estimators with Minimum Mean Square Error

The usual estimator for β under the model considered is $\hat{\beta}(.)$ where $\hat{\beta}(Y) = (X^TV^{-1}X)^{-1}X^TV^{-1}Y$. This estimator is the best linear unbiased estimator and the generalised least squares estimator (which minimises the residual sum of squares) $(Y - X\beta^*)^TV^{-1}(Y - X\beta^*)$.

If, in addition, we assume a normally distributed error vector $e \sim N_n(0, \Sigma)$ then $\hat{\beta}(.)$ is the maximium likelihood estimator. The usual estimator also satisfies various other well known and desirable criteria but we shall show that, in terms of the criteria in definitions 4.1 and 4.2 it may be improved upon.

The following theorem gives two more desirable properties of the usual estimator. Both results are known, but the first is much more widely known than the second although the second result is at least as important.

Theorem 1 For both of the partial oderings for estimators $\leq$ and $\leq_Q$ the generalised least squares estimator minimises the mean square error among estimators in the class of linear unbiased estimators and in the class of linear estimators with bounded mean square error.

Proof Consider the class of estimators of the form $\beta^*(.)$ with $\beta^*(Y) = AY + c$. The mean square error is

$$\begin{aligned} M(\beta^*(.), \beta) &= E[(AY + c - \beta)(AY + c - \beta)^T] \\ &= E[((AX - I)\beta + c + Ae)((AX - I)\beta + c + Ae)^T] \\ &= ((AX - I)\beta + c)((AX - I)\beta + c)^T + A\Sigma A^T \end{aligned}$$

since $E[e] = 0$ and $E[ee^T] = \Sigma$.

We shall now show that, for $\beta^*(.)$ to be in either of the classes considered, we must have $c = 0$ and $AX = I$. Firstly, $\beta^*(.)$ is unbiased if and only if for all β

$$E[\beta^*(Y)] = E[AY + c] = AX\beta + c = \beta .$$

This implies that $c = 0$ and $AX = I$.

Secondly, $\beta^*(.)$ has bounded mean square error if and only if $AX = I$ - for, if not, then we may choose λ such that $\lambda^T(AX - I) = \mu^T \neq 0^T$, and we may choose $\beta = t\mu$ in which case $\lambda^T((AX - I) + c)((AX - I) + c)^T\lambda = (t\mu^T\mu + \lambda^T c)^2$ which is unbounded as a function of t since $\mu^T\mu \neq 0$.

If Q is positive definite then we may write $Q = \sum_i \lambda_i\lambda_i^T$ and $\operatorname{tr} MQ = \sum_i \lambda_i^T M \lambda_i$. Since the λ_i form a basis for the column space of Q, there exists i such that $\lambda_i^T (AX - I) \neq 0$. We conclude that $AX \neq I$ implies unboundedness of the mean square error matrix and of the Q-weighted mean square error.

Now, if $AX = I$ then M and $\operatorname{tr} MQ$ will be minimised if $c = 0$. We now show that $A = (X^T V^{-1} X)^{-1} X^T V^{-1}$ gives the minimum mean square error of estimators in the class $\{\beta^*(.): \beta^*(Y) = AY \text{ and } AX = I\}$.

Let $A = (X^T V^{-1} X)^{-1} X^T V^{-1} + K$ and $KX = 0$.

Since $(X^T V^{-1} X)^{-1} X^T V^{-1} V K^T = (X^T V^{-1} X)^T X^T K^T = 0$ this gives

$$M(\beta^*(.), \beta) = \sigma^2(X^T V^{-1} X)^{-1} X^T V^{-1} V V^{-1} X(X^T V^{-1} X)^{-1} + \sigma^2 K V K^T .$$

For either partial ordering this is minimised if $K = 0$. This completes the proof.

Note that if $D \neq I$ then $D\hat{\beta}(.)$ has unbounded mean square error (assuming that D is a non-stochastic matrix).

4.4 Unrestricted Minimum Mean Square Error Estimation

The importance of the bounded mean square error part of theorem 4.1 is that it shows that no linear estimator may have uniformly smaller

mean square error than $\hat{\beta}(.)$. This is so because, if an estimator were to dominate $\hat{\beta}(.)$ then it would have bounded mean square error and this would contradict the conclusion of theorem 4.3.1. Therefore, for uniformly better estimators, we need to consider non-linear estimators. In the case of normally distributed errors it is necessary to reject the assumption of unbiasedness since the usual estimator is minimum variance unbiased.

Theil(1971) attempted to find an estimator of the form $\beta^*(.)$ where $\beta^*(Y) = AY$ with A such that this gives an estimator with uniformly minimum mean square error. Although we have just shown this to be impossible, it is nevertheless interesting to make the attempt. If we do then it turns out that the optimal value of A depends on β and Σ and does not therefore give an estimator for β. However we do obtain a lower bound on the mean square error of any linear estimator and also, substituting estimators for β and Σ into the expression for A, leads to the discovery of non-linear estimators which do have uniformly smaller mean square error than the usual estimator.

The mean square error of $\beta^*(.)$ is

$$\begin{aligned} M(\beta^*(.),\beta) &= (AX - I)\beta\beta^T(AX - I)^T + A\Sigma A^T \\ &= A(\Sigma + X\beta\beta^T X^T)A^T - \beta\beta^T X^T A^T - AX\beta\beta^T + \beta\beta^T \\ &= \{A-\beta\beta^T X^T(\Sigma+X\beta\beta^T X^T)^{-1}\}(\Sigma+X\beta\beta^T X^T)\{A-\beta\beta^T X^T(\Sigma+X\beta\beta^T X^T)^{-1}\} \\ &\quad + \beta\beta^T - \beta\beta^T X^T(\Sigma + X\beta\beta^T X^T)^{-1}X\beta\beta^T \\ &\geq \beta\beta^T - \beta\beta^T X^T(\Sigma + X\beta\beta^T X^T)^{-1}X\beta\beta^T. \end{aligned}$$

The above inequality also applies to $\geq_Q$ and equality holds if

$$A = \beta\beta^T X^T(\Sigma + (X\beta\beta^T X^T)^{-1}).$$

The only case for which this leads to an estimator for β occurs if $(X^T\Sigma^{-1}X)^{\frac{1}{2}}$ is known and $\Sigma = \sigma^2 V$ with V known. We shall see this when we have simplified the expression for A. This expression can be simplified using the formula for the inverse of the sum of matrices

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I + CA^{-1}B)^{-1}CA^{-1}.$$

Applying this formula gives

$$\begin{aligned} A &= \beta\beta^T X^T(\Sigma + (X\beta\beta^T X^T)^{-1}) \\ &= \beta\beta^T X^T\{\Sigma^{-1} - \Sigma^{-1}X\beta(1 + \beta^T X^T\Sigma^{-1}X\beta)^{-1}\beta^T X^T\Sigma^{-1}\} \\ &= \beta\beta^T X^T\Sigma^{-1} - \frac{\beta^T X^T\Sigma^{-1}X\beta}{1 + \beta^T X^T\Sigma^{-1}X\beta}\beta\beta^T X^T\Sigma^{-1} \\ &= \frac{1}{1 + \beta^T X^T\Sigma^{-1}X\beta}\beta\beta^T X^T\Sigma^{-1}. \end{aligned}$$

Substituting this value for A into the formula for the mean square error gives

$$M(\beta^*(.), \beta) = \beta\beta^T - \frac{1}{1 + \beta^T X^T \Sigma^{-1} X \beta} \beta\beta^T X^T \Sigma^{-1} X \beta\beta^T$$

$$= \frac{1}{1 + \beta^T X^T \Sigma^{-1} X \beta} \beta\beta^T .$$

Also we obtain

$$\beta^* = AY = \frac{\beta^T X^T \Sigma^{-1} X \hat{\beta}}{1 + \beta^T X^T \Sigma^{-1} X \beta} \beta$$

where $\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$ is the usual estimator for β. We shall write $D = X^T \Sigma^{-1} X$ so that

$$\beta^* = D^{-\frac{1}{2}} \frac{\gamma^T \hat{\gamma}}{1 + \gamma^T \gamma} \gamma \quad \text{where} \quad \gamma = D^{\frac{1}{2}} \beta \quad \text{and} \quad \hat{\gamma} = D^{\frac{1}{2}} \hat{\beta} .$$

Now let $C = X^T V^{-1} X = \sigma^2 D$. This gives $\beta^* = C^{-\frac{1}{2}} \frac{\gamma^T C^{\frac{1}{2}} \hat{\beta}}{1 + \gamma^T \gamma} \gamma$ which depends only on γ, C and $\hat{\beta}$. If γ and C are known then this implies that β^* is an estimator for β. If not then β^* depends on the unknown parameters and is not an estimator since it is not an observable random variable.

Some authors find an apparent contradiction in a random vector of the form AY depending on β, having minimum mean square error for β and yet *not* being equal to β. In fact there isno difficulty since β^* is not of minimum mean square error in the class of *all* random vectors which depend on β - it has minimum mean square error in a class of which β *is not a member.* To see this note that, unless $Y = X\beta$, $\beta \in \{\beta^*: \beta^* = AY\} \Rightarrow \exists\ A(\beta)$ such that $\beta = AY = AX\beta + Ae$ which is impossible unless $AX = I$ and $e = 0$. If, on the other hand, we consider the wider class $\{\beta^*: \beta^* = AY + c\}$ then we find that β *does* belong to this class and *does* have minimum mean square error.

We next consider the class of estimators of the form $\beta^\times = \alpha\hat{\beta} = \alpha(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$. If we substitute $\alpha(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}$ for A in the expression for the mean square error then we obtain

$$M(\beta^\times(.), \beta) = \alpha^2 (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \Sigma \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1}$$
$$+ \{\alpha(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} X - I\} \beta\beta^T \{\alpha(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} X - I\}$$
$$= \alpha^2 D^{-1} + (\alpha - 1)^2 \beta\beta^T .$$

This gives a weighted mean square error of

$$M_W(\beta^\times(.), \beta) = \alpha^2 \operatorname{tr} D^{-1} W + (\alpha - 1)^2 \beta^T W \beta .$$

It is clear that there is no uniform minimum for the former expression since $t^T M(\beta^{\times}(.), \beta)\, t = \alpha^2 t^T D^{-1} t + (\alpha-1)^2 (t^T \beta)^2$ has its minimum at $\alpha = \dfrac{(t^T \beta)^2}{t^T D^{-1} t + (t^T \beta)^2}$ which is not independent of t. However, for any given W, the latter expression does have a uniform minimum value of

$$\frac{\beta^T W \beta \operatorname{tr} D^{-1} W}{\operatorname{tr} D^{-1} W + \beta^T W \beta} \quad \text{at} \quad \alpha = \frac{\beta^T W \beta}{\operatorname{tr} D^{-1} W + \beta^T W \beta} .$$

An important special case occurs when $W = D$ so that

$$\alpha = \frac{\beta^T D \beta}{p + \beta^T D \beta} = \frac{\beta^T C \beta}{p\sigma^2 + \beta^T C \beta} .$$

We have now derived two random variables

$$\beta^* = \frac{\beta \beta^T C}{\sigma^2 + \beta^T C \beta} \hat{\beta} \quad \text{and} \quad \beta^{\times} = \frac{\beta^T C \beta}{p\sigma^2 + \beta^T C \beta} \hat{\beta}$$

the former being a matrix shrinkage of $\hat{\beta}$ and the latter a scalar shrinkage.

4.5 More Random Vectors

As noted previously, unless $D^{-\frac{1}{2}}\beta$ is known, it is necessary to substitute estimators for β and σ^2 into the expressions for A or α in order to turn β^* into an estimator for β. If we do then we can no longer be certain that the estimator has smaller mean square error then the usual estimator. This is because the above proof was based on the assumption that A and α are non-stochastic. The following example shows that the mean square error may be either increased or reduced by substituting estimators for unknown parameters.

Let X be a univariate random variable with expectation θ. The variate $\alpha X + (1-\alpha)\theta$ has expectation θ and mean square error α^2 var X. Substituting X for θ gives the estimator X which has mean square error var X. Thus the mean square error is increased if $|\alpha| < 1$ and reduced if $|\alpha| > 1$.

Returning to our original problem, we see that, when we substitute estimators for β and σ^2, we will have to recalculate the mean square error. A motivation for the use of shrinkage estimators is that $\hat{\beta}$ is too long on average, but, in a sense, has the right direction. This is clear because, as noted in Brook and Moore(1979), by Jensen's inequality $E[\|\hat{\beta}\|^2] \geq (E[\|\hat{\beta}\|])^2 \geq \|E[\hat{\beta}]\|^2$ for any norm.

A scalar shrinkage is therefore indicated. Now, if a scalar shrinkage of $\hat{\beta}$ is used in estimating A, then the resulting estimator for is also a scalar shrinkage of $\hat{\beta}$. Although a scalar shrinkage seems reasonable we shall also consider matrix shrinkages.

Let us write the expression for β^* as $\beta^* = \beta^\dagger + L\hat{\beta}$

where $\beta^\dagger = \dfrac{\beta^T D \beta}{1 + \beta^T D \beta}$ and $L = \dfrac{\beta\beta^T D - \beta^T D \beta I}{1 + \beta^T D \beta}$.

Now, if $\hat{\beta} = l\beta + \varepsilon$ then $\beta\hat{\beta}^T - \hat{\beta}\beta^T = \beta\varepsilon^T - \varepsilon\beta^T$

and $L\hat{\beta} = \dfrac{a^2}{1+a^2}(\delta\varepsilon^T - \varepsilon\delta^T)D\delta = \dfrac{a^2}{1+a^2}(\delta\delta^T D - I)\varepsilon$

where $a = (\beta^T D \beta)^{\frac{1}{2}}$ and $\delta = \dfrac{1}{a}\beta$.

It is possible to choose l and ε so that ε is orthogonal to β with respect to the inner product $\langle a,b\rangle_C = a^T C b$ in which case $L\hat{\beta}$ is orthogonal to β and $L\hat{\beta} = \dfrac{a^2}{1+a^2}(-\varepsilon)$. As we have no information about ε (except that, in a sense, $\hat{\beta}$ is in the right direction on average) it seems reasonable to set $L\hat{\beta}$ to zero and use the estimate $\beta^* = \beta^\dagger$.

Farebrother(1975) suggested another way of making β^* into a scalar shrinkage. This is done in such a way as to illustrate a connection between the minimum mean square error estimator and the ridge regression estimator. We shall present a slight generalisation. Let Q be a symmetric positive definite matrix and write

$$\beta^* = \frac{\beta\beta^T D}{1 + \beta^T D \beta}\left(D + \frac{Q}{\beta^T Q \beta}\right)\left(D + \frac{Q}{\beta^T Q \beta}\right)^{-1} D\hat{\beta}.$$

Putting $\beta^\ddagger = \left(D + \dfrac{Q}{\beta^T Q \beta}\right)^{-1} D\hat{\beta}$ gives $\beta^* = \dfrac{\beta\beta^T}{1 + \beta^T D\beta}\left(D + \dfrac{Q}{\beta^T Q\beta}\right)\beta^\ddagger$.

If $\beta^\ddagger$ is proportional to β then, putting $\beta^\ddagger = \alpha\beta$, we obtain

$$\beta^* = \alpha\frac{\beta\beta^T D\beta}{1 + \beta^T D\beta} + \alpha\frac{\beta\beta^T Q\beta}{(1 + \beta^T D\beta)\beta^T Q\beta}$$

$$= \alpha\beta = \beta^\ddagger.$$

In the case $p = 1$, $\beta^\ddagger$ is certainly proportional to β. Although this is not necessarily so in higher dimensions, $\beta^\ddagger$ may be suggested as a replacement for β^*. This is of the same form as a generalised ridge regression estimator for β the purpose of which is to give more stable estimators for β when the X matrix is ill conditioned. This variate is not orthogonal under orthogonal transformations of the

parameter space unless Q is proportional to D, but in this case $\beta^{\ddagger} = \beta^{\dagger}$ and this leads to numerically unstable estimators in ill conditioned problems.

We may divide β^{*} into components as follows:

$$\beta^{*} = \beta^{\ddagger} + K\beta^{\ddagger} \quad \text{where} \quad K = \left(\frac{\beta\beta^{T}D}{\beta^{T}D\beta} - I\right)\frac{\beta^{T}D\beta}{1+\beta^{T}D\beta} + \left(\frac{\beta\beta^{T}Q}{\beta^{T}Q\beta} - I\right)\frac{1}{1+\beta^{T}D\beta}.$$

Now $K\beta = 0$ so replacing $\beta^{\ddagger}$ by β in the second term gives $\beta^{*} \doteqdot \beta^{\ddagger}$. This further supports the idea of using $\beta^{\ddagger}$ as a replacement for β^{*}.

We have now produced four random variables from each of which we can construct estimators for β. They are

$$\beta^{*} = \frac{\beta\beta^{T}C}{\sigma^{2}+\beta^{T}C\beta}\hat{\beta} \qquad \beta^{\times} = \frac{\beta^{T}C\beta}{p\sigma^{2}+\beta^{T}C\beta}\hat{\beta}$$

$$\beta^{\dagger} = \frac{\beta^{T}C\beta}{\sigma^{2}+\beta^{T}C\beta}\hat{\beta} \qquad \beta^{\ddagger} = \left(C + \frac{Q\sigma^{2}}{\beta^{T}Q\beta}\right)^{-1}C\hat{\beta}.$$

An interesting way of choosing the shrinkage matrix in ridge regression was given by Strawderman(1978). He showed that his estimator is minimax for a different quadratic loss function from those used in practice. That is he used a form of weighted mean square error, the weighting matrix being C^{2} instead of C or I. Thus a consequence of using ridge regression, at least if Strawderman's form is used, is that the estimator is good for a form of loss function not likely to be considered by the user.

In the next section we shall discuss the estimation of the shrinkage factors in the first three of the random vectors given above.

4.6 Estimating the Shrinkage Factor

At each particular value of β and Σ the mean square errors of the random vectors in section 4.4 are the minimum attainable with linear estimators and they are only achieved when the shrinkage happens to take the ideal value. Estimating the shrinkage will result in a different value for the mean square error - usually a greater value. If we are interested in weighted mean square error with weighting matrix, $W = C^{-1}$, then it seems more sensible to study realisations of $\beta^{\times}$. We shall, however, consider β^{*} and $\beta^{\dagger}$ as well, but, as $\beta^{\ddagger}$ is more appropriate to a study of robustness, we shall consider it no further.

The naïve estimator for the shrinkage factos may be obtained by

substituting the usual estimators for β and σ^2 into them. Thus using $\hat{\beta} = C^{-1} X^T V^{-1} Y$ and $\hat{\sigma}^2 = \frac{1}{n-p} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta})$ gives the following estimators for β:

$$\hat{\beta}^\dagger = \hat{\beta}^* = \frac{\hat{\beta}^T C \hat{\beta}}{\hat{\beta}^T C \hat{\beta} + \frac{1}{n-p} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta})} \hat{\beta}$$

$$\hat{\beta}^\times = \frac{\hat{\beta}^T C \hat{\beta}}{\hat{\beta}^T C \hat{\beta} + \frac{p}{n-p} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta})} \hat{\beta} .$$

Our aim, of course, is to find the best estimator for β that we can, and to do so we estimate the shrinkage as accurately as possible. It is reasonable to suppose that better estimators of the numerator and denominator of the shrinkage factor will give better estimators for the shrinkage itself. However, while this may often be true there may also be efficient estimators for the components which lead to inefficient estimators for the ratio. Although $\hat{\beta}(.)$ and $\hat{\sigma}^2(.)$ are unbiased and quite efficient for β and σ^2 ($\frac{n-p}{n-p+2} \hat{\sigma}^2(.)$ is biased but more efficient for σ^2), they do not lead to efficient unbiased estimators for constants of the form $B\beta^T C \beta + A\sigma^2$. We shall divide the shrinkage into component parts in a number of ways and try to find improved estimators for these components. (Later in this chapter we shall briefly discuss another approach in which the bias in the numerator is adjusted to compensate for the bias in the denominator and vice-versa).

In the following expressions for β^* the parts contained in square brackets are to be treated as a whole for estimation purposes. If these parts contain $\hat{\beta}$ then this vector is treated as known since we are only trying to estimate the shrinkage (which, in the form given in this chapter does not contain $\hat{\beta}$). We may write

$$\beta^* = \left[\frac{\beta\beta^T}{\sigma^2 + \beta^T C \beta}\right] C\hat{\beta} = \left[\frac{\hat{\beta}^T C \beta}{\sigma^2 + \beta^T C \beta}\right][\beta] = \left[\frac{\beta^T C \beta}{\sigma^2 + \beta^T C \beta}\right]\hat{\beta} + \left[\frac{\beta\hat{\beta}^T - \hat{\beta}\beta^T}{\sigma^2 + \beta^T C \beta} C\beta\right]$$

$$= \frac{[\beta\beta^T]}{[\sigma^2 + \beta^T C \beta]} C\hat{\beta} = \frac{[\hat{\beta}^T C \beta]}{[\sigma^2 + \beta^T C \beta]}[\beta] = \frac{[\beta^T C \beta]}{[\sigma^2 + \beta^T C \beta]}\hat{\beta} + \left[\frac{\beta\hat{\beta}^T - \hat{\beta}\beta^T}{\sigma^2 + \beta^T C \beta} C\beta\right]$$

$$= \frac{[\beta\beta^T/\sigma^2]}{[1+\beta^T C \beta/\sigma^2]} C\hat{\beta} = \frac{[\beta^T C \beta/\sigma^2]}{[1+\beta^T C \beta/\sigma^2]}[\beta] = \frac{[\beta^T C \beta/\sigma^2]}{[1+\beta^T C \beta/\sigma^2]}\hat{\beta} + \left[\frac{\beta\hat{\beta}^T - \hat{\beta}\beta^T}{\sigma^2 + \beta^T C \beta} C\beta\right]$$

$$= \frac{[\beta\beta^T/\beta^T C \beta]}{1+[\sigma^2/\beta^T C \beta]} C\hat{\beta} = \frac{[\hat{\beta}^T C \beta/\beta^T C \beta]}{1+[\sigma^2/\beta^T C \beta]}[\beta] = \frac{1}{1+[\sigma^2/\beta^T C \beta]}\hat{\beta} + \left[\frac{\beta\hat{\beta}^T - \hat{\beta}\beta^T}{\sigma^2 + \beta^T C \beta} C\beta\right].$$

In addition we may consider the form of the minimum mean square error variate before simplification and obtain

$$\beta^* = [\beta\beta^T/\sigma^2]X^T(V + X[\beta\beta^T/\sigma^2]X^T)^{-1}Y$$
$$= [\beta\beta^T]X^T([\sigma^2]V + X[\beta\beta^T]X^T)^{-1}Y .$$

These have the advantage of treating β in the same way in the numerator and denominator (β always occurs in the factor $\beta\beta^T$ and never in the factor $\beta^T C \beta$).

We can decompose the scalar shrinkage factor in a similar way to obtain

$$\beta^\times = \left[\frac{\beta^T C \beta}{\sigma^2 + \beta^T C\beta}\right]\hat{\beta} = \frac{[\beta^T C \beta]}{[\sigma^2 + \beta^T C\beta]}\hat{\beta} = \frac{[\beta^T C \beta/\sigma^2]}{[1+\beta^T C\beta/\sigma^2]}\hat{\beta} = \frac{1}{1+[\sigma^2/\beta^T C\beta]}\hat{\beta} .$$

(Decomposition of $\beta^\dagger$ leads to the first terms in the last column in the expressions for the decomposition of β^*). In the decompositions of β^* which involve two terms, the first term is $\beta^\dagger$ and the second term will be estimated to be zero.

A further class of estimators may be obtained by using the improved estimators for β to improve the estimation of the shrinkage factor. Repeating the process leads to iterating to convergence. We shall do this in chapter 5.

From now on we shall not differentiate between an estimator and an estimate leaving it to the context to determine which we mean in each particular case. We now wish to estimate $\beta\beta^T$ and similar expressions. Using $\hat{\beta}\hat{\beta}^T$ we find that $E[\hat{\beta}\hat{\beta}^T] = \beta\beta^T + \sigma^2 C^{-1}$. We shall find that by removing some of the bias we can reduce the mean square error. Similar remarks apply to estimation of other expressions such as $\beta^T C \beta$ and $\beta^T C \beta/\sigma^2$. In the next section we shall discuss this in detail.

4.6.1 Estimating the Components of the Shrinkage

We first calculate $E[\hat{\beta}\hat{\beta}^T]$. We have

$$E[\hat{\beta}\hat{\beta}^T] = E[C^{-1}X^T V^{-1} Y Y^T V^{-1} X C^{-1}] = C^{-1}X^T V^{-1}(\sigma^2 V + X^T \beta\beta^T X)V^{-1}XC^{-1}$$
$$= \beta\beta^T + \sigma^2 C^{-1}.$$

It follows that

$$E[\hat{\beta}\hat{\beta}^T C] = \beta\beta^T C + \sigma^2 I \ , \quad E[C^{\frac{1}{2}}\hat{\beta}\hat{\beta}^T C^{\frac{1}{2}}] = C^{\frac{1}{2}}\beta\beta^T C^{\frac{1}{2}} + \sigma^2 I \quad \text{and}$$
$$E[\hat{\beta}^T C \hat{\beta}] = \beta^T C \beta + p\sigma^2.$$

An unbiased estimator for $\beta\beta^T C$ is therefore $\hat{\beta}\hat{\beta}^T C - \hat{\sigma}^2 I$ and an unbiased estimator for $\beta^T C \beta$ is $\hat{\beta}^T C \hat{\beta} - p\hat{\sigma}^2$.

The former may give estimates which are negative definite and the latter may give negative estimates. In the latter case the probability of negative estimates when $\beta^T C \beta = 0$ is approximately 0.5 and tends to zero as $\beta^T C \beta$ tends to infinity. It seems reasonable to replace a negative or negative definite estimate by zero. This gives a smoothed preliminary test estimator akin to the positive part version of the James-Stein estimator.

As we shall see, it is possible to find estimators with smaller mean square error than either the naïve estimators $\hat{\beta}\hat{\beta}^T C$ and $\hat{\beta}^T C \hat{\beta}$ or the unbiased estimators $\hat{\beta}\hat{\beta}^T C - \hat{\sigma}^2 I$ and $\hat{\beta}^T C \hat{\beta} - p\hat{\sigma}^2$. In a similar manner we may also discuss estimation of $\frac{1}{\sigma^2}\beta$, $\frac{1}{\sigma}\beta$, $\beta^T C\beta/\sigma^2$, $\sigma^2/\beta^T C\beta$, $\beta\beta^T C/\sigma^2$ and $\beta\beta^T C/\beta^T C\beta$. In order to calculate the mean square errors of the estimators of this section we need to make some distributional assumptions. We shall suppose that the error term, e, in the linear model is distributed as $e \sim N(0,\sigma^2 V)$. In this case $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ independently of $\hat{\beta}$.

When estimating the above functions of the parameters we should like to use estimators which are consistent. Unfortunately this is not always possible. In the next section we shall discuss another concept, relative consistency, which is a more desirable concept and is often achievable when consistency is not.

4.6.2 Relatively Consistent Estimators

Given a sequence of linear models $Y_n = X_n\beta + e_n$, $E[e_n] = 0_n$ var $e_n = \sigma^2 V_n$ we wish to estimate β and σ^2. Graybill(1976) gives a definition of consistency for this case which is not general enough for a discussion of consistency of estimation of $\beta^T X_n^T V_n^{-1} X_n \beta$ since this is not fixed as n increases. Graybill shows that $\hat{\beta}$ is mean square error consistent for β if and only if $(X_n^T V_n^{-1} X_n)^{-1} \to 0$ as $n \to \infty$, and that $\hat{\sigma}^2$ is always mean square error consistent for σ^2. If $(X_n^T V_n^{-1} X_n)^{-1} \to 0$ as $n \to \infty$, then $\beta^T X_n^T V_n^{-1} X_n \beta \to \infty$ as $n \to \infty$. We shall say that a sequence of estimators $\{\hat{\theta}_n\}$ is a mean square error consistent estimator for a sequence of parameters $\{\theta_n\}$ if and only if $E[(\hat{\theta}_n - \theta_n)^2] \to 0$ as $n \to \infty$.

Clearly this is not a very strict requirement if $\theta_n \to 0$ as $n \to \infty$, while it is much stricter if $\theta_n \to \infty$ as $n \to \infty$. A sequence of estimators $\{\hat{\theta}_n\}$ is said to be relatively consistent for the sequence of parameters $\{\theta_n\}$ if and only if $\frac{1}{\theta_n^2} E[(\hat{\theta}_n - \theta)^2] \to 0$ as $n \to \infty$. It is clear that the two definitions of consistency given above can be

interpreted as consistency with respect to the loss functions $\mathcal{L}(\hat{\theta}_n, \theta_n) = (\hat{\theta}_n - \theta_n)^2$ and $\mathcal{L}(\hat{\theta}_n, \theta_n) = \frac{1}{\theta_n^2}(\hat{\theta}_n - \theta_n)^2$.

These definitions easily extend to estimation of vectors and matrices of parameters by replacing the square by a norm, in the case of a vector of parameters the Euclidean norm, $\|a\| = (\Sigma a_i)^{\frac{1}{2}}$, and in the case of a matrix of parameters the Froebenius norm, $\|A\| = (\Sigma a_{ij})^{\frac{1}{2}}$.

We also define the relative variance of $\hat{\theta}_n$ to be $\frac{1}{\|\theta_n\|^2} E[\|\hat{\theta}_n - E[\hat{\theta}_n]\|^2]$ and the relative bias to be $\frac{1}{\|\theta_n\|}(E[\hat{\theta}_n] - \theta)$ where $\|\theta\| = |\theta|$ if θ is a scalar or the Euclidean or Froebenius norms if θ is a vector or a matrix. It is almost trivial that a sequence of estimators is relatively mean square error consistent if and only if the relative variance and relative bias tend to zero as $n \to \infty$.

4.6.3 Estimation of $B\beta^T C\beta + A\sigma^2$

We shall consider the cases $A, B \geq 0$. The expected value and variance of $b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2$ are

$$E[b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2] = b\beta^T C\beta + (a+bp)\sigma^2 \quad \text{and}$$
$$\text{var}(b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2) = 4b^2\sigma^2\beta^T C\beta + (2b^2p + \frac{2a^2}{n-p})\sigma^4.$$

These expressions are deduced from the moments of the central and non-central χ^2 distributions given in appendix 2. The mean square error will be

$$\begin{aligned} MSE &= 4b^2\sigma^2\beta^T C\beta + (2b^2p + \frac{2a^2}{n-p})\sigma^4 + (b-B)^2(\beta^T C\beta)^2 + (a+bp-A)^2\sigma^4 \\ &\quad + 2(b-B)(a+bp-A)\sigma^2\beta^T C\beta \\ &= \sigma^4\{[2b^2p + \frac{2a^2}{n-p} + (a+bp-A)^2] + [2(b-B)(a+bp-A) + 4b^2]\times 2\lambda \\ &\quad + (b-B)^2\times 4\lambda^2\} \end{aligned}$$

where $\lambda = \frac{\beta^T C\beta}{2\sigma^2}$. Now, for large λ this is minimised by putting $b = B$ and $a = \frac{n-p}{n-p+2}(A - Bp)$ (the latter minimises the first term which is already insignificant for large λ). For small λ we require that $a \simeq \frac{n-p}{n-p+2}(A - bp)$. It seems reasonable to take $b = B$ so that the mean square error remains bounded as $\lambda \to \infty$ and this gives, for small λ, $a = \frac{n-p}{n-p+2}(A - Bp)$. If $B = 0$ and $A = 1$ then we are estimating σ^2. Taking $b = 0$ and $a = \frac{n-p}{n-p+2}$ gives the well known result that $\frac{n-p}{n-p+2}\hat{\sigma}^2$ has minimum mean square error among all estimators of the form $a\hat{\sigma}^2$ for σ^2.

If $A = 0$ and $B = 1$ then we are estimating $\beta^T C \beta$ and taking $b = 1$ and $a = \frac{n-p}{n-p+2}(A - p)$ gives smaller mean square error than the naïve estimator. However, choosing a and b so that the first two terms in the expression for the mean square error are zero gives bounded relative mean square error as $\lambda \to 0$, otherwise it is unbounded. Now, if $C^{-1} \to 0$ as $n \to \infty$ then $\beta^T C \beta \to \infty$ as $n \to \infty$; thus if $B \neq 0$ then $b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2$ is relatively consistent for $b\beta^T C\beta + A\sigma^2$ if $b \to B$ as $n \to \infty$, while for $B \neq 0$ ordinary mean square error consistency is not attainable. If $B = 0$ then we have consistent estimation and relatively consistent estimation if $a \to A$ as $n \to \infty$.

4.6.4 Estimation of $\beta^T C\beta/\sigma^2$

Using the moments of the central and non-central χ^2-distribution in appendix 2 and the independence of $\hat{\beta}$ and $\hat{\sigma}^2$ we obtain

$$E[b\,\hat{\beta}^T C\hat{\beta}/\hat{\sigma}^2 + a] = b\,\frac{n-p}{n-p-2}\,(\beta^T C\beta/\sigma^2 + p) + a \qquad \text{and}$$

$$\operatorname{var}(b\,\beta^T C\beta/\sigma^2 + a) = b^2p^2\,\frac{2(n-p)^2}{p^2(n-p-2)}\left\{\frac{(p+2\lambda)^2}{(n-p-2)(n-p-4)} + \frac{p+4\lambda}{n-p-4}\right\}.$$

These results are also attainable from the mean and variance of the non-central F distribution.

We therefore obtain the mean square error

$$\mathrm{MSE} = \frac{2b^2(n-p)^2}{n-p-2}\left\{\frac{(p+2\lambda)^2}{(n-p-2)(n-p-4)} + \frac{p+4\lambda}{n-p-4}\right\} + \left\{\frac{b(n-p)}{n-p-2}(p+2\lambda) + a - 2\lambda\right\}^2.$$

Our estimator is unbiased if $b = \frac{n-p-2}{n-p}$ and $a = -p$. Only for this value of b can we minimise the mean square error by a suitable choice of a, this choice being $a = -p$. The resulting estimator has unbounded mean square error as $\lambda \to \infty$ and unbounded relative mean square error as $\lambda \to 0$. Since $\lambda \to \infty$ as $n \to \infty$ we have relative mean square error consistency so long as $b \to 1$ as $n \to \infty$.

4.6.5 Estimation of $\sigma^2/\beta^T C\beta$

Using appendix 2 again, we obtain

$$E[a\hat{\sigma}^2/\hat{\beta}^T C\,\hat{\beta} + b] = \frac{a}{p-2}\;{}_1F_1(1;\tfrac{1}{2}p;-\lambda) + b \qquad \text{and}$$

$$\operatorname{var}(a\hat{\sigma}^2/\hat{\beta}^T C\hat{\beta} + b) = a^2\left\{\frac{n-p+2}{n-p}\,\frac{1}{(p-2)(p-4)}\;{}_1F_1(2;\tfrac{1}{2}p;-\lambda) - \frac{1}{(p-2)^2}\;{}_1F_1(1;\tfrac{1}{2}p;-\lambda)^2\right\}.$$

We therefore obtain the mean square error

$$\mathrm{MSE} = a^2\left\{\frac{n-p+2}{n-p}\,\frac{1}{(p-2)(p-4)}\,F_1 - \frac{1}{(p-2)^2}\,F_2^2\right\} + \left\{\frac{a}{p-2}\,F_2 + b - \frac{1}{2\lambda}\right\}^2$$

where $F_1 = {}_1F_1(2;\tfrac{1}{2}p;-\lambda)$ and $F_2 = {}_1F_1(1;\tfrac{1}{2}p;-\lambda)$. As $\lambda \to \infty$ we may apply the asymptotic expansion for the confluent hypergeometric function and obtain

$$\text{MSE} \sim \frac{a^2}{4\lambda^2}\left\{\frac{n-p+2}{n-p}\ {}_2F_0(3-\tfrac{1}{2}p,2;;1/\lambda) - {}_2F_0(2-\tfrac{1}{2}p,1;;1/\lambda)^2\right\} + \frac{a}{2\lambda}\left\{{}_2F_0(2-\tfrac{1}{2}p,1;;1/\lambda) + b - \frac{1}{2\lambda}\right\}^2.$$

For small λ, therefore,

$$\text{MSE} \sim a^2\left\{\frac{n-p+2}{n-p}\,\frac{1}{(p-2)(p-4)} - \frac{1}{(p-2)^2}\right\} + \left(\frac{a}{p-2}+b\right)^2 + \frac{2a}{p(p-2)} - \left(\frac{a}{p-2}+b\right)/\lambda + \frac{1}{4\lambda^2}.$$

Since the final term does not depend on a or b we cannot prevent the mean square error from being unbounded. The relative mean square error will tend to 1. For large λ we have

$$\text{MSE} \sim \frac{a^2}{2\lambda^2(n-p)} + \left\{b - \frac{1}{2\lambda}(1-a)\right\}^2$$

$$= \left\{\frac{2a^2}{n-p} + (1-a)^2\right\}\frac{1}{4\lambda^2} - \frac{b(1-a)}{\lambda} + b^2.$$

This is minimised when $b = 0$. If we now choose a to minimise the relative mean square error then we require that $a = \dfrac{n-p}{n-p+2}$.

Now if $C^{-1} \to 0$ as $n \to \infty$ and $b \to 0$ then MSE $\to 0$ and the estimator is mean square error consistent. For relative mean square error consistency we require that $b \to 0$ and $a \to 1$ faster than $\lambda \to \infty$. If interest centres on estimating $A\sigma^2/\beta^TC\beta + B$ with $B \neq 0$ then absolute and relative consistency are equivalent if $C^{-1} \to 0$ as $n \to \infty$.

4.6.6 Estimation of $\frac{1}{\sigma}\beta$

We first calculate the mean and variance of $\frac{1}{\hat\sigma}\hat\beta$. We have

$$E\left[\frac{1}{\hat\sigma}\hat\beta\right] = \sqrt{\tfrac{1}{2}(n-p)}\,\frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p))}\,\frac{1}{\sigma}\beta \qquad \text{and}$$

$$E\left[\frac{1}{\hat\sigma^2}\hat\beta\hat\beta^T\right] = \frac{n-p}{n-p-2}\left\{\frac{1}{\sigma^2}\beta\beta^T + C^{-1}\right\}.$$

Therefore

$$\operatorname{var}\left(\frac{1}{\hat\sigma}\hat\beta\right) = \frac{n-p}{n-p-2}C^{-1} + \left\{\frac{n-p}{n-p-2} - \tfrac{1}{2}(n-p)\left(\frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p))}\right)^2\right\}\frac{1}{\sigma^2}\beta\beta^T.$$

This implies that $\frac{b}{\hat\sigma}\hat\beta$ is unbiased for $\frac{1}{\sigma}\hat\beta$ if $b = \dfrac{1}{\sqrt{\frac{1}{2}(n-p)}}\,\dfrac{\Gamma(\frac{1}{2}(n-p))}{\Gamma(\frac{1}{2}(n-p-1))}$.

The mean square error of $\frac{b}{\hat\sigma}$ is

$$\text{MSE} = b^2\left\{\frac{n-p}{n-p-2}\left(C^{-1} + \frac{1}{\sigma^2}\beta\beta^T\right)\right\} - 2b\sqrt{\tfrac{1}{2}(n-p)}\,\frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p))}\,\frac{1}{\sigma^2}\beta\beta^T + \frac{1}{\sigma^2}\beta\beta^T.$$

For small λ we require $b = 0$ for a minimum while for large λ we require $b = \frac{n-p-2}{n-p}\sqrt{\tfrac{1}{2}(n-p)}\,\frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p))}$.

If $C^{-1} \to 0$ as $n \to \infty$ then, for mean square error consistency, we require that $\frac{n-p}{n-p-2}b^2 - 2b\sqrt{\tfrac{1}{2}(n-p)}\,\frac{\Gamma(\frac{1}{2}(n-p-1))}{\Gamma(\frac{1}{2}(n-p))} + 1 \to 0$ as $n \to \infty$.

This will be so if $b \to 1$ as $n \to \infty$.

4.6.7 Estimation of $\frac{1}{\sigma^2}\beta$

Performing similar calculations to those in the last section we find

$$E[\frac{1}{\hat\sigma^2}\hat\beta] = \frac{n-p}{n-p-2}\frac{1}{\sigma^2}\beta\,, \quad E[\frac{1}{(\hat\sigma^2)^2}\hat\beta\hat\beta^T] = \frac{(n-p)^2}{(n-p-2)(n-p-4)}\left\{\frac{1}{\sigma^4}\beta\beta^T + C^{-1}\right\}$$

and the mean square error for $\frac{b}{\hat\sigma^2}\hat\beta$ is

$$MSE = b^2\left\{\frac{(n-p)^2}{(n-p-2)(n-p-4)}\left[C^{-1} + \frac{1}{\sigma^4}\beta\beta^T\right]\right\} - 2b\,\frac{n-p}{n-p-2}\frac{1}{\sigma^4}\beta\beta^T + \frac{1}{\sigma^4}\beta\beta^T.$$

For an unbiased estimator we must have $b = \frac{n-p-2}{n-p}$. For minimum mean square error we require $b = 0$ if λ is small and $b = \frac{n-p-4}{n-p}$ if λ is large.

Under the same conditions as in the last section we obtain a mean square error consistent estimator.

4.6.8 Means and Variances of Vectors and Matrices

Components of the shrinkage so far considered are scalars. When estimating $\frac{1}{\sigma^2}\beta$ and $\frac{1}{(\beta^TC\beta)^t}\beta$ for some power t we need the expected value and variance of a vector variable, while for estimation of $\beta\beta^T$, $\frac{1}{\sigma^2}\beta\beta^T$ and $\frac{1}{(\beta^TC\beta)^t}\beta\beta^T$ we need the expectation and variance of a variable matrix. In this section we shall compute the means and variances of the naïve estimators for the above expressions.

Noting that $E[C^{\frac{1}{2}}\hat\beta] = C^{\frac{1}{2}}\beta$ and $\operatorname{var}(C^{\frac{1}{2}}\hat\beta) = \sigma^2 I$, we obtain $C^{\frac{1}{2}}\hat\beta \sim N(C^{\frac{1}{2}}\beta, \sigma^2 I)$. This suggests the transformation

$$Y = \frac{1}{\sigma^2}C^{\frac{1}{2}}\hat\beta\,, \quad \eta = E[Y] = \frac{1}{\sigma^2}\beta \quad \text{so that} \quad Y \sim N(\eta\,, I)\,.$$

Now let $R^2 = Y^TY = \frac{1}{\sigma^2}\beta^TC\beta$. By an orthogonal transformation we may transform to the variable $Z = HY$ for which $E[Z_1] = (\eta^T\eta)^{\frac{1}{2}}$ and $E[Z_i] = 0$ if $i \neq 1$. This implies that $R = Z^TZ$.

The above transformation matrix is of the form H where

$$H^T = \left[\frac{1}{\sqrt{2\lambda}}\eta^T \quad H_1\right] \quad \text{with} \quad \lambda = \eta^T\eta\,, \; H_1\eta = 0\,, \; H_1H_1^T = I \quad \text{and}$$

$H_1^T H_1 = I - \frac{1}{2\lambda}\eta\eta^T$. The matrix H_1 is arbitrary apart from the given constraints and this arbitrariness can be characterised as follows: H_2 is another matrix satisfying the constraints if and only if there is an orthogonal matrix B such that $H_2 = B H_1$. (In fact if H_1 and H_2 are given then $B = H_2 H_1^T$).

When we calculate the variance of a matrix we could follow some authors in using the "stacking operator" to turn the matrix into a vector but it is more convenient to use tensor notation which is equivalent to working with the elements of the matrix. Thus we require the variances of the matrix elements and the covariances between them. For consistency we shall treat vectors in the same manner. We shall also use the summation convention so that if an expression contains a repeated subscript then this implies a summation over that subsript. Since we shall need many subscripts, some letters will have to take a dual rôle. This will cause no confusion if, when a letter is used as a subscript, it is not interpreted as taking its other meaning. For example, the letter p stands for the dimension of the parameter space but does not have that meaning when used as a subscript.

Now $E\left[\frac{Z_i}{(Z^TZ)^t}\right] = 0$ if $i \neq 1$,

$E\left[\frac{Z_iZ_j}{(Z^TZ)^{2t}}\right] = 0$ if $i \neq j$ (since $i \neq j \Rightarrow i \neq 1$ or $j \neq 1$)

and $E\left[\frac{Z_iZ_jZ_kZ_l}{(Z^TZ)^{4t}}\right] = 0$ if i, j, k and l are not all equal and not equal in pairs (since otherwise Z_i $i \neq 1$ occurs to an odd power in the numerator).

The expressions still to be evaluated are: $E[R^{-2t}Z_1]$, $E[R^{-4t}Z_1^2]$, $E[R^{-8t}Z_1^4]$, $E[R^{-4t}Z_i^2]$ $(i \neq 1)$, $E[R^{-8t}Z_1^2Z_i^2]$ $(i \neq 1)$, $E[R^{-8t}Z_i^4]$ $(i \neq 1)$ and $E[R^{-8t}Z_i^2Z_j^2]$ $(i \neq 1, j \neq 1, i \neq j)$.

Since $E[R^{-8t}(Z_i^2 + Z_j^2)^2] = E[R^{-8t}Z_i^4] + 2E[R^{-8t}Z_i^2Z_j^2] + E[R^{-8t}Z_j^4]$,

all of the above can be calculated as special cases of the following two theorems.

Theorem 2 If $U \sim \chi^2_s$ and $W \sim \chi^2_r(\lambda)$ independently then

$$E\left[\frac{U^aW^b}{(U+W)^c}\right] = \frac{2^{a+b-c}(\frac{1}{2}s)_a(\frac{1}{2}r)_b}{(a+b-c+\frac{1}{2}r+\frac{1}{2}s)_c} e^{-\lambda} {}_2F_2(b+\tfrac{1}{2}r, a+b-c+\tfrac{1}{2}r+\tfrac{1}{2}s; \tfrac{1}{2}r, a+b+\tfrac{1}{2}s+\tfrac{1}{2}r; \lambda).$$

Theorem 3 If $W \sim N(\mu, I)$ and $U \sim \chi^2_s$ idependently of W and if $2b$ is an iteger then

$$E\left[\frac{U^a W^{2b}}{(U+W^2)^c}\right] = \frac{2^{a+b-c}\,(\frac{1}{2}s)_a\,(\frac{1}{2})_b}{(a+b-c+\frac{1}{2}s+\frac{1}{2})_c}\,e^{-\lambda}\,{}_2F_2(b+\tfrac{1}{2},a+b-c+\tfrac{1}{2}+\tfrac{1}{2}s;\tfrac{1}{2},a+b+\tfrac{1}{2}s+\tfrac{1}{2};\lambda)$$

if $2b$ is even

$$= \sqrt{\lambda}\,\frac{2^{a+b-c}\,(\frac{1}{2}s)_a\,\Gamma(b+1)}{(a+b-c+\frac{1}{2}s+1)_c\,\Gamma(1\frac{1}{2})}\,e^{-\lambda}\,{}_2F_2(b+1,a+b-c+1+\tfrac{1}{2}s;1\tfrac{1}{2},\ a+b+\tfrac{1}{2}s+1;\lambda)$$

if $2b$ is odd.

Note (i) The case when $2b$ is even in theorem 3 is just the case when $r = 1$ of theorem 2.

(ii) If b is an integer then the expression in theorem 2, and if $2b$ is an integer then the expressions in theorem 3, can be written as finite sums of confluent hypergeometric functions.

Proof of theorem 2 Writing $V = W/U$ we have

$$E\left[\frac{U^a W^b}{(U+W)^c}\right] = E\left[\frac{U^{a+b}\,V^b}{\{U(1+V)\}^c}\right] = E\left[\frac{U^{a+b-c}\,V^b}{(1+V)^c}\right].$$

Now $$p(u,w) = \frac{u^{\frac{1}{2}s-1}e^{-\frac{1}{2}u}}{2^{\frac{1}{2}s}\,\Gamma(\frac{1}{2}s)}\,e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\,\frac{w^{\frac{1}{2}r+k-1}\,e^{-\frac{1}{2}w}}{2^{\frac{1}{2}r+k}\Gamma(\frac{1}{2}r+k)}.$$

Also $$du\,dw = \left|\left|\begin{matrix}\frac{\partial u}{\partial u} & \frac{\partial u}{\partial v}\\ \frac{\partial w}{\partial u} & \frac{\partial w}{\partial v}\end{matrix}\right|\right| du\,dv = \left|\left|\begin{matrix}1 & 0\\ v & u\end{matrix}\right|\right| = u\,du\,dv$$

so that

$$p(u,v) = \frac{u^{\frac{1}{2}s-1}\,e^{-\frac{1}{2}u}}{2^{\frac{1}{2}s}\,\Gamma(\frac{1}{2}s)}\,e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\,\frac{u^{\frac{1}{2}r+k}\,v^{\frac{1}{2}r+k-1}\,e^{-\frac{1}{2}uv}}{2^{\frac{1}{2}r+k}\,\Gamma(\frac{1}{2}r+k)}$$

$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\,\frac{u^{\frac{1}{2}s+\frac{1}{2}r+k-1}\,v^{\frac{1}{2}r+k-1}\,e^{-\frac{1}{2}u(1+v)}}{2^{\frac{1}{2}s+\frac{1}{2}r+k}\,\Gamma(\frac{1}{2}s)\,\Gamma(\frac{1}{2}r+k)}.$$

We then have

$$E\left[\frac{U^a\,W^b}{(U+W)^c}\right] = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty}\frac{v^{b+\frac{1}{2}r+k-1}}{(1+v)^c}\int_0^{\infty}\frac{u^{a+b-c+\frac{1}{2}r+k-1}\,e^{-\frac{1}{2}u(1+v)}}{2^{\frac{1}{2}s+\frac{1}{2}r+k}\,\Gamma(\frac{1}{2}s)\,\Gamma(\frac{1}{2}r+k)}\,du\,dv$$

$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty}\frac{2^{a+b-c}\,\Gamma(a+b-c+\frac{1}{2}s+\frac{1}{2}r+k)}{\Gamma(\frac{1}{2}s)\,\Gamma(\frac{1}{2}r+k)}\,\frac{v^{b+\frac{1}{2}r+k-1}}{(1+v)^{a+b+\frac{1}{2}s+\frac{1}{2}r+k}}\,dv$$

$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\,\frac{2^{a+b-c}\,\Gamma(a+b-c+\frac{1}{2}s+\frac{1}{2}r+k)}{\Gamma(\frac{1}{2}s)\,\Gamma(\frac{1}{2}r+k)}\,\frac{\Gamma(a+\frac{1}{2}s)\,\Gamma(b+\frac{1}{2}r+k)}{\Gamma(a+b+\frac{1}{2}s+\frac{1}{2}r+k)}$$

$$= \frac{2^{a+b-c}\,(\frac{1}{2}s)_a\,(\frac{1}{2}r)_b}{(a+b-c+\frac{1}{2}r+\frac{1}{2}s)_c}\,{}_2F_2(b+\tfrac{1}{2}r,a+b-c+\tfrac{1}{2}r+\tfrac{1}{2}s;\tfrac{1}{2}r,a+b+\tfrac{1}{2}s+\tfrac{1}{2}r;\lambda)\times e^{-\lambda}.$$

Proof of theorem 3 Writing $\lambda = \frac{1}{2}\mu^2$, the probability density of W is

$$p(w) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(w-\mu)^2} = \frac{1}{\sqrt{2\pi}}\,e^{-\lambda}\,e^{-\frac{1}{2}w^2}\,e^{\sqrt{2\lambda}\,w}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\sqrt{2\lambda})^k w^k}{k!} e^{-\frac{1}{2}w^2}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\lambda} \left\{ \sum_{k=0}^{\infty} \frac{(2\lambda)^k w^{2k}}{(2k)!} e^{-\frac{1}{2}w^2} + \sqrt{2\lambda} \sum_{k=0}^{\infty} \frac{(2\lambda)^k w^{2k+1}}{(2k+1)!} e^{-\frac{1}{2}w^2} \right\}$$

$$= e^{-\lambda} \sum_{i=0}^{1} (\lambda)^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{2^{k+\frac{1}{2}i} w^{2k+i} \lambda^k}{\sqrt{2\pi}\,(2k+i)!} e^{-\frac{1}{2}w^2} .$$

By the duplication formula for the gamma function,

$$\sqrt{2\pi}\,\Gamma(2z) = 2^{2z-\frac{1}{2}} \Gamma(z)\,\Gamma(z+\tfrac{1}{2})$$

we obtain

$$p(w) = e^{-\lambda} \sum_{i=0}^{1} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{w^{2k+i}}{2^{k+\frac{1}{2}i+\frac{1}{2}}\,\Gamma(k+i+\frac{1}{2})} e^{-\frac{1}{2}w^2} .$$

(This form of the normal distribution density function is convenient for calculating non-central moments and for deriving the density of the non-central χ^2 distribution).

Putting $V = W/U^{\frac{1}{2}}$ gives

$$du\,dw = \left\| \begin{matrix} \frac{\partial u}{\partial u} & \frac{\partial u}{\partial v} \\ \frac{\partial w}{\partial u} & \frac{\partial w}{\partial v} \end{matrix} \right\| du\,dv = \left\| \begin{matrix} 1 & 0 \\ \frac{v}{2u^{\frac{1}{2}}} & u^{\frac{1}{2}} \end{matrix} \right\| du\,dv = u^{\frac{1}{2}}\,du\,dv .$$

$$\text{Now } p(u,w) = \frac{u^{\frac{1}{2}s-1} e^{-\frac{1}{2}u}}{2^{\frac{1}{2}s}\,\Gamma(\frac{1}{2}s)} e^{-\lambda} \sum_{i=0}^{1} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{w^{2k+i} e^{-\frac{1}{2}w^2}}{2^{k+\frac{1}{2}i+\frac{1}{2}}\,\Gamma(k+i+\frac{1}{2})}$$

so that

$$p(u,v) = e^{-\lambda} \sum_{i=0}^{1} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{u^{\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k-1} v^{2k+i} e^{-\frac{1}{2}u(1+v^2)}}{2^{\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k}\,\Gamma(\frac{1}{2}s)\,\Gamma(i+\frac{1}{2}+k)}$$

and

$$E\left[\frac{U^a W^{2b}}{(U+W^2)^c}\right] = e^{-\lambda} \sum_{i=0}^{1} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \int_{-\infty}^{\infty} \frac{v^{2b+2k+i}}{(1+v^2)^c} \times \int_0^{\infty} \frac{u^{a+b-c+\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k-1} e^{-\frac{1}{2}u(1+v^2)}}{2^{\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k}\,\Gamma(\frac{1}{2}s)\,\Gamma(i+\frac{1}{2}+k)}\,du\,dv$$

$$= e^{-\lambda} \sum_{i=0}^{1} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \int_{-\infty}^{\infty} \frac{v^{2b+i+2k}\, 2^{a+b-c}\,\Gamma(a+b-c+\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k)}{(1+v^2)^{a+b+\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k}\,\Gamma(\frac{1}{2}s)\,\Gamma(i+\frac{1}{2}+k)} \times dv$$

$$= e^{-\lambda} \sum_{i=0}^{1} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{1+(-1)^{2b+i}}{2} B(b+\tfrac{1}{2}i+\tfrac{1}{2}+k, a+\tfrac{1}{2}s) \times \frac{2^{a+b-c}\,\Gamma(a+b-c+\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k)}{\Gamma(\frac{1}{2}s)\,\Gamma(i+\frac{1}{2}+k)}$$

$$= e^{-\lambda} \sum_{i=0}^{1} \frac{1+(-1)^{2b+i}}{2} \lambda^{\frac{1}{2}i} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{\Gamma(a+\frac{1}{2}s)\,\Gamma(b+\frac{1}{2}i+\frac{1}{2}+k)}{\Gamma(a+b+\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k)} G$$

where $G = \dfrac{2^{a+b-c}\,\Gamma(a+b-c+\frac{1}{2}s+\frac{1}{2}i+\frac{1}{2}+k)}{\Gamma(\frac{1}{2}s)\,\Gamma(i+\frac{1}{2}+k)}$. When $2b$ is even the term in which $i=0$ vanishes, while, when $2b$ is odd the term in which $i = 1$ vanishes. The remaining term in each case is easily seen to be the required expression and the result is proved.

