SIMULTANEOUS ESTIMATION OF THE PARAMETERS OF
THE DISTRIBUTIONS OF INDEPENDENT POISSON RANDOM VARIABLES

by

KAM-WAH TSUI
B.Sc., The Chinese University of Hong Kong, 1970
M.Sc., University of Windsor, 1974

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in
THE FACULTY OF GRADUATE STUDIES
Department of Mathematics
and
Institute of Applied Mathematics and Statistics

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June, 1978

© Kam-Wah Tsui, 1978
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics and Institute of Applied Mathematics and Statistics

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date June 30, 1978
ABSTRACT

This work is devoted to simultaneously estimating the parameters of the distributions of several independent Poisson random variables. In particular, we explore the possibility of finding estimators of the Poisson parameters which have better performance than the maximum likelihood estimator (MLE). We first approach the problem from a frequentist point of view, employing a generally scaled loss function, called the k-normalized squared error loss function

$$L_k(\hat{\lambda}, \lambda) = \frac{1}{\lambda K} \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i^k,$$

where k is a non-negative integer. The case k=0 is the squared error loss case, in which we propose a large class of estimators including those proposed by Peng [1975] as special cases. Estimators pulling the MLE towards a point other than zero as well as a point determined by the data itself are proposed, and it is shown that these estimators dominate the MLE uniformly. Under $L_k$ with $k \geq 1$, we obtain a class of estimators dominating the MLE which includes the estimators proposed by Clevenson and Zidek [1975].

We next approach the problem from a Bayesian point of view; a two-stage prior distribution is adopted and results for a large class of prior distributions are derived. Substantial savings in terms of mean squared error loss of the Bayes point estimators over the MLE are expected, especially when the Poisson parameters fall into a relatively narrow range.
An empirical Bayes approach to the problem is carried out along the line suggested by Clevenson and Zidek [1975]. Some results are obtained which parallel those of Efron and Morris [1973], who work under the assumption that the random variables are normally distributed.

We report the results of our computer simulation to quantitatively examine the performance of some of our proposed estimators. In most cases, the savings, under the appropriate loss functions, are an increasing function of the number of Poisson parameters. The simulation results indicate that our estimators are very promising. The savings of the Bayes estimators depend on the choice of prior hyperparameters, and hence proper choice leads to substantial improvement over the MLE.

Although most of the results in this work are derived under the assumption that only one observation is taken from each Poisson distribution, we extend some results to the case where possibly more than one observation is taken. We conclude with suggestions for further work.
TABLE OF CONTENTS

SECTION 1. INTRODUCTION

1.1 Background ......................................................... 1
1.2 Outline ............................................................... 5

SECTION 2. NOTATION AND FUNDAMENTALS

2.1 Notation ............................................................... 10
2.2 Basic Lemmas and Theorems ........................................ 11

SECTION 3. ESTIMATION UNDER SQUARED ERROR LOSS

3.1 Introduction ......................................................... 19
3.2 Notation ............................................................... 20
3.3 Shifting the MLE Towards k ....................................... 23
3.4 Adaptive Estimators ................................................ 36

SECTION 4. ESTIMATION UNDER K-NORMALIZED SQUARED ERROR LOSS

4.1 Introduction ......................................................... 41
4.2 Minimax Estimators ................................................ 41
4.3 Better Estimators Under $L_k$ ..................................... 50

SECTION 5. BAYESIAN ANALYSIS

5.1 Introduction ......................................................... 57
5.2 Estimates of the Parameters ....................................... 59
5.3 Marginal Posterior Density ........................................ 70
5.4 Summary ............................................................... 76
SECTION 6. EMPIRICAL BAYES ESTIMATION

6.1 Background ......................................................... 78
6.2 Relative Savings Loss in the Poisson Case ..................... 79
6.3 The Plus Rules ....................................................... 81
6.4 Bayes Rules with Respect to Other Priors ..................... 81
6.5 Truncated Bayes Rules ............................................... 83
6.6 The Risk Function of the Estimator \( \hat{\lambda}^* \) .............. 87
6.7 Simplification of \( R(b,\hat{\lambda}^*) \) ................................. 89
6.8 Risk Function of \( \hat{\lambda}^* \) as a Function of \( \lambda \) ........... 91
6.9 Risk Function of \( \hat{\lambda}' \) as a Function of \( \lambda \) ............ 94
6.10 Risk Function of the Clevenson-Zidek Estimators ............. 95
6.11 Summary .............................................................. 96

SECTION 7. COMPUTER SIMULATION

7.1 Introduction ......................................................... 98
7.2 Estimators Under \( k\)-NSEL ......................................... 100
7.3 Bayes Estimators ..................................................... 103
7.4 Estimators Under Squared Error Loss ............................. 109
7.5 Comparison of the Estimators ...................................... 112

SECTION 8. EXTENSIONS

8.1 Motivation ......................................................... 113
8.2 Estimators Under Generalized Squared Error Loss ............. 114
8.3 Estimators Under Generalized \( k\)-NSEL ......................... 117
8.4 An Application ...................................................... 120

SECTION 9. SUGGESTIONS FOR FURTHER RESEARCH ..................... 122

BIBLIOGRAPHY ............................................................ 123
<table>
<thead>
<tr>
<th>I.</th>
<th>Improvement Percentage of $\chi^2$ over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Narrow Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>Page 101</td>
</tr>
<tr>
<td>II.</td>
<td>Improvement Percentage of $\chi^2$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Wide Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>Page 101</td>
</tr>
<tr>
<td>III.</td>
<td>Improvement Percentage of $\lambda^1$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Narrow Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>Page 102</td>
</tr>
<tr>
<td>IV.</td>
<td>Improvement Percentage of $\lambda^1$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Wide Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>Page 102</td>
</tr>
<tr>
<td>V.</td>
<td>Improvement Percentage of $\lambda$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>104</td>
</tr>
<tr>
<td>VI.</td>
<td>Poisson-Distributed $\alpha$ ($u=0, v=0$) Narrow</td>
</tr>
<tr>
<td></td>
<td>Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>Page 104</td>
</tr>
<tr>
<td>VII.</td>
<td>Poisson-Distributed $\alpha$ ($u=0, v=0$) Effect of $\mu$</td>
</tr>
<tr>
<td></td>
<td>107</td>
</tr>
<tr>
<td>VIII.</td>
<td>Poisson-Distributed $\alpha$ ($u=0, v=0$) Wide Range</td>
</tr>
<tr>
<td></td>
<td>for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>107</td>
</tr>
<tr>
<td>IX.</td>
<td>Negative Binomial-Distributed $\alpha$ ($u=1, v=0, a=3.0$)</td>
</tr>
<tr>
<td></td>
<td>108</td>
</tr>
<tr>
<td>X.</td>
<td>Geometric-Distributed $\alpha$ ($u=1, v=0, a=1.0$)</td>
</tr>
<tr>
<td></td>
<td>108</td>
</tr>
<tr>
<td>XI.</td>
<td>Improvement Percentage of $\lambda(0)$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Narrow Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>110</td>
</tr>
<tr>
<td>XII.</td>
<td>Improvement Percentage of $\lambda(0)$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Wide Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>110</td>
</tr>
<tr>
<td>XIII.</td>
<td>Improvement Percentage of $\lambda[m]$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Narrow Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>111</td>
</tr>
<tr>
<td>XIV.</td>
<td>Improvement Percentage of $\lambda[m]$ over the MLE</td>
</tr>
<tr>
<td></td>
<td>Wide Range for $\lambda_i$'s</td>
</tr>
<tr>
<td></td>
<td>111</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor James Zidek for suggesting my thesis topic and guiding my research during the first year of my work. His keen insights were extremely helpful. I would also like to thank Professor S. James Press for supervising my work for the past two years, and for suggesting the topics studied in Sections 4 and 5. Both Professor Zidek and Professor Press provided encouragement and stimulation during the preparation of this thesis. Thanks are further extended to Professor Stanley Nash for meticulous reading of the thesis, and to Professor Fred Wan, who has been very helpful during my entire enrollment at the University of British Columbia.

