1982

Shrinkage estimators for multiple parameters

Richard E. Auer
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/8330
INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of “sectioning” the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106
Auer, Richard E.

SHRINKAGE ESTIMATORS FOR MULTIPLE PARAMETERS

Iowa State University

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106

Ph.D. 1982
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages
2. Colored illustrations, paper or print
3. Photographs with dark background
4. Illustrations are poor copy
5. Pages with black marks, not original copy
6. Print shows through as there is text on both sides of page
7. Indistinct, broken or small print on several pages
8. Print exceeds margin requirements
9. Tightly bound copy with print lost in spine
10. Computer printout pages with indistinct print
11. Page(s) lacking when material received, and not available from school or author.
12. Page(s) seem to be missing in numbering only as text follows.
13. Two pages numbered. Text follows.
14. Curling and wrinkled pages
15. Other
Shrinkage estimators for multiple parameters

by

Richard E. Auer

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1982
## TABLE OF CONTENTS

1. INTRODUCTION
   1.1 Criteria for Evaluating Estimators 1
   1.2 A Synopsis of Biased Estimation 4
   1.3 Review of the Work Done in the Dissertation 12

2. SHRINKAGE ESTIMATORS OF A MULTIVARIATE NORMAL MEAN VECTOR 16
   2.1 Introduction 16
   2.2 Shrinking Toward Data-Based Values with Differing Shrinking Factors 22
   2.3 Shrinking to Constant Values with a Common Shrinking Factor 30
   2.4 Shrinking to Data-Based Values with a Common Shrinking Factor 38
   2.5 The Monte Carlo Analysis 46

3. SHRINKAGE ESTIMATORS FOR THE MEAN OF A STRATIFIED NORMAL POPULATION 53
   3.1 Introduction 53
   3.2 The Development of the Estimators 55
   3.3 Component-Wise Estimators 56
   3.4 Estimators Minimizing $\text{MSE}[\sum_{i=1}^{p} \pi_i \delta_i]$ 60
      3.4.1 The common shrinking factor 60
      3.4.2 Varying shrinking factor 62
   3.5 Monte Carlo Results $p = 2$ 77

4. SHRINKAGE ESTIMATORS OF A MULTIVARIATE POISSON MEAN VECTOR 86
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.</td>
<td>SHRINKAGE ESTIMATORS OF A MULTIVARIATE GAMMA SCALE PARAMETER VECTOR</td>
<td>103</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>103</td>
</tr>
<tr>
<td>5.2</td>
<td>Bayes and Albert-Type Estimators</td>
<td>110</td>
</tr>
<tr>
<td>5.3</td>
<td>Thompson-Type Estimators</td>
<td>116</td>
</tr>
<tr>
<td>5.4</td>
<td>The Empirical Bayes Nature of the Thompson-Type Estimators</td>
<td>128</td>
</tr>
<tr>
<td>5.5</td>
<td>A Methods of Moments Empirical Bayes Estimator</td>
<td>135</td>
</tr>
<tr>
<td>5.6</td>
<td>The Monte Carlo Analysis</td>
<td>145</td>
</tr>
<tr>
<td>6.</td>
<td>SIMULTANEOUS ESTIMATION OF PARAMETERS IN EXPONENTIAL FAMILIES</td>
<td>158</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>158</td>
</tr>
<tr>
<td>6.2</td>
<td>Estimators in the Absolutely Continuous Case</td>
<td>160</td>
</tr>
<tr>
<td>6.3</td>
<td>The Normal Case with General Covariance Structure</td>
<td>171</td>
</tr>
<tr>
<td>7.</td>
<td>CONCLUSION</td>
<td>182</td>
</tr>
<tr>
<td>8.</td>
<td>BIBLIOGRAPHY</td>
<td>184</td>
</tr>
<tr>
<td>9.</td>
<td>ACKNOWLEDGEMENTS</td>
<td>187</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Values of \((c_2, c)\) for different choices of \(F_y, F_K, K\) .................................................. 35

Table 2. Mean square error of estimators \(\delta_{\bar{X}}\) and \(\delta_{\bar{X}_1}\), \(w_1 = w_2 = 1/2, V_1 = 3, V_2 = 6, p = 2, V(\bar{X}_1) + V(\bar{X}_2) = 9\) 47

Table 3. Ratio of total mean square error to \((V_1 + V_2)\) for 5 estimators where \(p = 2, \pi_1 = \pi_2 = 1/2, V_1\) and \(V_2\) known except for \(\delta_1\) and \(\delta_{1X_1}\) .................................................. 50

Table 4. Measurements of (i) contribution of bias to MSE, (ii) consistency of simulated results where \(\mu_1 = 2, \mu_2 = 4, \theta_1 = \theta_2 = 0, \pi_1 = \pi_2 = 1/2\) and \(V_1 = 3\) and \(V_2 = 6\) 52

Table 5. Ratios of MSE of proposed estimators to the variance of the usual stratified sample mean of the different choices of strata parameters 79

Table 6. Measurements of (i) bias to MSE, (ii) consistency of simulated results for estimation of \(\mu\) where \(\mu_1 = 2, \mu_2 = 4, \theta_1 = \theta_2 = 0, \pi_1 = \pi_2 = 1/2\) and \(V_1 = 3\) and \(V_2 = 6\) 85

Table 7. Thompson-type estimators for the gamma scale parameter, a list 117

Table 8. The performance of Bayesian-related estimators for large values of \(\beta\) .................................................. 146

Table 9. The effect of choosing special values for prior parameters .................................................. 148

Table 10. The performance of Bayesian-related estimators under general conditions 150
Table 11. Proportional MSEs of Thompson-type shrinkage estimators, the gamma case, where $\theta_1 = \theta_2 = 5.71$

and $\hat{\lambda}_1 = \lambda_{i\theta_i(\gamma+1)} \frac{X_i}{\gamma + 1}$ substituted for $\lambda_i$ in $c_i$,

$\hat{\lambda}_2 = \lambda_{i\theta_i} \frac{X_i}{\gamma}$ substituted for $\lambda_i$ in $c_i$,

$\hat{\lambda}_3 = \lambda_{i\theta_i} \frac{X_i}{\gamma}$ substituted for $\lambda_i$ in $c_i$.  

157
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Comparison of Thompson-type shrinking factors, the normal case</td>
<td>34</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Proportional MSEs of Thompson-type shrinkage estimators, the normal case</td>
<td>48</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Mean square error versus common stratum mean</td>
<td>84</td>
</tr>
<tr>
<td>Figure 4</td>
<td>MSE of Bayesian-related estimators, $V(\lambda) = 5.44$, $E(\lambda) = 5.71$, $\alpha = 40$, $\beta = 8$</td>
<td>152</td>
</tr>
<tr>
<td>Figure 5</td>
<td>MSE of Bayesian-related estimators, $V(\lambda) = \frac{1}{2}$, $E(\lambda) = 5.71$, $\alpha = 378$, $\beta = 67$</td>
<td>153</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Thompson-type estimators shrunk from $\frac{X_i}{Y}$, the gamma case, $\theta_1 = \theta_2 = 5.71$</td>
<td>154</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Thompson-type estimators shrunk from $\frac{X_i}{Y + 1}$, the gamma case, $\theta_1 = \theta_2 = 5.71$</td>
<td>155</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

1.1 Criteria for Evaluating Estimators

Suppose there is a random sample of size \( n \), \( \mathbf{X} = (X_1, X_2, \ldots, X_n) \), chosen from an infinite population with density function \( f(\mathbf{X} | \theta_1, \theta_2, \ldots, \theta_k) \). We are interested in the estimation of the unknown parameter \( \phi(\theta) \), where \( \theta = (\theta_1, \ldots, \theta_k) \). Note that this structure could easily be generalized to the case of a \( p \)-variate population by considering \( X_i(i = 1, 2, \ldots, n) \), and \( \phi \) as \( px1 \) vectors, but for now we consider the univariate problem.

Given an estimator of \( \phi \) based upon the \( n \) random variables, \( \hat{\phi} \), it is subject to evaluation through the utilization of many different criteria. Many of these criteria, commonly found in statistical textbooks, are discussed here to clearly lead into the motivation behind this dissertation.

Perhaps the most commonly noted property of an estimator is its bias. Under the notation given above, the bias of \( \hat{\phi} \) is

\[
E(\hat{\phi}) - \phi,
\]

where expectation is taken over the density of \( \mathbf{X} \). Should the bias of \( \hat{\phi} \) equal 0, making it an unbiased estimator, this would indicate that the center of mass for the distribution of \( \hat{\phi} \) is \( \phi \). Unbiasedness is a property of an estimator that is generally held in high esteem as evidenced by the high regard given to estimators referred to as minimum variance unbiased estimators and best linear
unbiased estimators.

The previous statement points out that, even within a group of unbiased estimators, other criteria are necessary before deciding that some particular estimator is optimal. One such criterion is $E[\hat{\phi} - \phi]^2$, a measure of the closeness of $\hat{\phi}$ to $\phi$ resulting from the distribution of $\hat{\phi}$. If $\hat{\phi}$ is unbiased, this criterion is called the variance of $\hat{\phi}$, otherwise it is referred to as the mean square error (MSE) of $\hat{\phi}$. A small variance for an unbiased $\hat{\phi}$ indicates a criterion of closeness of $\hat{\phi}$ to $\phi$. Clearly, an estimator which is unbiased and possesses the minimum possible variance of all unbiased estimators is judged highly. Also, if an estimator achieves the smallest possible variance among all linear unbiased estimators, it is again optimal with respect to both, the criteria of bias and variance.

Since $E[\hat{\phi} - \phi]^2 = E[\hat{\phi} - E(\hat{\phi})]^2 + [E(\hat{\phi}) - \phi]^2$, this criterion measures the variability of the estimator and the extent of its bias. With little variability and small bias being desired results, one hopes to find an estimator with small MSE.

This brings us to a very relevant question: "If a biased estimator has a smaller mean square error than the variance of an unbiased estimator, might the biased estimator be preferred?" Our answer to the question is affirmative and a great deal of the work in this dissertation is to propose biased estimators, to demonstrate their value and to study the estimators to see why they perform as they do.
To continue a review of criteria used to evaluate estimators, we introduce the meaning of estimators being admissible and minimax with respect to MSE. If one finds all the values of the parameter $\phi$ that makes the MSE for each of the possible estimators of $\phi$ the largest, the estimator that gives the smallest of these "worst cases" is called the minimax estimator of $\phi$. On the other hand, an estimator $\hat{\phi}$ is called admissible for $\phi$, with respect to MSE, if there does not exist an estimator that has an MSE no larger than that of $\hat{\phi}$ regardless of the value of the parameter $\phi$ and a smaller MSE for at least one value of $\phi$.

An estimator is called a maximum likelihood estimator if it can be substituted for the estimated parameter and result in maximizing the density function over all other possible data-based substitutions. In a sense, we say the maximum likelihood estimator maximizes the probability of observing the sample of size $n$ that was actually obtained. Clearly, this is another criterion of optimality worth considering.

If one were to consider that the value of the parameter $\phi$ not only be unknown, but somehow, variable or changing or if one had some prior information on the value of $\phi$ beyond the results of the vector $X$, the use of Bayesian estimation is appropriate.

This estimation procedure assumes the parameter of interest is, itself, a random variable having the density function $G(\phi)$. Using the original density of $X$ and the prior distribution $G(\phi)$, the
Bayes estimator is the \( \hat{\phi} \) that minimizes what is called the Bayes risk, namely \( \int (\hat{\phi} - \phi)^2 h(\phi | \mathbf{X}) d\phi \). Here \( h(\phi | \mathbf{X}) \) is the conditional distribution of \( \phi \) given \( \mathbf{X} \), otherwise known as the posterior distribution of \( \phi \). By considering the expectation of \( (\hat{\phi} - \phi)^2 \), the Bayesian results depend on what is called the quadratic loss function, which is the same loss function implicitly assumed when using the MSE.

If some of the prior information on \( \phi \) is uncertain in some way, one hopes the Bayes estimator is not overly dependent on this vague information. It may be best that the estimator relies mainly on the data, if the true value of \( \phi \) lies in a region where we have less reliable information. An estimator sensitive to uncertain prior information is called robust and a detailed discussion of robustness is presented by Berger (1980b).

1.2 A Synopsis of Biased Estimation

If we considered the normal density with unknown mean \( \mu \) and variance \( \sigma^2 \) as \( f(\mathbf{X} | \mu, \sigma^2) \), (\( \mu = 0, \sigma^2 = \sigma^2 \), and \( \mu = \mu \)), several accepted estimators of \( \sigma^2 \) exist, all of the form

\[
\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{k}
\]

where

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

and \( k \) is some constant.
If \( k = n - 1 \), the result is an unbiased estimator of \( \sigma^2 \) which leads to straight-forward distributional results used in methods of inference for \( \sigma^2 \). But if \( k = n \), the result is a biased maximum likelihood estimator and, if \( k = n + 1 \), the estimator minimizes the MSE as compared to any other estimator of the form \( \sigma^2 \).

This is a clear and well-known case of different estimators proving to be optimal depending upon the criterion of evaluation. In particular, it shows that unbiased estimators are not always optimal with respect to MSE.

Efron (1975) confronted the need of unbiasedness when he compared the merits of two estimators. Assuming the density of \( X_i \) is again normal, but with variance 1, the sample mean is an obvious choice as an estimator of the mean \( \theta \). But Efron proposed another estimator \( \hat{\theta}^1 \) and, upon checking its properties, found that unlike the sample mean, it wasn't unbiased, it wasn't invariant, wasn't minimax, and wasn't even admissible. But this estimator \( \hat{\theta}^1 \) will be closer to the true value of \( \theta \) more than half the time, no matter what \( \theta \) is.

Thompson (1968) again considered the estimation of the mean of a normal distribution but with no restriction on the population variance. Considering an estimator of the type \( \hat{\theta} = (1 - c)(\bar{X} - \theta) \), where \( \theta \) is some prechosen constant, he obtained the value of \( c \) that minimizes the MSE. The minimizing value of \( c \) involves \( \mu \) and \( \sigma^2 \), so in order to make the estimator useful in practice, sample estimators are substituted for the unknown parameters. Assuming \( \sigma^2 \) unknown, the
The final form of the estimator is

$$\theta + \left(1 - \frac{s^2/n}{(\bar{x} - \theta)^2 + s^2/n}\right)(\bar{x} - \theta),$$

where $s^2$ is the sample variance.

The MSE of this estimator is found to be less than that of the sample mean when $\mu$ is reasonably close to the value of $\theta$. Once again, a biased estimator is often preferable to $\bar{x}$.

Not only is Thompson's estimator another case of a biased estimator improving upon the generally accepted sample mean, but it also introduces a special form of an estimator studied extensively in this dissertation. It is a shrinkage estimator or a shrinker and it balances between two extremes, one of which is the sample mean. The amount of shrinkage away from $\bar{x}$ depends upon the value of the shrinking factor, $c$.

In Thompson's case, the estimator shrinks heavily to $\theta$ when the shrinking factor is nearly 1. If $\bar{x}$ results in being close to $\theta$ in value, this gives $\theta$ credibility. In this case, the shrinking factor is nearly 1 and the estimator complies with the well-chosen $\theta$. Should $\bar{x}$ be quite different from $\theta$, the factor nearly equals 0 and the estimator subsequently avoids $\theta$.

This type of utility is what offers the shrinkage estimator the opportunity to perform better than standard estimators. In the rest of this chapter and in the main chapters of this dissertation, this
utility of the shrinker will be of special interest and will always be noted.

Now considering the p-variate setting of the normal distribution where notationally we set the vector of sample means, $\overline{X} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)$, distributed as $N_p(\mu, \sigma^2 I)$. Stein (1956) showed that the maximum likelihood estimator of $\mu$, $\overline{X}$, is not admissible with respect to the sum of component MSEs as long as $p > 3$. Where Thompson's estimator showed improvement over $\overline{X}$ in the univariate case for part of the parameter space of $\mu$, Stein's result suggests that the vector $\overline{X}$ could be uniformly improved upon for any value of $\mu$ as long as $p > 3$. So even an unbiased maximum likelihood estimator is not optimal with respect to all criteria.

In 1961, James and Stein were able to specify an estimator which dominates the maximum likelihood estimator. The component form of the James-Stein estimator is

$$\left(1 - \frac{(p - 2)\sigma^2}{n||\overline{X}||^2}\right) \overline{X}_i$$

where

$$||\overline{X}||^2 = \sum_{i=1}^{p} \overline{X}_i^2 .$$

The form of this estimator is again that of a shrinkage estimator. If the sample means are quite different than 0, one would not likely want the estimator to shrink toward 0. In this case, the shrinker
does avoid such shrinking.

To insure that the James-Stein component estimator shrinks to a value between $\bar{X}_i$ and 0, James and Stein (1961) used a positive-part rule that replaces the estimator with $\bar{X}_i$ when the multiplier of $\bar{X}_i$ is negative. Baranchick (1964) showed that using the positive-part rule results in even further reduction of the MSE.

In the same report, Baranchick (1964) generalized the James-Stein estimator and Efron and Morris (1976) were able to find necessary and sufficient conditions that made Baranchick's estimator minimax.

Focusing on the Bayesian estimation of the same normal mean vector $\mu$, the standard prior distribution has $\mu_i \sim N(\mu_{i0}, \tau_i^2)$ for $i = 1, 2, \ldots, p$. The Bayes estimator of $\mu_i$ is then

$$\frac{n\bar{X}_i + \mu_{i0}}{\sigma^2 + \tau_i^2}.$$ Even this estimator can be expressed as a shrinkage estimator of the component form

$$\mu_{i0} + (1 - \frac{\sigma^2/n}{\sigma^2/n + \tau_i^2})(\bar{X}_i - \mu_{i0}). \quad (1.1)$$

Should the component prior variance $\tau_i^2$ be large, making the prior information in $\mu_{i0}$ weak, the Bayes estimator displays robustness as
it avoids shrinking to the prior mean $\mu_{10}$.

If the Bayesian framework was altered by assuming one or more of the prior parameters to be unknown, the question then arises whether it is possible to infer, from the set of values $\bar{X}_1, \ldots, \bar{X}_p$, the approximate form of the unknown $G$, or at least, in the present case of quadratic estimation, to approximate the value of the Bayes estimator. H. Robbins (1955) introduced his notion of empirical Bayes estimation. Using the data, through its marginal distribution, unknown prior parameters can be substituted for, and still new estimators formed.

There is a special example of an empirical Bayes estimator due to Efron and Morris (1973). They used the Bayes estimator in (1.1) where $\tau_i$ is assumed unknown and common for all $i$, $\sigma^2/n = 1$ and $\mu_{10} = \mu_{20} = \ldots = \mu_{p0} = 0$. The Bayes estimator $(1 - \frac{1}{1 + \tau^2}) \bar{X}_i$ can be made usable by considering that, marginally, $E(\bar{X}_i) = 0$ and $V(\bar{X}_i) = 1 + \tau^2$ for all $i$. This implies that

$$\frac{p}{\sum (\frac{\bar{X}_i^2}{1 + \tau^2})} \sim \chi^2(p)$$

and, since the expected value of the inverse of a chi-square random variable with $p$ degrees of freedom is $(p - 2)^{-1}$,

$$\frac{p - 2}{\sum \frac{\bar{X}_i^2}{1 + \tau^2}}$$
becomes the empirical substitute for $\frac{1}{1 + \tau^2}$. Since the ensuing estimator is exactly the James-Stein estimator, they succeeded in showing that the James-Stein estimator is an empirical Bayes estimator.

While the James-Stein estimator implicitly assumes a known prior mean and unknown prior variance, Lindley (1962) and Lindley and Smith (1972) proposed the reverse assumptions of a known prior mean $\mu_0$. By integrating out the parameter $\mu_0$ from the posterior distribution, the component estimator of $u_i$ becomes

$$\frac{n\bar{X}_i}{\sigma^2 + \bar{X}/\tau^2} = \bar{X} + \left(1 - \frac{\sigma^2/n}{\sigma^2/n + \tau^2}\right)(\bar{X}_i - \bar{X})$$

where

$$\bar{X} = \frac{1}{p} \sum_{i=1}^{p} \bar{X}_i.$$

This estimator is another shrinker and it also can be seen as an empirical Bayes estimator since $\bar{X}$ estimates $\mu_0$ through its marginal distribution.

Lindley sees this estimator as being valuable, because it compensates for extreme values of the sample means. Since extreme values of $\bar{X}_i$ tend to contradict the notion of employing a common prior mean, such compensation is reasonable.

Efron and Morris (1973) introduced a way of evaluating estimators
with regard to their Bayes risk. If we call the Bayes estimator $\delta_B$, the standard estimator $\delta_0$ and the proposed estimator $\delta$, the proposed "relative savings loss" (RSL) of $\delta$ is

$$RSL(\delta) = \frac{\text{Bayes risk (}\delta) - \text{Bayes risk (}\delta_B)}{\text{Bayes risk (}\delta_0) - \text{Bayes risk (}\delta_B)}.$$

This measurement gives the fraction of the gain found by using the Bayes estimator over the standard estimator that remains if the Bayes estimator is used instead of the proposed estimator. A small value for the RSL indicates a nearness of $\delta$ to $\delta_B$, which is known to minimize the Bayes risk.

Efron and Morris (1973) found the RSL of the James-Stein estimator to be $\frac{2}{p}$ while straight-forward two stage expectation yields an RSL of $\frac{1}{p}$ for the Lindley estimator, using the prior distribution for $\mu$ given earlier. This indicates a preference for a known prior variance over a known prior mean. With respect to Bayes risk, the data are evidently better able to estimate a prior mean through its marginal distribution.

Another estimation scheme was proposed by Albert (1981) which assumed the use of different prior parameters for each component in the estimation of the mean vector of a $p$-variate Poisson distribution. From the resulting Bayes estimator, written in the form of a shrinker, a constant common to all components is found, that minimizes the sum of the component MSEs. This constant is meant to restrict the
extent of shrinkage when the data tends to not support the information of the prior means. This is similar to the purpose of Thompson's shrinking factor and, in fact, the development of the estimators is similar. A major difference is that Albert's scheme assumes an initial Bayesian framework, so his estimator can more easily be evaluated through Bayesian considerations. Clearly Albert's estimators are more robust against misspecification of priors than the usual Bayes estimators.

While discussing biased and shrinkage estimators, it should be noted for completeness that some ridge regression estimators fall into this category. While they are developed quite differently than those estimators already noted, there is a relationship between them and empirical Bayes estimators. For a discussion of this relationship and of ridge regression estimators overall, we refer the reader to Thisted (1976).

1.3 Review of the Work Done in the Dissertation

In this dissertation, new shrinkage estimators are proposed for the multivariate setting of the normal, gamma, Poisson and other distributions within the exponential family.

Chapter 2 specifically considers the estimation of the mean vector of a p-variate normal distribution. A relationship between the Bayes estimator, Thompson's estimator and Albert's estimator is given for the univariate case. Then, simple extensions of
Thompson's estimator are developed and compared. These extensions involve considering a common shrinking factor for all components versus allowing them to differ. Also, two different statistics are used as the value the estimators shrink toward. The approach of the latter estimators parallel that taken by Lindley (1962).

Using a computer simulation study, an evaluation of these estimators demonstrates the improvements in MSE over the vector of sample means.

Chapter 3 assumes the multivariate normal distribution is actually the structure of a stratified population. Shrinkage estimators of the stratified mean are considered in ways similar to those of Chapter 2. When we consider shrinking to the sample stratified mean, the performance of the estimators depend upon the parametric structure of the strata in a subtle way. This subtleness is explained and then verified using simulation results. These results also demonstrate improvements over the usual stratified estimator under Neyman allocation.

Chapter 4 considers the estimation of the mean vector of a p-variate Poisson distribution. Again, extensions of Thompson's work are given. An improper Bayes estimator of Leonard's (1976) is shown equivalent to an empirical Bayes estimator and to a limiting Bayes estimator.

Chapter 5 uses Albert's (1981) technique to estimate the vector of scale parameters for a p-variate gamma distribution. Thompson-
type estimators are again constructed for the gamma distribution, and
an empirical Bayes interpretation is given to them.

Also in Chapter 5, an inverted gamma prior distribution is as­
sumed for the scale parameter of the gamma distribution. Using a
method of moments technique and the marginal distribution of the
data, another empirical Bayes estimator is constructed and then eval­
uated with respect to MSE and Efron and Morris' (1973) RSL. Then all
estimators in the chapter are studied through simulation results.

As one approaches the estimation of a parameter vector of a
multivariate distribution, \( \theta = (\theta_1, \theta_2, \ldots, \theta_p)' \), loss functions
can be considered other than the squared error loss function employed
through Chapter 5. In Chapter 6, loss functions

\[
L_d(\theta, a) = \sum_{i=1}^{p} d_i (\theta_i - a_i)^2
\]

and

\[
L_{d,m}(\theta, a) = \sum_{i=1}^{p} d_i (\theta_i - a_i)^2 / \theta_i^{m_i}
\]

where \( d = (d_1, d_2, \ldots, d_p)' \) and \( m = (m_1, m_2, \ldots, m_p)' \), are con­sidered. Here proposed estimators dominate standard estimators
simultaneously under losses of these forms for various constant
vectors, \( d \).

Brown (1975) and Shinozaki (1980) pursued such simultaneous
estimation for the mean of a multivariate normal distribution. We
generalize their results by proposing estimators for parameter vectors of continuous and discrete distributions within a subfamily of the general exponential family. We also propose estimators for the mean vector of a normal distribution with a more general covariance structure than the ones assumed earlier. Our method of proof uses integration by parts techniques similar to that originally discovered by Stein (1973), rather than the techniques utilized by Brown (1975) or Shinozaki (1980).

Before concluding this introductory chapter, there is an important point that needs to be made about the notation used for shrinkage estimators in Chapter 2 through 5. There is a flexibility we are assuming for this notation. Since the optimal forms of many of the shrinkage estimators require the knowledge of unknown parameters, it is necessary to consider these estimators with data-based substitutions used in place of these parameters. Although the resulting estimators are then different than the optimal shrinkage estimators, the same notation is used for both forms. Since there are so many estimators considered in the dissertation, it was decided, for the sake of simplicity, that the context of the use of the estimators would satisfactorily dictate which form we are considering. For all of the simulation studies, it should be noted, we clearly must use shrinkage estimators in their practical data-substituted form.
2. SHRINKAGE ESTIMATORS OF A MULTIVARIATE NORMAL MEAN VECTOR

2.1 Introduction

As noted in Chapter 1, many estimators for the mean vector of a multivariate distribution have appeared in the literature. Some of these estimators are Bayesian in nature, some are empirical Bayes estimators and some are not Bayesian at all. Many of these estimators are expressible as shrinkage estimators.

James Thompson (1968) devised a shrinkage estimator for the univariate population mean. In this chapter, we not only extend his estimator to the multivariate normal distribution, but also consider estimators of somewhat different forms than his. In some estimators, we allow the shrinking factor for each of the components to differ and in others consider a common shrinking factor for all components. We also utilize estimators that shrink to data-based values as opposed to some arbitrary constants.

Let us first note the usual Bayesian estimation procedure. Let the vector of sample means \( \overline{X} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)' \) be distributed p-variate normal with mean vector \( \mu = (\mu_1, \mu_2, \ldots, \mu_p)' \) and variance-covariance matrix \( \Sigma \). We assume \( \Sigma \) to be a diagonal matrix with \( i^{th} \) diagonal element \( \Sigma_{ii} = \delta_i^2 / n_i \). The multivariate normal prior distribution for \( \mu \) has mean vector \( \mu_0 = (\mu_{10}, \mu_{20}, \ldots, \mu_{p0})' \) and diagonal variance-covariance matrix \( T \) with \( i^{th} \) diagonal element \( \tau_i^2 \).
The familiar Bayesian estimator of $\mu_i$ can be expressed as a shrinkage estimator of the following form:

\[ (1 - \frac{V_i}{V_i + \tau_i^2})(\bar{x}_i - \mu_i) + \mu_i. \]

The estimator considered by Thompson assumes the value of $\mu_i$ to be nearly equal to a constant which we denote as $\theta_i$. His approach was to find the value of the shrinking factor $c_i$ that minimizes the mean square error of the estimator $(1 - c_i)(\bar{x}_i - \theta_i) + \theta_i$. The resulting estimator takes the form

\[ (1 - \frac{V_i}{V_i + (\mu_i - \theta_i)^2})(\bar{x}_i - \theta_i) + \theta_i, \]

which we denote as $\delta_{i\theta_i}$.

The notation used for this estimator demonstrates the notation employed in this and the following chapter. The first $i$ in the subscript denotes the estimation of the $i^{th}$ component mean. The second subscript, $\theta$ in this case, denotes what type of value the estimator is shrinking toward. The third subscript $i$ points out that the shrinking factor is different for each component. When a common shrinking factor is used, the third subscript is dropped.

Thompson showed that $\delta_{i\theta_i}$ has smaller mean square error (MSE) than the simple estimator $\bar{x}_i$. But a potential problem arises when we consider that the estimator depends upon the unknown value of $\mu_i$. 
The natural solution to this problem is to substitute $\bar{X}_i$ for $\mu_i$ in the estimator $5$. The ensuing estimator, while not dominating $\bar{X}_i$ as before, still performs better in many cases.

In many of the estimators created in this thesis, it will be necessary to consider substituting sample estimators for unknown parameters in order to have shrinkage estimators of practical value. It may even be necessary to substitute sample variances for unknown values of the $V_i$'s.

In the "substituted" form, however, the distributions of the resulting estimators are very hard to derive. This necessitates the use of computer simulation in studying the performance of the estimators.

Before going on to the newly constructed estimators, a relationship between the Bayesian and the Thompson estimators will be demonstrated.

If we utilize the marginal distribution derived from the previously described Bayesian framework, we find

$$\mathbb{E}\left[\frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \theta_i)^2\right] = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E}(X_{ij} - \theta_i)^2$$

$$= \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\mu(X_{ij} - \theta_i) + (\mathbb{E}(X_{ij} - \theta_i))^2\right]$$
\[
\frac{1}{n_i} \sum_{j=1}^{n_i} \left[ \text{VE}(X_{ij} - \theta_i) + \text{EV}(X_{ij} - \theta_i) + (\text{EE}(X_{ij} - \theta_i))^2 \right]
\]

\[
= \frac{1}{n_i} \sum_{j=1}^{n_i} \left[ v_{ij}^2 - e_{ij}^2 + (\epsilon_{ij}^2 + \sigma_i^2) \right]
\]

\[
= \frac{1}{n_i} \sum_{j=1}^{n_i} \left[ \tau_i^2 + \sigma_i^2 + 0 \right]
\]

\[
= \tau_i^2 + \sigma_i^2
\]

where \( X_{ij} \) is the \( j \)th observation taken from the \( i \)th component.

Since the maximum likelihood estimator of \( \sigma_i^2 \) is

\[
\frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_i)^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \theta_i + \theta_i - \bar{x}_i)^2
\]

\[
= \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \theta_i)^2 - (\bar{x}_i - \theta_i)^2
\]

we have

\[
\frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \theta_i)^2
\]
estimating $\tau_i^2 + \sigma_i^2$ in the marginal distribution sense and we have

$$\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \theta_i)^2 - (\bar{x}_i - \theta_i)^2$$

estimating $\sigma_i^2$ as the maximum likelihood estimator. Combining these two results, we have $(\bar{x}_i - \theta_i)^2$ as an estimator of $\tau_i^2$.

With this result $(\bar{x}_i - \theta_i)^2$ can be substituted for an unknown value of $\tau_i^2$ and $\theta_i$ can be viewed as a prior mean when considering the Bayesian estimator of $\mu_i$. But doing this changes the Bayesian estimator to the exact form of the Thompson estimator. Hence, the Thompson estimator can be seen as an empirical Bayes estimator as well as an MSE minimizing estimator.

James Albert (1981) considered a shrinkage estimator of a multivariate mean vector that, in a sense, used both the Bayesian and the Thompson approach. Upon expressing the Bayesian estimator as a shrinker using a different prior for each component mean, Albert found a constant multiplier of the shrinking factors that would lessen the MSE of the multivariate estimator.

In this case of the multivariate normal distribution, this would entail finding the value of $k$ that would minimize the following expression:

$$\sum_{i=1}^{p} \mathbb{E}[\left(1 - k \frac{V_i}{V_i + \tau_i^2}\right)(\bar{x}_i - \theta_i) + \theta_i - \mu_i]^2.$$
To perform this minimization, consider

\[
\frac{\partial}{\partial k} \left\{ \sum_{i=1}^{p} E \left[ (1 - k \frac{v_i}{v_i + \tau_i^2}) (x_i^\prime - \theta_i) + \theta_i - \mu_i \right]^2 \right\}
\]

\[
= \frac{\partial}{\partial k} \left\{ \sum_{i=1}^{p} \left[ v \left( (1 - k \bar{c}_i) (x_i^\prime - \theta_i) + \theta_i \right) \right. \right.
\]

\[
+ \left[ \text{bias} \left( (1 - k \bar{c}_i) (x_i^\prime - \theta_i) + \theta_i \right)^2 \right] \right\}
\]

\[
= \frac{\partial}{\partial k} \left\{ \sum_{i=1}^{p} \left[ (1 - k \bar{c}_i)^2 v_i + k^2 \bar{c}_i^2 (\mu_i - \theta_i)^2 \right] \right\}
\]

\[
= 2 \sum_{i=1}^{p} (1 - k \bar{c}_i) (-c_i v_i + k \bar{c}_i^2 (\mu_i - \theta_i)^2) = 0,
\]

where \( c_i \) is the Bayesian shrinking factor \( \frac{v_i}{v_i + \tau_i^2} \).

This implies that

\[
k = \frac{\sum_{i=1}^{p} c_i v_i}{\sum_{i=1}^{p} [c_i^2 (v_i + (\mu_i - \theta_i)^2)]}
\]

\[
= \frac{\sum_{i=1}^{p} v_i \left( \frac{v_i}{v_i + \tau_i^2} \right)}{\sum_{i=1}^{p} \left( \frac{v_i}{v_i + \tau_i^2} \right)^2 (v_i + (\mu_i - \theta_i)^2)}
\]
If the value of $\tau_i^2$ was unknown and again estimated by $(\bar{X}_i - \theta_i)^2$, k would take on the value 1 making the Albert-type estimator equivalent to the Thompson estimator.

2.2 Shrinking Toward Data-Based Values with Differing Shrinking Factors

When estimating the mean vector of a multivariate distribution, one needs to clearly state the criterion upon which the goodness of the estimator will be judged. In this chapter, the estimators are judged on the basis of the sum of its component MSEs, so estimators are developed that minimize this total.

If we consider estimating the component $\mu_i$'s with estimators of the type $(1 - c_i)(\bar{X}_i - \theta_i) + \theta_i$ where $\theta_i$ is again a constant believed to be nearly equal to $\mu_i$ in value, the result is an estimator already discussed. Taking the partial derivative of the total MSE with respect to the value $c_i$ results in making the $i^{th}$ component estimator the Thompson estimator, denoted earlier as $\delta_i \theta_i$.