Now $E\left[\dfrac{Y}{(Y^TY)^t}\right] = E\left[\dfrac{H^TZ}{(Z^TZ)^t}\right] = \dfrac{1}{\sqrt{2\lambda}}\,\eta\, E\left[\dfrac{Z_1}{(Z^TZ)^t}\right]$. If in theorem 3 we put $a = 0$, $b = \frac{1}{2}$, $c = t$ and $s = p-1$ then we obtain

$$E\left[\frac{Y}{(Y^TY)^t}\right] = \frac{e^{-\lambda}}{2^t\,(\frac{1}{2}p+\frac{1}{2}-t)_t}\;{}_1F_1(\tfrac{1}{2}p+1-t;\tfrac{1}{2}p+1;\lambda)\,\eta$$

$$= \frac{1}{2^t\,(\frac{1}{2}p+\frac{1}{2}-t)_t}\;{}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\,\eta\,.$$

Also

$$E\left[\frac{YY^T}{(Y^TY)^{2t}}\right] = E\left[\frac{H^TZZ^TH}{(Z^TZ)^{2t}}\right]$$

$$= \left[\frac{1}{\sqrt{2\lambda}}\,\eta \quad H_1^T\right]\begin{bmatrix} E\left[\dfrac{Z_1^2}{(Z^TZ)^{2t}}\right] & 0 \\ 0 & E\left[\dfrac{Z_i^2}{(Z^TZ)^{2t}}\right] I \end{bmatrix}\begin{bmatrix} \dfrac{1}{\sqrt{2\lambda}}\,\eta^T \\ H_1 \end{bmatrix} \qquad i \neq 1$$

$$= \frac{1}{2\lambda}\eta\eta^T\,E\left[\frac{Z_1^2}{(Z^TZ)^{2t}}\right] + (I - \frac{1}{2\lambda}\eta\eta^T)\,E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right]$$

$$= E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right] I + \frac{1}{2\lambda}\left\{E\left[\frac{Z_1^2}{(Z^TZ)^{2t}}\right] - E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right]\right\}\eta\eta^T.$$

Therefore

$$\operatorname{var}\left(\frac{Y}{(Y^TY)^t}\right) = (I - \frac{1}{2\lambda}\eta\eta^T)\,E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right] + \frac{1}{2\lambda}\eta\eta^T\left\{E\left[\frac{Z_1^2}{(Z^TZ)^{2t}}\right] - E\left[\frac{Z_1}{(Z^TZ)^t}\right]^2\right\}.$$

In theorem 2 we now put $b = 1$, $a = 0$, $c = 2t$, $r = 1$ and $s = p-1$ and obtain

$$E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right] = \frac{2^{1-2t}\,(\frac{1}{2})_1}{(1-2t+\frac{1}{2}p)_t}\,e^{-\lambda}\;{}_2F_2(1\tfrac{1}{2},1-2t+\tfrac{1}{2}p;\tfrac{1}{2},1+\tfrac{1}{2}p;\lambda)$$

and putting $a = 1$, $b = 0$, $c = 2t$, $s = 1$ and $r = p-1$ we obtain

$$E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right] = \frac{2^{1-2t}\,(\frac{1}{2})_2}{(1-2t+\frac{1}{2}p)_{2t}}\,e^{-\lambda}\;{}_2F_2(-\tfrac{1}{2}+\tfrac{1}{2}p,1-2t+\tfrac{1}{2}p;\tfrac{1}{2}p-\tfrac{1}{2},1+\tfrac{1}{2}p;\lambda).$$

We shall calculate the fourth moments of the $R^{-2t}\,Y_i$. We have

$$E[R^{-8t}Y_iY_jY_kY_l] = E[R^{-8t}h_{pi}Z_ph_{qj}Z_qh_{rk}Z_rh_{sl}Z_s]$$
$$= h_{pi}h_{qj}h_{rk}h_{sl}\,E[R^{-8t}Z_pZ_qZ_rZ_s].$$

Let $A = E[R^{-8t}Z_1^4]$, $B = E[R^{-8t}Z_i^4]$, $C = E[R^{-8t}Z_1^2Z_i^2]$ and $D = E[R^{-8t}Z_i^2Z_j^2]$ where $1 \neq i \neq j \neq 1$. We then have

$$E[R^{-8t}Y_iY_jY_kY_l] = Ah_{1i}h_{1j}h_{1k}h_{1l} + B\sum_{p\neq 1}h_{pi}h_{pj}h_{pk}h_{pl} + C\sum_{p>1}{}'h_{1i}h_{1j}h_{pk}h_{pl}$$
$$+ D\sum_{p>q>1}{}''h_{pi}h_{pj}h_{qk}h_{ql} + D\sum_{q>p>1}{}''h_{pi}h_{pj}h_{qk}h_{ql}$$

where Σ' means summation over all six rearrangements of the subscripts 1 and p and Σ'' means summation over all three rearrangements in which p occurs first. Thus

$$E[R^{-8t}Y_iY_jY_kY_l] = Ah_{1i}h_{1j}h_{1k}h_{1l} + (B-3D)\sum_{p>1}h_{pi}h_{pj}h_{pk}h_{pl}$$
$$+ C\sum_{p>1}{}'h_{1i}h_{1j}h_{pk}h_{pl} + D\sum_{p>1,q>1}{}''h_{pi}h_{pj}h_{qk}h_{ql}.$$

Since this expectation cannot depend on the transformation used to calculate it we must have $B = 3D$ (the last two terms are invariant to the transformation used). As a check we shall give a direct proof that $B = 3D$. We may write

$B = E[R^{-8t}Z_i^4] = E[T^4R^{-8t}E[(Z_i/T)^4|T^2]]$ where $T^2 = Z_i^2 + Z_j^2$ and

$D = E[R^{-8t}Z_i^2Z_j^2] = E[T^4R^{-8t}E[(Z_i/T)^2(Z_j/T)^2|T^2]]$. Changing to polar coordinates in the Z_i-Z_j plane we have $\cos\theta = Z_i/T$, $\sin\theta = Z_j/T$, $B = E[T^4R^{-8t}E[\cos^4\theta|T^2]]$ and $D = E[T^4R^{-8t}E[\sin^2\theta\cos^2\theta|T^2]]$.

Now, since the distribution of Z_iZ_j is spherically symmetric, θ is uniformly distributed over a circle of radius T. Therefore

$$E[\cos^4\theta|T^2] = \int_0^{2\pi}\cos^4\theta\,d\theta$$
$$= [-\sin\theta\cos^3\theta]_0^{2\pi} + 3\int_0^{2\pi}\sin^2\theta\cos^2\theta\,d\theta$$
$$= 3\int_0^{2\pi}\sin^2\theta\cos^2\theta\,d\theta$$
$$= 3\,E[\sin^2\theta\cos^2\theta|T^2].$$

Thus $B = 3D$ as required.

Now $h_{1i} = \frac{1}{\sqrt{2\lambda}}\eta_i$ and $\sum_{p>1}h_{pi}h_{pj} = \delta_{ij} - \frac{1}{2\lambda}\eta_i\eta_j$ so that

$$E[R^{-8t}\, Y_iY_jY_kY_l] = \frac{A}{4\lambda^2}\eta_i\eta_j\eta_k\eta_l + \frac{C}{2\lambda}\eta_i\eta_j(\delta_{kl} - \frac{1}{2\lambda}\eta_k\eta_l)$$
$$+ \text{rearrangements of } i, j, k \text{ and } l$$
$$+ D(\delta_{ij} - \frac{1}{2\lambda}\eta_i\eta_j)(\delta_{kl} - \frac{1}{2\lambda}\eta_k\eta_l)$$
$$+ \text{rearrangements of } i, j, k \text{ and } l$$
$$+ \frac{A-6C+3D}{4\lambda^2}\eta_i\eta_j\eta_k\eta_l + D(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$
$$+ \frac{C-D}{2\lambda}(\eta_i\eta_j\delta_{kl}+\eta_i\eta_k\delta_{jl}+\eta_i\eta_l\delta_{jk}+\eta_j\eta_k\delta_{il}+\eta_j\eta_l\delta_{ik}$$
$$+\eta_k\eta_l\delta_{ij}).$$

When $t = 0$ we have $A = 4\lambda^2 + 12\lambda + 3$, $B = 3$, $C = 1 + 2\lambda$ and $D = 1$ from elementary properties of the χ^2_1 and non-central χ^2_1 distributions. This gives, for this value of t,

$$E[Y_iY_jY_kY_l] = \eta_i\eta_j\eta_k\eta_l + (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$
$$+ (\eta_i\eta_j\delta_{kl} + \eta_i\eta_k\delta_{jl} + \eta_i\eta_l\delta_{jk} + \eta_j\eta_k\delta_{il} + \eta_j\eta_l\delta_{ik}$$
$$+ \eta_k\eta_l\delta_{ij})$$

and

$$\operatorname{cov}(Y_iY_jY_kY_l) = (\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}) + (\eta_i\eta_k\delta_{jl}+\eta_i\eta_l\delta_{jk}+\eta_j\eta_k\delta_{il}+\eta_j\eta_l\delta_{ik}).$$

Using theorem 2 with $a = 0$, $s = p-1$, $b = 2$, $r = 1$ and $c = 4t$ we obtain

$$A = \frac{2^{2-4t}\,(\tfrac{1}{2})_2}{(2-4t+\tfrac{1}{2}p)_{4t}}\, e^{-\lambda}\, {}_2F_2(2\tfrac{1}{2},2-4t+\tfrac{1}{2}p;\tfrac{1}{2},2+\tfrac{1}{2}p;\lambda)$$

while putting $a = 2$, $s = 1$, $b = 0$, $r = p-1$ and $c = 4t$ gives

$$B = 3D = \frac{2^{2-4t}\,(\tfrac{1}{2})_2}{(2-4t+\tfrac{1}{2}p)_{4t}}\, e^{-\lambda}\, {}_2F_2(\tfrac{1}{2}p-\tfrac{1}{2},2-4t+\tfrac{1}{2}p;\tfrac{1}{2}p-\tfrac{1}{2},2+\tfrac{1}{2}p;\lambda).$$

Now $(p-1)C = \sum_{i>1} E[R^{-8t}\, Z_1^2Z_i^2] = E[R^{-8t}\, Z_1^2 \sum_{i>1} Z_i^2]$

so putting $a = 1$, $s = p-1$, $b = 1$, $r = 1$ and $c = 4t$ we obtain

$$(p-1)C = \frac{2^{2-4t}\,(\tfrac{1}{2}p-\tfrac{1}{2})(\tfrac{1}{2})}{(2-4t+\tfrac{1}{2}p)_{4t}}\, e^{-\lambda}\, {}_2F_2(1\tfrac{1}{2},2-4t+\tfrac{1}{2}p;\tfrac{1}{2},2+\tfrac{1}{2}p;\lambda).$$

As a check note that $E[R^{-4t}\, Z_1^2] + (p-1)\, E[R^{-4t}\, Z_i^2] = E[R^{2-4t}]$

and $A + (p-1)B + 2(p-1)C + (p-1)(p-2)D = E[R^{4-8t}]$.

Putting $a = b = 0$, $r + s = p$ and $c = 1 - 2t$ in theorem 2 we see

$$E[R^{2-4t}] = \frac{2^{1-2t}}{(1-2t+\tfrac{1}{2}p)_{2t-1}}\, e^{-\lambda}\, {}_2F_2(\tfrac{1}{2}r,1-2t+\tfrac{1}{2}p;\tfrac{1}{2}r,\tfrac{1}{2}p;\lambda)$$

while putting $a = b = 0$, $r + s = p$ and $c = 1-2t$ in theorem 2 yields

$$E[R^{4-8t}] = \frac{2^{2-4t}}{(2-4t+\frac{1}{2}p)_{4t-2}} e^{-\lambda} {}_2F_2(\tfrac{1}{2}r, 2-4t+\tfrac{1}{2}p; \tfrac{1}{2}r, \tfrac{1}{2}p; \lambda).$$

From theorem A1.4.1 and equation A1.4.31 we may write

$$\begin{aligned} {}_2F_2(a+u,b;a,c;z) &= \sum_{i=0}^{\infty} \frac{[u]_i\ (b)_i\ z^i}{(a)_i\ (c)_i\ i!}\ {}_1F_1(b+i;c+i;z) \\ &= e^z \sum_{i=0}^{\infty} \frac{[u]_i\ (b)_i\ z^i}{(a)_i\ (c)_i\ i!}\ {}_1F_1(c-b;c+i;-z). \end{aligned}$$

Using these to simplify our expressions we obtain

$$\begin{aligned} E\left[\frac{Z_1^2}{(Z^TZ)^{2t}}\right] &= \frac{2^{-2t}}{(1-2t+\frac{1}{2}p)_{2t}} \left\{ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) + \frac{1-2t+\frac{1}{2}p}{\frac{1}{2}(1+\frac{1}{2}p)}\ \lambda\ {}_1F_1(2t;\tfrac{1}{2}p+2;-\lambda) \right\} \\ &= \frac{2^{-2t}}{(1-2t+\frac{1}{2}p)_{2t}} \left\{ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) - (p-2\lambda)\ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) \right. \\ &\qquad \left. + p\ {}_1F_1(2t;\tfrac{1}{2}p;-\lambda) \right\}, \end{aligned}$$

$$E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right] = \frac{2^{-2t}}{(1-2t+\frac{1}{2}p)_{2t}} \left\{ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) \right\} \quad \text{for} \quad i \neq 1 \quad \text{and}$$

$$\begin{aligned} E[(Z^TZ)^{1-2t}] &= \frac{2^{1-2t}}{(1-2t+\frac{1}{2}p)_{2t-1}} \left\{ {}_1F_1(2t-1;\tfrac{1}{2}p;-\lambda) \right\} \\ &= \frac{2^{1-2t}\ (\frac{1}{2}p)}{(1-2t+\frac{1}{2}p)_{2t}} \left\{ {}_1F_1(2t;\tfrac{1}{2}p;-\lambda) + \frac{\lambda}{\frac{1}{2}p}\ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) \right\} \\ &= \frac{2^{-2t}}{(1-2t+\frac{1}{2}p)_{2t}} \left\{ p\ {}_1F_1(2t;\tfrac{1}{2}p;-\lambda) + 2\lambda\ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) \right\}. \end{aligned}$$

From these we derive

$$\begin{aligned} E\left[\frac{Z_1^2}{(Z^TZ)^{2t}}\right] + (p-1)\ E\left[\frac{Z_i^2}{(Z^TZ)^{2t}}\right] &= \frac{2}{(1-2t+\frac{1}{2}p)_{2t}} \left\{ 2\lambda\ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda) \right. \\ &\qquad \left. + p\ {}_1F_1(2t;\tfrac{1}{2}p;-\lambda) \right\} \end{aligned}$$

which verifies the first check.

Rewriting the expressions for A, B, C, D and $E[(Z^TZ)^{2-4t}]$ gives

$$\begin{aligned} A = \frac{3\times 2^{-4t}}{(\frac{1}{2}p+2-4t)_{4t}} &\left\{ {}_1F_1(4t;\tfrac{1}{2}p+2;-\lambda) + \frac{2(\frac{1}{2}p+2-4t)\lambda}{\frac{1}{2}(\frac{1}{2}p+2)}\ {}_1F_1(4t;\tfrac{1}{2}p+3;-\lambda) \right. \\ &\left. + \frac{(\frac{1}{2}p+2-4t)_2\lambda^2}{(\frac{1}{2})_2\ (\frac{1}{2}p+2)_2}\ {}_1F_1(4t;\tfrac{1}{2}p+4;-\lambda) \right\}, \end{aligned}$$

$$B = 3D = \frac{3\times 2^{-4t}}{(\frac{1}{2}p+2-4t)_{4t}} \left\{ {}_1F_1(4t;\tfrac{1}{2}p+2;-\lambda) \right\},$$

$$C = \frac{2^{-4t}}{(\frac{1}{2}p+2-4t)_{4t}} \left\{ {}_1F_1(4t;\tfrac{1}{2}p+2;-\lambda) + \frac{(\frac{1}{2}p+2-4t)\lambda}{\frac{1}{2}(\frac{1}{2}p+2)} \; {}_1F_1(4t;\tfrac{1}{2}p+3;-\lambda) \right\}$$

and

$$E[(Z^TZ)^{2-4t}] = \frac{2^{2-4t}}{(\frac{1}{2}p+2-4t)_{4t-2}} \; {}_1F_1(4t-2;\tfrac{1}{2}p;-\lambda).$$

Now $$\frac{(\frac{1}{2}p+2-4t)_i}{(\frac{1}{2}p+2)_i} \frac{1}{(\frac{1}{2}p+2-4t)_{4t}} = \frac{(\frac{1}{2}p-4t)_{i+2}}{(\frac{1}{2}p)_{i+2}} \frac{1}{(\frac{1}{2}p-4t)_{4t}}$$

and $$\frac{1}{(\frac{1}{2}p+2-4t)_{4t-2}} = \frac{(\frac{1}{2}p-4t)_2}{(\frac{1}{2}p-4t)_{4t}} = \frac{(\frac{1}{2}p)_2}{(\frac{1}{2}p-4t)_{4t}} \frac{(\frac{1}{2}p-4t)_2}{(\frac{1}{2}p)_2} .$$

Let $a_i = \dfrac{(\frac{1}{2}p-4t)_i}{(\frac{1}{2}p)_i}$ and let $F_i = {}_1F_1(4t;\tfrac{1}{2}p+i;-\lambda)$.

We may then write

$$A = \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} (3a_2F_2 + 12a_3\lambda F_3 + 4a_4\lambda^2F_4) ,$$

$$B = 3D = \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} 3a_2F_2 ,$$

$$C = \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} (a_2F_2 + 2a_3\lambda F_3)$$

and

$$E[(Z^TZ)^{2-4t}] = \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} 2^2(\tfrac{1}{2}p)_2 a_2 \left\{ {}_1F_1(4t-2;\tfrac{1}{2}p;-\lambda) \right\}$$

$$= \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} 2^2(\tfrac{1}{2}p)_2 a_2 \left\{ {}_1F_1(4t-1;\tfrac{1}{2}p;-\lambda) + \frac{\lambda}{\frac{1}{2}p} \; {}_1F_1(4t-1;\tfrac{1}{2}p+1;-\lambda) \right\}$$

$$= \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} 2^2(\tfrac{1}{2}p)_2 a_2 \left\{ {}_1F_1(4t;\tfrac{1}{2}p;-\lambda) + \frac{\lambda}{\frac{1}{2}p} \; {}_1F_1(4t;\tfrac{1}{2}p+1;-\lambda) \right.$$

$$\left. + \frac{\lambda}{\frac{1}{2}p} \; {}_1F_1(4t;\tfrac{1}{2}p+1;-\lambda) + \frac{\lambda^2}{(\frac{1}{2}p)_2} \times {}_1F_1(4t;\tfrac{1}{2}p+2;-\lambda) \right\}$$

$$= \frac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}} 2^2 a_2 \{(\tfrac{1}{2}p)_2 F_1 + 2(\tfrac{1}{2}p+1)\lambda F_1 + \lambda^2F_2\} .$$

Now $$A + (p-1)B \propto 3p a_2F_2 + 12a_3\lambda F_3 + 4a_4\lambda^2F_4$$

$$= 3p a_2F_2 + 12a_3\lambda F_3 + 4\lambda a_3\{(\tfrac{1}{2}p+2)F_2 - (\tfrac{1}{2}p+2-\lambda)F_3\}$$

$$= \{3p + 4(\tfrac{1}{2}p+2-4t)\lambda\}a_2F_2 - 4(\tfrac{1}{2}p-1-\lambda)a_3F_3$$

and $$2C + (p-2)D \propto p a_2F_2 + 4a_3\lambda F_3$$

with the same constant of proportionality $\dfrac{2^{-4t}}{(\frac{1}{2}p-4t)_{4t}}$.

With this same constant we have

$$
\begin{aligned}
&A + (p-1)B + 2(p-1)C + (p-1)(p-2)D \\
&\quad \propto \{p^2 + 2p + 4(\tfrac{1}{2}p+2-4t)\lambda\}a_2F_2 + 4\{p-1-(\tfrac{1}{2}p-1-\lambda)\}a_3\lambda F_3 \\
&\quad = 4\{(\tfrac{1}{2}p)_2 + (\tfrac{1}{2}p+2-4t)\lambda\}a_2F_2 + 4(\tfrac{1}{2}p+\lambda)\{(\tfrac{1}{2}p+1)a_2F_1 - (\tfrac{1}{2}p+1-\lambda)a_2F_2\} \\
&\quad = 4(\tfrac{1}{2}p+1-4t)\lambda a_2F_2 + 4\lambda^2 a_2F_2 + 4(\tfrac{1}{2}p+\lambda)(\tfrac{1}{2}p+1-4t)a_1F_1 \\
&\quad = 4(\tfrac{1}{2}p+\lambda)(\tfrac{1}{2}p+1-4t)a_1F_1 + 4(\tfrac{1}{2}p+1-4t)\{\tfrac{1}{2}p\, a_1F_0 - (\tfrac{1}{2}p-\lambda)a_1F_1\} + 4\lambda^2 a_2F_2 \\
&\quad = 2^2 a_2\{(\tfrac{1}{2}p)_2F_0 + 2(\tfrac{1}{2}p+1)\lambda F_1 + \lambda^2 F_2\}
\end{aligned}
$$

and this verifies the second check.

$$\text{Now}\quad A - 6C + 3D = \frac{2^{2-4t}}{(\tfrac{1}{2}p-4t)_{4t}}\, a_4\lambda^2 F_4$$

and $$C - D = \frac{2^{1-4t}}{(\tfrac{1}{2}p-4t)_{4t}}\, a_3\lambda F_3\ ;$$

thus

$$
\begin{aligned}
E[R^{-8t}\, Y_iY_jY_kY_l] = \frac{2^{-4t}}{(\tfrac{1}{2}p-4t)_{4t}}\{&a_4F_4\eta_i\eta_j\eta_k\eta_l + a_3F_3(\eta_i\eta_j\delta_{kl}+\eta_i\eta_k\delta_{jl} \\
&+\eta_i\eta_l\delta_{jk}+\eta_k\eta_l\delta_{ij} \\
&+\eta_j\eta_l\delta_{ik}+\eta_j\eta_k\delta_{il}) \\
&+a_2F_2(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\}
\end{aligned}
$$

This agrees with the special case $t = 0$ which was calculated earlier.

We shall now write our results in the original coordinate system. Noting that $C^{\frac{1}{2}}\hat{\beta} = \sigma Y$ and $C^{\frac{1}{2}}\beta = \sigma\eta$ we may write

$$
\begin{aligned}
E\left[\frac{C^{\frac{1}{2}}\hat{\beta}}{(\hat{\beta}^TC\hat{\beta})^t}\right] &= E\left[\frac{Y}{(Y^TY)^t}\right]\sigma^{1-2t} \\
&= \frac{\sigma^{1-2t}}{2^t(\tfrac{1}{2}p+\tfrac{1}{2}-t)_t}\ {}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\,\eta \\
&= \frac{1}{(2\sigma^2)^t(\tfrac{1}{2}p+\tfrac{1}{2}-t)_t}\ {}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\ C^{\frac{1}{2}}\beta\ ,
\end{aligned}
$$

$$
\begin{aligned}
E\left[\frac{C^{\frac{1}{2}}\hat{\beta}\hat{\beta}^TC^{\frac{1}{2}}}{(\hat{\beta}^TC\hat{\beta})^{2t}}\right] &= E\left[\frac{YY^T}{(Y^TY)^{2t}}\right]\sigma^{2-4t} \\
&= \frac{\sigma^{2-4t}}{2^t(\tfrac{1}{2}p+1-2t)_{2t}}\left\{{}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda)\,I + \frac{\tfrac{1}{2}p+1-2t}{\tfrac{1}{2}p+1}\ {}_1F_1(2t;\tfrac{1}{2}p+2;-\lambda)\,\eta\eta^T\right\} \\
&= \frac{1}{(2\sigma^2)^{2t}(\tfrac{1}{2}p+1-2t)_{2t}}\left\{{}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda)\sigma^2 I\right. \\
&\qquad \left. + \frac{\tfrac{1}{2}p+1-2t}{\tfrac{1}{2}p+1}\ {}_1F_1(2t;\tfrac{1}{2}p+2;-\lambda)\,C^{\frac{1}{2}}\hat{\beta}\hat{\beta}^TC^{\frac{1}{2}}\right\}
\end{aligned}
$$

and

$$E\left[\frac{(\mu^TC^{\frac{1}{2}}\hat{\beta}\hat{\beta}^TC^{\frac{1}{2}}\nu)^2}{(\hat{\beta}^TC\hat{\beta})^{4t}}\right] = E\left[\frac{(\mu^TYY^T\nu)^2}{(Y^TY)^{4t}}\right]$$

$$= \frac{\sigma^{4-8t}}{2^{4t}(\frac{1}{2}p-4t)_{4t}}\Big[a_4F_4(\mu^T\eta\nu^T\eta)^2 + a_3F_3\{4\mu^T\eta\nu^T\eta\mu^T\nu + (\mu^T\eta)^2\nu^T\nu + (\nu^T\eta)^2\mu^T\mu\} + a_2F_2\{2(\mu^T\nu)^2 + \mu^T\mu\nu^T\nu\}\Big].$$

We therefore obtain

$$E\left[\frac{\hat\beta}{(\hat\beta^TC\hat\beta)^t}\right] = \frac{1}{(2\sigma^2)^t(\frac{1}{2}p+\frac{1}{2}-t)_t}\ {}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\,\beta\ ,$$

$$E\left[\frac{\hat\beta\hat\beta^T}{(\hat\beta^TC\hat\beta)^{2t}}\right] = \frac{1}{(2\sigma^2)^{2t}(\frac{1}{2}p+1-2t)_{2t}}\left\{{}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda)\sigma^2C^{-1} + \frac{\frac{1}{2}p+1-2t}{\frac{1}{2}p+1}\ {}_1F_1(2t;\tfrac{1}{2}p+2;-\lambda)\,\beta\beta^T\right\}$$

and

$$E\left[\frac{(\mu^T\hat\beta\hat\beta^T\nu)^2}{(\hat\beta^TC\hat\beta)^{4t}}\right] = \frac{1}{(2\sigma^2)^{4t}(\frac{1}{2}p+2-4t)_{4t}}\Big[{}_1F_1(4t;\tfrac{1}{2}p+2;-\lambda)\{2(\mu^TC^{-1}\nu)^2 + \mu^TC^{-1}\mu\nu^TC^{-1}\nu\}\sigma^4 + \frac{\frac{1}{2}p+2-4t}{\frac{1}{2}p+2}\ {}_1F_1(4t;\tfrac{1}{2}p+3;-\lambda) \times\{4\mu^T\beta\nu^T\beta\mu^TC^{-1}\nu + (\mu^T\beta)^2\nu^TC^{-1}\nu + (\nu^T\beta)^2\mu^TC^{-1}\mu\}\sigma^2 + \frac{(\frac{1}{2}p+2-4t)_2}{(\frac{1}{2}p+2)_2}\ {}_1F_1(4t;\tfrac{1}{2}p+4;-\lambda)(\mu^T\beta\nu^T\beta)^2\Big].$$

4.6.9 Estimation of $B\beta\beta^T + A\sigma^2C^{-1}$

We know that $B\hat\beta\hat\beta^T + (A-B)\hat\sigma^2C^{-1}$ is unbiased for $B\beta\beta^T + A\sigma^2C^{-1}$. We shall find the mean square error of the estimator $b\mu^T\hat\beta\hat\beta^T\nu + a\hat\sigma^2\mu^TC^{-1}\nu$ for $B\mu^T\beta\beta^T\nu + a\sigma^2\mu^TC^{-1}\nu$.

From the esults of the previous section we have

$$\begin{aligned}MSE &= b^2\{(\mu^T\beta\nu^T\beta)^2+(\mu^T\beta)^2\nu^TC^{-1}\nu\sigma^2+4\mu^T\beta\nu^T\beta\mu^TC^{-1}\nu\sigma^2+(\nu^T\beta)^2\mu^TC^{-1}\mu\sigma^2 \\ &\quad +\mu^TC^{-1}\mu\nu^TC^{-1}\nu\sigma^4+2(\mu^TC^{-1}\nu)^2\sigma^4-[\mu^T(\beta\beta^T+\sigma^2C^{-1})\nu]^2 \\ &\quad +\{b\mu^T\beta\beta^T\nu+b\sigma^2\mu^TC^{-1}\nu+a\sigma^2\mu^TC^{-1}\nu-B\mu^T\beta\beta^T\nu-A\sigma^2\mu^TC^{-1}\nu\}^2 \\ &\quad + \frac{2a^2}{n-p}\sigma^4(\mu^TC^{-1}\nu)^2 \\ &= b^2\{(\mu^T\beta)^2\nu^TC^{-1}\nu\sigma^2+2\mu^T\beta\nu^T\beta\mu^TC^{-1}\nu\sigma^2+(\nu^T\beta)^2\mu^TC^{-1}\mu\sigma^2 \\ &\quad +\mu^TC^{-1}\mu\nu^TC^{-1}\nu\sigma^4+(\mu^TC^{-1}\nu)^2\sigma^4\} \\ &\quad +\{(b-B)\mu^T\beta\beta^T\nu+(b+a-A)\mu^TC^{-1}\nu\sigma^2\}^2+\frac{2b^2}{n-p}\sigma^4(\mu^TC^{-1}\nu)^2.\end{aligned}$$

For large λ, $(b-B)^2(\mu^T\beta\beta^T\nu)^2$ becomes the dominant term and we must have $b = B$ to achieve a minimum. For small λ the dominant terms are $\{b^2[\mu^TC^{-1}\mu\nu^TC^{-1}\nu + (\mu^TC^{-1}\nu)^2] + (b+a-A)^2(\mu^TC^{-1}\nu)^2\}\sigma^4$ and their sum is minimised for $b = 0$, $a = A$. If we take $b = B$ to

minimise the mean square error for large λ, then the small λ mean square error is minimised by putting $a = \frac{n-p}{n-p+2}(A - B)$.

It is also clear that if $b = B$ then this value of a minimises the mean square error for all values of λ. Thus we suggest the estimator

$$B\hat{\beta}\hat{\beta}^T + \frac{n-p}{n-p+2}(A-B)\hat{\sigma}^2 \quad \text{for} \quad b\beta\beta^T + A\sigma^2 .$$

Among all estimators of the form $B\hat{\beta}\hat{\beta}^T + a\hat{\sigma}^2$ this minimises the mean square error for every bilinear form in the matrix to be estimated when the corresponding bilinear form in the estimator is used to estimate it.

If, as $n \to \infty$, $C^{-1} \to 0$ then an estimator of the form $b\hat{\beta}\hat{\beta}^T + a\hat{\sigma}^2 C^{-1}$ is relatively consistent for $B\beta\beta^T + A\sigma^2 C^{-1}$ if and only if a is bounded and $b \to B$ as $n \to \infty$ since in this case the relative mean square error is asymptotically

$$\frac{(b-B)^2(\mu^T\beta\nu^T\beta)^2}{B^2(\mu^T\beta\nu^T\beta)^2} = \left(\frac{b}{B} - 1\right)^2 .$$

4.6.10 Estimation of $\beta\beta^T/\sigma^2$

The estimator $b\hat{\beta}\hat{\beta}^T/\hat{\sigma}^2 + aC^{-1}$ will be unbiased if $b = \frac{n-p-2}{n-p}$ and $a = -1$ since $E\left[\frac{\hat{\beta}\hat{\beta}^T}{\hat{\sigma}^2}\right] = \frac{1}{\sigma^2}\frac{n-p}{n-p-2}(\beta\beta^T + \sigma^2 C^{-1})$. The mean square error of the general form of the estimator is

$$\begin{aligned} MSE = \frac{b^2}{\sigma^4}\frac{(n-p)^2}{(n-p-2)(n-p-4)} \{ & (\mu^T\beta)^2(\nu^T\beta)^2 + (\mu^T\beta)^2\nu^T C^{-1}\nu\sigma^2 + 4\mu^T\beta\nu^T\beta\mu^T C^{-1}\nu\sigma^2 \\ & + (\nu^T\beta)^2\mu^T C^{-1}\mu\sigma^2 + \mu^T C^{-1}\mu\nu^T C^{-1}\nu\sigma^4 + 2(\mu^T C^{-1}\nu)\sigma^4 \\ & - (\mu^T\beta\beta^T\nu + \sigma^2\mu^T C^{-1}\nu)^2 \} \end{aligned}$$

$$+ \left\{\left(\frac{n-p}{n-p-2}b - 1\right)\frac{\mu^T\beta\beta^T\nu}{\sigma^2} + \left(a + \frac{n-p}{n-p-2}b\right)\mu^T C^{-1}\nu\right\}^2 .$$

If λ is large then the dominant term is $\left(\frac{n-p}{n-p-2}b - 1\right)^2\left(\frac{\mu^T\beta\beta^T\nu}{\sigma^2}\right)^2$ so that $b = \frac{n-p-2}{n-p}$ minimises the mean square error. If λ is small then the mean square error has the dominant terms

$$\frac{(n-p)^2}{(n-p-2)(n-p-4)} b^2\{\mu^T C^{-1}\mu\nu^T C^{-1}\nu + (\mu^T C^{-1}\nu)^2\} + \left(a + \frac{n-p}{n-p-2}b\right)^2(\mu^T C^{-1}\nu)^2$$

whose sum is minimised when $b = 0$ and $a = 0$. On fixing b at the optimal value for large λ we find that the value $a = -1$ is optimal for small λ. Thus we are lead to the unbiased choice for a and b.

As in the last section we have relative consistency if $C^{-1} \to 0$, a is bounded and $b \to 1$ as $n \to \infty$.

4.6.11 Estimation of $\beta\beta^T/\beta^T C\beta$

The expected value of $\hat\beta\hat\beta^T/\hat\beta^T C\hat\beta$ is

$$\frac{1}{2\sigma^2\times\frac12 p}\left\{\frac{p}{p+2}\ {}_1F_1(1;\tfrac12 p+2;-\lambda)\beta\beta^T + {}_1F_1(1;\tfrac12 p+1;-\lambda)\sigma^2 C^{-1}\right\}$$

$$= \frac{{}_1F_1(1;\tfrac12 p+2;-\lambda)}{p+2}\frac{\beta\beta^T}{\sigma^2} + \frac{{}_1F_1(1;\tfrac12 p+1;-\lambda)}{p}C^{-1}$$

$$= \frac{2\lambda\ {}_1F_1(1;\tfrac12 p+2;-\lambda)}{p+2}\frac{\beta\beta^T}{\beta^T C\beta} + \frac{{}_1F_1(1;\tfrac12;\tfrac12 p+1;-\lambda)}{p}C^{-1}$$

$$= \frac{\beta\beta^T}{\beta^T C\beta} - {}_1F_1(1;\tfrac12 p+1;-\lambda)\left\{\frac{\beta\beta^T}{\beta^T C\beta} - \frac{1}{p}\right\}C^{-1}$$

the last line using a recurrence relation for the confluent hypergeometric function.

For small λ the bias is $\frac{1}{p}C^{-1} - \frac{\beta\beta^T}{\beta^T C\beta}$ and no estimator of the form $b\frac{\hat\beta\hat\beta^T}{\hat\beta^T C\hat\beta} + aC^{-1}$ can remove the bias. For large λ

$$E\left[\frac{\hat\beta\hat\beta^T}{\hat\beta^T C\hat\beta}\right] \sim \frac{\beta\beta^T}{\beta^T C\beta} - \frac{p}{2\lambda}\ {}_2F_0(1-\tfrac12 p,1;;1/\lambda)\left\{\frac{\beta\beta^T}{\beta^T C\beta} - \frac{1}{p}C^{-1}\right\}$$

and the bias tends to zero as $\lambda\to\infty$.

The mean square error of $b\frac{\mu^T\hat\beta\hat\beta^T\nu}{\hat\beta^T C\hat\beta} + a\mu^T C^{-1}\nu$ for $\frac{\mu^T\beta\beta^T\nu}{\beta^T C\beta}$ is

$$\begin{aligned}
\mathrm{MSE} = b^2\Bigg\{&\frac{1}{p(p+2)}\ {}_1F_1(2;\tfrac12 p+2;-\lambda)[2(\mu^T C^{-1}\nu)^2 + \mu^T C^{-1}\mu\nu^T C^{-1}\nu] \\
&+ \frac{1}{(p+2)(p+4)}\ {}_1F_1(2;\tfrac12 p+3;-\lambda)\Bigg[4\frac{\mu^T\beta}{\sigma}\frac{\nu^T\beta}{\sigma}\mu^T C^{-1}\nu + \frac{(\mu^T\beta)^2}{\sigma^2}\nu^T C^{-1}\nu \\
&\qquad + \frac{(\nu^T\beta)^2}{\sigma^2}\mu^T C^{-1}\mu\Bigg] \\
&+ \frac{1}{(p+4)(p+6)}\ {}_1F_1(2;\tfrac12 p+4;-\lambda)\left(\frac{\mu^T\beta\beta^T\nu}{\sigma^2}\right)^2\Bigg\} \\
&- b^2\left\{\left(\frac{1}{p+2}\ {}_1F_1(1;\tfrac12 p+2;-\lambda)\frac{\mu^T\beta\beta^T\nu}{\sigma^2} + \frac{1}{p}\ {}_1F_1(1;\tfrac12 p+1;-\lambda)\ \mu^T C^{-1}\nu\right)^2\right\} \\
&+ \left\{\frac{b}{p+2}\ {}_1F_1(1;\tfrac12 p+2;-\lambda)\frac{\mu^T\beta\beta^T\nu}{\sigma^2} + \frac{b}{p}\ {}_1F_1(1;\tfrac12 p+1;-\lambda)\mu^T C^{-1}\nu + a\mu^T C^{-1}\nu\right. \\
&\qquad \left. - \frac{\mu^T\beta\beta^T\nu}{\beta^T C\beta}\right\}^2 .
\end{aligned}$$

For small λ the dominant terms are

$$\frac{b^2}{p(p+2)}\ {}_1F_1(2;\tfrac12 p+2;-\lambda)\left\{2(\mu^T C^{-1}\nu)^2 + \mu^T C^{-1}\mu\nu^T C^{-1}\nu\right\} + \frac{2ab}{p}\ {}_1F_1(1;\tfrac12 p+1;-\lambda) \times(\mu^T C^{-1}\nu)^2$$

and their sum is minimised when $a = b = 0$. For large λ we have the dominant terms

$$\frac{b^2}{(p+4)(p+6)}\,{}_1F_1(2;\tfrac{1}{2}p+4;-\lambda)\left(\frac{\mu^T\beta\beta^T\nu}{\sigma^2}\right)^2+\left(a\mu^TC^{-1}\nu-\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}\right)^2$$

$$+\,2b\left\{\frac{1}{p+2}\,{}_1F_1(1;\tfrac{1}{2}p+2;-\lambda)\frac{\mu^T\beta\beta^T\nu}{\sigma^2}+\frac{1}{p}\,{}_1F_1(1;\tfrac{1}{2}p+1;-\lambda)\mu^TC^{-1}\nu\right\}$$
$$\times\left(a\mu^TC^{-1}\nu-\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}\right)$$

$$\sim b^2\left(\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}\right)^2+2b\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}\left(a\mu^TC^{-1}\nu-\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}\right)+\left(a\mu^TC^{-1}\nu-\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}\right)^2$$

$$=\left\{(b-1)\,\frac{\mu^T\beta\beta^T\nu}{\beta^TC\beta}+a\mu^TC^{-1}\nu\right\}^2.$$

In order to minimise this expression we require that $a = 0$ and $b = 1$. Note that when we have considered the case of large λ we have supposed that it is the length of β causing λ to be large. As $C^{-1} \to 0$, $\mu^T\beta\beta^T\nu/\beta^TC\beta \to 0$ (for fixed β) and the above results cease to be valid. To investigate the consistency of these estimators we need to consider the limit of the mean square error as $C^{-1} \to 0$. In this case the dominant term is $a^2(\mu^TC^{-1}\nu)^2$ and we require $a \to 0$ for consistency and $a \to 0$ faster than $1/\lambda \to 0$ for relative consistency. If the latter condition is satisfied then the next most dominant term must tend to zero and this means that $b \to 1$ as $n \to \infty$ for relative consistency.

4.6.12 Estimation of $(\beta^TC\beta)^{-t}\beta$ $0 \le t \le 1$

Taking the expected value of the naïve estimator we have

$$E\left[\frac{1}{(\hat\beta^TC\hat\beta)^t}\,\hat\beta\right]=\frac{\sigma^{-2t}}{2^t(\frac{1}{2}p+\frac{1}{2}-t)_t}\,{}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\,\beta$$

$$=\frac{\lambda^t}{(\frac{1}{2}p+\frac{1}{2}-t)_t}\,{}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\,\frac{1}{(\beta^TC\beta)^t}\,\beta\,.$$

When $t = 0$ this just gives β showing that $\hat\beta$ is unbiased for β. For no other value of t can we find an estimator of the form $\dfrac{b}{(\hat\beta^TC\hat\beta)^t}$ which minimises the length of the bias vector for all λ. For small λ the bias will be approximately zero if $b = 0$ and $t < 1$ while for large λ

$$E\left[\frac{1}{(\hat\beta^TC\hat\beta)^t}\,\hat\beta\right]\sim\frac{(\frac{1}{2}p+1-t)_t}{(\frac{1}{2}p+\frac{1}{2}-t)_t}\,{}_2F_0(t-\tfrac{1}{2}p,t;;1/\lambda)\,\frac{1}{(\beta^TC\beta)^t}\,\beta$$

which suggests taking $b=\dfrac{(\frac{1}{2}p+\frac{1}{2}-t)_t}{(\frac{1}{2}p+1-t)_t}=\dfrac{\Gamma(\frac{1}{2}p+\frac{1}{2})}{\Gamma(\frac{1}{2}p+1)}\dfrac{\Gamma(\frac{1}{2}p+1-t)}{\Gamma(\frac{1}{2}p+\frac{1}{2}-t)}$.

When $t = \frac{1}{2}$ this gives $b = \frac{(\Gamma(\frac{1}{2}p+\frac{1}{2}))^2}{\Gamma(\frac{1}{2}p)\Gamma(\frac{1}{2}p+1)}$ and when $t = 1$ this gives $b = \frac{\Gamma(\frac{1}{2}p)}{\Gamma(\frac{1}{2}p+1)} \frac{\Gamma(\frac{1}{2}p+\frac{1}{2})}{\Gamma(\frac{1}{2}p-\frac{1}{2})} = \frac{p-1}{p} = 1-\frac{1}{p}$.

We shall now consider the mean square error of $\frac{b}{(\hat{\beta}^T C\hat{\beta})^t}\hat{\beta}$. We have

$$\begin{aligned}
MSE = {} & \frac{b^2}{(2\sigma^2)^{2t}(\frac{1}{2}p+1-2t)_{2t}} \left\{ {}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda)\sigma^2 C^{-1} \right. \\
& \left. + \frac{\frac{1}{2}p+1-2t}{\frac{1}{2}p+1} {}_1F_1(2t;\tfrac{1}{2}p+2;-\lambda)\beta\beta^T \right\} \\
& - \left\{ \frac{b^2}{(2\sigma^2)^t(\frac{1}{2}p+\frac{1}{2}-t)_t} {}_1F_1(t;\tfrac{1}{2}p+1;-\lambda) \right\}^2 \beta\beta^T \\
& + \left\{ \frac{b}{(2\sigma^2)^t(\frac{1}{2}p+\frac{1}{2}-t)_t} {}_1F_1(t;\tfrac{1}{2}p+1;-\lambda) - \frac{1}{(\beta^T C\beta)^t} \right\}^2 \beta\beta^T \\
= {} & \frac{\sigma^2}{(\beta^T C\beta)^{2t}} \left\{ \frac{b^2\lambda^{2t}}{(\frac{1}{2}p+1-2t)_{2t}} \left[{}_1F_1(2t;\tfrac{1}{2}p+1;-\lambda)C^{-1} \right. \right. \\
& \left. + \frac{\frac{1}{2}p+1-2t}{\frac{1}{2}p+1} {}_1F_1(2t;\tfrac{1}{2}p+2;-\lambda)\frac{1}{\sigma^2}\beta\beta^T \right] \\
& \left. + \frac{1}{\sigma^2}\beta\beta^T - \frac{2b\lambda^t}{(\frac{1}{2}p+\frac{1}{2}-t)_t} {}_1F_1(t;\tfrac{1}{2}p+1;-\lambda)\frac{1}{\sigma^2}\beta\beta^T \right\}.
\end{aligned}$$

If λ is small then for $0 < t < \frac{1}{2}$ the first term is dominant and we require $b = 0$; for $t = \frac{1}{2}$ the first and third terms are dominant and we still require $b = 0$; while for $t > \frac{1}{2}$ only the third term is dominant and the mean square error is approximately independent of b.

For large λ we have

$$\frac{(\beta^T C\beta)^{2t}}{\sigma^2} MSE = b^2 C^{-1} + b^2 \frac{1}{\sigma^2}\beta\beta^T + \frac{1}{\sigma^2}\beta\beta^T - 2b \frac{\Gamma(\frac{1}{2}p+1)}{\Gamma(\frac{1}{2}p+\frac{1}{2})} \frac{\Gamma(\frac{1}{2}p+\frac{1}{2}-t)}{\Gamma(\frac{1}{2}p+1-t)} \frac{1}{\sigma^2}\beta\beta^T$$

and the last three terms are dominant. For a minimum we require that $b = \frac{\Gamma(\frac{1}{2}p+1)}{\Gamma(\frac{1}{2}p+\frac{1}{2})} \frac{\Gamma(\frac{1}{2}p+\frac{1}{2}-t)}{\Gamma(\frac{1}{2}p+1-t)}$. When $t = \frac{1}{2}$ this gives $b = \frac{\Gamma(\frac{1}{2}p)\Gamma(\frac{1}{2}p+1)}{\Gamma(\frac{1}{2}p+\frac{1}{2})^2}$ and for $t = 1$ we have $b = \frac{p}{p-1}$.

For relative consistency we again assume that $C^{-1} \to 0$ as $n \to \infty$ and we therefore consider the case when λ is large but β is not. For all b the estimator will be consistent but the relative mean square error will only tend to zero if $b^2 - 2b \frac{\Gamma(\frac{1}{2}p+1)}{\Gamma(\frac{1}{2}p+\frac{1}{2})} \frac{\Gamma(\frac{1}{2}p+\frac{1}{2}-t)}{\Gamma(\frac{1}{2}p+1-t)} + 1 = 0$ and this will be so if $t = 0$.

4.6.13 Summary of Estimators

We now give a summary of possible estimators suggested in the previous sections for various functions of the parameters. These are

given in table 2.

Table 2 Estimators of functions of β and σ

Function	Estimator	Best Unbiased Choice	Minimum Mean Square Error Choice
$\frac{1}{\sigma}\beta$	$\frac{b}{\hat{\sigma}}\hat{\beta}$	$b = \frac{\Gamma(\frac{1}{2}(n-p))}{\Gamma(\frac{1}{2}(n-p)-\frac{1}{2})} \times (\frac{1}{2}(n-p))^{-\frac{1}{2}}$	$b = \frac{\Gamma(\frac{1}{2}(n-p)-\frac{1}{2})}{\Gamma(\frac{1}{2}(n-p))} \times \frac{n-p-2}{n-p} (\frac{1}{2}(n-p))^{\frac{1}{2}}$ *
$\frac{1}{\sigma^2}\beta$	$\frac{b}{\hat{\sigma}^2}\hat{\beta}$	$b = \frac{n-p-2}{n-p}$	$b = \frac{n-p-4}{n-p}$ *
$\frac{1}{(\beta^T C\beta)^{\frac{1}{2}}}\beta$	$\frac{b}{(\hat{\beta}^T C\hat{\beta})^{\frac{1}{2}}}\hat{\beta}$	$b = \frac{(\Gamma(\frac{1}{2}p+\frac{1}{2}))}{\Gamma(\frac{1}{2}p)\,\Gamma(\frac{1}{2}p+1)}$ *	$b = \frac{\Gamma(\frac{1}{2}p)\,\Gamma(\frac{1}{2}p+1)}{(\Gamma(\frac{1}{2}p+\frac{1}{2}))}$ *
$\frac{1}{\beta^T C\beta}\beta$	$\frac{b}{\hat{\beta}^T C\hat{\beta}}\hat{\beta}$	$b = 1 - \frac{1}{p}$ *	$b = \frac{p}{p-1}$ *
$B\beta^T C\beta + A\sigma^2$	$b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2$	$b = B$ $a = A - Bp$	If $b = B$ then $a = \frac{n-p}{n-p+2}(A - Bp)$
$B\beta\beta^T + A\sigma^2 I$	$b\hat{\beta}\hat{\beta}^T C + a\hat{\sigma}^2 I$	$b = B$ $a = A - B$	If $b = B$ then $a = \frac{n-p}{n-p+2}(A - B)$
$\frac{\beta^T C\beta}{\sigma^2}$	$b\frac{\hat{\beta}^T C\hat{\beta}}{\hat{\sigma}^2} + a$	$b = \frac{n-p-2}{n-p}$ $a = -p$	If $b = \frac{n-p-2}{n-p}$ then $a = -p$
$\frac{\beta\beta^T}{\sigma^2}$	$b\frac{\hat{\beta}\hat{\beta}^T}{\hat{\sigma}^2} + aC^{-1}$	$b = \frac{n-p-2}{n-p}$ $a = -1$	$b = \frac{n-p-2}{n-p}$ $a = -1$
$\frac{\sigma^2}{\beta^T C\beta}$	$a\frac{\hat{\sigma}^2}{\hat{\beta}^T C\hat{\beta}} + b$	If $a = 1$ then $b = 0$ *	If $a = 1$ then $b = 0$ *
$\frac{\beta\beta^T}{\beta^T C\beta}$	$b\frac{\hat{\beta}\hat{\beta}^T}{\hat{\beta}^T C\hat{\beta}} + aC^{-1}$	$b = 1$ $a = 0$ *	$b = 1$ $a = 0$ *

* *for large* λ

In the next section we shall apply these results to the estimation of the shrinkage factor.

4.6.14 Estimators for the Shrinkage factor

By substituting the component estimators into the shrinkage factor $\frac{\beta\beta^T C}{\sigma^2 + \beta^T C\beta}$ and different rearrangements of this we obtain an estimator of the form $\frac{b\hat{\beta}\hat{\beta}^T C + a\hat{\sigma}^2 I}{d\hat{\beta}^T C\hat{\beta} + c\hat{\sigma}^2}$ for the shrinkage factor and this gives rise to an estimator for β of the form $\frac{b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2}{d\hat{\beta}^T C\hat{\beta} + c\hat{\sigma}^2}\hat{\beta}$. This is of

of the same form as the estimator obtained by substituting the component estimators into the shrinkage factor $\frac{\beta^T C\beta}{p\sigma^2 + \beta^T C\beta}$.

We may also substitute estimators for $\beta\beta^T$ and σ^2 into the last expression for β^* in section 4.6 or an estimator for $\sigma^{-2}\beta\beta^T$ into the previous expression in the same section. Putting $b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1}$ for $\beta\beta^T$ and $c\hat\sigma^2$ for σ^2 gives

$$\hat\beta^* = (b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1})X^T\{c\hat\sigma^2 V + X(b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1})X^T\}^{-1}Y$$

while putting $\frac{b}{\hat\sigma^2}\hat\beta\hat\beta^T + aC^{-1}$ for $\frac{1}{\sigma^2}\beta\beta^T$ gives

$$\hat\beta^* = (b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1})X^T\{\hat\sigma^2 V + X(b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1})X^T\}^{-1}Y .$$

Since these are of the same form we shall simplify the former expression.

Let $$A = c\hat\sigma^2 V + X(b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1})X^T$$
$$= c\hat\sigma^2 V + a\hat\sigma^2 XC^{-1}X^T + bX\hat\beta\hat\beta^T X^T .$$

Using the formula for the inverse of the sum of two matrices we obtain

$$A^{-1} = B^{-1} - B^{-1}X\hat\beta(b^{-1} + \hat\beta^T X^T B^{-1} X\hat\beta)^{-1}\hat\beta^T X^T B^{-1}$$

where $B = c\hat\sigma^2 V + a\hat\sigma^2 XC^{-1}X^T$.

The same formula gives

$$B^{-1} = \frac{1}{c\hat\sigma^2}V^{-1} - \frac{1}{(c\hat\sigma^2)^2}V^{-1}X\left(\frac{1}{a\hat\sigma^2}C + \frac{1}{c\hat\sigma^2}X^T V^{-1}X\right)^{-1}X^T V^{-1}$$
$$= \frac{1}{c\hat\sigma^2}V^{-1}\left(V - \frac{a}{a+c}XC^{-1}X^T\right)V^{-1} .$$

Therefore

$$X^T B^{-1} = \frac{1}{c\hat\sigma^2}\left(X^T V^{-1} - \frac{a}{a+c}X^T V^{-1}\right)$$
$$= \frac{1}{(a+c)\hat\sigma^2}X^T V^{-1}$$

and $$X^T A^{-1} = \frac{1}{(a+c)\hat\sigma^2}X^T V^{-1} - \left(\frac{1}{(a+c)\hat\sigma^2}\right)^2 C\hat\beta(b^{-1} + \frac{1}{(a+c)\hat\sigma^2}\hat\beta^T C\hat\beta)^{-1}\hat\beta^T X^T V^{-1}$$
$$= \frac{1}{(a+c)\hat\sigma^2}\left\{C - \frac{b}{(a+c)\hat\sigma^2 + b\hat\beta^T C\hat\beta}C\hat\beta\hat\beta^T C\right\}C^{-1}X^T V^{-1} .$$

This gives

$$\hat\beta^* = (b\hat\beta\hat\beta^T + a\hat\sigma^2 C^{-1})\frac{1}{(a+c)\hat\sigma^2}C\left\{1 - \frac{b\hat\beta^T C\hat\beta}{(a+c)\hat\sigma^2 + b\hat\beta^T C\hat\beta}\right\}\hat\beta$$
$$= \frac{b\hat\beta\hat\beta^T C + a\hat\sigma^2 I}{(a+c)\hat\sigma^2 + b\hat\beta^T C\hat\beta}\hat\beta$$
$$= \frac{a\hat\sigma^2 + b\hat\beta^T C\hat\beta}{c\hat\sigma^2 + a\hat\sigma^2 + \hat\beta^T C\hat\beta}\hat\beta .$$

This is a scalar shrinkage estimator of the same form as our other shrinkage estimators. The suggestion raised by this estimator is that the values of a and b suitable for estimating $\beta\beta^T$ by an estimator of the form $b\hat{\beta}\hat{\beta}^T + a\hat{\sigma}^2 I$ should be used and substituted into the estimator, $b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2$, for $\beta^T C\beta$ in the scalar form of the shrinkage factor. A similar interpretation is taken for the function $\frac{1}{\sigma^2}\beta\beta^T$.

It will be seen that this estimator for the shrinkage factor is democratic in the sense that each β occurring in the factor is treated in the same way.

Table 3 contains suggested values of a, b, c and d. We do not, however, recommend all these values. In table 4 the numbers in square brackets give a simplified form of the same shrinkage factor.

4.6.15 Consistency of Estimators for the Shrinkage Factor

We shall suppose as before that $C^{-1} \to 0$ as $n \to \infty$. In this case, as $n \to \infty$, $\lambda \to \infty$ and the ratio $\frac{\beta^T C\beta}{k\sigma^2 + \beta^T C\beta} \to 1$. Only if $b \sim d$ as $n \to \infty$ will $E\left[\frac{b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2}{d\hat{\beta}^T C\hat{\beta} + c\hat{\sigma}^2}\right]$ tend to 1 when $c > 0$. So long as the numerator vanishes for larger values of λ than does the denominator and we replace the ratio by zero when the numerator vanishes, this result applies if $c \leq 0$. The result follows from the fact that the value of $\hat{\lambda}$ exceeds any given bound with arbitrarily small probability as $n \to \infty$. In considering the matrix shrinkage we observe that, for an estimator to be consistent, its trace must be consistent for the trace of the shrinkage matrix. We are thus lead to the case above and we require that $b \sim d$ as $n \to \infty$ (if $c < 0$ we must replace the factor by zero for a negative numerator). We shall not investigate the sufficiency of the condition $a \sim c$ as $n \to \infty$, but it is intuitively clear that the variance will tend to zero if this holds.

4.6.16 Consistency of Estimators for β

We now wish to show that, under the conditions of the previous section, $\frac{b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2}{d\hat{\beta}^T C\hat{\beta} + c\hat{\sigma}^2}\hat{\beta}$ is consistent for β. Now, as $n \to \infty$, $\hat{\lambda} \to \infty$ in probability if $b/d \to 1$. Intuitively, this gives the result. More precisely $E\left[\frac{2b\hat{\lambda} + a}{2d\hat{\lambda} + c}\hat{\beta} - \beta\right] = E\left[\left(\frac{2b\hat{\lambda} + a}{2d\hat{\lambda} + c} - \frac{b}{d}\right)\hat{\beta}\right] + \frac{b}{d}E[\hat{\beta}] - \beta$. The last term tends to zero as $n \to \infty$ if and only if $b/d \to 1$. The integrand of the first term is $-\frac{b}{d}\hat{\beta}$ for small $\hat{\lambda}$ (with small probability) and $\frac{ad - bc}{d}\frac{1}{2d\hat{\lambda} + c}\hat{\beta}$ for large $\hat{\lambda}$. This tends to zero as $\hat{\lambda} \to \infty$ ($\hat{\beta}$ fixed).