The financial assistance of the University of British Columbia and the National Research Council of Canada, as well as the use of the computer facilities at the University of British Columbia and the University of California, Riverside, are gratefully acknowledged.

Finally, I would like to thank my wife for her steadfast encouragement and patient assistance.
SECTION 1. INTRODUCTION

1.1 Background

The literature on the problem of estimating the mean of a p-dimensional multivariate normal distribution \( \mathcal{N}(\theta, \Sigma) \), both when the covariance matrix \( \Sigma \) is known and unknown, has proliferated since Stein [1956] discovered the surprising result that when \( \Sigma = I \), the sample mean, which is the maximum likelihood estimator (MLE), is inadmissible under squared error loss when \( p \geq 3 \). Better estimators of the multivariate normal mean have been found by Alam [1973, 1975], Baranchik [1964, 1970], Berger [1976a, 1976b, 1976c], Berger and Bock [1976a, 1976b], Berger et al. [1977], Efron and Morris [1973, 1976], Haff [1976, 1977], Hudson [1974], James and Stein [1961], Lin and Tsai [1973], Stein [1974], Strawderman [1971, 1973], and others. Basically, these estimators are obtained by shrinking the MLE towards a known fixed point or a point determined by the data itself. When we consider the problem from another point of view, we see that for simultaneous estimating of several means of normal populations, it is inadmissible to estimate each population mean by its sample mean, even though the populations have no mathematical dependence and the observations between populations are independent.

Attention has also recently been brought to simultaneous estimation of the parameters of non-normal distributions, including the discrete distributions. Johnson [1971] shows that in the binomial case the usual estimator is admissible under squared error loss. This is due to the superb performance of the usual estimator near the boundaries of the parameter space. Also, attempts have been made to work with general members of the exponential family in the simultaneous estimation problem (Hudson [1974, 1977]).
In this study, we concentrate on the situation where the underlying distribution is Poisson. There are many practical situations that lead us to investigate the problem of simultaneous estimation of the parameters of several independent Poisson distributions. For example, a metropolitan area divided into several fire precincts might be interested in knowing the expected numbers of fires in each of the precincts in a fixed time period so that fire-fighting resources can be optimally allocated. During that period, the number of fires in a single precinct could be supposed to follow a Poisson distribution, with an intensity parameter which would be different across precincts.

In the study of the process of oilwell discovery by wildcat exploration in Alberta, Canada, Clévenson and Zidek [1975] suggest that the problem be treated as one of simultaneously estimating the parameters of the distributions of several independent Poisson random variables. They let each parameter represent the expected number of oilwell discoveries during a particular month. With their available data, the number of parameters is approximately 200.

Let \( x_1, \ldots, x_p \) be observations of \( p \) independent Poisson random variables \( X_1, \ldots, X_p \) with intensity parameters \( \lambda_1, \ldots, \lambda_p \), respectively and \( p \geq 2 \). Let \( x = (x_1, \ldots, x_p) \), \( X = (X_1, \ldots, X_p) \), and \( \lambda = (\lambda_1, \ldots, \lambda_p) \). The problem is to simultaneously estimate the parameters \( \lambda_1, \ldots, \lambda_p \) based on the data \( x \).

The usual estimator of \( \lambda \) is the MLE, \( \hat{X} \). One might conjecture that, as in the normal case, estimators uniformly better than the MLE can be found in the Poisson case when \( p \) is large. Using the normalized squared error loss function 
\[
L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} \frac{(\hat{\lambda}_i - \lambda_i)^2}{\lambda_i},
\]
Clevenson and Zidek [1975] do indeed obtain a large class of estimators dominating the MLE as long as \( p \geq 2 \).
Such a loss function was chosen partly because their $\lambda_i$'s were expected to be small and hence inaccurate estimation of small $\lambda_i$'s seemed highly undesirable; their loss function reflects the desire to penalize over-estimation of small $\lambda_i$'s. The loss function was chosen also because the usual estimator would be minimax in this case and because $\lambda_i$ is the $i^{th}$ population's variance. Their estimators shrink the MLE towards the origin. They give two reasons why one might expect that shrinking the usual estimator will yield a better estimator:

Firstly the loss function penalizes heavily for bad estimates when the $\lambda_i$'s are small, and in such cases it is only possible to produce bad overestimates. Secondly Stein's results suggest it is better to restrain random multivariate and hence chaotic observations by shrinking them toward some point, here zero which does play a distinguished role in view of the first reason above and because of its special nature as the extreme point of the parameter space.

The estimation problem under different error loss functions has been studied by others. Using the squared error loss function $L(\lambda, \hat{\lambda}) = \sum_{i=1}^{P} (\lambda_i - \hat{\lambda}_i)^2$, Hudson [1974] proposes some empirical Bayes estimators for $\lambda$. The estimators are expected to improve considerably on $X$ for a wide range of parameter values, especially when $\lambda_1, \ldots, \lambda_P$ are similar in value and are not too large. But when one parameter $\lambda_1$ is very large and the other parameters small, the estimators are inferior to the MLE $X$, though the loss is expected to be small. That is, the proposed estimators do not dominate the MLE uniformly. Hudson also derives an unbiased estimate $U$ for the improvement of the risk when the estimator $\hat{\lambda}$ is used instead of $X$, i.e.

$$\mathbb{E}_\lambda U(X) = R(\lambda, X) - R(\lambda, \hat{\lambda}).$$

A sufficient condition for $\hat{\lambda}$ to be uniformly better than $X$ is that $U(x) > 0$ for all $x$ and $U(x) > 0$ for some $x$. Estimators of the form
\hat{\lambda} = (I - C(X))X, are of considerable appeal. For, they are similar to those obtained in the normal case. However, Hudson [1974] shows that in order for the estimators of this form to yield \( U(x) \geq 0 \) for all \( x \), it is necessary that \( C(x) = 0 \).

When all the observations but one are zero, none of the observations can be moved any closer to the origin. This is because the dimension of the problem is then essentially one, and when \( p = 1 \), the MLE is known to be admissible under squared error loss (Hodges and Lehmann [1951]). Because of the difficulty of handling cases when all but one of the parameters are near zero, it was conjectured that it might be impossible to improve upon the MLE when squared error loss is the criterion. The results of Peng [1975] are therefore quite surprising. He shows that although the usual estimator is admissible when \( p \leq 2 \), it is inadmissible when \( p \geq 3 \) by proposing estimators that are actually superior to the usual one uniformly in \( \lambda \). One of his estimators shrinks the MLE towards the origin provided the number of non-zero observations is greater than two. He also proposes an estimator which shrinks all the non-zero components of the MLE towards zero, but which sometimes gives non-zero values for the zero components of the MLE. The same condition is required, namely the number of non-zero observations is at least three, in order that the proposed estimator be strictly better than the MLE.

So far, the estimators that are uniformly better than the MLE under certain loss functions shrink the MLE towards the origin only. There is no other point towards which the shrinkage is made. We propose some estimators which shrink the MLE towards an arbitrary prechosen nonnegative integer \( k \), and some estimators which shrink the MLE towards a point
determined by the data. These estimators dominate the usual one uniformly under squared error loss.

1.2 Outline

In Section 2, we introduce the notation to be used in the subsequent sections and given some basic lemmas that are useful. Essentially, the basic lemmas (Lemmas 2.2.2 and 2.2.7) provide identities of the "risk deterioration" under certain loss functions when an estimator \( \hat{\lambda} \) is used instead of the MLE \( X \). Two basic theorems will also be stated. The first one (Theorem 2.2.3), due to Peng [1975], provides estimators of the Poisson parameters which dominate the MLE uniformly under squared error loss. The other (Theorem 2.2.9), due to Clevenson and Zidek [1975], provides estimators which dominate the MLE uniformly under normalized squared error loss. An alternate proof for the latter theorem is provided.