So to truly develop new estimators, we consider shrinking to some data-based values. The first of these data-based values is the mean of the entire sample drawn, which is of the form

$$\bar{X} = \frac{p}{\Sigma n_i \bar{X}_i} / \Sigma n_i = \frac{p}{\Sigma w_i \bar{X}_i}$$

where $w_i = n_i / \Sigma n_j$. 
The estimator of the component $u_i$ is then $(1 - c_i)(\bar{x}_i - \bar{x}) + \bar{x}$ which we denote as $\delta_{i\bar{x}_i}$. Here the notation $\delta_{i\bar{x}_i}$ indicates a component estimator shrinking to $\bar{x}$ while utilizing a different shrinking factor than those for other components.

With estimators of this type, no pre-conceived values for the $u_i$'s are needed and no Bayesian prior is used either. Should the values of the $u_i$'s be nearly equal, one would expect that shrinking toward an overall mean $\bar{x}$ would prove beneficial. Should the choices of the $\theta_i$'s be very poor, one may expect $\delta_{i\bar{x}_i}$ to perform better than $\delta_{i\theta_i}$ even if the values of the $u_i$'s are widely dispersed.

In order to determine the optimal choice of the $c_i$'s for this estimator, one must minimize $\sum_{i=1}^{p} \text{MSE}(\delta_{i\bar{x}_i})$ where

$$\text{MSE}(\delta_{i\bar{x}_i}) = E[(1 - c_i)(\bar{x}_i - \bar{x}) + \bar{x} - u_i]^2$$

$$= E[(1 - c_i)^2(\bar{x}_i - \bar{x})^2] + E[\bar{x} - u_i]^2$$

$$+ 2(1 - c_i) E[(\bar{x}_i - \bar{x})(\bar{x} - u_i)].$$

We find

$$-\frac{\partial}{\partial c_i} \left[ \sum_{i=1}^{p} \text{MSE}(\delta_{i\bar{x}_i}) \right] = 2(1 - c_i) E(\bar{x}_i - \bar{x})^2$$

$$+ 2E[(\bar{x}_i - \bar{x})(\bar{x} - u_i)] = 0,$$
which implies

\[ 1 - c_i = \frac{E[(\bar{X}_i - \bar{X})(\mu_i - \bar{X})]}{E(\bar{X}_i - \bar{X})^2} \]

\[ = \frac{\mu_i(\mu_i - \bar{\mu}) - E[X(\bar{X}_i - \bar{X})]}{V(\bar{X}_i - \bar{X}) + [E(\bar{X}_i - \bar{X})]^2} \]

\[ = \frac{\mu_i(\mu_i - \bar{\mu}) - E(XX_i) + V(\bar{X}) + (E(\bar{X}))^2}{V(1 - w_i)X_i - \sum_{j \neq i} \sum_{j} w_j \bar{X}_j + \mu_i^2 + V(\bar{X})} \]

\[ = \frac{\mu_i(\mu_i - \bar{\mu}) - \sum_{j} w_j \mu_j - w_i \bar{X}_i + \mu_i^2 + V(\bar{X})}{(1 - w_i)^2 \bar{X}_i - \sum_{j \neq i} \sum_{j} w_j \bar{X}_i + \mu_i^2 + V(\bar{X})} \]

\[ = \frac{(\mu_i - \bar{\mu})^2 + V(\bar{X}) - \bar{X}_i \bar{X}_i}{(\mu_i - \bar{\mu})^2 + V(\bar{X}) + (1 - 2w_i)\bar{X}_i} \]

\[ = 1 - \frac{(1 - w_i)v_i}{(\mu_i - \bar{\mu})^2 + V(\bar{X}) + (1 - 2w_i)v_i} \], \quad (2.1) \]

where \( \bar{\mu} = \sum_{i=1}^{p} w_i \mu_i \).

To clearly demonstrate the superiority of using the component estimator \( \delta_{\bar{X}_i} \) over simply using \( \bar{X}_i \), consider
\[
\sum_{i=1}^{p} V(\bar{X}_i) - \sum_{i=1}^{p} \text{MSE}(\delta_i \bar{X}_i) = \sum_{i=1}^{p} V_i - \sum_{i=1}^{p} \left[ \text{MSE}\left(1 - c_i (1 - \omega_i) \bar{X}_i \right) \right]
\]

\[
= \sum_{i=1}^{p} V_i - \sum_{i=1}^{p} \left[ \text{MSE}\left(1 - c_i (1 - \omega_i) \bar{X}_i \right) \right]
\]

\[
+ c_i \sum_{j \neq i}^{p} w_j \bar{X}_j \]

\[
= \sum_{i=1}^{p} V_i - \sum_{i=1}^{p} \left[ (1 - c_i (1 - \omega_i))^2 V_i \right]
\]

\[
+ c_i \sum_{j \neq i}^{p} w_j^2 V_j + (\mu_i (1 - c_i (1 - \omega_i)))
\]

\[
+ c_i \sum_{j \neq i}^{p} w_j \mu_i - \mu_i^2 \]

\[
= \sum_{i=1}^{p} V_i - \sum_{i=1}^{p} \left[ \sum_{i=1}^{p} V_i - 2c_i (1 - \omega_i) V_i \right]
\]

\[
+ c_i (V_i (1 - \omega_i)^2 + \sum_{j \neq i}^{p} w_j^2 V_j + (\mu_i - \mu)^2) \]

\[
= \sum_{i=1}^{p} V_i - \sum_{i=1}^{p} V_i
\]
The final expression not only verifies the superiority of \( \hat{\beta}_{\overline{x}_1} \), but also gives us a measurement of the difference between the two multivariate estimators.

From the form of 1 - \( c_i \) for this estimator, we note that

\[(\mu_i - \overline{\mu})^2 + V(\overline{x}) - w_i V_i < (\mu_i - \overline{\mu})^2 + V(\overline{x}) + (1 - 2w_i) V_i \]

forces 1 - \( c_i < 1 \) and hence \( c_i > 0 \). Whenever \( w_i V_i < \sqrt{\overline{V}} + (\mu_i - \overline{\mu})^2 \), \( c_i \) is also bounded from above by 1 making \( \hat{\beta}_{\overline{x}_i} \) a true shrinkage estimator which shrinks between \( \overline{x}_i \) and \( \overline{x} \).

In the rare case of \( c_i \) exceeding 1, we simply define \( \hat{\beta}_{\overline{x}_i} \) as the estimator resulting by replacing \( c_i \) by 1.

The practice of forcing the value of the shrinking factor into the interval [0, 1] is not only an intuitively pleasing practice, but is also a common practice for shrinkage estimators in the literature.

When we replace unknown parameters in shrinking factors by sample estimators and study the MSE of the estimators through simulation results, we also will take the precaution of insuring the value of shrinking factors to be in the interval [0, 1].
We look at the special case of $p = 2$ in order to study the estimator structurally. In this case (2.1), becomes

\[ c_1 = \frac{(1 - w_1)V_1}{\left(\mu_1 - \bar{\mu}\right)^2 + V(\bar{X}) + (1 - 2w_1)V_1} \]

\[ = \frac{w_2V_1}{[w_2(\mu_1 - \mu_2)^2 + w_2V_1 + w_2V_2]} \]

\[ = \frac{V_1}{w_2[(\mu_1 - \mu_2)^2 + V_1 + V_2]} \]

and

\[ c_2 = \frac{V_2}{w_1[(\mu_1 - \mu_2)^2 + V_1 + V_2]} \]

From these expressions, one can see that when $V_i$ is small, $c_1$ is also small. Algebraically this means $\delta_{iX_1}$ will tend toward $\bar{X}_i$ which is a reasonable result as $\bar{X}_i$ is a good estimator of $\mu_i$ when $V_i$ is small.

When the $\mu_i$'s are of nearly equal value, the shrinking factors will become large, consequently shifting $\delta_{iX_1}$ toward $\bar{X}$. But, as mentioned earlier, one would suspect shrinking toward the overall mean to prove beneficial when the component means are nearly equal.

Another data-based value to consider shrinking towards in the case of $p = 2$ and $\mu_1 = \mu_2 = \mu$ is the estimator of the form $d_1\bar{X}_1 + d_2\bar{X}_2$
where \( d_1 + d_2 = 1 \). This unbiased estimator for a common mean was discussed in the literature by Graybill and Deal (1959). The value of \( d_1 \) that minimizes the variance is \( V_j/(V_1 + V_j) \) for \( i \neq j \). The subsequent variance is \( V_1 V_2/(V_1 + V_2) \).

We call this estimator \( \hat{\mu} \) and wish to compare it to the estimator \( \delta_{i \bar{X}_i} \) in the case of two equal component means:

\[
\delta_{i \bar{X}_i} = (1 - \frac{V_1}{w_2(V_1 + V_2)})(\bar{X}_1 - \Sigma_{i=1}^2 w_i \bar{X}_i) + \bar{X}
\]

\[
= \frac{(w_2 - 1)V_1 + w_2V_2}{w_2(V_1 + V_2)}(\bar{X}_1(1 - w_1) - w_2 \bar{X}_2) + \frac{V_1 + V_2}{V_1 + V_2} \bar{X}
\]

\[
= \frac{(w_2V_2 - w_1V_1)(\bar{X}_1 - \bar{X}_2) + (V_1 + V_2)(w_1 \bar{X}_1 + w_2 \bar{X}_2)}{V_1 + V_2}
\]

\[
= \frac{\bar{X}_1V_2(w_1 + w_2) + \bar{X}_2V_1(w_1 + w_2)}{V_1 + V_2}
\]

\[
= \frac{\bar{X}_1V_2}{V_1 + V_2} + \frac{\bar{X}_2V_1}{V_1 + V_2} = \hat{\mu}.
\]

Hence, the two estimators are equivalent.

Utilizing \( \hat{\mu} \) in the construction of another possible estimator of the normal mean vector, consider \((1 - c_1)(\bar{X}_1 - \hat{\mu}) + \hat{\mu}\).

Using the fact that
\[ \text{MSE}(1 - c_1)(\bar{x}_1 - \hat{\mu}) + \hat{\mu}] = (1 - c_1)^2 E(\bar{x}_1 - \hat{\mu})^2 + E(\hat{\mu} - \mu)^2 + 2(1 - c_1) E[(\hat{\mu} - \mu)(\bar{x}_1 - \hat{\mu})] , \]

the partial derivative of this MSE with respect to \( c_1 \) equals

\[ -2(1 - c_1) E(\bar{x}_1 - \hat{\mu})^2 - 2E[(\hat{\mu} - \mu)(\bar{x}_1 - \hat{\mu})] . \]

Setting this equal to zero implies that

\[ c_1 = \frac{E(\bar{x}_1 - \hat{\mu})^2 + \mu^2 - V(\hat{\mu}) + E(\hat{\mu})}{E(\bar{x}_1 - \hat{\mu})^2} \]

\[ = \frac{\mu^2 + v_1 + \mu^2 + V(\hat{\mu}) - 2(\mu^2 + \frac{v_1 v_2}{v_1 + v_2}) - \mu^2 - V(\hat{\mu}) + (\mu^2 + \frac{v_1 v_2}{v_1 + v_2})}{\mu^2 + v_1 + \mu^2 + V(\hat{\mu}) - 2(\mu^2 + \frac{v_1 v_2}{v_1 + v_2})} \]

\[ = \frac{v_1 - V(\hat{\mu})}{v_1 - V(\hat{\mu})} = 1. \]

When \( c_1 = 1 \) for \( i = 1, 2 \), the shrinkage estimator is simply \( \hat{\mu} \) itself. Therefore, such a shrinkage estimator is unable to improve upon \( \hat{\mu} \).
2.3 Shrinking to Constant Values with a Common Shrinking Factor

If the shrinking factors for all p components were restricted to be the same value, the form and performance of each of the previously studied estimators would change. Since the minimization would be conducted by taking the partial derivative with respect to a common shrinking factor, the final form of the factor would involve information from all p components. In many cases, this would restrict the flexibility of the shrinkage estimators.

The flexibility offered by allowing the shrinking factors to differ from component to component is most likely to be of special value when the individual components differ with respect to their parametric structure. Forcing the commonality of the shrinking factors would then be appropriate when the components are quite similar. In these cases, one would find the component means and variances to be of nearly equal value and we would define the components as being "balanced."

One estimator utilizing a common shrinking factor is the counterpart to the Thompson estimator which is denoted by $\delta_{1\theta}$ and is of the component form $(1 - c)(\bar{x}_i - \theta_i) + \theta_i$. To minimize

$$\sum_{i=1}^{p} \text{MSE}(\delta_{1\theta}) = \sum_{i=1}^{p} [(1 - c)^2 \nu_i + c^2 (\theta_i - \mu_i)^2]$$

$$= (1 - c)^2 \sum_{i=1}^{p} \nu_i + c^2 \sum_{i=1}^{p} (\theta_i - \mu_i)^2,$$
the partial derivative with respect to $c$ is taken and

$$\frac{3}{3c} \sum_{i=1}^{p} \text{MSE}(\hat{\theta}_{i\theta}) = -2(1 - c) \sum_{i=1}^{p} V_{i} + 2c \sum_{i=1}^{p} (\theta_{i} - \mu_{i})^{2} = 0$$

implies that

$$c = \frac{\sum_{i=1}^{p} V_{i}}{\sum_{i=1}^{p} V_{i} + \sum_{i=1}^{p} (\theta_{i} - \mu_{i})^{2}}.$$ 

Regardless of the choice for the $\theta_{i}$'s, estimation with component estimator $\hat{\delta}_{i\theta}$ gives smaller total MSE than does the use of the component sample means alone. Note that

$$\sum_{i=1}^{p} \frac{V_{i}(\bar{x}_{i})}{\text{MSE}(\hat{\theta}_{i\theta})} = \sum_{i=1}^{p} V_{i} - (1 - c)^{2} \sum_{i=1}^{p} V_{i} - c^{2} \sum_{i=1}^{p} (\theta_{i} - \mu_{i})^{2}$$

$$= \sum_{i=1}^{p} V_{i} - \frac{\sum_{i=1}^{p} (\theta_{i} - \mu_{i})^{2}^{2}}{\sum_{i=1}^{p} (V_{i} + (\theta_{i} - \mu_{i})^{2})^{2}}$$

$$- \frac{\sum_{i=1}^{p} V_{i}^{2}}{(\sum_{i=1}^{p} (V_{i} + (\theta_{i} - \mu_{i})^{2}))^{2}} \sum_{i=1}^{p} (\theta_{i} - \mu_{i})^{2}.$$
The previous expression shows that the worse the choice of the \( \theta_i \)'s, the less the improvement of using \( \delta_{i\theta} \) our \( \bar{X}_i \). For a poor choice of the \( \theta_i \)'s the value of \( c \) becomes very small and the component estimators tend to avoid the value of the \( \theta_i \)'s.

In order to compare estimating with \( \delta_{i\theta_1} \) and \( \delta_{i\theta_2} \), we consider the case \( p = 2 \) and assume the values of the component means and variances are known.

To enable a simple look at the comparable structures of these estimators, the following constants are defined:

\[
K_1 = \frac{(\mu_1 - \theta_1)^2}{V_1}
\]

\[
K_2 = \frac{(\mu_2 - \theta_2)^2}{V_1}
\]
Using these constants, the shrinking factors of $\delta_i\theta_i$ become

$$c_1 = \frac{V_1}{V_1 + (\mu_1 - \theta_1)^2} = \frac{V_1}{V_1 + K_1 V_1} = \frac{1}{1 + K_1}$$

and

$$c_2 = \frac{V_2}{V_2 + (\mu_2 - \theta_2)^2} = \frac{V_2}{V_2 + K_2 V_1} = \frac{F_V V_1}{F_V V_1 + K_2 V_1} = \frac{F_V}{F_V + K_2}$$

$$= \frac{F_V}{F_V + F_K K_1} .$$

The common shrinking factor of $\delta_i\theta_i$ is

$$c = \frac{V_1 + V_2}{V_1 + V_2 + K_1 V_1 + K_2 V_1} = \frac{V_1 (1 + F_V)}{V_1 (1 + F_V) + V_1 (K_1 + K_2)}$$

$$= \frac{1 + F_V}{1 + F_V + K_1 (1 + F_K)} .$$

The graphs on Figure 1 point out the values of $c_1$, $c_2$ and $c$ where combinations of $F_V$, $F_K$ and $K_1$ are chosen and Table 1 lists the numerical values of these factors.
Figure 1. Comparison of Thompson-type shrinking factors, the normal case. (Plotted points are \((c_2, c)\) for estimators \(\delta_i\theta_i\) and \(\hat{\delta}_i\theta_i\) and spiked values are \(c_1\) of \(\delta_i\theta_i\) for various choices of \(f_v, K_1, f_{K_1}\).)
Table 1. Values of \((c_2, \ c)\) for different choices of \(F_Y\), \(F_K\), \(K\)

<table>
<thead>
<tr>
<th>(F_K)</th>
<th>(K = \frac{1}{10}), (c_1 = .909)</th>
<th>(K = 2,\ c_1 = .333)</th>
<th>(K = 10,\ c_1 = .091)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\infty)</td>
<td>((0.000, 0.000))</td>
<td>((0.000, 0.000))</td>
<td>((0.000, 0.000))</td>
</tr>
<tr>
<td>10.00</td>
<td>((0.200, 0.532))</td>
<td>((0.012, 0.054))</td>
<td>((0.002, 0.011))</td>
</tr>
<tr>
<td>3.00</td>
<td>((0.454, 0.758))</td>
<td>((0.040, 0.135))</td>
<td>((0.008, 0.030))</td>
</tr>
<tr>
<td>1.50</td>
<td>((0.625, 0.833))</td>
<td>((0.077, 0.200))</td>
<td>((0.016, 0.048))</td>
</tr>
<tr>
<td>(F_Y = \frac{1}{4})</td>
<td>(1.00)</td>
<td>((0.714, 0.862))</td>
<td>((0.111, 0.238))</td>
</tr>
<tr>
<td>&amp; (0.50)</td>
<td>((0.833, 0.893))</td>
<td>((0.200, 0.200))</td>
<td>((0.048, 0.077))</td>
</tr>
<tr>
<td>&amp; (0.25)</td>
<td>((0.909, 0.909))</td>
<td>((0.333, 0.333))</td>
<td>((0.090, 0.091))</td>
</tr>
<tr>
<td>&amp; (0.00)</td>
<td>((1.000, 0.926))</td>
<td>((1.000, 0.385))</td>
<td>((1.000, 0.111))</td>
</tr>
<tr>
<td>(\infty)</td>
<td>((0.000, 0.000))</td>
<td>((0.000, 0.000))</td>
<td>((0.000, 0.000))</td>
</tr>
<tr>
<td>10.00</td>
<td>((0.500, 0.645))</td>
<td>((0.048, 0.083))</td>
<td>((0.010, 0.018))</td>
</tr>
<tr>
<td>3.00</td>
<td>((0.769, 0.833))</td>
<td>((0.143, 0.200))</td>
<td>((0.032, 0.048))</td>
</tr>
<tr>
<td>1.50</td>
<td>((0.869, 0.888))</td>
<td>((0.250, 0.286))</td>
<td>((0.062, 0.074))</td>
</tr>
<tr>
<td>(F_Y = 1)</td>
<td>(1.00)</td>
<td>((0.909, 0.909))</td>
<td>((0.333, 0.333))</td>
</tr>
<tr>
<td>&amp; (0.50)</td>
<td>((0.952, 0.930))</td>
<td>((0.500, 0.400))</td>
<td>((0.167, 0.118))</td>
</tr>
<tr>
<td>&amp; (0.25)</td>
<td>((0.976, 0.941))</td>
<td>((0.667, 0.444))</td>
<td>((0.286, 0.138))</td>
</tr>
<tr>
<td>&amp; (0.00)</td>
<td>((1.000, 0.952))</td>
<td>((1.000, 0.500))</td>
<td>((1.000, 0.143))</td>
</tr>
<tr>
<td>(\infty)</td>
<td>((0.000, 0.000))</td>
<td>((0.000, 0.000))</td>
<td>((0.000, 0.000))</td>
</tr>
<tr>
<td>10.00</td>
<td>((0.800, 0.820))</td>
<td>((0.167, 0.185))</td>
<td>((0.038, 0.077))</td>
</tr>
<tr>
<td>3.00</td>
<td>((0.930, 0.926))</td>
<td>((0.400, 0.385))</td>
<td>((0.117, 0.167))</td>
</tr>
<tr>
<td>1.50</td>
<td>((0.964, 0.952))</td>
<td>((0.571, 0.500))</td>
<td>((0.211, 0.222))</td>
</tr>
<tr>
<td>(F_Y = 4)</td>
<td>(1.00)</td>
<td>((0.976, 0.962))</td>
<td>((0.667, 0.556))</td>
</tr>
<tr>
<td>&amp; (0.50)</td>
<td>((0.988, 0.970))</td>
<td>((0.800, 0.625))</td>
<td>((0.444, 0.286))</td>
</tr>
<tr>
<td>&amp; (0.25)</td>
<td>((0.994, 0.976))</td>
<td>((0.888, 0.667))</td>
<td>((0.615, 0.308))</td>
</tr>
<tr>
<td>&amp; (0.00)</td>
<td>((1.000, 0.980))</td>
<td>((1.000, 0.714))</td>
<td>((1.000, 0.333))</td>
</tr>
</tbody>
</table>
The values chosen for $F_V$ are 1, $1/4$ and 4. The value of $F_V$ indicates the relative size of the component variances.

The values chosen for $F_K$ are $\infty$, 10, 3, 3/2, 1, $1/2$, $1/4$ and 0. The value of $F_K$ indicates the relative merit of the choices of the $\theta_i$'s.

The values chosen for $K_1$ are 10, 2 and $1/10$. $K_1$ represents a comparison between the first component variance and the appropriateness of $\theta_1$.

The constant $K_2$ is an automatic consequence of the values for $F_V$, $F_K$ and $K_1$ and, hence, its values are not specifically pointed out.

As the value of $c_2$ increases on any of the nine graphs making up Figure 1, the plotted points correspond to the descending values of $F_K$. For example, in graph 1 $(c_2, c) = (0, 0)$ is the left-most point. This corresponds to the case of $F_K = \infty$, which in turn signifies that $|\mu_2 - \theta_2|$ tends to infinity or that $\mu_1 - \theta_1$ tends to zero. Since $K_1 \neq 0$ for graph 1, the former is the case. It is reasonable that $c_2$ should equal 0 as this allows the estimator of $\mu_2$ to disregard the poor choice of $\theta_2$. The fact that $c = 0$ means that this infinitely poor choice of $\theta_2$ is avoided by $\delta_{29}$ as well.

Proceeding to larger values of $c_2$ in graph 1 and hence smaller values of $F_K$, eventually $F_K = 0$ is considered. Since $K_1 \neq \infty$, the value of $F_K$ equals 0 only when $\mu_2 = \theta_2$ in which case the estimators should shrink heavily to $\theta_2$. As $c_2 = 1$, $\delta_{29i}$ does shrink totally to this perfect choice of $\theta_2$. But $c \neq 1$ for the estimator $\delta_{29}$. This
demonstrates the utility of allowing different shrinking factors for each component.

Although the value of $\lambda_1 = 1/10$ is small, $\hat{\delta}_{1\theta}$ cannot allow $c$ to equal 1 as this would shrink $\overline{x}_1$ totally to $\theta_1$. The spike on the $c_2$ axis of graph 1 indicates the value of $c_1$ to be .926, showing that $\hat{\delta}_{1\theta_1}$ shrinks heavily toward the good choice of $\theta_1$.

The utility of $\hat{\delta}_{1\theta_1}$ can also be seen from graph 1 when considering that the value of $c_1$ is constant because $\frac{(\mu_1 - \theta_1)^2}{V_1}$ is held constant. Changing $(\mu_2 - \theta_2)^2$ and hence $F_K$, should then affect only the estimation of $\mu_2$. Since $c_1$ remains constant while $c_2$ and $F_K$ changes, the estimator $\hat{\delta}_{1\theta_1}$ does perform in this way. The estimator $\hat{\delta}_{1\theta}$, however, takes on different values of $c$ as $F_K$ changes.

When $F_V = F_K = 1$, the component variances are equal and the choice of $\theta_1$ and $\theta_2$ are equally good making the components balanced. In this case, one would suspect that allowing shrinking factors to vary over the components will not offer improvement over $\hat{\delta}_{1\theta}$. Graphs 4 through 6 and Table 1 show that $c_1 = c_2 = c$ in this case.

In such a balanced arrangement, assume it is known that $\mu_1 = \mu_2 = \mu$, $V_1 = V_2$, $\theta_1 = \theta_2 = \theta$ and that $w_1 = w_2 = 1/2$. Now the value of

$$c = \frac{V_1 + V_2}{2} = \frac{2V}{2 + 2(\mu - \theta)^2}$$
demonstrating the equality of the estimators.

With the above assumptions, the mean square error of \( \hat{\mu} \) would equal \( \frac{V}{2} \) and the \( \text{MSE}(\delta_i, \theta) = \frac{(u - \theta)^2 V}{V + (u - \theta)^2} \). The estimator \( \delta_i, \theta \), therefore, can even out-perform \( \hat{\mu} \) if \((u - \theta)^2 < V\). This happens if the choice of \( \theta \) is sufficiently good.

2.4 Shrinking to Data-Based Values with a Common Shrinking Factor

The estimator shrinking to \( \bar{\mu} \) through the use of a common shrinking factor for all components is denoted by \( \delta_\bar{\mu} \) and is of the component form \((1 - c)(\bar{X}_i - \bar{\mu}) + \bar{\mu}\). Note that

\[
\sum_{i=1}^{p} \text{MSE}(\delta_\bar{\mu}) = \sum_{i=1}^{p} \text{MSE}[(1 - c)\bar{X}_i + c\bar{\mu}]
\]

\[
= \sum_{i=1}^{p} \text{MSE}[(1 - c + cw_i)\bar{X}_i + c\sum_{j \neq i} w_j \bar{X}_j]
\]

\[
= \sum_{i=1}^{p} \text{MSE}[(1 - c(1 - w_i))\bar{X}_i + c\sum_{j \neq i} w_j \bar{X}_j]
\]

\[
= \sum_{i=1}^{p} [(1 - c(1 - w_i))^2 v_i + c^2 \sum_{j \neq i} w_j v_j]
\]
\[ + \sum_{i=1}^{p} \left[ (1 - c(1 - w_i)) \mu_i + c \sum_{j \neq 1}^{p} w_j \mu_j - \mu_i \right]^2 \]

\[ = \sum_{i=1}^{p} \left[ V_i - 2c(1 - w_i) V_i + c^2(1 - w_i)^2 V_i + c^2 \sum_{j \neq 1}^{p} w_j^2 V_j \right] \]

\[ + c^2 \sum_{i=1}^{p} \sum_{j=1}^{p} w_i \mu_j - \mu_i \right]^2 \]

\[ = \sum_{i=1}^{p} \left[ V_i - 2c(1 - w_i) V_i + c^2 V(X) + c^2 V_i(1 - 2w_i) \right] \]

\[ + c^2 \sum_{i=1}^{p} (\mu_i - \mu)^2 . \]

Now

\[ \frac{\partial}{\partial c} \sum_{i=1}^{p} \text{MSE}(\delta_i X) = -2 \sum_{i=1}^{p} (1 - w_i) V_i + 2c \sum_{i=1}^{p} V(X) + 2c \sum_{i=1}^{p} V_i(1 - 2w_i) \]

\[ + 2c \sum_{i=1}^{p} (\mu_i - \mu)^2 = 0 \]

implies that

\[ c = \frac{\sum_{i=1}^{p} (1 - w_i) V_i}{p V(X) + \sum_{i=1}^{p} (V_i(1 - 2w_i) + (\mu_i - \mu)^2)} . \]
This multivariate estimator of \( \mu \) gives smaller total MSE than the usual means as the following shows:

\[
\frac{\sum_{i=1}^{p} V(\bar{X}_i) - \sum_{i=1}^{p} \text{MSE}(\hat{\mu}_i)}{pV(\bar{X}) + \sum_{i=1}^{p} [V_i(1 - 2w_i) + (\mu_i - \bar{\mu})^2]} = \frac{\sum_{i=1}^{p} V_i - \sum_{i=1}^{p} V_i + 2c \sum_{i=1}^{p} (1 - w_i)V_i}{pV(\bar{X}) + \sum_{i=1}^{p} [V_i(1 - 2w_i) + (\mu_i - \bar{\mu})^2]} \]

\[
= \frac{2[\sum_{i=1}^{p} (1 - w_i)V_i]}{pV(\bar{X}) + \sum_{i=1}^{p} [V_i(1 - 2w_i) + (\mu_i - \bar{\mu})^2]} \]

\[
\geq 0. \quad (2.2)
\]
When one considers that this estimator does not have the flex-
ibility to shrink to \( \bar{X} \) differently for each component, it seems that
\( \delta_{\bar{X}} \) would have a greater total mean square error than does \( \delta_{\bar{X}_i} \).
This is verified for \( p = 2 \) as follows

\[
\begin{align*}
\sum_{i=1}^{2} \text{MSE}(\delta_{\bar{X}}) - \sum_{i=1}^{2} \text{MSE}(\delta_{\bar{X}_i}) &= \sum_{i=1}^{2} \left( \sum_{i=1}^{2} (1 - w_i) \text{V}_i \right)^2 \\
&= \sum_{i=1}^{2} \text{V}_i - \frac{\left( \sum_{i=1}^{2} (1 - w_i) \text{V}_i \right)^2}{2 \text{V}(\bar{X}) + \sum_{i=1}^{2} \left[ \text{V}_i (1 - 2w_i) + (\mu_i - \bar{\mu})^2 \right]} \\
&- \left( \sum_{i=1}^{2} \text{V}_i - \sum_{i=1}^{2} \frac{(1 - w_i)^2 \text{V}_i^2}{\text{V}(\bar{X}) + \text{V}_i (1 - 2w_i) + (\mu_i - \bar{\mu})^2} \right) \\
&= \frac{w_2^2 \text{V}_1^2}{w_2^2 (\mu_1 - \mu_2)^2 + \text{V}_1 (1 - 2w_1) + \text{V}(\bar{X})} \\
&+ \frac{w_1^2 \text{V}_2^2}{w_1^2 (\mu_1 - \mu_2)^2 + \text{V}_2 (1 - 2w_2) + \text{V}(\bar{X})} \\
&- \frac{(w_2 \text{V}_1 + w_1 \text{V}_2)^2}{2 \text{V}(\bar{X}) + \text{V}_1 (1 - 2w_1) + \text{V}_2 (1 - 2w_2) + w_2^2 (\mu_1 - \mu_2)^2 + w_1^2 (\mu_1 - \mu_2)^2}
\end{align*}
\]
\[
\frac{1}{(v_1^2 + v_2^2 + (\mu_1 - \mu_2)^2)} \left[ v_1^2 + v_2^2 - \frac{w_2^2v_1^2 + 2w_1w_2v_1v_2 + w_1^2v_2^2}{w_1^2 + w_2^2} \right]
\]

\[
= \frac{1}{(w_1^2 + w_2^2)(v_1^2 + v_2^2 + (\mu_1 - \mu_2)^2)} \left[ w_1^2v_1^2 + w_2^2v_2^2 - 2w_1w_2v_1v_2 \right]
\]

\[
= \frac{(w_1^2v_1^2 - w_2^2v_2^2)^2}{(w_1^2 + w_2^2)(v_1^2 + v_2^2 + (\mu_1 - \mu_2)^2)} > 0. \quad (2.3)
\]

The amount of improvement offered by \( \delta_{\overline{X}} \) decreases as the two components are more balanced with respect to the component variances and weights. This agrees with the similar findings when \( \delta_{\overline{X}} \) and \( \delta_{\overline{\theta}} \) were compared in a setting of balanced components.

To inspect the structure of the estimator \( \delta_{\overline{X}} \) closely, the case \( p = 2 \) is again utilized and the value of the shrinking factor becomes

\[
c = \frac{w_1v_1^2 + w_2v_2^2}{(w_1^2 + w_2^2)(v_1^2 + v_2^2 + (\mu_1 - \mu_2)^2)}.
\]

When \( \mu_1 = \mu_2 \), the value of \( c \) is larger. This is a reasonable result as \( \delta_{\overline{X}} \) should then shrink heavily toward the overall mean \( \overline{X} \).

In a setting of component balance, the component variances would both equal \( V \) and \( w_1 = w_2 = 1/2 \). The subsequent value of the shrinking factor in \( \delta_{\overline{X}} \) is
\[ c = \frac{\frac{1}{2}v + \frac{1}{2}v}{\left(\frac{1}{4} + \frac{1}{4}\right)(v + v + (\mu_1 - \mu_2)^2)} = \frac{v}{v + \frac{(\mu_1 - \mu_2)^2}{2}} \]

while, from (2.1),

\[ c_i = \frac{v}{\frac{1}{2}(v + v + (\mu_1 - \mu_2)^2)} = \frac{v}{v + \frac{(\mu_1 - \mu_2)^2}{2}} \]

for \( i = 1, 2 \) in \( \delta_{\bar{X}_1} \). This again demonstrates the equivalence of the estimators for balanced components.

To show the limited utility of \( \delta_{\bar{X}} \), assume \( V_1 = 0 \). The value of \( c_1 \) equals 0 allowing \( \delta_{\bar{X}_1} \) to utilize \( \bar{X}_1 \), a perfect estimator of \( \mu_1 \). The value of \( c \), on the other hand, does not allow \( \delta_{\bar{X}} \) to simply equal \( \bar{X}_1 \).

The estimator \( \delta_{\bar{X}_1} \) can be given a Bayesian interpretation by comparing it to an estimator proposed by D. V. Lindley (1962). Lindley suggested shrinking to a data-based value rather than an arbitrary constant while setting \( V_i = 1 \) for \( i = 1, 2, \ldots, p \). His estimator for \( \mu_i \) is

\[
1 - \frac{a_p}{\sum_{i=1}^{p} (\bar{X}_i - \bar{X})^2} (\bar{X}_i - \bar{X}) + \bar{X}
\]
where \( w_i = \frac{1}{p} \) for \( i = 1, 2, \ldots, p \) and \( a_p \) is a constant depending on \( p \).