Table 3 Suggested Coefficients for the Shrinkage Factor

a	b	c	d
$-p$	1	$1-p$	1
$-p$	1	0	1
$-\frac{n-p}{n-p+2}p$	1	$\frac{n-p}{n-p+2}(1-p)$	1
$-\frac{n-p}{n-p+2}p$	1	0	1
-1	1	$1-p$	1
$-\frac{n-p}{n-p+2}$	1	$\frac{n-p}{n-p+2}(1-p)$	1
$-\frac{n-p-2}{n-p}p$ $[-p]$	$\frac{n-p-2}{n-p}$ $[1]$	$1-\frac{n-p-2}{n-p}p$ $\left[\frac{n-p}{n-p-2}-p\right]$	$\frac{n-p-2}{n-p}$ $[1]$
$-\frac{n-p-2}{n-p}p$ $[-p]$	$\frac{n-p-2}{n-p}$ $[1]$	$\frac{2p}{n-p}$ $\left[\frac{2p}{n-p-2}\right]$	$\frac{n-p-2}{n-p}$ $[1]$
$-\frac{n-p-2}{n-p}$ $[-1]$	$\frac{n-p-2}{n-p}$ $[1]$	$1-\frac{n-p-2}{n-p}p$ $\left[\frac{n-p}{n-p+2}-p\right]$	$\frac{n-p-2}{n-p}$ $[1]$
0	1	p	1
0	1	1	1
0	1	$1-1/p$	1
0 $[0]$	$\frac{n-p-2}{n-p}$ $[1]$	$1-\frac{n-p-2}{n-p}p$ $\left[\frac{n-p}{n-p-2}-p\right]$	$\frac{n-p-2}{n-p}$ $[1]$
-1	1	$-\frac{2}{n-p+2}$	1
$-\frac{n-p-2}{n-p}$	1	$\frac{4}{(n-p)(n-p+2)}$	1
-1	$\frac{n-p-2}{n-p}$	0	$\frac{n-p-2}{n-p}$

Note that we have omitted the estimators in which $b \neq d$ since these are not consistent.

This proves that the bias tends to zero. In a similar way it is seen that the variance tends to zero.

4.7 Alternative Estimators

When estimating the shrinkage factor we have tried to cope with the bias in the naïve estimators for the components. Another approach is to compensate for the bias in the numerator by biasing the denominator, or vice-versa. We shall consider the general scalar shrinkage of the form $\frac{\beta^T C \beta}{k\sigma^2 + \beta^T C \beta}$. Using the bilinear shrinkage rule of the previous section we have $\frac{E[b\hat{\beta}^T C\hat{\beta} + a\hat{\sigma}^2]}{E[d\hat{\beta}^T C\hat{\beta} + c\hat{\sigma}^2]} = \frac{\beta^T C\beta}{k\sigma^2 + \beta^T C\beta}$ if and only if

$$(k\sigma^2 + \beta^T C\beta)(b\beta^T C\beta + (bp+a)\sigma^2) = \beta^T C\beta(d\beta^T C\beta + (dp+c)\sigma^2).$$

This implies $b = d$, $(bp+a)k = 0$ and $bp+a+bk = dp+c$. Thus $a = -bp$, $d = b$ and $d = b(k-p)$. With $b = 1$ this is the unbiased choice for numerator and denominator. If we wish to accept the bias in the numerator then we put $a = 0$. In this case there will be no exact solution. However, allowing a stochastic choice of values we have

$$(k + 2\lambda)(2b\lambda + bp + a) = (2\lambda)^2 d + (dp+c)\times 2\lambda$$

i.e. $$(2\lambda)^2(b-d) + (2\lambda)(bp+a+bk-dp-c) + (bp+a)k = 0$$

and if $b = d$ then this reduces to

$$2\lambda(bp+a-c) = -(bp+a)k .$$

The case $a = 0$ now gives $c = bp + \frac{bp}{2\lambda} = bp\left(\frac{2\lambda+1}{2\lambda}\right)$. This suggests replacing the constant d by the random variable $b\,p\frac{\hat{\beta}^T C\hat{\beta} + \hat{\sigma}^2}{\hat{\beta}^T C\hat{\beta}}$ giving $\frac{\hat{\beta}^T C\hat{\beta}}{(\hat{\beta}^T C\hat{\beta})^2 + (k+p)\hat{\sigma}^4 + p\hat{\beta}^T C\hat{\beta}\hat{\sigma}^2}$ as an estimator for β. As mentioned previously, we are interested in the cases $k = 1$ and $k = p$.

Similarly, accepting the bias in the denominator leads to $b = d = 1$, $c = k$ and $2\lambda(p+c-k) = -(p+a)k$ so that $a = \frac{(k-p)2\lambda - pk}{2\lambda + k}$. This suggests $\frac{(\hat{\beta}^T C\hat{\beta})^2 + (2k-p)\hat{\beta}^T C\hat{\beta}\hat{\sigma}^2 - pk\hat{\sigma}^4}{(\hat{\beta}^T C\hat{\beta} + k\hat{\sigma}^2)}$ as an estimator for β. Again we are interested in the two cases $k = 1$ and $k = p$.

We can easily apply the same approach to other representations of the shrinkage factor; for example $\frac{2b\hat{\lambda} + a}{2d\hat{\lambda} + c}$ gives the equation

$$(k+2\lambda)(2b\lambda\,\frac{n-p}{n-p-2} + bp + a) = 2\lambda(2d\lambda\,\frac{n-p}{n-p-2} + dp + c).$$

4.8 Risk Functions for Bilinear Shrinkage Estimators

The computational formulae for the risk have been delayed until chapter 6 so that the class of estimators of interest may first be determined. A formula for computing the risk of bilinear shrinkage estimators is given in theorem 6.7.7 . This formula, when the hypergeometric functions are computed from recurrence relations, gives the risk as the sum of a single infinite series. However, the hypergeometric functions have to be computed by their series expansions from time to time to avoid numerical instability in the recurrence formulae.

In the interest of generality, we did not use this formula in the final computation but based the calculation on the unbiased estimator for the risk given in equation 6.3.2 . Taking the expectation of the unbiased risk estimator requires the evaluation of a one-dimensional integral. This was done numerically. In order to compare the estimators of this chapter with those of the next, the risk functions of the two families of estimators are plotted on the same graphs at the end of chapter 5 (where a description of the program may be found).

Chapter 5

Iterative Improvement of the Minimum Mean Square Error Estimator

5.1 Introduction

Whenever we estimate the shrinkage factor (scalar or matrix) for the minimum mean square error variate we arrive at a new estimator for β. This new estimator can be used to re-estimate the shrinkage factor. It is of interest to know whether, on repeating the process indefinitely, we obtain a sequence of estimators which converges to a limit. If so, does the limit provide a good estimator for β? Hemmerle(1975) was the first to find a fixed point for the iteration and he gave conditions under which the fixed points were stable. Later Vinod(1976) compared the resulting estimator with other estimators by using a Monte-Carlo simulation. Vinod, in fact, gave two iterations, one using the usual estimator for the variance at each step in the iteration, the other basing each estimator for the variance on the latest estimate for β. In the next section we discuss these iterative processes.

5.2 Fixed Point Estimators

Consider the matrix shrinkage $\beta^* = \dfrac{\beta\beta^T C}{\sigma^2 + \beta^T C\beta}$. The usual estimator for σ^2 is $\hat{\sigma}^2 = \dfrac{1}{n-p}(Y - X\hat{\beta})^T V^{-1}(Y - X\hat{\beta})$. If we base our estimator for σ^2 on an improved estimator β_o^* we may use $\sigma^2{}_o^* = \dfrac{1}{\mu}(Y - X\beta_o^*)^T V^{-1}(Y - X\beta_o^*)$. We have used a different divisor from $\nu = n-p$ to allow for the fact that β_o^* is, hopefully, a better estimator for β than $\hat{\beta}$ (if β were known then we would use n as the divisor). We may combine the two estimation formulae by writing

$$\begin{aligned}\sigma^2{}_o^* &= \frac{1}{\mu}(Y - X\hat{\beta} - X(\beta_o^* - \hat{\beta}))^T V^{-1}(Y - X\hat{\beta} - X(\beta_o^* - \hat{\beta})) \\ &= \frac{1}{\mu}(Y - X\hat{\beta})^T V^{-1}(Y - X\hat{\beta}) + \frac{1}{\mu}(\beta_o^* - \hat{\beta})^T C(\beta_o^* - \hat{\beta}) \\ &= \frac{n-p}{\mu}\hat{\sigma}^2 + \frac{\alpha}{\mu}(\beta_o^* - \hat{\beta})^T C(\beta_o^* - \hat{\beta})\end{aligned}$$

where, $\alpha = 0$ gives the usual estimator and $\alpha = 1$ gives the estimator based on β_o^*, so long as μ is suitably chosen. Replacing $n-p$ by ν allows us to use different divisors for $\hat{\sigma}^2$ (remembering that the divisor $n-p+2$ gives minimum mean square error). Vinod refused to use a different divisor "in deference to the usual practice", However, we are estimating β not σ^2 and we have already abandonned "the usual

practice" in doing so. The iteration now procedes by substituting β_o^* for β and σ^{2*}_o for σ^2. This gives an estimator β_1^* which is substituted for β_o^*. It is clear that each iteration gives a resulting vector in the same direction as β_o^*.

Now consider the scalar shrinkage $\beta^\dagger = \dfrac{\beta^T C\beta}{p\sigma^2 + \beta^T C\beta}\,\beta$. Proceeding in the same way we see that, whatever the direction of $\beta_o^\dagger$, the first vector to be substituted for β, all iterates are in the direction of β.

In either case the value of μ should be revised at each step, but as it is difficult (if at all possible) to choose the best value, we shall choose μ after finding the fixed point as a function of μ (and of ν and α).

The scalar and vector shrinkages will be combined together by considering the shrinkage $\beta^* = \dfrac{\beta^T C\beta}{k\sigma^2 + \beta^T C\beta}$. The case $k = 1$ gives the matrix shrinkage, while the case $k = p$ gives the scalar shrinkage provided that we add the extra condition that the initial vector is in the direction of β.

A fixed point of the iteration can be found by solving the equation

$$(k\sigma^{2*}_\infty + \beta^{*T}_\infty C\beta^*_\infty)\,\beta^*_\infty = \beta^{*T}_\infty C\hat\beta\;\beta^*_\infty.$$

Substituting for σ^{2*}_∞ we obtain

$$\left\{\frac{k\nu}{\mu}\hat\sigma^2 + (1+\frac{k\alpha}{\mu})\,\beta^{*T}_\infty C\beta^*_\infty - (1+\frac{2k\alpha}{\mu})\,\beta^{*T}_\infty C\hat\beta + \frac{k\alpha}{\mu}\hat\beta^T C\hat\beta\right\}\beta^*_\infty = 0.$$

We see that one solution is $\beta^*_\infty = 0$ while the other solutions satisfy the equation

$$\beta^{*T}_\infty C\beta^*_\infty - \frac{\mu+2k\alpha}{\mu+k\alpha}\,\beta^{*T}_\infty C\hat\beta + \frac{k\alpha}{\mu+k\alpha}\,\hat\beta^T C\hat\beta + \frac{k\nu}{\mu+k\alpha}\hat\sigma^2 = 0.$$

Completing the square in this expression gives

$$(1)\quad \left\{\beta^*_\infty - (1-\frac{\frac12\mu}{\mu+k\alpha})\hat\beta\right\}^T C\left\{\beta^*_\infty - (1-\frac{\frac12\mu}{\mu+k\alpha})\hat\beta\right\} = \frac14\left(\frac{\mu}{\mu+k\alpha}\right)^2\hat\beta^T C\hat\beta - \frac{k\nu}{\mu+k\alpha}\hat\sigma^2.$$

Since C is positive definite, this has real roots if and only if $\hat\lambda = \dfrac{1}{2\hat\sigma^2}\,\hat\beta^T C\hat\beta > \dfrac{\mu+k\alpha}{\mu}\,\dfrac{2k\nu}{\mu}$. Thus, for small $\hat\lambda$, the only solution is $\beta^*_\infty = 0$, while there are three solutions for large $\hat\lambda$. The non-zero solutions (when they exist) are points on a hyperellipsoid such that $C^{\frac12}\beta^*_\infty$ lies on a hypersphere with centre $(1-\frac{\frac12\mu}{\mu+k\alpha})\,C^{\frac12}\hat\beta$ and radius $\left\{\frac12\left(\frac{\mu}{\mu+k\alpha}\right)^2 - \frac{k\nu}{\mu+k\alpha}\right\}^{\frac12}\hat\sigma$.

In order that the estimator should be continuous as a function of $\hat{\beta}$ and $\hat{\sigma}^2$, it seems desirable that, in the one real root case we use the real parts of the complex roots for our solution. In other words, if $\hat{\lambda} < \frac{\mu+k\alpha}{\mu}\frac{2k\nu}{\mu}$ then we use $(1-\frac{\frac{1}{2}\mu}{\mu+k\alpha})\hat{\beta}$. On the other hand we might be tempted - especially if $k = 1$ - to use $\beta_\infty^* = 0$ as a kind of preliminary test estimator.

In the next section we shall discuss how we may choose among the solutions found.

5.2.1 Sums of Squares Criteria of Choice

Vinod chose the fixed point which minimises the sum of squared residuals. We shall consider this choice as well as some others. In particular we might choose the solution with minimum length so as to minimise the danger of over-shrinking. Another approach is to maximise the expected length. If $\beta_\infty^* = 0$ is the only solution we might be interested in the closest approach to a solution of (5.1).

We shall let $u = 1 - \frac{\frac{1}{2}\mu}{\mu+k\alpha}$ and $v = \frac{1}{2}\left(\frac{\mu}{\mu+k\alpha}\right)^2 \hat{\lambda} - \frac{k\nu}{\mu+k\alpha}$ so that (5.1) may be rewritten

(1) $(\beta_\infty^* - u\hat{\beta})^T C (\beta_\infty^* - u\hat{\beta}) = v$.

5.2.2 Least Squares Criterion

In order to find the solution of (5.2.1.1) which minimises the sum of squared residuals we find a stationary value of

$$z = (Y - X\beta_\infty^*)^T V^{-1} (Y - X\beta_\infty^*) + \gamma[(\beta_\infty^* - u\hat{\beta})^T C (\beta_\infty^* - u\hat{\beta}) - v]$$

where γ is a Lagrange multiplier. Now

$$\frac{\partial z}{\partial \beta_\infty^*} = -2X^T V^{-1}(Y - X\beta_\infty^*) + 2\gamma C(\beta_\infty^* - u\hat{\beta})$$

$$= 2C\beta_\infty^* - 2C\hat{\beta} + 2\gamma C(\beta_\infty^* - u\hat{\beta})$$

$$= 2C[(1+\gamma)\beta_\infty^* - (1+\gamma u)\hat{\beta}]$$

and

$$\frac{\partial^2 z}{\partial \beta_\infty^{*T} \partial \beta_\infty^*} = 2(1+\gamma)C.$$

Thus $\frac{\partial z}{\partial \beta_\infty^*} = 0 \Leftrightarrow \beta_\infty^* = \frac{1+\gamma u}{1+\gamma}\hat{\beta} = u\hat{\beta} + \frac{1-u}{1+\gamma}\hat{\beta}$

and substituting into the constraint equation gives

$$[(1+\gamma u) - u(1+\gamma)^2]\hat{\beta}^T C \hat{\beta} = (1+\gamma)^2 v$$

$$\Leftrightarrow \quad (1-u)^2 \hat{\beta}^T C \hat{\beta} = (1+\gamma)^2 v.$$

This gives $(1+\gamma)^2 = (1-u)^2 \frac{1}{v} \hat{\beta}^T C \hat{\beta}$

and if we take the positive root for $1+\gamma$ we obtain a minimum for the residual sum of squares. This leads to the solution for β_∞^*

$$\beta_\infty^* = u\hat{\beta} + \sqrt{\frac{v}{\hat{\beta}^T C \hat{\beta}}}\ \hat{\beta}.$$

In view of the fact that

$$(Y - X\beta_\infty^*)^T V^{-1} (Y - X\beta_\infty^*) = \hat{\sigma}^2 + (\beta_\infty^* - \hat{\beta})^T C (\beta_\infty^* - \hat{\beta})$$

this value also minimises the distance from $\hat{\beta}$.

5.2.3 Maximum Length Solution

In order to maximise the length of β_∞^* we find a stationary value of

$$z = \beta_\infty^{*T} C \beta_\infty^* + \gamma[(\beta_\infty^* - u\hat{\beta})^T C (\beta_\infty^* - u\hat{\beta}) - v].$$

Now $\frac{\partial z}{\partial \beta_\infty^*} = 2C\beta_\infty^* + 2\gamma C(\beta_\infty^* - u\hat{\beta})$ and $\frac{\partial^2 z}{\partial \beta_\infty^{*T} \partial \beta_\infty^*} = 2(1+\gamma)$.

For stationary values,

$$\beta_\infty^* = (1+\gamma)^{-1} u\gamma\hat{\beta} = u\frac{\gamma}{1+\gamma}\hat{\beta} = u\hat{\beta} - \frac{u}{1+\gamma}\hat{\beta}.$$

Substituting this into the constraint equation gives $u^2\hat{\beta}^T C \hat{\beta} = (1+\gamma)\, v$. For a minimum we require the negative value of $1+\gamma$ and this gives the estimator

$$\beta_\infty^* = u\hat{\beta} + \sqrt{\frac{v}{\hat{\beta}^T C \hat{\beta}}}\ \hat{\beta}$$

which is the same solution as before and is in the direction of $\hat{\beta}$.

5.2.4 Mean Square Error Criterion

Vinod compromised his principle that "mean square error is a better proxy for closeness to the truth than the sum of squared residuals" slightly by minimising the latter quantity. A suggestion to minimise the former quantity seemed to be promising but was not completely successful. It is nevertheless interesting and we now examine the method.

We wish to minimise the function $E[(\beta^* - \beta)^T C (\beta^* - \beta)]$ subject to the constraint $(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) = v$. Unfortunately the expectation does not exist since the constraint cannot be satisfied if v is negative. Two alternative constraints may be considered

(i) $(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) = v_1$ where v_1 is the positive part of v (or a smooth version of it)

(ii) $\{(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) - v\}\beta^* = 0$.

The first constraint is equivalent to $\beta^* = u\hat{\beta}$ if $v < 0$ while, for $v < 0$, the latter is equivalent to $\beta^* = 0$. With consraint (i) and

Lagrange multiplier $\gamma(Y)$ we wish to find stationary values of

$$z = E[(\beta^* - \beta)^T C (\beta^* - \beta) + \gamma\{(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) - v_1\}] .$$

This is the usual caculus of variations technique for this kind of problem. Let $\beta^\dagger(Y,\alpha)$ be a parametrised class of functions with $\beta^*(Y) = \beta^\dagger(Y,0)$. For stationary values $\left.\frac{\partial z}{\partial \alpha}\right|_{\alpha=0} = 0$. This must be true for all such classes of funtions.

Writing $\frac{\partial \beta^*}{\partial \alpha} = \left.\frac{\partial \beta^\dagger}{\partial \alpha}\right|_{\alpha=0}$ we have

$$\left.\frac{\partial z}{\partial \alpha}\right|_{\alpha=0} = E[2 \frac{\partial \beta^{*T}}{\partial \alpha} C (\beta^* - \beta) + 2\gamma \frac{\partial \beta^{*T}}{\partial \alpha} C(\beta^* - u\hat{\beta})]$$

and if this is zero for each such family of functions then

$$C(\beta^* - \beta) + \gamma C(\beta^* - u\hat{\beta}) = 0.$$

This gives $\beta^* = \frac{1}{1+\gamma}\beta + \frac{\gamma u}{1+\gamma}\hat{\beta}$ from which we derive

$$\beta^* - u\hat{\beta} = \frac{1}{1+\gamma}\beta - \frac{u}{1+\gamma}\hat{\beta}$$

and substituting in the constraint equation gives

$$(1+\gamma)^2 v_1 = (\beta - u\hat{\beta})^T C (\beta - u\hat{\beta}).$$

If $v_1 = 0$ then this gives $\gamma = \infty$ and therefore $\beta^* = u\hat{\beta}$. If $v_1 > 0$ then

$$\beta^* = u\hat{\beta} \pm \sqrt{\frac{v_1}{(\beta - u\hat{\beta})^T C (\beta - u\hat{\beta})}} \; (\beta - u\hat{\beta}) .$$

This is not an estimator for β as it depends on β itself. However, it can be estimated. Putting $\hat{\beta}$ for β we obtain

$$\beta^* = u\hat{\beta} \pm \sqrt{\frac{v_1}{(1-u)^2 \hat{\beta}^T C \hat{\beta}}} \; (1-u)\hat{\beta}$$

$$= u\hat{\beta} \pm \sqrt{\frac{v_1}{\hat{\beta}^T C \hat{\beta}}}$$

as before.

An alternative, which is now a familiar trick, is to iterate to convergence. We require that

$$\sqrt{(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta})}\,(\beta^* - u\hat{\beta}) = \pm\sqrt{v_1}\,(\beta^* - u\hat{\beta})$$

with $\beta^* \neq u\hat{\beta}$. This is merely the constraint equation so we have solved nothing.

For constraint (ii) we must use a vector of Lagrange multipliers and find a stationary value of

$$z = E[(\beta^* - \beta)^T C (\beta^* - \beta) + \gamma^T C \beta^* \{(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) - v\}].$$

The condition for stationary values is

$$2C(\beta^* - \beta) + \{(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) - v\}C\gamma + 2\gamma^T C \beta^* C (\beta^* - u\hat{\beta}) = 0$$

which may be written

$$\beta^* - \beta + \tfrac{1}{2}\{(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) - v\}\gamma + \gamma^T C \beta^* (\beta^* - u\hat{\beta}) = 0.$$

If $v < 0$ then $\beta^* = 0$ from the constraint equation. We wish to solve for the case when $v > 0$. In fact $\beta^* = 0$ is always a solution of the constraint equation, but for $v > 0$ we are interested in other solutions which are solutions of $(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) = 0$ and this gives

$$(1) \quad (\beta^* - \beta) + \gamma^T C \beta^* (\beta^* - u\hat{\beta}) = 0.$$

From these we see that

$$(\beta^* - u\hat{\beta})^T C (\beta^* - \beta) + \gamma^T C \beta^* (\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) = 0$$

i.e.

$$v + (\beta^* - u\hat{\beta})^T C (u\hat{\beta} - \beta) + \gamma^T C \beta^* v = 0$$

from which we find $\gamma^T C \beta^*$ and substitute back into (1) to obtain

$$(\beta^* - \beta) - \{1 + \frac{1}{v}(\beta^* - u\hat{\beta})^T C (\beta - u\hat{\beta})\}(\beta^* - u\hat{\beta}) = 0.$$

Simplifying we obtain

$$(\beta^* - u\hat{\beta})^T C (\beta - u\hat{\beta})(\beta^* - u\hat{\beta}) = v(\beta - u\hat{\beta}).$$

Finally, we write this as

$$\{(\beta^* - u\hat{\beta})(\beta^* - u\hat{\beta})^T C - vI\}(\beta - u\hat{\beta}) = 0.$$

Thus we are lead to the result that $\beta - u\hat{\beta}$ must be an eigenvector of $(\beta^* - u\hat{\beta})(\beta^* - u\hat{\beta})^T C$ and v its non-zero eigenvalue (the latter statement is equivalent to the constraint equation).

Now, the only eigenvector of the matrix above, which corresponds to a non-zero eigenvalue, is $\beta^* - u\hat{\beta}$ and thus $\beta^* = \beta$. This is impossible as this value does not satisfy the constraint. Thus, assuming there is a stationary value other than $\beta^* = 0$ leads to a contradiction and therefore $\beta^* = 0$ is the only stationary value. We have thus shown that this approach leads nowhere.

5.2.5 Maximising the Expected Length

Consider the constraint in the form $(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) = v_1$. We wish to maximise $E[\beta^{*T} C \beta^*]$ under this constraint. By the method of the previous section we obtain

$$\beta^* + \gamma(\beta^* - u\hat{\beta}) = 0$$

for a stationary value, and this gives

$$\beta^* = \frac{\gamma u}{1+\gamma}\hat{\beta} = u\hat{\beta} - \frac{u}{1+\gamma}\hat{\beta}.$$

Substituting this into the constraint equation gives

$$u^2 \hat{\beta}^T C \hat{\beta} = (1+\gamma)^2 v_1.$$

Thus the solutions are

$$\beta^* = u\hat{\beta} \pm \sqrt{\frac{v_1}{\hat{\beta}^T C \hat{\beta}}}.$$

With the positive sign this agrees with the solution which maximises the length. It is easily seen that this must be so since, if the length is maximum for each value of Y, then the expected length must also be a maximum.

5.3 The Case for which there is No Solution

When the only fixed point of the iteration is $\beta^* = 0$ we have $\hat{\lambda} < \frac{k\nu}{\mu}\frac{1}{1-u}$. We have already suggested that $\beta^* = u\hat{\beta}$ or $\beta^* = 0$ should be taken as the solution according as we reject or accept the hypothesis $\beta = 0$ at about the 50% level.

Another argument for using the solution $\beta^* = u\hat{\beta}$ is that this gives, in a sense, the closest approach to a solution of (5.2.1.1). This is the closest apprach to a solution in the sense that it minimises $(\beta^* - u\hat{\beta})^T C (\beta^* - u\hat{\beta}) - v$.

An alternative is to note that, at each iteration, we may write $\beta_i^* = \theta_i \beta_o^*$. If there is no fixed point then we might choose the closest value to a fixed point in that the derivative of θ_{i+1} with respect to θ_i is unity. Unfortunately this leads to a quartic equation and as we already have a reasonable solution for this case we proceed no further.

Another alternative is to write $\beta_{i+1}^* = \theta_i \beta_i^*$ and, if we cannot find a solution for which $\theta_i = 1$ we maximise θ_i. For a maximum $\frac{\partial \theta_i}{\partial \beta_i^*} = 0$. We shall omit the subscripts so the quantity to be minimised is written $z = \frac{\beta^{*T} C \beta}{k\sigma^{2*} + \beta^{*T} C \beta^*}$ where

$$\sigma^{2*} = \frac{\nu}{\mu}\hat{\sigma}^2 + \frac{\alpha}{\mu}(\beta^* - \hat{\beta})^T C (\beta^* - \hat{\beta}).$$

Now $$\frac{\partial z}{\partial \beta^*} = \frac{(k\sigma^{2*} + \beta^{*T} C \beta^*) C\hat{\beta} - \frac{2k\alpha}{\mu}\beta^{*T} C \hat{\beta}\, C(\beta^* - \hat{\beta}) - 2\beta^{*T} C \hat{\beta}\, C\beta^*}{(k\sigma^{2*} + \beta^{*T} C \beta^*)^2}$$

and this is zero when

$$(k\sigma^{2*} + \beta^{*T} C \beta^{*}) C \hat{\beta} = \frac{2k\alpha}{\mu} \beta^{*T} C \hat{\beta} C(\beta^{*} - \hat{\beta}) + 2\beta^{*T} C \hat{\beta} C\beta^{*}.$$

This gives $\beta^{*} = h\hat{\beta}$ where

$$\frac{k\nu}{\mu} \hat{\sigma}^2 + \frac{k\alpha}{\mu} (h-1)^2 \hat{\beta}^T C \hat{\beta} + h^2 \hat{\beta}^T C \hat{\beta} = \frac{2k\alpha}{\mu} \hat{\beta}^T C \hat{\beta}(h^2-h) + 2\hat{\beta}^T C \hat{\beta} h^2 .$$

This quadratic equation for h gives

$$(1 + \frac{k\alpha}{\mu}) \hat{\lambda} h^2 = \frac{k\alpha}{\mu} \hat{\lambda} + \frac{k\nu}{2\mu} .$$

Clearly we require the positive root. Now $v = \frac{1}{2}\left(\frac{\mu}{\mu+k\alpha}\right)^2 \hat{\lambda} - \frac{k\nu}{\mu+k\alpha}$

and we wish to show that $\theta < 1$ when $v < 0$. If $v < 0$ then

$\hat{\lambda} < \frac{2(\mu+k\alpha)k\nu}{\mu^2}$ and therefore

$$\begin{aligned} h^2 &= \frac{k}{\mu+k\alpha} (\alpha + \frac{\nu}{2\hat{\lambda}}) \\ &> \frac{k\alpha}{\mu+k\alpha} + \frac{k\nu}{4(\mu+k\alpha)} \frac{\mu^2}{k\nu(\mu+k\alpha)} \\ &= \frac{k\alpha}{\mu+k\alpha} + \frac{\mu^2}{4(\mu+k\alpha)^2} \\ &= \frac{4k\alpha(\mu+k\alpha) + \mu^2}{4(\mu+k\alpha)^2} \\ &= \frac{(2k\alpha+\mu)^2}{4(\mu+k\alpha)^2} \\ &= \left(\frac{\frac{1}{2}\mu+k\alpha}{\mu+k\alpha}\right)^2 \\ &= u^2 . \end{aligned}$$

Thus $h > u$ with equality if $v = 0$.

Now

$$\theta = \frac{h\hat{\beta}^T C \hat{\beta}}{h^2 \hat{\beta}^T C \hat{\beta} + \frac{k\nu}{\mu} \hat{\sigma}^2 + \frac{k\alpha}{\mu} (h-1)^2 \hat{\beta}^T C \hat{\beta}}$$

and by the equation for h,

$$\begin{aligned} \theta &= \frac{h\hat{\beta}^T C \hat{\beta}}{\frac{2k\alpha}{\mu} \hat{\beta}^T C \hat{\beta} (h^2-h) + 2\hat{\beta}^T C \hat{\beta} h^2} \\ &= \frac{1}{2h(1 + \frac{k\alpha}{\mu}) - \frac{2k\alpha}{\mu}} \\ &< \frac{1}{2u(1 + \frac{k\alpha}{\mu}) - \frac{2k\alpha}{\mu}} \\ &= 1 . \end{aligned}$$

note however that as $\hat{\lambda} \to 0$, h increases. This is not a desirable property and the solution is not recommended.

5.4 Stability of Fixed Point Solutions

Writing $\beta^*_{i+1} = \theta\beta^*_i$ we obtain $\theta = \dfrac{\beta^{*T}_i C\hat{\beta}}{\sigma^{2*} + \beta^{*T}_i C\beta^*_i}$ and $\theta < 1$ if and only if

$$\sigma^{2*} + \beta^{*T}_i C\beta^*_i - \beta^{*T}_i C\hat{\beta} < 0.$$

The equation $(\sigma^{2*} + \beta^{*T}_i C\beta^*_i - \beta^{*T}_i C\hat{\beta})\beta^*_i = 0$ was solved in section 5.2. Similar simplification of the inequality leads to $\theta < 1$ if and only if

$$(\beta^*_i - u\hat{\beta})^T C(\beta^*_i - u\hat{\beta}) - v > 0.$$

Thus, if $\beta^{*T}_i C\beta > 0$ then $\theta > 0$ and the iteration converges if and only if $(\beta^*_i - u\hat{\beta})^T C(\beta^*_i - u\hat{\beta}) - v > 0$. This is because the components of the successive iterated vectors form a decreasing sequence bounded below by zero (for positive components - if a component is negative then its successive iterates form an increasing sequence bounded above by zero).

Completing the square in the denominator for θ gives

$$\frac{1}{1-w}\{(\beta^*_i - w\hat{\beta})^T C(\beta^*_i - w\hat{\beta}) + w(1-w)\hat{\beta}^T C\hat{\beta} + \frac{k\nu}{\mu}(1-w)\delta^2\}$$

where $w = \dfrac{k\alpha}{\mu+k\alpha} = 2u - 1$. Since $0 \le w < 1$ the denominator is positive and θ is continuous as a function of β^*_i (this is also clear from the original form of the denominator).

Now, if $\beta^{*T}_i C\beta < 0$ then $\beta^{*T}_{i+1} C\hat{\beta} = \dfrac{\beta^{*T}_i C\hat{\beta}}{\beta^{*T}_i C\beta^*_i + k\sigma^{2*}}\beta^{*T}_i C\hat{\beta} > 0.$

Thus, after the first iteration, $\beta^{*T} C\hat{\beta} \ge 0$ whatever the starting value and thus the iteration converges if for some starting value $\theta < 1$. If the iteration has only one fixed point then the limit must be $\beta^*_\infty = 0$.

Furthermore, writing $\beta^*_i = \phi_i\beta^*_o$ we have

$$\beta^*_{i+1} = \frac{\beta^{*T}_i C\hat{\beta}}{\beta^{*T}_i C\beta^*_i + k\sigma^{2*}_i}\beta^*_i = \frac{\phi_i^2\beta^{*T}_o C\hat{\beta}}{\phi_i^2\beta^{*T}_o C\beta^*_o + k\sigma^{2*}_i}\beta^*_o = \phi_{i+1}\beta^*_o$$

Thus the absolute value of the shrinkage factor is smaller than $\left|\dfrac{\beta^{*T}_o C\hat{\beta}}{\beta^{*T}_o C\beta^*_o}\right|$. The transfer function from ϕ_i to ϕ_{i+1} is shown in the

graph in figure 13. It is symmetric if $\alpha = 0$ and almost so for large values of $|\phi_i|$. We have drawn the graph for the case of three fixed points (if there is only one fixed point then the graph only crosses the line $\phi_{i+1} = \phi_i$ at the origin).

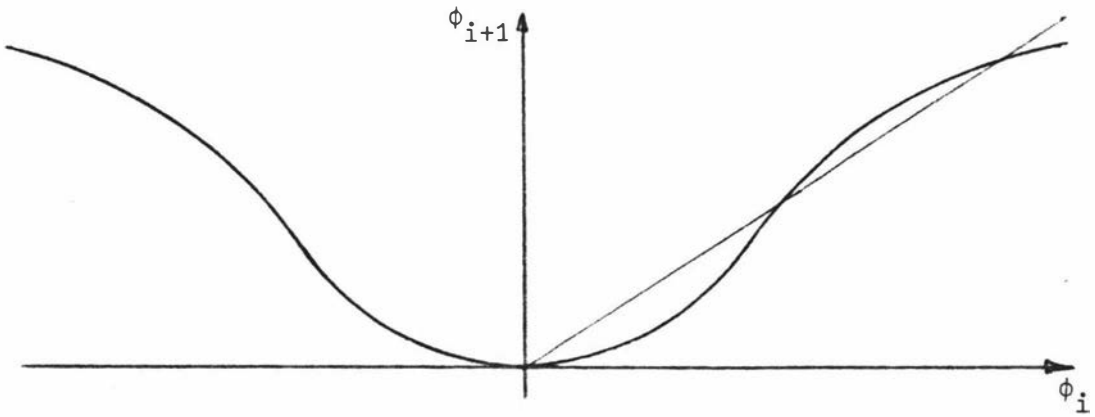

Figure 13 Transfer Function for Fixed Point Iteration of Bilinear Shrinkage Estimators

By the continuity of the function, the gradient must be greater than unity at the middle crossing and less than unity at the others. The middle solution is therefore unstable and the others are stable. This is another reason for preferring the positive root in the equation for β^*_∞ (this gives the upper crossing).

Finally note that, after the first iteration, the component of β^*_i in the direction of $\hat{\beta}$ is a shrinkage of $\hat{\beta}$. We see this as follows. Let $\beta^*_o = a\hat{\beta} + \delta$ and let $\delta^T C \hat{\beta} = 0$. Now

$$\beta^*_{i+1} = \frac{\phi_i^2 \beta_o^{*T} C \hat{\beta}}{\phi_i^2 \beta_o^{*T} C \beta^*_o + k\sigma^{2*}_i} \beta^*_o$$

$$= \frac{\phi_i^2 a^2 \hat{\beta}^T C \hat{\beta}}{\phi_i^2 a^2 \hat{\beta}^T C \hat{\beta} + \phi_i^2 \delta^T C \delta + k\sigma^{2*}_i} \hat{\beta} + \frac{\phi_i^2 a \hat{\beta}^T C \hat{\beta}}{\phi_i^2 a^2 \hat{\beta}^T C \hat{\beta} + \phi_i^2 \delta^T C \delta + k\sigma^{2*}_i} \delta$$

Now the modulus of the shrinkage factor for $\hat{\beta}$ is less than unity (but note that $a = 0$ implies $\beta^*_i = 0$ if $i > 0$).

5.5 Another Fixed Point Iteration

We argued in chapter 4 that we may be able to improve our estimation of the quadratic expressions in our shrinkage factor. If this is so then we might find an improved estimator for β by using an iterative version of the estimators in chapter 4. In that chapter we suggested estimators of the form $b\hat{\beta}\hat{\beta}^T C + a\sigma^2 I$ for $\beta\beta^T C$ and $b\hat{\beta}^T C \hat{\beta} + a\hat{\sigma}^2 I$ for $k\sigma^2 + \beta^T C \beta$. In the iterative process to be

considered we shall substitute β_i^* for $\hat{\beta}$ and σ^{2*} for $\hat{\sigma}^2$ in these estimators. Thus we take

$$\beta_{i+1}^* = \frac{b\beta_i^*\beta_i^{*T} C + a\sigma^{2*} I}{d\beta_i^{*T} C \beta_i^* + c\sigma^{2*}} \hat{\beta}$$

to define our iterative process. There may not be any best values of a, b, c or d at a particular step, but some reasonable choice needs to be made - therefore we should think of these as functions of β_i^*. However, in view of the complication of choosing these values, we prefer to keep them fixed but unknown. After finding the fixed points we shall try to choose these coefficients so as to minimise the risk. Another problem, which also occurred in chapter 4, is that the denominator may vanish if $c < 0$. To avoid this problem we shall take $b = d$ and $a < c$, then we shall set the shrinkage to zero when the numerator vanishes.

In order to find the fixed points of the iteration we shall, for the moment, ignore this problem.

5.5.1 Fixed Points

For β_∞^* to be a fixed point we require that

$$\beta_\infty^* = \frac{b\,\beta_\infty^*\beta_\infty^{*T} C + a\sigma^{2*} I}{d\beta_\infty^{*T} C \beta_\infty^* + c\sigma^{2*}} \hat{\beta}$$

i.e. $$(d\beta_\infty^{*T} C \beta_\infty^* + c\sigma^{2*}) \beta_\infty^* = b\beta_\infty^{*T} C \hat{\beta}\, \beta_\infty^* + a\sigma^{2*} \hat{\beta}.$$

Thus, if $a \neq 0$ then $\beta_\infty^* \propto \hat{\beta}$ - a result obtained without imposing any side conditions on our solutions as we had to do with the previous form of the shrinkage. Substituting the formula for σ^{2*} we obtain

$$\{(d + \frac{c\alpha}{\mu})\beta_\infty^{*T} C \beta_\infty^* - (b + \frac{2c\alpha}{\mu})\beta_\infty^{*T} C \hat{\beta} + \frac{c\alpha}{\mu} \hat{\beta}^T C \hat{\beta} + \frac{c\nu}{\mu} \hat{\sigma}^2\}\beta_\infty^*$$

$$= \{\frac{a\alpha}{\mu} \beta_\infty^{*T} C \beta_\infty^* - \frac{2a\alpha}{\mu} \beta^{*T} C \hat{\beta} + \frac{a\alpha}{\mu} \hat{\beta}^T C \hat{\beta} + \frac{a\alpha}{\mu} \hat{\sigma}^2\}\hat{\beta}.$$

Putting $\beta_\infty^* = h\hat{\beta}$ we obtain

$$(d + \frac{c\alpha}{\mu}) \hat{\lambda} h^3 - (b + \frac{2c\alpha}{\mu}) \hat{\lambda} h^2 + \frac{c\alpha}{\mu} \hat{\lambda} h + \frac{c\nu}{2\mu} h$$

$$= \frac{a\alpha}{\mu} \hat{\lambda} h^2 - \frac{2a\alpha}{\mu} \hat{\lambda} h + \frac{a\alpha}{\mu} \hat{\lambda} + \frac{a\alpha}{2\mu}.$$

Finally, after simplifying this we obtain

$$(d + \frac{c\alpha}{\mu}) h^3 - (b + \frac{2c\alpha}{\mu} + \frac{a\alpha}{\mu}) h^2 + (\frac{c\alpha}{\mu} + \frac{2a\alpha}{\mu} + \frac{c\nu}{2\mu\hat{\lambda}}) h + (\frac{a\alpha}{\mu} + \frac{a\alpha}{2\mu\hat{\lambda}}) = 0$$

which has three solutions for h, one of which must be real. The other

solutions are comlex conjugates if they are not real. In order to choose among these roots it seems sensible to use one of the criteria used in our previous choice of solution. By analogy with that section we shall choose the largest real root (which is always less than unity if $b = d$) or the largest real part among the roots.

5.5.2 Convergence of the Iteration

In order to investigate the manner of the convergence of the iteration we shall make the assumption that $c > 0$. We shall relax this assumtion later. The iterative process is defined by the equation

$$\beta^*_{i+1} = u_i \beta^*_i + v_i \hat{\beta}$$

where
$$u_i = \frac{b \beta^{*T}_i C \hat{\beta}}{d \beta^{*T}_i C \beta^*_i + \frac{c\nu}{\mu} \delta^2 + \frac{c\alpha}{\mu} (\beta^*_i - \hat{\beta})^T C (\beta^*_i - \hat{\beta})}$$

and
$$v_i = \frac{\frac{a\nu}{\mu} \delta^2 + \frac{a\alpha}{\mu} (\beta^*_i - \hat{\beta})^T C (\beta^*_i - \hat{\beta})}{c \beta^{*T}_i C \beta^*_i + \frac{c\nu}{\mu} \delta^2 + \frac{c\alpha}{\mu} (\beta^*_i - \hat{\beta})^T C (\beta^*_i - \hat{\beta})} .$$

Now we may write any starting vector β^*_o as

$$\beta^*_o = \theta_o \hat{\beta} + \phi_o \delta \quad \text{where} \quad \hat{\beta}^T C \delta = 0 \quad \text{and} \quad \delta^T C \delta = \hat{\beta}^T C \hat{\beta} \quad (\hat{\beta} \neq 0).$$

The values of θ_o, ϕ_o and δ are $\theta_o = \frac{\beta^{*T}_o C \hat{\beta}}{\hat{\beta}^T C \hat{\beta}}$,

$$\phi_o = \left\{ \frac{(\beta^*_o - \theta_o \hat{\beta})^T C (\beta^*_o - \theta_o \hat{\beta})}{\hat{\beta}^T C \hat{\beta}} \right\}^{\frac{1}{2}} \quad \text{and} \quad \delta = \frac{1}{\phi_o} (\beta^*_o - \theta_o \hat{\beta}) .$$

We show by induction that we may write $\beta^*_i = \theta_i \hat{\beta} + \phi_i \delta$. Assume this to be so. In that case

$$\begin{aligned} \beta^*_{i+1} &= u_i \beta^*_i + v_i \hat{\beta} \\ &= u_i(\theta_i \hat{\beta} + \phi_i \delta) + v_i \hat{\beta} \\ &= (u_i \theta_i + v_i) \hat{\beta} + u_i \phi_i \delta . \end{aligned}$$

We thus have the recurrence relation

$$\theta_{i+1} = u_i \theta_i + v_i , \qquad \phi_{i+1} = u_i \phi_i .$$

Thus all iterations give vectors lying in the same plane as $\hat{\beta}$ and β^*_o. Substituting for u_i and v_i we obtain the following recurrence relations for θ_i and ϕ_i

$$\theta_{i+1} = \frac{b(\theta_i\hat{\beta}+\phi_i\delta)^T C\hat{\beta}\theta_i + \frac{a\nu}{\mu}\hat{\sigma}^2 + \frac{a\alpha}{\mu}\{(\theta_i-1)\hat{\beta}+\phi_i\delta\}^T C\{(\theta_i-1)\hat{\beta}+\phi_i\delta\}}{d(\theta_i\hat{\beta}+\phi_i\delta)^T C(\theta_i\hat{\beta}+\phi_i\delta) + \frac{c\nu}{\mu}\hat{\sigma}^2 + \frac{c\alpha}{\mu}\{(\theta_i-1)\hat{\beta}+\phi_i\delta)\}^T C\{(\theta_i-1)\hat{\beta}+\phi_i\delta\}}$$

$$= \frac{b\,\theta_i^2 + \frac{a\alpha}{\mu}(\theta_i-1)^2 + \frac{a\alpha}{\mu}\phi_i^2 + \frac{a\nu}{2\mu\lambda}}{d\,\theta_i^2 + \frac{c\alpha}{\mu}(\theta_i-1)^2 + d\,\phi_i^2 + \frac{c\alpha}{\mu}\phi_i^2 + \frac{c\nu}{2\mu\lambda}}$$

and

$$\phi_{i+1} = \frac{b\,\theta_i\,\phi_i}{d\,\theta_i^2 + \frac{c\alpha}{\mu}(\theta_i-1)^2 + d\,\phi_i + \frac{c\alpha}{\mu}\phi_i^2 + \frac{c\nu}{2\mu\lambda}} .$$

If $\theta_i \neq 0$ then let $\psi_i = \frac{\phi_i}{\theta_i}$ so that

$$\psi_{i+1} = \frac{b\,\theta_i^2\,\psi_i}{b\,\theta_i^2 + \frac{a\alpha}{\mu}(\theta_i-1)^2 + \frac{a\alpha}{\mu}\phi_i^2 + \frac{a\nu}{2\mu\lambda}} .$$

We shall replace negative values of θ_i by zero which means that θ_{i+1} and ϕ_{i+1} are also zero if $a < 0$. If $a < 0$ and $\theta_{i+1} > 0$ then $|\psi_{i+1}| > |\psi_i|$. Thus if $\phi_i > 0$ then the $|\psi_i|$ form an increasing sequence and do not tend to zero. Therefore the ϕ_i cannot tend to zero and thus cannot converge (zero is the only fixed value for ϕ_i). On the other hand, if $a > 0$ then the $\psi_i \to 0$ and therefore $\phi_i \to 0$. If, when $a < 0$, a fixed point has $\theta_\infty > 0$ and $\phi_\infty = 0$ then it cannot therefore be stable. However a modified process in which we set each ϕ_i to zero is worth considering. We may argue, as for the previous iteration, that the middle fixed point is the only unstable one for the modified iteration (the iteration needs no modification if $a > 0$).

We now relax the restriction that $c > 0$. As suggested earlier, if the denominator of the shrinkage factor vanishes then so should the numerator. In this case we replace the shrinkage factor by zero. This is equivalent to setting θ_i and ϕ_i to zero. The previous arguments are unchanged so long as θ_i and ϕ_i remain positive.

5.6 Practical Estimators

In order to use the fixed point estimator of the previous section we need to solve a cubic equation. This is easily done using a numerical procedure. If the solutions are all real then we take the largest root; if two are complex conjugates then we take their real part or use $\beta^*_\infty = 0$ if we want to accept the hypothesis $\beta = 0$.

5.7 Graphs of Risk Functions for Fixed Point Estimators

Risk functions for the James-Stein estimator and its positive part version, positive part bilinear shrinkage estimator and the fixed point estimators were all computed using the formula 6.3.2 which, for the spherically symmetric case becomes

$$r_\delta(F) = E\left[\frac{p-2}{4F}\, v(2 - v) + (1 + c\, v)\,\frac{dv}{dF}\right]$$

where $\delta(X,S) = (1 - \frac{\tilde{c}\, v(F)}{F})X$, $c = \frac{p-2}{\nu+2}$, $\tilde{c} = \frac{\nu}{p}\, c$

and the expectation is with respect to the distribution of the usual F test statistic for testing $\beta = 0$ i.e. $F = \frac{1}{p}\,\|X\|^2/S$. In order that the integration should be over a finite range, this distribution was transformed to a non-central beta distribution using the transformation $U = \frac{pF}{\nu + pF}$. The integral was divided into four ranges, the lower and upper tails and two central areas on either side of the approximate mode. This ensured more rapid convergence and made certain that the integration procedure did not miss the narrow peak which occurs in the beta distribution when ν or λ is large. The procedure used repeated bisection of the interval farthest from the mode until the convergence criterion was satisfied, and then repeated the process for the next interval. A maximum of 203 evaluations was allowed in each of the four intervals and the values of the beta density function were saved for the evaluation of future risk functions. In certain cases the beta density is unbounded near $U = 0$ or $U = 1$ so these extreme values were not included in the range of integration. The range extended close enough to these end points for an upper bound to the integral over the neglected intervals to be less than a tolerable error threshold. The central regions extended to approximately three times the standard deviation either side of the mean (unless this led to values outside the interval $[0,1]$).

In order to choose the ranges of integration, the beta density function was plotted with a wide range of parameter values. On the same graphs an approximation to the density based on an approximation to the non-central F distribution given by Searle(1971) and various points were marked. These points corresponded to the mode and points of inflection of the approximate density, the transformed inflection points of the approximate F density, the mean and points one standard deviation each side for the approximate beta distribution,

and the transfornation of these points for the non-central F distribution.

In most cases it was observed that the approximate density and the true density were almost indistinguishable, and where they differed visibly the difference was not great. The mode of the approximate beta distribution was chosen as a cetral value and, as a measure of width, the standard deviation was multiplied by the mode divided by the mean. The other curves shown with the graphs of the density functions in figures 14-19 (which are a selection from the set of graphs plotted) are the cumulative curves calculated at four points by two different numerical routines and smoothed by fitting a piecewise cubic function to the cumulative frequencies and its derivative. These routines were not subsequently used in the evaluation of the risk functions.

When evaluating the risk estimate, the derivative of $v(F)$ is required. While this presented no dificulty for most of the estimators, the derivative was sometimes discontinuous. For the iterated estimators of this chapter the formula for the derivative of an implicit function was used.

Initially the cubic equation for the shrinkage factor was solved using the procedure REALPOLYZEROFINDER in the Burrough's numerals package. This proved to be too sloww and was replaced by a specially written procedure which proved to be ten times faster. This procedure used the well known algebraic solution in the case in which the equation has two complex roots. For the case of three real roots an iterative solution, based on the well known trigonometrical solution, was used. The numerals package was, however, used for the cubic spline subroutine and for the numerical integration procedure.

The program was written to be used interactively so that information from earlier plots could be used to help with later plots. In particular this enabled values for the parameters in the estimators to be chosen close to values which had shown promise, and also enabled the ranges of values for the axes to be chosen interacively. It was also possible to choose the number of graphs to be plotted per frame and which graphs to be so plotted. Interactive runs were used in order to gain experience with the program. Once choices of parameters etc. had been made the remainder of the computations were done in batch mode.