Section 3 examines the estimation problem when the familiar squared error loss function is used. Although Peng [1975] investigates the problem and finds estimators dominating the MLE when \( p \geq 3 \), the performance of his estimators is expected to be good only when all the underlying parameters are relatively small or when only some of the parameters are large. It is therefore of interest to see if further improvements are possible. In this section, we generalize Peng's results and propose estimators \( \hat{\lambda}^{(k)} \) which dominate the MLE uniformly in \( \lambda \). For a fixed non-negative integer \( k \), \( \hat{\lambda}^{(k)} \) pulls the MLE towards \( k \) whenever the number of observations greater than \( k \) is at least three. The estimator \( \hat{\lambda}^{(k)} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p) \) is given by

\[ \hat{\lambda}_i^{(k)} = X_i + f_i(X), \]

where
\[ f_i(x) = -\frac{r_k \phi(x) h(x_i)}{\sum_{j=1}^{p} h^2(x_j)}, \quad i = 1, \ldots, p \]

and \( r_k \) is the maximum of the number of observations greater than \( k \), less two, and zero. The conditions on \( \phi(x) \) and \( h(x_i) \) are given in Section 3.2. When the number of observations greater than \( k \) is less than three, \( \hat{\lambda}(k) \) gives the same estimate as the MLE. When \( k=0 \), \( \hat{\lambda}(k) \) reduces to one of Peng's estimators.

We also propose estimators \( \hat{\lambda}^{[m]} \) which shift the MLE towards a point determined by the data itself. These adaptive estimators, which dominate the MLE uniformly, are expected to perform well for a wide range of parameter values, including the case when the parameters are all relatively large but similar in value. The estimator \( \hat{\lambda}^{[m]} \) is similar in form to \( \hat{\lambda}(k) \) with the functions \( h(x_i) \), \( \phi(x) \), and \( r_k \) replaced by appropriate ones.

In Section 4, we examine the simultaneous estimation problem from a different angle. In addition to normalized squared loss, we employ a more general loss function, namely, \( k \)-normalized square error loss \((k\text{-NSEL})\)
\[ L_k(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i^k}, \quad \text{where } k \text{ is a positive integer.} \]

The case when \( k=1 \), \( L_1 \), is simply the normalized squared error loss function under which Clevenson and Zidek [1975] perform their analysis.

The usual estimator \( \hat{X} \) of \( \lambda \), under \( L_1 \), is minimax. Using this fact, a sufficient condition that an estimator \( \hat{\lambda} \) of \( \lambda \) be minimax under \( L_1 \) is that its risk is less than or equal to that of \( \hat{X} \) uniformly in \( \lambda \). Theorem 2.2.9 thus provides us with a class of minimax estimators of \( \lambda \). A typical member of this class is of the form \( \hat{\lambda}(X) = (1 - \phi(Z)/(Z+p-1))X \) where \( \phi(z) \) is a nondecreasing real-valued function and \( 0 \leq \phi(z) \leq 2(p-1) \).
In the first part of this section, we shall show that this class of mini-
max estimators can be enlarged in two ways. First, we include estimators
of the form

\[ \hat{\lambda}(X) = (1 - \frac{\phi(Z)}{Z + \psi(Z)})X \]

where (1) \( \psi(z) \geq b > 0 \) for some \( b \)

(2) \( \phi(z) \) is nondecreasing and \( 0 \leq \phi(z+1) \leq 2\text{Min}\{p-1, \psi(z)\} \)

for all \( z \geq 0 \)

(3) \( \psi(z) \) is nonincreasing and \( z + \psi(z) \) is nondecreasing.

We next include estimators of the form

\[ \hat{\lambda}(X) = (1 - \frac{\phi_t(Z)}{(Z+b)^{t+1}})X \]

where (1) \( t \geq 0, b > t+1 \)

(2) \( \phi_t(z) \) is nondecreasing

(3) \( \phi_t(z) \geq 0 \) and \( \phi_t(z) \neq 0 \)

(4) \( \phi_t(z)/(z+b)^t \leq \text{Min}\{2(p-t-1), 2(b-t-1)\} \).

Moreover, when \( b = p-1 \), condition (4) can be replaced by

\[ \phi_t(z)/(z+p-1)^t \leq 2(p-t-1) \]

In the remainder of the section, we provide motivation for the use
of \( k\)-NSEL, \( L_k \), \( (k \geq 2) \) and derive estimators \( \hat{\lambda} = (\hat{\lambda}_1, ..., \hat{\lambda}_p) \) dominating
the MLE under \( L_k \). Basically, the estimators have the form

\[ \hat{\lambda}_i = X_i - \frac{\phi(Z)X_i(X_i-1) \cdots (X_i-k+1)}{\sum_{j=1}^{p} (X_j+1)(X_j+2) \cdots (X_j+k) + X_i(X_i-1) \cdots (X_i-k+1)} \]
where (1) $0 \leq \phi(z) \leq 2k(p-1)$

(2) $\phi(z)$ is nondecreasing in $z$.

Because the MLE $X$ is a Bayes estimator if prior knowledge of the intensity parameters is vague and the parameters are independent, it is hoped that substantial prior knowledge, when it is available, can be incorporated in a Bayesian manner to obtain significantly better estimators of $\lambda$ than $X$. In Section 5, we take a Bayesian approach to the problem of simultaneously estimating the $p$ intensity parameters when mean squared error is the loss criterion. The intensity parameters are assumed exchangeable in the sense of de Finetti [1964], and to jointly follow a two-parameter gamma distribution, a priori. For the second stage of the prior distribution we adopt a vague prior density for one of the gamma distribution parameters, and a generalized hypergeometric distribution for the other. The joint and marginal posterior densities of the Poisson intensity parameters are developed, and Bayesian point estimators are found.

In Section 6, we focus on estimators of $\lambda$ of the form $\hat{\lambda} = (1 - \hat{b}(Z))X$ and use empirical Bayes methods to perform analysis along the line suggested by Clevenson and Zidek [1975]. Although no new estimators are found in this section, the approach will be helpful in gaining insights about some of the estimators. We calculate the "relative savings loss" in the Poisson case under normalized square error loss and use it as a tool to obtain "plus rules" ($\hat{\lambda}^+ = (1 - \hat{b}(Z))_+X$) and "truncated Bayes rules". We also calculate the risk $R(\lambda, \lambda^S)$ of the Clevenson-Zidek estimators

$$\hat{\lambda}^S = [1 - (s(p-1) / (Z+p-1))]X$$
as a function of $\lambda$, where $0 \leq s \leq 2$. We find that $R(\lambda, \lambda^s)$, which is actually a function of $\Lambda = \sum_{i=1}^{p} \lambda_i$, is increasing and concave in $\Lambda$.

Section 7 contains the results of a computer simulation used to quantitatively compare the MLE with some of our estimators, as well as those of Clevenson and Zidek [1975] and Peng [1975]. For each estimator $\lambda$, the percentage of the savings in risk using $\lambda$ instead of the MLE is calculated as $\left(\frac{R(\lambda, X) - R(\lambda, \hat{\lambda})}{R(\lambda, X)}\right) \times 100\%$, using the loss function under which $\lambda$ was derived. In most of the cases, the improvement percentage is seen to be an increasing function of $p$, the number of distributions of the independent Poisson random variables. For the non-Bayes estimators, the improvement percentage generally decreases as the magnitude of the $\lambda_i$'s increases. The improvement percentage of the Bayes estimators depends on the choice of prior hyperparameters, and hence proper choice leads to substantial improvement over the MLE.