The expected value of the denominator of his shrinking factor is

\[
E\left[ \sum_{i=1}^{p} \left( \bar{X}_i - \bar{X} \right)^2 \right] = \sum_{i=1}^{p} \left[ E\left( \bar{X}_i^2 \right) + E\left( \bar{X}^2 \right) - 2E\left( \bar{X}_i \bar{X} \right) \right] 
\]

\[
= \sum_{i=1}^{p} \left[ \mu_i^2 + v_i + V(\bar{X}) + \mu^2 - 2\left( \frac{1}{p} (\mu_i^2 + v_i) \right) + \frac{1}{p} \sum_{j \neq i}^{p} u_{ij} \right] 
\]

\[
= pV(\bar{X}) + \sum_{i=1}^{p} \left[ v_i (1 - \frac{2}{p}) + (\mu_i - \mu)^2 \right] 
\]

which is exactly the denominator of the shrinking factor, \( c \) in the estimator \( \hat{\delta}_{i\bar{X}} \). The numerator of \( c \) is

\[
\sum_{i=1}^{p} \left( 1 - \frac{1}{p} \right) v = \sum_{i=1}^{p} \left( 1 - \frac{1}{p} \right) = (p - 1) 
\]

which is a constant involving \( p \). With these results in mind, the Lindley estimator is merely the empirical version of \( \hat{\delta}_{i\bar{X}} \).

Considering shrinking to \( \hat{\mu} \), the estimator \( (1 - c)(\bar{X}_i - \hat{\mu}) + \hat{\mu} \)

is now studied for the case \( p = 2 \) and \( \mu_1 = \mu_2 = \mu \). The total of the component MSEs is

\[
\sum_{i=1}^{2} \left( 1 - c \right)^2 E\left( \bar{X}_i - \hat{\mu} \right)^2 + \sum_{i=1}^{2} E\left( \hat{\mu} - \mu \right)^2 + 2 \sum_{i=1}^{2} \left( 1 - c \right) E\left( \bar{X}_i - \hat{\mu} \right)(\hat{\mu} - \mu) 
\]
\[
= (1 - c)^2 \sum_{i=1}^{2} V(\bar{x}_i - \hat{\mu}) + 2V(\hat{\mu}) + 2(1 - c) \sum_{i=1}^{2} [E(\hat{\mu|x}_i) - E(\hat{\mu})^2] \\
- \mu E[\bar{x}_i - \hat{\mu}] \\
= (1 - c)^2 \sum_{i=1}^{2} V(\bar{x}_i - \hat{\mu}) + 2V(\hat{\mu}) + 2(1 - c) \sum_{i=1}^{2} [E(\hat{\mu|x}_i) - E(\hat{\mu})^2] \\
\]

Now
\[
\frac{\partial}{\partial c} \sum_{i=1}^{2} \text{MSE}((1 - c)(\bar{x}_i - \hat{\mu}) + \hat{\mu}) = 2(1 - c) \sum_{i=1}^{2} V(\bar{x}_i - \hat{\mu}) \\
+ 2 \sum_{i=1}^{2} [E(\hat{\mu|x}_i) - V(\hat{\mu}) - \mu^2] = 0
\]
implies
\[
1 - c = \frac{2V(\hat{\mu}) + 2\mu^2 - \sum_{i=1}^{2} E(\hat{\mu|x}_i)}{2 \sum_{i=1}^{2} V(\bar{x}_i - \hat{\mu})} \\
= \frac{(\frac{-v_2}{v_1 + v_2})v_1 + (\frac{-v_1}{v_1 + v_2})v_2 + 2V(\hat{\mu})}{v_1 + v_2 + 2V(\hat{\mu}) - 2[\frac{v_2}{v_1 + v_2}v_1 + \frac{v_1}{v_1 + v_2}v_2]} = 0.
\]

Again there is no possible improvement by shrinking to the data-based value \(\hat{\mu}\).
2.5 The Monte Carlo Analysis

To demonstrate the performance of the estimators when sample statistics are substituted in the shrinking factors for unknown parameters, a simulation study is utilized where 5000 component sample means and variances are generated. Assuming \( w_1 = w_2 = 1/2 \) and \( p = 2 \), MSEs and expected values of the estimators are found.

It has been discovered from the simulation runs that the estimators using estimated component variances give a mean square error only 1 to 5 percent larger than the estimators using known component variances. Unless otherwise noted, the results of this section assume known component variances.

Table 2 allows us to compare the estimators \( \delta_{\overline{X}} \) and \( \delta_{\overline{X}_1} \) by studying the totals of the MSEs for different choices of \((\mu_1, \mu_2)\) where \( V_1 = 3 \) and \( V_2 = 6 \).

It was pointed out previously that shrinking toward a common mean \( \overline{X} \) seems appropriate when the component means are nearly equal. It is in this case that the estimators \( \delta_{\overline{X}} \) and \( \delta_{\overline{X}_1} \) outperform the means \( \overline{X}_1 \) and \( \overline{X}_2 \). When the estimators \( \delta_{\overline{X}}, \delta_{\overline{X}_1}, \) and \( \overline{X}_1 \) were compared in (2.2) and (2.3), the difference in their MSEs was minimized when the component means were widely dispersed. When the component sample means are substituted for the \( \mu_i \)'s, it becomes clear from Table 2 that these earlier properties need not remain entirely true.
Table 2. Mean square error of estimators $\delta_{i\bar{x}}$ and $\delta_{i\bar{x}_1}$; $w_1 = w_2 = 1/2$, $V_1 = 3$, $V_2 = 6$, $p = 2$, $V(\bar{X}_1) + V(\bar{X}_2) = 9$

<table>
<thead>
<tr>
<th>$(\mu_1, \mu_2)$</th>
<th>$\text{MSE}(\delta_{i\bar{x}})$</th>
<th>$\text{MSE}(\delta_{i\bar{x}_1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 2)</td>
<td>7.186</td>
<td>7.006</td>
</tr>
<tr>
<td>(2, 0)</td>
<td>7.171</td>
<td>6.963</td>
</tr>
<tr>
<td>(2, 4)</td>
<td>7.186</td>
<td>7.006</td>
</tr>
<tr>
<td>(4, 2)</td>
<td>7.171</td>
<td>6.963</td>
</tr>
<tr>
<td>(2, 8)</td>
<td>9.705</td>
<td>9.803</td>
</tr>
<tr>
<td>(8, 2)</td>
<td>9.668</td>
<td>9.719</td>
</tr>
<tr>
<td>(2, 10)</td>
<td>10.012</td>
<td>10.155</td>
</tr>
<tr>
<td>(10, 2)</td>
<td>9.955</td>
<td>10.053</td>
</tr>
</tbody>
</table>

Table 3 lists the ratio of the total MSE of the estimators to $V_1 + V_2$ for different choices of $(\mu_1/\sqrt{V_1}, \mu_2/\sqrt{V_2})$. When the ratio exceeds 1, the estimator performs more poorly than using the component sample means. The ratios also demonstrate when certain shrinkage estimators perform better than others.

Since the estimators $\delta_{i\theta_1}$ and $\delta_{i\theta}$ are shrinking to the constants $(\theta_1, \theta_2) = (0, 0)$, one would expect their performance to be best when $(\mu_1/\sqrt{V_1}, \mu_2/\sqrt{V_2})$ is near $(0, 0)$.

When $\mu_1$ is nearly equal to $\mu_2$, the values of $\mu_1/\sqrt{V_1}$ will be nearly equal to $\mu_2/\sqrt{V_2}$. In this case, one would expect $\delta_{i\bar{x}_1}$ to perform well as it is shrinking toward a common mean $\bar{X}$.
Figure 2. Proportional MSEs of Thompson-type shrinkage estimators, the normal case.

\[ \frac{\text{MSE}(\delta_1) + \text{MSE}(\delta_2)}{V_1 + V_2} \]

(2.1) \( \frac{\text{MSE}(\delta_1) + \text{MSE}(\delta_2)}{V_1 + V_2} \), only \( V_1, V_2 \) known.

(2.2) \( \frac{\text{MSE}(\delta_1) + \text{MSE}(\delta_2)}{V_1 + V_2} \), no parameters known.

(2.3) \( \frac{\text{MSE}(\delta_1) + \text{MSE}(\delta_2)}{V_1 + V_2} \), only \( V_1, V_2 \) known, i.e., evaluation of the Thompson estimator (1968).
When $\mu_1/\sqrt{V_1} \neq \mu_2/\sqrt{V_2}$, a sense of balance does not exist. This makes $\delta_{i\Theta_1}$ the likely choice over $\delta_{i\Theta}$ as this estimator gives the special utility of shrinking differently for each component.

All the conjectures of the last three paragraphs are confirmed by the values in Table 3.

Figure 2 gives a graphical display of the performance of the estimators considered in Table 3 for the case of $\mu_1/\sqrt{V_1} = \mu_2/\sqrt{V_2}$. The estimators $\delta_{i\Theta}$ and $\delta_{i\Theta_1}$ outperform $\bar{X}_i$ when the component means are close to $(\Theta_1, \Theta_2)$ relative to the component variation. They never perform extremely poorly and, as the means become very distant from $(\Theta_1, \Theta_2)$, their performances tend to match the $\bar{X}_i$'s.

The estimators $\delta_{i\bar{X}}$ and $\delta_{i\bar{X}_1}$ do much better than $\bar{X}_1$, but note that a setting of component balance is in effect here.

One also can see that little is lost in mean square error when the component variances are also assumed unknown. The ratios of total MSE to $V_1 + V_2$ for $\delta_{i\Theta}$ and $\delta_{i\bar{X}_1}$ under these conditions are given in Figure 2.

In simulation studies, the results depend, to some degree, on the generated random variables. To demonstrate the consistency of our study of 5000 generated random variables, 16 simulations were run while only changing the starting point of the computer routine. By finding the standard deviation of the 16 resulting expected values of a given estimator, a measure of the consistency of simulated results is found. The standard deviations, found in such a way, are...
Table 3. Ratio of total mean square error to \((V_1 + V_2)\) for 5 estimators where \(p = 2, \pi_1 = \pi_2 = 1/2, V_1\) and \(V_2\) known except for \(\delta_{i\theta}^l\) and \(\delta_{iX_i}^l\)

<table>
<thead>
<tr>
<th>(\mu_1 / \sqrt{V_1})</th>
<th>(\mu_2 / \sqrt{V_2})</th>
<th>(\delta_{i\theta})</th>
<th>(\delta_{i\theta}^l)</th>
<th>(\delta_{i\theta_i})</th>
<th>(\delta_{iX_i})</th>
<th>(\delta_{iX_i}^l)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>.377</td>
<td>.393</td>
<td>.459</td>
<td>.7177</td>
<td>.7300</td>
</tr>
<tr>
<td>0</td>
<td>1.155</td>
<td>.618</td>
<td>.629</td>
<td>.652</td>
<td>.832</td>
<td>.840</td>
</tr>
<tr>
<td>.288</td>
<td>.288</td>
<td>.413</td>
<td>.430</td>
<td>.488</td>
<td>.7177</td>
<td>.7300</td>
</tr>
<tr>
<td>1.155</td>
<td>1.155</td>
<td>.773</td>
<td>.782</td>
<td>.852</td>
<td>.7177</td>
<td>.7300</td>
</tr>
<tr>
<td>1.155</td>
<td>2.309</td>
<td>.985</td>
<td>.991</td>
<td>1.039</td>
<td>.832</td>
<td>.840</td>
</tr>
<tr>
<td>1.155</td>
<td>4.619</td>
<td>1.046</td>
<td>1.052</td>
<td>.988</td>
<td>1.103</td>
<td>1.108</td>
</tr>
<tr>
<td>1.155</td>
<td>5.774</td>
<td>1.035</td>
<td>1.043</td>
<td>.966</td>
<td>1.094</td>
<td>1.099</td>
</tr>
<tr>
<td>1.732</td>
<td>1.732</td>
<td>.963</td>
<td>.970</td>
<td>1.002</td>
<td>.7177</td>
<td>.7300</td>
</tr>
<tr>
<td>2.309</td>
<td>2.309</td>
<td>1.040</td>
<td>1.047</td>
<td>1.225</td>
<td>.7177</td>
<td>.7300</td>
</tr>
<tr>
<td>4.619</td>
<td>4.619</td>
<td>1.028</td>
<td>1.035</td>
<td>1.127</td>
<td>.7177</td>
<td>.7300</td>
</tr>
<tr>
<td>6.928</td>
<td>6.928</td>
<td>1.007</td>
<td>1.010</td>
<td>1.048</td>
<td>.7177</td>
<td>.7300</td>
</tr>
</tbody>
</table>
listed in the last column of Table 4, for many of our proposed estimators. Their small sizes indicate a dependable simulation routine.

Also given in Table 4 are the proportions of total MSE accounted for by the square of the bias of component estimators. Since all of our estimators are generally biased estimators, these proportions indicate the extent of the bias. The results of the 16 simulations with different starting points are used to find these proportions. Note that the small size of these proportions indicate a minimal
Table 4. Measurements of (i) contribution of bias to MSE, (ii) consistency of simulated results where \( \mu_1 = 2, \mu_2 = 4, \theta_1 = \theta_2 = 0, \pi_1 = \pi_2 = 1/2 \) and \( V_1 = 3 \) and \( V_2 = 6 \)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( V_1, V_2 ) assumed</th>
<th>((\text{Bias})^2 / \text{Total MSE})</th>
<th>Standard deviations of simulated values of the estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_{1\theta} )</td>
<td>known</td>
<td>0.0331</td>
<td>0.0168</td>
</tr>
<tr>
<td>( \delta_{2\theta} )</td>
<td>known</td>
<td>0.0958</td>
<td>0.0314</td>
</tr>
<tr>
<td>( \delta_{1\theta} )</td>
<td>unknown</td>
<td>0.0351</td>
<td>0.0180</td>
</tr>
<tr>
<td>( \delta_{2\theta} )</td>
<td>unknown</td>
<td>0.1054</td>
<td>0.0325</td>
</tr>
<tr>
<td>( \delta_{1\theta i} )</td>
<td>known</td>
<td>0.0298</td>
<td>0.0182</td>
</tr>
<tr>
<td>( \delta_{2\theta i} )</td>
<td>known</td>
<td>0.0419</td>
<td>0.0330</td>
</tr>
<tr>
<td>( \delta_{1\theta i} )</td>
<td>unknown</td>
<td>0.0104</td>
<td>0.0180</td>
</tr>
<tr>
<td>( \delta_{2\theta i} )</td>
<td>unknown</td>
<td>0.0109</td>
<td>0.0301</td>
</tr>
<tr>
<td>( \delta_{1\bar{X}} )</td>
<td>known</td>
<td>0.0133</td>
<td>0.0185</td>
</tr>
<tr>
<td>( \delta_{2\bar{X}} )</td>
<td>known</td>
<td>0.0140</td>
<td>0.0245</td>
</tr>
<tr>
<td>( \delta_{1\bar{X}} )</td>
<td>unknown</td>
<td>0.0132</td>
<td>0.0185</td>
</tr>
<tr>
<td>( \delta_{2\bar{X}} )</td>
<td>unknown</td>
<td>0.0139</td>
<td>0.0295</td>
</tr>
<tr>
<td>( \delta_{1\bar{X} i} )</td>
<td>known</td>
<td>0.0065</td>
<td>0.0189</td>
</tr>
<tr>
<td>( \delta_{2\bar{X} i} )</td>
<td>known</td>
<td>0.0263</td>
<td>0.0294</td>
</tr>
<tr>
<td>( \delta_{1\bar{X} i} )</td>
<td>unknown</td>
<td>0.0063</td>
<td>0.0209</td>
</tr>
<tr>
<td>( \delta_{2\bar{X} i} )</td>
<td>unknown</td>
<td>0.0242</td>
<td>0.0292</td>
</tr>
<tr>
<td>( \bar{X}_1 )</td>
<td></td>
<td>0.00000</td>
<td>0.0219</td>
</tr>
<tr>
<td>( \bar{X}_2 )</td>
<td></td>
<td>0.00001</td>
<td>0.0336</td>
</tr>
</tbody>
</table>
3. SHRINKAGE ESTIMATORS FOR THE MEAN OF A STRATIFIED NORMAL POPULATION

3.1 Introduction

In the previous chapter, various shrinkage estimators were suggested for the estimation of the mean vector of a multivariate normal distribution. If the components of this multivariate population were considered as strata for a stratified normal population, the estimation of the stratified mean is of interest.

Once again shrinkage estimators are proposed as a form of the estimator. Notationally, the estimators will be distinguished from those in Chapter 2 by omitting the first subscript, i, which was used to define an estimator of the $i^{th}$ component mean. Once these shrinkage estimators are found their structural properties will be noted and, through a simulation study, their mean square errors will be studied. It should be noted that the most complete results and the simulation study deal with the case of two strata, however generalizations are given in most cases.

We assume that each of the p strata can be modelled by a normal distribution. Hence, $X_{ij}$ denotes the $j^{th}$ observation from the $i^{th}$ stratum where $X_{ij}$ comes from a normal population with mean $\mu_i$ and variance $\sigma_i^2$. Furthermore, it is assumed that the stratum weights, the proportions that the strata represent of the total population, are known and equal to $\pi_i$ for $i = 1, 2, \ldots, p$. The
stratified mean, which is to be estimated, is then

\[ \bar{\mu} = \frac{\sum_{i=1}^{p} n_i \mu_i}{\Sigma n_i} . \]

The usual estimator of \( \bar{\mu} \) is

\[ \bar{X} = \frac{\sum_{i=1}^{p} n_i \bar{X}_i}{\Sigma n_i} \]

where

\[ \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \]

and \( n_i \) is the sample size from the \( i \)th stratum. Again the notation \( V_i \) is used to stand for \( \frac{\sigma_i^2}{n_i} \), the variance of \( \bar{X}_i \).

In order to estimate \( \mu_i \), the component part of \( \bar{\mu} \), we consider estimators of the following forms:

(a) \( (1 - c_i)(\bar{X}_i - \theta_i) + \theta_i \)

(b) \( (1 - c_i)(\bar{X}_i - \bar{X}) + \bar{X} \)

(c) \( (1 - c_i)(\bar{X}_i - \hat{\mu}) + \hat{\mu} . \) \hspace{1cm} (3.1)

In the first case, \( \theta_i \) is some pre-conceived value believed to be near the true value \( \mu_i \). In the second case, we shrink toward the overall stratified mean and lastly to another accepted estimator, \( \hat{\mu} \), which will be introduced later. In any case, the value of the shrinking factor \( c_i \) determines whether the estimator is nearly the same as \( \bar{X}_i \) or if its value is nearer the value to which \( \bar{X}_i \) is shrunk to-
wards. With this flexibility built into the estimator of $\mu_1$, one hopes to find the MSE of our estimators to be less than the variance of $\bar{X}$.

An added restriction could be put on these estimators by requiring the shrinking factors to take on a common value, i.e., $c_i = c$ for $i = 1, 2, \ldots, p$. Such a restriction leads to quite different shrinking factors from those of before and the resulting estimators of $\bar{\mu}$ can prove valuable in different settings than the estimators in (3.1).

One should note that the values of the shrinking factors will depend upon unknown $\mu_1$'s and possibly unknown $V_1$'s. The performance of the estimators, when we must substitute estimates of these unknown parameters, will be analyzed in section 5 by the use of a Monte Carlo study. In sections 3 and 4, the estimators are studied in their presubstitution form.

3.2 The Development of the Estimators

Before proceeding with finding these estimators, we must decide on a criterion upon which the estimators will be judged. Since we are generally dealing with biased estimators of $\bar{\mu}$ and $\mu_1$, we attempt to minimize the MSE. We must find the value of $c$ or else of $c_i$ that insures such a minimization. This is done by differentiating the MSE with respect to the shrinking factor, setting the result equal to zero and solving for the factor.

If we define $\delta_1$ as the subsequent estimator of $\mu_1$, we are dealing
with $\sum_{i=1}^{p} \pi_i \delta_i$ as the estimator of $\overline{\mu}$. We can develop the estimator $\delta_i$ which minimizes the component MSE of $\delta_i$ or else we can choose to minimize the MSE of $\sum_{i=1}^{p} \pi_i \delta_i$ directly. If we note that

$$\text{MSE}(\sum_{i=1}^{p} \pi_i \delta_i) = \sum_{i=1}^{p} \pi_i^2 \text{MSE}(\delta_i) + \sum_{i=1}^{p} \pi_i \pi_j E(\delta_i - \mu_i)(\delta_j - \mu_j), \quad (3.2)$$

we can see that minimizing the component MSEs only assures us of minimizing the first sum on the right-hand side of (3.2). But minimizing the MSE of $\sum_{i=1}^{p} \pi_i \delta_i$ assures us of minimizing the entire right-hand side of (3.2).

Both of these possible criteria are pursued and comparisons of the results are made. When dealing with minimizing the component MSEs, the resulting estimators are called the component-wise estimators.

### 3.3 Component-Wise Estimators

Considering the three forms in (3.1), we now determine the shrinking factors that minimize the component MSEs. As we are considering each component separately, there is no need to consider a common factor.

The first estimator of $\overline{\mu}$ to be considered is given by

$$\delta_{\theta c} = \sum_{i=1}^{p} \pi_i [(1 - c_i)(X_i - \theta_i) + \theta_i].$$
The subscript \( c \) denotes component-wise minimization and the subsequent \( \delta_i \) is the Thompson estimator. Thompson found the minimizing value of \( c_i \) as

\[
\frac{V_i}{V_i + (\mu_i - \theta_i)^2}
\]

and found that this \( \delta_i \) uniformly beats \( \bar{X}_i \) in estimating \( \mu_i \). This does not mean, however, that \( \sum_{i=1}^{p} \pi_i \delta_i = \delta_{\Theta C} \) beats \( \bar{X} \) in terms of overall MSE.

In order to show that \( \delta_{\Theta C} \) does often improve upon \( \bar{X} \), note that

\[
\text{MSE} \left[ \sum_{i=1}^{p} \pi_i \left( 1 - \frac{V_i}{V_i + (\mu_i - \theta_i)^2} \right) (\bar{X}_i - \theta_i) + \theta_i \right] \]

\[
= \sum_{i=1}^{p} \pi_i \left[ \frac{(\mu_i - \theta_i)^2 V_i}{V_i + (\mu_i - \theta_i)^2} + 2 \sum_{i<j} \pi_i \pi_j \frac{V_i V_j (\mu_i - \theta_i) (\mu_j - \theta_j)}{V_i + (\mu_i - \theta_i)^2} \right].
\]

Defining

\[
R_i = \frac{1}{V_i + (\mu_i - \theta_i)^2},
\]

we find that

\[
\bar{X} - \text{MSE}(\delta_{\Theta C}) = \sum_{i=1}^{p} \pi_i^2 V_i (1 - (\mu_i - \theta_i)^2 R_i) - 2 \sum_{i<j} \pi_i \pi_j V_i V_j R_i R_j (\mu_i - \theta_i) (u_j - \theta_j)
\]
\[
= \sum_{i=1}^{p} \pi_i^2 \left(1 - (p - 1)R_i (\mu_i - \theta_i)^2\right)
\]
\[
+ \sum_{i<j} \left[\pi_i V_i R_i (\mu_i - \theta_i) - \pi_j V_j R_j (\mu_j - \theta_j)\right]^2.
\]

From this, one can see that \( \delta_{\theta_c} \) is better than \( \bar{X} \) if one of the following conditions is met:

(a) \( p = 2 \)

(b) \( |\theta_i - \mu_i| < \sqrt{\frac{V_i}{p-2}} \) for \( i = 1, 2, \ldots, p \), meaning that our prior guess for \( \mu_i \) is good or else the variation within a stratum is quite large when \( \theta_i \) is not close to \( \mu_i \),

(c) \( \frac{\pi_i^2}{V_i + (\mu_i - \theta_i)^2} \) is especially small for the strata

where \( |\theta_i - \mu_i| > \sqrt{\frac{V_i}{p-2}} \).

Now consider \( \delta_{\bar{X}_c} = \sum_{i=1}^{p} \pi_i \left[(1 - c_i) (\bar{X}_i - \bar{X}) + \bar{X}\right] \). With this estimation, no prior guesses are needed in order to estimate \( \bar{u} \). If all the \( \mu_i \)'s are nearly equal, one would expect \( \delta_{\bar{X}_c} \) to do well as it shrinks to a common value \( \bar{X} \). Should the \( \mu_i \)'s be widely dispersed, it may still be better to shrink to \( \bar{X} \) than an arbitrary set of \( \theta_i \)'s if these \( \theta_i \)'s are poorly chosen.
The superiority of $\delta_{\bar{X}}$ to $\bar{X}$ will be shown explicitly when an equivalent estimator is developed later.

Minimizing the component MSE of $\delta_{\bar{X}}$ is equivalent to the procedure taken to find the shrinking factor of $\delta_{\bar{X}_i}$ in Chapter 2. We found from (2.1) that minimizing the component MSE leads to a shrinking factor

$$c_i = \frac{(1 - \pi_i) V_i}{(\mu_i - \bar{\mu})^2 + V(\bar{X}_i - \bar{X})}.$$  \hspace{1cm} (3.3)

When the difference $\bar{X}_i - \bar{X}$ has a large variance or when the variance of $\bar{X}_i$ is small or when $\mu_i$ is not near $\bar{\mu}$, the value of $c_i$ will be nearly equal to zero which, in turn, makes the estimator of $\mu_i$ shrink very little to $\bar{X}$. This is an expected result because in all three cases, the value of $\bar{X}_i$ is likely to be a better estimator of $\mu_i$ than $\bar{X}$.

The value of $c_i$ is always greater than zero and usually less than one, but to be sure we are dealing with a true shrinkage estimator, we could further define $c_i$ to be

$$\min \left[ 1, \frac{(1 - \pi_i) V_i}{(\mu_i - \bar{\mu})^2 + V(\bar{X}_i - \bar{X})} \right].$$

This precaution is taken in the simulation studies.

The final component-wise estimator to be considered involves shrinking to
\[ \hat{\mu} = \frac{V_2}{V_1 + V_2} \bar{X}_1 + \frac{V_1}{V_1 + V_2} \bar{X}_2 \]

where \( p = 2 \).

Finding the value of \( c_i \) that minimizes the MSE of \((1 - c_i)(\bar{X}_i - \hat{\mu})\)
+ \( \hat{\mu} \) is again equivalent to a minimization that was performed in Chapter 2. It was found that \( c_i = 1 \) for \( i = 1, 2 \) when considering the estimator \( \delta_i \). So once again we are unable to improve upon \( \hat{\mu} \) with our shrinkage estimator and the unbeatable variance of \( \hat{\mu} \) is then \( \frac{V_1 V_2}{V_1 + V_2} \).

In comparing the estimator \( \hat{\mu} \) to \( \delta_{Xc} \) in the case of the common stratum means, we find from (3.3) that the shrinking factor for \( \delta_{Xc} \) is \( c_i = \frac{V_i}{V_j(V_i + V_j)} \) for \((i, j) = (1, 2)\) or \((i, j) = (2, 1)\). A straightforward simplification shows that \( \delta_{Xc} \) is \( \hat{\mu} \) itself, making it the minimum variance unbiased estimator of the common mean \( \mu \) of the form \( d_1 \bar{X}_1 + d_2 \bar{X}_2 \).

3.4 Estimators Minimizing MSE of \( \sum_{i=1}^{p} \pi_i \delta_i \)

3.4.1 The common shrinking factor

Let us find shrinkage estimators of the form \( \sum_{i=1}^{p} \pi_i \delta_i \) that not only minimize the total MSE, but also assume a common shrinking factor, \( c \). Notationally, let the three forms in (3.1) be \( \delta_\theta \), \( \delta_{Xc} \) and \( \hat{\mu} \) respectively.

Considering the MSE of \( \delta_\theta \), we find
\[ \text{MSE}(\hat{\theta}) = \text{MSE}\left( \sum_{i=1}^{P} \pi_i ((1 - c)(\bar{X}_i - \theta_i) + \theta_i) \right) \]

\[ = E[(1 - c)(\bar{X} - \bar{\theta}) + \bar{\theta} - \bar{\mu}]^2 \]

where \( \bar{\theta} = \sum_{i=1}^{P} \pi_i \theta_i \). But this is simply a special case of the univariate Thompson estimator where

\[ c = \frac{\sqrt{\bar{X}}}{\sqrt{\bar{X}} + (\bar{\mu} - \bar{\theta})^2} . \]

The MSE of \( \hat{\delta}_X \) gives

\[ \text{MSE}\left( \sum_{i=1}^{P} \pi_i [(1 - c)(\bar{X}_i - \bar{X}) + \bar{X}] \right) = \text{MSE}[(1 - c) \sum_{i=1}^{P} \pi_i (\bar{X}_i - \bar{X}) + \bar{X}] \]

\[ = \text{MSE}(\bar{X}) = \sqrt{\bar{X}} . \]

Hence, any constant value of \( c \) leads to \( \bar{X} \) itself showing that we are unable to improve upon \( \bar{X} \) with the form \( \hat{\delta}_X \).

And furthermore,

\[ \text{MSE}(\hat{\mu}) = \text{MSE}\left( \sum_{i=1}^{2} \pi_i [(1 - c)(\bar{X}_i - \hat{\mu}) + \hat{\mu}] \right) \]

\[ = \text{MSE}[(1 - c)(\bar{X} - \hat{\mu}) + \hat{\mu}] \]
gives us a minimizing value of $c$ equal to 1. Once again we are unable to improve upon $\mu$ with a shrinkage estimator.

### 3.4.2 Varying shrinking factors

Again we minimize the MSE of $\sum_{i=1}^{p} \pi_i \delta_i$, as in the previous section, but in allowing the shrinking factors to differ from stratum to stratum the results are not nearly as simple and automatic. It is of interest to compare the results of this section with those dealing with the component-wise estimators as both sets of estimators utilize shrinking factors that vary across the strata.

Notationally, we define $\delta_{\theta i}$, $\delta_{X i}$ and $\delta_{\mu i}$ as the estimators for the three forms in (3.1) respectively, where the subscript $i$ denotes the fact that the shrinking factor, $c_i$, differs for each stratum.

Beginning with $\delta_{\theta i}$, we must minimize $\text{MSE}(\delta_{\theta i})$ where

$$
\text{MSE}(\delta_{\theta i}) = \mathbb{E}\left[ \sum_{i=1}^{p} \pi_i \left[ (1 - c_i)(\bar{X}_i - \theta_i) + \theta_i - \mu_i \right]^2 \right]
$$

$$
= \mathbb{E}\left[ \sum_{i=1}^{p} \pi_i \left[ (1 - c_i)(\bar{X}_i - \theta_i) + \theta_i - \mu_i \right]^2 \right]
$$

$$
+ \mathbb{E}\left[ \sum_{i \neq j} \pi_i \pi_j \left[ (1 - c_i)(\bar{X}_i - \theta_i) \right] \right]
$$

$$
+ \theta_i - \mu_i \left[ (1 - c_j)(X_j - \theta_j) + \theta_j - \mu_j \right]
$$
\[
\begin{align*}
&= \sum_{i=1}^{p} \pi_i^2 [(1 - c_i)^2 E(\bar{x}_i - \theta_i)^2 + 2(1 - c_i) E(\bar{x}_i - \theta_i)(\theta_i - \mu_i)] \\
&\quad + (\theta_i - \mu_i)^2] + \sum_{i \neq j}^{p} \pi_i \pi_j [(1 - c_i)(1 - c_j) E(\bar{x}_i - \theta_i)(\bar{x}_j - \theta_j)] \\
&\quad + (1 - c_i) E(\bar{x}_i - \theta_i)(\theta_j - \mu_j)] \\
&\quad + (1 - c_j) E(\bar{x}_j - \theta_j)(\theta_i - \mu_i)] + (\theta_i - \mu_i)(\theta_j - \mu_j)] \\
&= \sum_{i=1}^{p} \pi_i^2 [(1 - c_i)^2 (V_i + (\mu_i - \theta_i)^2) - 2(1 - c_i) (\mu_i - \theta_i)^2] \\
&\quad + (\theta_i - \mu_i)^2] + \sum_{i \neq j}^{p} \pi_i \pi_j [(1 - c_i)(1 - c_j) (\mu_i - \theta_i)(\mu_j - \theta_j)] \\
&\quad + (1 - c_i) (\mu_i - \theta_i)(\theta_j - \mu_j) + (1 - c_j)(\mu_j - \theta_j)(\theta_i - \mu_i)] \\
&\quad + (\theta_i - \mu_i)(\theta_j - \mu_j)] \\
&= \sum_{i=1}^{p} \left[ \pi_i^2 (1 - c_i)^2 (V_i + (\mu_i - \theta_i)^2) \right] + \left[ \sum_{i=1}^{p} \pi_i (\theta_i - \mu_i) \right]^2 \\
&\quad + \sum_{i \neq j}^{p} \pi_i \pi_j (1 - c_i)(1 - c_j)(\mu_i - \theta_i)(\mu_j - \theta_j)] \\
&\quad + 2 \left( \sum_{i=1}^{p} \pi_i (\theta_i - \mu_i) \right) \left( \sum_{i=1}^{p} \pi_i (1 - c_i)(\mu_i - \theta_i) \right)
\end{align*}
\]
\[
\begin{align*}
= \sum_{i=1}^{p} \left[ \pi_{i}^{2} (1 - c_{i})^{2} (v_{i} + (\mu_{i} - \theta_{i})^{2}) \right] + \left[ \bar{\theta} - \bar{\mu} \right]^{2} \\
+ \sum_{i \neq j}^{p} \left[ \pi_{i} \pi_{j} (1 - c_{i})(1 - c_{j})(\mu_{i} - \theta_{i})(\mu_{j} - \theta_{j}) \right] \\
+ 2(\bar{\theta} - \bar{\mu}) \sum_{i=1}^{p} \pi_{i} (1 - c_{i})(\mu_{i} - \theta_{i}) .
\end{align*}
\]

Now,
\[
\frac{\partial}{\partial c_{i}} \text{MSE} = 2\pi_{i}^{2} (1 - c_{i}) (v_{i} + (\mu_{i} - \theta_{i})^{2}) \\
+ 2 \sum_{i \neq j} \left[ \pi_{i} \pi_{j} (\mu_{i} - \theta_{i})(\mu_{j} - \theta_{j})(1 - c_{j}) \right] \\
+ 2(\bar{\theta} - \bar{\mu}) \pi_{i} (\mu_{i} - \theta_{i}) = 0 .
\]

Hence,
\[
1 - c_{i} = \frac{\pi_{i} [\omega_{i} - \theta_{i})(\mu_{i} - \theta_{i}) - (\mu_{i} - \theta_{i}) \sum_{j(\neq i)}^{p} \pi_{j} (\mu_{j} - \theta_{j})(1 - c_{j})]}{\pi_{i}^{2} v_{i} + (\mu_{i} - \theta_{i})^{2}}
\]

for \( i = 1, 2, \ldots, p \) which is a system of \( p \) equations and \( p \) unknowns.