The risk function was plotted for seven equally spaced values of

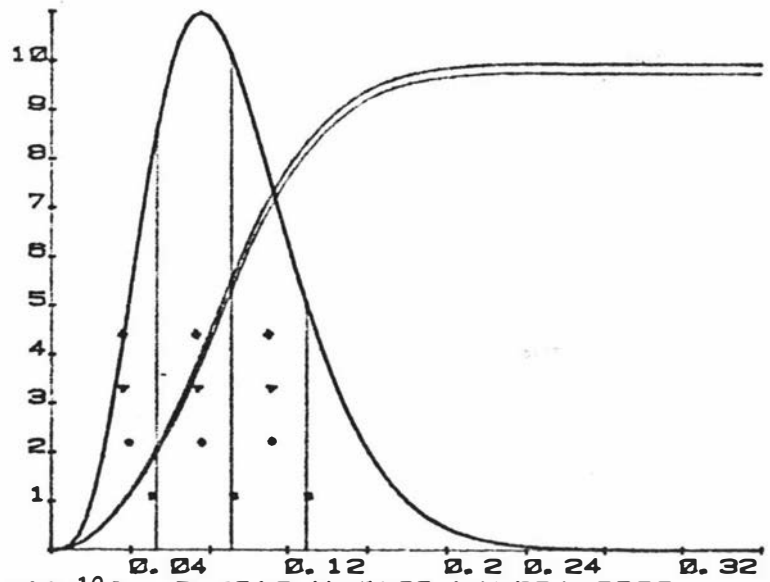

Figure 19a P=0010 N=0100 LAMBDA=0000

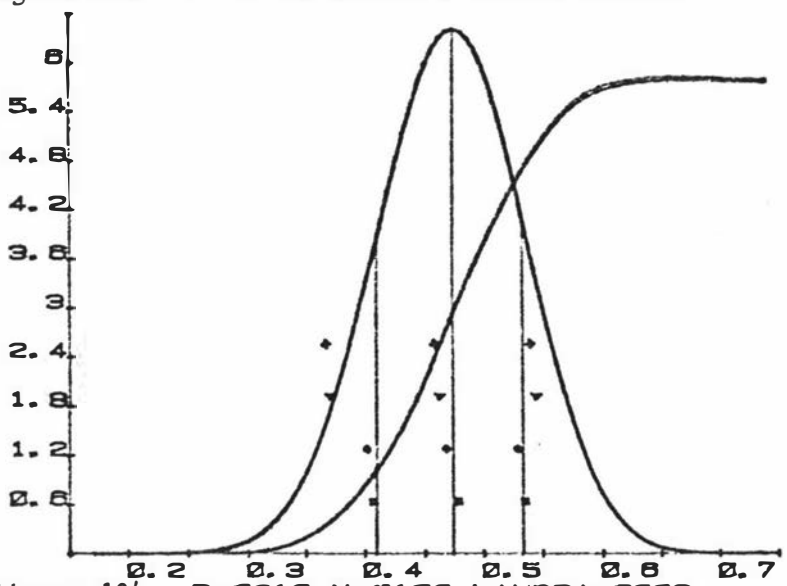

Figure 19b P=0010 N=0100 LAMBDA=0002

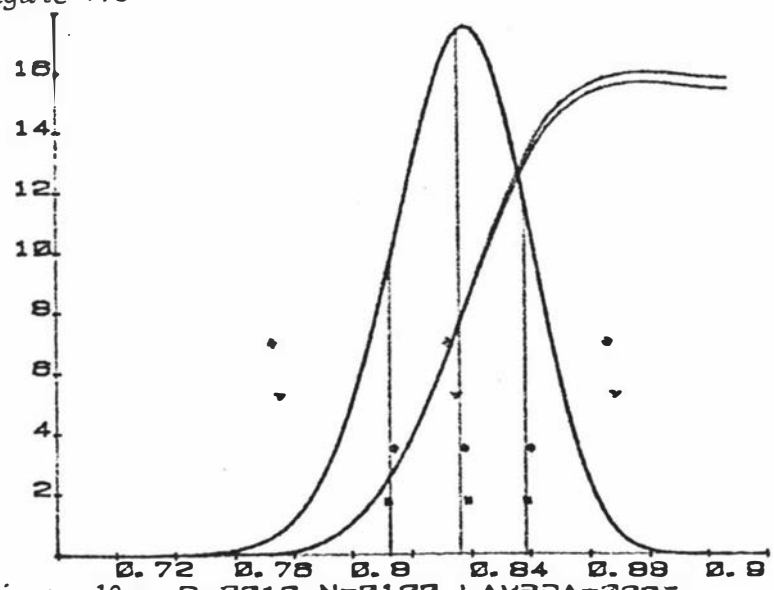

Figure 19c P=0010 N=0100 LAMBDA=0005

Figure 19 Non-central Beta Distributions

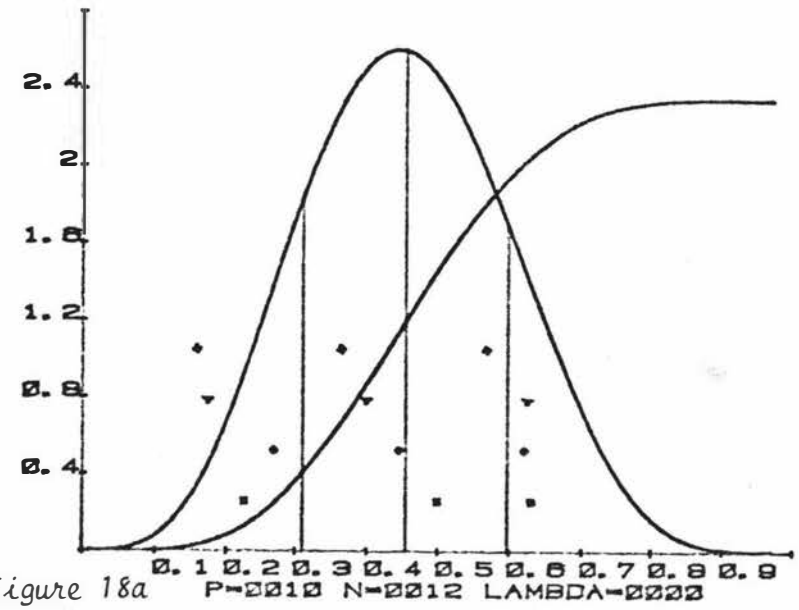

Figure 18a

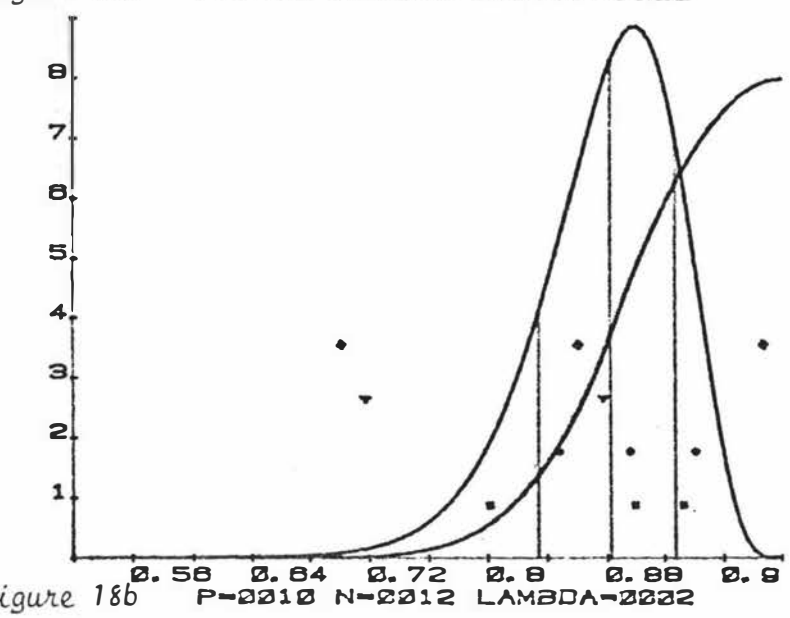

Figure 18b

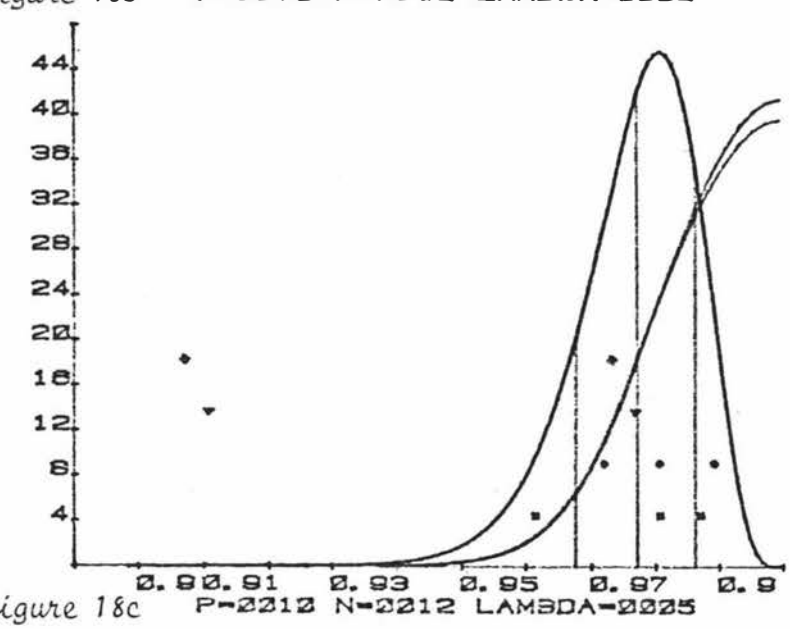

Figure 18c

Figure 18 Non-central Beta Distributions

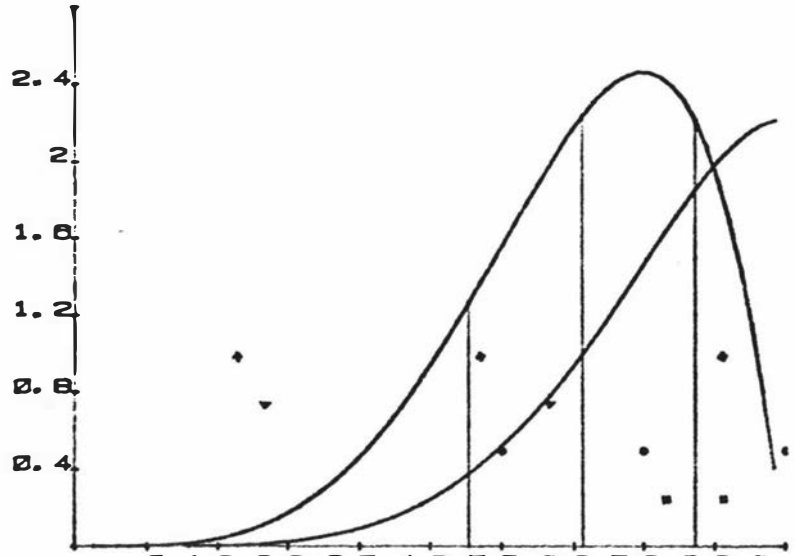

Figure 17a

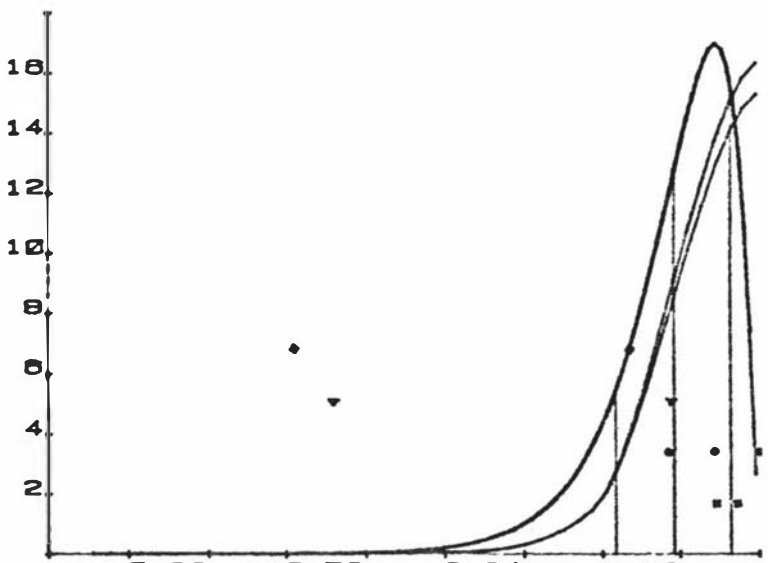

Figure 17b

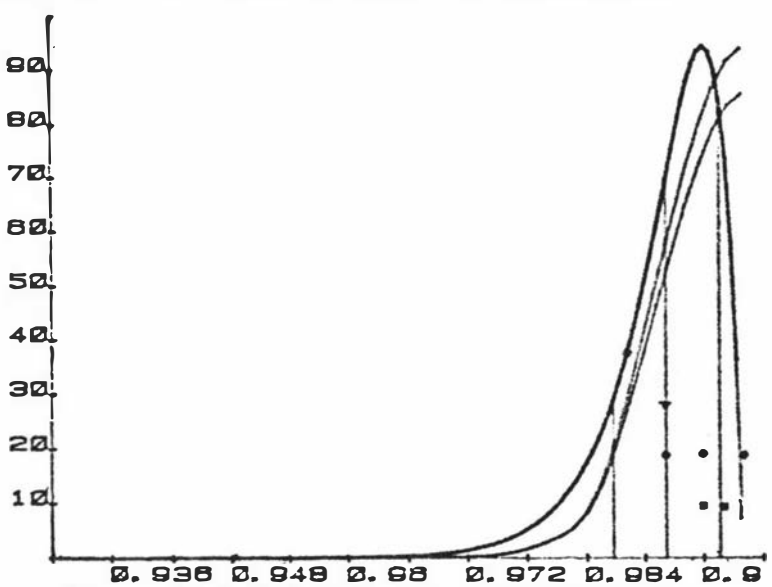

Figure 17c

Figure 17 Non-central Beta Distributions

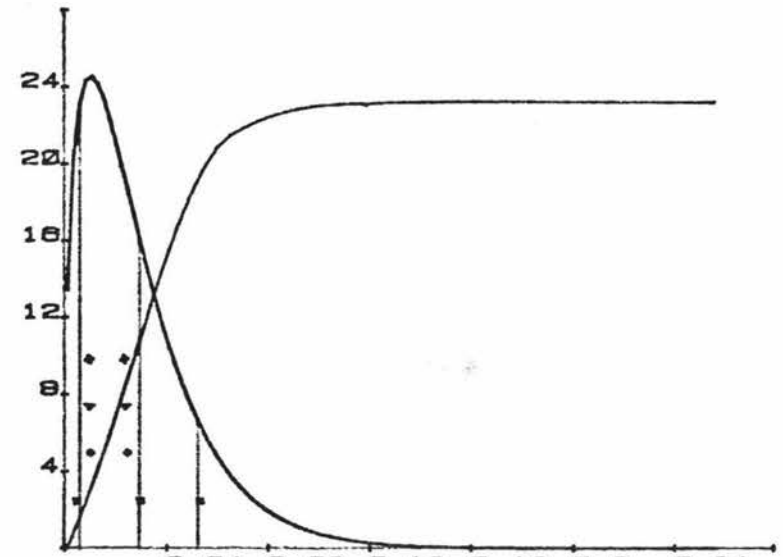

Figure 16a

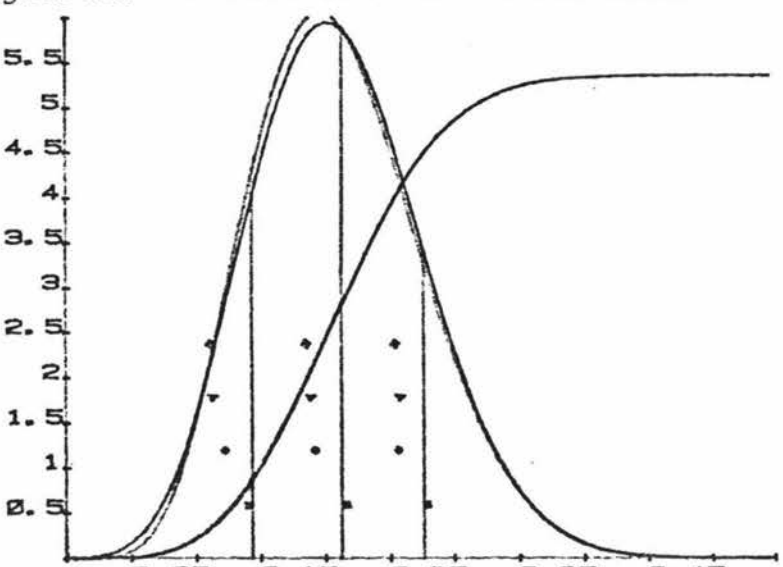

Figure 16b

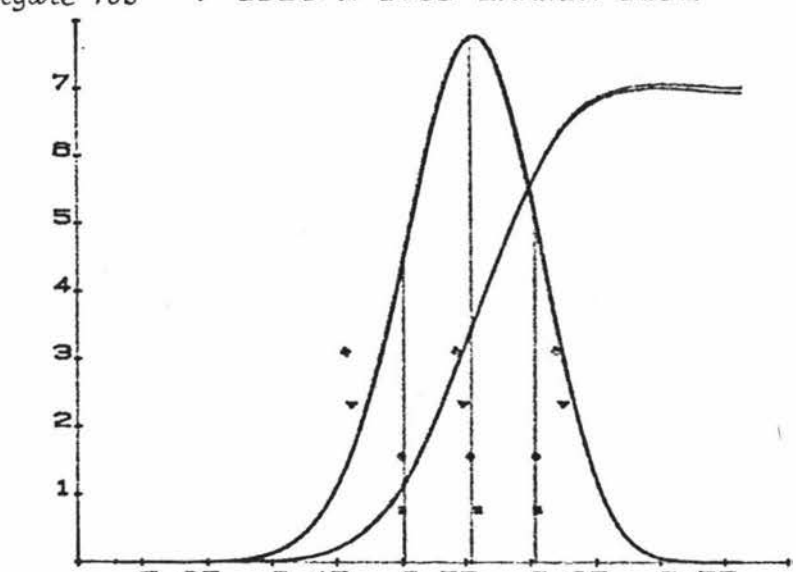

Figure 16c

Figure 16 Non-central Beta Distributions

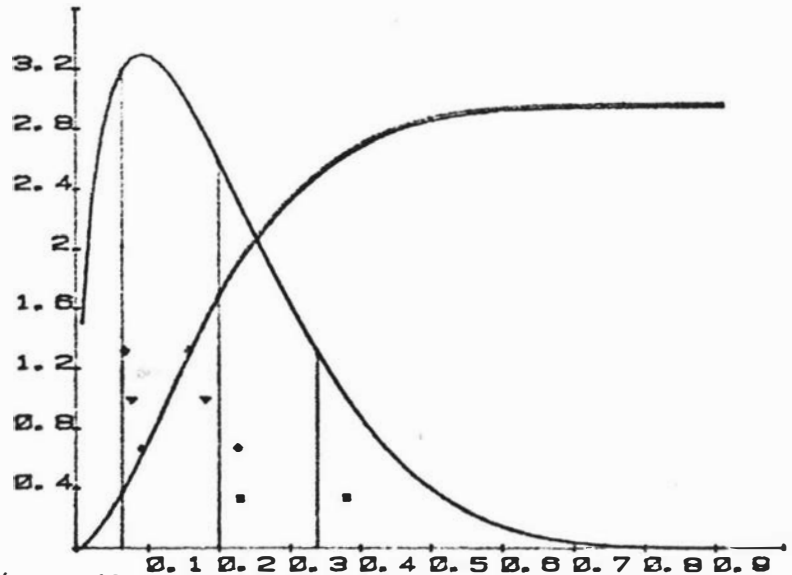

Figure 15a

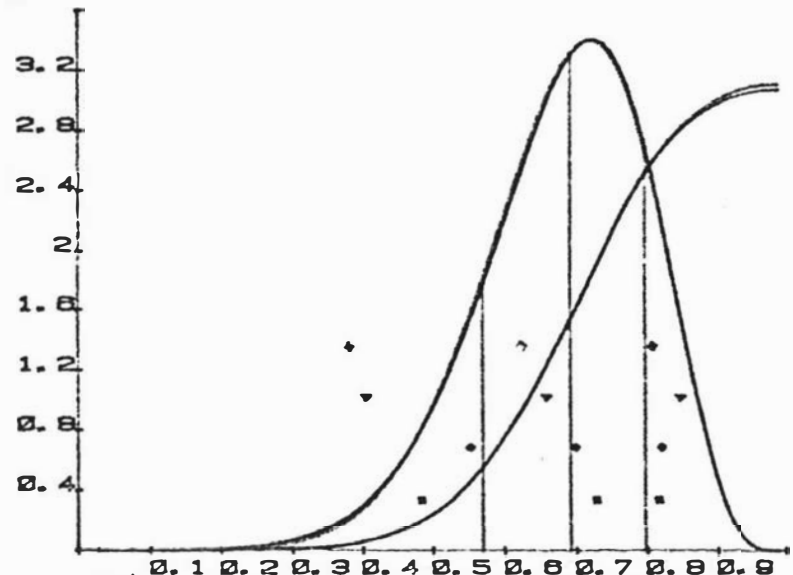

Figure 15b

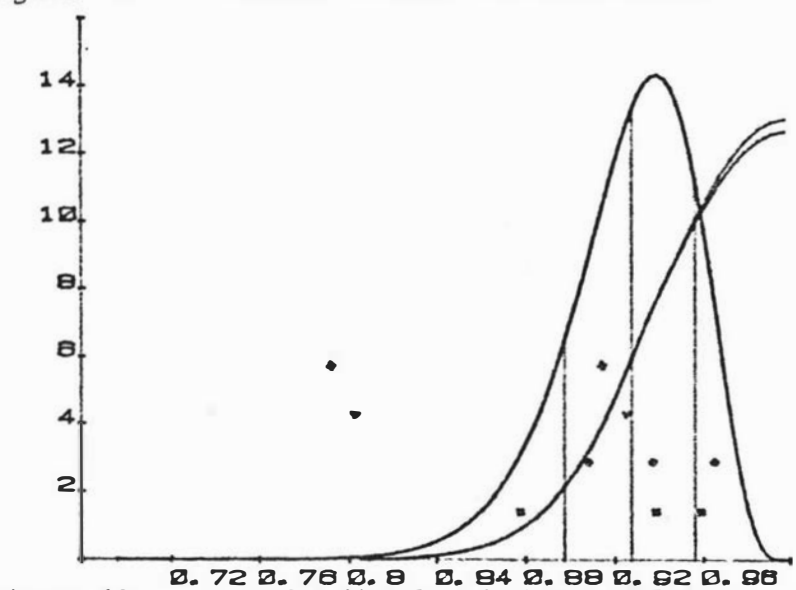

Figure 15c

Figure 15 Non-central Beta Distributions

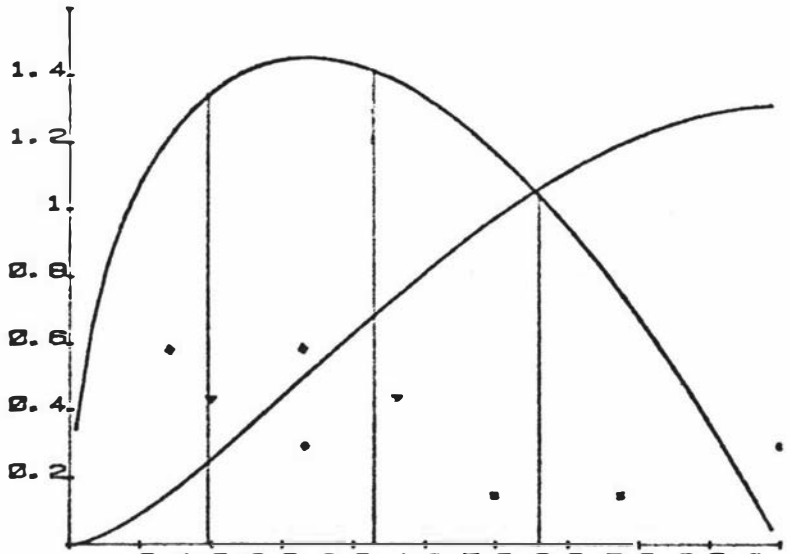

Figure 14a

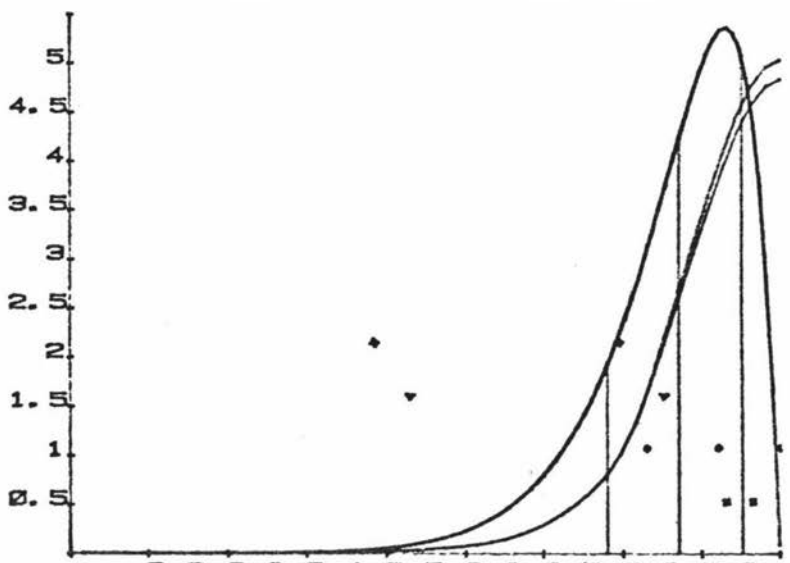

Figure 14b

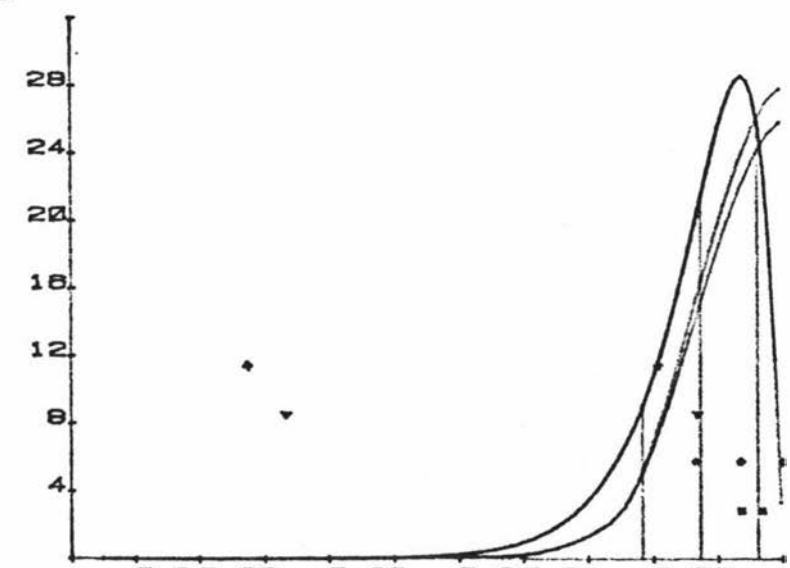

Figure 14c

Figure 14 Non-central Beta Distributions

$\phi = \sqrt{\frac{\lambda}{r}}$ between zero and three - the number of points and the range being chosen interactively when the program was run. A cubic spline was used to smoothly fill in the intermediate values. In cases in which the risk fluctuates wildly, this cubic spline is a poor fit and gives graphs which oscillate in the extreme. An aberrant spline was therefore taken to mean that the estimator was also poor. An example of this behaviour is shown in figure 27.

For the estimator $\delta(X,S) = (1 - \frac{\tilde{c}\alpha}{F})X$ and its positive part, the difference between the risk and the James-Stein risk was plotted. The graph of the differnce between the risk of the estimator in question and the risk of the Efron and Morris version of the positive part James-Stein estimator (taking $\alpha = \min(\frac{p-0.66}{p-2}\frac{n+2}{n-0.66}, 2)$) was then plotted. We have already presented these in chapter 2. This difference was also used for all the other estimators. The risk for the James-Stein estimator was found to agree with the risk calculated in chapter 2 thus verifying the accuracy of both programs (at least when the shrinkage has zero derivative).

The risk functions for the estimators of chapters 4 and 5 were plotted for values of p and ν given in table 4.

Table 4 Values of p and ν for which the Risk Functions Have Been Computed

p	3	3	3	6	6	6	10	10	10
ν	4	10	20	4	10	20	4	20	40

For each of these values the risk functions of the bilinear shrinkage estimators $\frac{a + bF}{c + dF}X$ for values of a, b, c and d in table 5 were plotted. On the same graph as each bilinear shrinkage estimator were plotted the risk functions for the corresponding iterative versions with α, the weighting factor in the variance estimate, taking values 0, 0.5 and 1.

As it had been found that changing the value of μ, the divisor for the variance estimate, had little effect, we kept $\mu = \nu$ when running the program in batch mode.

A sample of these graphs is shown in figures 20-49 and a key to the plotting symbols for these curves is shown in table 6. For all values of p and ν some of the estimators performed comparably with the Efron and Morris version of the James-Stein estimator, but

only when $p = 3$ were any of the estimators uniformly better. In this case they were significantly better. It is to be noted that only when $p = 3$ that the Efron and Morris estimator differs markedly from the preliminary test level of 50% and this could be an explanation for the improvement. If this is the case then some of the estimators for other values of p and ν in table 1 (section 2.2.1) which improve on the Efron and Morris rule except near the origin, might be capable of being tuned to give improved estimators everywhere.

Table 5 Values of a, b, c and d for Which Shrinkage Estimators of Bilinear Type and Their Iterative Versions were Plotted

a	b	c	d
$-p$	p	0	p
$-\frac{n}{n+2}p$	p	0	p
-1	p	$1-p$	p
-1	p	$\frac{n}{n-2}-p$	p
$-\frac{n}{n+2}$	p	$\frac{n}{n+2}(1-p)$	p
$-p$	p	$1-p$	p
$-p$	p	$\frac{n}{n-2}-p$	p
$-p$	p	$\frac{2}{n-2}p$	p
$-\frac{n}{n+2}p$	p	$\frac{n}{n+2}(1-p)$	p
0	p	1	p
0	p	$1-\frac{1}{p}$	p
0	p	$1-\frac{1}{2p}$	p
0	p	p	p
0	p	$\frac{n-2}{n}-p$	p
-1	p	$-\frac{2}{n+2}$	p
$-\frac{n-2}{n}$	p	$\frac{4}{n(n+2)}$	p
$-\frac{n}{n-2}$	p	0	p

One feature of the graphs is that when the bilinear shrinkage estimator does well its iterative version usually does well also. This does not mean that the iterative versions are better although this is often the case.

It is not at all clear from these plots which estimator is the best to use in practice except when $p = 3$. The Efron and Morris estimator compares favourably with the more complicated estimators of this chapter and may be worth considering for its simplicity. On the other hand, there are estimators of this chapter which sacrifice some of the saving in risk near the origin to save in risk for larger values of λ. It could be the case that some users would prefer these. Examples appear in figures 21,22,23,24,25,26,32,37 and 44. Unfortunately there is no consistency from one pair of values of p and ν to another in the choice of good values for the parameters of the estimator.

Table 6 Key to the Plotting Symbols in Graphs of Figures 20-49

△	maximum likelihood estimator
□	bilinear shrinkage
○	iterated bilinear shrinkage $\alpha = 0.5$
▽	iterated bilinear shrinkage $\alpha = 1$
◇	iterated bilinear shrinkage $\alpha = 0$

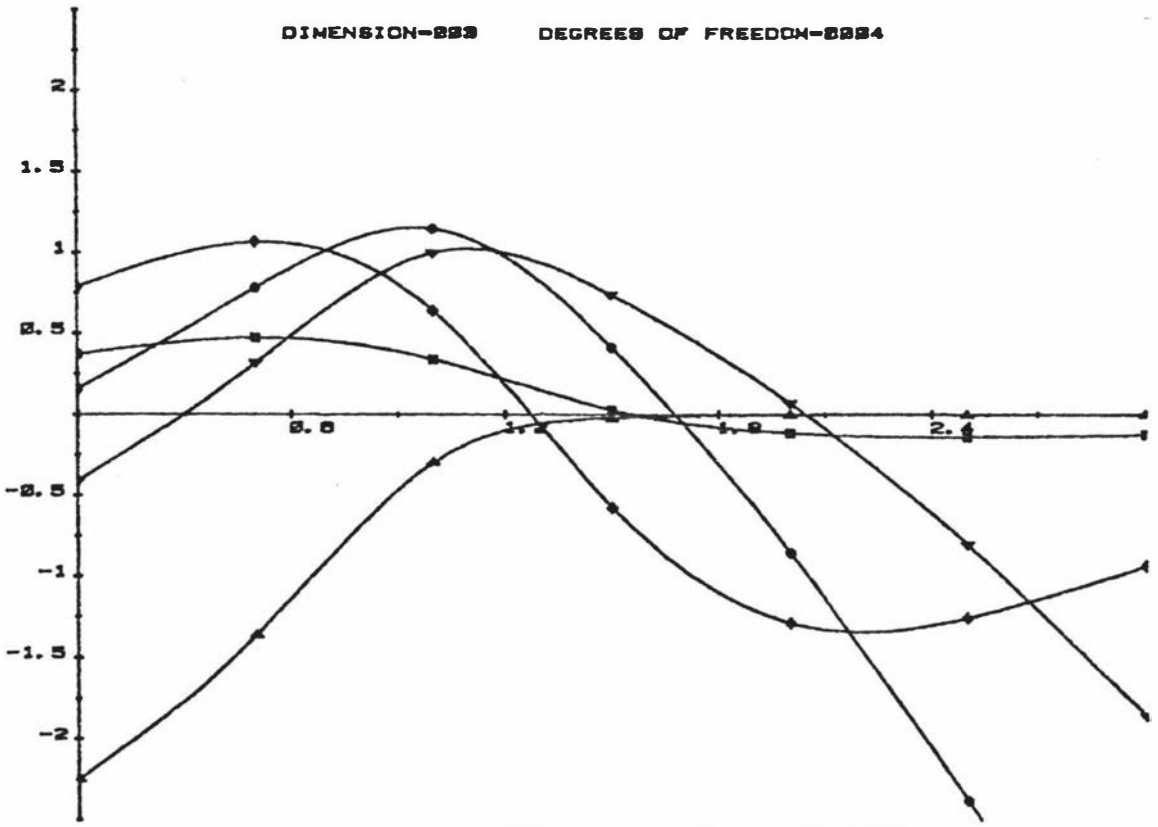

Figure 20 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with a = 0 , b = p , c = p and d = p

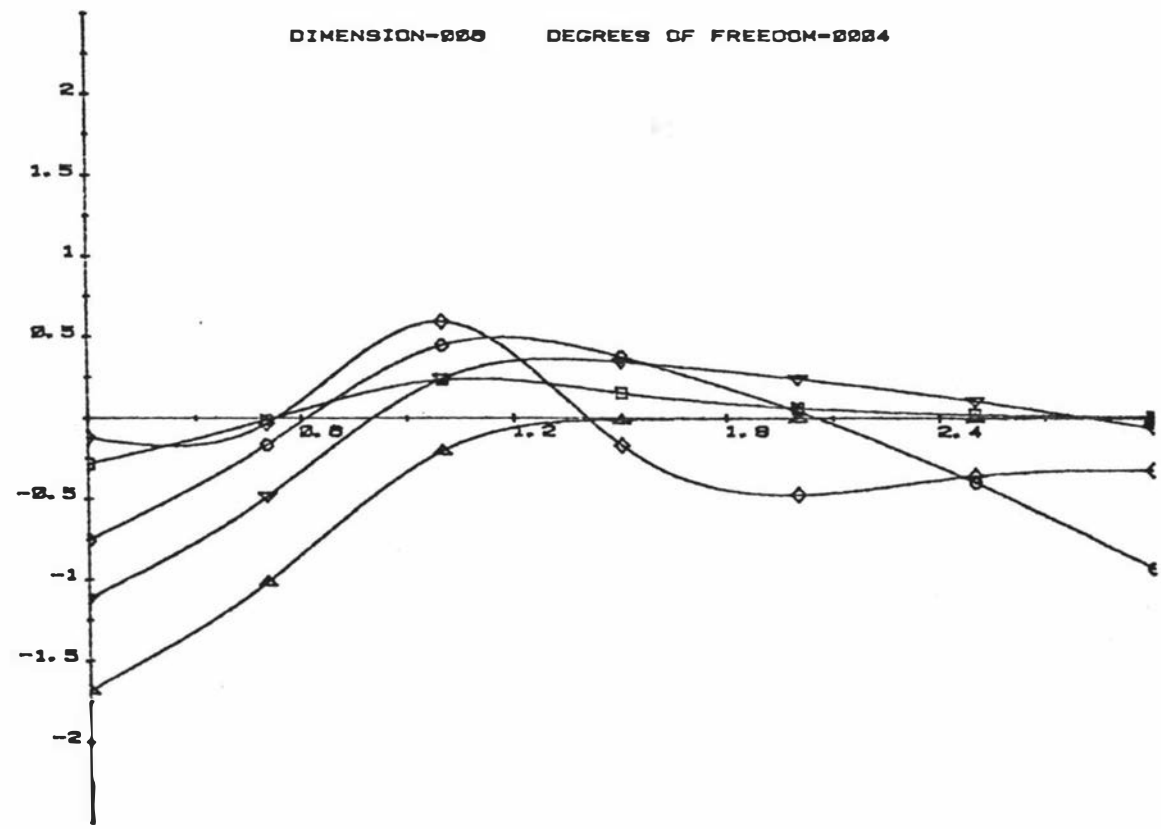

Figure 21 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with a = 0 , b = p , c = p and d = p

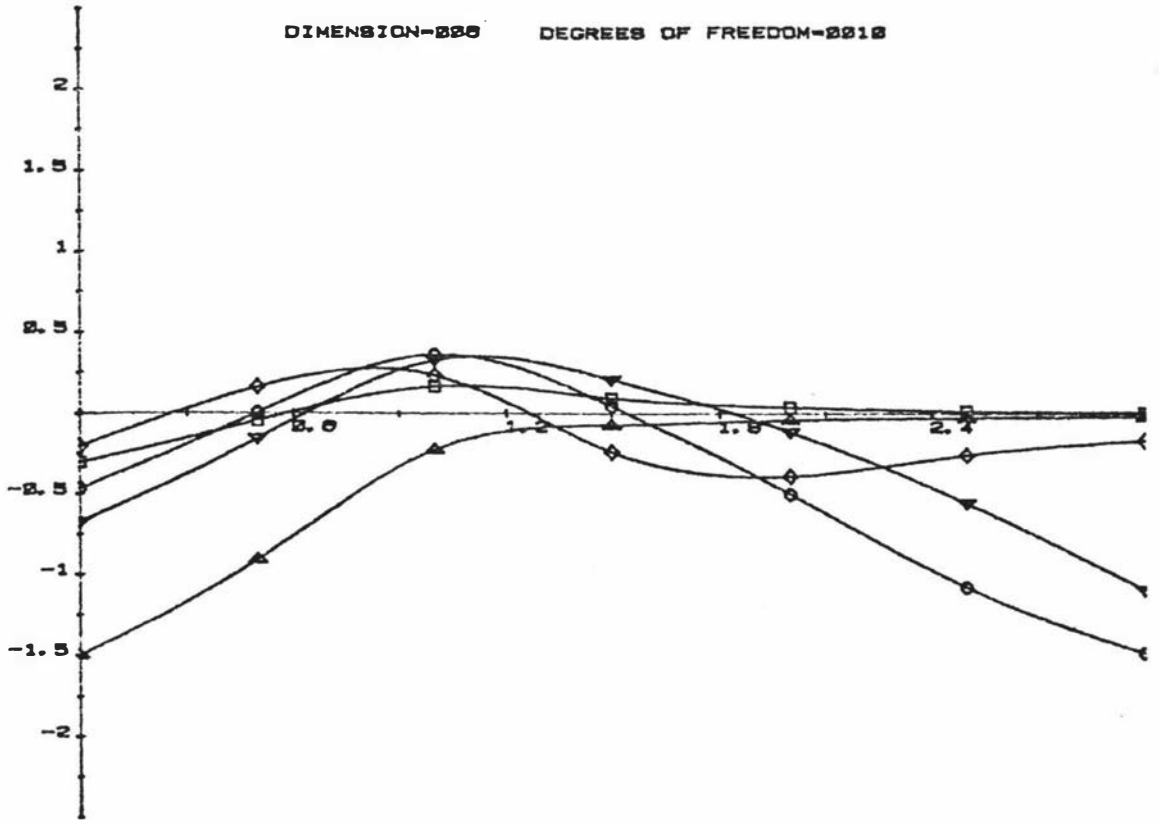

Figure 22 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = p$ *and* $d = p$

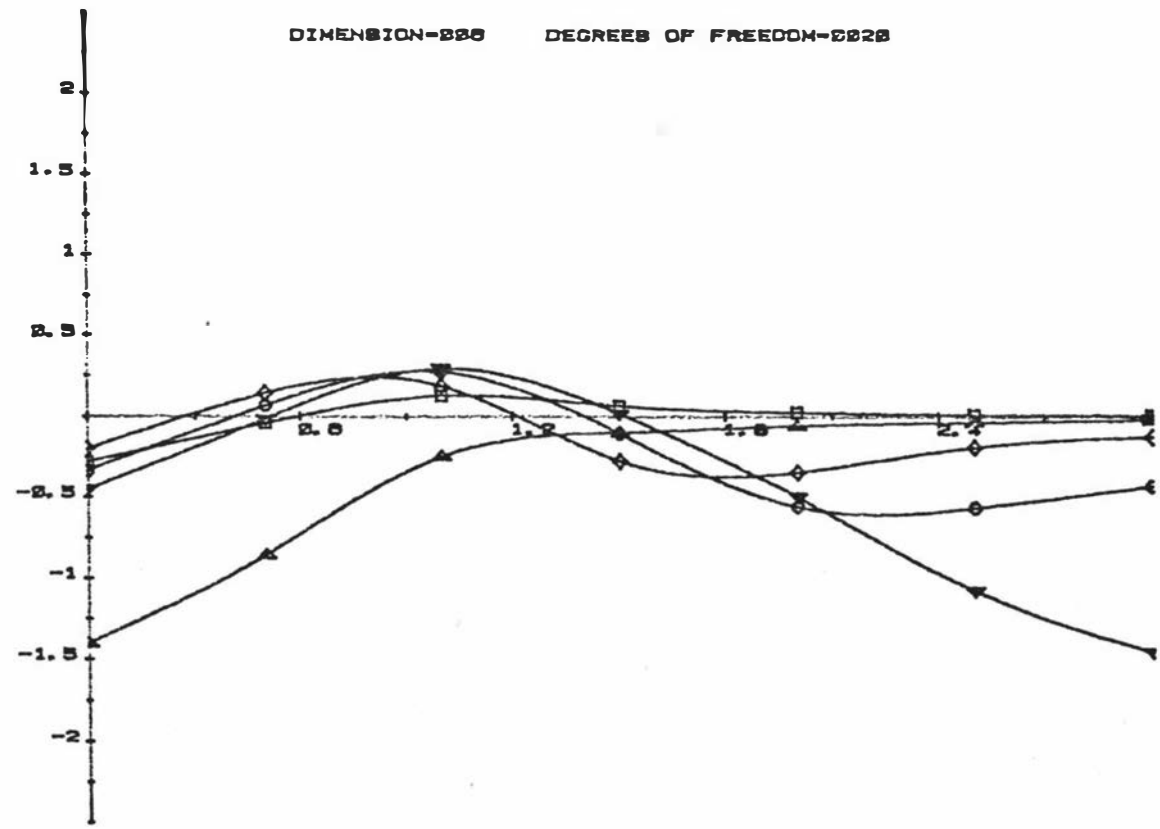

Figure 23 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = p$ *and* $d = p$

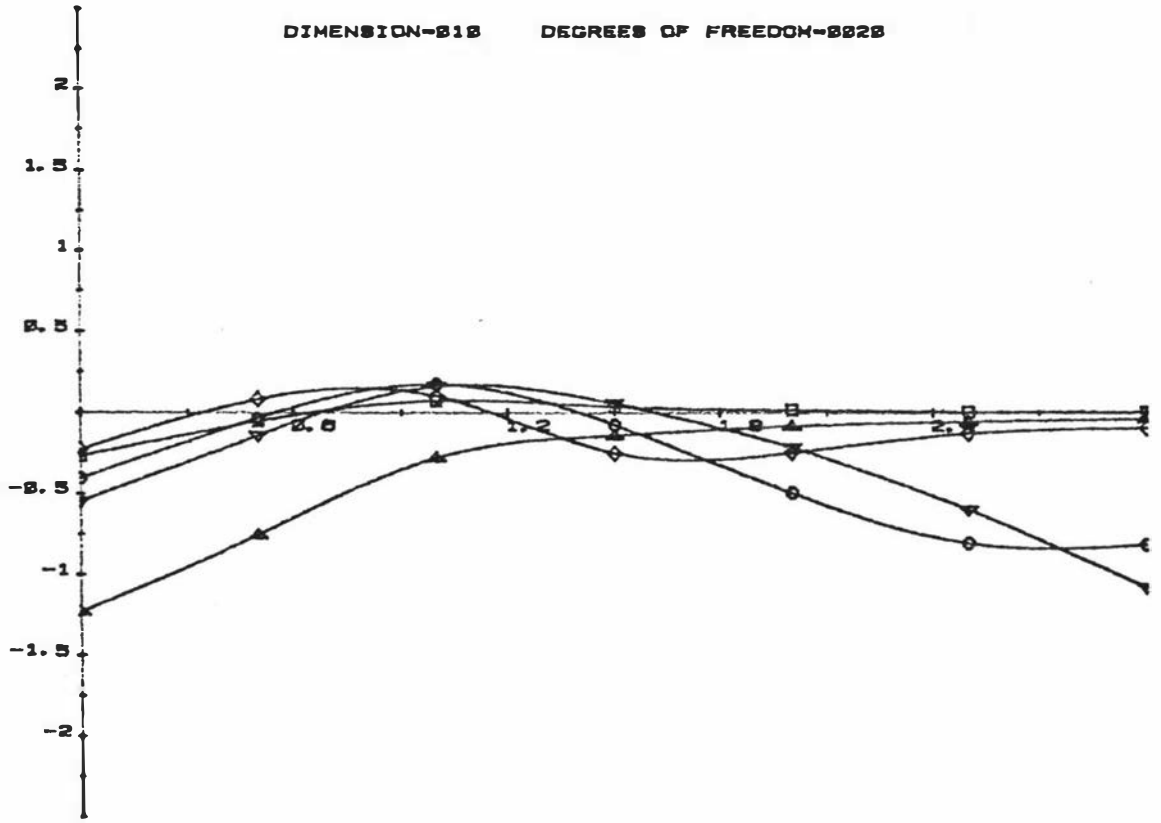

Figure 24 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = p$ *and* $d = p$

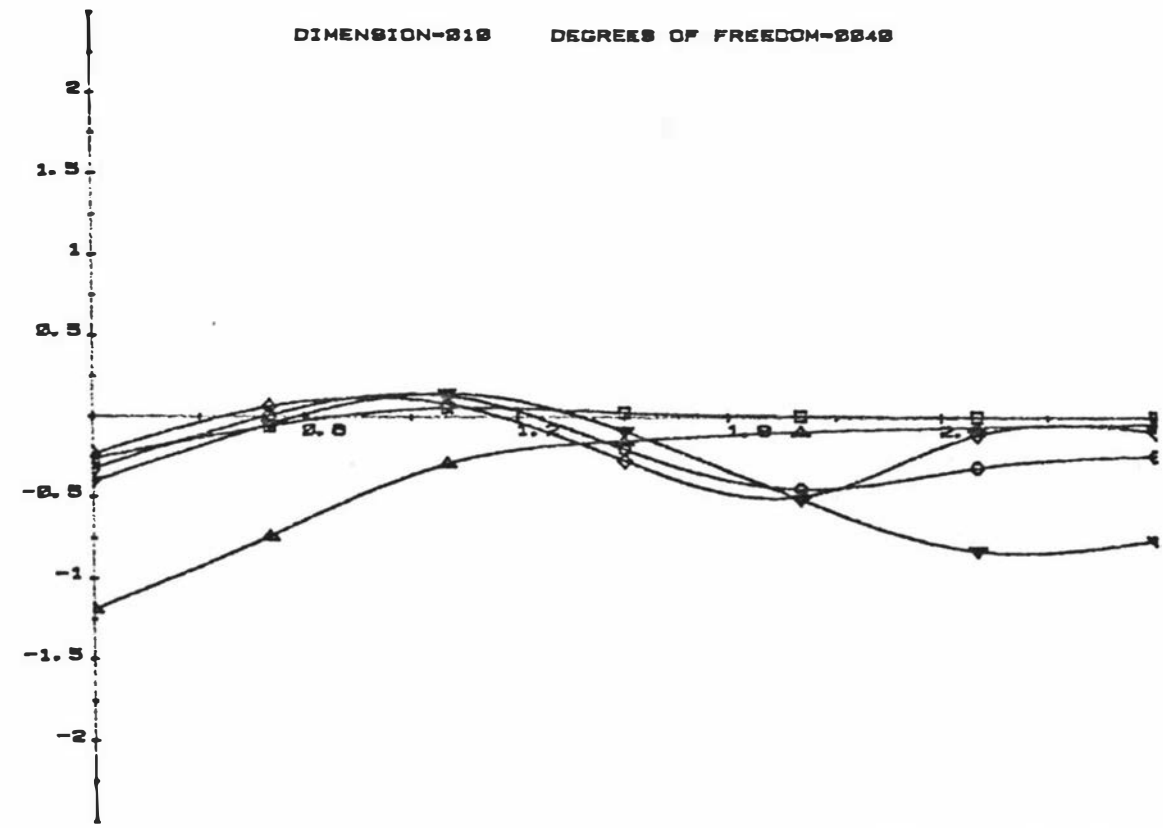

Figure 25 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = p$ *and* $d = p$

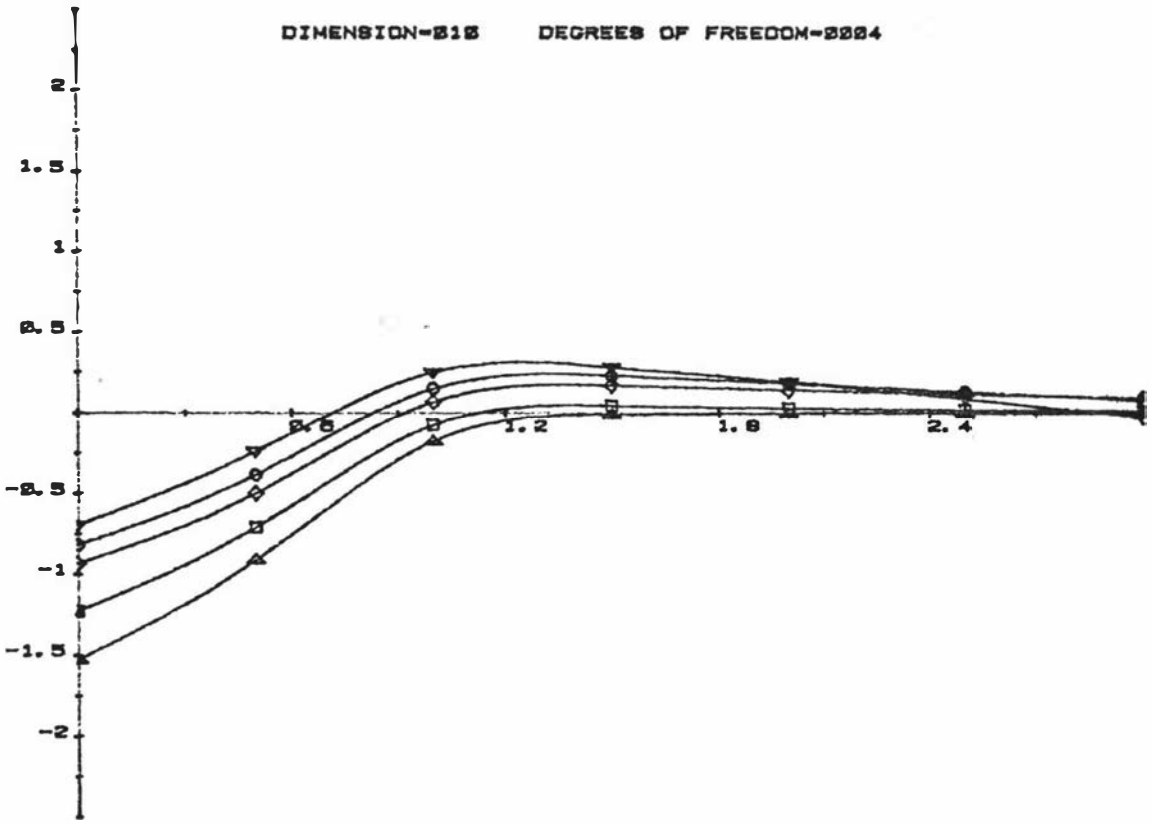

Figure 26 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = 1$ *and* $d = p$

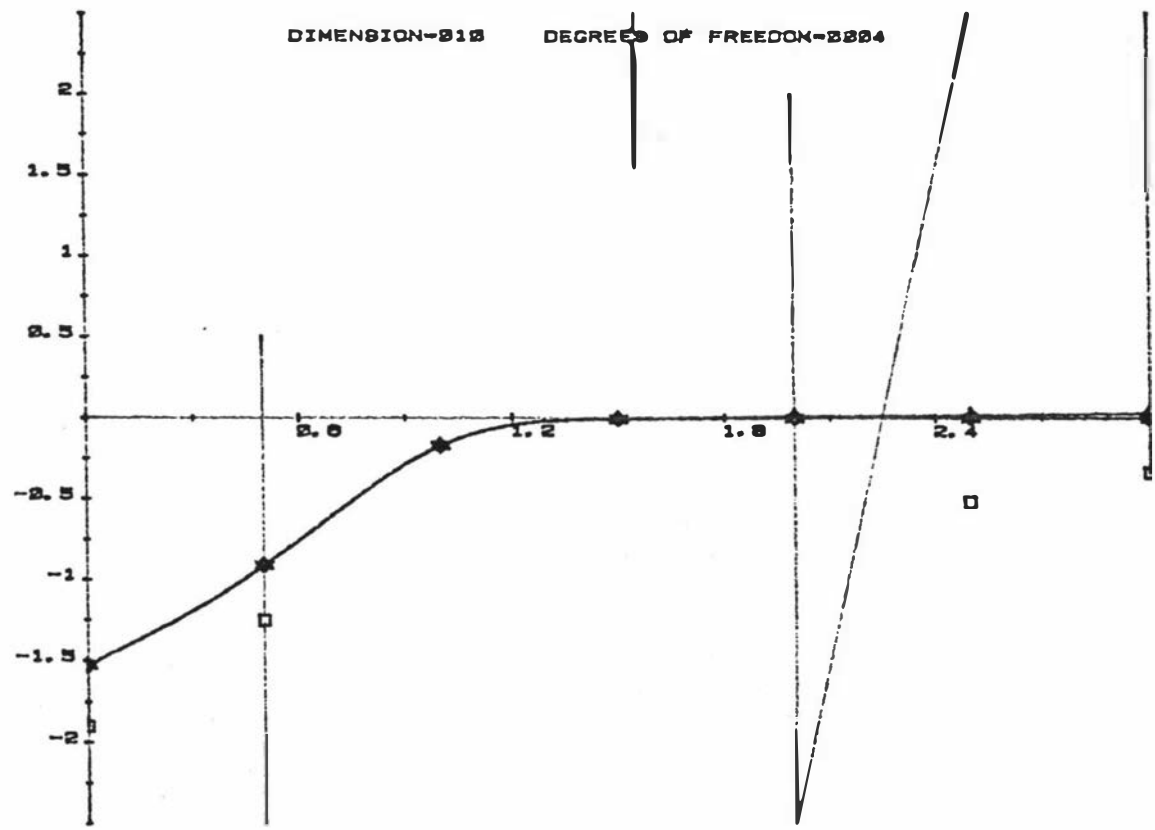

Figure 27 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = \frac{n-2}{n} - p$ *and* $d = p$

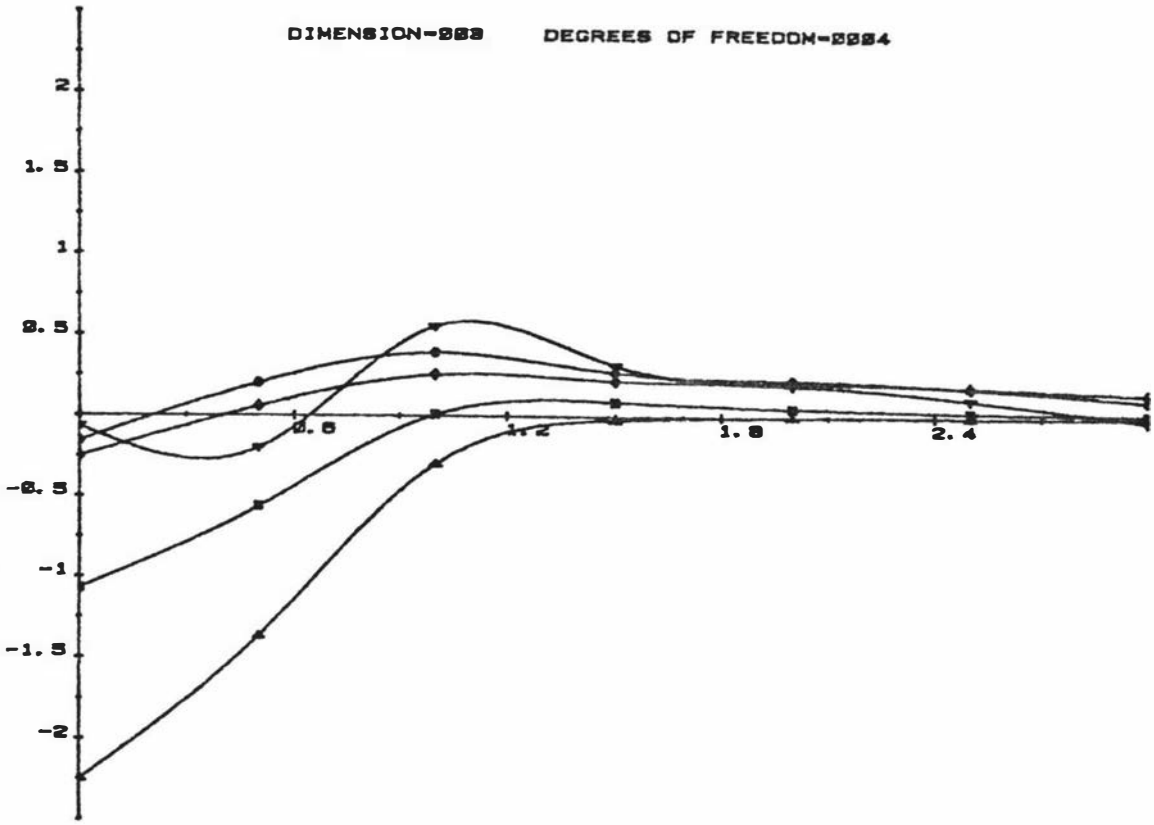

Figure 28 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = 1 - 1/p$ *and* $d = p$

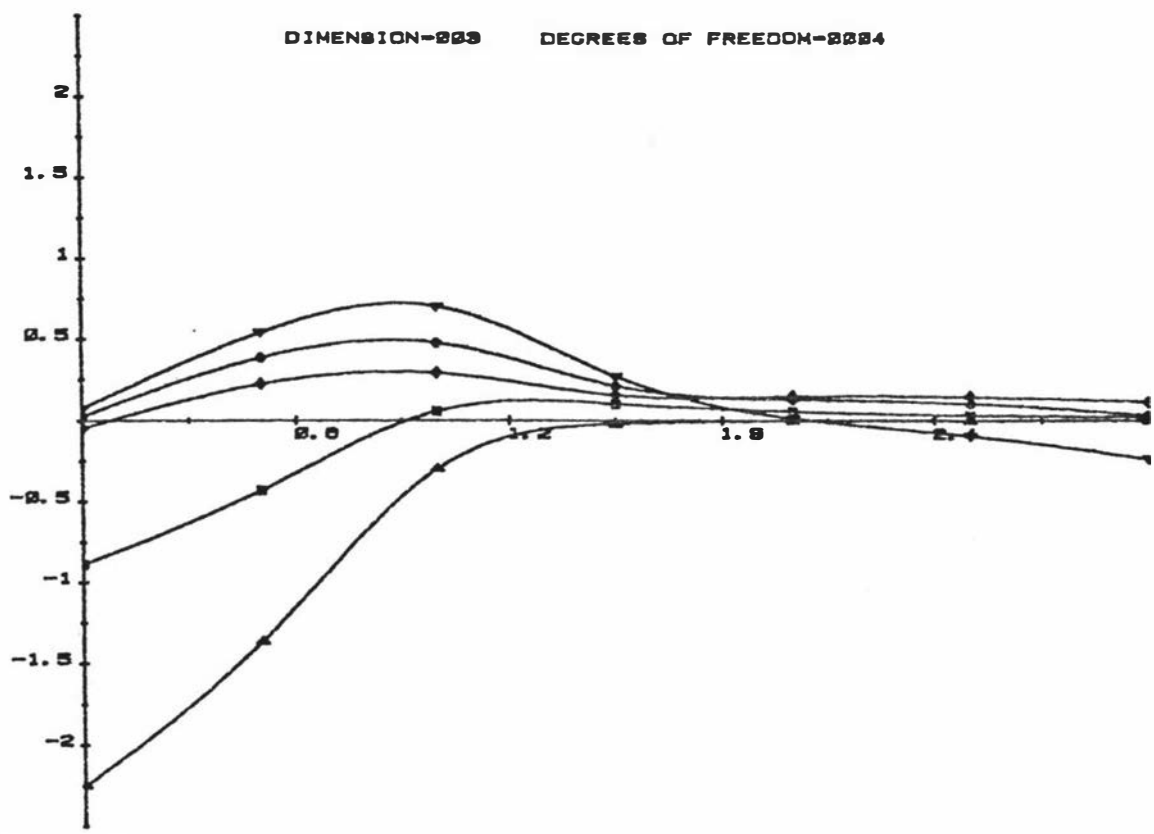

Figure 29 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = 0$, $b = p$, $c = 1 - \frac{1}{2p}$ *and* $d = p$