In Section 8, we return to the search for better estimators than the MLE under various loss functions, but alter the setting by allowing different numbers of observations to be taken from each population. We are led to consider weighted loss functions, an example of which is the generalized k-NSEL function $L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} c_i (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i^k$. More estimators dominating the MLE uniformly in $\lambda$ are proposed, and an application to estimation of the parameters of $p$ independent Poisson processes is provided.

Finally, Section 9 consists of proposals for further research.
SECTION 2. NOTATION AND FUNDAMENTALS

Let $X_1, \ldots, X_p$ be $p$ independent Poisson random variables with intensity parameters $\lambda_1, \ldots, \lambda_p$, respectively ($p \geq 2$). For easy reference, we introduce here some notation that will be used throughout this paper. We also include some basic lemmas and theorems which prove to be useful in ensuing sections.

2.1 Notation

Definitions

(1) "(Random variable name) ~ (Distribution name)" indicates that the random variable has the specified distribution with given parameter(s).

(2) $X = (X_1, \ldots, X_p); x = (x_1, \ldots, x_p)$ is the vector of observations of $X$;
   $\lambda = (\lambda_1, \ldots, \lambda_p); \Lambda = \sum_{i=1}^{p} \lambda_i$.

(3) $J$ = the set of all integers; $J^+$ = the set of all nonnegative integers.

(4) $f_i: J^p \rightarrow \mathbb{R}, i = 1, \ldots, p$, are functions from the $p$-fold Cartesian product $J^p$ of $J$ with itself to the set of real numbers $\mathbb{R}$.

(5) $f(X) = (f_1(X), \ldots, f_p(X)). \quad E_{\lambda} |f_i(X)| < \infty, i = 1, \ldots, p$.

(6) $Y^{(k)} = Y(Y-1) \ldots (Y-k+1)$, where $k$ is a positive integer and $Y$ is a real number.

(7) $Z = \sum_{i=1}^{p} X_i; z = \sum_{i=1}^{p} x_i$.

(8) $\phi(z)$ is a non-decreasing real-valued function.

(9) $[S]^w$ = the $w^{th}$ power of $S$. 


(10) \( e_i \) = the \( p \)-vector whose \( i \)th coordinate is one and the rest of whose coordinates are zero.

(11) \( R(\lambda, \hat{\lambda}) \) = the risk of the estimator \( \hat{\lambda} \) of \( \lambda \).

(12) \( NB(\mu, p) \) = the negative binomial distribution with probability mass function \( \Pr(z|b) = \binom{z+p-1}{z} b^p (1-b)^z, \ z \in J^+ \).

(13) \( \ell = \max \{x_i\}; m = \min \{x_i\} \).

(14) \( N_j = \# \{ x_i : x_i = j \} \).

(15) \( N = (N_0, \ldots, N_p) \).

(16) \( (y)_+ = \max \{ y, 0 \} \).

2.2 Basic Lemmas and Theorems

Let \( \hat{\lambda} \) be an estimator of \( \lambda \) and \( R(\lambda, \hat{\lambda}) \) be the risk of \( \hat{\lambda} \). Most of our estimators considered in subsequent sections will be of the form \( X + f(X) \) where \( f \) is as defined in subsection 2.1.

Define the risk improvement of \( \hat{\lambda} \) over the MLE, \( X \), to be

\[ I = R(\lambda,X) - R(\lambda,\hat{\lambda}). \]

Hudson [1974] furnishes a useful identity which we state as Lemma 2.2.1 below. Based on that identity, Hudson derives a basic identity about the unbiased risk improvement estimate under squared error loss: \( I = E \lambda U(X) \), where \( U \) is a function of \( X \) only. Peng [1975] defines risk deterioration \( D = -I = E \lambda (-U(X)) \) and shows that his estimators satisfy the inequality \( -U(x) \leq 0 \) for all \( x \) with strict inequality for some \( x \). This implies that the proposed estimators dominate the usual one under squared error loss.

The identity of the risk deterioration will be stated as Lemma 2.2.2 below.
Use of the same identity leads to the discovery of still more estimators of $\lambda$ better than the MLE under squared error loss. We shall show in Section 3 that to any fixed nonnegative integer $k$, there corresponds a class of estimators of $\lambda$ which shrinks the MLE towards $k$ as long as the number of variables is at least three. These estimators will give an estimate different from the MLE if the number of observations that is greater than $k$ is at least three; otherwise they give the same estimate as the MLE does. Adaptive estimators will also be proposed in Section 3.

**Lemma 2.2.1.** (Hudson [1974], Peng [1975]).

Suppose $Y \sim \text{Poisson } (\mu)$ and $G : R \to R$ is a measurable function such that $E_{\mu} |G(Y)| < \infty$ and $G(y) = 0$ if $y < 0$. Then $E_{\mu} G(Y) = E_{\mu} YG(Y-1)$.

The following lemma gives the unbiased estimate $\Delta_0$ of the deterioration in risk of $\hat{\lambda} = X + f(X)$ as compared to the risk of the MLE $X$.

**Lemma 2.2.2.** (Hudson [1974], Peng [1975]).

Suppose $X$ is a random vector with independent Poisson random variables as coordinates, and $\lambda$ is the corresponding Poisson parameter vector. Then the deterioration in risk $D$ of $\hat{\lambda} = X + f(X)$ as compared to the risk of the MLE is $D = R(\lambda, \hat{\lambda}) - R(\lambda, X) = E_{\lambda} \Delta_0$ where

$$\Delta_0 = \sum_{i=1}^{p} f_i^2(X) + 2 \sum_{i=1}^{p} X_i [f_i(X) - f_i(X - e_i)].$$

**Proof:** Use Lemma 2.2.1 (see Peng [1975]).

We see that a sufficient condition for an estimator of the form $X + f(X)$ to have smaller risk than $X$ (under squared error loss) is that
\[ \Delta_0 = \sum_{i=1}^{p} f_i^2(x) + 2 \sum_{i=1}^{p} x_i [f_i(x) - f_i(x - e_i)] \]

\[ \leq 0 \text{ for all } x \in J^p \] with strict inequality for some \( x \in J^p \).

The following theorem is due to Peng [1975].

**Theorem 2.2.3.**

Let \( X_1, \ldots, X_p \) be independent Poisson random variables with unknown expectations \( \lambda_1, \ldots, \lambda_p \), and let the loss function \( L \) be given by

\[ L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2. \]

The estimator

\[ X = [ (p-N_0-2)_+ / S ] H \]

dominates the MLE, \( X \) if \( p \geq 3 \). Here

\[ X = (X_1, \ldots, X_p), \]
\[ \lambda = (\lambda_1, \ldots, \lambda_p) \in [0, \infty]^p, \]
\[ H_i = \sum_{k=1}^{X_i} (1/k) \text{ for } i = 1, \ldots, p \text{ (} H_i = 0 \text{ if } X_i = 0), \]
\[ H = (H_1, \ldots, H_p), \]
\[ S = ||H||^2 = \sum_{i=1}^{p} H_i^2, \]
\[ N_0 = \# \{ X_i : X_i = 0 \} = \text{the number of zero observations, and } (p-N_0-2)_+ = \text{Max } \{p-N_0-2,0\}. \]

**Proof:** Use Lemma 2.2.2 (see Peng [1975]).
In addition to the squared error loss function, we shall consider other loss functions. We derive an identity similar to that of Lemma 2.2.2 for the case of "k-normalized squared error loss" (k-NSEL),

$$L_k(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} \left( \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i} \right)^k$$

where $k$ is a positive integer. The motivation for considering k-NSEL is discussed in Section 4.3. The derivation of the identity will be decomposed into the following four lemmas.

**Lemma 2.2.4.**

Suppose $Y \sim$ Poisson ($\mu$) and $h: J \rightarrow \mathbb{R}$ is a real-valued function such that

1. $E_{\mu} |h(Y)| < \infty$
2. $h(j) = 0$ if $j < 0$.