To clearly see the form of the shrinking factors, we study the case where \( p = 2 \) and solve for \( 1 - c_{i} \).
\[
1 - c_i = \frac{\pi_i (\mu_i - \theta_i)(\mu - \bar{\theta}) - \pi_j (\mu_j - \theta_j)(1 - c_j)}{\pi_i [V_i (\mu - \theta)^2]}
\]

\[
= \frac{(\mu_i - \theta_i)(\mu - \bar{\theta}) - \pi_j (\mu_j - \theta_j)[(\mu - \theta) - \pi_i (\mu_i - \theta_i)(1 - c_i)]}{\pi_i [V_i (\mu - \theta)^2]}
\]

leading to

\[
1 - c_i = \frac{V_i (\mu_i - \theta_i)(\mu - \bar{\theta})}{\pi_i [V_i V_j + V_i (\mu - \theta)^2 + V_j (\mu_i - \theta_i)^2]}
\]

for \(i, j = (1, 2)\) or \(i, j = (2, 1)\).

When \(\bar{\mu} = \bar{\theta}\), giving our prior guess the mark of perfection, we find \(1 - c_i\) to be equal to zero for \(i = 1\) and \(2\). Thus \(\delta_{\theta_1}\) becomes simply \(\bar{\theta}\).

Since \(\delta_{\theta_1}\) minimizes the total MSE while \(\delta_{\theta_2}\) only minimizes the sum of the component MSEs, we should expect the former to obviously dominate the latter. With \(p = 2\) and \(\theta_1 = \theta_2 = 0\), we find

\[
\text{MSE}(\delta_{\theta_2}) - \text{MSE}(\delta_{\theta_1}) = \frac{V_1 V_2 \mu_1^2 + V_1^2 \mu_2^2 + V_1^2 \mu_2^2}{V_1 V_2 + V_1^2 \mu_2^2 + V_2^2 \mu_1^2 + \mu_1^2 \mu_2^2}
\]

\[
- \frac{V_1 V_2^2 \mu_1^2 + V_2 V_1^2 \mu_2^2 + V_1 V_2^2 \mu_2^2}{(V_1^2 + V_1^2 \mu_2^2 + V_2^2 \mu_1^2)^2}
\]
To show that both $\delta_{\theta_1}$ and $\delta_{\theta_c}$ perform better than $\bar{X}$ for $p = 2$, we consider

\[\bar{X} - \text{MSE}(\delta_{\theta_c}) = \bar{X} - \frac{\frac{V_1 V_2}{V_1^2 + V_1 \mu_2^2 + V_2 \mu_2^2} \cdot \frac{V_1 V_2 \bar{X}}{V_1 V_2 + V_1 \mu_2^2 + V_2 \mu_2^2}}{\mu_1 \mu_2 V_1^2 \left(\pi_1 V_1 \mu_2^2 + \pi_2 V_2 \mu_2^2\right)^2 + V_1 V_2 \bar{X}} \frac{> 0}{V_1 V_2 + V_1 \mu_2^2 + V_2 \mu_2^2}\]

Because of the extensive algebraic difficulties in dealing with the estimator $\delta_{\theta_1}$, the case of $p = 2$ is initially considered. Generalizations will be commented upon for $p > 2$ later in the chapter.

The problem at hand is simplified when considering a special
form of $\text{MSE}(\delta_{X_1})$.

$$\text{MSE}(\delta_{X_1}) = E\left[\gamma_1(1-c_1)(\overline{X}_1 - \overline{X}) + \gamma_2(1-c_2)(\overline{X}_2 - \overline{X}) - \overline{u}\right]^2$$

$$= E\left[\gamma_1(1-c_1)(\overline{X}_1 - \overline{X}) + \gamma_2(1-c_2)(\overline{X}_2 - \overline{X}) + (\overline{X} - \overline{u})\right]^2$$

$$= E\left[\gamma_1(1-c_1)(\overline{X}_1 - \overline{X}) + \gamma_2(1-c_2)(\overline{X}_2 - \overline{X}) + (\overline{X} - \overline{u})\right]^2$$

$$= E\left[\gamma_1\gamma_2(1-c_1)(\overline{X}_1 - \overline{X}) + \gamma_2\gamma_1(1-c_2)(\overline{X}_2 - \overline{X}) + (\overline{X} - \overline{u})\right]^2$$

$$= E\left[\gamma_1\gamma_2(1-c_1)(\overline{X}_1 - \overline{X}) + \gamma_2\gamma_1(1-c_2)(\overline{X}_2 - \overline{X}) + (\overline{X} - \overline{u})\right]^2$$

Here we see that only the minimizing value of $c_2 - c_1$ is needed. Any selection of $c_2$ and $c_1$ which gives the minimizing value for the difference would make $\delta_{X_1}$ the optimal estimator. If the solutions for $c_1$ and $c_2$ were found by differentiating the MSE with respect to $c_1$ and $c_2$ individually, a system of two equations would lead to the same result as considering the difference $c_2 - c_1$.

Now,

$$\frac{\partial}{\partial(c_2-c_1)} \text{MSE}(\delta_{X_1}) = 2\gamma_1^2\gamma_2^2(c_2 - c_1)(V_1 + V_2 + (\mu_1 - \mu_2)^2)$$

$$+ 2\gamma_1\gamma_2(\mu_1 - \mu_2)(V_1 - V_2) = 0$$

implies
\[ c_2 - c_1 = \frac{(\pi_2 V_2 - \pi_1 V_1)}{\pi_1 \pi_2 (V_1 + V_2 + (\mu_1 - \mu_2)^2)}. \quad (3.4) \]

The minimized MSE is

\[ \text{MSE}(\delta_{X_1}) = V(X) + \frac{2^2}{\pi_1 \pi_2} \left( \frac{(\pi_2 V_2 - \pi_1 V_1)^2(V_1 + V_2 + (\mu_1 - \mu_2)^2)}{\pi_1 \pi_2 (V_1 + V_2 + (\mu_1 - \mu_2)^2)^2} \right. \]

\[ + \left. 2 \frac{\pi_1^2}{\pi_1 \pi_2} \frac{(\pi_2 V_2 - \pi_1 V_1)(\pi_1 V_1 - \pi_2 V_2)}{\pi_1 \pi_2 (V_1 + V_2 + (\mu_1 - \mu_2)^2)} \right) \]

\[ = V(X) - \frac{(\pi_2 V_2 - \pi_1 V_1)^2}{V_1 + V_2 + (\mu_1 - \mu_2)^2}. \quad (3.5) \]

This expression not only shows the obvious superiority of \( \delta_{X_1} \) to \( \bar{X} \), but also expresses the amount of superiority. The improvement is maximized with respect to the stratum means if \( \mu_1 = \mu_2 \), an intuitively pleasing result as this is when \( \bar{X} \) would seem to be the best expression to shrink toward. But the question of improving upon \( \bar{X} \) versus not improving lies in the values of \( \pi_1, \pi_2, V_1, \) and \( V_2 \). These are the terms that, in a sense, dictate the relative weighting of the two strata.

To demonstrate how the structure of \( \delta_{X_1} \) dictates how it works, let us recall that the form \( \delta_{X_1} \) was shown to be
\[ \pi_1 \pi_2 (c_2 - c_1)(\bar{X}_1 - \bar{X}_2) + \bar{X} \] or more particularly

\[ \pi_1 \pi_2 \left[ \frac{\pi_2 V_2 - \pi_1 V_1}{\pi_1 \pi_2 (V_1 + V_2 + (\mu_1 - \mu_2)^2)} \right] (\bar{X}_1 - \bar{X}_2) + \bar{X} \]

\[ = \frac{\pi_2 V_2 - \pi_1 V_1}{(V_1 + V_2 + (\mu_1 - \mu_2)^2)} (\bar{X}_1 - \bar{X}_2) + \bar{X}. \quad (3.6) \]

Should we want to express \( \delta_{X_1} \) as an estimator of \( \bar{u} \) that is shrinking between \( \bar{X}_1 \) and \( \bar{X}_2 \), it is a simple result to find that \( \delta_{X_1} \) equals \((1 - K)\bar{X}_1 + k\bar{X}_2\) where

\[ K = \frac{V_1 + \pi_2 (\mu_1 - \mu_2)^2}{V_1 + V_2 + (\mu_1 - \mu_2)^2} \quad (3.7) \]

and \( 0 < K < 1 \).

Each of the preceding forms gives us reasonable insight into \( \delta_{X_1} \).

From (3.7) we see that \( \delta_{X_1} \) shrinks more heavily toward \( \bar{X}_2 \) if \( V_1 \) is relatively larger than \( V_2 \) or if \( \pi_2 \) is larger than \( \pi_1 \). This is intuitively pleasing because we would expect smaller error when shrinking toward the sample stratum mean with smaller variance.

Furthermore, since we are shrinking toward \( \bar{X} \) in \( \delta_{X_1} \), it also makes sense that we are shrinking more heavily toward the stratum mean with larger stratum weight as this tends to reduce the bias in
estimating $\bar{\mu}$. Also noteworthy is that, as the difference between $\mu_1$ and $\mu_2$ expands to infinity, $\delta_{X_1}$ goes to $\bar{X}$. This demonstrates that the sensitive relationships of stratum variances and weights no longer matters in the face of such different stratum means.

We also study the structure of $\delta_{X_1}$ with respect to how its value compares with $\bar{X}$ by using expression (3.6). If we assume, without loss of generality, that $\bar{X}_1$ is less than $\bar{X}_2$, then the difference $\bar{X}_1 - \bar{X}_2$ is negative. In the case of equal stratum weights, $\delta_{X_1}$ takes a value between $\bar{X}_1$ and $\bar{X}$ when $V_2$ is larger than $V_1$. Again this is a reasonable result as $\bar{X}_1$ will contribute less variance to the final measurement of MSE.

If we now assume equal values of $V_1$ and $V_2$ and a value of $\pi_2$ greater than $1/2$, this will force $\bar{X}$ closer to $\bar{X}_2$ than $\bar{X}_1$. When we inspected (3.7), we saw where such a value of $\pi_2$ would also force $\delta_{X_1}$ closer to $\bar{X}_2$ than $\bar{X}_1$. But here we use (3.6) to determine on which side of $\bar{X}$ the value of $\delta_{X_1}$ will fall. Although $\delta_{X_1}$ is nearer $\bar{X}_2$ than $\bar{X}_1$, we also see that $\delta_{X_1}$ is on the $\bar{X}_1$ side of $\bar{X}$. The estimator $\delta_{X_1}$, in attempting to reduce MSE by being less biased, also attempts to minimize MSE by shrinking to the side of $\bar{X}$ where the smaller stratum weight will reduce variance. So $\delta_{X_1}$ is balancing variance and bias in its attempt to minimize MSE.

The comparison of $\delta_{X_C}$ and $\delta_{X_1}$ is now a relevant issue. When $p = 2$, it can be shown from (3.3) that
\[ c_i = \frac{v_i}{\pi_i (v_1 + v_2 + (\mu_1 - \mu_2)^2)} \]

where \((i, j) = (1, 2)\) or \((i, j) = (2, 1)\). With these values for the two shrinking factors, we find the value of \(c_2 - c_1\) to be

\[ \frac{\pi_2 v_2 - \pi_1 v_1}{\pi_1 \pi_2 (v_1 + v_2 + (\mu_1 - \mu_2)^2)}. \]

But this is the minimizing value for \(c_2 - c_1\) for the estimator \(\hat{\xi}_{\text{Xi}}\). This implies that, for \(p = 2\), the component-wise values of \(c_1\) and \(c_2\) not only minimize the sum of the component MSEs but the sum of the cross product terms in (3.2) as well. This can be shown explicitly:

\[
\frac{\partial}{\partial c_1} \left[ \sum_{i \neq j} \pi_i \pi_j E[((1 - c_1)(\overline{x}_i - \overline{x}) + \overline{x} - \mu_1)((1 - c_j)(\overline{x}_j - \overline{x}) + \overline{x} - \mu_j)] \right]
\]

\[
= \frac{\partial}{\partial c_1} \left[ 2\pi_1 \pi_2 E[((1 - c_1)(\overline{x}_1 - \overline{x}) + \overline{x} - \mu_1)((1 - c_2)(\overline{x}_2 - \overline{x}) + \overline{x} - \mu_2)] \right] = 0.
\]

Hence,
\[ 1 - c_2 = \frac{-E(\bar{X}_1 \bar{X}) + \mu_1 \mu_2 + \bar{\mu}^2 + V(\bar{X}) - \mu_2 \bar{\mu}}{\mu_1 \mu_2 - E(\bar{X}_1 \bar{X}) - E(\bar{X}_2 \bar{X}) + V(\bar{X}) + \bar{\mu}^2} \]

\[ - \frac{\mu_1 \bar{\mu} - \pi_1 V_1 + \mu_1 \mu_2 + \bar{\mu}^2 + V(\bar{X}) - \mu_2 \bar{\mu}}{\mu_1 \mu_2 - \mu_1 \bar{\mu} - \pi_1 V_1 - \mu_2 \bar{\mu} - \pi_2 V_2 + V(\bar{X}) + \bar{\mu}^2} \]

\[ = \frac{(\mu_1 - \bar{\mu})(\mu_2 - \bar{\mu}) + \pi_2(\pi_2 V_2 - \pi_1 V_1)}{-(\mu_1 - \bar{\mu})(\mu_2 - \bar{\mu}) - \pi_1 \pi_2 (V_1 + V_2)} \]

\[ = \frac{-\pi_1 \pi_2 (\mu_1 - \mu_2)^2 + \pi_2(\pi_2 V_2 - \pi_1 V_1)}{-\pi_1 \pi_2 (\mu_1 - \mu_2)^2 - \pi_1 \pi_2 (V_1 + V_2)} \]

\[ = 1 - \frac{V_2}{\pi_1 (V_1 + V_2 + (\mu_1 - \mu_2)^2)}. \]

Similarly,

\[ 1 - c_1 = 1 - \frac{V_1}{\pi_2 (V_1 + V_2 + (\mu_1 - \mu_2)^2)}. \]

These are the values of 1 - c_1 and 1 - c_2 used in the estimator \( \delta_{X_c} \).

To study the case of \( \delta_{X_1} \) for \( p > 2 \) and to learn more about the relationship between \( \delta_{X_1} \) and \( \delta_{X_c} \) we now consider:
\[
\text{MSE}(\delta_{X_1}) = E\left[ \sum_{i=1}^{p} \pi_i \left[ (1 - c_i)(\bar{X}_i - \bar{X}) + \bar{X} - \mu_1 \right]^2 \right]
\]

\[
= E\left[ \sum_{i=1}^{p} \pi_i \left[ (1 - c_i)(\sum_{j \neq i} \pi_j X_j) - \sum_{j \neq i} \pi_j X_j \right]^2 \right]
\]

\[
= E\left[ \sum_{i=1}^{p} \pi_i (1 - c_i)(\sum_{j \neq i} \pi_j X_j) + \bar{X} - \mu \right]^2
\]

Again we must concentrate on solving for the minimizing values of the differences of shrinking factors. Differentiating with respect to \(c_2 - c_1\),

\[
\frac{\partial}{\partial (c_2 - c_1)} \text{MSE}(\delta_{X_1}) = 2\pi_1 \pi_2 (c_2 - c_1) E(\bar{X}_1 - \bar{X}_2)^2 + 2\pi_1 \pi_2 E[(\bar{X}_1 - \bar{X}_2)(\bar{X} - \mu)]
\]

\[
+ 2\pi_1 \pi_2 \sum_{k \neq 1}^{p} \pi_k (c_k - c_1) E[(\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_k)]
\]

\[
+ 2\pi_1 \pi_2 \sum_{k \neq 1, 2}^{p} \pi_k (c_k - c_2) E[(\bar{X}_1 - \bar{X}_2)(\bar{X}_2 - \bar{X}_k)]
\]
This represents one equation in a system of \( \binom{p}{2} \) equations with 
\( p - 1 \) unknown values of \( c_i - c_j \) (\( i = j + 1, j = 1, 2, \ldots, p - 1 \)).

To utilize the above system of equations in solving for \( c_i - c_j \),
we must, in practice, substitute estimates for the unknown parameters.
If the system does offer unique solutions for the differences in
factors, they can be found using a computer routine and then substituted
into the estimator \( \delta_{X_1} \).

One wonders, now, if the equivalence between \( \delta_{X_C} \) and \( \delta_{X_1} \)
can be extended to the case \( p > 2 \). Should \( p = 3 \), \( \pi_1 = \pi_2 = \pi_3 = 1/3 \),
\( \mu_1 = \mu_2 = \mu_3 = \mu \) and \( V_i = i \), the system can be used to solve for the
value of \( c_2 - c_1 \):

\[
c_2 - c_1 = \frac{V_2 - V_1 + \pi V_2 \sum_{k \neq 1}^{p} (c_k - c_2) - V_1 \sum_{k \neq 1}^{p} (c_k - c_1)}{\pi (V_1 + V_2)}
\]

\[
= \frac{1 + 1/3[2(c_3 - c_2) - 1(c_3 - c_1)]}{1/3[1 + 2]}
\]

Finally,
\[ c_2 - c_1 = \frac{3(2 - 1)}{1 \cdot 2 + 1 \cdot 3 + 2 \cdot 3} = \frac{3}{11} \]

upon substitution of \( c_3 - c_2 \) and \( c_3 - c_1 \).

The values of \( c_1 \) and \( c_2 \) for the similar set of parameters gives

\[ c_1 = \frac{V_1(1 - \pi)}{V_1(1 - 2\pi) + \pi^2(V_1 + V_2 + V_3)} \]

\[ = \frac{1 \cdot 2/3}{1 \cdot 1/3 + 1/9(1 + 2 + 3)} = \frac{2}{3} \]

\[ c_2 = \frac{V_2(1 - \pi)}{V_2(1 - 2\pi) + \pi^2(V_1 + V_2 + V_3)} \]

\[ = \frac{2 \cdot 2/3}{2 \cdot 1/3 + 1/9(1 + 2 + 3)} = 1 \]

and, hence, \( c_2 - c_1 = 1/3 \) when utilizing the estimator \( \hat{\sigma}_{Xc} \).

This simple counterexample proves the lack of equivalence between \( \hat{\sigma}_{Xc} \) and \( \hat{\sigma}_{X_1} \). But now we must consider the reason for the equivalence when \( p = 2 \) and its lack of equivalence for \( p > 2 \).

In our previous discussion, it was pointed out how \( \hat{\sigma}_{X_1} \) uses the relationships of the different strata in determining the values of the differences of the shrinking factors. But when \( p = 2 \), the expression \( (1 - c_1)(\overline{X}_1 - \overline{X}) + \overline{X} \) in \( \hat{\sigma}_{Xc} \) immediately involves, through
the simplicity of its form, the special relationship between the two strata and hence $\delta_{XC}$ is able to account for the cross-product terms in (3,2) in this way. But $(1 - c_1)(\bar{X}_1 - \bar{X}) + \bar{X}$ does not sufficiently weigh the relationships between all the strata when $p > 2$, and hence there is a discrepancy between $\delta_{XC}$ and $\delta_{X_1}$ for $p > 2$.

The possible use of $\hat{\mu}$ in a shrinkage estimator is again studied where the MSE to be minimized is

$$\text{MSE}(\hat{\mu}_{11}) = E[\pi_1[(1 - c_1)(\bar{X}_1 - \hat{\mu}) + \hat{\mu}]]$$

$$+ \pi_2[(1 - c_2)(\bar{X}_2 - \hat{\mu}) + \hat{\mu}] - \mu]^2$$

$$= E[\pi_1(1 - c_1)(\bar{X}_1 - \hat{\mu}) + \pi_2(1 - c_2)(\bar{X}_2 - \hat{\mu}) + \hat{\mu} - \mu]^2$$

$$= E[\pi_1(1 - c_1)(\bar{X}_1 - \hat{\mu}) + \pi_2(1 - c_2)(\bar{X}_2 - \hat{\mu}) + \hat{\mu} - \mu]^2$$

$$= E[\pi_1(1 - c_1)(\bar{X}_1 - \hat{\mu}) + \pi_2(1 - c_2)(\bar{X}_2 - \hat{\mu}) + \hat{\mu} - \mu]^2$$

$$= E[\pi_1(1 - c_1)(\bar{X}_1 - \hat{\mu}) + \pi_2(1 - c_2)(\bar{X}_2 - \hat{\mu}) + \hat{\mu} - \mu]^2$$

$$= E[\pi_1(1 - c_1)(\bar{X}_1 - \hat{\mu}) + \pi_2(1 - c_2)(\bar{X}_2 - \hat{\mu}) + \hat{\mu} - \mu]^2$$
where
\[
K^\mu = \frac{\tau_1(1 - c_1)V_1}{V_1 + V_2} + \frac{\tau_2(1 - c_2)V_2}{V_1 + V_2}
\]
and \( \mu \) is the common stratum mean.

Now,
\[
\frac{\partial}{\partial K^\mu} \text{MSE}(\mu^\mu) = \frac{\partial}{\partial K^\mu} \left[ \frac{K^\mu}{\mu} \left( E(\bar{X}_1 - \bar{X}_2)^2 + 2K^\mu E[(\bar{X}_1 - \bar{X}_2)(\hat{\mu} - \mu)] + E(\hat{\mu} - \mu)^2 \right) \right]
\]
\[
= 2K^\mu \frac{E(\bar{X}_1 - \bar{X}_2)^2}{E(\bar{X}_1 - \bar{X}_2)^2} + 2E[(\bar{X}_1 - \bar{X}_2)(\hat{\mu} - \mu)] = 0.
\]

Hence,
\[
K^\mu = \frac{E[(\bar{X}_1 - \bar{X}_2)(\mu - \hat{\mu})]}{E(\bar{X}_1 - \bar{X}_2)^2} = \frac{-E[(\bar{X}_1 - \bar{X}_2)(\hat{\mu} - \mu)]}{E(\bar{X}_1 - \bar{X}_2)^2}
\]
\[
= \frac{-[V_2E(\bar{X}_1)^2 - V_1E(\bar{X}_2)^2 - V_2\mu^2 + V_1\mu^2]}{(V_1 + V_2)E(\bar{X}_1 - \bar{X}_2)^2}
\]
\[
= \frac{-[V_2V_1 + V_2\mu^2 - V_1V_2 - V_1\mu^2 - V_2\mu^2 + V_1\mu^2]}{(V_1 + V_2)E(\bar{X}_1 - \bar{X}_2)^2} = 0.
\]

For this value of \( K^\mu \), the estimator \( \mu^\mu \) is merely \( \hat{\mu} \). Again, shrinking to \( \hat{\mu} \) does not offer any improvement over \( \hat{\mu} \) alone.

3.5 Monte Carlo Results for \( p = 2 \)

So far, we have discussed shrinkage estimators of \( \mu \) that perform better than the usual stratified sample mean. These estimators,
however, depend upon unknown population parameters. In fact, they
depend upon the stratum means which are essentially what we hope to
estimate. In order to use these estimators in practice, sample
statistics must be substituted for the unknown parameters. The
ensuing estimators are not necessarily going to maintain all the
good properties that we have discussed, but will still show merit
themselves. Because the distribution of these more practical estimators
are difficult to derive, we must depend upon a simulation study in
order to judge performance.

Five thousand sets of sample stratum means and variances are
generated (for the case \( p = 2 \)), from which the expected values of the
estimators and their MSEs are approximated. We find that the
simulated MSEs of the estimators, when stratum variances are assumed
unknown, are only from 1 to 4 percent larger than the
the stratum means are assumed unknown.

Let us first inspect the performance of \( \hat{\delta}_{X_1} \). First we check its
behavior for different settings of stratum variances and weights.
Then, from these results, we compare \( \hat{\delta}_{X_1} \) to both \( \delta_{\theta_c} \) and \( \delta_{\theta_1} \) where we
assume \( (V_1, V_2) \) is known, \( \theta_1 = \theta_2 = 0 \) and \( n_1 = n_2 = 10 \). To do this,
the information in Table 5 is used.

When \( (\mu_1, \mu_2) = (0, 2) \) the value of \( \bar{X}_1 \) will clearly be less than
\( \bar{X}_2 \) most of the time, so the setting of the discussion on (3.6) and
(3.7) exists. If we let the location of the star (*) on the graph
show where the average simulation value of \( \hat{\delta}_{X_1} \) lies, we can see that
Table 5. Ratios of MSE of proposed estimators to the variance of the usual stratified sample mean for different choices of strata parameters

<table>
<thead>
<tr>
<th>(μ₁, μ₂)</th>
<th>(V₁, V₂)</th>
<th>(π₁, π₂)</th>
<th>μ₂</th>
<th>MSEδ̂_{X_1}</th>
<th>MSEδ̂_{θ_1}</th>
<th>MSEδ̂_{θ_c}</th>
<th>MSEδ̂_{θ}</th>
<th>Line graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 2)</td>
<td>(5.0, 1.0)</td>
<td>(1/4, 3/4)</td>
<td>1.5</td>
<td>4/16</td>
<td>.9589</td>
<td>.9841</td>
<td>.9238</td>
<td>1.0520</td>
</tr>
<tr>
<td>(0, 2)</td>
<td>(3.0, 3.0)</td>
<td>(1/4, 3/4)</td>
<td>1.5</td>
<td>36/16</td>
<td>.9211</td>
<td>.7925</td>
<td>.8123</td>
<td>.8842</td>
</tr>
<tr>
<td>(0, 2)</td>
<td>(1.0, 5.0)</td>
<td>(1/4, 3/4)</td>
<td>1.5</td>
<td>196/16</td>
<td>.7708</td>
<td>.6619</td>
<td>.7140</td>
<td>.7543</td>
</tr>
<tr>
<td>(0, 2)</td>
<td>(1.0, 5.0)</td>
<td>(1/2, 1/2)</td>
<td>1.0</td>
<td>64/16</td>
<td>.8487</td>
<td>.6727</td>
<td>.6774</td>
<td>.7025</td>
</tr>
<tr>
<td>(0, 2)</td>
<td>(3.0, 3.0)</td>
<td>(1/2, 1/2)</td>
<td>1.0</td>
<td>0/16</td>
<td>.9750</td>
<td>.7004</td>
<td>.6577</td>
<td>.7025</td>
</tr>
</tbody>
</table>

- a The vertical line on the line graph marks μ for each set of parameters.
- b The star (*) indicates the position of the simulated expected value of the estimator δ̂_{X_1} for each set of parameters.
- c The line connecting the star to the vertical line indicates the simulated size of the bias of the estimator δ̂_{X_1}.
this location is on the left side of the respective value of $\mu$ when $\pi_2 V_2 - \pi_1 V_1$ is positive and on the right otherwise. This agrees with our discussion involving (3.6).

Also note that the starred location is closer to 2 than 0 whenever $\pi_2$ is larger than 1/2. Furthermore, this location seems to be tugged to the left when $V_1$ is small and to the right when $V_2$ is small. This shows the estimator is favoring, to some degree, the stratum with smaller variance. On the fifth line, the starred location is at .973 which is very reasonable considering that $\mu = 1$ and that the strata show balanced weighting with respect to the $\pi_1$'s and the $V_i$'s. The comments of this paragraph agree with our discussion involving expression (3.7).

When $(\pi_2 V_2 - \pi_1 V_1)^2$ is large, $\bar{\delta}_{X_1}$ does its best improving upon $\bar{VX}$ since

$$\text{MSE}(\bar{\delta}_{X_1}) = \bar{VX} - \frac{(\pi_2 V_2 - \pi_1 V_1)^2}{V_1 + V_2 + (\mu_1 - \mu_2)^2}.$$  

This is upheld in the simulation studies as columns 5 and 6 in Table 5 show. Note that, even under total balance of the strata, $\bar{\delta}_{X_1}$ dominates $\bar{X}$.

Also note that the horizontal lines give some indication of the bias of the estimator $\bar{\delta}_{X_1}$.

To compare $\bar{\delta}_{X_1}$ to the estimators $\bar{\delta}_{C}$ and $\bar{\delta}_{O}$, we see from
columns 7 and 8 of Table 5 that $\delta_{\overline{X}_1}$ usually performs worse than either
of the other two. This result is not surprising considering that
$\theta_1 = \theta_2 = 0$ was used and that the true means $(\mu_1, \mu_2) = (2, 0)$ are
reasonably close to $(0, 0)$.

When the values of the means are not close to $(0, 0)$ relative to
the variation within the strata, we can expect $\delta_{\theta_c}$ and $\delta_{\theta_i}$ to lose
their dominance over $\delta_{\overline{X}_1}$.

If we set $(\mu_1, \mu_2) = (3, 2), (\nu_1, \nu_2) = (3, 6), (\pi_1, \pi_2) =
(1/2, 1/2), (\Theta_1, \Theta_2) = (0, 0)$ and $(n_1, n_2) = (10, 10), we find

\[
\frac{\text{MSE}(\delta_{\overline{X}_1})}{\text{MSE}(\delta_{\theta_c})} = .8090
\]

and

\[
\frac{\text{MSE}(\delta_{\overline{X}_1})}{\text{MSE}(\delta_{\theta_i})} = .7528 .
\]

In relation to $(\nu_1, \nu_2)$, the stratum means are far from $(\Theta_1, \Theta_2)$ and
subsequently $\delta_{\overline{X}_1}$ not only dominates $\overline{X}$ but also dominates $\delta_{\theta_c}$ and $\delta_{\theta_i}$.

If we set $(\mu_1, \mu_2) = (4, 2)$ and $(\nu_1, \nu_2) = (48, 60)$ while holding
the other values the same as before, we find

\[
\frac{\text{MSE}(\delta_{\overline{X}_1})}{\text{MSE}(\delta_{\theta_c})} = 1.7947
\]

and
Here the means are more distant from \((\theta_1, \theta_2)\) than before, yet they are much closer relative to the stratum variances.

To more graphically represent the performance of these estimators, Figure 3 considers the case of equal stratum means, \((V_1, V_2) = (3.75, 5), (\pi_1, \pi_2) = (1/2, 1/2)\) and \((n_1, n_2) = (8, 12)\). If we choose to utilize Neyman allocation for these choices of \((V_1, V_2)\) and \((\pi_1, \pi_2)\), we would find \((8, 12)\) to be the optimal choice of \((n_1, n_2)\). Therefore, \(\bar{X}\) not only represents the classical stratified estimator, but also the optimal choice of \(\bar{X}\) based upon Neyman allocation. From this graph, we now can make clear statements on the performance of the estimators.

It was known previous to the simulation work that \(\delta_{\theta_1}\) dominates \(\delta_{\theta_2}\) which dominates \(\bar{X}\). From Figure 3, it is clear that the estimators perform much worse when \((\mu_1, \mu_2)\) is assumed unknown. But also assuming unknown stratum variances adds very little further loss.

As to be expected, \(\delta_{\theta_1}\) and \(\delta_{\theta_2}\) perform well when \((\mu_1, \mu_2)\) is near the value \((\theta_1, \theta_2) = (0, 0)\).

Also note that \(\text{MSE}(\bar{X})\) does not change at all as the common value of \(\mu\) changes. Since Figure 3 assumes \(\mu_1 = \mu_2\) and constant values for the stratum variances and weights and from (3.5), this result is expected.

Finally, note that the simple Thompson estimator \((1 - c)(\bar{X} - \bar{\theta}) + \bar{\theta}\) does as well or better than many of the estimators in this
chapter for many values of the common mean $\mu$. Again it must be pointed out that the stratum weights and variances are either equal or near equal. The advantage of these estimators over Thompson's is that a different shrinking factor is employed for every stratum. So one would expect the best improvement when the strata are quite different in weights and variances. Columns 7, 8, and 9 on Table 5 show clearly that the Thompson estimator of $\bar{\mu}$ is consistently beaten when the strata are reasonably different in this way.

As in Chapter 2, the consistency of simulated results are demonstrated by the standard deviation of 16 generated expected values of our shrinkage estimators. The component parameters used are the same as those used in Table 4 with the results appearing in Table 6.

Also listed on Table 6 are the proportions of MSE due to the biasedness of our estimators. The larger proportions are due to the shrinkage to 0, but, overall, the bias is not dominating the MSE.

From our Monte Carlo study, we see that our estimators do improve upon $\bar{X}$ for at least a portion of the parameter space. The estimators $\delta_{\Theta_1}$ and $\delta_{\Theta_c}$ perform well if the value of $(\Theta_1, \Theta_2)$ is relatively near the actual $(\mu_1, \mu_2)$. The estimators also fare well against a Thompson-type estimator of $\bar{\mu}$ and even the estimator $\bar{X}$ under Neyman allocation.
(3.0) $\text{MSE}(\pi_1 \bar{X}_1 + \pi_2 \bar{X}_2) = 2.1875$
(3.1) $\text{MSE}(\delta_{\theta_1})$, all parameters known
(3.2) $\text{MSE}(\delta_{\theta_2})$, only $V_1, V_2$ known
(3.3) $\text{MSE}(\delta_{\theta_3})$, only $V_1, V_2$ known
(3.4) $\text{MSE}(\delta_{X_1}) = 2.1429$, all parameters known

+ $\text{MSE}(\delta_{\theta_c})$, all parameters known.
* $\text{MSE}(\delta_{\theta_1})$, no parameters known.

NOTE: $\text{MSE}(\delta_{X_1}) = 2.1524$, only $V_1, V_2$ known. $\text{MSE}(\delta_{X_1}) = 2.1668$, no parameters known.

Figure 3. Mean square error versus common stratum mean
Table 6. Measurements of (i) bias to MSE, (ii) consistency of simulated results for estimation of μ where \( \mu_1 = 2, \mu_2 = 4, \theta_1 = \theta_2 = 0, \pi_1 = \pi_2 = 1/2 \) and \( V_1 = 3 \) and \( V_2 = 6 \)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( V_1, V_2 ) assumed</th>
<th>(Bias)(^2) of simulated values of the estimator</th>
<th>Standard deviations of simulated values of the estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_\theta )</td>
<td>known</td>
<td>.1284</td>
<td>.0185</td>
</tr>
<tr>
<td>( \delta_\theta )</td>
<td>unknown</td>
<td>.1669</td>
<td>.0167</td>
</tr>
<tr>
<td>( \delta_{\theta_1} )</td>
<td>known</td>
<td>.2092</td>
<td>.0190</td>
</tr>
<tr>
<td>( \delta_{\theta_1} )</td>
<td>unknown</td>
<td>.1942</td>
<td>.0189</td>
</tr>
<tr>
<td>( \delta_{\bar{X}} = \delta_{\bar{X}_1} )</td>
<td>known</td>
<td>.0056</td>
<td>.0190</td>
</tr>
<tr>
<td>( \delta_{\bar{X}} = \delta_{\bar{X}_1} )</td>
<td>unknown</td>
<td>.0048</td>
<td>.0200</td>
</tr>
</tbody>
</table>
4. SHRINKAGE ESTIMATORS OF A MULTIVARIATE POISSON MEAN VECTOR

Let $X_1, X_2, \ldots, X_p$ be $p$ independent Poisson variables with respective means $\lambda_1, \lambda_2, \ldots, \lambda_p$. This chapter deals with the estimation of $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$. Restriction to a single observation from each distribution does not involve any loss of generality because if $X_{i1}, X_{i2}, \ldots, X_{in_i}$ denotes a random sample of size $n_i$ from a Poisson distribution with mean $\lambda_i$ ($i = 1, 2, \ldots, p$), then the minimal sufficient statistic for $\lambda_i$ is $\sum_{j=1}^{n_i} X_{ij}$, which is again Poisson ($\lambda_i'$) with $\lambda_i' = n_i \lambda_i$, $i = 1, 2, \ldots, p$.