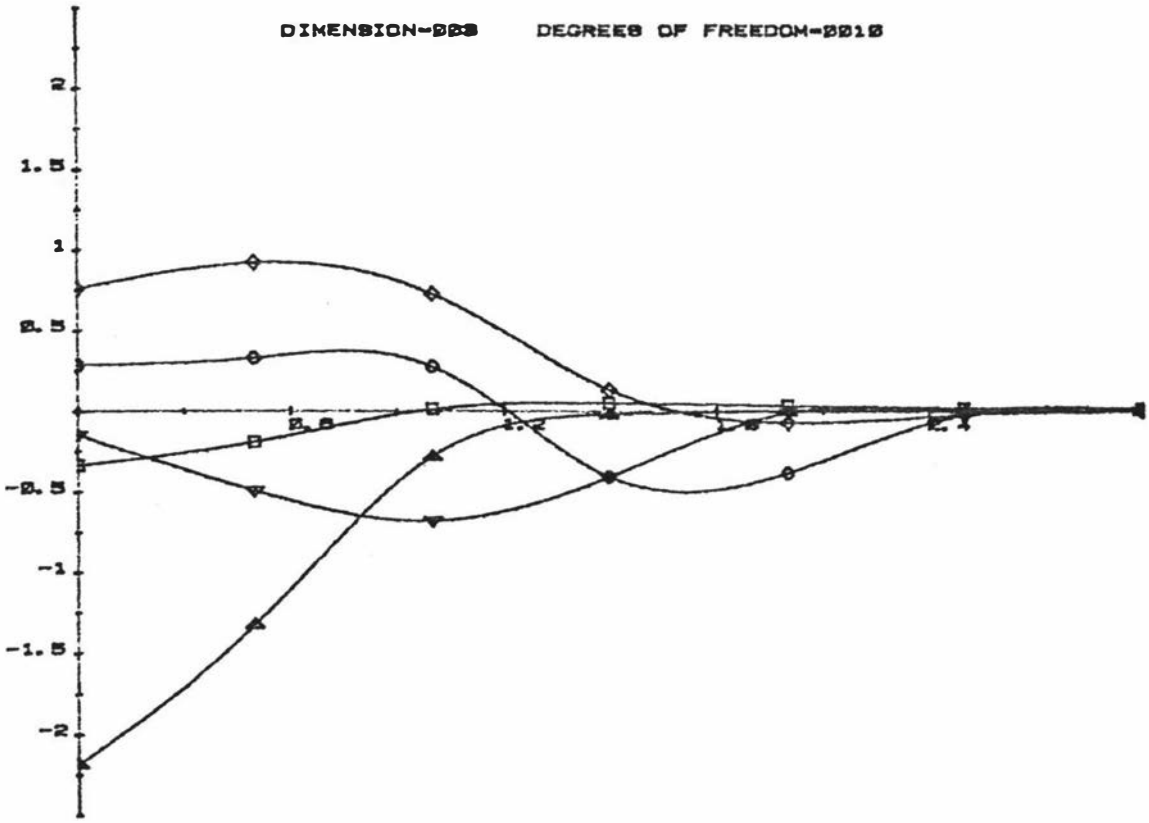

Figure 30 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n-2}{n}$, $b = p$, $c = \frac{4}{n(n+2)}$ and $d = p$

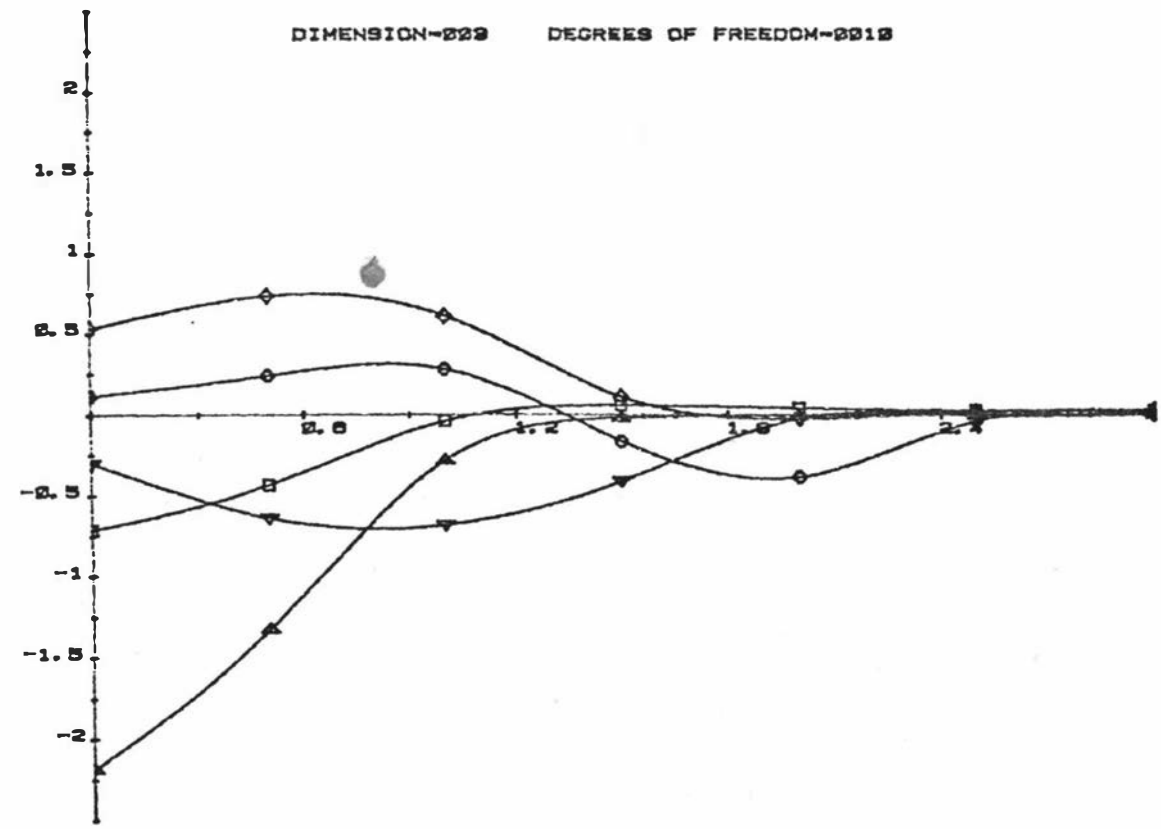

Figure 31 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n}{n-2}$, $b = p$, $c = 0$ and $d = p$

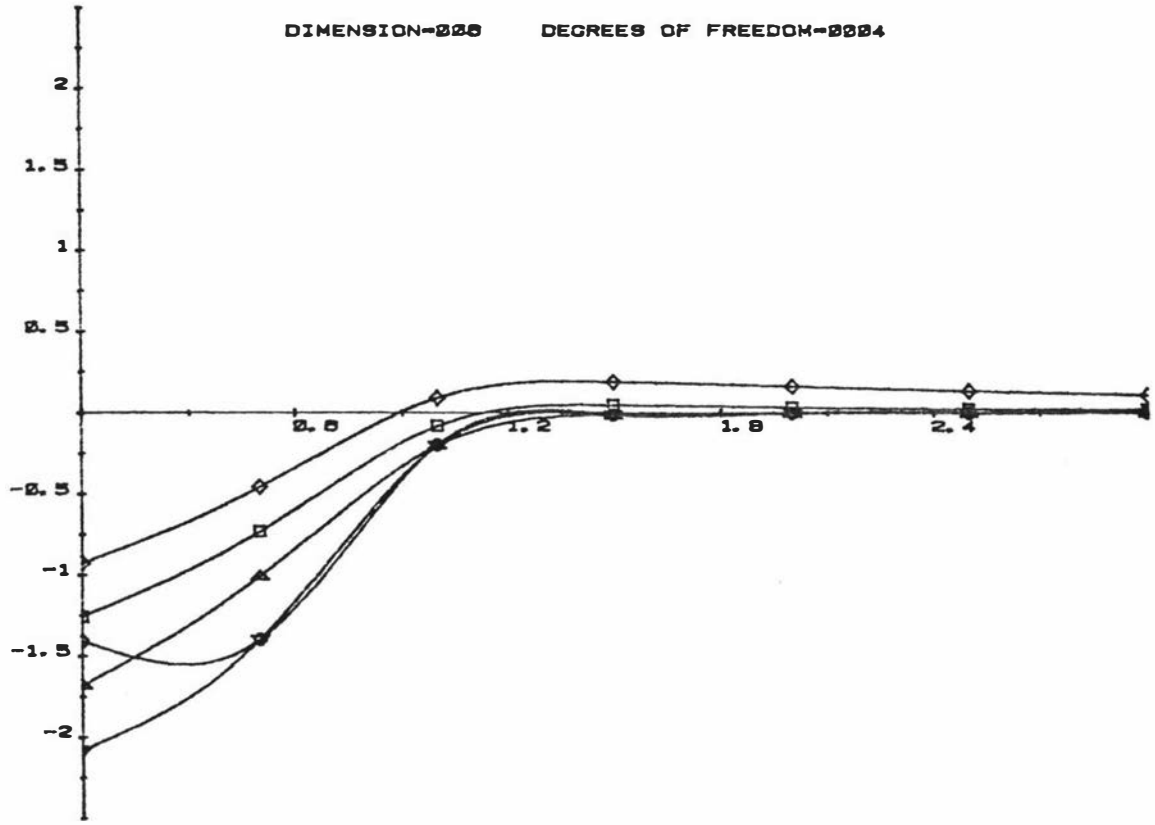

Figure 32 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n-2}{n}$, $b = p$, $c = \frac{4}{n(n+2)}$ *and* $b = p$

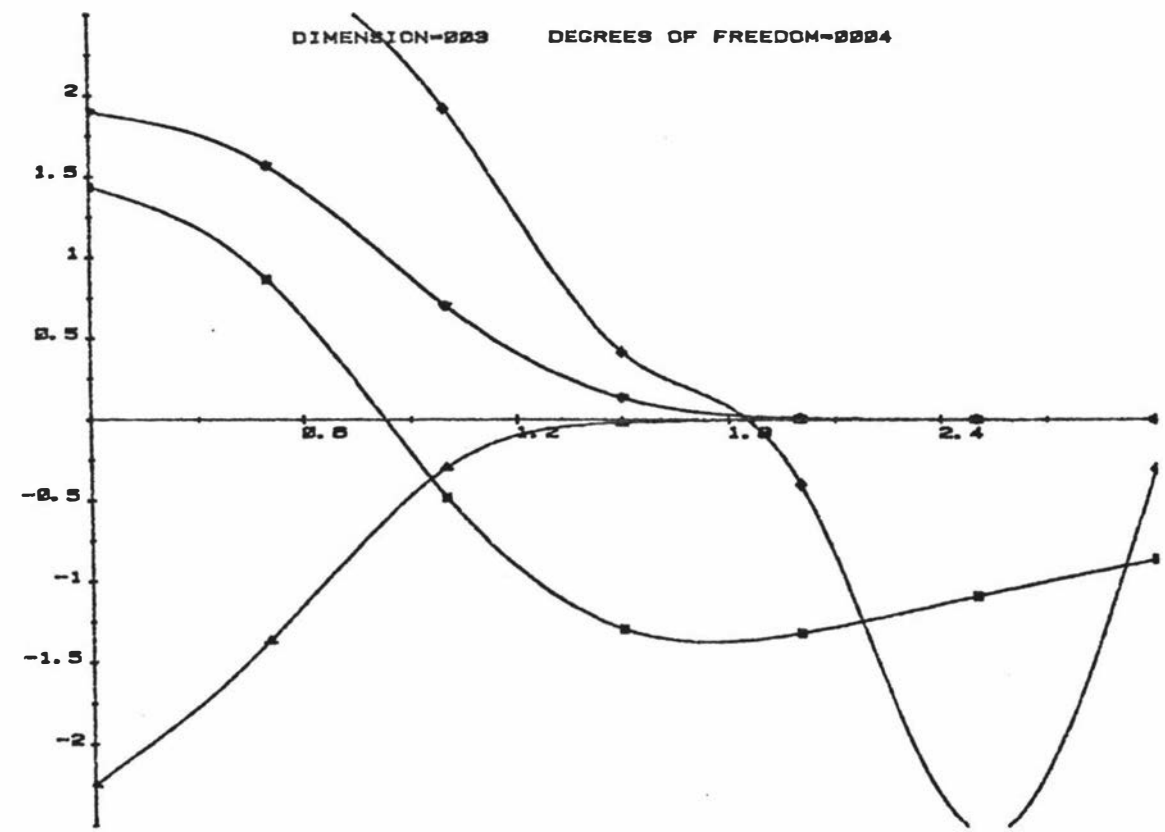

Figure 33 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = p$ *and* $d = p$

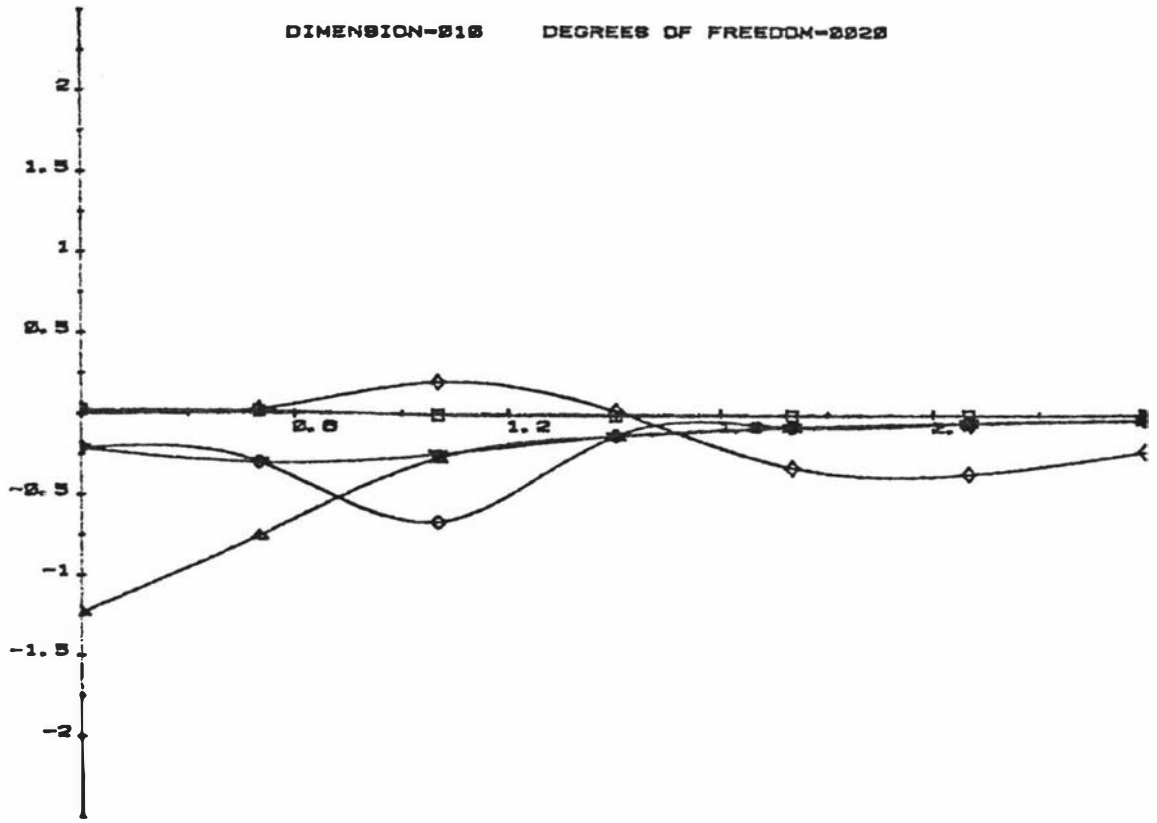

Figure 34 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = \frac{2}{n-2}p$ *and* $d = p$

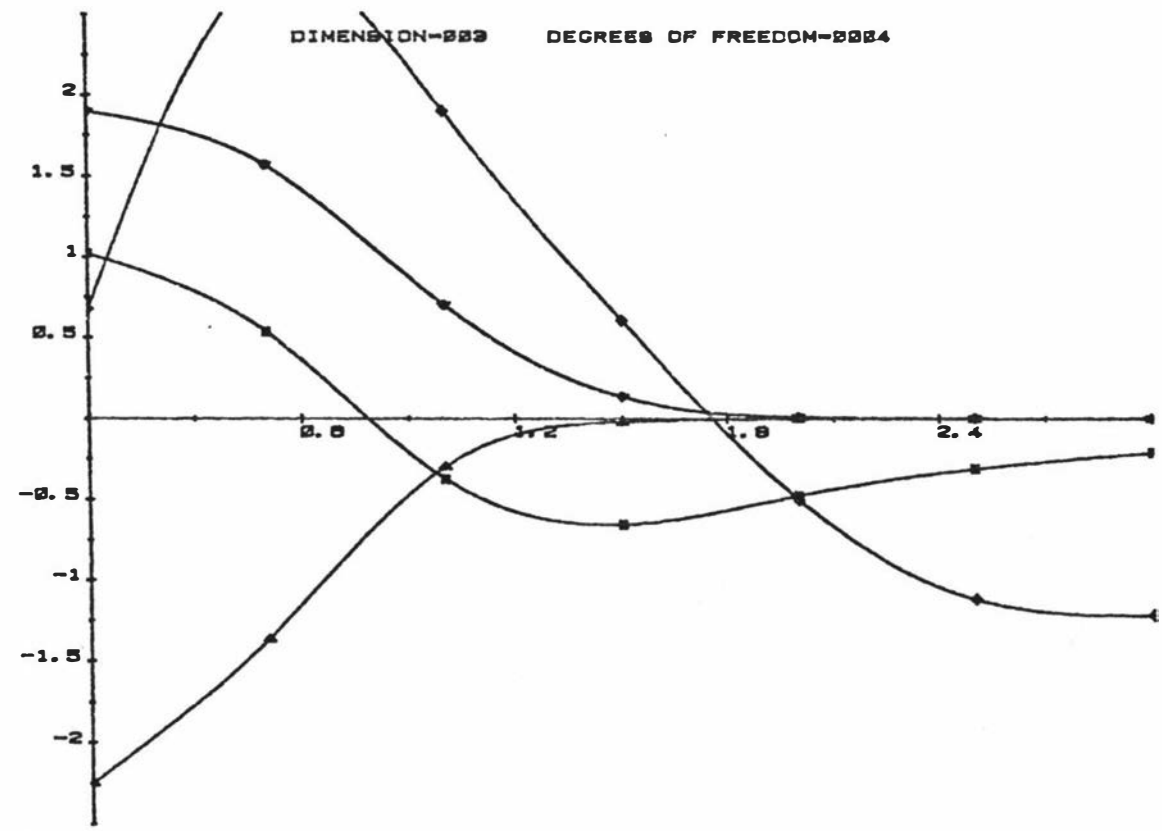

Figure 35 Risk Reduction For Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = 0$ *and* $d = p$

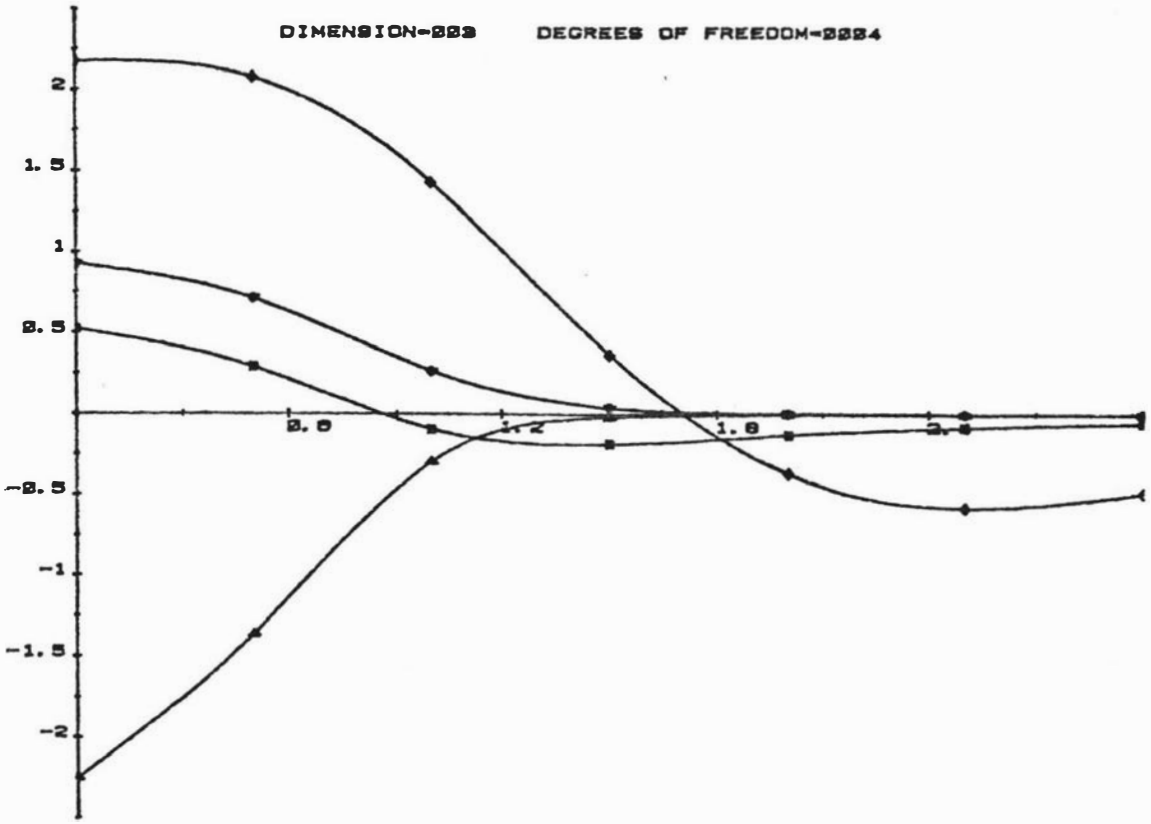

Figure 36 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n}{n+2}\,p$, $b = p$, $c = 0$ and $d = p$

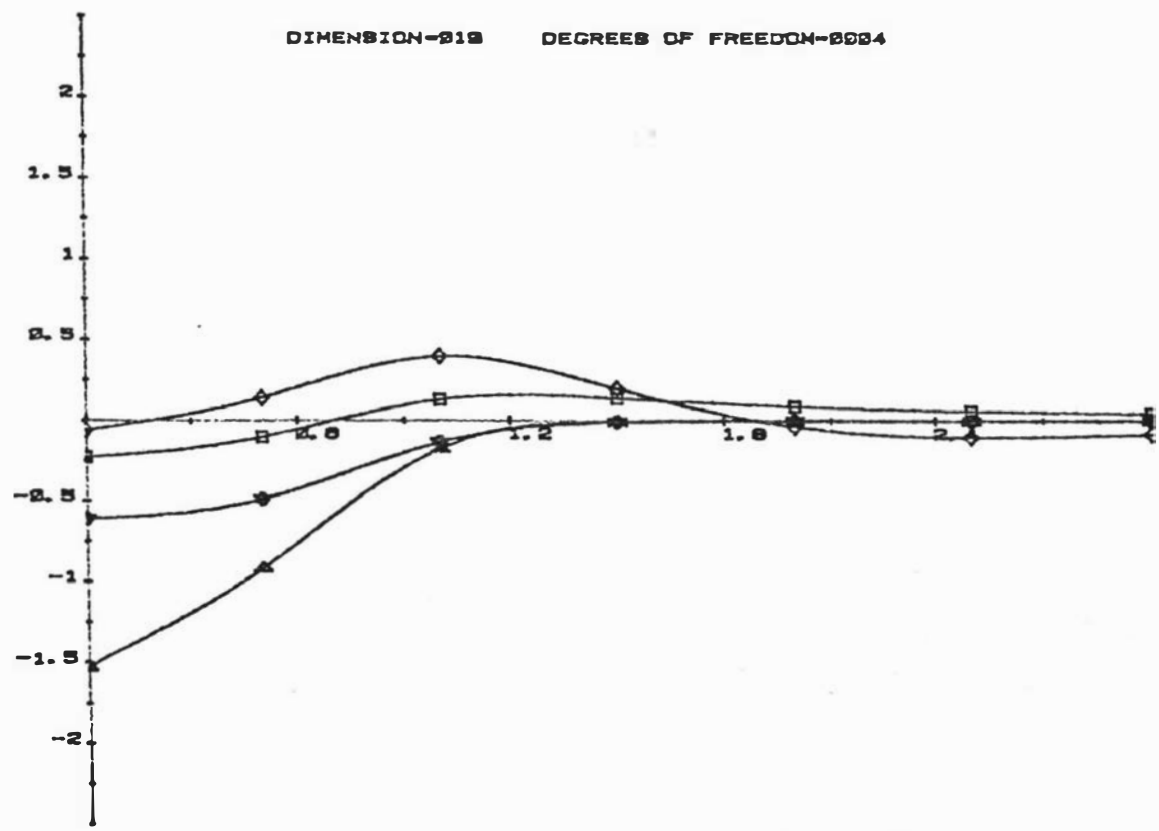

Figure 37 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n}{n+2}\,p$, $b = p$, $c = 0$ and $d = p$

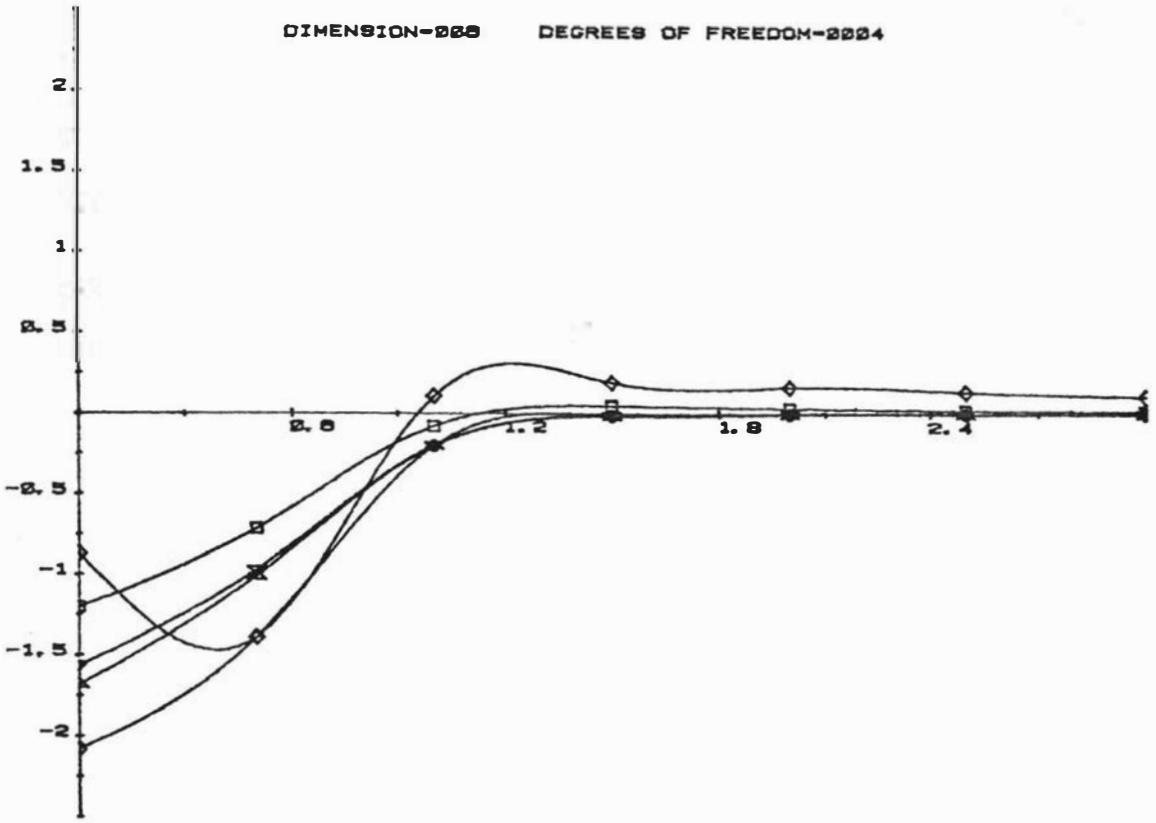

Figure 38 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -1$, $b = p$, $c = -\frac{2}{n+2}$ *and* $d = p$

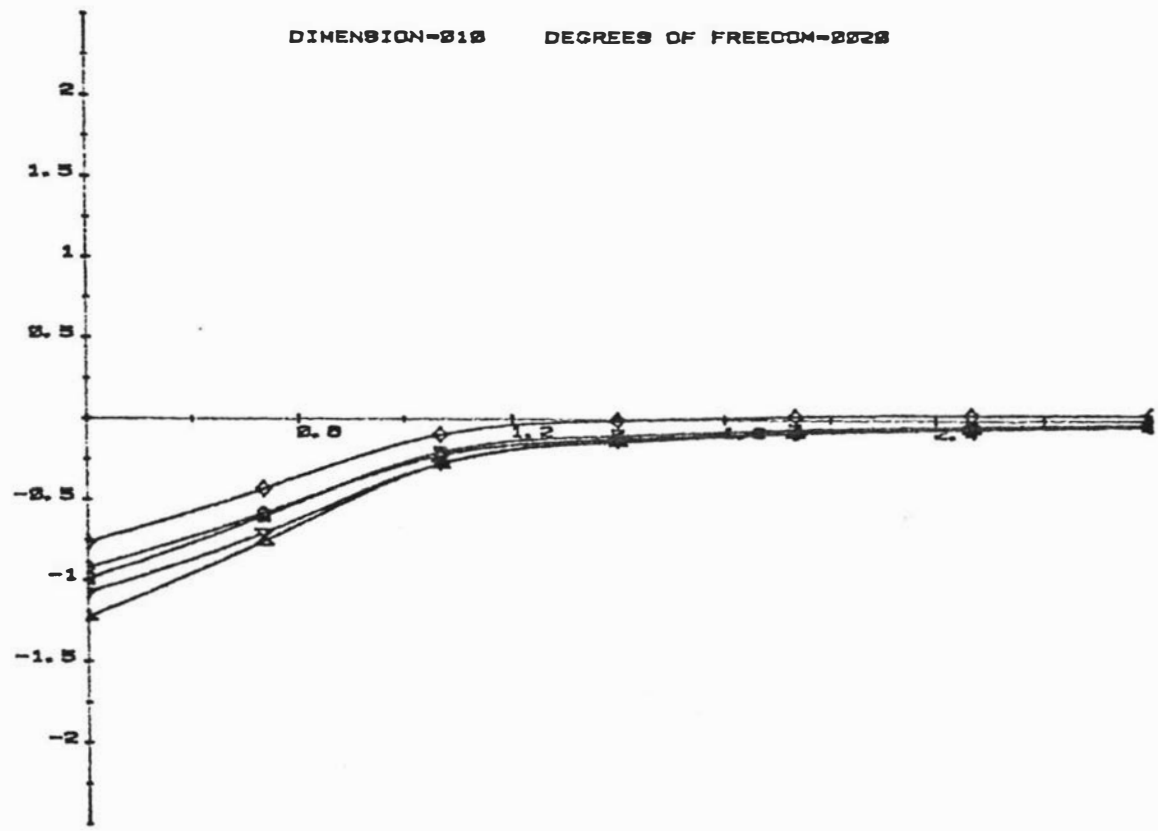

Figure 39 Risk reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -1$, $b = p$, $c = -\frac{2}{n+2}$ *and* $d = p$

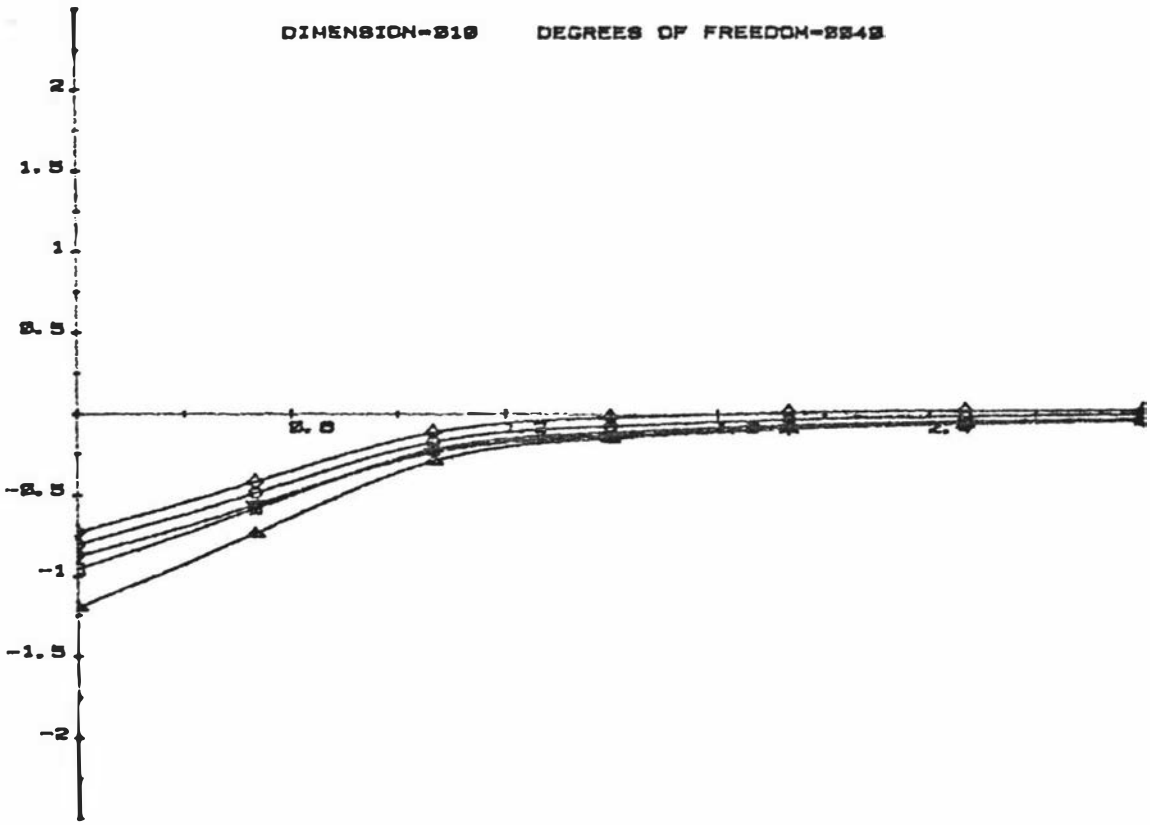

Figure 40 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -1$, $b = p$, $c = -\frac{2}{n+2}$ *and* $d = p$

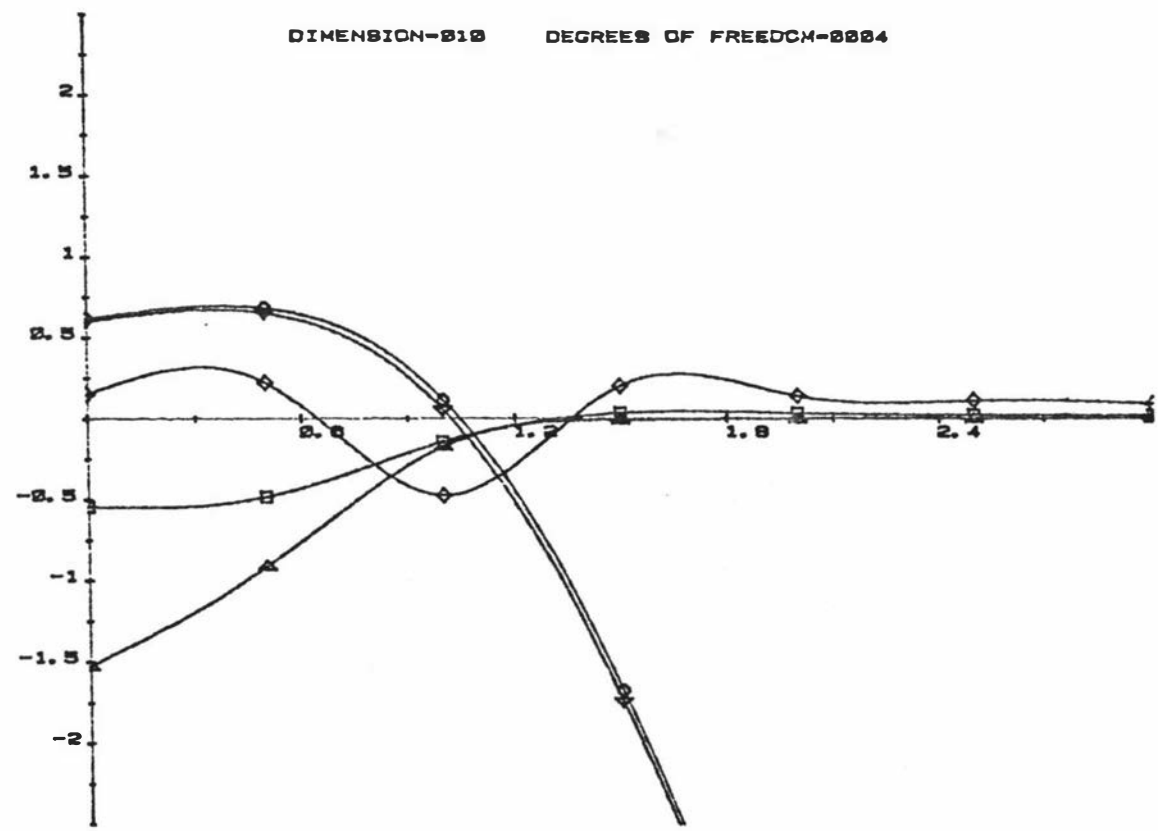

Figure 41 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = 1-p$ *and* $d = p$

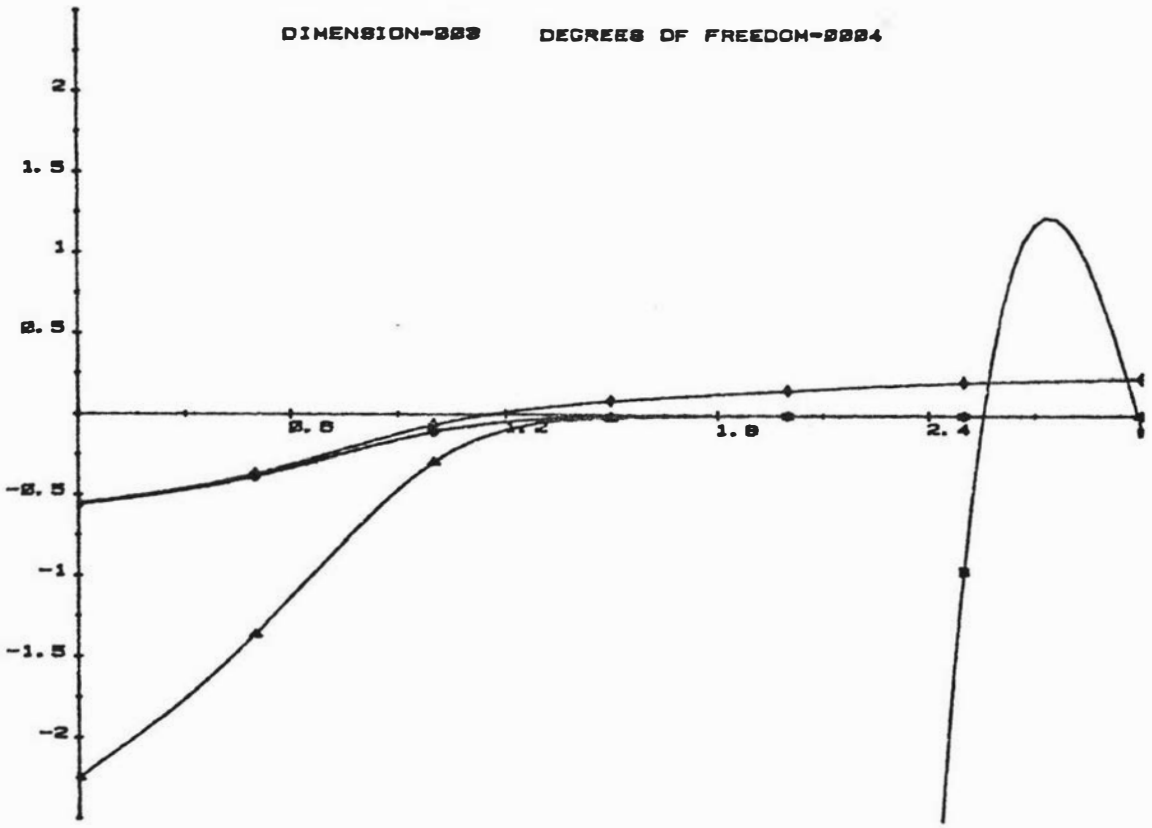

Figure 42 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -1$, $b = p$, $c = 1-p$ *and* $d = p$

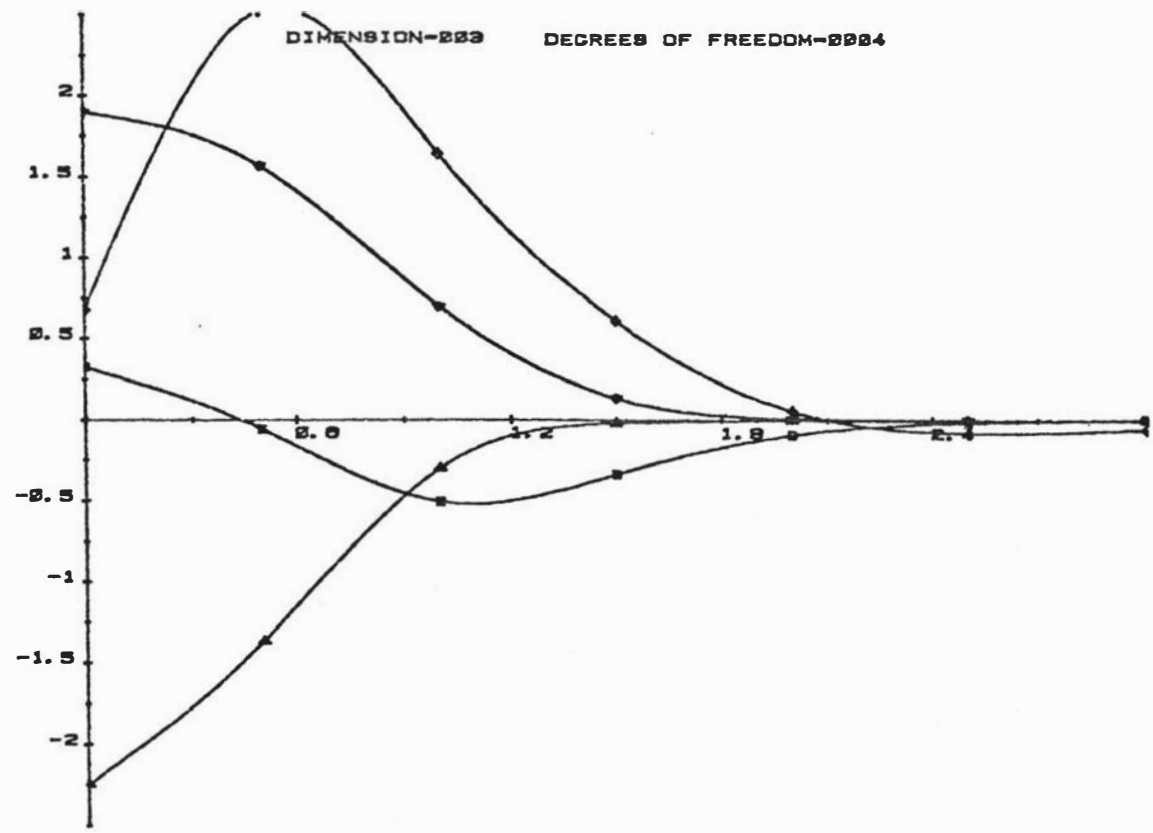

Figure 43 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = 1-p$ *and* $d = p$

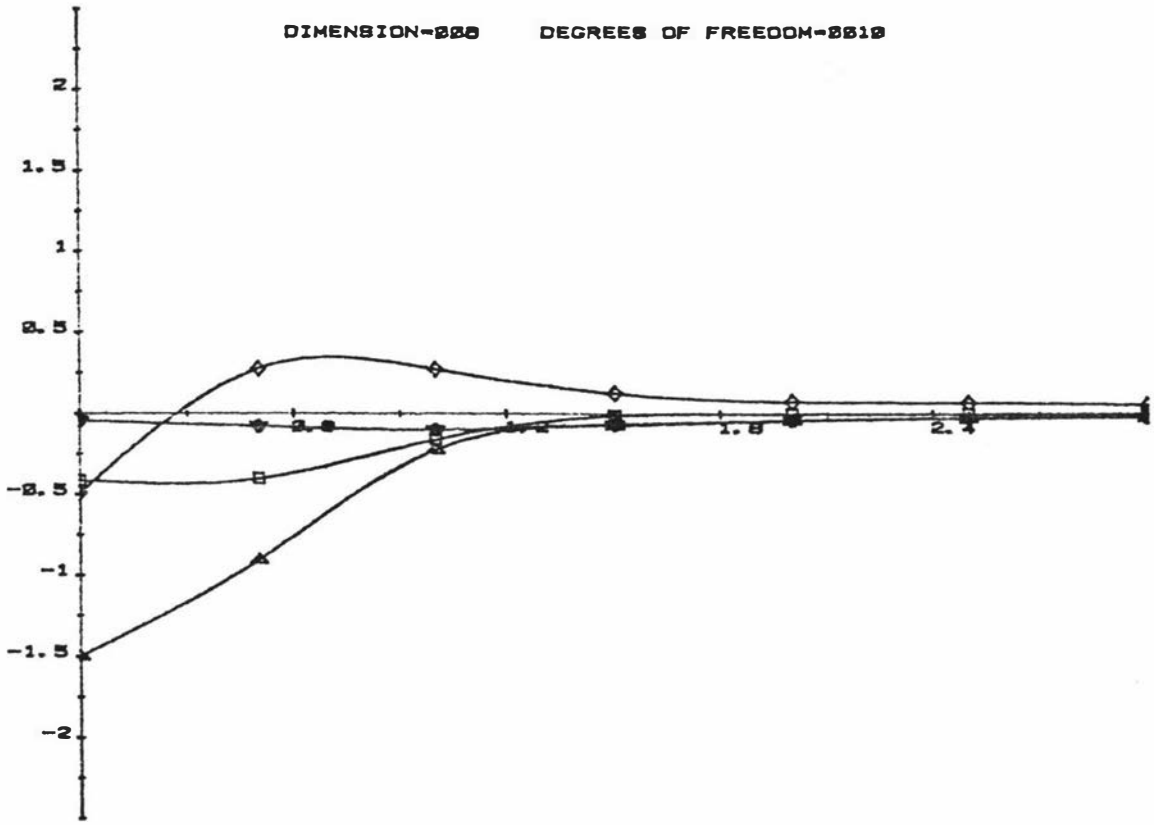

Figure 44 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with a = -p , b = p , c = 1-p and d = p

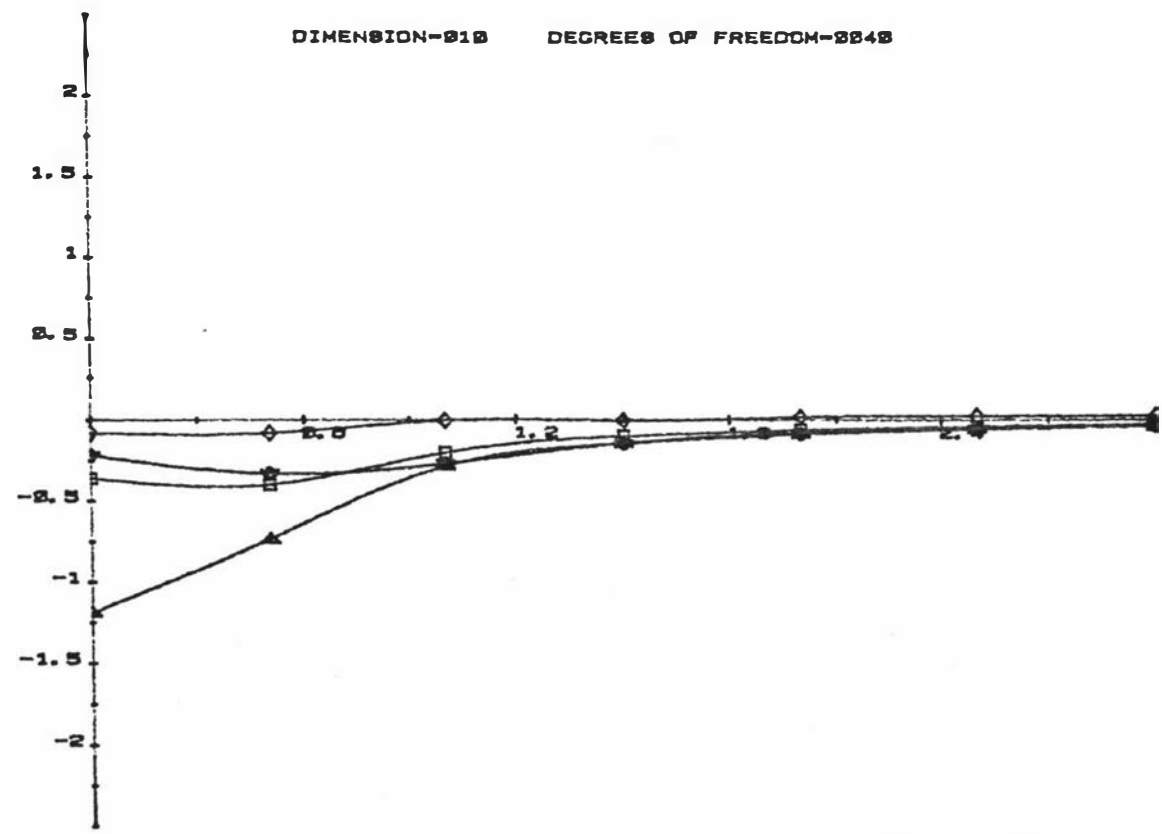

Figure 45 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with a = -p , b = p , c = 1-p and d = p

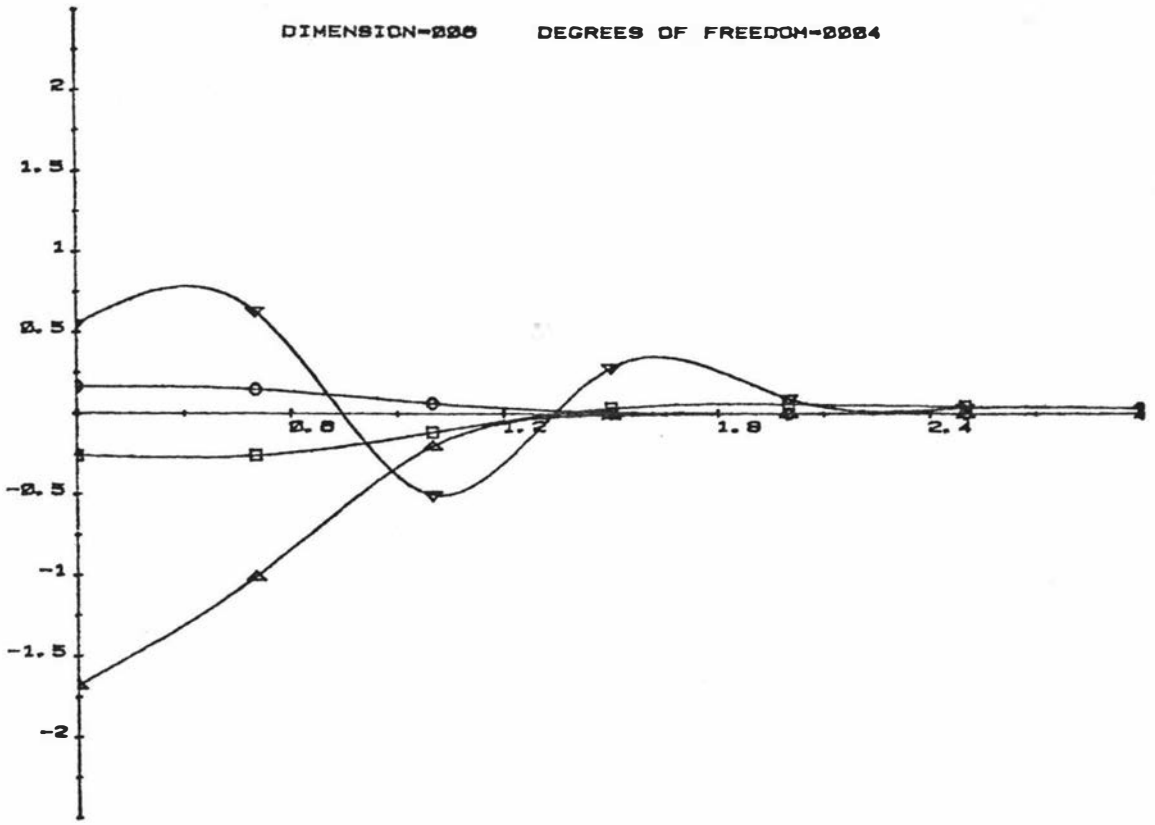

Figure 46 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = 1-p$ *and* $d = p$

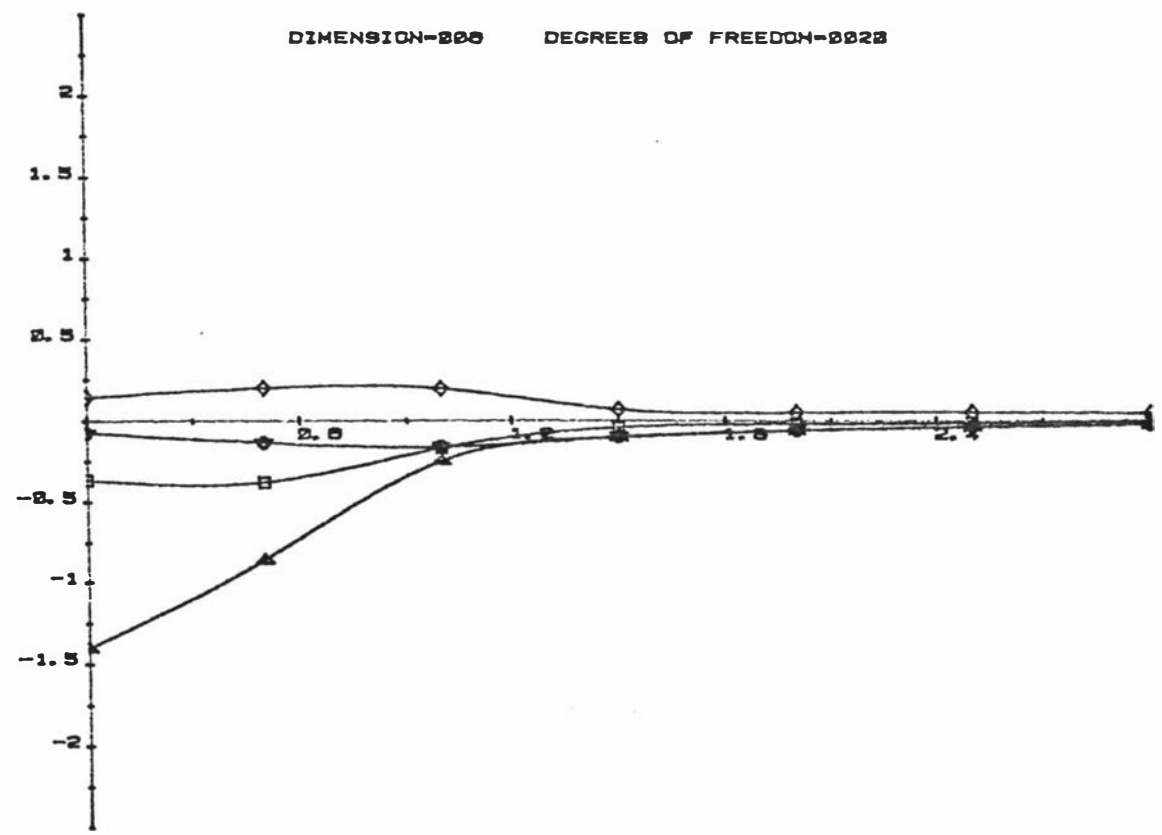

Figure 47 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -p$, $b = p$, $c = 1-p$ *and* $d = p$

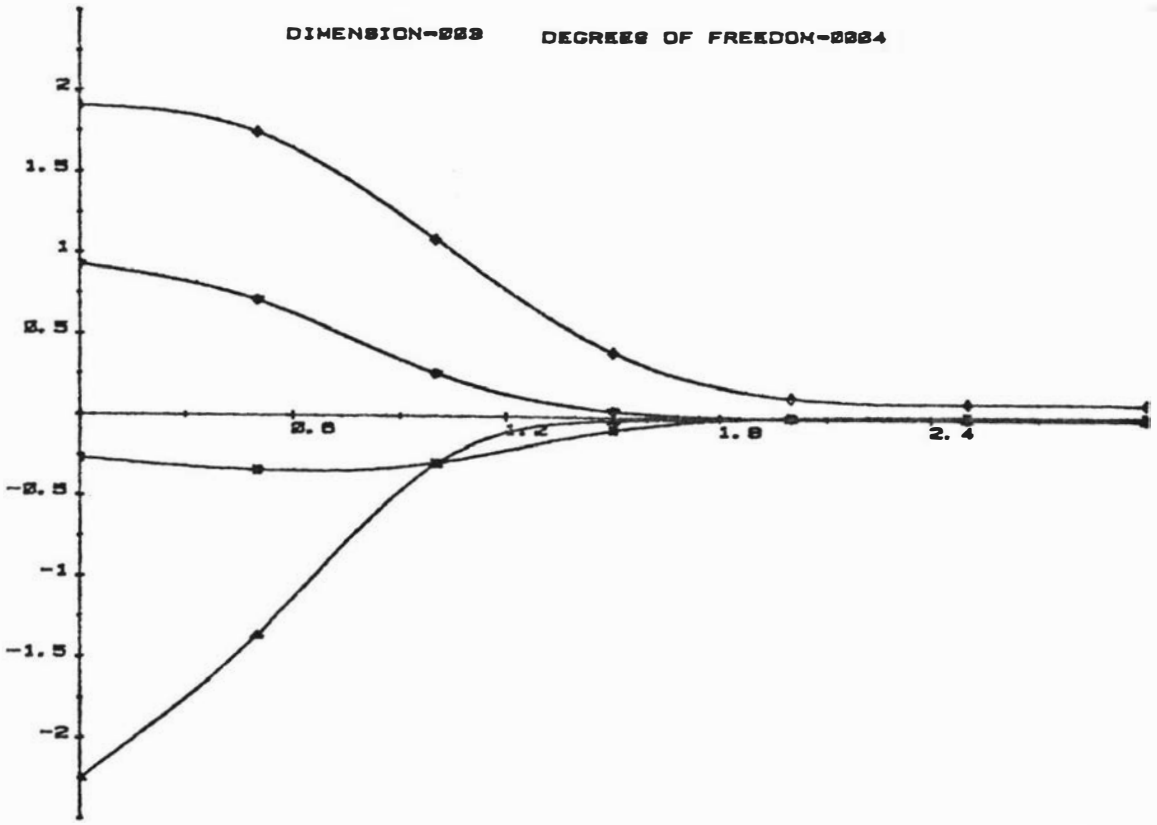

Figure 48 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n}{n+2}p$, $b = p$, $c = \frac{n}{n+2}(1-p)$ *and* $d = p$

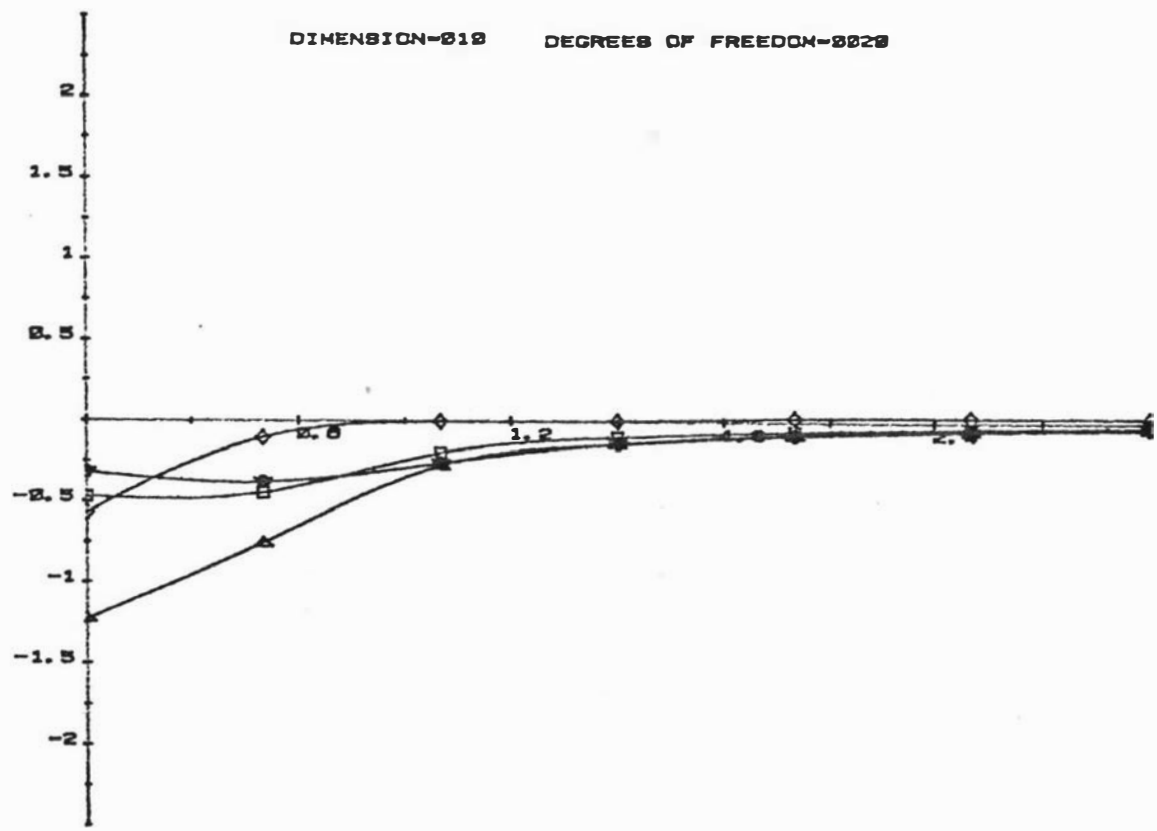

Figure 49 Risk Reduction for Bilinear and Iterative Bilinear Shrinkage Estimators with $a = -\frac{n}{n+2}p$, $b = p$, $c = \frac{n}{n+2}(1-p)$ *and* $d = p$

Chapter 6
Risk Functions for Shrunken Estimators

6.1 Introduction

We shall consider the following problem. Given that $X \sim N_p(\xi,\sigma^2 I)$ and $S \sim \frac{1}{n}\sigma^2 \chi^2_n$ with X and S independent, and given the loss $L(\hat{\xi},\xi,\sigma^2) = \sigma^{-2}\|\hat{\xi} - \xi\|^2$ we wish to calculate the risk function for estimators for ξ of the form $\delta(X,S) = h(W,S)X$ where $W = \frac{1}{p}\|X\|^2$. An estimator of this form will be said to be spherically symmetric. If, in addition, the shrinkage factor, $h(W,S)$, depends only on the ratio $F = \frac{W}{S}$ then the estimator is invariant to transformations of the form $X \to \alpha X$, $S \to \alpha^2 S$; a property which we shall call scale invariance.