Then

$$E_{\mu} \frac{h(Y)}{\mu} = E_{\mu} \frac{h(Y+1)}{Y+1}.$$

**Proof:**

$$E_{\mu} \frac{h(Y)}{\mu} = \sum_{y=0}^{\infty} \frac{h(y)}{\mu} \cdot \frac{e^{-\mu}}{y!} \cdot \frac{\mu^y}{y!}$$

$$= \frac{h(0)e^{-\mu}}{\mu} + \sum_{y=1}^{\infty} \frac{h(y)}{\mu} \cdot \frac{e^{-\mu}}{y!} \cdot \frac{\mu^y}{y!}$$

$$= 0 + \sum_{y=0}^{\infty} \frac{h(y+1)}{(y+1)!} \cdot \frac{e^{-\mu}}{y!} \cdot \frac{\mu^{y+1}}{y+1!}$$

$$= \sum_{y=0}^{\infty} \frac{h(y+1)}{y+1} \cdot \frac{e^{-\mu}}{y!} \cdot \frac{\mu^y}{y!}$$

$$= E_{\mu} \frac{h(Y+1)}{Y+1}.$$

Q.E.D.

The next lemma is an immediate consequence of Lemma 2.2.4.
Lemma 2.2.5.

Suppose $Y \sim \text{Poisson} (\mu)$, $k$ is a positive integer and $h : J \rightarrow \mathbb{R}$ is a real-valued function such that

1. $E_{\mu} |h(Y+j)| < \infty$, $j = 0, \ldots, k-1$
2. $h(j) = 0$ if $j < k$.

Then

$$E_{\mu} \frac{h(Y)}{\mu^k} = E_{\mu} \frac{h(Y+k)}{(Y+k)(k)} = E_{\mu} \frac{h(Y+k)}{(Y+k)(Y+k-1) \cdots (Y+1)}.$$ 

Proof: Induction on $k$ and application of Lemma 2.2.4.

Lemma 2.2.6 below is a generalization of Lemma 2.2.2 to the vector case.

Lemma 2.2.6.

Suppose $X_i \sim \text{Poisson} (\lambda_i)$, $i = 1, \ldots, p$, $p \geq 2$, and $k$ is a positive integer. If $f_i : J^p \rightarrow \mathbb{R}$, $i = 1, \ldots, p$, are functions on the $p$-fold Cartesian product of $J$, such that

1. $E \lambda |f_i(X + je_i)| < \infty$, $j = 0, \ldots, k-1$
2. $f_i(x) = 0$ if $x_i < k$,

then

$$E_{\lambda} \frac{f_i(X)}{\lambda_i^k} = E_{\lambda} \frac{f_i(X + ke_i)}{(X_i + k)(k)}.$$ 

Proof: Condition on $X_j : j \neq i$ and apply Lemma 2.2.5.

Let $\hat{\lambda} = X + f(X)$ be an estimator of $\lambda$, where $f(X) = (f_1(x), \ldots, f_p(X))$ and the $f_i$'s satisfy the conditions in Lemma 2.2.6. The next lemma gives an unbiased estimate of $D_k$, the deterioration in risk of $\hat{\lambda}$ as compared to the risk of $X$. 

Lemma 2.2.7.

Under k-NSEL, the deterioration in risk of \( \hat{\lambda} \) is

\[
D_k = R(\lambda, \hat{\lambda}) - R(\lambda, X) = E_{\lambda} \Delta_k.
\]

where

\[
\Delta_k = \sum_{i=1}^{p} \frac{f_i^2(X + k\epsilon_i)}{(X_i + k)^k} + 2 \sum_{i=1}^{p} \frac{f_i(X + k\epsilon_i) - f_i(X + (k-1)\epsilon_i)}{(X_i + k)^{(k-1)}}.
\]

Proof:

\[
R(\lambda, \hat{\lambda}) = E_{\lambda} \sum_{i=1}^{p} \frac{(\lambda_i - X_i - f_i(X))^2}{\lambda_i^k}
\]

\[
= E_{\lambda} \sum_{i=1}^{p} \frac{(\lambda_i - X_i)^2}{\lambda_i^k} + X f_i(X) - 2E_{\lambda} \sum_{i=1}^{p} \frac{f_i(X)}{\lambda_i^k} - 2E_{\lambda} \sum_{i=1}^{p} \frac{(X_i - \lambda_i) f_i(X)}{\lambda_i^k}
\]

\[
= R(\lambda, X) + E_{\lambda} \sum_{i=1}^{p} \frac{f_i^2(X)}{\lambda_i^k} + 2E_{\lambda} \sum_{i=1}^{p} \frac{X f_i(X)}{\lambda_i^k} - 2E_{\lambda} \sum_{i=1}^{p} \frac{f_i(X)}{\lambda_i^{k-1}}
\]

\[
= R(\lambda, X) + E_{\lambda} \left[ \sum_{i=1}^{p} \frac{f_i^2(X + k\epsilon_i)}{(X_i + k)^k} + 2 \sum_{i=1}^{p} \frac{f_i(X + k\epsilon_i) - f_i(X + (k-1)\epsilon_i)}{(X_i + k)^{(k-1)}} \right].
\]

The last equality follows from Lemma 2.2.6. The result follows immediately.

Q.E.D.

The special case of Lemma 2.2.7 when \( k = 1 \) is the case when normalized squared error loss is considered. The result is stated as a corollary below.
Corollary 2.2.8.

Under the normalized squared error loss function, the deterioration in risk of $\hat{\lambda}$ is $D_1 = R(\hat{\lambda}, \lambda) - R(\lambda, X) = E_\lambda \Delta_1$, where

$$\Delta_1 = \sum_{i=1}^{p} \frac{f_i^2(X+e_i)}{X_i+1} + 2 \sum_{i=1}^{p} [f_i(X+e_i) - f_i(X)].$$

The following theorem of Clevenson and Zidek [1975] provides a class of estimators $\hat{\lambda}$ of $\lambda$ dominating the MLE under the normalized squared error loss function.

Theorem 2.2.9.

Let $X_1, \ldots, X_p$ be independent Poisson random variables with unknown parameters $\lambda_1, \ldots, \lambda_p$ ($p \geq 2$). Let $\lambda = (\lambda_1, \ldots, \lambda_p)$, $X = (X_1, \ldots, X_p)$, and $Z = \sum_{i=1}^{p} X_i$. Let the loss function be the normalized squared loss function

$$L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2/\lambda_i.$$ 

Then, for all $\lambda$, the risk using the estimator

$$\hat{\lambda}^* = [1 - (\phi(Z)/(Z+p-1))]X$$

is less than or equal to the risk using $\lambda^0 = X$ when $\phi : [0, \infty) \rightarrow [0, 2(p-1)]$ is nondecreasing and not identically zero.

Proof:

Let $f_i(x) = -\phi(z)x_i/(z+p-1)$ if $x_i > 0$

$$= 0 \hspace{1cm} \text{if } x_i < 0, \ i = 1, \ldots, p.$$
By Corollary 2.2.8, the deterioration in risk is

\[ \Delta_1 = \sum_{i=1}^{p} \frac{\phi^2(z+1)(x_i+1)}{(z+p)^2} - 2 \sum_{i=1}^{p} \frac{\phi(z+1)(x_i+1)}{z+p} + 2 \sum_{i=1}^{p} \frac{\phi(z)x_i}{z+p-1} \]

\[ \leq \frac{\phi^2(z+1)}{z+p} - 2\phi(z+1) + 2 \frac{\phi(z+1)z}{z+p-1} \quad \text{(since } \phi \text{ is nondecreasing)} \]

\[ = \phi(z+1) \left( \frac{\phi(z+1)}{z+p} - \frac{2(p-1)}{z+p-1} \right) \]

\[ \leq 0 \quad \text{(since } 0 \leq \phi(z+1) \leq 2(p-1)) \].