The case $p = 1$ was considered by Thompson (1968). He wanted to obtain an estimator for $\lambda_1$, of the type $\lambda_{10} + (1 - c)(X_1 - \lambda_{10})$, where $\lambda_{10}$ is a prechosen constant. In Thompson's case, $c$ was obtained as follows.

First consider an estimator of type $\lambda_{10} + (1 - c)(X_1 - \lambda_{10})$ for $\lambda_1$. Such an estimator has mean square error (MSE) $E_\lambda \lambda_{10} + (1 - c)^2 (X_1 - \lambda_{10})^2 = (1 - c)^2 \lambda_1 + c^2 (\lambda_1 - \lambda_{10})^2$. Such a MSE is minimized at $c = \lambda_1 / [\lambda_1 + (\lambda_1 - \lambda_{10})^2]$. Since $\lambda_1$ is unknown, substituting the estimator $X_1$ for $\lambda_1$, one gets the estimator $\lambda_{10} + (1 - X_1 / [X_1 + (X_1 - \lambda_{10})^2]) (X_1 - \lambda_{10}) = \lambda_{10} + (X_1 - \lambda_{10})^3 / [X_1 + (X_1 - \lambda_{10})^2]$ as proposed by Thompson (he used $n$ observations instead of 1).

In order to generalize the results of Thompson to the $p$-dimensional
case, we use a common shrinking factor that involves information from each of the p univariate populations. Such an estimator can be written as $\lambda_1 = \lambda_{i0} + (1 - c)(X_1 - \lambda_{i0})$. We want to minimize

$$\sum_{i=1}^{p} \text{MSE}(\lambda_1) = \sum_{i=1}^{p} \left[ (1 - c)(X_1 - \lambda_{i0}) - \lambda_1 \right]^2$$

$$= \sum_{i=1}^{p} (1 - c)^2 \lambda_1 + c^2 \sum_{i=1}^{p} (\lambda_i - \lambda_{i0})^2.$$

Setting

$$\frac{\partial}{\partial c} \sum_{i=1}^{p} \text{MSE}(\lambda_1) = -2(1 - c) \sum_{i=1}^{p} \lambda_i + 2c \sum_{i=1}^{p} (\lambda_i - \lambda_{i0})^2$$

equal to zero, one gets

$$c = \frac{\sum_{i=1}^{p} \lambda_1}{\left[ \sum_{i=1}^{p} \lambda_i + \sum_{i=1}^{p} (\lambda_i - \lambda_{i0})^2 \right]}.$$

Again substituting $X_1$ for $\lambda_1$, we obtain the estimator

$$\left[ \lambda_{10} + (1 - c)(X_1 - \lambda_{10}), \ldots, \lambda_{p0} + (1 - c)(X_p - \lambda_{p0}) \right]$$

with

$$\hat{c} = \frac{\sum_{i=1}^{p} X_i}{\sum_{i=1}^{p} X_i + \sum_{i=1}^{p} (X_i - \lambda_{i0})^2}$$

$$= \frac{\bar{X}}{\bar{X} + \frac{1}{p} \sum_{i=1}^{p} (X_i - \lambda_{i0})^2},$$

where
We could also consider a Thompson-type estimator \( \hat{\lambda}_{1X}, \hat{\lambda}_{2X}, \ldots \) \( \hat{\lambda}_{pX} \) with \( \hat{\lambda}_{iX} = \bar{x} + (1 - c)(x_i - \bar{x}) \) which shrinks toward the mean of the entire sample. We then must consider

\[
\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iX}) = \sum_{i=1}^{p} \mathbb{E}[x_i + (1 - c)(x_i - \bar{x}) - \lambda_i]^2
\]

\[
= \sum_{i=1}^{p} \mathbb{E}[(1 - c)x_i + c\bar{x} - \lambda_i]^2
\]

\[
= \sum_{i=1}^{p} [(1 - c)^2 x_i + c^2 \bar{x} + 2c(1 - c) \text{ Cov}(x_i, \bar{x})] + [(1 - c)\lambda_i - \lambda_i + c \frac{\sum_{i=1}^{p} \lambda_i}{p}]^2
\]

\[
= \sum_{i=1}^{p} [(1 - c)^2 \lambda_i + c^2 \frac{\sum_{i=1}^{p} \lambda_i}{p} + 2c(1 - c) \frac{\lambda_i}{p}] + \sum_{i=1}^{p} c^2 [\lambda_i - \frac{\sum_{i=1}^{p} \lambda_i}{p}]^2
\]

\[
= \sum_{i=1}^{p} \lambda_i \left[\frac{c^2}{p} + (1 - c)^2 + \frac{2c(1 - c)}{p}\right] + c^2 \sum_{i=1}^{p} [\lambda_i - \frac{\sum_{i=1}^{p} \lambda_i}{p}]^2.
\]

Setting
\[
\frac{\partial \sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iX})}{\partial c} = \frac{p}{\sum \lambda_i} \frac{2c}{p} - 2(1 - c) + \frac{2(1 - 2c)}{p} \\
+ 2c \sum_{i=1}^{p} \left[\lambda_i - \frac{\sum \lambda_i}{p}\right]^2
\]

equal to zero, one gets
\[
c = \frac{p-1}{p} \frac{\sum \lambda_i}{[\sum \lambda_i] - \frac{p}{\sum \lambda_i} \left[\frac{\sum \lambda_i}{p}\right]^2} = \frac{\bar{\lambda}}{\lambda + s_{\lambda}^2}
\]

where
\[
\bar{\lambda} = \frac{\sum \lambda_i}{p}
\]

and
\[
s_{\lambda}^2 = \frac{\sum (\lambda_i - \bar{\lambda})^2}{p-1}
\]

Again \(X_i\) would be substituted for \(\lambda_i\) in \(c\) to obtain a usable estimator.

Comparing the form of \(\hat{\lambda}_{iX}\) and its MSE minimizing shrinking factor to that of \(\hat{\delta}_{iX}\) and its factor in Chapter 2, there is much similarity between them. If we considered \(n_i = 1\) for all \(i\) and, hence, \(w_i = \frac{1}{p}\) for all \(i\), the shrinking factor of \(\hat{\delta}_{iX}\) is again the mean of the component variances divided by the sum of this mean and the variance of the component means. So the amount of shrinkage depends on the variability of the component means and the average amount of component.
variation whether the underlying multivariate distribution is normal, Poisson or, as we will see in Chapter 5, gamma.

If we allowed the shrinking factor of this estimator to differ for each component, we would have the component estimator

\[ \hat{\lambda}_{iX_i} = \bar{X} + (1 - c_i)(X_i - \bar{X}). \]

In order to minimize

\[
\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iX_i}) = \sum_{i=1}^{p} \mathbb{E}[(1 - c_i)X_i + c_i\bar{X} - \lambda_i]^2
\]

\[
= \sum_{i=1}^{p} \left[ (1 - c_i)^2\bar{X} + c_i^2\bar{X} \right] + 2c_i(1 - c_i)\text{Cov}(X_i, \bar{X}) + \sum_{i=1}^{p} \left[ (1 - c_i)\lambda_i - \lambda_i + c_i \frac{\lambda_i}{p} \right]^2
\]

we set

\[
\frac{\partial}{\partial c_i} \sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iX_i}) = -2(1 - c_i)\lambda_i + 2c_i \frac{\lambda_i}{p} + 2(1 - c_i)\frac{\lambda_i}{p} + 2c_i(\lambda_i - \frac{\lambda_i}{p})^2
\]
equal to zero. This implies

\[ c_i = \frac{1 - \frac{p-1}{p}}{\frac{1}{\lambda_i} + \frac{\sum_{i=1}^{p} \lambda_i}{p^2} - 2\frac{1}{p} + (\lambda_i - \frac{1}{p})^2} \]

\[ = \frac{(p-1)\lambda_i}{(p-1)\lambda_i - (\lambda_i - \frac{1}{p}) + p(\lambda_i - \frac{1}{p})^2} \]

One would expect that shrinking to the mean of the entire sample would be most reasonable when the \( \lambda_i \)'s are of nearly equal value. If \( \lambda_i = \lambda \) for \( i = 1, 2, ..., p \), the value of \( c_i \), above, becomes 1 for each component. This forces the estimator \( \hat{\lambda}_i \) to simply be \( \bar{x} \) for each component.

The Bayes estimator of \( \lambda_i \) is now studied where \( \lambda_i \) is assumed to have come from a gamma distribution with parameters \( \alpha_i \) and \( \beta_i \). Note that the Bayes estimate is

\[ B[\lambda_i/\bar{x}_i = x_i] = \frac{(x_i + \alpha_i)\beta_i}{\beta_i + 1} \]

where \( \alpha_i > 0 \) and \( \beta_i > 0 \). Writing the prior mean \( \alpha_i\beta_i = \lambda_{i0} \) and letting \( (\beta_i + 1)^{-1} = c_i \), such an estimate can also be written as
\[ \lambda_{10} + (1 - c_1 \lambda \delta_1 - \lambda_{10}) \]

Albert (1981) finds the shrinking factor \( k \) that minimizes

\[
\sum_{i=1}^{p} E\left[ \lambda_{10} + (1 - c_1 k)(X_i - \lambda_{10}) - \lambda_i \right]^2 =
\]

\[
\sum_{i=1}^{p} (1 - c_1 k)^2 + k^2 \sum_{i=1}^{p} c_i^2 \lambda_i - \lambda_{10}^2 .
\]

The minimizing constant \( k \) turns out to be

\[
k = \frac{\sum_{i=1}^{p} c_i \lambda_i}{\sum_{i=1}^{p} c_i^2 \lambda_i + \sum_{i=1}^{p} c_i^2 (\lambda_i - \lambda_{10})^2} .
\]

Substituting the estimator \( \hat{X}_i \) for the unknown \( \lambda_i \), one obtains the estimator \( [\lambda_{10} + (1 - c_1 k)(X_i - \lambda_{10}), \ldots, \lambda_p + (1 - c_p k)(X_p - \lambda_{10})] \) for \( \lambda \) where

\[
\hat{k} = \frac{\sum_{i=1}^{p} c_i X_i}{\left[ \sum_{i=1}^{p} c_i^2 X_i + \sum_{i=1}^{p} c_i^2 (X_i - \lambda_{10})^2 \right]}.
\]

Finally, in this section we obtain some empirical Bayes estimators of \( \lambda \). With this end, first find the marginal distributions of the \( X_i \)'s when \( \alpha_1 = \alpha_2 = \ldots = \alpha_p = \alpha \) and \( \beta_1 = \beta_2 = \ldots = \beta_0 = \beta \).
\[ \beta_p = \beta. \] Note that

\[
f(x_i/\alpha, \beta) = \int_0^\infty f(x_i/\alpha, \beta, \lambda) d\lambda
\]

\[
= \int_0^\infty e^{-\lambda(\beta+1)/\beta} \frac{\lambda^{x_i-1}}{x_i^\alpha \Gamma(\alpha)} d\lambda
\]

\[
= \frac{1}{x_i^\alpha \Gamma(\alpha)} \int_0^\infty e^{-y} \left(\frac{\beta y}{\beta+1}\right)^{x_i-1} \frac{\beta}{\beta+1} dy
\]

\[
= \frac{\beta \Gamma(\alpha+x_i)}{x_i^\alpha (\beta+1) \Gamma(\alpha)}
\]

\[
= \frac{\Gamma(\alpha+x_i)}{x_i^\alpha \Gamma(\alpha)} \left(\frac{1}{\beta+1}\right)^\alpha (1 - \frac{1}{\beta+1}) x_i
\]

which is a generalized negative binomial distribution (since \( \alpha \) need not be an integer) with parameters \( \frac{1}{\beta+1}, \alpha \). Then

\[
\mathbb{E}(X_i) = \frac{(\alpha \beta / (\beta+1))}{(1/\beta+1)} = \alpha \beta
\]

and

\[
\mathbb{V}(X_i) = \frac{\alpha \beta / (\beta+1)}{1/(\beta+1)^2} = \alpha \beta (\beta+1).
\]

Note that the maximum likelihood estimate based on \( x_1, x_2 \ldots x_p \)
is obtained as follows. The log-likelihood function is given by

$$\log L(\beta) = \sum_{i=1}^{p} \log \left( \frac{x_i + \alpha - 1}{\alpha - 1} \right) - px \log (\beta+1)$$

$$+ \left( \sum_{i=1}^{p} x_i \right) \left[ \log \beta - \log (\beta+1) \right].$$

Hence,

$$\frac{\partial \log L(\beta)}{\partial \beta} = -px \frac{1}{\beta+1} + \frac{\sum_{i=1}^{p} x_i}{\beta+1} - \frac{\sum_{i=1}^{p} x_i}{\beta+1},$$

$$= \frac{\sum_{i=1}^{p} x_i - p\alpha}{\beta(\beta+1)} > 0$$

accordingly as

$$\beta > \frac{\sum_{i=1}^{p} x_i}{p\alpha} = \frac{x}{\alpha}.$$

Thus, the maximum likelihood estimator of $\beta$ is $\frac{x}{\alpha}$.

Substituting the above estimator of $\beta$ in the Bayes estimator

$$\frac{(x_i + \alpha)^{\beta}}{\beta + 1}$$

of $\lambda_i$, one gets the resulting empirical Bayes estimator

$$\frac{(x_i + \alpha)x}{x + \alpha}.$$

Leonard (1976) assumes that $\beta$ is unknown, but, instead of using
an estimator of $\beta$ in its place, he puts an improper prior $f(\beta)\alpha$ for $\beta$. Now the improper Bayes estimate of $\frac{\beta}{\beta+1}$ is given by

\[
E_{\frac{\beta}{\beta+1}/X} = \frac{\int_0^\infty \frac{\beta}{\beta+1} \frac{1}{\beta} \pi \left( \frac{x_i+\alpha-1}{\alpha-1} \right) \left( \frac{1}{\beta+1} \right)^{x_i} \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i} d\beta}{\int_0^\infty \frac{1}{\beta} \pi \left( \frac{x_i+\alpha-1}{\alpha-1} \right) \left( \frac{1}{\beta+1} \right)^{x_i} \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i} d\beta}
\]

\[
= \frac{\sum_{i=1}^p x_i \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i} \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i} d\beta}{\sum_{i=1}^p \frac{x_i}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i} \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i} d\beta}
\]

\[
= \frac{\int_0^\infty \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i+\alpha+1} \pi \left( \frac{1}{\beta+1} \right)^{x_i+\alpha+1} d\beta}{\int_0^\infty \frac{1}{\beta+1} \left( \frac{1}{\beta+1} \right)^{x_i+\alpha+1} \pi \left( \frac{1}{\beta+1} \right)^{x_i+\alpha+1} d\beta}
\]

\[
\frac{\Gamma(p\alpha)\Gamma\left( \sum_{i=1}^p x_i + 1 \right)}{\Gamma(p\alpha + \sum_{i=1}^p x_i + 1)} = \frac{\int_0^1 y^{p\alpha-1}(1-y)^{x_i+1} dy}{\int_0^1 y^{p\alpha-1}(1-y)^{x_i+1} dy} = \frac{\Gamma(p\alpha)\Gamma\left( \sum_{i=1}^p x_i \right)}{\Gamma(p\alpha + \sum_{i=1}^p x_i)}
\]
\[
\frac{\sum_{i=1}^{p} x_i}{\sum_{i=1}^{p} x_i + p\alpha} = \frac{x}{x + \alpha}
\]

where the last integrals are integrals of beta density functions.

The two-stage estimator is thus \( \frac{(x_1 + \alpha)\bar{x}}{\alpha + \bar{x}} \), which is exactly the same as the previous empirical Bayes estimator.

Next we find a proper prior distribution for \( \beta \) that will lead us to an estimator which, in its limiting form, is the same as that from the previous two methods. Similar results are considered by Ghosh and Parsian (1981) for a different loss.

Using the following beta distribution as the proper prior on \( \beta \),

\[
f(\beta) = \frac{1}{\beta^{m+1}} (\frac{\beta}{\beta+1})^{n-1} \Gamma(m) \Gamma(n) / \Gamma(m+n)
\]

where \( m > 0, n > 0 \) and

\[
\Gamma(m) \Gamma(n) = \frac{\Gamma(m+n)}{\Gamma(m) \Gamma(n)},
\]

we need to find

\[
E[\lambda_1/x] = E_\beta [E[\lambda_1/x, \alpha, \beta]/m, n]
\]
Focusing our attention on $E_{\beta-\beta+1/m, n, x}$, we find

\[E_{\beta-\beta+1/m, n, x} = \left( x_i + \alpha \right) E_{\beta-\beta+1/m, n, x} .\]

\[= \mathcal{E}_{\beta-\beta+1/m, n, x} \]

\[\mathcal{E}_{\beta-\beta+1/m, n, x} = \frac{\int_0^\infty \left( \frac{1}{\beta+1} \right)^m \left( \frac{\beta}{\beta+1} \right)^{n-1} \frac{p}{n-1} \prod_{i=1}^p x_i^{\alpha-1} \right) \frac{(\beta+1)^x}{B(m, n)} d\beta}{\int_0^\infty \left( \frac{1}{\beta+1} \right)^m \left( \frac{\beta}{\beta+1} \right)^{n-1} \frac{p}{n-1} \prod_{i=1}^p x_i^{\alpha-1} \right) \frac{(\beta+1)^x}{B(m, n)} d\beta} \]

\[= \frac{\int_0^\infty \left( \frac{1}{\beta+1} \right)^{m+p\alpha+2-1} \frac{\beta}{\beta+1}^{n+1+\sum x_i-1} d\beta}{\int_0^\infty \left( \frac{1}{\beta+1} \right)^{m+p\alpha+2-1} \frac{\beta}{\beta+1}^{n+1+\sum x_i-1} d\beta} \]

\[= \frac{\int_0^1 y^{m+p\alpha-1} (1-y)^{\sum x_i+n+1-1} dy}{\int_0^1 y^{m+p\alpha-1} (1-y)^{\sum x_i+n+1-1} dy} \]

\[= \frac{\Gamma(p\alpha+m) \Gamma(\sum x_i+n+1)}{\Gamma(p\alpha+m+\sum x_i+n)} \quad \frac{n^p \sum_{i=1}^{p} x_i}{m+n+p\alpha+ \sum_{i=1}^{p} x_i} .\]
\[
\bar{x} + \frac{n}{p} \cdot \frac{\bar{x} + \alpha + \frac{m+n}{p}}{\bar{x} + \alpha + \frac{m+n}{p}}.
\]

So
\[
E[\lambda_i/X] = (X_i + \alpha)(\frac{\bar{x} + \frac{n}{p}}{\bar{x} + \alpha + \frac{m+n}{p}})
\]
\[
= (1 - \frac{\frac{m}{p} + \alpha}{\frac{m+n}{p} + \bar{x} + \alpha})(X_i + \alpha),
\]

and, as \(m\) and \(n\) both go to zero, this two-stage proper Bayesian estimator goes to \((X_i + \alpha)(\frac{\bar{x}}{\alpha + \bar{x}})\). Thus, a proper prior on \(\beta\) can be chosen so that the Bayes estimator, in the limiting form, is exactly the same as the empirical Bayes or the improper Bayes estimator found earlier.
5. SHRINKAGE ESTIMATORS OF A MULTIVARIATE GAMMA SCALE PARAMETER VECTOR

5.1 Introduction

In this chapter, attention is focused on the multivariate gamma case with mean vector \( \mathbf{\mu} = (\gamma_1, \gamma_2, \ldots, \gamma_p) \) where \( \mathbf{\mu} = (\lambda_1, \lambda_2, \ldots, \lambda_p) \) is unknown and is to be estimated, while \( \gamma \) is known. We will discuss Thompson-type shrinkage estimators, Bayes estimators and empirical Bayes estimators of several types. Many of our estimators are similar to those in the previous chapters and several are totally new. After deriving the estimators using different viewpoints, they will be evaluated using a simulation study in the final section of this chapter. Furthermore, we should note that all these estimators are expressed as shrinkage estimators and, to insure the shrinking factors fall in the interval \([0, 1]\), the simulation study utilizes a positive-part rule similar to that discussed in Chapter I.

As a forerunner to the derivations, let us note a simple result which we utilize in the chapter. When we have a Bayes estimator (BE), we would like to express it as a shrinker which shrinks toward the prior mean (PM) and from the maximum likelihood estimator or the best invariant estimator which we call the "usual" estimator (UE). Notationally we write \( \text{BE} = \text{PM} + (1 - c)(\text{UE} - \text{PM}) \).

Solving for the shrinking factor \( c \) is now simple and it can be expressed as
In general, this $c$ will depend on the observations; but in many instances, we shall find that $c$ is a constant.

To motivate the subsequent discussions in this chapter, we first look at the multivariate gamma distribution with $\gamma = 1$. This is merely the exponential case and much of the necessary algebra is simpler to carry out.

Assume the density

$$f(x_i | \lambda_i) = \frac{1}{\lambda_i} e^{-x_i/\lambda_i};$$

$x_i > 0, \lambda_i > 0, i = 1, \ldots, p$. Here $E(X_i) = \lambda_i$, $V(X_i) = \lambda_i^2$ and $E(X_i^2) = 2\lambda_i^2$. Let us assume $\lambda_i$ has an inverted gamma prior distribution with parameters $\alpha_i > 0$ and $\beta_i > 1$. That is

$$g(\lambda_i | \alpha_i, \beta_i) = \left(\frac{\beta_i - 1}{\lambda_i}\right)^{\beta_i - 1} \alpha_i e^{-\alpha_i/\lambda_i}/\Gamma(\beta_i)\lambda_i^{2}$$

for $\lambda_i > 0$, and $i = 1, 2, \ldots, p$.

Now

$$E(\lambda_i) = \frac{\alpha_i}{\beta_i - 1}, \quad V(\lambda_i) = \frac{\alpha_i^2}{(\beta_i - 1)(\beta_i - 2)},$$

and

$$c = \frac{UE - BE}{UE - EM}. \quad (5.1)$$
These facts are used extensively later in the chapter.

The posterior distribution of $\lambda_i$ given $X_i = x_i$ is

$$g(\lambda_i | x_i = x_i, \alpha_i, \beta_i) = \frac{g(\lambda_i, x_i)}{f(x_i)}$$

$$= \frac{f(x_i | \lambda_i) g(\lambda_i | \alpha_i, \beta_i)}{f(x_i)}$$

$$= \frac{\beta_i}{\Gamma(\beta_i)} \frac{1}{\lambda_i^{\beta_i+2}} e^{-\frac{x_i + \alpha_i}{\lambda_i}}$$

$$= \frac{\beta_i}{\Gamma(\beta_i)} \frac{1}{\lambda_i^{\beta_i+2}} e^{-\frac{x_i + \alpha_i}{\lambda_i}} \int_0^{\infty} \frac{\alpha_i}{\Gamma(\beta_i)} \frac{1}{\lambda_i^{\beta_i+2}} e^{-\frac{x_i + \alpha_i}{\lambda_i}} \, d\lambda_i$$

$$= \int_0^{\infty} e^{-\frac{y_i}{x_i + \alpha_i}} \frac{1}{x_i + \alpha_i^{\beta_i+2}} \left( \frac{1}{y_i} \right)^{\beta_i+2} x_i^{\alpha_i+1} \, dy_i$$
where
\[ y_i = \frac{x_i + \alpha_i}{\lambda_i} \]

and
\[ dy_i = \frac{-x_i}{\lambda_i} \frac{1}{\lambda_i} \]  

Assuming squared error loss, this leads to the Bayes estimate of \( \lambda_i \) given by

\[ E[\lambda_i | X_i = x_i] = \int_0^\infty \frac{-(\alpha_i + x_i)}{\lambda_i} \frac{1}{\lambda_i} \frac{e^{-x_i/\lambda_i}}{\Gamma(\beta_i+1)} d\lambda_i \]

\[ = \int_0^\infty y_i \frac{\beta_i+1}{\Gamma(\beta_i+1)} \frac{e^{-x_i/\lambda_i}}{\Gamma(\beta_i+1)} \gamma_i\]

\[ = \frac{x_i + \alpha_i}{\Gamma(\beta_i+1)} \frac{\Gamma(\beta_i)}{\beta_i} = \frac{x_i + \alpha_i}{\beta_i} \]
The Bayes estimator is expressible as a shrinker by utilizing the convention given in (5.1). Thus,

\[
BE = PM + (1 - c_i)(UE - PM) = \frac{\alpha_i}{\beta_i - 1} + (1 - c_i)(X_i - \frac{\alpha_i}{\beta_i - 1})
\]

making

\[
c_i = \frac{X_i - \frac{X_i + \alpha_i}{\beta_i}}{X_i - \frac{\alpha_i}{\beta_i - 1}} = \frac{\beta_i - 1}{\beta_i}.
\]

This estimator is a true shrinker as \(c_i\) falls in the interval \([0, 1]\).

Albert (1981) used such a Bayesian shrinkage estimator for Poisson parameters and found a second constant which utilized information in the other components in the multivariate setting.

Applying his procedure to the gamma case, we want to find the constant \(k\) that minimizes

\[
\sum_{i=1}^{p} \text{MSE}\left[\frac{\alpha_i}{\beta_i - 1}\right] + (1 - k)\frac{\beta_i - 1}{\beta_i}(X_i - \frac{\alpha_i}{\beta_i - 1})]. \tag{5.4}
\]

Using the notation \(\theta_i\) for \(\frac{\alpha_i}{\beta_i - 1}\) and \(c_i\) for \(\frac{\beta_i - 1}{\beta_i}\), this sum is expressible as
\[ \sum_{i=1}^{p} [(1 - c_i k)^2 \lambda_i^2 + (\theta_i - \lambda_i) + (1 - k c_i)(\lambda_i - \theta_i)]^2 \]

\[ = \sum_{i=1}^{p} [(1 - c_i k)^2 \lambda_i^2 + k^2 c_i^2(\lambda_i - \theta_i)^2] \]

\[ = \sum_{i=1}^{p} \lambda_i^2 - 2k \sum_{i=1}^{p} c_i \lambda_i^2 + k^2 \sum_{i=1}^{p} c_i^2 \lambda_i^2 + \sum_{i=1}^{p} c_i^2(\lambda_i - \theta_i)^2. \]

Setting the derivative of this quantity with respect to \( k \),

\[ -2 \sum_{i=1}^{p} c_i \lambda_i^2 + 2k \sum_{i=1}^{p} c_i \lambda_i^2 + \sum_{i=1}^{p} c_i^2(\lambda_i - \theta_i)^2, \]

equal to 0, we get

\[ k = \sum_{i=1}^{p} c_i \lambda_i^2 / \sum_{i=1}^{p} c_i^2 \lambda_i^2 \frac{\sum_{i=1}^{p} c_i^2(\lambda_i - \theta_i)^2}{\sum_{i=1}^{p} c_i^2(\lambda_i - \theta_i)^2}. \]

It is easy to see that such a \( k \) minimizes the MSE given in (5.4)

but in order to use the resulting estimator, data-based substitutions

must be made for the unknown \( \lambda_i \)'s in \( k \).

Since \( X_i \) is unbiased for \( \lambda_i \), it is a reasonable substitute.

There exists, however, a substitute for \( \lambda_i \) that attains smaller MSE

than \( X_i \) does.
If we let \( d \) be some constant and note that

\[
\text{MSE}(dX^i) = d^2 \lambda^2 + (d\lambda_i - \lambda_i)^2 = d^2 \lambda^2 + (d - 1)^2 \lambda_i^2,
\]

straightforward differentiation leads to a minimizing value of \( d = 1/2 \).

This means \( \frac{1}{2}X_i \) is another choice to consider as a substitute for \( \lambda_i \).

Such substitutions are used when the estimators are evaluated with the Monte Carlo analysis at the end of the chapter.

As the final introductory idea with the multivariate exponential distribution, let us consider a Thompson-type estimator of \( \lambda \) by finding the \( c^i \)'s that minimize

\[
\sum_{i=1}^{p} \text{MSE[}\theta_i + (1 - c_i)(X_i - \theta_i)]
\]

where \( \theta_i \) is a prechosen constant believed close to \( \lambda_i \). The above sum equals

\[
\sum_{i=1}^{p} [(1 - c_i)^2 \lambda_i^2 + c_i^2(\lambda_i - \theta_i)^2].
\]

Equating the derivative of this sum with respect to \( c_i \) to 0, we find

\[
c_i = \frac{\lambda_i^2}{\lambda_i^2 + (\lambda_i - \theta_i)^2}
\].
and a subsequent component estimator that would be the Thompson shrinkage estimator for a gamma scale parameter.

But, if we note that \( E(X_i^2) = 2\lambda_i^2 \) and \( E(X_i - \theta_i)^2 = \lambda_i^2 + (\lambda_i - \theta_i)^2 \), a different Thompson-type shrinkage estimator is given by

\[
\theta_i + (1 - \frac{X_i^2}{2(X_i - \theta_i)^2})(X_i - \theta_i).
\]

With this work on the exponential case, we have pertinent questions to pose regarding the shrinkage estimators for the multivariate gamma parameters:

1. Should \( \lambda_i \) be estimated in shrinking factors by unbiased or minimum MSE estimators?
2. Should the shrinkers shrink from an unbiased estimator or a MSE minimizing estimator?
3. Should \( \lambda_i^2 \) be estimated by \( \frac{X_i^2}{2} \) or \( \frac{X_i^2}{2} \)?
4. Should our Thompson-type estimators employ a common shrinking factor for all components?
5. Should the parameters of the prior be assumed known?

5.2 Bayes and Albert-Type Estimators

For the gamma setting, let the density of \( X_i, i = 1, 2, \ldots, p \), be
for $\gamma, \lambda_i, x_i > 0$, where

$$E(X_i) = \gamma \lambda_i, E(X_i^2) = \gamma(\gamma + 1)\lambda_i^2$$

and

$$V(X_i) = \gamma \lambda_i^2. \quad (5.5)$$

Let the prior distribution of $\lambda_i$ be that in (5.2) with moments described in (5.3).

The posterior distribution of $\lambda_i$ given $X_i = x_i$ is then

$$g(\lambda_i | X_i = x_i, \alpha_i, \beta_i) = \frac{\frac{\beta_i}{\Gamma(\gamma)\Gamma(\beta_i)} \frac{1}{\lambda_i^{\beta_i+\gamma+1}} \frac{-(\alpha_i+x_i)}{\lambda_i} x_i^{\gamma-1}}{\int_0^\infty \frac{\alpha_i}{\Gamma(\gamma)\Gamma(\beta_i)} \frac{1}{\lambda_i^{\beta_i+\gamma+1}} e^{-\frac{\alpha_i+x_i}{\lambda_i}} x_i^{\gamma-1} d\lambda_i}.$$
The Bayes estimate is then

\[ E[\lambda_i | x_i = x_i] = \frac{1}{\Gamma(\beta_i + \gamma)} \int_0^\infty \frac{(x_i + \alpha_i)^{\beta_i + \gamma}}{\lambda_i^{\beta_i + \gamma + 1}} e^{-\frac{(x_i + \alpha_i)}{\lambda_i}} d\lambda_i \]
\[ \frac{(x_i + a_i)^{\beta_i + \gamma}}{\Gamma(\beta_i + \gamma)} \int_0^\infty \frac{1}{\lambda_i} e^{-\frac{x_i + a_i}{\lambda_i}} \, d\lambda_i = \frac{(x_i + a_i)^{\beta_i + \gamma}}{\Gamma(\beta_i + \gamma)} \int_0^\infty -y_i \left( \frac{1}{x_i + a_i} \right)^{\beta_i + \gamma} \left( \frac{x_i + a_i}{y_i} \right)^{\beta_i + \gamma} \, dy_i \]

\[ = \frac{(x_i + a_i)^{\beta_i + \gamma}}{\Gamma(\beta_i + \gamma)} \Gamma(\beta_i + \gamma - 1) = \frac{x_i + a_i}{\beta_i + \gamma - 1}. \]

Notationally, we will call the Bayes estimator, \( \frac{x_i + a_i}{\beta_i + \gamma - 1}, \lambda_i \beta \).

The

\[ \text{MSE}(\lambda_i \beta) = \text{Var}(\lambda_i \beta) + (\text{bias}(\lambda_i \beta))^2 \]

\[ = \frac{\gamma \lambda_i^2}{(\beta_i + \gamma - 1)^2} + \frac{\gamma \lambda_i + a_i}{(\beta_i + \gamma - 1) - \lambda_i} \]

\[ = \frac{1}{(\beta_i + \gamma - 1)^2} \left[ \gamma \lambda_i^2 + (\alpha_i - (\beta_i - 1)\lambda_i)^2 \right]. \]
This Bayes estimator is expressible as a shrinkage estimator, using (5.1), with shrinking factor

\[ c_i = \frac{X_i/\gamma - \frac{X_i + \alpha_i}{\beta_i + \gamma - 1}}{X_i/\gamma - \frac{\alpha_i}{\beta_i - 1}} = \frac{\beta_i - 1}{\beta_i + \gamma - 1}. \]

So the estimator is

\[ \hat{\lambda}_{i\beta} = \frac{\alpha_i}{\beta_i - 1} + (1 - c_i)(X_i/\gamma - \frac{\alpha_i}{\beta_i - 1}) \]

where \( 0 < c_i < 1 \).

Extending Albert's (1981) work to the general gamma setting, we want to find a second constant \( k \) so that the Bayes estimator can be altered to

\[ \hat{\lambda}_{iA} = \theta_i + (1 - c_i k)(X_i/\gamma - \theta_i), \]

where this component estimator of \( \lambda_i \) minimizes \( \sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iA}) \) with

\[ \theta_i = \frac{\alpha_i}{\beta_i - 1}. \]

With
\[ \sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iA}) = \sum_{i=1}^{p} \left[ (1 - c_i k)^2 \frac{\lambda_i^2}{Y} + ((\theta_i - \lambda_i) + (1 - c_i k)(\lambda_i - \theta_i))^2 \right] \]
\[ = \sum_{i=1}^{p} \left[ (1 - c_i k)^2 \frac{\lambda_i^2}{Y} + k^2 c_i^2 (\lambda_i - \theta_i)^2 \right] \]

and setting
\[ \frac{\partial}{\partial k} \sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iA}) = 0, \]

we find
\[ \sum_{i=1}^{p} \left[ -2(1 - c_i k)c_i \frac{\lambda_i}{Y} + 2kc_i^2 (\lambda_i - \theta_i)^2 \right] = 0 \]

making
\[ k = \frac{\sum_{i=1}^{p} \frac{c_i^2 \lambda_i^2}{Y} + \sum_{i=1}^{p} c_i^2 (\lambda_i - \theta_i)^2}{\sum_{i=1}^{p} c_i \lambda_i} \]

As we saw earlier, estimators for \( \lambda_i \) must be substituted for \( \lambda_i \) in \( k \) and once again the unbiased estimator, \( \frac{X_i}{Y} \), does not minimize
the MSE.