It is convenient to define $g(W,S) = 1 - h(W,S)$, $v(W,S) = \frac{1}{\tilde{c}} F\, g(W,S)$ and $w(W,S) = v(W,S) - 1$ where $\tilde{c} = \frac{p-2}{p}\frac{n}{n+2}$. The estimator $\delta(X,S)$ may then be written in any of the forms

$$\delta(X,S) = h(W,S)X = [1 - g(W,S)]X$$
$$= [1 - \tfrac{1}{F}\tilde{c}\, v(W,S)]X = [1 - \tfrac{1}{F}\tilde{c}\{1 + w(W,S)\}]X.$$

Note We have defined $S \sim \frac{1}{n}\sigma^2\chi^2_n$ rather than $S \sim \sigma^2\chi^2_n$ because this allows the case of unknown variance to be conveniently treated as a special case by putting $S = \sigma^2$ and formally writing $n = \infty$. In order that the notation be consistent we have defined $W = \frac{1}{p}\|X\|^2$ so that $W \sim \frac{1}{p}\chi^2_n(\lambda)$ where $\lambda \sim \tfrac{1}{2}\sigma^{-2}\|\xi\|^2$.

In the next section we collect together some lemmas which will be useful in deriving some of the many expressions for the risk function.

6.2 Some Identities Involving Expectations

The first identity is a well known result given, for example, in Rao(1973) and in Lindley(1965).

Lemma 1 If X is a scalar random variable for which $E[X]$ exists then $E[X] = a - \int_{-\infty}^{a} F(x)\,dx + \int_{a}^{\infty}(1 - F(x))\,dx$.

where $F(x)$ is the distribution function of X and a is a constant.

Proof Integrating by parts we see that

$$\int_a^T x\,dF(x) = [-x(1 - F(x))]_a^T + \int_a^T (1 - F(x))dx$$

and $$\int_{-U}^{a} x\,dF(x) = [\,x\,F(x)\,]_{-U}^{a} - \int_{-U}^{a} F(x)\,dx$$

so that $$\int_{-U}^{T} x\, dF(x) = a - \int_{-U}^{a} F(x)\, dx + \int_{a}^{T} (1 - F(X))\, dx + U\,F(U) - T(1 - F(T)).$$

Now $$T(1 - F(T)) = T \int_{T}^{\infty} dF(x) \leq \int_{T}^{\infty} x\, dF(x)$$

and this tends to zero as $T \to \infty$ since $E[X]$ exists.

Similarly $$-U\,F(U) = -U \int_{-\infty}^{-U} dF(x) \leq \int_{-\infty}^{-U} -x\, dF(x)$$

tends to zero as $U \to \infty$.

The next lemmas are of similar form to this but relate to particular distributions. At the same time we can calculate the expectations of more complicated functions than X.

Lemma 2 If $X \sim N(\xi,\sigma^2)$ and $h(.)$ is absolutely continuous then $E[X\,h(X)] = \xi\,E[h(X)] + E[\frac{d}{dX} h(X)]$ so long as $h'(X) = \frac{d}{dX} h(X)$ exists almost everywhere and both of $E[X]h(X)$ and $E[h'(X)]$ are finite.

Proof The density of X is $p(x) = A \exp\{-\frac{1}{2}\sigma^{-2}(x-\xi)^2\}$ so that $\frac{d}{dx} p(x) = -\sigma^{-2}(x-\xi)\,p(x)$ and thus $-\sigma^2\,p(x) = \int (x-\xi)\,p(x)\,dx$.

Integrating by parts we obtain

$$\int_{-U}^{T} (x-\xi)\,p(x)\,h(x)\,dx = [-\sigma^2\,p(x)\,h(x)]_{-U}^{T} + \int_{-U}^{T} p(x)\,h'(x)\,dx.$$

Also $$E[(X-\xi)\,h(X)] = \lim_{U,T\to\infty} \int_{-U}^{T} (x-\xi)\,p(x)\,h(x)\,dx$$

and $$E[h'(X)] = \lim_{U,T\to\infty} \int_{-U}^{T} p(x)\,h'(x)\,dx.$$

We therefore need to prove that $\lim_{x\to\pm\infty} p(x)\,h(x) = 0$.

Now $\int_{\xi}^{\infty} (x-\xi)\,h(x)\,p(x)\,dx$ and $\int_{\xi}^{\infty} h'(x)\,p(x)\,dx$

are both finite so that $\lim_{x\to\infty} p(x)\,h(x)$ is finite. If this limit is non-zero then we may, without loss of generality, suppose that $\forall\, x > U$, $p(x)\,h(x) > l > 0$. We then have

$$\int_{U}^{T} (x-\xi)\,h(x)\,p(x)\,dx \geq (U-\xi) \int_{U}^{T} h(x)\,p(x)\,dx \geq (U-\xi)(T-U)l.$$

Since the left hand side is bounded as $T \to \infty$ it is impossible that $\ell > 0$ since otherwise the right hand side is unbounded.

Similarly, we may show that $p(x)\, h(x) \to 0$ as $x \to -\infty$.

Lemma 3 If $X \sim N(\xi,\sigma^2)$ and $h(.)$ is absolutely continuous then, so long as $E[X\, h(X)]$ is finite,

$$E[X\, h(X)] = \xi\, E[h(X)] + \sigma^2 \frac{\partial}{\partial \xi} E[h(X)].$$

Proof Differentiating the density function $p(x)$ in the proof of lemma 2 partially with respect ξ we obtain $\frac{\partial}{\partial \xi} p(x) = \frac{x-\xi}{\sigma^2} p(x)$.

Therefore
$$E[(X-\xi)\, h(X)] = \int_{-\infty}^{\infty} \sigma^2 \{\frac{\partial}{\partial \xi} p(x)\}\, h(x)\, dx$$
$$= \sigma^2 \frac{\partial}{\partial \xi} \int_{-\infty}^{\infty} p(x)\, h(x)\, dx$$
$$= \sigma^2 \frac{\partial}{\partial \xi} E[h(X)].$$

The next lemma appears in Efron and Morris(1976). The proof is given here as Efron and Morris do not prove convergence. This lemma is also a consequence of (4) below, but that proof assumes that $h(.)$ is continuous at the origin, whereas this proof does not.

Lemma 4 If $S \sim a\, \chi^2_n$ (or more generally $S \sim \gamma(\frac{1}{2}n, 2a)$ with n not necessarily an integer), $h(.)$ is absolutely continuous with derivative existing almost everywhere then

$$E[S\, h(S)] = a\, n\, E[h(S)] + 2a\, E[S\, h'(S)]$$

so long as $S\, h(S)$ and $S\, h'(S)$ both have finite expectation.

Proof The density function of S is $p(\xi) = A\, s^{\frac{1}{2}n-1} \exp(-\frac{s}{2a})$ so that $\frac{\partial}{\partial s}(s\, p(s)) = \frac{1}{2}n\, p(s) - \frac{1}{2a}\, s\, p(s)$

and $\int (s - a\, n)\, p(s)\, ds = -2a\, s\, p(s)$.

Integrating by parts we obtain

$$\int_{\varepsilon}^{T} (s-an)\, p(s)\, h(s)\, ds = [-2as\, p(s)\, h(s)]_{\varepsilon}^{T} + 2a \int_{\varepsilon}^{T} s\, p(s)\, h'(s)\, ds.$$

Dividing the integrals from ε to T as the sum of integrals from ε to $a\,n$ and from $a\,n$ to T we see that if these tend to finite limits as $T \to \infty$ and as $\varepsilon \to 0$ then $s\, p(s)\, h(s)$ also tend to finite limits as $s \to 0$ and as $s \to \infty$. If the limits are non-zero then we may, without loss of generality, assume that they are positive: if not, consider $-h(s)$. Suppose that $\lim_{T\to\infty} s\, p(s)\, h(s) = \ell > 0$ so that there exists $U > a\,n$ for which $\forall\, s > U$, $s\, p(s)\, h(s) > \ell$

and $$\int_U^T (s-an)\, p(s)\, h(s) ds \geq (U - a\,n) \int_U^T p(s)\, h(s)\, ds$$

which is greater than or equal to $(U - a_n)\, l \int_U^T \frac{1}{s}\, ds$.

Similarly, if $\lim_{\varepsilon \to 0} s\, p(s)\, h(s) = l > 0$ then $\exists\ \delta < a_n$ such that $\forall\ s < \delta \quad s\, p(s)\, h(s) > l$

and $$\int_\varepsilon^\delta (a_n - \delta)\, p(s)\, h(s)\, ds \geq (a_n - \delta) \int_\varepsilon^\delta p(s)\, h(s)\, ds$$

$$\geq (a_n - \delta)\, l \int_\varepsilon^\delta \frac{1}{s}\, ds.$$

In both cases the right hand sides converge by assumption. This contradicts the assumption that the other limits are non-zero and proves the lemma.

The next lemma collects together some results for the multivariate normal distribution which are mostly generalisations of lemmas 2 and 3. We shall not find all of them directly useful but, in the order given the later results are derived from the earlier ones. The first result is quoted in Efron and Morris(1976); the sixth is adapted from Stein(1966).

Lemma 5 Suppose that $X \sim N_p(\xi, \sigma^2 I)$ and $h(.)$ is an absolutely continuous row vector (or scalar) function independent of ξ. Suppose also that in each of the following expressions the left hand side is finite and the last term on the right hand side exists and is finite. (In expressions involving derivatives of h with respect to X this implies that the derivative exists almost everywhere). The following results then apply:

(1) $E[X\, h(X)] = \xi\, E[h(X)] + \sigma^2\, E[\frac{\partial}{\partial \xi} h(X)]$

(2) $E[X\, h(X)] = \xi\, E[h(X)] + \sigma^2 \frac{\partial}{\partial \xi} E[h(X)]$

(3) $E[X\, h(X)] = \sigma^2\, e^{-\lambda} \frac{\partial}{\partial \xi}\{e^{\lambda}\, E[h(X)]\}$ where $\lambda = \frac{1}{2\sigma^2} \| \xi \|^2$

(4) $E[X\, h(\|X\|^2)] = \xi\, E[h(\|X\|^2)] + 2\sigma^2\, E[X\, h'(\|X\|^2)]$

(5) $E[X\, h(\|X\|^2)] = \xi\, E[h(\|X\|^2)] + \xi \frac{\partial}{\partial \lambda} E[h(\|X\|^2)]$

(6) $E[X\, h(\|X\|^2)] = \xi\, e^{-\lambda} \frac{\partial}{\partial \lambda}\{e^{\lambda}\, E[h(\|X\|^2)]\}$

(7) $E[\|X\|^2 h(X)] = \xi^T E[X\, h(X)] + p\sigma^2\, E[h(X)] + \sigma^2\, E[X^T \frac{\partial}{\partial X} h(X)]$

(8) $E[\|X\|^2 h(\|X\|^2)] = \xi^T E[X\, h(\|X\|^2)] + p\sigma^2\, E[h(\|X\|^2)]$
$+ 2\sigma^2\, E[\|X\|^2 h'(\|X\|^2)]$.

Proof Note first that if any of the above results is true when $h(.)$ is a scalar then it is also true when $h(.)$ is a row vector.

We see this by applying the result to each component of $h(.)$. Writing E_i for the expectation with respect to X_i (the i th component of X) we have, by lemmas 2 and 3, for scalar $h(.)$:

$$E_i[(X_i - \xi_i)h(X)] = \sigma^2 E_i[\frac{\partial}{\partial X_i} h(X)] = \sigma^2 \frac{\partial}{\partial \xi_i} E_i[h(X)].$$

Thus, on taking expectations with respect to the other coordinates we have,

$$E[(X_i - \xi_i)h(X)] = \sigma^2 E_i[\frac{\partial}{\partial X_i} h(X)] = \sigma^2 \frac{\partial}{\partial \xi_i} E[h(X)]$$

thus proving (1) and (2).

Now, for any function $f(\xi)$, differentiable with respect to ξ,

$$\frac{\partial}{\partial \xi}\{e^{\lambda} f(\xi)\} = \frac{\partial \lambda}{\partial \xi} e^{\lambda} f(\xi) + e^{\lambda} \frac{\partial}{\partial \xi} f(\xi)$$

$$= \sigma^{-2} \xi e^{\lambda} f(\xi) + e^{\lambda} \frac{\partial}{\partial \xi} f(\xi).$$

Since $E[h(X)]$ is a function of ξ differentiable with resect to ξ, (2) and (3) are equivalent.

Now suppose that a function $f(y_1,\ldots,y_p)$ depends only on the length of $y = [y_1,\ldots,y_p]^T$. Taking polar coordinates r,θ where

$= [\theta_1,\ldots,\theta_{p-1}]^T$ we obtain

(a) $$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial \theta}{\partial y}\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial r}\frac{\partial r}{\partial y} \quad \text{since} \quad \frac{\partial f}{\partial \theta} = 0.$$

Applying this to $h(||X||^2)$ which depends only on the length of X we have $\frac{\partial}{\partial X} h(||X||^2) = \frac{\partial ||X||^2}{\partial X}\frac{\partial}{\partial ||X||^2} h(||X||^2) = 2X\, h'(||X||^2)$ and so (1) implies (4).

Also, $||X||^2$ has a non-central χ^2 distibution which therefore depends on ξ only through $||\xi||^2 = 2\lambda\sigma^2$ and so $E[h(||X||^2)]$ depends only on λ and we may apply (a) to obtain

$$\frac{\partial}{\partial \xi} E[h(||X||^2)] = \frac{\partial \lambda}{\partial \xi}\frac{\partial}{\partial \lambda} E[h(||X||^2)] = \sigma^{-2}\xi \frac{\partial}{\partial \lambda} E[h(||X||^2)]$$

and so (2) implies (5).

By the same argument (3) implies (6) or alternatively we may show the equivalence of (5) and (6) in the same way that we showed the equivalence of (2) and (3).

Writing $X^T h(X)$ instead of $h(X)$ in (1) we obtain

$$E[X X^T h(X)] = \xi E[X^T h(X)] + \sigma^2 E[\frac{\partial}{\partial X}(X^T h(X))]$$

$$= \xi E[X^T h(X)] + \sigma^2 I E[h(X)] + \sigma^2 E[\frac{\partial}{\partial X} h(X)\, X^T]$$

for h a scalar function of X. Taking the trace of both sides of this expression proves (7).

We may now prove that (7) implies (8) in the same way that we showed that (1) implies (4).

One final comment on the proof is necessary. If $E[X\ h(X)]$ is finite then so is $E[h(X)]$ and if $E[\|X\|^2 h(X)]$ is finite then so are $E[X\ X^T\ h(X)]$, $E[X^T\ h(X)]$ and $E[h(X)]$.

We may now prove some similar results for non-central χ^2 and F disributions (the degrees of freedom not necessarily being integers). They are more complicated since the derivatives of the density functions are related to density functions in the same distributional family but different numbers of degrees of freedom. Since these densities are mixtures of the corresponding central densities, we may also derive expressions relating to expectations with respect to the mixing distribution which is Poisson with parameter λ.

We first make some remarks concerning the notation for these mixtures of distributions.

Suppose that T is a random variable and for each t there is a random variable X_t. If $\pi(t)$ is the density function of T and $p_t(x)$ is the density function of X_t then the joint density $p_t(x)\ \pi(t)$ is the density of a random variable (X,T) where the conditional distribution of X given $T = t$ is $p_t(x)$. The marginal density of X is $\int p_t(x)\ \pi(t)\ dt$. We use the notation X_T for a random variable with this density. The random variable X_T is defined as follows: observe $T = t$ from the density $\pi(t)$ then observe X_t from the density $p_t(x)$. This variable may be thought of as a mixture of the X_t random variables with weighting function $\pi(t)$.

With the above notation we may write χ^2_{p+2K} for a random variable with a non-central $\chi^2_p(\lambda)$ distribution where K has a Poisson distribution with parameter λ. A similar remark applies to the unweighted non central F distribution (i.e. the ratio of a non-central χ^2 distribution to an independent central χ^2 distribution - not divided by their degrees of freedom) which is a Poisson weighted mixture of unweighted central F distributions.

We may now state and prove a lemma concerning central and non-central χ^2 distributions.

Lemma 6 If $U_i \sim$ a χ^2_{p+2i}, $W_i \sim$ a $\chi^2_{p+2i}(\lambda)$ then, if K has a Poisson distribution with parameter λ, $W_i = U_{i+K}$ and

$$(9)\quad E[U_i^\alpha\ h(U_i)] = (\tfrac{p}{2} + i)_\alpha\ (2a)^\alpha E[h(U_{i+\alpha})] \quad \text{if either side is finite}$$

$$(10)\quad E[h(W_{i+1})] = e^{-\lambda}\ \frac{\partial}{\partial\lambda}\{e^{\lambda}\ E\ h(W_i)\}$$

(11) $E[W_i\, h(W_i)] = (p + 2i)\, a E[h(W_{i+1})] + 2a\lambda E[h(W_{i+2})]$

(12) $E[K\, h(U_{i+K})] = \lambda E[h(U_{i+1+K})] = \lambda E[h(W_{i+1})]$

Proof Let the density function of U_i be $p_i^{\circ}(u)$ and let the density of W_i be $p_i(w)$. These densities are given by

$$p_i^{\circ}(u) = \frac{u^{\frac{1}{2}p+i-1}\exp(-\frac{u}{2a})}{(2a)^{\frac{1}{2}p+i}\,\Gamma(\frac{1}{2}p+i)} \quad \text{and} \quad p_i(w) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\,p_{i+k}^{\circ}(w)$$

and the joint density of U_{i+K} and K is given by

$$p_i(u,k) = e^{-\lambda}\frac{\lambda^k}{k!}\,p_{i+k}^{\circ}(u).$$

Since $u^{\alpha}\,p_i^{\circ}(u) = (2a)^{\alpha}(\frac{1}{2}p+i)_{\alpha}\,p_{i+\alpha}^{\circ}(u)$ we have

$$\begin{aligned} E[U_i^{\alpha}\,h(U_i)] &= \int_0^{\infty} u^{\alpha}\,h(u)\,p_i^{\circ}(u)\,du \\ &= (2a)^{\alpha}(\tfrac{1}{2}p+i)_{\alpha}\int_0^{\infty} h(u)\,p_{i+\alpha}^{\circ}(u)\,du \\ &= (2a)^{\alpha}(\tfrac{1}{2}p+i)_{\alpha}\,E[h(U_{i+\alpha})] \end{aligned}$$

which proves (9).

$$\begin{aligned} \text{Now } e^{-\lambda}\frac{\partial}{\partial\lambda}\{e^{\lambda}E[h(W_i)]\} &= e^{-\lambda}\frac{\partial}{\partial\lambda}\{\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty} h(u)\,p_{i+k}^{\circ}(u)\,du\} \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^{k-1}}{k!}\int_0^{\infty} h(u)\,p_{i+k}^{\circ}(u)\,du \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty} h(u)\,p_{i+1+k}^{\circ}(u)\,du \\ &= E[h(W_{i+1})] \end{aligned}$$

which proves (10).

$$\begin{aligned} \text{Also } E[W_i\,h(W_i)] &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty} u\,h(u)\,p_{i+k}^{\circ}(u)\,du \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty}(2a)(\tfrac{1}{2}p+i+k)\,h(u)\,p_{i+1+k}^{\circ}(u)\,du \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\int_0^{\infty}(2a)(\tfrac{1}{2}p+i)\,h(u)\,p_{i+k+1}^{\circ}(u)\,du \\ &\quad + e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^k}{k!}\int_0^{\infty}(2a)\,h(u)\,p_{i+k+1}^{\circ}(u)\,du \\ &= (p+2i)\,a E[h(W_{i+1})] + 2a\lambda e^{-\lambda}\frac{\partial}{\partial\lambda}\{e^{\lambda}E[h(W_{i+1})]\} \\ &= (p+2i)\,a E[h(W_{i+1})] + 2a\lambda E[h(W_{i+2})] \end{aligned}$$

Proving (11).

Finally, $E[K\, h(U_{i+K})] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} \int_0^{\infty} h(u)\, p^{\circ}_{i+k}(u)\, du$

$= E[h(W_{i+1})] = \lambda\, E[h(U_{i+1+K})]$

proving (12).

Note that by (10) $\frac{\partial}{\partial\lambda} E[h(W_i)] = E[h(W_{i+1})] - E[h(W_i)]$.

The next lemma refers to the unweighted F and non-central F distributions, or to multiples of them. We note that an unweighted $F_{m,n}(\lambda)$ distribution is the same as an inverse non-central beta distribution with parameters $\frac{1}{2}m$, $\frac{1}{2}n$ and non-centrality parameter λ, and whose density function is

$$p(u) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{1}{B(\frac{1}{2}m+k, \frac{1}{2}n)} \frac{u^{\frac{1}{2}m+k-1}}{(1+u)^{\frac{1}{2}m+\frac{1}{2}n+k}}$$

Lemma 7 If $F_{ij} \sim a\,\beta_2(\frac{1}{2}p+i, \frac{1}{2}n+j)$ then

$$(13)\quad E[F_{ij}^{\alpha}(a+F_{ij})^{-\beta} h(F_{ij})] = a^{\alpha-\beta} \frac{(\frac{1}{2}p+i)_{\alpha}(\frac{1}{2}n+j)_{\alpha-\beta}}{(\frac{1}{2}p+\frac{1}{2}n+i+j)_{\beta}} E[h(F_{i+\alpha, j+\alpha-\beta})]$$

If G_{ij} a $(\frac{1}{2}p+i, \frac{1}{2}n+j,)$ then $G_{ij} = F_{i+K,j}$ where K has a Poisson distribution with parameter λ and

$$(14)\quad E[G_{ij}^{\alpha}(a+G_{ij})^{-\beta} h(G_{ij})] = a^{\alpha-\beta} E\left[\frac{(\frac{p}{2}+i+K)_{\alpha}(\frac{n}{2}+j)_{\alpha-\beta}}{(\frac{1}{2}p+\frac{1}{2}n+i+j)_{\beta}} h(F_{i+\alpha+K, j+\beta-\alpha})\right].$$

Proof The probability density of F_{ij} is given by

$$p^{\circ}_{ij}(u) = \frac{a^{\frac{1}{2}n+j}}{B(\frac{1}{2}p+i, \frac{1}{2}n+j)} \frac{u^{\frac{1}{2}p+i-1}}{(a+u)^{\frac{1}{2}p+\frac{1}{2}n+i+j}}.$$

Therefore $\frac{u^{\alpha}}{(a+u)^{\beta}} p^{\circ}_{ij} = \frac{B(\frac{1}{2}p+i+\alpha, \frac{1}{2}n+j+\beta-\alpha)}{B(\frac{1}{2}p+i, \frac{1}{2}n+j)\, a^{\beta-\alpha}} p^{\circ}_{i+\alpha, j+\beta-\alpha}(u)$

$= a^{\alpha-\beta} \frac{(\frac{1}{2}p+i)_{\alpha}(\frac{1}{2}n+j)_{\beta-\alpha}}{(\frac{1}{2}p+\frac{1}{2}n+i+j)_{\beta}} p^{\circ}_{i+\alpha, j+\beta-\alpha}(u).$

It follows that

$$E\left[\frac{F_{ij}^{\alpha}}{(a+F_{ij})^{\beta}} h(F_{ij})\right] = a^{\alpha-\beta} \frac{(\frac{1}{2}p+i)_{\alpha}(\frac{1}{2}n+j)_{\beta-\alpha}}{(\frac{1}{2}p+\frac{1}{2}n+i+j)_{\beta}} \int_0^{\infty} h(u)\, p^{\circ}_{i+\alpha, j+\beta-\alpha}(u)\, du$$

Proving (13).

The probability density of G_{ij} is $p_{ij}(u) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} p^{\circ}_{i+k,j}(u)$

and the joint density of $(F_{i+K,j}, K)$ is $p_{ij}(u,k) = e^{-\lambda} \frac{\lambda^k}{k!} p^{\circ}_{i+k,j}(u)$.

Therefore

$$E\left[\frac{G_{ij}^{\alpha}}{(a+G_{ij})^{\beta}}\,h(G_{ij})\right] = E\left[E\left[\frac{F_{i+K,j}^{\alpha}}{(a+F_{i+K,j})^{\beta}}\,h(F_{i+K,j})\,|K\right]\right]$$

$$= E\left[a^{\alpha-\beta}\,E\left[\frac{(\tfrac{1}{2}p+i+K)_{\alpha}(\tfrac{1}{2}n+j)_{\beta-\alpha}}{(\tfrac{1}{2}p+\tfrac{1}{2}n+i+j+K)_{\beta}}\,h(F_{i+\alpha+K,j+\beta-\alpha})\,|K\right]\right]$$

$$= a^{\alpha-\beta}\,(\tfrac{1}{2}n+j)_{\beta-\alpha}\,E\left[\frac{(\tfrac{1}{2}p+i+K)_{\alpha}}{(\tfrac{1}{2}p+\tfrac{1}{2}n+i+j+K)_{\beta}}\,h(F_{i+\alpha+K,j+\beta-\alpha})\right]$$

proving (14).

6.3 An Unbiased Estimator for the Risk

We shall denote the risk function of the estimator $\delta(X,S)$ by $R_{\delta}(\xi,\sigma^2)$. The following theorem generalised from Efron and Morris(1976) gives a formula for the risk of $\delta(X,S) = (1 - \frac{1}{F}\,\tilde{c}\,v(W,S))\,X$ in terms of W and S.

Theorem 1 If $v(W,S)$ is absolutely continuous with partial derivatives existing almost everywhere, if $c = \frac{p}{n}$ $\tilde{c} = \frac{p-2}{n+2}$ and if each term under the expectation operator in the right hand side of (1) is finite then

(1) $$R_{\delta}(\xi,\sigma^2) = p - 4\tilde{c}\left\{E\left[\frac{1}{F}\,\frac{p-2}{4}\,v(2-v) + W\,\frac{\partial v}{\partial W} - c\,v\,S\,\frac{\partial v}{\partial S}\right\}\right].$$

Proof We may write

$$R_{\delta}(\xi,\sigma^2) = \sigma^{-2}\,E\left[\,\|(X - \frac{\tilde{c}\,v}{F}\,X) - \xi\,\|^2\,\right]$$

$$= E\left[\sigma^{-2}\|X-\xi\|^2 - 2\sigma^{-2}\,\frac{\tilde{c}\,v}{F}\,X^{T}(X-\xi) + \frac{\tilde{c}^2 v^2}{\sigma^2 F}\,\|X\|^2\right].$$

Now using (7) of lemma 6.2.5 applied to $Y = X - \xi$ (which has zero expectation) and taking $h(Y) = \sigma^{-2}$ we have the well known result that $E[\sigma^{-2}\,\|X-\xi\|^2\,] = p$.

Using (8) of lemma 6.2.5 where $h(X) = \frac{v}{\sigma^2 F}$ and afterwards taking expectations with respect to S we obtain

$$\sigma^{-2}\,E\,[\frac{v}{F}\,X^{T}(X-\xi)] = p\,E[E[\frac{v}{F}|S]] + 2E[E[\,\|X\|^2\frac{\partial}{\partial\|X\|^2}(\frac{v}{F})|S]]$$

$$= p\,E[\frac{v}{F}\,] + 2E[W\,\frac{\partial}{\partial W}\,(\frac{v}{F})]$$

$$= p\,E[\frac{v}{F}] - 2E[\frac{v}{F}] + 2E[\frac{1}{F}\,W\,\frac{\partial v}{\partial W}]$$

where the second equality is due to the scale invariance of the operator $y\,\frac{\partial}{\partial y}$ and the third is due to the fact that

$$W\,\frac{\partial}{\partial W}\,(\frac{v}{F}) = w\,\frac{\partial}{\partial W}\,(\frac{Sv}{W}) = S\,\frac{\partial v}{\partial W} - \frac{Sv}{W} = \frac{W}{F}\,\frac{\partial v}{\partial W} - \frac{v}{F}\,.$$

Finally, using lemma 6.2.4 with $h(S) = \frac{v^2\,\|X\|^2}{\sigma^2 WF} = \frac{p\,v^2}{\sigma^2 F}$ and $a = \frac{\sigma^2}{n}$ and taking expectations with respect to X last we obtain

$$\sigma^{-2}[E\,\frac{v^2}{F}\,\|X\|^2] = E[\frac{p\,v^2}{\sigma^2 F}S] = E[E[\frac{p\,v^2}{\sigma^2 F}S \mid W]]$$

$$= p\,E[\frac{v^2}{F}] + \frac{2p}{n}E[S\,\frac{\partial}{\partial S}\,(\frac{v^2}{F})]$$

$$= p\,E[\frac{v^2}{F}] + \frac{2p}{n}E[\frac{v^2}{F}] + \frac{4p}{n}E[\frac{v}{F}\,S\,\frac{\partial v}{\partial S}].$$

The last equality follows from the fact that

$$S\,\frac{\partial}{\partial S}\,(\frac{v^2}{F}) = S\,\frac{\partial}{\partial S}\,(\frac{v^2 S}{W}) = \frac{S\,v^2}{W} + \frac{2S^2}{W}\,v\,\frac{\partial v}{\partial S}$$

$$= \frac{v^2}{F} + \frac{2\,v}{F}\,S\,\frac{\partial v}{\partial S}.$$

Combining these terms together gives

$$R_\delta(\xi,\sigma^2) = p - E[\check{c}(p-2)\frac{v}{F}(2 - \frac{p}{n}\,\frac{n+2}{p-2}\,\check{c}\,v) + \frac{4\check{c}}{F}\,W\,\frac{\partial v}{\partial W}$$

$$- \frac{4\check{c}^2 p\,v}{n\,F}\,S\,\frac{\partial v}{\partial S}].$$

This is true no matter what value we take $\check{c}$ to have. It is convenient to take $\check{c} = \frac{p-2}{p}\,\frac{n}{n+2}$ and $c = \frac{p}{n}\,\check{c}$ in which case

$$R_\delta(\xi,\sigma^2) = p - 4\check{c}\,E\left[\frac{1}{F}\left\{\frac{p-2}{4}\,v(2-v) + W\,\frac{\partial v}{\partial W} - c\,v\,S\,\frac{\partial v}{\partial S}\right\}\right].$$

Note that, for the proof to be valid, each term must have finite expectation since the proof calculated the risk as the sum of the expectations of several terms.

Remark If $a = \frac{\sigma^2}{n}$ then lemma 4 remains formally true when $n = \infty$ and $S = \sigma^2$ since $\frac{1}{n}\,\chi^2_n$ converges strongly to 1 as $n \to \infty$. In this case the last term is zero (if $h'(\sigma^2)$ exists) so that, in the theorem, $c = 0$. This gives the formula for known variance

$$R_\delta(\xi,\sigma^2) = p - 4\check{c}\,E\left[\frac{1}{F}\left\{\frac{p-2}{4}\,v(2-v) + W\,\frac{\partial v}{\partial W}\right\}\right].$$

We will now express the risk function in terms of F and S. This will be seen to be a special case of the transformation

$$U = W^\alpha S^\beta, \qquad V = W^\gamma S^\delta \quad \text{where} \quad \det\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix} \neq 0.$$

We let v represent the *value* of the function occurring in the shrinkage factor so that we may write $v(U,V) = v(W,S)$ and use subscripts to denote the variable held constant in partial derivatives.

We have

$$\begin{bmatrix}\frac{\partial}{\partial W}\\ \frac{\partial}{\partial S}\end{bmatrix} = \begin{bmatrix}\frac{\partial U}{\partial W} & \frac{\partial V}{\partial W}\\ \frac{\partial U}{\partial S} & \frac{\partial V}{\partial S}\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial U}\\ \frac{\partial}{\partial V}\end{bmatrix} = \begin{bmatrix}\alpha\frac{U}{W} & \gamma\frac{V}{W}\\ \beta\frac{U}{S} & \delta\frac{V}{S}\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial U}\\ \frac{\partial}{\partial V}\end{bmatrix}$$

therefore $W(\frac{\partial v}{\partial W})_S = \alpha\,U(\frac{\partial v}{\partial U})_V + \gamma\,V(\frac{\partial v}{\partial V})_U$

and $S(\frac{\partial v}{\partial S})_W = \beta\,U(\frac{\partial v}{\partial U})_V + \delta\,V(\frac{\partial v}{\partial V})_U$.

This gives the risk function in terms of U and V

$$R_\delta(\xi,\sigma^2) = p - 4\tilde{c}\, E\left[\frac{1}{F}\left\{\frac{p-2}{4}\, v(2-v) + (\alpha-\beta c v) U\left(\frac{\partial v}{\partial U}\right)_V + (\gamma-\delta c v) V\left(\frac{\partial v}{\partial V}\right)_U\right\}\right].$$

As this does not involve ξ or σ^2 explicitly this provides an unbiased estimator for the risk

$$\hat{R}_\delta(U,V) = p - 4\tilde{c}\, \frac{1}{F}\left\{\frac{p-2}{4}\, v(2-v) + (\alpha-\beta c v) U\left(\frac{\partial v}{\partial U}\right)_V + (\gamma-\delta c v) V\left(\frac{\partial v}{\partial V}\right)_U\right\}.$$

If we define $\quad r_\delta(\xi,\sigma^2) = \frac{1}{4\tilde{c}}\,(p - R_\delta(\xi,\sigma^2))$

and $\quad \hat{r}_\delta(U,V) = \frac{1}{4\tilde{c}}\,(p - \hat{R}(U,V))$

then these give the scaled reduction in risk over that of the maximum likelihood estimator and an unbiased estimator of this quantity.

We are particularly interested in the result of putting $U = F$, and $V = S$ given by $\alpha = 1$, $\beta = -1$, $\gamma = 0$ and $\delta = 1$. We quote the result as a theorem.

Theorem 2 If $v(F,S)$ is absolutely continuous with finite partial derivatives almost everywhere, if $c = \frac{p}{n}\,\tilde{c} - \frac{p-2}{n+2}$ and if each term on the right hand side of (2) has finite expectation then an unbiased estimator for the scaled reduction in risk over the maximum likelihood estimator is given by

$$(2)\quad \hat{r}_\delta(F,S) = \frac{1}{F}\left\{\frac{p-2}{4}\, v(2-v) + (1+c v)F\,\frac{\partial v}{\partial F} - c v S\,\frac{\partial v}{\partial S}\right\}.$$

(The special case in which v is independent of S - except for the dependence on S implicit in the dependence of v on F - was given in Efron and Morris(1976)).

As an application of this result we take $v = t$ so that we have

$$\hat{r}_\delta(F,S) = \frac{1}{F}\,\frac{p-2}{4}\, t(2-t)$$

which is positive if $0 < t < 2$ with a maximum of $\frac{p-2}{4F}$ at $t = 1$. Thus we may achieve a uniform reduction in risk by using this estimator which is the James-Stein estimator. In chapter 7 we shall show that there is no spherically symmetric estimator with uniformly smaller risk estimator than the James-Stein estimator with $t = 1$. However, we shall see in the next section that the positive part James-Stein estimator does have uniformly smaller risk than the unmodified version. This demonstrates that we need more powerful results to prove some domination theorems.

Remark Efron and Morris(1976) showed that when v depends on F alone the unbiased estimator of the risk is unique. We shall show that the same is true in the more general case considered here. This follows from the completeness of S as a function of σ^2 and from the

completeness of W as a function of λ. Suppose $R^*(F,S)$ is another unbiased estimator of the risk. We have

$$E[R^*(F,S) - R(\hat{F},S)] = E[R^*(F,S)] - E[\hat{R}(F,S)]$$
$$= R_\delta(\xi,\sigma^2) - R_\delta(\xi,\sigma^2) = 0.$$

Therefore

$$E[R^*(F,S) - \hat{R}(F,S)] = E[E[R^*(F,S) - \hat{R}(F,S) \mid W]] = 0$$

which, by the completeness of W (which is independent of S), implies

$$E[R^*(F,S) - \hat{R}(F,S) \mid W] = 0 \quad \text{almost everywhere.}$$

6.4 Explicit Expressions for the Risk

We shall first calculate the risk in terms of W and S. This will be done in terms of the shrinkage factor h as well as in terms of g and v.

We note that the risk depends on λ and σ^2, the dependence on ξ being only through λ. We shall henceforth write the risk as $R(\lambda,\sigma^2)$.

We then have the following formula for the risk:

$$R(\lambda,\sigma^2) = \sigma^{-2} E[\|h(W,S)X - \xi\|^2]$$
$$= \sigma^{-2} E[\|X\|^2 h^2(W,S)] - 2\sigma^{-2}\xi^T E[X\, h(W,S)] + \sigma^{-2}\|\xi\|^2.$$

Now, by 6.2.6 and by 6.2.10,

$$E[X\, h(W,S)] = \xi\, e^{-\lambda} \frac{\partial}{\partial\lambda}\{e^{\lambda} E[h(W,S)]\}$$
$$= \xi\, E[h(W_1,S)]$$

where $W_i \sim \frac{\sigma^2}{p}\chi^2_{p+2i}(\lambda)$ and in particular $W_0 = W$, $W_1 \sim \frac{\sigma^2}{p}\chi^2_{p+2}(\lambda)$.

Thus we have

$$(1)\quad R(\lambda,\sigma^2) = p\sigma^{-2} E[W\, h^2(W,S)] - 4\lambda E[h(W_1,S)] + 2\lambda.$$

Substituting $h = 1-g$ in (1) gives

$$(2)\quad R(\lambda,\sigma^2) = p\sigma^{-2} E[W - 2W g(W,S) + W g^2(W,S)] - 2\lambda + 4\lambda E[g(W_1,S)]$$

and since $E[W] = \frac{\sigma^2}{p}(p+2\lambda)$ by 6.2.11 (or by a well known result) we have

$$(3)\quad R(\lambda,\sigma^2) = p + p\sigma^{-2} E[Wg(W,S)\{g(W,S)-2\}] + 4\lambda E[g(W_1,S)].$$

Using 6.1.1 again we obtain

$$(4)\quad R(\lambda,\sigma^2) = p + p\sigma^{-2} E[Wg(W,S)\{g(W,S)-2\}] + 2p\sigma^{-2} E[W_{-1}g(W_{-1},S)] - 2(p-2) E[g(W,S)].$$

Substituting $g = \frac{\tilde{c}\, v}{F} = \frac{\tilde{c}\, v\, S}{W}$ we obtain

$$(5)\quad R(\lambda,\sigma^2) = p + \frac{p\tilde{c}^2}{\sigma^2} E[\frac{S^2}{W} v^2(W,S)] - 2(p-2)\tilde{c}\, E[\frac{S}{W} v(W,S)] + \frac{2p\tilde{c}}{\sigma^2}\{E[Sv(W_{-1},S)] - E[Sv(W,S)]\}$$

which we may write as

$$(6)\quad R(\lambda,\sigma^2) = p + \frac{p\tilde{c}^2}{\sigma^2} E[\frac{S^2}{W} v(W,S)\{v(W,S)-2\}] + \frac{2p\tilde{c}^2}{\sigma^2} E[\frac{S^2}{W} v(W,S)] - 2(p-2)\tilde{c}\, E[\frac{S}{W} v(W,S)] + \frac{2p\tilde{c}}{\sigma^2}\{E[Sv(W_{-1},S)] - E[Sv(W,S)]\}.$$

Applying the central χ^2 case of 6.2.11 with $i = 0$, n and $\frac{\sigma^2}{n}$ substituted for p and a respectively, and putting $S_i \sim \frac{\sigma^2}{n}\chi^2_{n+2i}$ we obtain

$$(7)\quad R(\lambda,\sigma^2) = p + \frac{p\tilde{c}^2}{\sigma^2}\,E\left[\frac{S^2}{W}v(W,S)\{v(W,S)-2\}\right] + 2p\,\tilde{c}^2\,E\left[\frac{S_1}{W}\,v(W,S_1)\right]$$
$$- 2(p-2)\,\tilde{c}\,E\left[\frac{S}{W}\,v(W,S)\right] + \qquad \{E[S\,v(W_{-1},S)] - E[S\,v(W,S)]\}.$$

Applying the same formula once more gives

$$(8)\quad R(\lambda,\sigma^2) = p + p\tilde{c}^2E\left[\frac{S_1}{W}\,v(W,S_1)\{v(W,S_1)-2\}\right] + 2\,p\,\tilde{c}^2E\left[\frac{S_1}{W}\,v(W,S_1)\right]$$
$$- 2(p-2)\,\tilde{c}\,E\left[\frac{S}{W}\,v(W,S)\right] + 2\,p\,\tilde{c}\{E[v(W_{-1},S_1)] - E[v(W,S_1)]\}$$

which is an interesting result sice it does not contain λ or σ^2 explicitly and shows some simiarity with the unbiased risk estimator of section 6.3. This expression does not however provide an unbiased estimator for the risk since it contains S_1 and W_{-1} which may only be transformed to S and W by transformations involving unknown parameters.

Using 6.2.11 again for the cases $i = 0$ and $i = 1$ gives

$$(9)\quad R(\lambda,\sigma^2) = p + (p-2)\,\tilde{c}\,\sigma^2\,E\left[\frac{1}{W}\,v(W,S_2)\{v(W,S_2)-2\}\right]$$
$$+ 2(p-2)\sigma^2\left\{E\left[\frac{1}{S_2}\,v(W_{-1},S_2)\right] - E\left[\frac{1}{S_2}\,v(W,S_2)\right]\right\}$$
$$+ 2(p-2)\,\tilde{c}\,\sigma^2\left\{E\left[\frac{1}{W}\,v(W,S_2)\right] - E\left[\frac{1}{W}\,v(W,S_1)\right]\right\}$$

which, by rearranging terms may be written

$$(10)\quad R(\lambda,\sigma^2) = p + (p-2)\,\tilde{c}\,\sigma^2E\left[\frac{1}{W}\,v(W,S_1)\{v(W,S_1)-2\}\right]$$
$$+ 2(p-2)\sigma^2\left\{E\left[\frac{1}{S_2}\,v(W_{-1},S_2)\right] - E\left[\frac{1}{S_2}\,v(W,S_2)\right]\right\}$$
$$+ (p-2)\,\tilde{c}\,\sigma^2\left\{E\left[\frac{1}{W}\,v^2(W,S_2)\right] - E\left[\frac{1}{W}\,v^2(W,S_1)\right]\right\}.$$

Now let K have a Poisson distribution with parameter λ and let $U_i = \frac{\sigma^2}{p}\chi^2_{p+2i}$. We then have $U_{i+K} \sim \frac{\sigma^2}{p}\chi^2_{p+2i}(\lambda)$ and from (9) we obtain

$$(11)\quad R(\lambda,\sigma^2) = p + (p-2)\,\tilde{c}\,\sigma^2E\left[\frac{1}{U_K}\,v(U_K,S_2)\{v(U_K,S_2)-2\}\right]$$
$$+ 2(p-2)\sigma^2\left\{E\left[\frac{1}{S_2}\,v(U_{K-1},S_2)\right] - E\left[\frac{1}{S_2}\,v(U_K,S_2)\right]\right\}$$
$$+ 2(p-2)\,\tilde{c}\,\sigma^2\left\{E\left[\frac{1}{U_K}\,v(U_K,S_2)\right] - E\left[\frac{1}{U_K}\,v(U_K,S_1)\right]\right\}$$

while from (10) we deduce

$$(12)\quad R(\lambda,\sigma^2) = p + (p-2)\,\check{c}\,\sigma^2 E\left[\frac{1}{U_K}\, v(U_K,S_1)\{v(U_K,S_1)-2\}\right]$$
$$+ 2(p-2)\,\check{c}\,\sigma^2\left\{E\left[\frac{1}{S_2}\, v(U_{K-1},S_2)\right] - E\left[\frac{1}{S_2}\, v(U_K,S_2)\right]\right\}$$
$$+ (p-2)\,\check{c}\,\sigma^2\left\{E\left[\frac{1}{U_K}\, v^2(U_K,S_2)\right] - E\left[\frac{1}{U_K}\, v^2(U_K,S_1)\right]\right\}.$$

Computing conditionally on K and taking expectations with respect to K last and using 6.2.9 we obtain

$$(13)\quad R(\lambda,\sigma^2) = p + p\,\check{c}\,E\left[\frac{p-2}{p-2+2K}\, v(U_{K-1},S_2)\{v(U_{K-1},S_2)-2\}\right]$$
$$+ 2p\,\check{c}\,E\left[\frac{p-2}{p-2+2K}\,\{v(U_{K-1},S_2) - v(U_{K-1},S_1)\}\right]$$
$$+2p\,\check{c}\{E[v(U_{K-1},S_1)] - E[v(U_K,S_1)]\}$$

from equation (11) using the form of the third term occurring in (8). Similarly from (12) we obtain

$$(14)\quad R(\lambda,\sigma^2) = p + p\,\check{c}\,E\left[\frac{p-2}{p-2+2K}\, v(U_{K-1},S_1)\{v(U_{K-1},S_1)-2\}\right]$$
$$+ p\,\check{c}\,E\left[\frac{p-2}{p-2+2K}\,\{v^2(U_{K-1},S_2) - v^2(U_{K-1},S_1)\}\right]$$
$$+ 2p\,\check{c}\{E[v(U_{K-1},S_1)] - E[v(U_K,S_1)]\}.$$

In Efron and Morris(1976) a formula for the risk is given which only applies when the shrinkage is dependent only on F. It may be derived from (1) using 6.2.12 and 6.2.13. The expression is

$$(15)\quad R(\lambda,\sigma^2) = E\left[(n+p+2K)\,\frac{F}{\frac{n}{p}+F}\, h^2(F) - 4K\,h(F) + 2K\right].$$

Some of the above expressions for the risk have been given before in the case in which h depends on F alone and a few have been given in the general case. The following articles contain expressions similar to these, however, they are all written in a form which makes explicit reference to K. The articles are: Alam(1973), Stein(1966), Baranchik(1970), Strawderman(1971),(1973) and Sclove, Morris and Radhakrishnan(1972).

Another expression valid when v only depends on S may be derived either from (8) or from the risk estimate. The latter is easier but makes unnecessary assumptions about v. We first give that derivation and then check the result using (8).

From the risk estimate we have

$$\frac{R(\lambda,\sigma^2)-p}{p\check{c}} = \frac{4}{p}\,E\left[\frac{1}{F}\left\{\frac{p-2}{4}\, v(v-2) - (1+c\,v)\,F\frac{dv}{dF}\right\}\right]$$
$$= E\left[\frac{p-2}{p}\,\frac{v(v-2)}{F}\right] - \frac{2}{pc}\,E\left[\frac{d}{dF}\,(1+c\,v)^2\right].$$

Integrating the second term by parts and noting that $\frac{dp(F)}{dF} = \frac{p}{2}\{p_{-1,1}(F) - p_{0,1}(F)\}$, where $p_{ij}(F)$ is the density function of F_{ij}, we obtain

$$\frac{R(\lambda,\sigma^2)-p}{p\tilde{c}} = E\left[\frac{p-2}{p}\frac{v(v-2)}{F}\right] - \frac{2}{pc}[(1+cv)^2 p(F)]_0^\infty + \frac{2}{pc}\int_0^\infty (1+cv)^2 \frac{dp(F)}{dF}\,dF$$

$$= E\left[\frac{p-2}{p}\frac{v(v-2)}{F}\right] + \frac{1}{c}E_{-1,1}[(1+cv)^2] - \frac{1}{c}E_{0,1}[(1+cv)^2]$$

so that

$$(16)\quad \frac{R(\lambda,\sigma^2)-p}{p\tilde{c}} = E\left[\frac{p-2}{p}\frac{v(v-2)}{F}\right] + E_{-1,1}[2v] - E_{0,1}[2v] + cE_{-1,1}[v^2] - cE_{0,1}[v^2]$$

where $E_{i,j}[.]$ is the expectation with respect to the probability distribution of F_{ij}.

Now the expression in (8) may be written

$$\frac{R(\lambda,\sigma^2)-p}{p\tilde{c}} = \tilde{c}E_{0,1}\left[\frac{v(v-2)}{F}\right] + E_{-1,1}[2v] - E_{0,1}[2v] + \tilde{c}E_{0,1}\left[\frac{2v}{F}\right] - \frac{p-2}{p}E\left[\frac{2v}{F}\right]$$

$$= \tilde{c}E_{0,1}\left[\frac{v^2}{F}\right] - \frac{p-2}{p}E\left[\frac{2v}{F}\right] + E_{-1,1}[2v] - E_{0,1}[2v]$$

$$= \frac{p-2}{p}E\left[\frac{v^2-2v}{F}\right] + E_{-1,1}[2v] - E_{0,1}[2v] + \tilde{c}E_{0,1}\left[\frac{v^2}{F}\right] - \frac{p-2}{p}E\left[\frac{v^2}{F}\right].$$

This is equivalent to (16) if and only if

$$\tilde{c}E_{0,1}\left[\frac{v^2}{F}\right] - \frac{p-2}{p}E\left[\frac{v^2}{F}\right] = cE_{-1,1}[v^2] - \tilde{c}E_{0,1}[v^2].$$

Putting $\alpha = \frac{1}{2}p$ and $\beta = \frac{1}{2}n$ this is eqivalent to the equation

$$E_{0,1}\left[\frac{\beta+\alpha F}{\alpha F}v^2\right] = \frac{\beta+1}{\alpha}E\left[\frac{v^2}{F}\right] + E_{-1,1}[v^2].$$

Now when v^2 does not depend explicitly on S we may integrate with respect to F alone. The condition is then equivalent to the equation $\frac{\beta+\alpha F}{\alpha F}p_{0,1}(F) = \frac{\beta+1}{\alpha F}p(F) + p_{-1,1}(F)$ which can easily be checked by writing out the density functions which are of the form

$$p_{ij}(F) = \frac{\alpha^{\alpha+i}\beta^{\beta+j}}{B(\alpha+i,\beta+j)}\frac{F^{\alpha+i-1}}{(\beta+\alpha F)^{\alpha+\beta+i+j}}e^{-\lambda}{}_1F_1(\alpha+\beta+i+j;\alpha+i;\frac{\alpha F}{\beta+\alpha F}\lambda);$$

the result following from a recurrence relation for the confluent hypergeometric function.