Therefore, \( R(\lambda, X) \geq R(\lambda, \hat{\lambda}) \) for all \( \lambda \). In other words, \( \hat{\lambda} \) dominates \( X \) uniformly in \( \lambda \).

Q.E.D.

In Section 4, we propose estimators \( \hat{\lambda} \) of the form \( X + f(X) \) which satisfy the inequality \( \Delta_k \leq 0 \) for all \( x \) and \( \Delta_k \neq 0 \). Those estimators therefore dominate the MLE under k-NSEL.
SECTION 3. ESTIMATION UNDER SQUARED ERROR LOSS

3.1 Introduction

Probably the most extensively studied loss function used in estimation problems is the squared error loss function. Its popularity can be ascribed to its mathematical tractability and also to the fact that "it is an acceptable approximation in a wide variety of situations" (DeGroot [1970], p. 228) where the loss depends solely on the difference between a parameter and its estimate. If we allow $k$ to be equal to zero, the squared error loss function can be thought of as a special case of $k$-NSEL.

This section will be limited in scope to simultaneous estimation of Poisson parameters under the squared error loss function. The setting is as described previously. We suppose that $X_i \sim \text{Poisson} (\lambda_i), \ i = 1, \ldots, p$ and that one observation is taken from each population. As mentioned in Sections 1 and 2, this problem has been investigated by Peng [1975], who succeeded in finding estimators dominating the MLE when $p \geq 3$ (Theorem 2.2.3). Basically, his proposed estimators pull the MLE towards the origin whenever the number of non-zero observations exceeds two. The performance of his estimators is expected to be good when the underlying parameters $\lambda_i$ are relatively small. When some of the parameters are large, however, very little improvement over the MLE is anticipated. In this situation, some very large observations are likely to occur and both Peng's estimator (Theorem 2.2.3) and the MLE give virtually the same estimate. In order to remedy this situation, Peng [1975] uses Stein's method [1974] to modify his estimator.

If all the parameters $\lambda_i$ are relatively large, none of the estimators proposed by Peng will give noticeable improvement over the MLE. Basically,
this is due to the fact that those estimators are biased toward the origin, a point far away from the true $\lambda$. Estimators that shift the observations towards a point in a neighbourhood of the true underlying parameter would be expected to give better estimates in this case. For each non-negative integer $k$, we show that there is a family of estimators $\hat{\lambda}^{(k)}$ of $\lambda$ such that $\hat{\lambda}^{(k)}$ dominates the MLE uniformly in $\lambda$ under the squared error loss function. For a fixed $k$, the estimator $\hat{\lambda}^{(k)}$ has the property that it pulls the MLE towards the integer $k$ whenever the number of observations $x_i$ greater than $k$ is at least three. Otherwise, it gives the same estimate as the MLE. In the case when $k=0$, we have Peng's result.

Estimators that shift the MLE towards a point determined by the data itself will also be proposed. These adaptive estimators are expected to perform well for a wide range of parameter values, including the case when the parameters are all relatively large but similar in value. Most of the results of Peng [1975] will be generalized. In the next subsection, we will introduce the notation that is used in this section. Essentially, we employ the notation used by Peng [1975].

3.2 Notation

Let $x = (x_1, \ldots, x_p)$ be a vector of observations of the random vector $(X_1, \ldots, X_p)$, where the $X_i$'s are mutually independent Poisson random variables with parameters $\lambda_1, \ldots, \lambda_p$, respectively. Let $f = (f_1, \ldots, f_p)$ be as defined in Section 2.1 and satisfy the following conditions:

1. $f_i(x) = 0$ if $x$ has a negative coordinate

2. $E_\lambda |f_i(X + j e_i)| < \infty$ for $j = 0, 1, \ldots$

The notation employed here is listed below for reference.
Definitions:

(1) \( N_j = \#\{x_i : x_i = j\} \), i.e. the number of \( x_i \)'s that are equal to \( j \).

(2) \( \ell = \max_{i=1}^{p} x_i; \ m = \min_{i=1}^{p} x_i \).

(3) \( N = (N_0, \ldots, N_k) \).

(4) For any non-negative integer \( k \), \( r_k = \sum_{n=0}^{k} n^{2} \).

(5) \( h : J \to \mathbb{R} \) is a real-valued function such that \( h(y) = 0 \) if \( y < 0 \).

(6) \( S = \sum_{i=1}^{p} h^2(x_i); \ \bar{S} = \sum_{i=1}^{p} \frac{h^2(x_i)}{n} \).

(7) Write \( f_j(x) = \psi_j(N) \) if \( x_i = j \).

(8) \( \phi : J^p \to \mathbb{R} \) is a real-valued function satisfying the following properties:
   (i) \( \phi \) is nondecreasing in each argument \( x_i \)
       whenever \( x_i \geq k \).
   (ii) \( \phi \) is nonincreasing in each argument \( x_i \)
        whenever \( x_i \leq k \).
   (iii) There is a real number \( B > 0 \) such that
         \( 0 < \phi(x) \leq 2B \) and \( \phi(x) \neq 0 \).

We are interested in functions \( h \) which satisfy the properties listed in Lemma 3.2.2 below. Before stating the lemma, we provide a representative example of the functions \( h \) we want.
Example 3.2.1.

(a) For \( k \geq 2 \)

\[
h(y) = 1 + \sum_{n=2}^{y-k} \frac{1}{k+n} \quad \text{if } y = k+2, k+3, \ldots
\]

\[
= 1 \quad \text{if } y = k+1
\]

\[
= 0 \quad \text{if } y = k \text{ or } y < 0
\]

\[
= -b \sum_{n=1}^{k-y} \frac{1}{k+1-n} \quad \text{if } y = 0, \ldots, k-1
\]

where \( b \) is a positive number to be determined such that (8) of Lemma 3.2.2 below holds. One such \( b \) is \( b = \sqrt{3} \left( \sum_{n=1}^{k} \frac{1}{k+1-n} \right)^{-1} \).

(b) For \( k = 1 \)

\[
h(y) = 1 + \sum_{n=2}^{y-k} \frac{1}{k+n} \quad \text{if } y = k+2, k+3, \ldots
\]

\[
= 1 \quad \text{if } y = k+1
\]

\[
= 0 \quad \text{if } y = k \text{ or } y < 0
\]

\[
= -b \quad \text{if } y = 0
\]

where \( b \) is any positive number.

(c) For \( k = 0 \)

\[
h(y) = \sum_{n=1}^{y} \frac{1}{n} \quad \text{if } y = 1, 2, \ldots
\]

\[
= 0 \quad \text{if } y < 0.
\]

The following lemma gives the properties of \( h \). For simplicity, we denote \( h_j = h(j) \).
Lemma 3.2.2.

Let $h$ be as defined in Example 3.2.1. Then $h$ satisfies the following properties:

1. $h_j^2 - h_{j-1}^2$ is nonincreasing in $j$ for $j > k+1$.
2. $j[h_j - h_{j-1}]$ is nondecreasing in $j$ for $j > k$ and
   \[ \lim_{j \to \infty} j[h_j - h_{j-1}] = B \] for some $B > 0$.
3. $h_j > h_{j-1}$, $j = 1, 2, \ldots$
4. $h_k = 0$.
5. $h_j > 0$ if $j \geq k+1$.
6. If $k > 0$, then $h_j < 0$ for $j < k$.
7. $h_{k+1} = B$.
8. $3B^2 \times h_0$ provided $k > 0$.
9. $h_{k+1} \geq j[h_j - h_{j-1}]$ for $j \geq k+2$.
10. $h_j^2 - h_{j-1}^2 \leq h_{j+1}^2 - h_j^2$ for $1 \leq j < k$ if $k \geq 0$.

The proof of the lemma is straightforward and is omitted.