Here,

\[ \text{MSE}(d_{X_i}) = d^2 \gamma \lambda_i^2 + (d \gamma \lambda_i - \lambda_i)^2 \]

\[ = d^2 \gamma \lambda_i^2 + \lambda_i^2 (d \gamma - 1)^2 \]

and

\[ \frac{\partial}{\partial d} \text{MSE}(d_{X_i}) = 0 \]

which gives the MSE minimizing estimator \( \frac{X_i}{\gamma + 1} \). Note that \( \text{MSE}(\frac{X_i}{\gamma}) = \lambda_i^2 \)

and \( \text{MSE}(\frac{X_i}{\gamma + 1}) = \frac{\lambda_i^2}{\gamma + 1} \).

For the estimators proposed in Section 5.3, both substitutions for \( \lambda_i \) are utilized.

5.3 Thompson-Type Estimators

In this section, Thompson-type shrinkage estimators of \( \lambda \) are studied. Six such estimators will shrink toward an arbitrary guess of \( \lambda \) and the remaining to a mean

\[ \bar{X} = \frac{\sum X_i / \ell}{p} \]

where \( \ell \) equals \( \gamma \) or \( \gamma + 1 \). Within each category, we will consider both \( \frac{X_i}{\gamma} \) and \( \frac{X_i}{\gamma + 1} \), as the "usual" estimators. Finally, five of the
estimators will employ a common shrinking factor, while the rest allow the factors to change from component to component.

For easy reference, Table 7 lists the notation used for shrinkage estimators considered in this section.

<table>
<thead>
<tr>
<th>Shrunken from</th>
<th>( \frac{X_i}{\gamma} ) for ( \ell = \gamma )</th>
<th>( \frac{X_i}{\gamma + 1} ) for ( \ell = \gamma + 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shrunken to</td>
<td>( c_i = c )</td>
<td>( c_i = c )</td>
</tr>
<tr>
<td>( \theta_i )</td>
<td>( \hat{\lambda}_{i\theta} )</td>
<td>( \hat{\lambda}_{i\theta_i} )</td>
</tr>
<tr>
<td>( \bar{x}/\ell )</td>
<td>( \hat{\lambda}_{i\bar{x}} )</td>
<td>( \hat{\lambda}_{i\bar{x_i}} )</td>
</tr>
</tbody>
</table>

As in Chapter 2, the first subscript indicates an estimator of the \( i \)th component and the second indicates what the estimator is shrinking toward. If the third subscript is \( i \), it means the components have different shrinking factors while the absence of an \( i \) means a common shrinking factor is employed. Estimators with the \( (\gamma + 1) \) subscript shrink from \( \frac{X_i}{\gamma + 1} \), the MSE minimizing estimator.

Beginning with the case of shrinking to \( \theta_i \), we shall consider the estimator
\[ \lambda_i \theta_i = \theta_i + (1 - c_i)(\frac{x_i}{\gamma} - \theta_i). \]

Hence,

\[
\sum_{i=1}^{p} \text{MSE}(\lambda_i \theta_i) = \sum_{i=1}^{p} \left[ (1 - c_i)^2 \frac{\lambda_i^2}{\gamma} + (1 - c_i)(\lambda_i - \theta_i) + \theta_i - \lambda_i \right]^2.
\]

\[
= \sum_{i=1}^{p} (1 - c_i)^2 \frac{\lambda_i^2}{\gamma} + \sum_{i=1}^{p} c_i^2(\lambda_i - \theta_i)^2.
\]

The above MSE is minimized at

\[ c_i = \frac{\lambda_i^2}{\gamma(\lambda_i - \theta_i)^2}. \]

Estimation of \( \lambda_i \) by \( \frac{x_i}{\gamma} \) leads to the shrinkage estimator

\[ \theta_i + (1 - \frac{x_i^2}{x_i^2 + \gamma(x_i - \gamma \theta_i)^2})(\frac{x_i}{\gamma} - \theta_i). \]

while estimation of \( \lambda_i \) by \( \frac{x_i}{\gamma + 1} \) leads to the shrinkage estimator

\[ \theta_i + (1 - \frac{x_i^2}{x_i^2 + \gamma(x_i - (\gamma + 1) \theta_i)^2})(\frac{x_i}{\gamma} - \theta_i). \]
Yet another alternative shrinkage estimator is constructed as follows. Observe that

\[ E(X_i^2) = \gamma (\gamma + 1) \lambda_i^2 \]

while

\[ E(X_i - \gamma \theta_i)^2 = \gamma \lambda_i^2 + \gamma^2 (\lambda_i - \theta_i)^2. \]

Accordingly, the \( c_i \) given in (5.6) can be estimated as

\[
\frac{X_i^2 \gamma^{-1} (\gamma + 1)^{-1}}{\gamma^{-1} (X_i - \gamma \theta_i)^2} = \frac{X_i^2}{(\gamma + 1)(X_i - \gamma \theta_i)^2}
\]

giving

\[
\hat{\lambda}_{i \theta_i} = \theta_i + (1 - \frac{X_i^2}{(\gamma + 1)(X_i - \gamma \theta_i)^2}) \frac{X_i}{\gamma} \theta_i,
\]

(5.7)

The estimator in (5.7) will be given an interesting empirical Bayes interpretation in Section 5.4 of this chapter.

Finding \( \hat{\lambda}_{i \theta} \) involves the same steps used in finding \( \hat{\lambda}_{i \theta_i} \), but with a common shrinking factor. We get
\[
\frac{\partial}{\partial \gamma} \sum_{i=1}^{p} \text{MSE}(\lambda_i \theta) = 2 \sum_{i=1}^{p} (1 - \gamma) (\lambda_i^2) + 2 \sum_{i=1}^{p} \gamma \sum_{i=1}^{p} c(\lambda_i - \theta_i)^2 = 0
\]

implying

\[
c = \frac{\sum_{i=1}^{p} \lambda_i^2}{\sum_{i=1}^{p} \lambda_i^2 + \gamma \sum_{i=1}^{p} (\lambda_i - \theta_i)^2}.
\]

Estimation of \( \lambda_i \) by \( \frac{X_i}{\gamma} \) leads to the shrinkage estimator

\[
\theta_i + (1 - \frac{\sum_{i=1}^{p} X_i^2}{\sum_{i=1}^{p} X_i^2 + \gamma \sum_{i=1}^{p} (X_i - \gamma \theta_i)^2})(\frac{X_i}{\gamma} - \theta_i)
\]

while estimation of \( \lambda_i \) by \( \frac{X_i}{\gamma + 1} \) leads to the shrinkage estimator

\[
\theta_i + (1 - \frac{\sum_{i=1}^{p} X_i^2}{\sum_{i=1}^{p} X_i^2 + \gamma \sum_{i=1}^{p} (X_i - (\gamma + 1) \theta_i)^2})(\frac{X_i}{\gamma} - \theta_i)
\]

Another alternative shrinkage estimator is constructed as follows.

Observe that

\[
\mathbb{E}(\sum_{i=1}^{p} X_i^2) = \gamma(\gamma + 1) \sum_{i=1}^{p} \lambda_i^2
\]
while

$$E\left[\sum_{i=1}^{p} (X_i - \gamma \theta_i)^2\right] = \gamma \sum_{i=1}^{p} \lambda_i^2 + \gamma^2 \sum_{i=1}^{p} (\lambda_i - \theta_i)^2.$$  

The factor $c$ in (5.8) can then be estimated as

$$\frac{\sum_{i=1}^{p} X_i^2 \gamma^{-1}(\gamma + 1)^{-1}}{\gamma^{-1} \sum_{i=1}^{p} (X_i - \gamma \theta_i)^2} = \frac{\sum_{i=1}^{p} X_i^2}{(\gamma + 1) \sum_{i=1}^{p} (X_i - \gamma \theta_i)^2}.$$  

The resulting shrinkage estimator,

$$\theta_i + (1 - \frac{\sum_{i=1}^{p} X_i^2}{(\gamma + 1) \sum_{i=1}^{p} (X_i - \gamma \theta_i)^2}) \frac{X_i}{\gamma} - \theta_i),$$  

(5.9)

is shown to have an empirical Bayes interpretation in Section 5.4.

Shrinking from $\frac{X_i}{\gamma + 1}$, we get

$$\sum_{i=1}^{p} \text{MSE}(\lambda_i \theta_i(\gamma + 1)) = \sum_{i=1}^{p} \text{MSE}[\theta_i (1 - c_i)(\frac{X_i}{\gamma + 1} - \theta_i)]$$

$$= \sum_{i=1}^{p} \left[(1 - c_i)^2 \frac{\gamma \lambda_i}{(\gamma + 1)} + \left(1 - c_i \right) \frac{\gamma \lambda_i}{(\gamma + 1)} \right]$$
\[ - \theta_i + \theta_i - \lambda_i^2 \]

and

\[ \frac{\partial}{\partial c_i} \sum_{i=1}^{P} \text{MSE}(\hat{\lambda}_i \theta_i(\gamma+1)) = -2(1 - c_i) \frac{\gamma \lambda_i^2}{(\gamma + 1)^2} + 2(1 - c_i)(\frac{\gamma \lambda_i}{\gamma + 1} - \theta_i) \]

\[ + \theta_i - \lambda_i \left[ \theta_i - \frac{\gamma \lambda_i}{\gamma + 1} \right] = 0 \]

implies

\[ 1 - c_i = \frac{(\lambda_i - \theta_i)(\frac{\gamma \lambda_i}{\gamma + 1} - \theta_i)}{\gamma \lambda_i^2 + (\frac{\gamma \lambda_i}{\gamma + 1} - \theta_i)^2} \cdot \]

Estimating \( \lambda_i \) by \( X_i / \gamma \) leads to the shrinkage estimator

\[ \hat{\theta}_i + (1 - \frac{X_i}{\gamma} - \theta_i)(\frac{X_i}{\gamma + 1} - \theta_i) \]

\[ \frac{X_i}{\gamma(\gamma + 1)^2} + (\frac{X_i}{\gamma + 1} - \theta_i)^2 \]

In a very similar manner as for finding \( \hat{\lambda}_i \theta_i(\gamma+1) \), we get the shrinking factor
\[1 - c = \frac{\sum_{i=1}^{p} (\lambda_i - \theta_i)(\frac{\gamma \lambda_i}{\gamma + 1} - \theta_i)}{\sum_{i=1}^{p} \frac{\gamma \lambda_i^2}{(\gamma + 1)^2} + \sum_{i=1}^{p} (\frac{\gamma \lambda_i}{\lambda_i + 1} - \theta_i)^2}\]

and the shrinkage estimator

\[\theta_i + (1 - \frac{\sum_{i=1}^{p} \frac{X_i}{\gamma} (\frac{X_i}{\gamma + 1} - \theta_i)}{\sum_{i=1}^{p} \frac{X_i^2}{\gamma(\gamma + 1)^2} + \sum_{i=1}^{p} (\frac{X_i}{\gamma + 1} - \theta_i)^2})(\frac{X_i}{\gamma + 1} - \theta_i)\]

when substituting \(\frac{X_i}{\gamma}\) for \(\lambda_i\) in the shrinking factor.

In each of the last two shrinkage estimators, \(\frac{X_i}{\gamma + 1}\) could have been substituted for \(\lambda_i\) as well.

Concentrating now on shrinking to \(\frac{X}{\lambda}\) for \(\ell = \gamma, \gamma + 1\), we see

\[\sum_{i=1}^{p} \text{MSE}(\lambda_i, \frac{X}{\lambda}) = \sum_{i=1}^{p} \text{MSE}(\frac{X}{\gamma} + (1 - c_i)(\frac{X_i}{\gamma} - \frac{X}{\gamma}))\]

\[= \sum_{i=1}^{p} \left[(1 - c_i)^2 \frac{\lambda_i}{\gamma} + \frac{c_i^2}{\gamma} \sum_{j=1}^{p} \lambda_j^2 + 2c_i(1 - c_i) \text{Cov}(\frac{X_i}{\gamma}, \frac{X}{\gamma}) \right]\]

\[+ \left[(1 - c_i)\lambda_i + c_i \frac{\sum_{j=1}^{p} \lambda_j}{p} - \lambda_i \right]^2\]
\[
\begin{align*}
&= \sum_{i=1}^{p} (1 - c_i)^2 \frac{\lambda_i^2}{\gamma} + \sum_{i=1}^{p} \frac{c_i^2}{\gamma p^2} \left( \sum_{j=1}^{p} \lambda_j^2 \right) \\
&+ \frac{2 \gamma}{\gamma p} \sum_{i=1}^{p} \lambda_i^2 c_i (1 - c_i) + \frac{p}{\sum_{i=1}^{p} c_i^2} (\lambda_i - \frac{i=1}{p})^2.
\end{align*}
\]

Now
\[
\frac{\partial}{\partial c_i} \sum_{i=1}^{p} \text{MSE}(\lambda_i|x_i) = 2(1 - c_i) \left( \frac{-\lambda_i^2}{\gamma} \right) + \frac{2c_i}{\gamma p} \sum_{i=1}^{p} \lambda_i^2 \\
+ \frac{2}{\gamma p} \lambda_i^2 (1 - 2c_i) + 2c_i (\lambda_i - \frac{i=1}{p})^2 = 0
\]

implies
\[
c_i = \frac{\lambda_i^2 (1 - \frac{1}{p})}{\lambda_i^2 (1 - \frac{2}{p}) + \sum_{i=1}^{p} \left( \frac{\lambda_i^2}{p} \right) + \gamma (\lambda_i - \frac{i=1}{p})^2}
\]

for \( p > 2 \). Substituting \( \frac{X_i}{\gamma} \) for \( \lambda_i \) in the shrinking factor the shrinkage estimator becomes
\[
\frac{\bar{X}}{\gamma} + \left( 1 - \frac{X_i^2 (1 - \frac{1}{p})}{\sum_{i=1}^{p} X_i^2} \right) \left( \frac{1}{\gamma} - \frac{X}{\gamma} \right).
\]

\[
\frac{X_i^2 (1 - \frac{2}{p}) + \sum_{i=1}^{p} \left( \frac{X_i^2}{p^2} \right) + \gamma (X_i - \frac{i=1}{p})^2}{X_i (1 - \frac{2}{p}) + \sum_{i=1}^{p} \left( \frac{X_i}{p^2} \right) + \gamma (X_i - \frac{i=1}{p})^2}
\]
For a similar estimator where $c_i = c$ for $i = 1, 2, \ldots, p$, we get

\[
\begin{align*}
\frac{\partial}{\partial c} \text{MSE}(\hat{\lambda}_i|X) &= \frac{-2(1-c)}{Y} \sum_{i=1}^{p} \frac{\lambda_i^2}{\gamma_p} + \frac{2c}{Y} \sum_{i=1}^{p} \frac{\lambda_i^2}{\gamma_p^2} + \frac{2c(1-c)}{Y} \sum_{i=1}^{p} \gamma_{\lambda_i^2} + \\
&+ 2c \sum_{i=1}^{p} \left( \lambda_i - \frac{i-1}{p} \right)^2 = 0
\end{align*}
\]

implies
where

\[ s^2 = \frac{1}{p-1} \sum_{i=1}^{p} (\lambda_i - \bar{\lambda})^2 \]

and

\[ \bar{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p} \]

Substituting \( \frac{X_i}{\gamma} \) for this shrinking factor, \( \hat{\lambda}_1 \) becomes

\[ \frac{\bar{X}}{\gamma} + (1 - \frac{1}{p} \sum_{i=1}^{p} \frac{X_i^2}{\gamma} + \frac{\gamma}{p-1} \sum_{i=1}^{p} (X_i - \bar{X})^2 \) \left( \frac{X_i}{\gamma} - \frac{\bar{X}}{\gamma} \right). \]

Note that the pre-substitution form of the shrinking factor is structurely the same as that for \( \hat{\lambda}_1 \) in Chapter 4 and \( \delta_{1\bar{X}} \) in Chapter 2.

Also note that \( \frac{X_i}{\gamma + 1} \) could have been substituted for \( \lambda_i \) in the shrinking factors of the previous two shrinkage estimators.
We will not specifically derive the estimators \( \hat{\lambda}_{i \mathbf{x}}(Y + 1) \) and \( \hat{\lambda}_{i \mathbf{x}}(Y^+1) \), because their shrinking factors are the same as those for \( \hat{\lambda}_{i \mathbf{x}} \) and \( \hat{\lambda}_{i \mathbf{x}} \) and because their MSEs are those of \( \hat{\lambda}_{i \mathbf{x}} \) and \( \hat{\lambda}_{i \mathbf{x}} \) multiplied by \( \left( \frac{\gamma}{\gamma + 1} \right)^2 \). The relationship between considering \( \frac{X}{Y} \) and \( \frac{X}{Y + 1} \) can be seen when noting

\[
\hat{\lambda}_{i \mathbf{x}}(Y + 1) = \frac{X}{\gamma + 1} + (1 - c_i) \left( \frac{X_i}{\gamma + 1} - \frac{X}{\gamma + 1} \right)
\]

\[
= \frac{\gamma}{\gamma + 1} \left[ \frac{X}{\gamma} + (1 - c_i) \left( \frac{X_i}{\gamma} - \frac{X}{\gamma} \right) \right]
\]

\[
= \frac{\gamma}{\gamma + 1} \hat{\lambda}_{i \mathbf{x}}.
\]

Finally, it should be noted that the estimation of the mean vector \( Y \lambda = \gamma \lambda \) could have been considered in place of \( \lambda \). In this case, the shrinking factors would be the same as those we have found, but

\[
\sum_{i=1}^{p} \text{MSE(estimator of } Y \lambda_i \text{)} = \sum_{i=1}^{p} \text{MSE(} Y(\text{estimator of } \lambda_i)\text{)}
\]

\[
= \gamma^2 \sum_{i=1}^{p} \text{MSE(estimator of } \lambda_i)\text{).}
\]
5.4 The Empirical Bayes Nature of the Thompson-Type Estimators

As we now move our attention to Bayesian estimators, we would like to show a relationship between the empirical Bayes estimator and Thompson-type estimators. This was done for the normal case in Chapter 2. Given that the constants $\theta_i$, $i = 1, 2, \ldots, p$, take the place of the prior means in the Bayesian framework, we only need to show an equivalence between the shrinking factors.

To prove this equivalence, a major tool is the usage of the marginal distribution of $X_i$, which is now derived. Noting that

$$f(x | \lambda_i) = \frac{1}{\prod_{i=1}^{p} \lambda_i} \frac{1}{\Gamma(\gamma)} x_i^{\gamma-1} e^{-x_i/\lambda_i}$$

for $\gamma \geq 1$, $\lambda_i$ and $x_i > 0$ and

$$f(x) = \int_0^{\infty} \int_0^{\infty} \ldots \int_0^{\infty} f(x, \lambda_1) \, d\lambda_1 \, d\lambda_2 \ldots \, d\lambda_p$$

$$= \prod_{i=1}^{p} \int_0^{\infty} f(x_i, \lambda_i) \, d\lambda_i$$

$$= \prod_{i=1}^{p} \frac{x_i^{\gamma-1} \beta_i}{\Gamma(\gamma) \Gamma(\beta_i)} \int_0^{\infty} \frac{1}{\lambda_i} e^{-x_i/\lambda_i} d\lambda_i$$

$$= \prod_{i=1}^{p} \frac{x_i^{\gamma-1} \beta_i}{\Gamma(\gamma) \Gamma(\beta_i)} \int_0^{\infty} \frac{y_i}{x_i + \alpha_i} e^{-(x_i+\alpha_i)/y_i} \frac{y_i}{2} dy_i$$
The marginal distribution is a multivariate type II Beta distribution with independence among the components as seen by the factoring in the joint distribution.

To find the marginal moments of $X_i$, we first find $E(X_i + \alpha_i)$ and $E(X_i + \alpha_i)^2$.
\[ E(x_i + \alpha_i) = \int_0^\infty \frac{(\beta_i + \gamma - 1)!}{(\gamma - 1)!(\beta_i - 1)!} \frac{x_i^{\gamma-1}}{x_i} \frac{\alpha_i}{x_i + \alpha_i}^{\beta_i+\gamma-1} \, dx_i \]

\[ = \int_0^\infty \frac{(\beta_i + \gamma - 1)!}{(\gamma - 1)!(\beta_i - 1)!} \frac{x_i^{\gamma-1}}{x_i} \frac{\alpha_i}{x_i + \alpha_i}^{\beta_i+\gamma-1} \, dx_i \]

\[ = \int_0^{\gamma-1} \frac{(\beta_i + \gamma - 1)!}{(\gamma - 1)!(\beta_i - 1)!} \left( \frac{\alpha_i(1 - y_i)}{\alpha_i y_i} \right)^{\gamma-1} \frac{\beta_i+\gamma-1}{y_i} \, \frac{1}{y_i^2} \, dy_i \]

where

\[ y_i = \frac{\alpha_i}{x_i + \alpha_i} \]

and

\[ dx_i = \frac{-\alpha_i}{y_i^2} \, dy_i \]

\[ = \alpha_i \int_0^{(\gamma + \beta_i - 1)!} \frac{(\gamma - 1)!}{(\gamma - 1)!(\beta_i - 1)!} (1 - y_i)^{\gamma-1} \frac{\beta_i-2}{y_i} \, dy_i \]

\[ = \alpha_i \int_0^{\gamma + \beta_i - 1} \frac{\Gamma(\gamma + \beta_i - 1)}{\Gamma(\gamma)\Gamma(\beta_i)} \frac{(\beta_i-1)-1}{y_i} (1 - y_i)^{\gamma-1} \, dy_i \]

\[ = \alpha_i \frac{\Gamma(\gamma + \beta_i - 1)}{\Gamma(\beta_i)} \frac{\Gamma(\beta_i - 1)}{\Gamma(\gamma)\Gamma(\beta_i - 1)} \int_0^{\gamma + \beta_i - 1} \frac{(\beta_i-1)-1}{y_i} (1 - y_i)^{\gamma-1} \, dy_i \]
since a beta-density was formed which integrates to 1.

In similar fashion, we find

\[
E(x_1 + \alpha_i)^2 = \int_0^\infty \frac{(\beta_i + \gamma - 1)!}{(\gamma - 1)! (\beta_i - 1)!} \frac{x_1^\gamma}{x_i} \left( \frac{\alpha_i}{x_1 + \alpha_i} \right)^{\beta_i + \gamma - 2} a_i^2 \, dx_i
\]

\[
= \alpha_i^2 \int_0^1 \frac{(\beta_i + \gamma - 1)!}{(\gamma - 1)! (\beta_i - 1)!} \frac{1 - y_i}{y_i} \frac{\alpha_i^2 y_i}{y_i} \frac{1}{y_i^{\beta_i + \gamma - 2}} \, dy_i
\]

\[
= \alpha_i^2 \int_0^1 \frac{(\beta_i + \gamma - 1)!}{(\gamma - 1)! (\beta_i - 1)!} \frac{\beta_i - 3}{y_i} (1 - y_i)^{\gamma - 1} \, dy_i
\]

\[
= \frac{\alpha_i^2 (\beta_i + \beta_i + \gamma - 3)!}{(\beta_i - 1)! (\beta_i + \gamma - 3)! (\beta_i + \gamma - 3)!}
\]

\[
= \alpha_i^2 \frac{(\beta_i + \gamma - 1)(\beta_i + \gamma - 2)}{(\beta_i - 1)(\beta_i - 2)}.
\]

Now

\[
E(x_1) = \frac{\alpha_i (\beta_i + \gamma - 1)}{\beta_i - 1} = \alpha_i = \frac{\alpha_i \gamma}{\beta_i - 1}
\]

and
\[ E(X_1^2) = -\alpha_i^2 - 2\alpha_i \ E(X_1) + \alpha_i^2 \frac{(\beta_i + \gamma - 1)(\beta_i + \gamma - 2)}{(\beta_i - 1)(\beta_i - 2)} \]

\[ = \alpha_i^2 \left[ -1 - 2 \frac{\gamma}{\beta_i - 1} + \frac{(\beta_i + \gamma - 1)(\beta_i + \gamma - 2)}{(\beta_i - 1)(\beta_i - 2)} \right] \]

\[ = \frac{\alpha_i^2 \gamma(\gamma + 1)}{(\beta_i - 1)(\beta_i - 2)} \]

and

\[ V(X_1) = E(X_1^2) - (E(X_1))^2 = \frac{\alpha_i^2 \gamma(\gamma + 1)}{(\beta_i - 1)(\beta_i - 2)} - \frac{\alpha_i^2 \gamma^2}{(\beta_i - 1)^2} \]

\[ = \frac{\alpha_i^2 \gamma(\beta_i + \gamma - 1)}{(\beta_i - 1)^2(\beta_i - 2)} \cdot \]

We are now able to estimate \( \alpha_i^2 \) from the second moment of \( X_i \) through its marginal distribution as

\[ X_i^2(\beta_i - 1)(\beta_i - 2) \]

\[ \frac{\gamma(\gamma + 1)}{\gamma(\gamma + 1)} \cdot \]

Starting with the shrinking factor of the Bayesian estimator,
\[
\frac{\beta_i - 1}{\beta_i + \gamma - 1} = \frac{\alpha_i^2}{\gamma (\beta_i - 1)(\beta_i - 2)} + \frac{\alpha_i^2}{(\beta_i - 1)^2 (\beta_i - 2)}
\]

\[
= \frac{\alpha_i^2}{(\beta_i - 1)(\beta_i - 2)} + \frac{\gamma \alpha_i^2}{(\beta_i - 1)^2 (\beta_i - 2)}
\]

\[
= \frac{E(X_i^2)}{\gamma (\gamma + 1)} + \frac{E(X_i - \gamma \theta_i)^2}{\gamma}
\]

\[
= \frac{E(X_i^2)}{(\gamma + 1) E(X_i - \gamma \theta_i)^2}
\]

which is precisely the shrinking factor of the Thompson-type estimator in (5.7). This gives the estimator in (5.7) an empirical Bayes nature.

The final steps, above, result by taking the two-state expectations:
\[ E(X_i - \gamma \theta_i)^2 = \gamma E(\lambda_i^2) + \gamma^2 \nu(\lambda_i) \]

\[ = \frac{\gamma \alpha_i^2}{(\beta_i - 1)(\beta_i - 2)} + \gamma \frac{\alpha_i^2}{(\beta_i - 1)^2(\beta_i - 2)} \]

and

\[ E(x_i^2) = \gamma(\gamma + 1)E(\lambda_i^2) \]

\[ = \frac{\gamma(\gamma + 1)\alpha_i^2}{(\beta_i - 1)(\beta_i - 2)} \]

where \( \theta_i \) is the prior mean \( \frac{\alpha_i}{\beta_i - 1} \).

If we assumed \( \alpha_i = \alpha \) and \( \beta_i = \beta \) for \( i = 1, 2, \ldots, p \), the Bayesian shrinking factor would be \( \frac{\beta}{\beta + \gamma - 1} \), unchanging regardless of the \( \lambda_i \) being estimated. This notion agrees with that for the estimator \( \hat{\lambda}_{10} \) in (5.9). Again the Thompson-type estimator can be shown to have an empirical Bayes nature by considering the shrinking factor as follows:

\[ \frac{\beta - 1}{\beta + \gamma - 1} = \frac{\alpha^2/(\beta - 1)(\beta - 2)}{(\beta - 1)(\beta - 2) + \frac{\gamma \alpha^2}{(\beta - 1)^2(\beta - 2)}} \]
Again two-stage expectations are involved as before.

5.5 A Methods of Moments Empirical Bayes Estimator

Setting $\alpha_1 = \alpha$ and $\beta_1 = \beta$, for $i = 1, 2, ..., p$, and assuming $\alpha$ to be unknown, we can estimate the unknown prior parameter $\alpha$ from the marginal distribution attained in the previous section and develop an empirical form of the Bayes estimator $\hat{\lambda}_{1B}$. Since we are applying the method of moments technique to estimate $\alpha$, the resulting empirical Bayes estimator is called the methods of moments estimator with component part $\hat{\lambda}_{imm}$.

Since $E(X_i) = \frac{\alpha \gamma}{\beta - 1}$ from the marginal distribution of $X_i$, we find

$$E(\bar{X}) = E\left(\frac{1}{p} \sum_{i=1}^{p} X_i\right) = \frac{\alpha \gamma}{\beta - 1}$$
as well. Using $\frac{\beta - 1}{\gamma} \bar{X}$ to estimate $\alpha$ in $\hat{\lambda}_{imm}$, we see the estimator is utilizing information from all components in the estimation of any $\lambda_i$.

Defining the component estimator

$$\hat{\lambda}_{imm} = \frac{x_i + \hat{\alpha}}{\beta + \gamma - 1} = \frac{x_i + \frac{\beta - 1}{\gamma} \bar{X}}{\beta + \gamma - 1}$$

we find

$$E(\hat{\lambda}_{imm}) = \frac{\sum_{i=1}^{p} \lambda_i}{\beta + \gamma - 1}$$

and that $\hat{\lambda}_{imm}$ is unbiased for $\lambda_i$ only if

$$\lambda_i = \frac{\sum_{i=1}^{p} \lambda_i}{p}.$$

Since

$$\text{Var}(\hat{\lambda}_{imm}) = \text{Var}\left[\frac{x_i (1 + \frac{\beta - 1}{\gamma} \overline{X}) + \frac{\beta - 1}{\gamma} \sum_{j \neq i} x_j}{\beta + \gamma - 1}\right]$$

$$= \frac{(1 + \frac{\beta - 1}{\gamma})^2 \gamma \lambda_i^2 + (\frac{\beta - 1}{\gamma})^2 \sum_{j \neq i} \lambda_j^2}{(\beta + \gamma - 1)^2}$$
we find

\[
\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{imm}) = \sum_{i=1}^{p} \left[ \nu(\hat{\lambda}_{imm}) + \text{bias}(\hat{\lambda}_{imm}) \right]^2
\]

\[
= \gamma^2 \left( 1 + \frac{2(\beta - 1)}{\gamma p} \lambda_i^2 + \frac{2(\beta - 1)}{\gamma^2 p} \sum_{i=1}^{p} \lambda_i^2 \right) \lambda_i^2
\]

\[
= \gamma^2 \left( 1 + \frac{2(\beta - 1)}{\gamma p} \lambda_i^2 + \frac{2(\beta - 1)}{\gamma^2 p} \sum_{i=1}^{p} \lambda_i^2 \right) \lambda_i^2
\]

\[
= \frac{1}{(\beta + \gamma - 1)^2} \left[ \gamma^2 \left( 1 + \frac{2(\beta - 1)}{\gamma p} \lambda_i^2 + \frac{2(\beta - 1)}{\gamma^2 p} \sum_{i=1}^{p} \lambda_i^2 \right) \lambda_i^2 \right.
\]

\[
+ \frac{p}{(\beta + \gamma - 1)^2} \left[ \sum_{i=1}^{p} \lambda_i^2 \right] \left[ \sum_{i=1}^{p} \lambda_i^2 - \frac{\sum_{i=1}^{p} \lambda_i^2}{p} \right] \lambda_i^2
\]

\[
= \frac{p}{(\beta + \gamma - 1)^2} \left[ \gamma^2 \left( 1 + \frac{2(\beta - 1)}{\gamma p} \lambda_i^2 + \frac{2(\beta - 1)}{\gamma^2 p} \sum_{i=1}^{p} \lambda_i^2 \right) \lambda_i^2 \right.
\]

\[
+ \frac{p}{(\beta + \gamma - 1)^2} \left[ \sum_{i=1}^{p} \lambda_i^2 \right] \left[ \sum_{i=1}^{p} \lambda_i^2 - \frac{\sum_{i=1}^{p} \lambda_i^2}{p} \right] \lambda_i^2
\]

\[
= \frac{p}{(\beta + \gamma - 1)^2} \left[ \gamma^2 \left( 1 + \frac{2(\beta - 1)}{\gamma p} \lambda_i^2 + \frac{2(\beta - 1)}{\gamma^2 p} \sum_{i=1}^{p} \lambda_i^2 \right) \lambda_i^2 \right.
\]

\[
+ \frac{p}{(\beta + \gamma - 1)^2} \left[ \sum_{i=1}^{p} \lambda_i^2 \right] \left[ \sum_{i=1}^{p} \lambda_i^2 - \frac{\sum_{i=1}^{p} \lambda_i^2}{p} \right] \lambda_i^2
\]
where

\[ \lambda = \frac{\sum_{i=1}^{p} \lambda_i}{p} . \]

One way of evaluating the goodness of \( \lambda_{i_{\text{imm}}} \) versus \( \frac{X_i}{\gamma + 1} \) or 
\( \frac{X_i}{\gamma} \) is by using the "relative savings loss" of Efron and Morris (1973), 
which was discussed in Chapter 1. To evaluate the RSL, the second 
stage expectation must be found for \( \sum_{i=1}^{p} \text{MSE(}\text{estimator of } \lambda_i\text{)} \) using 
the moments of the prior distribution in (5.3):

\[
E[\text{MSE}(\lambda^A_{i_B})] = E[\text{MSE}(\frac{X_i + \alpha}{\beta + \gamma - 1})]
\]

\[
= \frac{1}{(\beta + \gamma - 1)^2} \left[ \gamma E(\lambda_i^2) + (\alpha^2 - 2\alpha(\beta - 1)E(\lambda_i) + 
+ (\beta - 1)^2 E(\lambda_i^2) \right]
\]

\[
= \frac{1}{(\beta + \gamma - 1)^2} \left[ \frac{\alpha^2}{(\beta - 1)(\beta - 2) + (\alpha^2 - 2\alpha(\beta - 1) \frac{\alpha}{(\beta - 1)}} \right.
\]

\[
+ (\beta - 1)^2 \frac{\alpha^2}{(\beta - 1)(\beta - 2)} \right] \]
\[
\frac{1}{(\beta + \gamma - 1)^2} \left[ \frac{\gamma^2}{(\beta - 1)(\beta - 2)} + \frac{\alpha^2}{\beta - 2} \right]
\]
\[
= \frac{\alpha^2}{(\beta + \gamma - 1)^2} \left[ \frac{\gamma}{\beta - 1} + 1 \right] \frac{1}{\beta - 2}
\]
\[
= \frac{\alpha^2}{(\beta + \gamma - 1)(\beta - 2)(\beta - 1)}
\]

so

\[
E[\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iB})] = \frac{p\alpha^2}{(\beta + \gamma - 1)(\beta - 2)(\beta - 1)},
\]

\[
E[\sum_{i=1}^{p} \text{MSE}(\frac{X_i}{Y})] = E[\sum_{i=1}^{p} \frac{\lambda_i^2}{Y}] = \frac{p\alpha^2}{\gamma(\beta - 1)(\beta - 2)},
\]

\[
E[\sum_{i=1}^{p} \text{MSE}(\frac{X_i}{Y + 1})] = E[\sum_{i=1}^{p} \frac{\lambda_i^2}{Y + 1}] = \frac{p\alpha^2}{(\gamma + 1)(\beta - 1)(\beta - 2)},
\]

and

\[
E[\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{i,\text{imm}})] = \frac{1}{\gamma^2} \left[ 1 + \frac{2(\beta - 1)}{\gamma p} + \frac{(\beta - 1)^2}{\gamma^2 p} \right]
\]

\[
= \frac{p}{\sum_{i=1}^{p} \frac{\alpha^2}{(\beta - 1)(\beta - 2)(\beta + \gamma - 1)^2}}
\]
\[
\begin{align*}
&+ \frac{(\beta - 1)^2}{(\beta + \gamma - 1)^2} \mathbb{E}\left[ \frac{p}{\sum_{i=1}^{p} (\lambda_i - p\lambda)^2} \right] \\
&= \frac{\alpha^2(p\gamma^2 + 2(\beta - 1)\gamma + (\beta - 1)^2)}{\gamma(\beta - 1)(\beta - 2)(\beta + \gamma - 1)^2} + \frac{(\beta - 1)^2}{(\beta + \gamma - 1)^2} \\
&= \frac{\alpha^2}{\gamma(\beta + \gamma - 1)^2(\beta - 1)(\beta - 2)} \left[ p\gamma^2 + 2\gamma(\beta - 1) + (\beta - 1)^2 \gamma(\beta - 1) \right] \\
&+ (\beta - 1)^2 + (p - 1)\gamma(\beta - 1) \\
&= \frac{\alpha^2}{\gamma(\beta + \gamma - 1)^2(\beta - 1)(\beta - 2)} \left[ p\gamma^2 + (p - 1) \gamma(\beta - 1) + (\beta - 1)^2 \right].
\end{align*}
\]

When we consider the "usual" estimator of \( \lambda_i \) to be \( \frac{X_i}{\gamma + 1} \),

\[
\text{RSL} = \frac{\text{Bayes risk usual} - \text{Bayes risk usual}}{\text{Bayes risk usual} - \text{Bayes risk usual}}
\]
When we consider the "usual" estimator of $\lambda_1$ to be $\frac{X_i}{\gamma}$,

$$RSL = \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)} \left[ p Y^2 + (p+1)Y(\beta - 1) + (\beta - 1)^2 \right] - \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)(\beta - 1)}$$

$$= \frac{\alpha^2}{(\beta + \gamma - 1)(\beta - 2)} - \frac{\alpha^2}{(\beta + \gamma - 1)(\beta - 2)(\beta - 1)}$$

$$= \frac{\alpha^2}{(\beta + \gamma - 1)(\beta - 2)} \left[ p Y^2 + (p+1)Y(\beta - 1) + (\beta - 1)^2 \right] - \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)(\beta - 1)}$$

$$= \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)} \left[ p Y^2 + (p + 1)(\beta - 1) + (\beta - 1)^2 \right] - \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)(\beta - 1)}$$

$$= \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)} \frac{p}{\gamma} \frac{\beta - 1}{\beta + \gamma - 1}$$

$$= \frac{\alpha^2}{\gamma(\beta + \gamma - 1)(\beta - 2)} \frac{p}{\gamma} \frac{\beta - 1}{p(\beta - 1)} = \frac{\beta - 1}{p} = \frac{1}{p} \cdot$$
These values for the RSL show that, in terms of Bayes risk, \( \hat{\lambda}_{im} \) nears the performance of \( \hat{\lambda}_{iB} \) as the number of components \( p \) increases. They also show that \( \hat{\lambda}_{im} \) is always an improvement over \( \frac{X_i}{Y} \) and often is an improvement over \( \frac{X_i}{Y + 1} \) as well.