Another interesting expression was given by Stein in the discussion to Efron and Morris(1973b). This is a forerunner of

Stein's (1973) unbiased estimator for the risk. The formula is

$$E[\| X+\frac{\partial}{\partial X} \log g(X)-\xi \|^2] = p - E[\| \frac{\partial}{\partial X} \log g(X)\|^2-\frac{2}{g(X)}\sum_{i=1}^{p} \frac{\partial^2}{\partial X_i^2} g(X)]$$

valid when $\sigma^2 = 1$.

A similar expresion, namely

$$E[\| X - h(X) - \xi \|^2]= p + E[\| h(X)\|^2] - 2\sigma^2 \text{ tr } E[\frac{\partial}{\partial X} h^T(X)]$$

may be derived from 6.2.1.

In section 6.6 we wish to derive some sufficient conditions for domination of one estimator over another. These conditions include a condition under which $E[X] < E[Y]$ for a pair of random variables X and Y. The next section therefore considers this problem.

6.5 Some Inequalities Concerning Expectations

The following two lemmas give alternative sufficient conditions for the inequality $E[X] < E[Y]$. They are standard results.

Lemma 1 If X and Y are random variables, jointly distributed such that $P(X < Y) = 1$ (i.e. X is stochastically less than Y) then $E[X] < E[Y]$ and if $h(.)$ is an increasing function then $E[h(X)] < E[h(Y)]$ if both sides exist.

Proof Writing $E[X]$ in terms of the joint density and noting that the contribution to the integral from the part of the space for which $X \geq Y$ is zero we have $E[X] = \int_{X<Y} x \, dP_{XY}(x,y)$ and similarly, $E[Y] = \int_{X<Y} y \, dP_{XY}(x,y)$. Since over this subset $x<y$ the result follows. The same argument may be applied to $h(X)$ and $h(Y)$ viz: $E[h(X)] = \int_{X<Y} h(x) \, dP_{XY}(x,y) < \int_{X<Y} h(y) \, dP_{XY}x,y)$ since when $x<y$, $h(x) < h(y)$. Alternatively we may apply the theorem to $H = h(X)$ and $K = h(Y)$ for which $P(H < K) = 1$.

Lemma 2 If a random variable X has distribution function P_X and Y has distribution function P_Y and if $P_X(u) > P_Y(u)$ for almost all u, then $E[X] < E[Y]$. Also, if $h(.)$ is non-decreasing then $E[h(X)] \leq E[h(Y)]$ with strict inequality if h is strictly increasing on a set which has non-zero measure in both X and Y. These inequalities apply so long as the expectations exist.

Proof We have

$$E[X] = \int_{-\infty}^{\infty} x \, dP_X(x)$$

$$= -\int_{-\infty}^{0} P_X(x) \, dx + \int_{0}^{\infty} (1-P_X(x)) \, dx.$$

This follows from lemma 6.2.1. Using the assumed inequality for the distribution functions we obtain

$$E[X] < -\int_{-\infty}^{0} P_Y(x)\,dx + \int_{0}^{\infty} (1 - P_Y(x))\,dx$$

$$= \int_{-\infty}^{\infty} y\,dP_Y(y)$$

$$= E[Y].$$

If $H = h(X)$ and $K = h(Y)$ then $P_X(u) > P_Y(u) \Rightarrow P_H(v) \geq P_K(v)$ where $v = h(u)$. The inequality is strict if, for some set of non-zero measure bounded above by u, $h(x)$ is strictly increasing. Applying the above proof to H and K we have $E[h(X)] \leq E[h(Y)]$. Strict inequality clearly applies when h is strictly increasing on some set of non-zero measure in X and Y.

Note that the result of lemma 2 depends only on properties of the two marginal distributions and not on any assumed joint distribution of X and Y. However, we shall show that if X and Y are jointly distributed such that $P(X < Y) = 1$ then $P_X(u) > P_Y(u)$ for almost all u.

Firstly, $P_Y(u) = P(Y < u)$

$$= P(X < Y \wedge Y < u \vee X \geq Y \wedge Y < u)$$

$$= P(X < Y \wedge Y < u) + P(X \geq Y \wedge Y < u)$$

$$= P(X < Y)\,P(Y < u \mid X < Y) + P(X \geq Y)\,P(Y < u \mid X \geq Y)$$

$$= P(Y < u \mid X < Y)$$

Secondly, $P_X(u) = P(X < u)$

$$= P(X < Y \wedge Y < u \vee X < u \wedge Y \geq u \vee X < u \wedge X \geq Y)$$

$$= P(X < Y \wedge Y < u) + P(X < u \wedge Y \geq u) + P(X < u \wedge X \geq Y)$$

$$\geq P(X < Y \wedge Y < u)$$

$$= P(X < Y)\,P(Y < u \mid X < Y)$$

$$= P(Y < u \mid X < Y)$$

$$= P_Y(u).$$

The inequality will be strict if $P(X < u \wedge Y \geq u) > 0$.

Since the condition $Y \geq u$ is independent of X this condition is equivalent to $P(X < u)\,P(Y \geq u) > 0$ which is true if and only if $P(X < u) > 0$ and $P(Y \geq u) > 0$. This proves the result.

For the reverse connection we need to make some assumptions. It is not true that if $P_X(u) > P_Y(u)$ then X is stochastically less than Y. It is true, however, that $P_X(u) > P_Y(u)$ implies that there exists a joint density of X and Y with P_X and P_Y as the

marginals for which the result is true. To prove this note that, given a variate X with distribution function $P_X(.)$ we may define a random variable Y such that $P_Y(Y) = P_X(X)$ at points where P_Y is continuous. At points of discontinuity we choose $Y = \inf_{P_Y(Z) \geq P_X(X)} Z$.

This makes Y functionally dependent on X and it is easy to see that $P(X < Y) = 1$.

In the next section we shall use lemmas 1 and 2 to prove some theorems on domination of estimators.

6.6 Ordering Among Estimators

If $S_k \sim \frac{\sigma^2}{n} \chi^2_{n+2k}$ then, for $i < j$, $P_{S_i}(u) > P_{S_j}(u)$ for all u. This is easy to see by examining the density functions. Alternatively we may easily define a joint density for S_i and S_j with P_{S_i} and P_{S_j} as the marginal distribution functions. We do this as follows. Let U_k $k = 1,2,\ldots,j$ be distributed as $\frac{\sigma^2}{n} \chi^2_1$. The joint distribution of S_i and S_j given by $S_i = \sum_{k=1}^{i} U_k$ and $S_j = \sum_{k=1}^{j} U_k$ will then be such that S_i is stochastically less than S_j and S_i and S_j have the required marginal distributions. A similar result applies to $W_k \sim \frac{\sigma^2}{p} \chi^2_{p+2k}(\lambda)$. In this case we take $U_1 \sim \frac{\sigma^2}{p} \chi^2_1(\lambda)$ and $U_k \sim \frac{\sigma^2}{p} \chi^2_1$ $k = 2,3,\ldots,j$.

The next theorem is a generalisation of a formula in Strawderman (1973) which in turn is a generalisation of a formula in Baranchik (1970). Our proof is simpler than that given by Strawderman.

Theorem 1 If $0 \leq v(W,S) \leq 2$ and $v(.,.)$ is a non-decreasing function of the first variable and a non-increasing function of the second then, with the notation of section 6.4, $\delta(X,S)$ is a minimax estimator for ξ.

Proof We need to show that the risk for $\delta(X,S)$ is less than p since the maximum likelihood estimator, whose risk is p, is minimax. In 6.4.9 the second term is negative or zero if $0 \leq v \leq 2$. Since W_{-1} is stochastically less than W the third term is negative or zero if v increases with W. Finally the last term is negative or zero if v is decreasing in S since S_2 is stochastically greater than S_1.

An alternative proof, valid only when the risk estimator exists, is even simpler. From 6.3.1

$$R_\delta(\lambda,\sigma^2) = p - 4\check{c}\,E\left[\frac{1}{F}\left\{\frac{p-2}{4}v(2-v) + W\frac{\partial v}{\partial W} - c\,v\,S\frac{\partial v}{\partial S}\right\}\right]$$

and under the conditions of the theorem each term under the expectation operator is negative or zero.

Corollary 1 (Strawderman's theorem). With the notation of the theorem, if v is written as a function of F and S, $0 \le v \le 2$ and v is non-decreasing in F and non-increasing in S then $\delta(X,S)$ is minimax.

Proof Writing $v(F,S) = v(\frac{W}{S},S)$ we see that, for fixed S, v is increasing in F implies that v is increasing in W; while, for fixed W, v decreases in S under the conditions of Strawderman's theorem. Thus the conditions of Strawdermans theorem imply the conditions of theorem 1 and the result follows. (Again under more stringent conditions, this result is deducible from 6.3.2, i.e. from

$$R_\delta(\lambda,\sigma^2) = p - 4\check{c}\,E\left[\frac{1}{F}\left\{\frac{p-2}{4}v(2-v) + (1+cv)F\frac{v}{F} - c\,v\,S\frac{\partial v}{\partial S}\right\}\right]).$$

Corollary 2 (Baranchik's theorem). If v in the theorem depends only on F, $0 \le v \le 2$ and v is a non-decreasing function of F then $\delta(X,S)$ is minimax.

Proof This clearly follows from Strawderman's theorem as a special case. Alternatively, $v(F) = v(\frac{W}{S})$ is clearly non-decreasing in W and non-increasing in S and the result follows from theorem 1.

Another inequality was given in Stein(1966) and we shall in chapter 7 that it cannot be proved using only the unbiased estimator for the risk. We shall give a slight generalisation of the result as a theorem.

Theorem 2 Let $\delta(X,S) = h(W,S)X$. If $h(.,.)$ is negative on a set of non-zero measure, then, under the assumptions of section 6.4, the estimator $\delta^+(X,S) = h^+(W,S)\,X$, where $h^+(W,S) = h(W,S)$ if $h(W,S) > 0$ and $h^+(W,S) = 0$ if $h(W,S) \le 0$, has smaller risk than $\delta(X,S)$.

Proof In 6.4.1, if h is negative then the first two terms may be reduced by replacing h by zero.

We now wish to find classes of estimators known not to be minimax. One obvious such class is obtained by adapting the conditions of theorem 1. If $v(W,S) \le 0$ or $v(W,S) \ge 2$ and if $v(.,.)$ is non-decreasing in the second variable and non-increasing in the first, then $\delta(X,S)$ is not minimax except in the trivial cases $v(W,S) = 0$ or $v(W,S) = 2$.

Another condition has been given by Efron and Morris(1973a) for the case in which v depends only on W. Their result is that if $v(W)$ is non-decreasing and if $v(W) > 2$ for some W then as $\delta \to \infty$ the risk is greater than p. This result was proved by using a prior density for

λ, $\lambda \sim \frac{\tau^2}{p}\chi^2_p$ (derived from a normal prior distribution for ξ, $\xi \sim N(0,\tau^2)$) and showing that the Bayes risk for this prior is greater than p. This turns out to be equivalent to showing that if $W \sim a\,\chi^2_p$ then, for large a, $E\left[\frac{(p-2)a}{W} v^2(W) - 2v(W)\right] > 0$. Efron and Morris do not prove the latter assertion which in fact holds under more general conditions on $v(.)$. We shall derive the analogous result for the case in which v depends on both W and S under more general conditions. Incidentally, we conjecture that a sufficient condition on $v(.,.)$ for this result to hold is that there exist $\delta > 0$, w_o and s_o such that $v(w,s) > 2+\delta$ for all $w > w_o$ and $s < s_o$. We cannot replace δ by zero since the estimator with $v(w) = w^{-1}(1 + \cos w)$ is not minimax.

In the following theorem we shall give the formula for the Bayes risk under the prior distribution $\lambda \sim \frac{\tau^2}{p}\chi^2_p$.

Theorem 3 If $\lambda \sim \frac{\tau^2}{p}\chi^2_p$ then under the conditions of section 6.4 the Bayes risk of $\delta(X,S)$ is

$$\bar{R}(\tau^2,\sigma^2) = p + \frac{p^2}{p+2\tau^2} E\left[\tilde{c}\, v^2(W_{-1},S_2) - \frac{2p}{p+2\tau^2}\,\frac{W_{-1}}{S_2}\, v(W_{-1},S_2)\right]$$

$$= p + \frac{p^2\tilde{c}}{p+2\tau^2} E\left[v^2(W_{-1},S_2) - 2v(W,S_1)\right]$$

where $W_i \sim \frac{\sigma^2(p+2\tau^2)}{p^2}\chi^2_{p+2i}$, $W = W_1$, $S_i \sim \frac{\sigma^2}{n}\chi^2_{n+2i}$.

Proof We shall show that if $U \sim a\,\chi^2_r(\lambda)$ and $\lambda \sim b\,\chi^2_s$ then the marginal distribution of U is $a(1+2b)\,\chi^2_r$ if $s = r$, and a mixture of $a(1+2b)\,\chi^2_r$ and $a(1+2b)\,\chi^2_{r+2}$ with weights $\frac{1}{1+2b}$ and $\frac{2b}{1+2b}$ if $s = r+2$. In general

$$p(u) = \sum_{k=0}^{\infty} \frac{u^{\frac{1}{2}r+k-1}\exp(-\frac{u}{2a})}{(2a)^{\frac{1}{2}r+k}\Gamma(\frac{1}{2}r+k)} \int_0^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\,\frac{\lambda^{\frac{1}{2}s-1}\exp(-\frac{\lambda}{2b})}{(2b)^{\frac{1}{2}s}\Gamma(\frac{1}{2}s)}$$

$$= \sum_{k=0}^{\infty} \frac{u^{\frac{1}{2}r+k-1}\exp(-\frac{u}{2a})\,\Gamma(\frac{1}{2}s+k)\,(\frac{2b}{2b+1})^{\frac{1}{2}s+k}}{(2a)^{\frac{1}{2}r+k}\,\Gamma(\frac{1}{2}r+k)\,k!\,(2b)^{\frac{1}{2}s}\,\Gamma(\frac{1}{2}s)}$$

$$= \frac{u^{\frac{1}{2}r-1}\exp(-\frac{u}{2a})}{(2a)^{\frac{1}{2}r}\,\Gamma(\frac{1}{2}r)(1+2b)^{\frac{1}{2}s}} \sum_{k=0}^{\infty} \frac{(\frac{1}{2}s)_k}{(\frac{1}{2}r)_k}\,\frac{(\frac{2bu}{2a(1+2b)})^k}{k!}$$

$$= \frac{u^{\frac{1}{2}r-1}\exp(-\frac{u}{2a})}{(2a)^{\frac{1}{2}r}\,\Gamma(\frac{1}{2}r)(1+2b)^{\frac{1}{2}s}}\; {}_1F_1(\tfrac{1}{2}s;\tfrac{1}{2}r;\frac{2bu}{2a(1+2b)}).$$

In the special case $s = r$ this reduces to

$$p(u) = \frac{u^{\frac{1}{2}r-1}\exp(-\frac{u}{2a})}{\{2a(1+2b)\}^{\frac{1}{2}r}\Gamma(\frac{1}{2}r)}\exp\left(\frac{2bu}{2a(1+2b)}\right) = \frac{u^{\frac{1}{2}r-1}\exp(-\frac{u}{2a(1+2b)})}{\{2a(1+2b)\}^{\frac{1}{2}r}\Gamma(\frac{1}{2}r)},$$

while when $s = r+2$ the expression gives

$$p(u) = \frac{u^{\frac{1}{2}r-1}\exp(-\frac{u}{2a})}{\{2a(1+2b)\}^{\frac{1}{2}r}\Gamma(\frac{1}{2}r)(1+2b)} \sum_{k=0}^{\infty} \frac{\frac{1}{2}r+k}{\frac{1}{2}r} \frac{\left(\frac{2bu}{2a(1+2b)}\right)^k}{k!}$$

$$= \frac{1}{1+2b}\frac{u^{\frac{1}{2}r-1}\exp(-\frac{u}{2a(1+2b)})}{\{2a(1+2b)\}^{\frac{1}{2}r}\Gamma(\frac{1}{2}r)} + \frac{2b}{1+2b}\frac{u^{(\frac{1}{2}r+1)-1}\exp(-\frac{u}{2a(1+2b)})}{\{2a(1+2b)\}^{\frac{1}{2}r+1}\Gamma(\frac{1}{2}r+1)}.$$

These are the required results.

From 6.4.9 we obtain

$$\bar{R}(\tau^2,\sigma^2) = p + (p-2)\sigma^2 E\left[\frac{\tilde{c}}{W}\{v^2(W,S_2)-2v(W,S_1)\} + \frac{2}{S_2}\{v(W_{-1},S_2)- v(W,S_2)\}\right]$$

where the expectation is with respect to the distributions

$W_i \sim \frac{\sigma^2}{p}\chi^2_{p+2i}(\lambda)$, $S_i \sim \frac{\sigma^2}{n}\chi^2_{n+2i}$ and $\lambda \sim \frac{\tau^2}{p}\chi^2_p$. Putting $a = \frac{\sigma^2}{p}$ $s = p$, $b = \frac{\tau^2}{p}$ and $r = p+2i$ $i = 0,-1$ in the previous result gives

$$\bar{R}(\tau^2,\sigma^2) = p + (p-2)\sigma^2 E\left[\frac{\tilde{c}}{W}\{v^2(W,S_2)- 2v(W,S_1)\}-\frac{2}{S_2}v(W,S_2) + \frac{2}{S_2}\left\{\frac{p}{p+2\tau^2}v(W_{-1},S_2)+\frac{2\tau^2}{p+2\tau^2}v(W,S_2)\right\}\right]$$

$$= p + (p-2)\sigma^2 E\left[\frac{\tilde{c}}{W}\{v^2(W,S_2)-2v(W,S_1)\} + \frac{2p}{p+2\tau^2}\frac{1}{S_2}\{v(W_{-1},S_2) -v(W,S_2)\}\right]$$

where the expectation is with respect to

$$W \sim \frac{\sigma^2(p+2\tau^2)}{p^2}\chi^2_p, \quad W_{-1} \sim \frac{\sigma^2(p+2\tau^2)}{p^2}\chi^2_{p-2} \quad \text{and} \quad S_1 \text{ and } S_2.$$

Applying 6.2.9 to this expression gives

$$\bar{R}(\tau^2,\sigma^2) = p + (p-2)\sigma^2 E\left[\frac{\tilde{c}}{W}v^2(W,S_2) - \frac{2(n+2)\tilde{c}\sigma^2}{nWS_2}v(W,S_2) + \frac{2(p-2)\sigma^2}{pWS_2}v(W,S_2) - \frac{2}{S_2}\frac{p}{p+2\tau^2}v(W,S_2)\right]$$

$$= p + (p-2)\sigma^2 E\left[\frac{\tilde{c}}{W}v^2(W,S_2) - \frac{2}{S_2}\frac{p}{p+2\tau^2}v(W,S_2)\right]$$

$$= p + E\left[\frac{\tilde{c}p^2}{p+2\tau^2}v^2(W_{-1},S_2) - \frac{2p^3}{(p+2\tau)^2}\frac{W_{-1}}{S_2}v(W_{-1},S_2)\right]$$

$$= p + E\left[\frac{\tilde{c}p^2}{p+2\tau^2}v^2(W_{-1},S_2) - \frac{2\tilde{c}p^2}{p+2\tau^2}v(W,S_1)\right].$$

We now wish to find sufficient conditions for

$$E[v^2(W_{-1},S_2) - 2v(W,S_1)] = E\left[v^2(W_{-1},S_2) - \frac{2p}{(p+2\tau^2)\tilde{c}}\frac{W_{-1}}{S_2}v(W_{-1},S_2)\right]$$

to be positive. The following two theorems give sufficient conditions for this to be so.

Theorem 4 If $W_{-1} \sim \frac{a}{p}\chi^2_{p-2}$, $S_2 \sim \frac{b}{n}\chi^2_{n+4}$ and $V = v(W_{-1},S_2)$ then a sufficient condition that $E = E[V^2 - \frac{2b}{ac}\frac{W_{-1}}{S_2}V] > 0$ for sufficiently large a and small b is that there exists $\delta > 0$ such that $v(w,s) > 1 + \frac{1}{\sqrt{c}} + \delta$ for sufficiently large w and small s.

Proof

$$E = E[V^2 - 2V - 2(\frac{b}{ac}\frac{W_{-1}}{S_2} - 1)V]$$

$$= \text{var } V + E[V](E[V]-2) - \frac{2b}{ac}\text{cov}(\frac{W_{-1}}{S_2}, V)$$

$$> \text{var } V + E[V](E[V]-2) - \frac{2b}{ac}\sqrt{\text{var}\frac{W_{-1}}{S_2}\ \text{var } V}$$

$$= \text{var } V + E[V](E[V]-2) - 2\sqrt{\frac{2(n+p)}{p(n+2)c}}\sqrt{\text{var } V}.$$

Now if $E[V] > 2$ then the above expression is positive if and only if

$$(\text{var } V)^2 + 2\{E[V](E[V]-2) - \frac{4(n+p)}{n(p-2)}\}\text{var } V + \{E[V](E[V]-2)\}^2 > 0.$$

This holds for all values of var V if

$$\left(\frac{4(p+n)}{n(p-2)}\right)^2 < \frac{8(p+n)}{n(p-2)}E[V](E[V]-2)$$

which, when $E[V] > 2$, holds if and only if

$$E[V] > 1 + \sqrt{1+\frac{2(p+n)}{n(p-2)}} = 1 + \sqrt{\frac{p(n+2)}{n(p-2)}} = 1 + \frac{1}{\sqrt{c}}.$$

We must now show that if $v(w,s) > 1 + \frac{1}{\sqrt{c}} + \delta$ for large enough w and small enough s then $E[V] > 1 + \frac{1}{\sqrt{c}}$ for large enough a and small enough b. Let $l = 1 + \frac{1}{\sqrt{c}}$. Given $\varepsilon > 0$ there are x_o, y_o such that $\int_0^{ax_o} p_{W_{-1}}(x)\,dx < \varepsilon$ and $\int_{by_o}^{\infty} p_{S_2}(y)\,dy < \varepsilon$ and there are x_1, y_1 such that $v(w,s) > l + \delta$ for $w > x_1$ and $s > y_1$. Suppose $V \geq 0$. We then have, if $a > \frac{x_1}{x_o}$ and $b < \frac{y_1}{y_o}$

$$E[V] = l + \delta + E[V-l-\delta]$$

$$> l + \delta + \int_0^{x_1}\int_{y_1}^{\infty}(v(x,y)-l-\delta)p_{W_{-1}}(x)p_{S_2}(y)\,dx\,dy$$

$$> l + \delta + \int_0^{ax_o}\int_{by_o}^{\infty}(v(x,y)-l-\delta)p_{W_{-1}}(x)p_{S_2}(y)\,dx\,dy$$

$$> l + \delta + (l+\delta)\{1 - (1-\varepsilon)^2\}$$

$$= (l+\delta)(1-\varepsilon)^2$$

$$> (l+\delta)(1-2\varepsilon)$$

and this is greater than ℓ if and only if $\varepsilon < \frac{\delta}{2(\ell+\delta)}$. Now if $V < 0$ on some set of non-zero measure then replacing V by V^+ increases E. We may apply the above result to V^+. This completes the proof.

In our application of this theorem we put $a = \sigma^2\left(\frac{p+2\tau^2}{p}\right)$ and $b = \sigma^2$. We then choose b small enough first and then τ^2 may be chosen so that a is sufficiently large.

Our other theorem concerns the case in which $v(w,s) \to \ell < \varepsilon$ as $w \to \infty$ and $s \to 0$.

Theorem 5 Let $W_{-1} \sim \frac{a}{p}\chi^2_{p-2}$, $S_2 \sim \frac{b}{n}\chi^2_{n+4}$, $W \sim \frac{a}{p}\chi^2_p$ and $S_1 \sim \frac{b}{n}\chi^2_{n+2}$. Let $V = v(W_{-1},S_2)$ and $V_1 = v(W,S_1)$. If $v(w,s) \to \ell$ where $0 < \ell < \infty$ as $w \to \infty$ and $s \to 0$ and if $v(w,s)$ is bounded above by ℓ_1 then for large enough a and small enough b, $E = E[V-2V_1] < 0$.

Proof Given $\varepsilon < 0$ there are x_o, y_o such that

$$\int_0^{ax_o} p_W(x)\,dx < \int_0^{ax_o} p_{W_{-1}}(x)\,dx < \varepsilon$$

and $$\int_{by_o}^{\infty} p_{S_1}(y)\,dy < \int_{by_o}^{\infty} p_{S_2}(y)\,dy < \varepsilon$$

and there are x_1, y_1 such that $\ell-\varepsilon < v(x,y) < \ell+\varepsilon$ if $x > x_1$ and $y < y_1$. Suppose that $a > \frac{x_1}{x_o}$ and $b < \frac{y_1}{y_o}$ then

$$\begin{aligned}
E &= (\ell-\varepsilon)^2 - 2(\ell+\varepsilon) + E[V - (\ell-\varepsilon)^2 + 2(\ell+\varepsilon) - 2V_1] \\
&> (\ell-\varepsilon)^2 - 2(\ell+\varepsilon) + \int_0^{x_1}\int_{y_1}^{\infty}\{v^2(x,y)-(\ell-\varepsilon)^2\}p_{W_{-1}}(x)\,p_{S_2}(y)\,dx\,dy \\
&\qquad + 2\int_0^{x_1}\int_{y_1}^{\infty}\{\ell+\varepsilon-v(x,y)\}p_W(x)\,p_{S_1}(y)\,dx\,dy \\
&> (\ell-\varepsilon)^2 - 2(\ell+\varepsilon) + \int_0^{ax_o}\int_{by_o}^{\infty}\{v^2(x,y)-(\ell-\varepsilon)^2\}p_{W_{-1}}(x)\,p_{S_2}(y)\,dx\,dy \\
&\qquad + 2\int_0^{ax_o}\int_{by_o}^{\infty}\{\ell+\varepsilon-v(x,y)\}p_W(x)\,p_{S_1}(y)\,dx\,dy \\
&> (\ell-\varepsilon)^2 - 2(\ell+\varepsilon) - \varepsilon(\ell-\varepsilon)^2 + 2\varepsilon(\ell-\ell_1+\varepsilon) \\
&> \ell^2 - 2\varepsilon\ell - 2\ell - 2\varepsilon - \varepsilon\ell^2 + 2\varepsilon\ell - 2\varepsilon\ell_1 \qquad \text{if } \varepsilon < \ell \\
&= \ell^2 - 2\ell - \varepsilon(\ell^2 + 2\ell_1 + 2).
\end{aligned}$$

If $\varepsilon < \frac{\ell^2-2\ell}{\ell^2+2\ell_1+2}$ then this expression is positive.

As in the application of theorem 4 we shall put $a = \sigma^2\left(\frac{p+2\tau^2}{p}\right)$

and $b = \sigma^2$.

We shall now summarise these results in a theorem.

Theorem 6 Under the definitions of section 6.4, the estimator $\delta(X,S)$ is not minimax if $v(w,s)$ converges to a limit greater than two as $w \to \infty$ and $s \to 0$ or if $v(w,s) > 1 + \frac{1}{\sqrt{c}} + \delta$ for some $\delta > 0$ for large w and small s.

We remark that there are analogues of theorems 4 and 5 as $a \to 0$ and as $b \to \infty$ for the case in which $v(w,s) < 0$ for small w and large s. However these do not help in a non-minimaxity proof since $a = \sigma^2 \frac{p+2\tau^2}{p}$ and as $\tau^2 \to 0$ $a \to \sigma^2$ which does not approach zero. However, we can state that if $v(w,s) < 0$ or if $v(w,s) > \frac{w}{\tilde{c}s}$ then the Bayes risk can be reduced. The first assertion can be seen from the form of the Bayes risk in theorem 3; the second from the fact that in this case $h(w,s) < 0$ and by theorem 2, the risk can be reduced.

6.7 Risk Functions for some Special Families of Estimators

A special case of the family of estimators in this chapter is given by $v(w,s) = t$. The risk function will be

$$R(\lambda,\sigma^2) = p + p(p-2)\tilde{c}\,t(t-2)\,E\left[\frac{1}{p-2+2K}\right]$$

where K has a Poisson distribution with parameter λ. This result is given in James and Stein(1960) and easily follows from 6.4.14. The case $t = 1$ achieves minimum risk and is the James-Stein estimator. Values of t between 0 and 2 achieve minimaxity. The value $t = 0$ gives the maximum likelihood estimator, while $t = 2$ gives the same risk. Efron and Morris(1973a) show that in order to dominate the James-Stein estimator the conditions of theorem must be violated.

Another class of estimators known to be minimax when $0 \leq t \leq 2$ is the class of estimators given by $v(w,s) = t$ if $w > t\tilde{c}s$ and $v(w,s) = \frac{w}{\tilde{c}s}$ if $w \leq t\tilde{c}s$. This is the class of positive part James-Stein estimators and its members dominate the corresponding estimator with $v(w,s) = t$ by virtue of theorem 2. These estimators are non-comparable for different values of t as can be seen by inspecting their risk functions for small and large values of λ.

We shall now consider an extended family of estimators. let $\delta(X,S) = (1 - \frac{\tilde{c}\alpha}{d+F})X$. The special case $d = 0$, $e = 1$ is the James-Stein estimator already considered. The case $d = \tilde{c}$ gives the estimator $\delta(X,S) = \frac{F}{\tilde{c}+F}X$ given by Alam and Thompson(1964). The

The risk function for estimators of this class can be written in terms of hypergeometric functions of two variables and is given in the next theorem.

Theorem 7 The risk of the estimator $\delta(X,S) = (1 - \frac{\tilde{c}\alpha}{d+F})X$ is

$$R(\lambda,\sigma^2) = p + \frac{p\tilde{c}^2\alpha^2}{d}\{e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{(a+k)_2}{a+b+2+k}\,{}_2F_1(1,b+1;a+b+3+k;1-\tfrac{a}{b}d)$$

$$- e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{(a+k)^2}{a+b+1+k}\,{}_2F_1(1,b+1;a+b+2+k;1-\tfrac{a}{b}d)\}$$

$$- 2p\tilde{c}\alpha\{e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{a+k}{a+b+1+k}\,{}_2F_1(1,b+1;a+b+2+k;1-\tfrac{a}{b}d)$$

$$- e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{k}{a+b+k}\,{}_2F_1(1,b+1;a+b+1+k;1-\tfrac{a}{b}d)$$

where $a = \frac{1}{2}p$ and $b = \frac{1}{2}n$.

Proof From 6.4.8 we have

$$\frac{R-p}{p} = \tilde{c}^2 E_{0,1}\left[\frac{v^2}{F}\right] + 2\tilde{c}\,E_{-1,1}[v] - 2\tilde{c}\,E_{0,1}[v] - 2\frac{p-2}{p}\tilde{c}\,E_{0,0}\left[\frac{v}{F}\right]$$

where $v = \alpha\frac{F}{d+F}$. Now $\frac{v}{F} = \alpha\frac{1}{d+F}$ and $\frac{v^2}{F} = \alpha^2\frac{F}{(d+F)^2}$

$$= \frac{\alpha^2}{d}\frac{F}{d+F} - \frac{\alpha^2}{d}\frac{F^2}{(d+F)^2}.$$

Using the formula in appendix 2

$$E\left[\frac{X^m}{(\gamma+X)^n}\right] = e^{-\lambda}(b)_{n-m}\sum_{k=0}^{\infty}\frac{(a+k)_m}{(a+b+k)_n}\,{}_2F_1(n,b+n-m;a+b+n+k;1-\gamma)$$

where X has a non-central inverse beta distribution $\beta_2(a,b,\lambda)$ we obtain

$$\frac{R-p}{p} = \frac{\tilde{c}^2\alpha^2}{d}\{e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{a+k}{a+b+1+k}\,{}_2F_1(1,b+1;a+b+2+k;1-\tfrac{a}{b}d)$$

$$- e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{(a+k)_2}{(a+b+1+k)_2}\,{}_2F_1(2,b+1;a+b+3+k;1-\tfrac{a}{b}d)\}$$

$$- 2\tilde{c}\alpha\{e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{a+k}{a+b+1+k}\,{}_2F_1(1,b+1;a+b+2+k;1-\tfrac{a}{b}d)$$

$$- e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{a-1+k}{a+b+k}\,{}_2F_1(1,b+1;a+b+1+k;1-\tfrac{a}{b}d)$$

$$+ \frac{a-1}{a}\frac{a}{b}\,b\,e^{-\lambda}\sum_{k=0}^{\infty}\frac{1}{a+b+k}\,{}_2F_1(1,b+1;a+b+1+k;1-\tfrac{a}{b}d)\}.$$

Now using the recurrence relation for the hypergeometric function

$$ {}_2F_1(2,b+1;a+b+3+k;1 - \frac{a}{b}\,d) = (a+b+2+k)\ {}_2F_1(1,b+1;a+b+2+k;1 - \frac{a}{b}\,d) $$
$$ - (a+b+1+k)\ {}_2F_1(1,b+1;a+b+1+k;1 - \frac{a}{b}\,d) $$

and simplifying gives the result.

As a check we shall show that the James-Stein risk may be obtained from this by letting d tend to zero. Using Gauss's theorem we see that the term in $\frac{1}{d}$ is of the form $\frac{0}{0}$ in the limit so we apply de l'Hôpital's rule. The derivative of a hypergeometric function is another one. Applying Gauss's theorem to the derivative gives

$$ R(\lambda,\sigma^2) = p + p\,\tilde{c}\,\alpha a\ \tilde{c}\,\alpha \frac{b+1}{b} - 2\,\frac{a-1}{a}\ e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{a-1+k} \frac{\lambda^k}{k!} $$

which is the result obtained earlier. Note that the minimum occurs at $\alpha = 1$ giving

$$ R(\lambda,\sigma^2) = p - (p-2)\,\frac{n}{n+2}\ e^{-\lambda} \sum_{k=0}^{\infty} \frac{a-1}{a-1+k} \frac{\lambda^k}{k!}\ . $$

Unless d is very small, in which case the algorithm in theorem 7 may be numerically unstable, this algorithm gives an efficient way to evaluate the risk. Using a recurrence relation for the hypergeometric function allows us to calculate each function $F(1,b+1;a+b+i+k;1 - \frac{a}{b}d)$ from the previous two values. Thus, in theory, only the values $F(1,b+1;a+b+1;1 - \frac{a}{b}d)$ and $F(1,b+1;a+b+2;1 - \frac{a}{b}\,d)$ need be calculated using the series expansion. In practice, errors tend to accumulate and it is better to calculate new values every about twenty terms. In addition it is possible to use a recurrence relation to calculate the hypergeometric functions for different numbers of degrees of freedom (differing by an even integer) but this is inefficient in storage and the saving in computation is small.

Chapter 7
Risk Estimate Optimality of Shrunken Estimators

7.1 Introduction

The unbiased estimator for the risk which we discussed in chapter 6 leads to an optimality property which,in some cases, is easier to handle mathematically than admissibility. If $\delta_1(X,S)$ and $\delta_2(X,S)$ are two estimators for which unbiased risk estimators exist, and if the risk estimator for $\delta_1(.,.)$ is uniformly less than that for $\delta_2(.,.)$ then the estimator $\delta_1(.,.)$ dominates the estimator $\delta_2(.,.)$ in terms of risk estimate and hence also in terms of risk. If no estimator dominates δ_1 in terms of risk estimate then δ_1 is said to be *risk estimate optimal.* Efron and Morris(1976) showed that a certain class of estimators, whose minimaxity was not previously known, was in fact a class of minimax estimators. This was done by using the concept of risk estimate dominance. Thus the risk estimate is useful for proving certain dominance results, although, as we shall see, we cannot prove all such results this way. In particular we shall show that the James-Stein estimator is optimal in terms of risk estimate so that the positive part version, which dominates it in terms of risk, cannot be shown to dominate it using the risk estimate alone. We shall also show that the positive part version of the James-Stein estimator is risk estimate optimal. These results, which first appeared in Moore and Brook(1978), will only be shown rigorously to be true in the class of scale invariant spherically symmetric estimators.

First we shall discuss the general problem of risk estimate dominance in the class of estimators for which we have previously calculated the unbiased risk estimator.

7.2 Risk Estimate Dominance

Suppose $X \sim N_p(\xi,\sigma^2 I)$ and $S \sim \frac{\sigma^2}{n}\chi^2_n$ independently of X. We wish to compare the estimators

$$\xi^*(X,S) = \left(1 - \frac{\tilde{c}(1+w(W,S))}{F}\right) X$$

and

$$\xi^*_f(X,S) = \left(1 - \frac{\tilde{c}(1+f(W,S))}{F}\right) X$$

by comparing their risk estimators (which we assume to exist). Using the transformation $U = W^{\alpha}S^{\beta}$, $V = W^{\gamma}S^{\delta}$ of section 6.3 and using the expression in that section for the unbiased estimator for the risk, we find that the difference between the risk estimates for ξ^* and ξ^*_f

is proportional to

$$\frac{p-2}{4}\,\Delta_f = \hat{r}_{\xi^*}(U,V) - \hat{r}_{\xi^*_f}(U,V)$$

$$= \frac{1}{F}\Big\{\frac{p-2}{4}(f^2-w^2) + (\alpha-\beta c-\beta c w)\,U\frac{\partial w}{\partial U} - (\alpha-\beta c-\beta f)\,U\frac{\partial f}{\partial U}$$

$$+ (\gamma-\delta c-\delta c w)\,V\frac{\partial w}{\partial V} - (\gamma-\delta c-\delta c f)\,V\frac{\partial f}{\partial V}\Big\}.$$

We shall write $a = \frac{4}{p-2}$ and $b = \frac{4}{n+2}$ in which case

$$(1)\quad \Delta_f = \frac{1}{F}(f^2-w^2) + (\alpha a-\beta b-\beta b w)\,U\frac{\partial w}{\partial U} - (\alpha a-\beta b-\beta b f)\,U\frac{\partial f}{\partial U}$$

$$+ (\gamma a-\delta b-\delta b w)\,V\frac{\partial w}{\partial V} - (\gamma a-\delta b-\delta b f)\,V\frac{\partial f}{\partial V}.$$

In order for ξ^* to dominate ξ^*_f in terms of risk estimate we require that $\Delta_f \geq 0$, the condition being necessary and sufficient. We are thus lead to the problem of solving the differential inequality

$$(2)\quad (A+Bw)\,U\frac{\partial w}{\partial U} + (C+Dw)\,V\frac{\partial w}{\partial V} - w^2 = g(f) + \psi^2,\quad \psi \geq 0$$

where $g(f) = (A+Bf)\,U\frac{\partial f}{\partial U} + (C+Df)\,V\frac{\partial w}{\partial V} - f^2$

and $A = \alpha a-\beta b$, $B = -\beta b$, $C = \gamma a-\delta b$ and $D = -\delta b$.

The characteristic equations of (2) are

$$(3)\quad \frac{dU}{U(A+Bw)} = \frac{dV}{V(C+Dw)} = \frac{dw}{w^2+g(f)+\psi}.$$

Any given function ψ may be written in terms of U, V and w since f is assumed to be known. In this case, the condition for (3) to have an integrating factor depending only on w given in Ince(1963) is

$$(4)\quad (A+Bf)\,U\frac{\partial f}{\partial U} + (C+Df)\,V\frac{\partial f}{\partial V} - f^2 = -t^2,\quad t^2 \text{ real.}$$

We can show that t^2 is positive, otherwise the estimator ξ^*_f would dominate the James-Stein estimator in terms of risk estimate and this we shall show to be impossible.

In particular (4) is satisfied by $f = t$, and any estimator ξ^*_f with f satisfying (4) has the same risk as ξ^*_t (where ξ^*_t is the estimator ξ^*_f with $f = t$ = constant). This means that it is easier to compare an estimator with ξ^*_t than with any other (apart from ξ^*_f where f satisfies (4)). In particular we may compare an estimator with the maximum likelihood estimator, ξ^*_{-1} or with the James-Stein estimator, ξ^*_o.

Let us choose α, β, γ and δ so that $U = W$ and $V = S$. We therefore put $\alpha = 1$, $\beta = 0$, $\gamma = 0$, $\delta = 1$ obtaining $A = a$, $b = 0$, and $C = D = -b$ giving the partial differential equation for w

(5) $a\,W\,\frac{\partial w}{\partial W} - b(1+w)\,S\,\frac{\partial w}{\partial S} = w^2 - t^2 + \psi,\ \psi \geq 0$

which has characteristic equations

$$\frac{dW}{aW} = -\frac{dS}{bS(1+w)} = \frac{dw}{w^2-t^2+\psi}.$$

Suppose that there are W_o, S_o such that $w_o = w(W_o,S_o) > t$. In this case as W increases, w increases and S decreases along a characteristic through (W_o,S_o,w_o). For a positive increment dW in W, $dS < 0$ and

$$\frac{dW}{aW} \leq \frac{dw}{w^2-t^2}, \qquad -\frac{dS}{bS} \leq \frac{1+w}{w^2-t^2}\,dw.$$

Assuming $t \neq 0$ and integrating these inequalities from (W_o,S_o,w_o) in the direction of increasing W gives

$$\frac{1}{a}\log\frac{W}{W_o} \leq \frac{1}{2|t|}\log\frac{w-|t|}{w+|t|}\,\frac{w_o+|t|}{w_o-|t|} \qquad \text{and}$$

$$\frac{1}{b}\log\frac{S_o}{S} \leq \frac{1}{2|t|}\log\frac{w-|t|}{w+|t|}\,\frac{w_o+|t|}{w_o-|t|} + \frac{1}{2}\log\frac{w^2-t^2}{w_o^2-t^2}$$

that is

$$\frac{w-|t|}{w+|t|} \geq \frac{w_o-|t|}{w_o+|t|}\left(\frac{W}{W_o}\right)^{\frac{2|t|}{a}} \qquad \text{and}$$

$$\left(\frac{S_o}{S}\right)^{\frac{2|t|}{b}} \leq \left(\frac{w-|t|}{w_o-|t|}\right)^{|t|+1}\left(\frac{w+|t|}{w_o+|t|}\right)^{|t|-1}.$$

The first equation shows that

$$w \to \infty \quad \text{as} \quad W \to W_2 < W_o\left(\frac{w_o+|t|}{w_o-|t|}\right)^{\frac{a}{2|t|}}$$

and the second equation shows that w is unbounded as $S \to 0$ along a characteristic. They also give bounds on the rate of convergence

$$\frac{w}{|t|} \geq \left\{1 + \frac{w_o-|t|}{w_o+|t|}\left(\frac{W}{W_o}\right)^{\frac{2|t|}{a}}\right\}\Big/\left\{1 - \frac{w_o-|t|}{w_o+|t|}\left(\frac{W}{W_o}\right)^{\frac{2|t|}{a}}\right\}$$

and $w^2 > t^2 + (w_o^2 - t^2)\left(\frac{S_o}{S}\right)^{\frac{2}{b}}$. The case $t = 0$ is also easily solved but we omit the result which is qualitively the same.

Given $M > 0$ we can find a point (W_1,S_1,w_1) on the characteristic for which $w_1 > M$. Consider the level curve through this point. Its equations are (5) and

(6) $0 = dw = \frac{\partial w}{\partial W}\,dW + \frac{\partial w}{\partial S}\,dS.$

From (5) and (6) we find that

$$a W \frac{\partial w}{\partial W} + b(1+w) S \frac{\partial w}{\partial W} \frac{dW}{dS} = w^2 - t^2 + \psi .$$

If $\frac{dW}{dS} > -\frac{a}{b(1+w)} \frac{W}{S}$ then $\frac{\partial w}{\partial W} > 0$, otherwise $\frac{\partial w}{\partial W} < 0$. However if $\frac{\partial w}{\partial W} < 0$ then there is a level curve between the characteristic (along which w is increasing) and the line $S = S_1$ so that $\frac{\partial w}{\partial W} > 0$. We have thus shown that $\frac{\partial w}{\partial W} > 0$ and, therefore, that there is a region in the W, S plane for which $w > M$ given by the characteristic base curve through (W_1,S_1), the part of the W axis for which $W > W_2$ and the part of the line $S = S_1$ for which $W > W_1$. This implies that $w > M$ in the region $S < S_1$, $W > W_2$. Thus, by theorem 6.6.6 the estimator ξ^* is not minimax. Now if $|t| \leq 1$ then ξ^* has smaller risk estimate than the maximum likelihood estimator and is therefore minimax. Thus ξ^* cannot dominate ξ^*_t in terms of risk estimate (or in terms of risk).

We have thus shown that, for $|t| \leq 1$, if $w > |t|$ at some point then ξ^* does not dominate ξ^*_t in terms of risk estimate.

If we wish to give a similar argument for the case $|t| > 1$ then we cannot use theorem 6.6.6. Since this is not an important special case we shall not give a precise argument in the general case in which w depends on both F and S. We shall give a precise argument for the spherically symmetric case in the next section. A rough argument is as follows. It is easily seen that the characteristics meet the W axis at right angles. Therefore, close to the axis, the characteristic may be replaced, approximately, by its tangent at $S = 0$. The solution of the inequality for S shows that, with the characteristic replaced by the tangent, S increases too quickly for the convergence of the integral of w^2 with respect to S. Since this occurs for each value of W greater than W_2 the double integral cannot converge. Since a condition for the existence of the risk estimate is that this integral should converge the result follows by contradiction. We believe that this argument can be made precise.

We should also like to show that the existence of a point for which $w < -|t|$ leads to a contradiction but we have had less success. Theorem 6.6.6 does not help even when $|t| < 1$ but the behaviour of w along the characteristics through the points for which $w < -|t|$ suggests that the double integral for w^2 cannot converge.

In the next section precise arguments shall be given for the class of scale invariant estimators, that is in the case for which w depends only on F.

7.2.1 Risk Estimate Dominance in the Class of Scale Invariant Estimators

When w depends on F alone the inequality to be solved is

$$(1)\quad (a+b+b\,w)\,F\,\frac{dw}{dF} = w^2 - t^2 + \psi\,, \qquad \psi \geq 0.$$

Now, in order that the expectation of w^2 shall not exist, it is sufficient that $w^{-1}(F) = O(F^{-1/b})$ as $F \to \infty$ or that $w^{-1}(F) = O(F^{1/a})$ as $F \to 0$. If there is a point F_o for which $w_o = w(F_o) > |t|$ then $\left.\frac{dw}{dF}\right|_{F_o} > 0$ and therefore $\frac{dw}{dF} > 0$ for $F > F_o$. It then follows from (1) that

$$(a+b+b\,w)\,F\,\frac{dw}{dF} \geq w^2 - t^2$$

for $F > F_0$. Integrating we see that

$$\frac{F}{F_o} \leq \left(\frac{w-|t|}{w_o-|t|}\,\frac{w_o+|t|}{w+|t|}\right)^{\frac{a+b}{2|t|}} \left(\frac{w^2-t^2}{w_o^2-t^2}\right)^{\frac{1}{2}b} \qquad \text{for} \quad F > F_o.$$

This means that w is unbounded as $F \to \infty$ and therefore that $w^{-1}(F) = O(F^{-1/b})$ as $F \to \infty$. This contradicts the existence of the expectation of w^2 and shows that there does not exist F_o such that $w(F_o) > |t|$. Our solutions assumed that $t \neq 0$. It is easy to see that the same result applies if $t = 0$.

We now wish to see what happens if there exist values of w which are less than $-|t|$. There are two possible cases if $|t| < 1 + \frac{a}{b}$

(i) $\exists\, F_o$ such that $-1 - \frac{a}{b} < w(F_o) < -|t|$

(ii) $\exists\, F_o$ such that $-1 - \frac{a}{b} > w(F_o)$.

If $t \geq 1 + \frac{a}{b}$ then only the second case may occur.

In the second case we may argue as previously that $w \to -\infty$ as $F \to \infty$ and that $-w^{-1}(F) = O(F^{-1/b})$ as $F \to \infty$ (only the signs of w and $\frac{dw}{dF}$ are changed in this argument).

In the first case, w is increasing. We integrate from $F < F_o$ to F_o and obtain

$$\frac{F_o}{F} \leq \left(\frac{w_o-|t|}{w-|t|}\,\frac{w+|t|}{w_o+|t|}\right)^{\frac{a+b}{2|t|}} \left(\frac{w_o^2-t^2}{w^2-t^2}\right)^{\frac{1}{2}b} \qquad \text{for} \quad F < F_o.$$

The inequality reverses if w reaches the value $-1-\frac{a}{b}$ as $F \to \infty$ decreases. This must occur, for, with the direction of the inequality unchanged the value of w is unbounded below as $F \to 0$. Let $w(F_1) = -1-\frac{a}{b}$. For $F < F_1$ we cannot continue the solution curve continuously so case (i) is impossible.

Thus, in order to dominate the shrunken estimator ξ^*_t in terms of risk estimate, we must use ξ^*_w where $-|t| \le w \le |t|$. In the case of the James-Stein estimator, $t = 0$ and $w = 0$ is the only solution. Therefore no estimator can dominate the James-Stein estimator in risk estimate. However, we have already seen that the positive part version of the James-Stein estimator dominates the unmodified version. Its risk estimate must be sometimes greater and sometimes less.

We now turn our attention to the positive part estimator. We prove part of the result for ξ^*_f in general. In order to dominate ξ^*_f in risk estimate we require that

$$(a+b+b\,w)\,F\,\frac{dw}{dF} = (a+b+b\,f)\,F\,\frac{df}{dF} + w^2 - f^2 + \psi, \qquad \psi \ge 0.$$

If $w^2 > f^2$ then this implies that

$$(a+b+b\,w)\frac{dw}{dF} > (a+b+b\,f)\frac{df}{dF}.$$

Therefore, if $w_o > f_o > 0$ at F_o, then, for $F > F_o$

$$(a+b+b\,w)^2 - (a+b+b\,f)^2 > (a+b+b\,w_o)^2 - (a+b+b\,f_o)^2.$$

It is easy to see that $w - f \to 0$ as $F \to \infty$ is only possible if $f \to \infty$ as $F \to \infty$. If f is bounded then $\inf_{F>F_o}(w-f) > 0$. Now, if $-w < -|f|$ and $a+b+b\,w > 0$ then

$$\frac{dw}{dF} > \frac{a+b+b\,f}{a+b+b\,w}\frac{df}{dF} > \frac{df}{dF}$$

and so, if $-w_o < -f_o$ at F_o then $f - w$ increases as F decreases below F_o. Finally, if $a+b+b\,w < 0$, $a+b+b\,f > 0$ and $\frac{df}{dF} > 0$ then $\frac{dw}{dF} < 0$.

Applying these results to the estimator $\xi^{*+}_t = \xi^*_f$

where $f(F) = t$ if $F \ge \tilde{c}(1+t)$

$= \frac{1}{\tilde{c}}F - 1$ $F \le \tilde{c}(1+t)$

we see that, if $w > |t|$ for $F_o \ge \tilde{c}(1+t)$, then $w > |t|$ for all $F > F_o$; that, if $w_o = w(F_o) < \min(-|t|, \frac{1}{\tilde{c}}F - 1)$ and $a+b+b\,w_o > 0$ then for some $F_1 < F_o$, $w(F_1) < -1$ and, if $w_o = w(F_o) < \min(-|t|, \frac{1}{\tilde{c}}F - 1)$ and $a+b+b\,w_o < 0$ then for all

$F > F_o$, $w(F) < w(F_o)$. It now follows from the previous section that, for $F \geq \tilde{c}(1+t)$, we need $w^2 > t^2$ in order to dominate the estimator ξ^{*+}_t in terms of risk estimate. Now if ξ^*_w dominates ξ^*_t in risk estimate then it also dominates it in risk and dominates the maximum likelihood estimator in risk estimate. Thus $-1 \leq w \leq 1$. However, we have already seen that if $w < \min(-|t|, \frac{1}{\tilde{c}} F - 1)$ and $a+b+b\,w > 0$ then there are values of $w < -1$. This is a contradiction. Also if $w > \frac{1}{\tilde{c}} F - 1$ for $F < \tilde{c}(1+t)$ then ξ^*_w can be improved upon in risk by taking the positive part estimator. Now, taking $t = 0$ we have the contradiction that

$$\xi^+_t = \xi^{*+}_w < \xi^*_w \leq \xi^{*+}_t$$

since the positive part version of ξ^*_w is ξ^+_t. Thus the positive part James-Stein estimator (with $t = 0$) cannot be improved upon in risk estimate.

7.3 A Condition for Risk Estimate Dominance over ξ^*_t

Efron and Morris(1976) gave a condition for risk estimate dominance of an estimator, for which an unbiased risk estimate exists, over the maximum likelihood estimator. We have already given a condition for risk estimate dominance of such an estimator over ξ^*_t. In this section we shall give a condition which is similar to that given by Efron and Morris. The condition is only given for ξ^*_w in the case that w depends only on F.

In order to find the condition we shall solve equation 7.2.1.1 again. Writing $\psi = (w^2 - t^2) F \frac{d\phi}{dF}$ we find that

$$\frac{a+b+b\,w}{w^2-t^2} \frac{dw}{dF} = \frac{1}{F}\left(1 + F \frac{d\phi}{dF}\right) = \frac{1}{F} + \frac{d\phi}{dF}$$

where ϕ is non-decreasing if $w^2 > t^2$ and non-increasing if $w^2 < t^2$. Integrating gives

$$\phi = -\log F + \frac{a+b}{2|t|} \log \frac{w-|t|}{w+|t|} + \frac{b}{2} \log (w^2 - t^2).$$

The complete solution is therefore that

$$\exp \phi = \frac{1}{F}\left(\frac{w-|t|}{w+|t|}\right)^{\frac{a+b}{2|t|}} (w^2-t^2)^{\frac{1}{2}b}$$

is non-decreasing when $w^2 > t^2$ and non-increasing when $w^2 < t^2$. We have already seen that the former case is impossible. Although we shall not prove it here, it is fairly easy to see that the above condition implies that if $w(F_o) = |t|$ then $w(F) = |t|$ for $F > F_o$ and that for $|t| < \frac{a+b}{b} = 1 + \frac{1}{c}$ if $w(F_1) = -|t|$ then $w(F) = -|t|$

for $F < F_1$. This result is contained in Moore and Brook(1978) and slightly generalises the similar theorem in Efron and Morris(1976).

Chapter 8

Distribution of Studentised Shrunken Estimators

8.1 Introduction

The methods of chapter 4 allow for the computation of the first, second and fourth moments of the James-Stein estimator in terms of hypergeometric functions. Similar methods lead to the computation of the third moment. The computation is complicated by the fact that the shrinkage factor depends on both X and s (and is homogeneous in $s^{-1}X$) but the other factor is just X. Thus the estimator is neither a homogeneous function of X nor of $s^{-1}X$. The Studentised shrunken estimator, $s^{-1}\xi^*$, on the other hand, is homogeneous in $s^{-1}X$. This makes the computation of the density function almost a triviality since, putting $Z = s^{-1}\xi^*$ and $T = s^{-1}X$, the distribution of Z is just a transformation of the distribution of T (which has a multivariate t-distribution with parameter λ).

If we Studentised the estimator in a different way, by dividing by the trace of its variance matrix, then we just have a linear multiple of Z - the factor being dependent on λ. This multiple is easily found by the methods of chapter 6, but we prefer to define Studentisation in the former manner.

Noting that the shrinkage factor only depends on the length of the vector $s^{-1}X$, we find it is easier to work in polar coordinates. Accordingly, in the next section we transform the mutivariate normal and t distributions to polar form.

8.2 Polar Coordinates

We first transform the coordinate system $(x_1,\ldots,x_p)$ to the system $(y_1,\ldots,y_p)$ by an orthogonal transformation in such a way that the y_p-axis is in the direction of ξ. Dividing by σ then gives a coordinate system $(z_1,\ldots,z_p)$ in which the point $(\xi_1,\ldots,\xi_k)$ in the original system is given by $(0,0,\ldots,0,\lambda)$ in the z-coordinates.