3.3 Shifting the MLE Towards $k$

We use the notation defined in 3.2 and define

\[
  f_i(x) = -\frac{r_i \phi(x) h(x_i)}{s_i}, \quad i = 1, \ldots, p. \tag{3.3.1}
\]

We shall show that the estimator $\hat{\lambda}^{(k)} = X + f(X)$ of $\lambda$ dominates $X$ uniformly in $\lambda$ under the squared error loss function when $p \geq 3$. The
estimator $\hat{\lambda}^{(k)}$ shifts the coordinates of the MLE towards the integer $k$ provided the number of observations greater than $k$ is at least three.

Recall that the risk deterioration of the estimator $\hat{\lambda} = X + f(X)$ as compared to $X$ is, by Lemma 2.2.2,

$$R(\hat{\lambda}, \lambda) - R(\lambda, X) = E_\lambda \Delta$$

where

$$\Delta = \sum_{i=1}^{P} f_i^2(X) + 2 \sum_{i=1}^{P} X_i [f_i(X) - f_i(X-e_i)].$$

(3.3.2)

In terms of $N$ and $\psi_j(N)$, $\Delta$ can be rewritten as

$$\Delta = \sum_{j=0}^{l} N \psi_j(N) + 2 \sum_{j=1}^{l} jN_j [\psi_j(N) - \psi_{j-1}(N - \delta_j + \delta_{j-1})]$$

(3.3.3)

where $\delta_j$ is an $(l+1)$-vector with the $j^{th}$ coordinate equal to one and the other coordinates zero. To show that $\hat{\lambda}^{(k)} = X + f(X)$ dominates $X$ under the squared error loss, it suffices to show that $\Delta(x) \leq 0$ for all $x \in J^+$. The proof will use the following series of lemmas. For convenience, we define

$$A_j = jN_j [\psi_j(N) - \psi_{j-1}(N - \delta_j + \delta_{j-1})]$$

(3.3.4)

$$= x_i [f_i(x) - f_i(x-e_i)], \text{ with } x_i = j).$$

Lemma 3.3.5.

1. For $k \geq 2$,

$$A_j \leq - \frac{j[h_i - h_{i-1}] \phi(x) r_k N_j}{S} \left[ \frac{S - h_j (h_j + h_{j-1})}{S - h_j^2 + h_{j-1}^2} \right] \text{ if } 1 \leq j < k.$$  

2. $A_k \leq \frac{\phi(x) k N_k r_k h_{k-1}}{S + h_{k-1}^2}.$
(3) \[ A_j \leq -\frac{j[h_j-h_{j-1}]}{S} r_k \phi(x) N_j \left[ \frac{S - h_j (h_j + h_{j-1})}{S - h_j^2 + h_{j-1}^2} \right] \]

\[ \leq -\frac{\ell[h_k-h_{k-1}]}{S} r_k \phi(x) N_j \left[ \frac{S-2h_j^2}{S - h_k^2 + h_{k-1}^2} \right] \] for \( k+2 \leq j \leq \ell \).

**Proof:**

(1) is true because \( \phi(x) \) is nonincreasing for \( x_1 \leq k, i = 1, \ldots, p \).

(2) is true because \( \phi(x) \) is nonincreasing for \( x_1 \leq k \) and \( h_{k-1} < 0 \).

(3) is true for the following reasons:

(i) If \( r_k > 0 \), then \( S > 2h_j^2 \) for \( k+2 \leq j \leq \ell \) and \( N_j \neq 0 \).

(ii) \( h_j^2 - h_{j-1}^2 \) is nonincreasing by (1) of Lemma 3.2.2.

(iii) \( h_j > h_{j-1} \).

(iv) \( \phi(x) \) is nondecreasing when \( x_1 \geq k, i = 1, \ldots, p \).

Q.E.D.

**Lemma 3.3.6**

\[ 2h_{k+1}^2 \geq h_k^2 - h_{k-1}^2 \] where \( \ell \geq k+2 \).

**Proof:**

By (9) of Lemma 3.3.2, \( h_{k+1} \geq j[h_j - h_{j-1}] \) for \( j \geq k+2 \). Also, since \( h_j - h_{j-1} \) is decreasing,

\[ \frac{h_j + h_{j-1}}{j} \leq \frac{2h_j}{j} \leq \frac{2[h_{k+1} + (h_{k+2} - h_{k+1}) + \cdots + (h_1 - h_{j-1})]}{j} \]

\[ \leq \frac{2(j-k)h_{k+1}}{j} \leq 2h_{k+1}. \]

Hence \( 2h_{k+1}^2 \geq j[h_j - h_{j-1}] \frac{[h_j + h_{j-1}]}{j} \)
\[ \begin{aligned}
&= h_j^2 - h_{j-1}^2 \quad \text{for } j \geq k+2. \\
\text{Q.E.D.}
\end{aligned} 

Lemma 3.3.7.

\[ \{N_{k+1} + \sum_{j=k+2}^{N} \frac{S - 2h_j^2}{S - h_{k+1}^2 + h_j^2 - 1} \} \geq r_k, \text{ if } r_k > 0. \]

Proof:

The left hand side of the above inequality is

\[
N_{k+1} + \sum_{j=k+2}^{N} \frac{S - 2h_j^2}{S - h_{k+1}^2 + h_j^2 - 1}
\]

\[
= N_{k+1} \sum_{j=k+2}^{N} \frac{N_j}{S - h_{k+1}^2 + h_j^2 - 1} + \sum_{j=k+2}^{N} \frac{S - 2h_j^2}{S - h_{k+1}^2 + h_j^2 - 1}
\]

\[
\geq \frac{r_k}{S - h_{k+1}^2 + h_j^2 - 1} \quad \text{(since } 2h_{k+1}^2 \geq h_j^2 - h_{k+1}^2 \text{ by Lemma 3.3.6)}
\]

\[
\geq r_k \quad \text{(since } r_k > 0). \\
\]

\text{Q.E.D.}

Theorem 3.3.8.

\[
\text{With } f_i(x) = -\frac{r_i \phi(x)h(x)}{S}, \quad i = 1, \ldots, p, \quad p \geq 3, \quad \text{and } h \text{ as described in Lemma 3.2.2, } \Delta \leq 0.
\]
Proof:

Recall that \( \Delta = \sum_{j=0}^{\ell} N_j \psi_j^2(N) + 2 \sum_{j=1}^{\ell} jN_j \psi_j(N) - \psi_{j-1}(N-\delta_j+\delta_{j-1}) \).

Case 1. \( r_k = 0 \). Then \( \Delta \equiv 0 \leq 0 \).

Case 2. \( r_k > 0 \).

(i) \( \sum_{j=0}^{\ell} N_j \psi_j^2(N) = \frac{r_k^2 \phi(x)}{S} \)

(ii) \( 2 \sum_{j=1}^{k-1} A_j = 2 \sum_{j=1}^{k-1} A_j + 2A_k + 2A_{k+1} + 2 \sum_{j=k+2}^{\ell} A_j \)

with the understanding that \( \sum_{j=1}^{k-1} A_j = 0 \) if \( k \leq 1 \).

By Lemma 3.3.5,

\[
2 \sum_{j=1}^{k-1} A_j \leq - \frac{\phi(x) r_k}{S} \sum_{j=1}^{k-1} j[h_j-h_{j-1}]N_j \left[ \frac{S - h_j(h_j+h_{j-1})}{S - h_j^2 + h_{j-1}^2} \right]
\]

since \( h_j < 0 \) for \( j < k \) and \( h_j > h_{j-1} \), so \( h_jh_0 > h_jh_{j-1} \) for \( 1 \leq j \leq k-1 \).