To compare \( \hat{\lambda}_{iB} \) and \( \hat{\lambda}_{im} \) on the basis of MSE, it is clear that such a comparison will depend greatly on the choice of the prior parameters. But let us view the MSE(\( \hat{\lambda}_{iB} \)) as

\[
\frac{1}{(\beta_i + Y - 1)^2} \left[ \alpha_i^2 + (\beta_i - 1)^2 \left( \frac{\alpha_i}{\beta_i - 1} - \lambda_i \right)^2 \right].
\]

It is obvious that \( \hat{\lambda}_{iB} \) will perform best if both \( \beta_i \) is large and if \( \frac{\alpha_i}{\beta_i - 1} \) equals \( \lambda_i \). This means that the prior mean is perfectly chosen and the large \( \beta_i \) allows little variance of \( \lambda_i \) through the prior distribution.

Writing the methods of moments estimator as a shrinker, we obtain

\[
\hat{\lambda}_{im} = \frac{X}{Y} + (1 - \frac{\beta - 1}{\beta + Y - 1}) \left( \frac{X_i}{Y} - \frac{X}{Y} \right)
\]

and remembering \( \hat{\lambda}_{iB} \) as a shrinker gives

\[
\hat{\lambda}_{iB} = \frac{\alpha_i}{\beta_i - 1} + (1 - \frac{\beta_i - 1}{\beta_i + Y - 1}) \left( \frac{X_i}{Y} - \frac{\alpha_i}{\beta_i - 1} \right).
\]

Now we can look at the effect of allowing \( \beta_i \) to be large on
both $\hat{\lambda}_{iB}$ and $\hat{\lambda}_{iB}$. In fact, we will let $\beta_i$ go to $+\infty$ for all $i$. This forces the shrinking factors of $\hat{\lambda}_{iB}$ and $\hat{\lambda}_{iB}$ to go to 1 and the estimators become $\hat{\lambda}_{iB} = \frac{\bar{X}}{\gamma}$ and $\hat{\lambda}_{iB} = 0$. Now

$$\text{MSE}(\hat{\lambda}_{im}) = \frac{1}{p} \frac{\sum_{i=1}^{p} \lambda_i^2}{\gamma} + [\lambda_i - \bar{X}]^2$$

and

$$\text{MSE}(\hat{\lambda}_{iB}) = 0 + (0 - \lambda_i)^2 = \lambda_i^2.$$ 

When $\lambda_i$ is large, the Bayes estimator will do poorly. This is reasonable because, as $\beta_i \to +\infty$, the prior distribution gives a mean of 0 and is very uncompromising as its variance also becomes 0.

Since $\hat{\lambda}_{iB}$ still depends on the data in this case, one would expect $\hat{\lambda}_{iB}$ to outperform $\hat{\lambda}_{iB}$. This is verified by considering

$$\frac{\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iB})}{i=1} - \frac{\sum_{i=1}^{p} \text{MSE}(\hat{\lambda}_{iB})}{i=1} = \frac{\sum_{i=1}^{p} \lambda_i^2}{\gamma p} - \frac{\sum_{i=1}^{p} \lambda_i^2}{\gamma p}$$

$$= \frac{\sum_{i=1}^{p} \lambda_i^2}{\gamma p} - \frac{\sum_{i=1}^{p} \lambda_i^2}{\gamma p}$$

$$= \frac{\sum_{i=1}^{p} \lambda_i^2}{\gamma p} - \frac{1}{\gamma p}.$$
This shows that, as we allow $\beta_i$ to be very large for all $i = 1, 2, \ldots, p$, $\hat{\beta}_{\text{mm}}$ performs better than $\hat{\beta}_B$ where we drop the $i$ subscript to denote vector estimators.

One may wonder how these estimators perform in another special case, that being when the component $\lambda$'s are both large and equal. To study this, we assume $\beta_i = \beta$ and $\lambda_i = \lambda$ for $i = 1, 2, \ldots, p$. As $\lambda$ goes to infinity, we find

$$
\frac{\sum_{i=1}^{p} \text{MSE}_{\hat{\beta}_{\text{mm}}}^i}{\sum_{i=1}^{p} \text{MSE}_{\hat{\beta}_B}^i} = \frac{\gamma \beta + 2(\beta - 1) + (\beta - 1)^2}{\gamma \beta + (\beta - 1)^2 p}.
$$

This ratio exceeds 1 only when

$$
\beta < \frac{\gamma(p + 2) - 1}{\gamma p - 1}.
$$

Considering that a restriction on $\beta$ has it greater than 2, the ratio automatically less than 1 when $\gamma > \frac{1}{p - 2}$. When $p = 2$, $\beta < \frac{4\gamma - 1}{2\gamma - 1}$ ensures $\hat{\beta}_B$ performs better than $\hat{\beta}_{\text{mm}}$. Therefore, we see that the method of moments estimator usually improves upon the Bayes estimator for large and equal component $\lambda$'s. This is because of $\hat{\beta}_{\text{mm}}$'s extra dependency on the data.
5.6 The Monte Carlo Analysis

Evaluating most of the estimators of this chapter is very difficult considering we know little of their true distributional properties once the shrinking factors have unknown parameters substituted by data-based estimators. So again, as in Chapters 2 and 3, we must turn to a simulation study where 5000 random vectors are generated but from a multivariate gamma distribution now. From these random vectors, we are able to approximate the MSEs of all of our estimators. Many of the important relative qualities of these estimators can be seen with the case $p = 2$ and $\gamma = 4$, so we concentrate on this case alone.

Beginning with an analysis of the Bayesian-related component estimators, namely $\hat{\lambda}_{iB}$, $\hat{\lambda}_{imm}$ and $\hat{\lambda}_{iA}$, we wish to compare these with each other and with the generally accepted estimator $X_i / \gamma + 1$.

In Section 5.5 of this chapter, we discussed these estimators in the case of large $\beta_i$'s. As $\beta_i$ is increased, the $i$th prior variance decreases which shifts the prior distribution more towards its mean. If the true mean is nearly equal to the prior mean, this would be a pleasing result, but not otherwise.

With $\beta_1 = \beta_2 = \beta$, $\lambda_1 = \lambda_2 = \lambda$ and $\alpha_1 = \alpha_2 = \alpha$, we study the performance of the Bayesian-related estimators for large values of $\beta$ in Table 8. Notationally, we write the prior mean and variance as $E(\lambda)$ and $V(\lambda)$, respectively.
Table 8. The performance of Bayesian-related estimators for large values of $\beta$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\beta$</th>
<th>$\alpha$</th>
<th>$E(\lambda)$</th>
<th>$V(\lambda)$</th>
<th>$\sum_{i=1}^{2} \text{MSE}(\hat{\lambda}_{iB})$</th>
<th>$\sum_{i=1}^{2} \text{MSE}(\hat{\lambda}_{iMM})$</th>
<th>$\sum_{i=1}^{2} \text{MSE}(\hat{\lambda}_{iA})$</th>
<th>$\sum_{i=1}^{2} \text{MSE}(\frac{\lambda_i}{\gamma+1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.71</td>
<td>31</td>
<td>171.3</td>
<td>5.71</td>
<td>1.12</td>
<td>.28</td>
<td>8.26</td>
<td>5.00</td>
<td>13.04</td>
</tr>
<tr>
<td>9.71</td>
<td>31</td>
<td>171.3</td>
<td>5.71</td>
<td>1.12</td>
<td>25.57</td>
<td>23.90</td>
<td>25.99</td>
<td>37.71</td>
</tr>
<tr>
<td>5.71</td>
<td>101</td>
<td>571</td>
<td>5.71</td>
<td>.33</td>
<td>.02</td>
<td>8.16</td>
<td>5.00</td>
<td>13.04</td>
</tr>
<tr>
<td>9.71</td>
<td>101</td>
<td>571</td>
<td>5.71</td>
<td>.33</td>
<td>29.66</td>
<td>23.61</td>
<td>25.99</td>
<td>37.71</td>
</tr>
<tr>
<td>5.71</td>
<td>176</td>
<td>999</td>
<td>5.71</td>
<td>.18</td>
<td>.01</td>
<td>8.16</td>
<td>5.00</td>
<td>13.04</td>
</tr>
<tr>
<td>9.71</td>
<td>176</td>
<td>999</td>
<td>5.71</td>
<td>.18</td>
<td>30.63</td>
<td>23.58</td>
<td>25.99</td>
<td>37.71</td>
</tr>
</tbody>
</table>
When \( \lambda \) equals \( E(\lambda) \), Table 8 shows the Bayes estimator dominating \( \hat{\lambda}_{mm} \), which agrees with our previous discussion. When \( V(\lambda) \) is made smaller, the improvement is more pronounced. But when the prior mean is not near \( \lambda \), \( \hat{\lambda}_{mm} \) dominates both \( \hat{\lambda}_B \) and \( \hat{\lambda}_A \), two estimators shrinking to this prior mean.

We can view \( \hat{\lambda}_A \), the Alberts-type estimator, as being somewhat Bayesian and also utilizing information from all components as \( \hat{\lambda}_{mm} \) does. Because of this, it is not surprising to see it performing better than one of \( \hat{\lambda}_{mm} \) and \( \hat{\lambda}_B \), and worse than the other.

In all cases, the three estimators perform better than using the component estimator \( \frac{X_i}{1 + \gamma} \).

In Section 5.5, we also discussed \( \hat{\lambda}_{mm} \) and \( \hat{\lambda}_B \) when \( \lambda_1 \) and \( \lambda_2 \) are very large and \( \beta_1 = \beta_2 \). This discussion demonstrated the way the choice of prior parameters can dictate the performance of Bayesian-related estimators. Applying the inequality in (5.10) we should find that \( \hat{\lambda}_B \) outperforms \( \hat{\lambda}_{mm} \) when \( \beta < \frac{2\gamma - 1}{2\gamma - 1} = \frac{15}{7} \).

Choosing \( \beta < \frac{15}{7} \) and allowing the prior means to both be 5.71, the claim of the discussion is put to a test when letting \( \lambda_1 = \lambda_2 = \lambda \) equal 40 and 90, two values far from the prior mean 5.71. Table 9 lists the results for the Bayesian-related estimators under these conditions, while setting \( \alpha_1 = \alpha_2 = \alpha \).

We can see from Table 9 that \( \hat{\lambda}_B \) performs better than \( \hat{\lambda}_{mm} \). In fact, \( \hat{\lambda}_B \) does better than using \( \frac{X_i}{\gamma + 1} \) or \( \hat{\lambda}_A \), as well. But it must be pointed out that \( \hat{\lambda}_B \) is being studied under fabricated conditions.
Table 9. The effect of choosing special values for prior parameters

<table>
<thead>
<tr>
<th>λ</th>
<th>E(λ)</th>
<th>V(λ)</th>
<th>α</th>
<th>β</th>
<th>2 MSE(λ^A_i)</th>
<th>2 MSE(λ^A_i,NN)</th>
<th>2 MSE(λ^A_i,K)</th>
<th>2 MSE(λ^A_i,Σ)</th>
<th>X_i/γ+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>5.71</td>
<td>326</td>
<td>6.28</td>
<td>2.10</td>
<td>601.52</td>
<td>646.06</td>
<td>636.98</td>
<td>640</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>5.71</td>
<td>3260</td>
<td>5.77</td>
<td>2.01</td>
<td>605.51</td>
<td>654.98</td>
<td>636.97</td>
<td>640</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>5.71</td>
<td>326</td>
<td>6.28</td>
<td>2.10</td>
<td>3152.40</td>
<td>3270.68</td>
<td>3251.96</td>
<td>3240</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>5.71</td>
<td>3260</td>
<td>5.77</td>
<td>2.01</td>
<td>3159.12</td>
<td>3315.83</td>
<td>3251.95</td>
<td>3240</td>
<td></td>
</tr>
</tbody>
</table>
Once again, despite the large discrepancy between $\lambda$ and $E(\lambda)$, all three of the Bayesian-related estimators perform better than using the component estimator $\frac{X_1}{Y+1}$. Some of this can be attributed to the large prior variance.

Now we study these same estimators under less specific conditions. We simply compare them for different values of $\lambda_1$, $\lambda_2$, $E(\lambda_1)$, $E(\lambda_2)$, $V(\lambda_1)$ and $V(\lambda_2)$. Assuming $E(\lambda_1) = 1.71$ and a different value $E(\lambda_2) = 5.71$, we refer to Table 10.

When a component value of $\lambda_1$ equals $E(\lambda_1)$ and the value of $V(\lambda_1)$ is small, very little is contributed to the total MSE of $\hat{A}_B$. A poor choice of $E(\lambda_1)$, however, is not disastrous if accompanied by a forgivingly large $V(\lambda_1)$. But when such a poor choice of $E(\lambda_1)$ is coupled with a small $V(\lambda_1)$, the MSE of $\hat{A}_B$ suffers as seen in the sixth row of Table 10.

Again the performance of $\hat{A}_A$ stands between the other two, except when a wise choice of $E(\lambda_1)$ and $E(\lambda_2)$ is coupled with large prior variances. Then $\hat{A}_A$ performs better than $\hat{A}_B$ and $\hat{A}_{mm}$.

Since $\hat{A}_{mm}$ shrinks toward a common $\frac{X}{Y}$, it is not surprising to see it sometimes beaten by the other estimators when $\lambda_1 \neq \lambda_2$.

Other simulation results have shown that, when $E(\lambda_1)$ and $E(\lambda_2)$ are quite far from 1.71 and 5.71 respectively, all three Bayesian-related estimators do worse than using the component estimator $\frac{X_1}{Y+1}$. 
Table 10. The performance of Bayesian-related estimators under general conditions

<table>
<thead>
<tr>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\nu(\lambda_1)$</th>
<th>$\nu(\lambda_2)$</th>
<th>$\frac{2}{\sum_{i=1}^{2} \text{MSE}^{\wedge}(\lambda^{\wedge}_{1B})^{i=1}}$</th>
<th>$\frac{2}{\sum_{i=1}^{2} \text{MSE}^{\wedge}(\lambda^{\wedge}_{1mm})^{i=1}}$</th>
<th>$\frac{2}{\sum_{i=1}^{2} \text{MSE}^{\wedge}(\lambda^{\wedge}_{1A})^{i=1}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.71</td>
<td>5.71</td>
<td>1/2</td>
<td>1/2</td>
<td>0.018</td>
<td>1.160</td>
<td>0.469</td>
</tr>
<tr>
<td>1.71</td>
<td>5.71</td>
<td>5.44</td>
<td>5.44</td>
<td>0.205</td>
<td>1.038</td>
<td>0.491</td>
</tr>
<tr>
<td>1.71</td>
<td>5.71</td>
<td>98</td>
<td>98</td>
<td>0.710</td>
<td>1.066</td>
<td>0.460</td>
</tr>
<tr>
<td>9.71</td>
<td>5.71</td>
<td>1/2</td>
<td>1/2</td>
<td>1.132</td>
<td>0.836</td>
<td>0.918</td>
</tr>
<tr>
<td>9.71</td>
<td>5.71</td>
<td>5.44</td>
<td>5.44</td>
<td>0.721</td>
<td>0.976</td>
<td>0.838</td>
</tr>
<tr>
<td>9.71</td>
<td>5.71</td>
<td>1/2</td>
<td>98</td>
<td>1.312</td>
<td>0.856</td>
<td>1.093</td>
</tr>
<tr>
<td>9.71</td>
<td>5.71</td>
<td>98</td>
<td>1/2</td>
<td>0.694</td>
<td>1.033</td>
<td>0.781</td>
</tr>
</tbody>
</table>
Setting $\lambda_1 = \lambda_2 = \lambda$ once again, the performance of the Bayesian-related estimators are graphed on Figures 4 and 5. Both assume $E(\lambda_1) = E(\lambda_2) = 5.71$, but Figure 4 also sets $V(\lambda_1) = V(\lambda_2) = 5.4$ while Figure 5 sets these prior variances to $1/2$.

From these figures, we note that $X^\lambda$ performs best for $\lambda$ near the prior mean $5.71$. All three of the estimators show improvement over using $X_i / (\gamma + 1)$ for most of the parameter space of $\lambda$ charted. Also note that the area of improvement for $\hat{A}_B$ is smaller when $V(\lambda) = 1/2$ as compared to when $V(\lambda) = 5.4$.

Let us now consider the Thompson-type shrinkage estimators of the multivariate vector $\lambda$. Figures 6 and 7 display graphs of the MSEs of these shrinkage estimators where $\lambda_1 = \lambda_2 = \lambda$ and $\theta_1 = \theta_2 = 5.71$. Figure 6 represents the estimators shrinking from $X_i$, while Figure 7 represents those shrinking from $X_i / (\gamma + 1)$.

Since $\lambda_1 = \lambda_2$, one would expect the estimators employing a common shrinking factor for both component estimators to prove more valuable. In almost all cases, this holds true.

So now we wish to find the shrinkage estimators that fare the best over all regions of the parameter space of $\lambda$. Note that, when $\lambda$ is 3 or less, none of the estimators are really able to improve upon $X_i / (\gamma + 1)$. But in the region from 3.71 to 11.71, the best shrinkers are those that shrink from the MSE minimizing estimator $X_i / (\gamma + 1)$. This means that, if one suspects the values of $\theta_1$ and $\theta_2$ to be close to those of $\lambda_1$ and $\lambda_2$, shrinking from $X_i / (\gamma + 1)$ proves most beneficial.
Figure 4. MSE of Bayesian-related estimators, $V(\lambda) = 5.44$, $E(\lambda) = 5.71$, $\alpha = 40$, $\beta = 8$
Figure 5. MSE of Bayesian-related estimators, $V(\lambda) = \frac{1}{2}$, $E(\lambda) = 5.71$, $\alpha = 378$, $\beta = 67$. 
Figure 6. Thompson-type estimators shrunk from $\frac{1}{\gamma}$, the gamma case, $\theta_1 = \theta_2 = 5.71$. 

(6.0) $\text{MSE}(\frac{1}{\gamma+1})$ $X$

(6.1) $\text{MSE}(\frac{\hat{\lambda}_{\theta}}{\gamma})$, substituted for $\lambda_i$

(6.2) $\text{MSE}(\frac{\hat{\lambda}_{\theta}}{\gamma})$, substituted for $\lambda_i$

(6.3) $\text{MSE}(\frac{\hat{\lambda}_{\theta}}{\gamma})$, substituted for $\lambda_i$

(6.4) $\text{MSE}(\frac{\hat{\lambda}_{\theta}}{\gamma})$, substituted for $\lambda_i$

(6.5) $\text{MSE}(\frac{1}{\gamma})$
Figure 7. Thompson-type estimators shrunk from $\frac{X_i}{Y + 1}$, the gamma case, $\theta_1 = \theta_2 = 5.71$. 
But once the common $\lambda$ is greater than 12, the estimator $\hat{X}_{i\theta}$ dominates all other Thompson-type shrinkage estimators, as long as the unbiased estimator of $\lambda$, $\frac{X_i}{Y}$, is substituted into the shrinking factors.

What we see here is a complex arrangement of optimal estimators. Sometimes it depends on what one is shrinking from and sometimes the type of substitution used for the shrinking factor makes the difference.

Of all the estimators considered for the case $\lambda_1 = \lambda_2 = \lambda$, it appears $\hat{X}_{mn}$ performs best for $\lambda$'s somewhat larger than the value 5.71. One must realize that $\hat{X}_{mn}$ utilizes information from all components by shrinking to $\frac{X}{Y}$ and, hence, is not overly dependent on any prior means or constants $\theta_1$ and $\theta_2$.

For $\lambda$'s near 5.71, $\hat{X}_B$ appears to dominate, but its dependency on two prior parameters and a prior distribution means much more specification than what is necessary for a well behaved estimator $\hat{X}_{i\theta\gamma+1}$ or $\hat{X}_{i\theta\gamma+1}$.

Considering the case of $\lambda_1 \neq \lambda_2$, shrinking from $\frac{X_i}{Y + 1}$ loses its appeal. Otherwise, the performance of the estimators tends to be better if the shrinking factors are allowed to differ. Should $\lambda_1$ and $\lambda_2$ differ by a lot, the estimator $\hat{X}_{i\theta_i}$ seems optimal when $\lambda_i$ is substituted by $\frac{X_i}{Y}$. Table 11 considers the proportion of MSE of the most promising Thompson-type estimators to that of using the component estimator $\frac{X_i}{Y + 1}$ and verifies the considerations given for the case $\lambda_1 \neq \lambda_2$. 
Table 11. Proportional MSEs of Thompson-type shrinkage estimators, the gamma case, where \( \theta_1 = \theta_2 = 5.71 \) and \( \hat{\lambda}_1 = \hat{\lambda}_1 \theta_i (\gamma + 1) \)

with \( \frac{X_i}{\gamma + 1} \) substituted for \( \lambda_1 \) in \( c, \quad \hat{\lambda}_2 = \hat{\lambda}_2 \theta_i \) with \( \frac{X_i}{\gamma} \)

substituted for \( \lambda_1 \) in \( c, \quad \hat{\lambda}_3 = \hat{\lambda}_3 \theta_i \) with \( \frac{X_i}{\gamma} \) substituted for \( \lambda_1 \) in \( c \).

<table>
<thead>
<tr>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \frac{2}{\sum \text{MSE}(\hat{\lambda}_1)} )</th>
<th>( \frac{2}{\sum \text{MSE}(\hat{\lambda}_2)} )</th>
<th>( \frac{2}{\sum \text{MSE}(\hat{\lambda}_3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.71</td>
<td>.71</td>
<td>1.802</td>
<td>.859</td>
<td>.751</td>
</tr>
<tr>
<td>9.71</td>
<td>1.71</td>
<td>1.229</td>
<td>.884</td>
<td>.786</td>
</tr>
<tr>
<td>9.71</td>
<td>3.71</td>
<td>.592</td>
<td>.720</td>
<td>.776</td>
</tr>
<tr>
<td>9.71</td>
<td>5.71</td>
<td>.455</td>
<td>.617</td>
<td>.689</td>
</tr>
<tr>
<td>9.71</td>
<td>9.71</td>
<td>.553</td>
<td>.688</td>
<td>.759</td>
</tr>
<tr>
<td>7.71</td>
<td>1.71</td>
<td>1.408</td>
<td>.832</td>
<td>.622</td>
</tr>
<tr>
<td>7.71</td>
<td>3.71</td>
<td>.440</td>
<td>.606</td>
<td>.666</td>
</tr>
<tr>
<td>7.71</td>
<td>7.71</td>
<td>.344</td>
<td>.512</td>
<td>.606</td>
</tr>
<tr>
<td>5.71</td>
<td>1.71</td>
<td>2.051</td>
<td>.918</td>
<td>.623</td>
</tr>
<tr>
<td>5.71</td>
<td>3.71</td>
<td>.374</td>
<td>.566</td>
<td>.639</td>
</tr>
<tr>
<td>5.71</td>
<td>5.71</td>
<td>.170</td>
<td>.384</td>
<td>.510</td>
</tr>
<tr>
<td>3.71</td>
<td>1.71</td>
<td>4.674</td>
<td>1.237</td>
<td>1.130</td>
</tr>
<tr>
<td>3.71</td>
<td>3.71</td>
<td>.864</td>
<td>.844</td>
<td>.571</td>
</tr>
<tr>
<td>1.71</td>
<td>1.71</td>
<td>24.436</td>
<td>1.578</td>
<td>2.043</td>
</tr>
</tbody>
</table>
6. SIMULTANEOUS ESTIMATION OF PARAMETERS IN EXPONENTIAL FAMILIES

6.1 Introduction

So far, this dissertation has proposed estimators for parameter vectors of the multivariate normal, Poisson and gamma distributions, which are generally biased shrinkage estimators. Whether the estimators were evaluated in terms of MSE or in terms of Bayes risk, the squared error loss function was employed and the proposed shrinkage estimators were shown to improve upon standard estimators for at least a reasonable range of values of the parameters being estimated.

Under the assumption of \( \mathbf{X} = (X_1, X_2, \ldots, X_p)' \sim N(\mathbf{\Theta}, I) \), where \( p \geq 3 \) and \( \mathbf{\Theta} = (\theta_1, \theta_2, \ldots, \theta_p)' \), we have stated in Chapter I that Stein (1956) proved the inadmissibility of \( \mathbf{X} \) under squared error loss and that later James and Stein (1961) proposed an explicit estimator that dominates \( \mathbf{X} \). In this chapter, we consider the estimation of \( \mathbf{\Theta} \), where \( \mathbf{X} \) has a more general \( p \)-variate distribution depending on \( \mathbf{\Theta} \), and we also assume a more general loss function. Using a method of proof based upon the integration of parts technique (or its discreet analogue) of Stein (see Stein (1973), Stein (1981) or Hwang (1979)), we propose estimators of parameter vectors which again dominate \( \mathbf{X} \).

The loss function considered initially in this chapter is

\[
L_d(\mathbf{\Theta}, \mathbf{a}) = \sum_{i=1}^{p} d_i (\theta_i - a_i)^2
\] (6.1)
where \(d_i > 0, i = 1, 2, \ldots, p\) and \(d = (d_1, d_2, \ldots, d_p)'\). Not only do we wish to find estimators that dominate \(\bar{x}\) with respect to the loss \(L_d\), but with respect to \(L_d\) for various values of \(d\) simultaneously.

The concept of using loss functions like \(L_d\) for estimating the vector of a \(p\)-variate normal distribution was discussed by Brown (1975). If one used \(L_d\) where \(d = (1, 1, \ldots, 1)'\), this would simply be the usual squared error loss function. In wondering why \(\bar{x}\) is inadmissible with respect to squared error loss while any \(X_i\) is, by itself, an admissible estimator of \(\theta_i\), Brown gave the following justification.

By assuming a loss function which totals individual squared error losses, one implicitly assumes that each of the \(p\) component losses are equally important. This implicit assumption destroys the independence of the \(p\) estimation problems. Hence, it is reasonable that the component estimators utilize information from all \(p\) components. If the \(p\) component losses are not to be equally weighted, the loss function of \(L_d\) applies and still, the resulting component estimator of \(\theta_i\) should utilize information from all components as long as the weights, the \(d_i\)'s, do not differ by a great deal.

Where Brown (1975) and Shinozaki (1980) discuss simultaneous estimation with \(L_d\) for the case of \(X \sim N_p(\Theta, I)\), our work extends their results in several directions. First, in Section 6.2, we consider estimation of the natural parameter vector from \(p\) independent distributions each belonging to the one-parameter expo-
nential family absolutely continuous with respect to Lebesgue measure. Assuming losses of the form (6.1), a class of estimators is produced dominating the minimum variance unbiased estimator or some constant multiple of it simultaneously for various d. The general results are illustrated with the estimation of the natural parameter vector for the normal and the gamma distributions. Next, in this section, a subfamily of the general exponential family of distributions is considered and a class of estimators dominating the minimum variance unbiased estimator of the mean vector is produced, again using losses of the form (6.1).

In section 6.3, the normal case is considered with a more general covariance structure than the one considered by Brown (1975) and Shinozaki (1980), and a class of estimators dominating the maximum likelihood estimator of the mean vector is produced.

In the final section, the discrete exponential family of distributions is considered and a general class of estimators dominating the minimum variance unbiased estimator is produced in a manner similar to that of the previous sections.

6.2 Estimators in the Absolutely Continuous Case

Consider p independent random variables $X_i$, $i = 1, 2, \ldots, p$, where the density function of $X_i$ is

$$f(x_i | \theta_i) = \pi_i(\theta_i) \rho_i(x_i) e^{-\theta_i r_i(x_i)}$$  \hspace{1cm} (6.2)
for \( i = 1, 2, \ldots, p \), defined with respect to Lebesgue measure on \((a_1, b_1), a_i\) and \(b_i\) being possibly infinite. We wish to estimate the parameter vector \( \theta \). This problem is addressed in Hudson (1978), Berger (1980) and Ghosh and Parsian (1980).

For any estimator \( \hat{\theta}^0(X) = (\hat{\theta}_1^0(X), \hat{\theta}_2^0(X), \ldots, \hat{\theta}_p^0(X))' \) of \( \theta \), consider the competitor \( \hat{\theta}(X) = (\hat{\theta}_1(X), \hat{\theta}_2(X), \ldots, \hat{\theta}_p(X))' \) with \( \delta_i(X) = \delta_i^0(X) - q_i(X)\phi_i(X) \), where

\[
q_i(X) = r_i(X_i)e^{s_i(X)/p_i(X_i)},
\]

and \( s_i(X) \) is defined by

\[
s_i^{(1)}(X) = \frac{\partial s_i(X)}{\partial x_i} = \delta_i^0(X)r_i(X_i).
\]

These expressions are subject to the following regularity assumptions:

(a) \( E_{\theta}q_i^2(X)\phi_i^2(X) < \infty \),

(b) \( \phi_i(x) \) is absolutely continuous as a function of \( x_i \) in any compact subset of \((a_i, b_i)\) for almost all \( x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \),

\[
\lim_{x_i \to a_i} (s_i(x) - \theta_i r_i(x_i)) = \lim_{x_i \to b_i} (s_i(x) - \theta_i r_i(x_i)) = 0 \text{ for all } \theta_i.
\]
These conditions are usually satisfied and are easy to verify as shown by Ghosh and Parsian (1980). Under the loss \((6.1)\), the above conditions when joined with

\[(e) \quad R(\theta, \delta^0) < \infty \text{ for all } \theta,\]

were used by Berger (1980) to imply that

\[R(\theta, \delta) - R(\theta, \delta^0) = E_0 \Delta(X) \tag{6.3}\]

\[\Delta(X) = 2 \sum_{i=1}^{p} d_i(q_i(x)/r_i(x))\phi_i^{(1)}(x)\]

\[+ \sum_{i=1}^{p} d_i q_i^2(x) \phi_i^2(x)\]

for

\[\phi_i^{(1)}(x) = \frac{\partial \phi_i(x)}{\partial x_i}.\]

Note that the minimum variance unbiased estimator of \(\theta_i\) is

\[k_i(X_i) = (r_i'(X_i))^{-1}\left(\frac{\rho_i'(X_i)}{\rho_i(X_i)}\right) + \frac{d}{dx_i} (r_i'(X_i))^{-1}.\]

Note that for any specified \(c_i > 1\), \(c_i k_i(X_i)\) is inadmissible, it being dominated by \(k_i(X_i)\). Consider \(\delta_i^0(X) = c_i k_i(X_i)\) for some
specified $c_i \in (0, 1]$, $i = 1, 2, \ldots, p$. The calculations of Ghosh and Parsian (1980) show that when $r_i(x_i)$ is increasing in $x_i$,

$$q_i(x) = \left(\frac{r_i(x_i)}{\rho_i(x_i)}\right)^{1-c_i}$$  \hspace{1cm} (6.4)

for $i = 1, 2, \ldots, p$, while if $r_i(x_i)$ is decreasing in $x_i$,

$$q_i(x) = -\left(-\frac{r_i(x_i)}{\rho_i(x_i)}\right)^{1-c_i}$$  \hspace{1cm} (6.5)

for $i = 1, 2, \ldots, p$. In either case, we can write

$$\Delta(x) = 2 \sum_{i=1}^{p} d_i v_i(x_i) \phi_{i1}^1(x) + \sum_{i=1}^{p} d_i w_i(x_i) \phi_{i2}^2(x),$$  \hspace{1cm} (6.6)

for some appropriately defined $v_i(x_i)$ and $w_i(x_i)$. We are now in a position to prove the main result of this section.