We shall transform to polar coordinates through a sequence of transformations. Let $r_1 = z_1$,

$$z_1 = r_2 \cos\theta_1 \qquad -\pi \le \theta_1 < \pi.$$
$$z_2 = r_2 \sin\theta_1$$

We then transform the other coordinates successively by the transformations

$$z_{i+1} = r_{i+1} \sin \theta_i$$
$$r_i = r_{i+1} \cos \theta_i \qquad -\tfrac{1}{2}\pi \le \theta_i \le \tfrac{1}{2}\pi$$

for $i = 2,3,\ldots,p-1$. Putting $r = r_p$, we finally obtain the coordinate system $(\theta_1,\ldots,\theta_{p-1},r)$ which may be written directly in terms of $(z_1,\ldots,z_p)$ by the relations

$$z_i = r \cos \theta_{p-1} \cos \theta_{p-2} \ldots \cos \theta_{i+1} \sin \theta_i \qquad i=2,3,\ldots,p-1$$
$$z_1 = r \cos \theta_{p-1} \cos \theta_{p-2} \ldots \cos \theta_1 .$$

Using this sequence of transformations makes it easy to find the Jacobian of the combined transformation which is

$$\frac{\partial(z_1,\ldots,z_p)}{\partial(\theta_1,\ldots,\theta_{p-1},r)} = r_2\, r_3 \ldots r_{p-1}\, r_p$$
$$= r^{p-1} \cos^0\theta_1 \cos^1\theta_2 \cos^2\theta_3 \ldots \cos^{p-1}\theta_{p-1} .$$

Now the normal density function is given by

$$p(z) = \frac{1}{(2\pi)^{\frac{1}{2}p}} \exp\left\{-\tfrac{1}{2}(z_p-\lambda)^2 + \sum_{i=1}^{p-1} z_i^2\right\}$$
$$= \frac{1}{(2\pi)^{\frac{1}{2}p}} e^{-\frac{1}{2}\lambda} e^{\lambda z_p} e^{-\frac{1}{2}z^T z} .$$

Transforming to polar coordinates gives

$$p(\theta_1,\ldots,\theta_{p-1},r) = \frac{1}{(2\pi)^{\frac{1}{2}p}} e^{-\frac{1}{2}\lambda^2} e^{\lambda r \sin \theta_{p-1}} e^{-\frac{1}{2}r^2} \cos \theta_2 \cos^2\theta_3 \ldots \cos^{p-2} \theta_{p-1}.$$

As might be expected, this factorises into a density involving r and θ_{p-1} and densities involving $\theta_1,\theta_2,\ldots,\theta_{p-2}$. Thus r and θ_{p-1} are not independent but are independent of the other variables which are also mutually independent. For each value of r and θ_{p-1}, the conditional density is, in fact, uniform over a $p-2$ dimensional sphere. Our shrunken estimators shrink in the direction of r and leave all the θ-coordinates unchanged.

Having transformed to this coordinate system it is almost a trivial matter to derive the density function for the non-central χ^2 distribution. We merely write

$$e^{\lambda r \sin \theta_{p-1}} = \sum_{k=0}^{\infty} \frac{(\lambda r)^k}{k!} \cos^k\theta_{p-1}$$

and integrate with respect to $\theta_1,\ldots,\theta_{p-1}$. Our concern here, though, is the joint density of r and θ_{p-1}.

Since $\int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \cos^i\theta \, d\theta = \frac{\Gamma(\frac{1}{2}i+\frac{1}{2})\,\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}i+1)}$ we have

$$\prod_{i-2}^{p-2} \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \cos^{i-1}\theta_i \, d\theta_i = \frac{\pi^{\frac{1}{2}p-\frac{1}{2}}}{\Gamma(\frac{1}{2}p-\frac{1}{2})} = \frac{(2\pi)^{\frac{1}{2}p}}{(2\pi)2^{\frac{1}{2}p-1}\,\Gamma(\frac{1}{2}p-1)}.$$

Writing $\phi = \theta_{p-1}$, we obtain the density for r and ϕ

$$p(r,\phi) = \frac{e^{-\frac{1}{2}\lambda^2}}{2^{\frac{1}{2}p}\,\Gamma(\frac{1}{2}p-\frac{1}{2})}\, r^{p-1}\, e^{-\frac{1}{2}r^2}\, e^{\lambda r \sin\phi}\, \cos^{p-2}\phi.$$

It is now an easy matter, in the case of known variance, to transform r by the shrinkage transformation. Let $u = h(r)\,r$ and assume that $h(.)$ is one-to-one so that $r = f(u)$ for some function $f(.)$. The joint density of u and ϕ is then

$$p(u,\phi) = \frac{e^{-\frac{1}{2}\lambda^2}}{2^{\frac{1}{2}p}\,\Gamma(\frac{1}{2}p-\frac{1}{2})}\, (f(u))^{p-1}\, e^{-\frac{1}{2}(f(u))^2}\, e^{\lambda f(u)\sin\phi}\, f'(u)\, \cos^{p-2}\phi.$$

If required it is then possible to multiply by the joint density of $\theta_2,\ldots,\theta_{p-2}$ and invert the transformation. However, it is probably better to leave the density in polar coordinates.

8.3 Unknown Variance

If the variance is unknown and estimated by s then $S = \frac{ns^2}{\sigma^2}$ has a χ^2-distribution on n degrees of freedom. The joint density of r, ϕ and S is therefore

$$p(r,\phi,S) = \frac{e^{-\frac{1}{2}\lambda^2}}{2^{\frac{1}{2}p}\,\Gamma(\frac{1}{2}p-\frac{1}{2})\,2^{\frac{1}{2}n}\,\Gamma(\frac{1}{2}n)}\, r^{p-1}\, e^{-\frac{1}{2}r^2}\, e^{\lambda r\sin\phi}\, \cos^{p-2}\phi\, S^{\frac{1}{2}n-1}\, e^{-\frac{1}{2}S}.$$

We require the joint density of $\frac{r\sigma}{s}$ and ϕ, that is, of $\sqrt{\frac{n}{S}}\,r$. If we transformed this density back to the original coordinate system then we would obtain the multivariate non-central t density in catesian form. This is already well known. Alternatively, we could have started with the known form of that density and transformed it in the manner of the previous section and thus avoiding the integration with respect to S.

Putting $t = \sqrt{\frac{n}{S}}\,r$ we have $\frac{\partial(t,S)}{\partial(r,S)} = \sqrt{\frac{n}{S}}$ and

$$p(t,\phi,S) = \frac{e^{-\frac{1}{2}\lambda^2}\, n^{1-\frac{1}{2}p}}{2^{\frac{1}{2}p}\,\Gamma(\frac{1}{2}p-\frac{1}{2})\,2^{\frac{1}{2}n}\,\Gamma(\frac{1}{2}n)}\, S^{\frac{1}{2}p-1}\, t^{p-1}\, e^{-\frac{St^2}{2n}}\, \exp(\tfrac{\lambda}{\sqrt{n}}S^{\frac{1}{2}}t\sin\phi) \times \cos^{p-2}\phi\, S^{\frac{1}{2}n-1}\, e^{-\frac{1}{2}S}.$$

Expanding $\exp(\frac{\lambda}{\sqrt{n}}S^{\frac{1}{2}}t\sin\phi)$ as a power series we have

$$p(t,\phi,S) = \frac{e^{-\frac{1}{2}\lambda^2}\, n^{1-\frac{1}{2}p}}{2^{\frac{1}{2}p}\,\Gamma(\frac{1}{2}p-1)\,2^{\frac{1}{2}n}\,\Gamma(\frac{1}{2}n)} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{\sin^k\phi\,\cos^{p-2}\phi}{n^{\frac{1}{2}k}}\, t^{p-1+k} \times S^{\frac{1}{2}p+\frac{1}{2}k-1} e^{-\frac{1}{2}(1+t^2/n)S}.$$

It is now easy to integrate with respect to S and obtain

$$p(t,\phi) = \frac{e^{-\frac{1}{2}\lambda^2}}{2^{\frac{1}{2}p+\frac{1}{2}n}\,\Gamma(\frac{1}{2}p-1)\Gamma(\frac{1}{2}n)} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{\sin^k\phi\,\cos^{p-2}\phi}{n^{\frac{1}{2}k}} \times \frac{t^{p+k-1}\,2^{\frac{1}{2}p+\frac{1}{2}k}\,\Gamma(\frac{1}{2}p+\frac{1}{2}k)}{(1+t^2/n)^{\frac{1}{2}p+\frac{1}{2}k}}.$$

If desired, then this can be transformed back. On putting $u = f(t)$ we can find the distribution of the shrunken estimator as in the last section. The formula is not really suited to analytical manipulation but can be dealt with numerically.

A p p e n d i x 1

Gamma Beta and Hypergeometric Functions

A1.1 Introduction

In order to make referenced formulae easier to find we list the main properties of the gamma beta and hypergeometric functions and various generalisations of the hypergeometric function. Most of these properties may be found in Erdélyi(1953), Slater(1960) and Slater(1966).

A1.2 The gamma Function

We define $(a)_z = \lim_{x\to a}\lim_{n\to\infty} \frac{x(x+1)(x+2)\ldots(x+n-1)}{(x+z)(x+z+1)\ldots(x+z+n-1)}\, n^z$

(when the limit exists),

$$[a]_z = (a-z+1)_z$$

and $$\Gamma(z) = (1)_{z-1}.$$

The following properties hold

(1) $$\Gamma(z) = \frac{1}{z}\lim_{N\to\infty} \frac{N!}{(z+1)(z+2)\ldots(z+N-1)}\, N^z$$

and is analytic except at simple poles at $z=0,-1,-2,-3,\ldots$

(2) $$\Gamma(z+1) = z\Gamma(z)$$

(3) $$(a)_z = \lim_{x\to a} \frac{\Gamma(x+z)}{\Gamma(x)}$$

(4) $$[a]_z = \lim_{x\to a} \frac{\Gamma(x+1)}{\Gamma(x+n-1)}$$

(5) $$(a)_z \sim a^z \quad \text{as} \quad a\to\infty$$

(6) $$[a]_z \sim a^{-z} \quad a\to\infty$$

(7) $$(a)_{z+\zeta} = \lim_{x\to a} (x)_z\,(x+z)_\zeta$$

(8) $$[a]_{z+\zeta} = \lim_{x\to a} [x]_z\,[x+z]_\zeta$$

in fact (5) and (7) together with $(a)_1 = a$ characterise $(a)_z$

(9) $$(a)_o = 1$$

(10) $$[a]_o = 1$$

(11) $$(a)_{-z} = \frac{1}{[a-1]_z} \quad \text{if it exists}$$

(12) $$[a]_{-z} = \frac{1}{(a+1)_z} \quad \text{if it exists}$$

(13) $$(a)_n = (-1)^n[-a]_n \quad n \text{ an integer}$$

(14) $$[a]_n = (-1)^n(-a)_n \quad n \text{ an integer}$$

(15) $$(a)_n = a(a+1)(a+2)\ldots(a+n-1) \quad n=1,2,3,\ldots$$

(16) $$[a]_n = a(a-1)(a-2)\ldots(a-n+1) \quad n=1,2,3\ldots$$

multiplication rule

$$(17)\qquad \Gamma(nz) = \frac{n^{nz-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}n-\frac{1}{2}}} \prod_{r=0}^{n-1} \Gamma\left(z + \frac{r}{n}\right)$$

special case - the duplication formula

$$(18)\qquad \Gamma(2z) = \frac{2^{2z-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \Gamma(z)\,\Gamma(z+\tfrac{1}{2})$$

$$(19)\qquad \Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

$$(20)\qquad \int_0^\infty x^{z-1} e^{-ax}\,dx = \frac{1}{a^z}\,\Gamma(z) \quad \text{if the real part of } z \text{ is positive.}$$

Stirling's expansion

$$(21)\qquad \log \Gamma(z) \sim (z-\tfrac{1}{2}) \log z - z + \tfrac{1}{2}\log(2\pi) + \sum_{r=1}^{\infty} \frac{(-1)^{r-1} B_r}{2r(2r-1)z^{2r-1}}$$

$$\text{as } z \to \infty \quad |\arg z| \le \pi - \Delta$$

where B_r is the r th Bernouilli number.

$$(22)\qquad \Gamma(z) \sim e^{-z} z^{z-\frac{1}{2}} (2\pi)^{\frac{1}{2}} \left\{1 + \frac{1}{12z} + \frac{1}{288z^2} - \frac{139}{51840z^3} - \frac{571}{2488320z^4} + O\left(\frac{1}{z^5}\right)\right\}.$$

A1.3 The Beta Function

The beta function is defined in terms of the gamma function as $B(z,\zeta) = \frac{\Gamma(z)\,\Gamma(\zeta)}{\Gamma(z+\zeta)}$. The following formulae, valid when the real parts of z and ζ are positive, are important

$$(1)\qquad B(z,\zeta) = \int_0^1 x^{z-1} (1-x)^{\zeta-1}\,dx$$

$$(2)\qquad B(z,\zeta) = \int_0^\infty \frac{x^{z-1}}{(1+x)^{z+\zeta}}\,dx$$

$$(3)\qquad B(z,\zeta) = 2\int_0^{\frac{1}{2}} \sin^{2z-1}\theta \cos^{2\zeta-1}\theta\,d\theta$$

A1.4 The Hypergeometric Function

We use the term "hypergeometric function" to include certain generalisations of the Gaussian hypergeometric function defined below. Within the circle of convergence we define

$${}_PF_Q(a_1,a_2,\ldots,a_P;b_1,b_2,\ldots,b_Q;z) = \sum_{r=0}^{\infty} \frac{(a_1)_r (a_2)_r \ldots (a_P)_r}{(b_1)_r (b_2)_r \ldots (b_Q)_r} \frac{z^r}{r!}$$

and by analytic continuation outside. We shall often omit the subscripts P and Q on F. The radius of convergence is ∞, 1 or 0 according

as $P \leq Q$, $P = Q + 1$ or $P > Q + 1$. In the latter case the series is an asymptotic expansion as $z \to 0$ of a convergent hypergeometric series (Slater(1966)) which may be identified with it. The special cases ${}_2F_1(a,b;c;z)$ and ${}_1F_1(a;c;z)$ are respectively the Gaussian hypergeometric function and the confluent hypergeometric function. The latter is a limiting case of the former since

$$ {}_1F_1(a;c;z) = \lim_{b\to\infty} {}_2F_1(a,b;c;\frac{z}{b}) . $$

Writing $F = F(a,b;c;z)$, $F(a+) = F(a+1,b;c;z)$, $F(a-) = F(a-1,b;c;z)$ etc. we have the following recurrence relations for the Gaussian hypergeometric function

(1) $$\{(c-2a)-(b-a)z\}F + a(1-z)F(a+) - (c-a)F(a-) = 0$$

(2) $$(b-a)F + aF(a+) - bF(b+) = 0$$

(3) $$(c-a-b)F + a(1-z)F(a+) - (c-b)F(b-) = 0$$

(4) $$c\{a-(c-b)z\}F - ac(1-z)F(a+) + (c-a)(c-b)zF(c+) = 0$$

(5) $$(c-a-1)F + aF(a+) - (c-1)F(c-) = 0$$

(6) $$(c-a-b)F - (c-a)F(a-) + b(1-z)F(b+) = 0$$

(7) $$(b-a)(1-z)F - (c-a)F(a-) + (c-b)F(b-) = 0$$

(8) $$c(1-z)F - cF(a-) + (c-b)zF(c+) = 0$$

(9) $$\{(a-1)-(c-b-1)z\}F + (c-a)F(a-) - (c-1)(1-z)F(c-) = 0$$

(10) $$\{(c-2b)+(b-a)z\}F + b(1-z)F(b+) - (c-b)F(b-) = 0$$

(11) $$c\{b-(c-a)z\}F - bc(1-z)F(b+) + (c-a)(c-b)zF(c+) = 0$$

(12) $$(c-b-1)F + bF(b+) - (c-1)F(c-) = 0$$

(13) $$c(1-z)F - cF(b-) + (c-a)zF(c+) = 0$$

(14) $$\{(b-1)-(c-a-1)z\}F + (c-b)F(b-) - (c-1)(1-z)F(c-) = 0$$

(15) $$c\{(c-1)-(2c-a-b-1)z\}F + (c-a)(c-b)F(c+) - c(c-1)(1-z)F(c-) = 0$$

Further recurrence relations deducible from these, but more easily proved directly are

(16) $$F(a,b+1;c;z) - F(a,b;c;z) = \frac{az}{c} F(a+1,b+1;c+1;z)$$

(17) $$F(a+1,b;c;z) - F(a,b;c;z) = \frac{bz}{c} F(a+1,b+1;c+1;z)$$

(18) $$(c-1)F(a,b;c-1;z) + F(a,b;c;z) = \frac{abz}{c} F(a+1,b+1;c+1;z).$$

Also

(19) $$\frac{d}{dz} F(a,b;c;z) = \frac{ab}{c} F(a+1,b+1;c+1;z).$$

Gauss's theorem

(20) $$F(a,b;c;1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} = \frac{(c-a)_a}{(c-a-b)_a} = \frac{(c-b)_b}{(c-a-b)_b} .$$

Euler's theorem

(21) $$F(a,b;c;z) = (1-z)^{c-a-b} F(c-a,c-b;c;z).$$

For the confluent hypergeometric function we have the following recurrence relations

(22) $$(c-2a-z)F + aF(a+) - (c-a)F(a-) = 0$$

(23) $$c(a+z)F - acF(a+) - (c-a)F(c+) = 0$$

(24) $$(c-a-1)F + aF(a+) - (c-1)F(c-) = 0$$

(25) $$cF - cF(a-) - zF(c+) = 0$$

(26) $$(a-1+z)F + (c-a)F(a-) - (c-1)F(c-) = 0$$

(27) $$c(c-1+z)F - (c-a)zF(c+) - c(c-1)F(c-) = 0.$$

Also

(28) $$F(a+1;c;z) - F(a;c;z) = \frac{z}{c} F(a+1;c+1;z)$$

(29) $$(c-1)F(a;c-1;z) + F(a;c;z) = \frac{az}{c} F(a+1;c+1;z)$$

(30) $$\frac{d}{dz} F(a;c;z) = \frac{a}{c} F(a+1;c+1;z)$$

(31) $$e^{-z}F(a;c;z) = F(c-a;c;-z)$$

(32) $$e^{-z}F(a;c;z) \sim \frac{\Gamma(c)}{\Gamma(a)} z^{a-b} \, {}_2F_0(1-a,b-a;;\frac{1}{z}) \quad \text{as} \quad z \to \infty$$

(33) $$F(a;c;-z) = \frac{\Gamma(c)}{\Gamma(c-a)} z^{-a} \, {}_2F_0(1-c+a,a;;\frac{1}{z}) \quad \text{as } z \to \infty .$$

The ${}_2F_2$ function may be written in terms of ${}_1F_1$ functions. We give the proof as the result is not so well known as the others. We first prove a useful lemma.

Lemma 1 $$\sum_{i=0}^{\infty} \frac{[u]_i[k]_i}{(a)_i \, i!} = \frac{(a+u)_k}{(a)_k} = \frac{(a+k)_u}{(a)_u} .$$

Proof Using the definition of the Gaussian hypergeometric function and Gauss's theorem we have

$$\sum_{i=0}^{\infty} \frac{[u]_i[k]_i}{(a)_i \, i!} = \sum_{i=0}^{\infty} \frac{(-u)_i(-k)_i}{(a)_i \, i!}$$

$$= {}_2F_1(-u,-k;a;1) = \frac{\Gamma(a)\Gamma(a+u+k)}{\Gamma(a+u)\Gamma(a+k)} .$$

This may also be proved by induction if either k or u is a positive integer. In our application to the next theorem k will be a positive integer.

Theorem 1 $$ {}_2F_2(a+u,b;a,c;z) = \sum_{i=0}^{\infty} \frac{[u]_i(b)_i z^i}{(a)_i(c)_i i!} \, {}_1F_1(b+i;c+i;z)$$

Proof $$ {}_2F_2(a+u,b;a,c;z) = \sum_{k=0}^{\infty} \frac{(a+u)_k}{(a)_k} \frac{(b)_k}{(c)_k} \frac{z^k}{k!} .$$

Thus
$$
\begin{aligned}
{}_2F_2(a+u,b;a,c;z) &= \sum_{k=0}^{\infty}\sum_{i=0}^{\infty} \frac{[u]_i[k]_i(b)_k}{(a)_i(c)_k\, i!}\frac{z^k}{k!} \\
&= \sum_{i=0}^{\infty}\sum_{k=i}^{\infty} \frac{[u]_i[k]_i(b)_k}{(a)_i(c)_k\, i!}\frac{z^k}{k!} \\
&= \sum_{i=0}^{\infty} \frac{[u]_i(b)_i z^i}{(a)_i(c)_i i!} \sum_{k=0}^{\infty} \frac{(b+i)_k}{(c+i)}\frac{z^k}{k!} \\
&= \sum_{i=0}^{\infty} \frac{[u]_i}{i!}\frac{(b)_i z^i}{(a)_i(c)_i}\ {}_1F_1(b+i;c+i;z).
\end{aligned}
$$

If u is a positive integer then this sum is finte.

A1.5 Hypergeometric Functions of Two Variables

The Gaussian hypergeometric series can be generalised to the case of two variables in a variety of ways. We shall only give the Appell series which form one set of generalisations. We define

$$F_1(a;b,b',c;x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_m(b')_n}{(c)_{m+n}\, m!\ n!}\, x^m y^n$$

$$F_2(a;b,b';c,c';x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_m(b')_n}{(c)_m(c')_n\, m!\ n!}\, x^m y^n$$

$$F_3(a,a';b,b';c;x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_m(a')_n(b)_m(b')_n}{(c)_{m+n}\, m!\ n!}\, x^m y^n$$

and
$$F_4(a;b;c,c';x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_{m+n}}{(c)_m(c')_n\, m!\ n!}\, x^m y^n .$$

Note that defining

$$F_0(a;b;c;x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_{m+n}}{(c)_{m+n}\, m!\ n!}\, x^m y^n$$

and
$$F_5(a,a';b,b';c,c';x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_m(a')_n(b)_m(b')_n}{(c)_m(c')_n\, m!\ n!}\, x^m y^n$$

do not define new functions since the former is equal to $F(a,b;c;x+y)$ and the latter is equal to $F(a,b;c;x)F(a',b';c';y)$.

Appendix 2
Distributions

A2.1 Introduction

In this appendix we briefly state the definitions of the density functions we have used and calculate the moments and expected values of certain random variables for which these values have been quoted earlier in the thesis.

A2.2 Non-central Beta and Gamma Distributions

The non-central gamma density with parameters α, β and non-centrality parameter λ is defined to be the function

$$p(x) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{x^{\alpha+k-1} e^{-x/\beta}}{\beta^{\alpha+k} \Gamma(\alpha+k)}$$

and is denoted by $\gamma(\alpha,\beta,\lambda)$ a notation which will be interpreted as standing for the distribution or for a random variable with that distribution. Putting $\alpha = \frac{1}{2}n$ and $\beta = 2$ gives the non-central γ distribution denoted by $\gamma(n,\lambda)$ which is the density of the sum of squares of n independent $N(\mu_i,1)$ variables with $\frac{1}{2}\Sigma_{i=1}^{n} \mu_i^2 = \lambda$. The special cases in which each μ_i is zero (i.e. $\lambda = 0$) are the central γ and χ^2 distributions denoted respectively by $\gamma(\alpha,\beta)$ and χ^2_n.

It is clear that the $\gamma(\alpha,\beta,\lambda)$ distribution is the marginal density of X from the joint distribution

$$p(x,k) = e^{-\lambda} \frac{\lambda^k}{k!} \frac{x^{\alpha+k-1} e^{-x/\beta}}{\beta^{\alpha+k} \Gamma(\alpha+k)}$$

the conditional desity given $K = k$ being $\gamma(\alpha+k,\beta)$. The marginal density of K is the Poisson distribution with parameter λ.

The ratio of independent $\gamma(\mu,\beta,\lambda)$ and $\gamma(\nu,\beta)$ variates is the non-central inverse beta distribution $\beta_2(\mu,\nu,\beta,\lambda)$ with the density function

$$p(x) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{x^{\mu+k-1}}{B(\mu+k,\nu)(1+x)^{\mu+\nu+k}}$$

which can be interpreted as a marginal density in the same way as the non-central gamma distribution.

The ratio of independent $\frac{1}{\mu}\gamma(\mu,\beta,\lambda)$ and $\frac{1}{\nu}\gamma(\nu,\beta)$ variates, or, which is the same, the ratio of $\frac{1}{2\mu}\chi^2(2\mu,\lambda)$ and $\frac{1}{2\nu}\chi^2_{2\nu}$ has the density function

$$p(x) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{m^{\frac{1}{2}m+k} n^{\frac{1}{2}n}}{B(\frac{1}{2}m+k, \frac{1}{2}n)} \frac{x^{\frac{1}{2}m+k-1}}{(n+mx)^{\frac{1}{2}m+\frac{1}{2}n+k}} \qquad \text{where } m=2\mu \text{ and } n=2\nu.$$

If m and n are positive integers then this is the non-central F distribution denoted by $F(m,n,\lambda)$. We shall use the same notation regardless of whether m and n are integers. Regarding this as the marginal density of a joint distribution, we note that the conditional density given $K = k$ is the density of $\frac{m+2k}{m} F_{m+2k,m}$ where $F_{a,b}$ denotes the random variable $F(a,b,0)$.

The transformations $y = \frac{x}{1+x}$ and $y = \frac{mx}{m+nx} = \frac{\mu x}{\mu+\nu x}$ respectively transform $\beta_2(\mu,\nu,\lambda)$ and $F(2\mu,2\nu,\lambda)$ to the distribution with density

$$p(y) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{y^{\mu+k-1} (1-y)^{\nu-1}}{B(\mu+k,\nu)}$$

which is defined to be the non-central beta distribution denoted by $\beta_1(\mu,\nu,\lambda)$ or by $\beta(\mu,\nu,\lambda)$.

A2.3 Moments of Non-central Beta and Gamma Distributions

The p th moment about zero of the non-central gamma distribution is

$$\begin{aligned}
E[X^p] &= \int_0^{\infty} e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{x^{\alpha+p+k-1} e^{-x/\beta}}{\beta^{\alpha+k} \Gamma(\alpha+k)} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{\beta^{\alpha+p+k}}{\beta^{\alpha+k}} \frac{\Gamma(\alpha+p+k)}{\Gamma(\alpha+k)} \\
&= \beta^p e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\alpha)_p (\alpha+p)_k}{(\alpha)_k} \\
&= \beta^p (\alpha)_p \; {}_1F_1(\alpha+p;\alpha;\lambda) \\
&= \beta^p (\alpha)_p \sum_{k=0}^{\infty} \frac{(-p)_k}{(\alpha)_k} \frac{(-\lambda)^k}{k!} \\
&= \beta^p (\alpha)_p \sum_{k=0}^{\infty} \frac{[p]_k \lambda^k}{(\alpha)_k k!} \; .
\end{aligned}$$

If p is a positive integer then this series terminates at $k = p$. The mean and variance are easily seen to be

$$E[X] = (\alpha+\lambda)\beta \quad \text{and} \quad \text{var}(X) = (\alpha+2\lambda)\beta^2 .$$

Putting $\beta = 2$ and $\alpha = \frac{1}{2}n$ gives the mean and variance of the $\chi^2(n,\lambda)$ distributions respectively as

$$n + 2\lambda \qquad \text{and} \qquad 2n + 8\lambda .$$

The p th moment about zero of the $\beta_1(\mu,\nu,\lambda)$ distribution is

$$E[X^p] = \int_0^1 e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{1}{B(\mu+k,\nu)} x^{\mu+p+k-1}(1-x)^{\nu-1}\, dx$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{B(\mu+p+k,\nu)}{B(\mu+k,\nu)}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\mu+k)_p}{(\mu+\nu+k)_p}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\mu+\nu)_k(\mu+p)_k}{(\mu)_k(\mu+\nu+p)_k} \frac{(\mu)_p}{(\mu+\nu)_p}$$

$$= \frac{(\mu)_p}{(\mu+\nu)_p} e^{-\lambda}\, {}_2F_2(\mu+\nu,\mu+p;\mu,\mu+\nu+p;\lambda).$$

This does not simplify to a finite sum when p is a positive integer.

The p th moment about zero for the $\beta_2(\mu,\nu,\lambda)$ distribution is

$$E[X^p] = \int_0^{\infty} e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{1}{B(\mu+k,\nu)} \frac{x^{\mu+p+k-1}}{(1+x)^{\mu+p+k+\nu-p}}\, dx$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{B(\mu+p+k,\nu-p)}{B(\mu+k,\nu)} \qquad \text{if } p < \nu$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\mu+k)_p}{(\nu-p)_p}$$

$$= \frac{1}{[\nu-1]_p} e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\mu)_p(\mu+p)_k}{(\mu)_k}$$

$$= \frac{(\mu)_p}{[\nu-1]_p} e^{-\lambda}\, {}_1F_1(\mu+p;\mu;\lambda)$$

$$= \frac{(\mu)_p}{[\nu-1]_p}\, {}_1F_1(-p;\mu;-\lambda)$$

$$= \frac{(\mu)_p}{[\nu-1]_p} \sum_{k=0}^{\infty} \frac{[p]_k \lambda^k}{(\mu)_k k!}$$

and this does terminate if p is a positive integer.

After a little simplification we find the mean and variance are

$$E[X] = \frac{1}{\nu-1}(\mu+\lambda) \qquad \text{and} \qquad \text{var}(X) = \frac{1}{\nu-1}\left\{\frac{(\mu+\lambda)^2}{(\nu-1)(\nu-2)} + \frac{\mu+2\lambda}{\nu-2}\right\}.$$

This gives the mean and variance of the non-central F distribution as

$$\frac{n}{n-2}\left(1 + \frac{2\lambda}{m}\right) \qquad \text{and} \qquad \frac{2n}{m^2(n-2)}\left\{\frac{(m+2\lambda)^2}{(n-2)(n-4)} + \frac{m+4\lambda}{n-4}\right\}.$$

By similar reasoning it is easy to show that, for the $\beta_1(\mu,\nu,\lambda)$ distribution

$$E[X^m(1-X)^n] = \frac{(\mu)_m(\nu)_n}{(\mu+\nu)_{m+n}} e^{-\lambda}\, {}_2F_2(\mu+\nu,\mu+m;\mu,\mu+\nu+m+n;\lambda)$$

and for the $\beta_2(\mu,\nu,\lambda)$ distribution

$$E\left[\frac{X^m}{(1+X)^n}\right] = \frac{(\mu)_m(\nu)_{n-m}}{(\mu+\nu)_n} e^{-\lambda} {}_2F_2(\mu+\nu,\mu+m;\mu,\mu+\nu+n;\lambda).$$

For the $\beta_2(\mu,\nu,\lambda)$ distribution we can also caculate $E\left[\frac{X^m}{(d+X)^n}\right]$ a result which is helpful for calculating the risk function for the bilinear shrinkage estimators of chapter 4. We have

$$E\left[\frac{X^m}{(d+X)^n}\right] = E\left[\frac{X^m}{(1+X)^n}\left\{1 - \frac{1-d}{1+X}\right\}^{-n}\right]$$

$$= E\left[\sum_{r=0}^{\infty} \frac{(n)_r}{r!} \frac{X^m}{(1+X)^n}\left(\frac{1-d}{1+X}\right)^r\right] \quad \text{which converges if } |1-d| < 1$$

$$= \sum_{r=0}^{\infty} \frac{(n)_r}{r!}(1-d)^r \frac{(\mu)_m(\nu)_{n+r-m}}{(\mu+\nu)_{n+r}} e^{-\lambda} \times {}_2F_2(\mu+m,\mu+\nu;\mu,\mu+\nu+n+r;\lambda)$$

$$= e^{-\lambda} \sum_{r=0}^{\infty}\sum_{k=0}^{\infty} \frac{(n)_r(\mu)_m(\nu)_{n+r-m}}{(\mu+\nu)_{n+r}} \frac{(\mu+m)_k(\mu+\nu)_k}{(\mu)_k(\mu+\nu+n+r)_k} \frac{(1-d)^k}{r!} \frac{k}{k!}$$

$$= e^{-\lambda} \frac{(\mu)_m(\nu)_{n-m}}{(\mu+\nu)_n} \sum_{r=0}^{\infty}\sum_{k=0}^{\infty} \frac{(n)_r(\nu+n-m)_r(\mu+m)_k(\mu+\nu)_k}{(\mu)_k(\mu+\nu+n)_{r+k}} \frac{(1-d)^r}{r!} \frac{\lambda^k}{k!}$$

$$= e^{-\lambda} \frac{(\mu)_m(\nu)_{n-m}}{(\mu+\nu)_n} F(n,\nu+n-m,\mu+m,\mu+\nu;\mu,\mu+\nu+n;1-d,\lambda)$$

Where F is one of the many possible hypergeometric functions of two variables and of order 2,2 in λ and 2,1 in $1-d$. By analytic continuation, the result also holds if $|1-d| \geq 1$. In the case $m = 0$ the order of the hypergeomtric function reduces to 1,1 in λ and 2,1 in $1-d$ and the case $m = 1$ may be reduced to the case $m = 0$ since

$$E\left[\frac{X}{(d+X)^m}\right] = E\left[\frac{1}{(d+X)^{m-1}}\right] - d\,E\left[\frac{1}{(d+X)^m}\right].$$

A2.4 Expectations with Respect to the Joint Density which Gives Rise to the Non-central Inverse Beta Distribution

Given the joint density

$$p(u,k) = e^{-\lambda} \frac{\lambda^k}{k!} \frac{u^{\mu+k-1}}{B(\mu+k,\nu)(1+u)^{\mu+\nu+k}}$$

we shall find $E\left[\frac{(\delta+K)_m}{(\gamma+K)_n} r(U)\right]$ in terms of $V \sim \beta_2(\mu,\nu)$. We have

$$E\left[\frac{(\delta+K)_m}{(\gamma+K)_n} r(U)\right] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\delta+k)_m}{(\gamma+k)_n} \int_0^{\infty} r(U) \frac{u^{\mu+k-1}}{B(\mu+k,n)(1+u)^{\mu+\nu+k}}\, du$$

We thus have

$$E\left[\frac{(\delta+K)_m}{(\gamma+K)_n}\, r(U)\right] = \int_0^\infty r(u)\, \frac{u^{\mu-1}}{B(\mu,\nu)(1+u)^{\mu+\nu}} \times \left\{e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{(\delta+k)_m}{(\gamma+k)_n} \frac{(\mu+\nu)_k}{(\mu)_k} \left(\frac{u}{1+u}\right)^k\right\} du$$

$$= \frac{(\delta)_m}{(\gamma)_n} e^{-\lambda} E[r(V)\ {}_3F_3(\delta+m,\gamma,\mu+\nu;\delta,\gamma+n,\mu;\tfrac{V\lambda}{1+V})].$$

For certain special cases, for example if $n = 0$, the order of the hypergeometric function is reduced.

A2.5 The Poisson Distribution

The first few central moments of the Poisson distribution (see Kendal and Stuart(1977-79) are

$$\mu_o(\lambda) = 1 \qquad \mu_1(\lambda) = 0$$
$$\mu_2(\lambda) = \lambda \qquad \mu_3(\lambda) = \lambda$$
$$\mu_4(\lambda) = \lambda + 3\lambda^2 \qquad \mu_5(\lambda) = \lambda + 10\lambda^2$$
$$\mu_6(\lambda) = \lambda + 10\lambda^2 + 30\lambda^3 \qquad \mu_7(\lambda) = \lambda + 54\lambda^2 + 105\lambda^3 .$$

The highest power of λ in $\mu_n(\lambda)$ is $[\frac{1}{2}n]$ where the square brackets indicate the "integer part" function. The coefficients in these expansions become large very quickly as n increases.

We wish to calculate $E\left[\frac{1}{\alpha+K}\right]$ for $\alpha > 0$ which we have already done exactly as

$$E\left[\frac{1}{\alpha+K}\right] = \frac{1}{\alpha} e^{-\lambda}\ {}_1F_1(\alpha;\alpha+1;\lambda) = \frac{1}{\alpha}\ {}_1F_1(1;\alpha+1;-\lambda)$$

and asymptotically as

$$E\left[\frac{1}{\alpha+K}\right] \sim \frac{1}{\lambda}\ {}_2F_0(1-\alpha,1;;\tfrac{1}{\lambda}).$$

Stein(1966) gives an approximation which we shall extend to give an asymptotic expansion. Expanding $\frac{1}{\alpha+K}$ as

$$\frac{1}{\alpha+K} = \frac{1}{\alpha+\lambda}\left(1 + \frac{K-\lambda}{\alpha+\lambda}\right)^{-1}$$

$$= \frac{1}{\alpha+\lambda} \sum_{r=0}^{n} (-1)^r \left(\frac{K-\lambda}{\alpha+\lambda}\right)^r + (-1)^{n+1} \left(\frac{K-\lambda}{\alpha+\lambda}\right)^{n+1} \frac{1}{\alpha+K}$$

and taking expectations, we have

$$(1)\quad E\left[\frac{1}{\alpha+K}\right] = \frac{1}{\alpha+\lambda}\left\{\sum_{r=0}^{n} \frac{(-1)^r \mu_r(\lambda)}{(\alpha+\lambda)^r} + (-1)^{n+1} \frac{1}{(\alpha+\lambda)^n} E\left[\frac{(K-\lambda)^{n+1}}{K+\alpha}\right]\right\}.$$

According to Erdélyi(1956), a series $\Sigma_{r=1}^{n} a_r\phi_r(x) + R_n(x)$ is an asymptotic expansion as $x \to x_o$ if $\phi_{r+1}(x) = o(\phi_r(x))$ as $x \to x_o$ and $R_n(x) = o(\phi_n(x))$ as $x \to x_o$.

It is easy to see that the series in (1) does not converge as, after n terms with n odd the remainder is

$$\sum_{k=0}^{\infty} \left(\frac{k-\lambda}{\alpha+\lambda}\right)^{n+1} \frac{1}{\alpha+k} \frac{\lambda^k}{k!}$$

and for any $k_o > \alpha+2\lambda$, $\left(\frac{k_o-\lambda}{\alpha+\lambda}\right) > 1$ so that $\left(\frac{k_o-\lambda}{\alpha+\lambda}\right)^{n+2} \to \infty$ as $n \to \infty$. Since all the other terms are positive the remainder is greater and tends to infinity also. We wish to show that the series is an asymptotic expansion for $E\left[\frac{1}{\alpha+K}\right]$ as $\lambda \to 0$ and as $\lambda \to \infty$.

In terms of Erdélyi's definition this is not strictly true as pairs of terms of the series are of the same order as $\lambda \to \infty$. Let us group the terms in pairs and write

$$E\left[\frac{\alpha+\lambda}{\alpha+K}\right] = \sum_{r=0}^{n} \left\{ \frac{\mu_{2r}(\lambda)}{(\alpha+\lambda)^{2r}} - \frac{\mu_{2r+1}(\lambda)}{(\alpha+\lambda)^{2n+1}} \right\} + \frac{1}{(\alpha+\lambda)^{2n+1}} E\left[\frac{(K-\lambda)^{2n+2}}{K+\alpha}\right]$$

$$= 1 + \frac{(\alpha-1+\lambda)\lambda}{(\alpha+\lambda)^3} + \frac{(\alpha-1+\lambda)+(3\alpha-10+3\lambda)\lambda^2}{(\alpha+\lambda)^5} + \ldots + R_n .$$

For small λ it is easy to see that the ratio of adjacent terms is asyptotically $1/\alpha^2$ which does not tend to zero as $\lambda \to 0$. The expansion is not, in terms of Erdélyi's definition, an asymptotic expansion near $\lambda = 0$, but the terms do become rapidly smaller as r increases if $\alpha > 1$ (at least until the high order terms in λ start to dominate). We shall show however that $R_n \to 0$ as $\lambda \to 0$. We have

$$R_n < \frac{1}{(\alpha+\lambda)^{2n+1}} E\left[\frac{(K-\lambda)^{2n+2}}{\alpha}\right] = \frac{\mu_{2n+2}(\lambda)}{\alpha(\alpha+\lambda)^{2n+1}}$$

and since $\mu_{2n+2}(\lambda)$ has λ as a factor this tends to zero as $\lambda \to 0$. Thus we can caculate $E\left[\frac{\alpha+\lambda}{\alpha+K}\right]$ near $\lambda = 0$ with this series even though it is not, strictly speaking an asymptotic expansion.

For large λ we wish to show that

$$\frac{(\alpha+\lambda)\,\mu_{2r+2}(\lambda) - \mu_{2r+3}(\lambda)}{(\alpha+\lambda)^{2r+3}} = o\left(\frac{(\alpha+\lambda)\,\mu_{2r}(\lambda) - \mu_{2r+1}(\lambda)}{(\alpha+\lambda)^{2r+1}}\right)$$

and that

$$\frac{1}{(\alpha+\lambda)^{2n+1}} E\left[\frac{(K-\lambda)^{2n+2}}{K+\alpha}\right] = o\left(\frac{(\alpha+\lambda)\,\mu_{2n}(\lambda) - \mu_{2n+1}(\lambda)}{(\alpha+\lambda)^{2n+1}}\right)$$

in which case the expansion will be asymptotic in Erdelyi's sense as $\lambda \to \infty$. The first relation is equivalent to $\left(\frac{1}{\lambda^{r+1}}\right) = o\left(\frac{1}{\lambda^n}\right)$ which is true. We have to show that

$E\left[\frac{(K-\lambda)^{2n+2}}{(K+\alpha)\lambda^{2n+1}}\right] = o\left(\frac{1}{\lambda^n}\right)$, that is that $E\left[\frac{(K-\lambda)^{2n+2}}{(K+\alpha)}\right] = o\left(\lambda^{n+1}\right)$.

Now $E\left[\frac{(K-\lambda)^{2n+2}}{K+\alpha}\right] < E\left[\frac{(K-\lambda)^{2n+2}}{K-\lambda}\right] = E[(K-\lambda)^{2n+1}] = o(\lambda^{n+1})$.

Thus the expansion is an asymptotic expansion as $\lambda \to \infty$. We now wish to find an upper bound on the relative error of the expansion. Firstly $R_n > 0$ and putting $n = 0$ shows that $E\left[\frac{1}{\alpha+K}\right] > \frac{1}{\alpha+\lambda}$. The relative error $\frac{1}{\alpha+\lambda} R_n / E\left[\frac{1}{\alpha+K}\right]$ is therefore less than

$$R_n = E\left[\frac{(K-\lambda)^{2n+2}}{(\alpha+\lambda)^{2n+1}(\alpha+K)}\right]$$

$$= \frac{(\alpha+\lambda)\,\mu_{2n+2}(\lambda) - \mu_{2n+3}(\lambda)}{(\alpha+\lambda)^{2n+3}} + E\left[\frac{(K-\lambda)^{2n+4}}{(\alpha+\lambda)^{2n+4}(\alpha+K)}\right]$$

$$< \frac{(\alpha+\lambda)\,\mu_{2n+2}(\lambda) - \mu_{2n+3}(\lambda)}{(\alpha+\lambda)^{2n+3}} + \frac{\mu_{2n+4}(\lambda)}{\alpha(\alpha+\lambda)^{2n+4}} \quad .$$

We could have replaced $\alpha+K$ by α in the expression for R_n but preferred to expand to one more term and replace $\alpha+K$ by α in R_{n+1}. From either formula we can, in theory, find an upper bound for the relative error. Using Stein's approximation and the simpler error formula, we see that the relative error is less than $\frac{\lambda(1+3\lambda)}{\alpha(\alpha+\lambda)^3}$ (and probably quite close to $\frac{\lambda(1+3\lambda)}{(\alpha+\lambda)^4}$). If α is greater than $1/3$ then an upper bound on the relative error is $\frac{3\lambda}{\alpha(\alpha+\lambda)^2} = \frac{3}{\alpha(\alpha+\lambda)} - \frac{3}{(\alpha+\lambda)^2}$ giving a crude bound of $\frac{3}{\alpha^2}$ for all λ. By using better approximations to the relative error bound given above, we may obtain bounds which are not so crude as this. Note, however, that this bound is very good for large α.

Appendix 3
Some Complete Families of Distributions

A3.1 Introduction

As proofs of the completeness of the non-central χ^2 and F distributions as functions of the non-centrality parameter do not appear in many texts on statistics, we append them here. These results are proved by showing that the distributions belong to a more general complete class of distributions than the exponential family. We first prove this family to be complete.

A3.2 A Complete Family of Densities

The first result associates some complete families of densities with a complete family of densities depending on a discrete parameter and conversely.

Theorem 1 Let $q(x,\omega) = a(\omega) \sum_{k=0}^{\infty} b(k) \frac{\omega^k}{k!} p(x,k)$ where $b(k)\omega^k > 0$ for all k and for all ω in some region Ω. Suppose that $q(x,\omega)$ and $\{p(x,k)\}$ are density functions on a sample space $\${X}$.

The family of densities $\{q(x,\omega):\omega \in \Omega\}$ is complete if and only if the family $\{p(x,k):k=0,1,2,...\}$ is complete so long as Ω is a set with a point of accumulation (e.g. an uncountable set).

Proof If $E_k[f(X)]$ exists for all k then

$$\sum_{k=0}^{K} E_k\left[b(k) \frac{\omega^k}{k!} f(X)\right] \quad \text{exists for all } K$$

and by the monotone convergence theorem

$$E_\omega[f(X)] \text{ exists on } \Omega \text{ and } E_\omega[f(X)] = a(\omega) \sum_{k=0}^{\infty} b(k) \frac{\omega^k}{k!} E_k[f(X)].$$

Conversely, if $E_\omega[f(X)]$ exists on Ω then

$$a(\omega) \sum_{k=0}^{K} b(k) \frac{\omega^k}{k!} p(x,k) < q(x,\omega)$$

and by the theorem on bounded convergence

$$E_\omega[f(X)] = a(\omega) \sum_{k=0}^{\infty} b(k) \frac{\omega^k}{k!} E_k[f(X)] \text{ and } E_k[f(X)] \text{ exist for all } k.$$

Suppose that $E_k[f(X)] = 0$ for all k implies $f(x) = 0$.

This implies that if $E_\omega[f(X)] = a(\omega) \sum_{k=0}^{\infty} b(k) \frac{\omega^k}{k!} E_k[f(X)] = 0$

for $\omega \in \Omega$ and if Ω is a set with a point of accumulation, then by the uniqueness theorem for power series (Rudin(1966))

$$E_k[f(X)] = 0 \text{ for all } k \text{ and therefore } f(x) = 0.$$

Conversely, suppose that $E_\omega[f(X)] = 0 \quad \forall\, \omega \in \Omega$ implies $f(x) = 0$. This implies that if $E_k[f(X)] = 0$ for all k then

$$E_\omega[f(X)] = a(\omega) \sum_{k=0}^{\infty} b(k) \frac{\omega^k}{k!} E_k[f(X)] = 0$$

and therefore $f(x) = 0$. This completes the proof.

Corollary Let $q(x,\omega) = a(\omega) \sum_{k=0}^{\infty} \frac{\omega^k}{k!} p(x,k)$ where $p(x,k) > 0 \quad \forall\, k$ be a complete family of densities on a parameter space Ω which has a point of accumulation. If $b(k)\omega^k p(x,k) > 0$ then the density

$$Q(x,\omega) = a_1(\omega) \sum_{k=0}^{\infty} b(k) \frac{\omega^k}{k!} p(x,k)$$

is complete on Ω.

Proof The $p(x,k)$ can be normalised to be a family of density functions which will be complete if $\{q(x,\omega)\}$ is complete. The completeness of $\{Q(x,\omega)\}$ now follows from the completeness of the normalised family $\{p(x,k)\}$.

A3.3 Applications - The non-central χ^2 and F Distributions

We shall consider the completeness of these densities as functions of the non-centrality parameter λ and shall regard the degrees of freedom as fixed.

The non-central χ^2 distribution with 2ν degrees of freedom has the density function

$$f(x,\lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{x^{\nu+k-1} e^{-\frac{1}{2}x}}{2^{\nu+k} \Gamma(\nu+k)}$$

and the non-central F distribution with 2μ and 2ν degrees of freedom has the density function

$$f(x,\lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \frac{\mu^{\mu+k} \nu^{\nu}}{B(\mu+k,\nu)} \frac{x^{\mu+k-1}}{(\nu+\mu x)^{\mu+\nu+k}} .$$

Both of these expressions are of the form

$$f(x,\lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} t^k(x)\, p(x)\, a_k .$$

Introducing a coefficient $\frac{\alpha(\lambda)}{a_k}$ before each term gives

$$g(x,\lambda) = \alpha(\lambda)\, p(x)\, e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda t(x))^k}{k!}$$

which, according to our corollaries, is complete if and only if $f(x,\lambda)$ is complete. Now $g(x,\lambda) = \alpha(\lambda)\, p(x)\, e^{\lambda t(x)}$ which is a

member of the exponential family of densities and hence complete if the parameter space contains an interval.

Incidently, we can now use the theorem to show that the central χ^2, $\{\chi^2(2\nu+k):k=0,1,2,...\}$ and central F, $\{F(2\mu+k,2\nu)\}$ are complete families as fuctions of the parameter k.

The above proof also applies to non-central beta and gamma distributions, in fact they are slightly easier since the coefficients are less complicated.

Finally we note that the result in our corollary shows that the exponential family is complete on any set of parameters which has a point of accumulation. This follows from the fact that we may choose our constants in such a way that the two series in the corollary are identical.

A p p e n d i x 4
Projections and Generalised Inverses

Rao and Mitra(1971) give properties of the generalised inverse of a matrix and properties of matrices whih are almost generalised inverses in a sense which we shall make precise.

A matrix G is defined to be a generalised inverse of a matrix A if and only if $AGA = A$. A generalised inverse is not unique unless A is non-singular in which case $G = A^{-1}$. We often denote a generalised inverse of A by the symbol A^-. The following extra conditions define the unique Penrose inverse of A:

(1) (i) $GAG = G$ (ii) $(GA)^T = GA$ (iii) $(AG)^T = AG$.

A generalised inverse satisfying (ii) or (iii) gives a kind of projection matrix. More general conditions will be given.

Suppose we have a semi-inner product $\langle .,. \rangle_M$ defined by $\langle a,b \rangle_M = a^T M b$ where M is symmetric and non-negative definite. This defines a semi-norm $\|.\|$ defined by $\|a\|_M = (\langle a,a \rangle_M)^{\frac{1}{2}}$.

The necessary and sufficient condition for $x = Gb$ to be a minimum semi-norm solution to the consistent equation $Ax = b$ is that

(2) (i) $AGA = A$ and (ii) $(GA)^T M = MGA$.

One such matrix is given by

(3) $G = (M + A^T A)^- A^T \{ A(M + A^T A)^- A^T \}^-$.

If the column space of A is contained in the column space of M then a simpler solution may be taken to be

(4) $G = M^- A^T (A M^- A^T)^-$.

A minimum semi-norm solution is by definition a projection of the origin onto the solution space of the equation $Ax = b$. In general, a solution minimising $\|y - x\|_M$ is defined to be a projection of y onto the solution space and is given by

(5) $x = Gb + (I - GA)y$.

Using a semi-norm $\|.\|_N$ in the column space of A we can find the nearest vector to a solution of inconsistent equations $Ax = b$ - "the nearest" being taken to mean that the semi-norm of the residual, $b - Ax$ is to be minimised. The solution is $x = Gb$ where G satisfies

(6) (i) $NAGA = NA$ and (ii) $(AG)^T N = NAG$.

Unless N is positive definite, condition (i) is weaker than the condition for a generalised inverse. One form for G is

(7) $G = (A^T N A)^- A^T N$.

The vector AGb is the projection of b onto the column space of A.

In general, a projection, z, of y onto the column space of A is defined to be a vector, $z = Ax$, for some x, for which $\|z - y\|_N$ is a minimum. Such a projection is given by $z = P_A y$ where P_A is a matrix for which

(8) $P_A{}^T N P_A = N P_A$, $N P_A A = N A$ and rank P_A = rank A.

This can always be computed by taking $P_A = AG$ where G is a generalised inverse defined by (6) above.

If N is positive definite then the projection is an orthogonal projection in the sense that $\langle z - y, Ax \rangle_N = 0$ for all x.

REFERENCES

Alam, K. and J.R. Thompson (1964). An estimate of the mean of a multivariate normal distribution. *Technical Report*, Indiana University.

Alam, K. (1973). A family of admissible minimax estimators of the mean of a multivariate normal distribution. *Annals of Statistics* 1,517-525.

Baranchik, A.J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Annals of mathematical statistics*, 41, 642-645.

Box, G.E.P and G.C. Tiao (1973). *Bayesian Inference in statistical analysis*. Reading, Massachusetts, Addison-Wesley.

Brook, R.J. and T. Moore (1980). On the expected length of the least squares coefficient vector. *Journal of Econometrics* 12, 245-246.

Brook, R.J. and T. D. Wallace (1973). A note on extraneous information in regression. *Journal of Econometrics* 1, 315-316

Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics* 42, 855-903.

Dickey, J.M. (1968). Three multidimensional-integral identities with Bayesian applications. *Annals of Mathematical Statistics* 39, 1615-1627.

Efron, B, and C. Morris (1972). Limiting the risk of Bayes and empirical Bayes estimators - part II: the empirical Bayes case. *Journal of the American Statistical Association* 67, 130-139.

Efron, B. and C. Morris (1973 a). Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association* 68, 117-130.

Efron, B. and C. Morris (1973 b). Combining possible related estimation problems. *Journal of the Royal Statistical Society* 379-421.

Efron, B. and C. Morris (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Annals of Statistics* 4, 11-21.

Erdélyi, A. (ed) (1953). *Higher Transcendental Functions, Bateman Manuscript project*. New York, McGraw-Hill.

Erdélyi, A. (1956). *Asymptotic expansions*. New York, Dover.

Farebrother, R.W. (1975). The minimum mean square error linear estimator and ridge regression. *Technometrics* 17, 127-128.

Ferguson, T.S. (1967). *Mathematical statistics: a decision theoretic approach*. New York and London, Academic Press.

Graybill, F.A. (1976). *Theory and Applications of the linear model.* North Scituate: Massachusetts, Duxbury Press.

Hemmerle, W.J. (1975). An explicit solution for generalised ridge regression. *Technometrics* 17, 309-314.

Ince, E.L. (1939,1963). *Integration of ordinary differential equations* Edinburgh, Oliver and Boyd.

James, W. and C. Stein (1960). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 1, 361-379, Berkeley and los Angeles, 1961, University of California Press.

Jeffreys, H. (1961). *Theory of probability,(3rd Edition).* Oxford, Clarendon Press.

Kendall, M.G. and A. Stuart (1977-79). *The advanced theory of statistics.* London, Griffin.

Lehman, E.L. (1959). *Testing statistical hypotheses.* New York, Wiley.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint.* Cambridge, University Press.

Lindley, D.V. and A.F.M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society,* 1 - 41.

Moore, T. and R. J. Brook (1978). Risk estimate optimality of James-Stein estimators. *Annals of Statistics* 6, 917-919.

Pringle, R.M. and A. A. Rayner (1971). *Generalised inverse matrices with applications to statistics.* London, Griffin.

Rao, C.R. (1973). *Linear statistical inference and its applications: second edition.* New York, Wiley.

Rudin, W. (1966). *Real and complex analysis.* New York, McGraw-Hill.

Sclove, S.L., C. Morris and R. Radhakrishnan (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* 43, 1481-1490.

Searle, S.R. (1971). *Linear Models.* New York, Wiley.

Slater, L.J. (1960). *Confluent hypergeometric functions.* Cambridge, C.U.P.

Slater, L.J. (1966). *Generalised hypergeometric functions.* Cambridge, CUP.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1,* 197-206. Berkeley and Los Angeles, University of California Press.

Stein, C.M. (1962). Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society (B)* 2, 265-296.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs, *Research Papers in Statistics: Festschrift for J. Neyman*, (editor F. N. David), 351-366. New York, Wiley.

Stein, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proceedings of the Prague Symposium on Asymptotic Statistics*, 345-381.

Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics* 42, 385-388.

Strawderman, W.E. and A. Cohen (1971). Admissibility of the mean vector of a multivariate normal distribution with quadratic loss. *Annals of Mathematical Statistics* 42, 270-296.

Strawderman, W. E. (1973). Proper Bayes Minimax Estimators of the multivariate normal mean for the case of common unknown variances. *Annals of Statistics* 1, 1189-1194.

Strawderman, W. E. (1978). Minimax adaptive generalised ridge regression estimators. *Journal of the American Statistical Association* 73, 623-627.

Theil, H. (1971). *Principles of Econometrics*. New York, Wiley.

Tiao, G.C. and A. Zellner (1964). Bayes' theorem and the use of prior knowledge in regression analysis. *Biometrika* 51, 219-230.

Vinod, H.D. (1976). Simulation and extension of a minimum mean squared error estimator in comparison with Stein's. *Technometrics* 18, 491-496.

Zellner, A (1971). *An introduction to Bayesian inference in econometrics*. New York, Wiley.

Zellner, A and W. Vandaele (1972). Bayes-Stein estimates for k-means, regression and simultaneous equation models. *Presented at the Third NBER-NSF Symposium on Bayesian Inference in Econometrics.*