**Theorem 1.** Assume $v_i^{-1}(x_i)$ is integrable with $g_i(x_i) = v_i^{-1}(x_i)$. Let

$$s = \sum_{i=1}^{p} |g_i(x_i)|^\beta$$

for some $\beta > 0$. It is assumed that

(a) there exists a constant $K > 0$ such that

$$\sum_{i=1}^{p} w_i(x_i) g_i^2(x_i) \leq KS;$$  \hspace{1cm} (6.7)

(b) there exists constants $a_i(> 0)$ such that

$$\inf_{x \in \mathbb{D}} \left[ \left( \sum_{i=1}^{p} d_i a_i - \beta \max_{1 \leq i \leq p} (d_i a_i) \right) / \max_{1 \leq i \leq p} (d_i a_i) \right] = a_0,$$

$$a_0 > 0,$$  \hspace{1cm} (6.8)
where \( D \) is some subset of \((R^+)^p\) for which \((6.8)\) holds.

Then for any \( \tau(S) \) satisfying

(c) \( \tau(S) \) increasing in \( S \);
(d) \( 0 < \tau(S) < 2K^{-1}a_0 \),

defining

\[
\phi_i(x) = \frac{-a_i \tau(S)}{S} g_i(x_i),
\]

\( i = 1, 2, \ldots, p \), for \( \phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_p(x)) \), provides a solution to the differential inequality \( \Delta(x) < 0 \). Consequently, under losses of the form \((6.1)\) with \( d \in D \), \( \delta^0(X) \) is improved by

\[
\delta(X) = (\delta^0_1(X) - q_1(X) \phi_1(X), \ldots, \delta^0_p(X) - q_p(X) \phi_p(X)).
\]

Two points should be made before proving Theorem 1. First, assumptions (c) and (d) can always be made by taking \( \tau(S) \) to be any constant in the interval \((0, 2K^{-1}a_0)\), for example. Also, for any given set of \( d_i \)’s, it is not necessarily possible to find \( a_i \)’s such that \((6.8)\) holds for \( d_i > 0 \) for \( i = 1, 2, \ldots, p \). To see an example of this, see Shinozaki (1980).

**Proof of Theorem 1.** First note that

\[
\phi_i^{(1)}(x) = \frac{-a_i \tau(S)}{S} g_i'(x_i) + a_i \beta_i(X) \left[ \frac{S \tau'(S) - \tau(S)}{S^2} \right] \beta S \frac{\beta}{\beta x_i} g_i(x_i)
\]

\[
= \frac{-a_i \tau(S)}{S} g_i'(x_i) + a_i \left( \frac{-\tau'(S)}{S} + \frac{\tau(S)}{S^2} \right) \beta |g_i(x_i)| \beta_i(X).
\]

\((6.9)\)

Hence, using, \((6.6)\), \((6.7)\), \((6.8)\), \((6.9)\), (c) and (d) one gets,
\[ \Delta(x) = 2 \sum_{i=1}^{p} d_i \frac{1}{g_i(x_i)} \phi_i^{(1)}(x) + \sum_{i=1}^{p} d_i w_i(x_i) g_i^2(x) \]

\[ = 2 \sum_{i=1}^{p} d_i a_i \left( -\frac{\tau(S)}{s} + \frac{\tau(S)}{s^2} - \frac{\tau^*(S)}{s} \right) \beta |g_i(x_i)|^\beta \]

\[ + \sum_{i=1}^{p} d_i a_i^2 w_i(x_i) g_i^2(x_i) \frac{\tau^2(S)}{s^2} \]

\[ \leq \frac{-2\tau(S)}{s} \sum_{i=1}^{p} d_i a_i + 2\beta \sum_{i=1}^{p} d_i a_i |g_i(x_i)|^\beta \left( \frac{\tau(S)}{s} \right) \]

\[ + \frac{\tau^2(S)}{s^2} \sum_{i=1}^{p} d_i a_i^2 w_i(x_i) g_i^2(x_i) \]

\[ \leq \frac{-2\tau(S)}{s} \sum_{i=1}^{p} d_i a_i + 2\beta \max_{1 \leq i \leq p} |d_i a_i| \frac{\tau(S)}{s} + \left( \max_{1 \leq i \leq p} d_i a_i^2 \right) \left( \frac{\tau^2(S)}{s^2} \right) \]

\[ = \frac{-\tau(S)}{s} \left( \sum_{i=1}^{p} d_i a_i - 2\beta \max_{1 \leq i \leq p} d_i a_i \right) - \tau(S) K \left( \max_{1 \leq i \leq p} d_i a_i^2 \right) \]

\[ = \frac{-\tau(S)}{s} \left( \max_{1 \leq i \leq p} d_i a_i^2 \right) - \tau(S) K \left( \max_{1 \leq i \leq p} d_i a_i \right) \]

\[ \leq \frac{-\tau(S)}{s} \left( \max_{1 \leq i \leq p} d_i a_i^2 \right) (2a_0 - \tau(S)) < 0, \]  

\[ 1 \leq i \leq p \]

(6.11)
completing the proof of Theorem 1.

Two important applications of Theorem 1 follow:

Suppose $X_1, X_2, \ldots, X_p$ are independent normal variables with means $\theta_1, \theta_2, \ldots, \theta_p$ and unit variances. Then $r_i(x_i) = -x_i$ and $s^0(x_i) = x_i$, so that $s_i^1(x) = -x_i$. Hence, $s_i(x_i) = -\frac{1}{2}x_i^2$. Also, since

$$\rho_i(x_i) = e^{-\frac{1}{2}x_i^2},$$

$q_i(x) = -1$. So $v_i(x_i) = w_i(x_i) = 1$ and, furthermore, $\xi_i(x_i) = x_i$. Consequently, (6.7) holds with $S = \sum_{i=1}^{p} x_i^2$ and $K = 1$. Now, take

$$\phi_i(x) = \frac{-a_i \tau(S)}{S} x_i$$

where the $a_i$'s satisfy (6.8). In order that the assumptions (a) through (e) hold, it is assumed that $\tau(S)$ is differentiable, and $\mathbb{E}|\tau'(S)| < \infty$ for all $S$. If, in addition, $\tau(S)$ is increasing in $S$ and $0 < \tau(S) < 2a_0$, $(X_1 + \phi_1(X), X_2 + \phi_2(X), \ldots, X_p + \phi_p(X))$ dominates $X$ under losses of the form (6.1) simultaneously for $d \in D$.

This example includes as a special case Theorem 1 of Shinozaki (1980). Also, if $D = [d: d_1 = d_2 = \ldots = d_p = d > 0]$, taking $a_1 = a_2 = \ldots = a_p = 1$, so that $a_0 = p - 2$, one gets a class of estimators dominating $X$ as proposed by Baranchick (1970) and Strawderman (1971).

The second illustration lets $X_1, X_2, \ldots, X_p$ be $p$ independent variables having density function
where \( x_i > 0, \theta_i > 0, \alpha_i > 2 \) and \( i = 1, 2, \ldots, p \). The minimum mean square error estimator of \( \theta_i \) in the class of all estimators of the form \( k_i/X_i \) is \((\alpha_i - 2)/X_i\) as shown in Chapter V. Such estimators are admissible in one dimension under the quadratic loss function.

Now, in this case, \( r_i(x_i) = x_i \) and \( p_i(x_i) = x_i^{\alpha_i-1} \), so that \( q_i(x_i) = x_i^{-1} \) and \( w_i(x_i) = x_i^{-2} \). With \( g_i(x_i) = \frac{1}{2}x_i^2 \), (6.8) holds with \( S = \frac{1}{2} \sum_{i=1}^{p} x_i^2 \) and \( K = \frac{1}{2} \). Hence, for \( \alpha_i \)'s satisfying (6.8), define

\[
\phi_i(x) = \frac{-a_i \tau(S)}{S} \left( \sum_{i=1}^{p} \frac{x_i^2}{2x_i^2} \right) = \frac{-a_i \tau(S)x_i}{p \sum_{i=1}^{p} x_i^2},
\]

where

(a) \( \tau(S) \) is increasing in \( S \),

(b) \( E(\tau'(S)) < \infty \) and

(c) \( 0 < \tau(S) < 4a_0 \) with \( a_0 \) defined in (6.8).

Now

\[
\left( \frac{\alpha_1 - 2}{X_1^2}, \frac{\alpha_2 - 2}{X_2^2}, \ldots, \frac{\alpha_p - 2}{X_p^2} \right)
\]

is dominated by
\[
\sum_{i=1}^{p} \frac{a_i \tau(S)}{\sum X_i^2} X_i, \quad \sum_{i=1}^{p} \frac{a_i \sigma(S)}{\sum X_i^2} X_i, \quad \ldots
\]

under losses of the form (6.1) simultaneously for \( d \in \mathbb{D} \).

Finally, in this section, we consider the following subfamily of the general exponential family of distributions with density functions given by

\[
f(x|\theta) = e^{(u(\theta)b(x) - \chi(\theta))} n^{-1}(x)e^{-\int n^{-1}(x)dx} (6.13)
\]

where \( u(\theta) = E_\theta(X) \) and \( b'(x) = n^{-1}(x) \). This particular subfamily was considered by Hudson (1978). Suppose now \( X_1, X_2, \ldots, X_p \) are independent, \( X_i \) having density function given in (6.13) with support \((w_1, w_2)\) where \( w_1 \) and \( w_2 \) can be possibly infinite. The aim, here, is to improve on the minimum variance unbiased estimator \( \bar{x} \) or \( \theta \).

Hudson (1978) has shown that under the conditions

\[
\lim_{x_i \to w_1} \phi_1(x) = \frac{(u(\theta_1)b(x_1) - \int x_1 n^{-1}(x_1)dx_1)}{n \Delta_1(x_1)} = 0
\]

\[
\lim_{x_i \to w_2} \phi_1(x) = \frac{(u(\theta_2)b(x_1) - \int x_1 n^{-1}(x_1)dx_1)}{n \Delta_1(x_1)} = 0
\]
for $i = 1, 2, \ldots, p$ and for almost all $x_1, x_2, \ldots$, $x_{i-1}, x_{i+1}, \ldots, x_p$, and

$(g) \quad E_\Theta |n(x_1)\phi_i(1)(x)| < \infty, \ i = 1, 2, \ldots, p, \ \text{for all} \ \Theta,$

one gets the identity

$$E_\Theta [(X_i - \mu(\Theta_i))\phi_i(x)] = E_\Theta [n(x_1)\phi_i(1)(x)],$$

$i = 1, 2, \ldots, p.$ \quad (6.14)

Hence, if $\delta^*_i(x) = (\delta^*_1(x), \delta^*_2(x), \ldots, \delta^*_p(x))$ with $\delta^*_i(x) = x_i + \phi_i(x)$,

then

$$\sum_{i=1}^{p} d_\Theta E_\Theta [\delta^*_i(x) - \mu(\Theta_i)]^2 = \sum_{i=1}^{p} d_\Theta [X_i - \mu(\Theta_i)]^2 = E_\Theta \Delta(x),$$

(6.15)

where

$$\Delta(x) = 2 \sum_{i=1}^{p} d_\Theta (x_i) \phi_i(1)(x) + \sum_{i=1}^{p} d_\Theta \phi_i^2(x).$$

(6.16)

It follows that (6.16) is also of the form (6.6) with $v_i(x_i) = n^{-1}(x_i)$ and $w_i(x_i) = 1$. Let $g_i(x_i) = \int v_i^{-1}(x_i)dx = b(x_i)$. Now, taking

$$S = \sum_{i=1}^{p} b^2(x_i),$$
it follows that (6.7) holds with $K = 1$. If (6.8) holds now with

$$
\beta = 1, \text{ define } \phi_i(X) = \frac{-a_i\tau(S)}{S} X_i, \ i = 1, 2, \ldots, p,
$$

where $\tau(S)$ is increasing in $S$, $E[\tau'(S)] < \infty$ and $0 < \tau(S) < 2a_0$. Then, $(X_1 + \phi_1(X), X_2 + \phi_2(X), \ldots, X_p + \phi_p(X))$ dominates $X$ under losses of the form (6.1) when $d \in \mathcal{D}$.

To see an application of the above result, consider the following example which appears in Hudson (1978), and in Ghosh and Parsian (1980).

Let $X_1, X_2, \ldots, X_p$ be independent, $X_i$ having density function

$$
f(x_i | \theta_i) = e^{\frac{\theta_i x_i^\theta_i - 1}{\theta_i}}, \ x_i > 0, \theta_i > 0,
$$

$i = 1, 2, \ldots, p$. (6.17)

In this case, $E_{\theta_i}(X_i) = \theta_i$ and the above class of density functions if of the form (6.13) with $b(x_i) = \log(x_i)$ and $n^{-1}(x_i) = x_i^{-1}$. Thus, (6.8) holds with $K = 1$ and $S = \sum_{i=1}^p (\log(x_i))^2$. Hence, for any $\tau(S)$ increasing in $S$ with $E[\tau'(S)] < \infty$ and $0 < \tau(S) < 2a_0$, $(X_1 + \phi_1(X), X_2 + \phi_2(X), \ldots, X_p + \phi_p(X))$ dominates $X$ with $\phi_i(X) = \frac{-a_i\tau(S)}{S} \log X_i$ for $i = 1, 2, \ldots, p$.

6.3 The Normal Case with General Covariance Structure

Let $X$ be distributed as $N_p(\theta, \sigma^2 V)$, where $V$ is a known positive definite matrix of order $p \geq 3$. We assume $\sigma^2$ is unknown, but an estimator $\hat{\sigma}^2$ of $\sigma^2$ is observed and $\hat{\sigma}^2 \sim \sigma^2 \chi^2_n/(n + 2)$ and is dis-
tributed independently of $X$. We use the multiplier $(n + 2)^{-1}$ rather than $n^{-1}$ so to simplify subsequent calculations.

For estimating $\theta$, consider losses of the form

$$L_d(\theta, u) = \sum_{i=1}^{p} d_i (\theta_i - u_i)^2 / \sigma^2$$

$$= \sigma^{-2} (\theta - u)' D^{-1} (\theta - u), \quad (6.18)$$

where $D = \text{diag}(d_1, d_2, \ldots, d_p)$ and $d_i > 0$ is known for $i = 1, 2, \ldots, p$. This loss is the same as the loss given in (6.1) with the exception of the additional multiplier $\sigma^{-2}$.

The usual maximum likelihood estimator of $\theta$ is $\hat{\theta}(X) = X$. In this section, we want to obtain estimators improving on $X$ under losses of the form (6.18) simultaneously for various $d \in \mathcal{D}$, where $\mathcal{D}$ is once again some suitable subset of $\mathbb{R}^p$. Since $V$ is positive definite, without any loss of generality, one can take $V = I_p$, the identity matrix of order $p$. This is true because, otherwise, using the simultaneous diagonalization theorem for matrices, there exists a nonsingular $F$ such that $F V F' = I_p$ and $F D^{-1} F' = D_0^{-1}$ where $D_0^{-1}$ is also a known diagonal matrix with strictly positive elements in the diagonals. It is now easy to see that

$$E[\hat{\theta}(X, \hat{\sigma}^2) - \theta]' D [\hat{\theta}(X, \hat{\sigma}^2) - \theta]$$

$$= E[\hat{\theta}(F^{-1} Z, \hat{\sigma}^2) - \theta]' (F' D_0 F) [\hat{\theta}(F^{-1} Z, \hat{\sigma}^2) - \theta]$$
\[
E[(\delta(F X, \theta^2) - \theta)'] D_0 E[(\delta(F^{-1} Z, \theta^2) - \theta)]
\]

\[= E[(n(Z, \theta^2) - F \theta)', D_0 (n(Z, \theta^2) - F \theta)]\]

where \(n(Z, \theta^2) = F \delta(F^{-1} Z, \theta^2)\). Since \(Z \sim N_p(F \theta, \sigma^2 I_p)\), a one-to-one correspondence can be set up for estimating \(\theta\) by \(X\) and \(F \theta\) by \(Z\) by redefining the loss (6.18) with \(D_0\) replacing \(D\) and \(F \theta\) replacing \(\theta\). For related discussions, see Strawderman (1978).

Thus, consider the situation when \(X \sim N_p(\theta, \sigma^2 I_p)\), and consider the competing estimators \(\delta(X) = (\delta_1(X), \delta_2(X), \ldots, \delta_p(X))\) of \(\theta\) where

\[
\delta_i(X) = (1 - \frac{a_i \tau(S/\sigma^2)}{S/\sigma^2})X_i,
\]

\(i = 1, 2, \ldots, p\), and \(S = X'X\). Then the risk difference is given by

\[
R(\theta, \delta) - R(\theta, \delta^0) = \sum_{i=1}^{p} d_i E_{\theta} \sigma^2 [X_i - \theta_i - \frac{a_i \tau(S/\sigma^2)}{S/\sigma^2} X_i]^2 / \sigma^2
\]

\[= -\sum_{i=1}^{p} d_i E_{\theta} \sigma^2 [X_i - \theta_i]^2 / \sigma^2 \]

\[+ \sum_{i=1}^{p} d_i^2 E_{\theta} \sigma^2 \frac{\tau(S/\sigma^2)}{S^2/\sigma^2} X_i^2 / \sigma^2. \]

(6.20)
We are now in a position to prove the main result of this section.

**Theorem 2.** Assume that (6.8) holds and \( \tau(S) \) is increasing in \( S \) with \( \mathbb{E}(\tau'(S)) < \infty \), and that \( 0 < \tau(S) < 2a_0 \) for \( a_0 \) defined as in (6.8).

Then \( \delta(X) \) dominates \( \delta^0(X) = X \) simultaneously for all \( d \in \mathcal{D} \).

**Proof of Theorem 2.** Since, \( \mathbb{E}(\tau'(S)) < \infty \), using Stein's integration by parts technique

\[
\mathbb{E}_q, \sigma^2 \left[ \frac{\tau(S/\sigma^2)}{S/\sigma^2} X_1(X_1 - \theta_1) \right] = \sigma^2 \mathbb{E}_q, \sigma^2 \left[ \frac{\frac{\partial}{\partial X_1}}{S/\sigma^2} \left\{ \frac{\tau(S/\sigma^2)}{S/\sigma^2} X_1 \right\} \right]
\]

\[
= \sigma^2 \mathbb{E}_q, \sigma^2 \left[ \frac{\partial^2}{\partial^2 S} \tau(S/\sigma^2) \right.
\]

\[
+ \frac{\partial^2}{\partial S} \tau'(S/\sigma^2) \frac{2X_1^2}{\sigma^4} - \frac{\partial^2}{\partial S} \tau(S/\sigma^2)(2X_1^2) \]

\[
\geq \sigma^2 \mathbb{E}_q, \sigma^2 \left[ \frac{\partial^2}{\partial S} \tau(S/\sigma^2) \right.
\]

\[
- \frac{\partial^2}{\partial S} \tau(S/\sigma^2)(2X_1^2) \right], \quad (6.21)
\]

since \( \tau'(S) \geq 0 \).

Again,
Using the integration by parts technique of Efron and Morris (1976), it follows that, writing $F = S/\sigma^2$,

\[
E_\theta, \sigma^2 \left[ \frac{\partial^2}{\partial \sigma^2} \frac{\tau^2(S/\sigma^2)}{S/\sigma^2} \right]
\]

\[
= \frac{n}{n + 2} E_\theta, \sigma^2 \left[ \frac{\tau^2(F)}{F} \right] + \frac{2}{n + 2} E_\theta, \sigma^2 \left[ \sigma^2 \frac{3}{\partial^2} \left( \frac{\tau^2(F)}{F} \right) \right]
\]

\[
= \frac{n}{n + 2} E_\theta, \sigma^2 \left[ \frac{\tau^2(F)}{F} \right] + \frac{2}{n + 2} E_\theta, \sigma^2 \left[ \sigma^2 \left( \frac{\tau^2(F)}{F} \right) \left( - \frac{F \tau'(F)}{F^2} \right) \right]
\]

\[
+ \frac{2}{n + 2} E_\theta, \sigma^2 \left[ \sigma^2 \left( \frac{2\tau(F)\tau'(F)}{F} \right) \left( - \frac{F}{\partial^2} \right) \right]
\]

\[
= \frac{n}{n + 2} E_\theta, \sigma^2 \left[ \frac{\tau^2(F)}{F} \right] + \frac{2}{n + 2} E_\theta, \sigma^2 \left[ \frac{\tau^2(F)}{F} \right]
\]

\[
= E_\theta, \sigma^2 \left[ \frac{\tau^2(F)}{F} \right],
\]

since $\tau(F) > 0$ and $\tau(F)$ is increasing in $F$. Now combining (6.20) - (6.23), it follows that for $d \in \mathcal{D}$
\[ R(\theta, \delta) - R(\theta, \delta_0) \]

\[ \leq \mathbb{E}_{\theta} \sigma^2 \left[ (-2 \sum_{i=1}^{p} d_i a_i + 4 \max_{1 \leq i \leq p} d_i a_i) \right. \]

\[ + \left. \left( \max_{1 \leq i \leq p} d_i a_i \right)^2 \right] \frac{\tau(F)}{F} \]

\[ < \max_{1 \leq i \leq p} \left( d_i a_i \right)^2 \left[ -2a_0 + 2a_0 \right] \frac{\tau(F)}{F} = 0. \quad (6.24) \]

The proof of Theorem 2 is complete.

Instead of defining \( S = X'X \), one can define \( S = X' \Sigma^{-1} X \)
where \( \Sigma = \text{diag}(c_1, c_2, \ldots, c_p) \) is a known diagonal matrix with positive diagonal elements. Then, one can improve on \( \bar{X} \) by \( \delta(X) = (\delta_1(X), \delta_2(X), \ldots, \delta_p(X)) \) with

\[ \delta_i(X) = -a_i \left( \frac{\tau'(S/\delta^2)}{S/\delta^2} \right) X_i \]

with this redefined \( S \) similar as before under the same conditions as \( a_i \)'s and \( \tau \).

6.4 Estimators in the Discrete Exponential Family

Let \( X_1, X_2, \ldots, X_p \) be \( p \) independent random variables where \( X_1 \) has the density function

\[ f(x_i | \theta_i) = \pi_i(\theta_i) t_i(x_i) \theta_i x_i, x_i = 0, 1, \ldots \quad (6.25) \]
where \( x_i = 0, 1, \ldots; \) and \( i = 1, 2, \ldots, p. \) The problem is to estimate \( \theta \) under losses of the form

\[
L_d(\theta, u) = \sum_{i=1}^{p} \frac{d_i(\theta - u_i)^2}{\theta_i^{m_i}},
\]

(6.26)

where the \( m_i \)'s are known positive integers.

The minimum variance unbiased estimator of \( \theta_i \) is

\[
\delta_i^0(x_i) = \frac{t_i(x_i - 1)}{t_i(x_i)},
\]

where \( t_i(-1) \) is defined as 0. Our goal in this section is to improve on the estimator \( \delta^0(X) = (\delta_1^0(X_1), \delta_2^0(X_2), \ldots, \delta_p^0(X_p))' \) of \( \theta \) under losses of the form (6.26) simultaneously for various \( d \in \mathcal{D} \), where \( \mathcal{D} \) is some suitable subset of \((\mathbb{R}^+)^p\). Consider the competitor estimators of the form \( \delta(X) = (\delta_1(X_1), \delta_2(X_2), \ldots, \delta_p(X_p))' \), where

\[
\delta_i(X) = \delta_i^0(X_i) + \phi_i(X),
\]

\( i = 1, 2, \ldots, p. \) If \( E\theta_i^2(X) < \infty \) for all \( i = 1, 2, \ldots, p, \) it will follow that, under (6.26) (see Hwang (1979)),

\[
R(\theta, \delta) - R(\theta, \delta^0) = 2E_U(X),
\]

where

\[
U(X) = \sum_{i=1}^{p} d_i v_i(x_i) \Delta_i \psi_i(x) + \sum_{i=1}^{p} d_i w_i(x_i) \psi_i^2(x),
\]

(6.27)

\[
v_i(x_i) = t_i(x_i + m_i - 1)/t_i(x_i), \quad w_i(x_i) = \frac{1}{2} t_i(x_i + m_i)/t_i(x_i),
\]

\[
\psi_i(x) = \phi_i(x + m_i e_i), \quad \text{and where } e_i \text{ is a vector with } i^{th} \text{ element 1}
\]

and the rest 0, and \( \Delta_i \psi_i(x) = \psi_i(x) - \psi_i(x - e_i). \) We want to obtain
solutions to the difference inequality $U(x) < 0$.

Assume that $v_i(x_i) > 0$ for all $x_i = 0, 1, \ldots$. One can see that this condition holds in important special cases of the Poisson and negative binomial distributions. We define

$$h_i(x_i) = \sum_{k=0}^{x_i} v_i^{-1}(k), \quad x_i = 0, 1, \ldots \quad (6.28)$$

and let $m_i(x_i) = b_i h_i(x_i)$ for some $b_i > 0$, $M = \sum_{i=1}^{p} m_i(x_i) + b_0$ and

$$M = \sum_{j=1}^{p} m_j(x_j) + m_i(x_i - 1) + b_0 = M - \Delta m_i(x_i) \quad (5.28).$$

Suppose now, there exists $a_1, a_2, \ldots, a_p$ each greater than 0, such that

$$\inf_{d \in D} \frac{p \sum_{i=1}^{p} d_i a_i - \max_{1 \leq i \leq p} d_i a_i}{\max_{1 \leq i \leq p} d_i a_i^{2}} = a_0 > 0 \quad (6.29).$$

If the elements of $D$ are bounded away from 0, then the existence of such $a_i$'s is automatically guaranteed. In this case, the class is much wider than in the previous two sections. Define now

$$\psi_i(x) = -ca_i h_i(x_i)/M \quad (6.30)$$

for $i = 1, 2, \ldots, p$.

The main result of this section is now as follows.

Theorem 3. Assume that (6.29) and the following condition
\[
\sum_{i=1}^{p} w_i(x_i)h_i^2(x_i) \leq \frac{1}{2} M
\]  
(6.31)

hold, then \(\psi(x) = (\psi_1(x), \psi_2(x), \ldots, \psi_p(x))\), defined in (6.30),

provides a solution to \(U(x) < 0\), where \(U(x)\) is defined in (6.27).

Consequently, \(\delta(X) = (\delta_1(X), \delta_2(X), \ldots, \delta_p(X))\)

+ \(\phi_p(X)\) dominates \(\delta^0(X)\), where \(\phi_i(X) = \psi_i(X - 1_i e_i)\), \(i = 1, 2, \ldots, p\).

Proof of Theorem 3. In view of (6.31),

\[
\sum_{i=1}^{p} d_i w_i(x_i)\psi_i^2(x) = c^2 \sum_{i=1}^{p} d_i a_i^2 w_i(x_i)h_i^2(x_i)/M^2 \\
\leq c^2 \left( \max_{1 \leq i \leq p} d_i a_i^2 \right) \sum_{i=1}^{p} w_i(x_i)h_i^2(x_i)/M^2 \\
\leq \frac{c^2}{2M} \left( \max_{1 \leq i \leq p} d_i a_i^2 \right). \tag{6.32}
\]

Also,

\[
\Delta_i \psi_i(x) = \psi_i(x) - \psi_i(x - e_i) \\
= -ca_i h_i(x_i) \frac{M}{M_i} - \left( \frac{-ca_i h_i(x_i - 1)}{M_i} \right) \\
= ca_i \left[ \frac{h_i(x_i - 1)}{M_i} - \frac{h_i(x_i)}{M} \right]
\]
\[
\begin{align*}
&= \frac{c a_i}{M} \left[ \frac{M h_i(x_i - 1) - h_i(x_i) M_i}{M_i} \right] \\
&= \frac{c a_i}{M} \left[ \frac{M h_i(x_i - 1) - \left[ h_i(x_i - 1) + \Delta_i h_i(x_i) \right] M_i}{M_i} \right] \\
&= \frac{c a_i}{M} \left[ \frac{h_i(x_i - 1)(M - M_i) - M_i \Delta_i h_i(x_i)}{M_i} \right] \\
&= \frac{c a_i}{M} \left[ \frac{h_i(x_i - 1) \Delta_i m_i(x_i) - M_i \Delta_i h_i(x_i)}{M_i} \right].
\end{align*}
\]

Now,
\[
\begin{align*}
p \sum_{i=1}^{p} d_i \psi_i(x_i) \Delta_i \psi_i(x) &= \sum_{i=1}^{p} \frac{c d_i a_i}{M M_i} \left[ v_i(x_i) h_i(x_i - 1) \Delta_i m_i(x_i) \right] \\
&= \frac{c}{M} \sum_{i=1}^{p} d_i a_i \left[ M_i^{L-1} v_i(x_i) h_i(x_i - 1) \Delta_i m_i(x_i) - 1 \right] \\
&= \frac{c}{M} \sum_{i=1}^{p} d_i a_i + \frac{c}{M} \sum_{i=1}^{p} d_i a_i (b_i h_i(x_i) - b_i h_i(x_i - 1)) \\
&= \frac{c}{M} \sum_{i=1}^{p} d_i a_i + \frac{c}{M} \sum_{i=1}^{p} b_i d_i a_i M_i^{L-1} h_i(x_i - 1) \\
& \leq \frac{c}{M} \left[ \sum_{i=1}^{p} d_i a_i - \left( \max_{1 \leq i \leq p} d_i a_i \right) \sum_{i=1}^{p} b_i h_i(x_i - 1) M_i^{L-1} \right]
\end{align*}
\]
where one uses the monotonicity of \( m_i(x_i) \) and \( h_i(x_i) \) in order to obtain the last inequality. Combining (6.31) and (6.33), and using (6.29), one gets

\[
U(x) \leq -c M^{-1} \left( \sum_{i=1}^{p} d_i a_i - \max_{1 \leq i \leq p} d_i a_i \right) + \frac{1}{2} c M^{-1} \max_{1 \leq i \leq p} d_i a_i^2
\]

\[
- \frac{2}{M} \max_{1 \leq i \leq p} d_i a_i^2 \left( \sum_{i=1}^{p} d_i a_i - \max_{1 \leq i \leq p} d_i a_i \right)
\]

\[
\leq - \frac{c}{2} M^{-1} \max_{1 \leq i \leq p} d_i a_i^2 (2a_0 - c) < 0 .
\]

This completes the proof of the theorem.

As an example of the application of this theorem, let \( X_1, X_2, \ldots, X_p \) be independent where \( X_i \sim \text{Poisson}(\theta_i) \), \( i = 1, 2, \ldots, p \). Consider the loss (6.26) with \( m_1 = m_2 = \ldots = m_p = 1 \). This is a generalized version of the loss considered by Clevenson and Zidek (1975), and Ghosh and Parsian (1981). Note that in this case, \( v_i(x_i) = 1 \) so that \( h_i(x_i) = x_i + 1 \) and \( w_i(x_i) = \frac{1}{2} (x_i + 1)^{-1} \). Taking \( b_i = 1 \), so that \( m_i(x_i) = h_i(x_i) = x_i + 1 \), it follows that (6.31) holds. Now, if (6.29) holds,
\begin{equation}
\psi_i(x) = \frac{-ca_i}{(b_0 + \sum_{i=1}^{p} (x_i + 1))} (x_i + 1), \quad i = 1, 2, \ldots, p,
\end{equation}

provide a solution to \( U(x) < 0 \). Consequently, \( (X_1 + \phi_1(x), X_2 + \phi_2(x), \ldots, X_p + \phi_p(x)) \) with

\begin{equation}
\phi_i(x) = \frac{-ca_i}{(b_0 + p - 1 + \sum_{i=1}^{p} x_i)} x_i
\end{equation}

dominates \( \bar{X} \) under losses of the form (6.26) simultaneously for \( d \in D \). This includes in particular the corresponding results of Clevenson and Zidek (1975) and of Ghosh and Parsian (1981).
7. CONCLUSION

New estimators for multivariate populations have been considered by extending the basic concept of Thompson (1968). Since these estimators have proved preferable over the classical estimators in many cases, the major questions become: When and why?

Since the estimators exist in the form of a shrinker, we have often been able to answer both questions at once. Depending on whether the shrinking factor was held constant over all components and depending on what the estimator was shrinking toward, the answers were found to change. So these estimators are proposed, not as much so to improve extensively from already existing estimators, but more to understand how the structure of the estimators, their shrinking factors and the components of their underlying distributions dictate the manner in which they work. This structure often allowed us to compare the estimators to Bayes, near Bayes and empirical Bayes estimators.

The results obtained in the realm of Bayesian estimation are due to the extension of already existing techniques and procedures.

A two-stage prior distribution considered by Ghosh and Parsian (1981) and the basic concept of empirical Bayes estimation allowed us to determine an equivalence between three types of Bayesian-related estimators in the Poisson distribution.
Another extension of the empirical Bayes concept and a special application of Albert's (1981) type of estimation allowed the construction of two estimators for the scale parameters of a multivariate gamma distribution.

In considering simultaneous estimation, estimators were constructed that improved upon standard estimators once again. The work of Brown (1975) and Shinozaki (1980) considered such estimation but for less general distributions. And the basic method we employed for constructing our estimators was used by Hudson (1978), Berger (1980b), and Ghosh and Parsian (1980), but they did not consider simultaneous estimation. Hence, our work studies the intersection of these two lines of research.
8. BIBLIOGRAPHY


9. ACKNOWLEDGEMENTS

First, I want to acknowledge the help and encouragement of the three professors under whom I have worked. Dr. Chein-Pai Han, who always made time to help me, and Dr. Malay Ghosh, from whose office I always exited with a feeling of determination and confidence, I heartily thank you for all you have done. I must also acknowledge the care and the consideration I received from the late Dr. B. V. Sukhatme, who I respected not only as a scholar but also as a person.

To my sweet and lovely wife, Cheryl, thank you for the patience and the understanding through many low times. Your love and support were essential.

To Drs. H. A. David and Dean Isaacson, I thank you for the opportunities and the listening ear you respectively provided me with.

I must express my appreciation to Drs. Vince Sposito, David Harville, Howard Meeks, and Robert Stephenson for the time they contributed to the reading and finalization of the dissertation.

For their friendship which kept me at an even keel, I am grateful to many close friends and especially Chris, Bill, Harold, Dick and Nagaraja.

If this dissertation can be viewed as a notch in my win column, Frances Bradley certainly deserves an official save for her organizing and her typing that brought a manuscript in my handwriting to such an excellent final form in such a short time. Thank you.
Finally, I must thank my parents who never pushed me yet always inspired me through 24 years of school. And I thank them for teaching me, through examples, how to appreciate and love life.