Now \( r_k > 0 \) implies that \( S \geq 3h_{k+1}^2 > 3B^2 \). By (8) of Lemma 3.2.2, we have \( S - h_j^2 - h_jh_0 > 0 \) for \( N_j \neq 0 \). Hence \( \sum_{j=1}^{k-1} A_j \leq 0 \). Again by Lemma 3.3.5,

\[
A_k \leq \frac{\phi(x) kN_k r_k h_{k-1}}{S + h_{k-1}^2} \quad \text{and} \quad A_{k+1} = - \frac{(k+1)N_k r_k \phi(x) h_{k+1}}{S}
\]

(because \( h_k = 0 \)).
As a result,
\[
2 \sum_{j=1}^{\ell} A_j \leq \frac{2kN r_k \phi(x) h_{k-1}}{S + h_{k-1}^2} - \frac{2r_k \phi(x)}{S} \left[ (k+1)N_{k+1} h_{k+1} + \ell [h_{k} - h_{k-1}] \right] \sum_{j=k+2}^{\ell} N_j \frac{S - 2h_j^2}{S - h_{k+1}^2 + h_{k-1}^2}.
\]

(by Lemma 3.3.5. (3)). Thus,
\[
2 \sum_{j=1}^{\ell} A_j \leq \frac{2kN r_k \phi(x) h_{k-1}}{S + h_{k-1}^2} - \frac{2\phi(x) r_k N_{k+1} h_{k+1}}{S} - \frac{2r_k \phi(x)}{S} \left[ h_{k+1} + \ell [h_{k} - h_{k-1}] \right] \sum_{j=k+2}^{\ell} N_j \frac{S - 2h_j^2}{S - h_{k+1}^2 + h_{k-1}^2}.
\]

(since \( h_{k+1} \geq \ell [h_{k} - h_{k-1}] \))
\[
\leq \frac{2kN r_k \phi(x) h_{k-1}}{S + h_{k-1}^2} - \frac{2\phi(x) r_k N_{k+1} h_{k+1}}{S} - \frac{2r_k \phi(x) B}{S}.
\]

(by Lemma 3.3.7 and since \( \ell [h_{k} - h_{k-1}] \geq B \)).

Consequently,
\[
\Delta \leq \frac{2kN r_k \phi(x) h_{k-1}}{S + h_{k-1}^2} - \frac{2kN r_k \phi(x) h_{k+1}}{S} - \frac{r_k^2 \phi(x)}{S} \left[ 2B - \phi(x) \right] \quad (3.3.9)
\]

< 0 since \( 0 \leq \phi(x) \leq 2B, h_{k-1} < 0, \) and \( h_{k+1} > 0. \)

Q.E.D.
Our results show that to every fixed non-negative integer $k$, there corresponds a class of estimators $\hat{\lambda}(k)$ of $\lambda$ which has the property that all its members pull the coordinates of the MLE towards $k$ when the number of observations that are greater than $k$ (i.e. $\sum_{j=k+1}^{l} N_j$) is at least three. If the number of observations greater than $k$ is less than three, then the estimators $\hat{\lambda}(k)$ give the same estimate as the MLE $\lambda$. The choice of $k$ is critical. It may depend on prior information available about the range of the $\lambda_i$'s. When $k=0$, the estimators $\hat{\lambda}(0)$ shrink all the non-zero observations towards zero as long as the number of non-zero observations is at least three. Such a value of $k$ should be chosen only when we have some prior knowledge about the $\lambda_i$'s indicating that they are all close to zero. $\hat{\lambda}(0)$ is expected to perform well in this situation, especially when all the $\lambda_i$'s are close to one another. If the prior information suggests that the $\lambda_i$'s are likely to be large and within the range $[a,b]$ with $0 < a < b$, then a large value of $k$ somewhere around $a$ may be chosen. Choosing $k$ in this manner will likely lead to considerable improvement over the MLE. Hence, having some prior knowledge about the parameters $\lambda_i$ is advantageous to the estimation problem. In fact, we see that, according to our simulation results reported in Section 7, our Bayesian analysis in Section 5 does lead to substantial improvement over the usual procedure when the parameters are known to be close to one another.

Notice that the bound (3.3.9) for the unbiased estimate $\Delta$ of the deterioration in risk of $\hat{\lambda}(k)$ depends on $k$, $N_k$ and $N_{k+1}$. Hence, savings in risk would be greater if $N_k$ and $N_{k+1}$ are large. In other words, when more observations are close to the chosen integer $k$ the risk will be reduced much more. One might say that reliable prior information
can be profitably exploited in our estimation problem. The dependency of
the bound for $\Delta$ on $k$, $N_k$, and $N_{k+1}$ further implies that the estimators
$\hat{l}(k)$ for various $k \in J^+$ are competitive; one cannot dominate the other.

In the case of simultaneous estimation of $p$ normal means, the
existence of an estimator which shrinks the MLE towards zero and
dominates the MLE implies the existence of another estimator shrinking
towards any fixed point and still dominating the MLE. This is due to
the translation invariance of the normal density and the squared error
loss function. However, in our case, we do not have this invariance
property for the Poisson probability function, and hence the existence
of better estimators shrinking towards a point other than zero is not
automatic even though a better estimator shrinking towards zero exists.
Thus our results are not obvious consequences of Peng's.

Since our estimators depend on $h$, it is interesting to find more
examples of functions $h$ which have the properties in Lemma 3.2.2. Below
are some examples.

**Example 3.3.10.** ($k = 0$)

Let $h(y) = \ln(ay)$ if $y > 1$

$= 0$ if $y < 1$

where $a \geq 4$. Let $B = 1$. We check below that this function $h$ satisfies
the properties listed in Lemma 3.2.2.

(a) In order to show that property (1) holds, it is sufficient
to show that $h^2_j \geq \frac{1}{2} (h^2_{j+1} + h^2_{j+1})$ for $j \geq 2$. This is certainly true since
the function $G(y) = [\ln(ay)]^2$ is concave for $y \geq 1$.

(b) Property (2): The function $F(y) = y[\ln(ay) - \ln a(y-1)]$

$= y[\ln y - \ln(y-1)]$ is decreasing when $y \geq 2$ since the derivative
\[ F'(y) = \ln(1 + \frac{1}{y-1}) - \frac{1}{y-1} \leq 0 \text{ for } y \geq 2. \] Also, \( h_1 - h_0 = \ln a \geq 2[\ln(2a) - \ln a] \) whenever \( a \geq 4 \). Moreover, \( \lim_{j \to \infty} j[\ln j - \ln(j-1)] = 1 = B. \)

(c) Properties (3), (4), (5), (7), and (9) clearly hold, and properties (6), (8), and (10) are not applicable here because \( k = 0 \) in this example.

Example 3.3.11. \((k = 0)\)

Let \( h : J \to \mathbb{R} \) be any function such that \( h_j = 0 \) if \( j \leq 0 \) and
\[
h_j = -\sum_{n=1}^{j} \frac{1}{g_n} \quad \text{for } j = 1, 2, \ldots \quad \text{Here } \{g_n\} \text{ is a sequence of real numbers satisfying}
\]

(1) \( g_1 = 1 \)

(2) \( g_{n+1} - g_n > 1, \) for \( n = 1, 2, \ldots \)

(3) \( \{\frac{n}{g_n}\} \text{ is nonincreasing and } \lim_{j \to \infty} \frac{1}{g_j} = B > 0. \)

Properties (1) through (10) of Lemma 3.2.2 are checked as follows:

(a) Property (1) holds since
\[
2h_{j}^2 - (h_{j-1}^2 + h_{j+1}^2) = \frac{1}{g_j g_{j+1}} \left( 2[g_{j+1} - g_j] \left( \sum_{n=1}^{j-1} \frac{1}{g_n} \right) - 2 + \frac{g_{j+1}}{g_j} - \frac{g_1}{g_{j+1}} \right) \geq 0
\]
for \( j \geq 2. \)

(b) Property (2) holds since \( \{g_n\} \) satisfies requirement (3) above.

(c) All the other properties clearly hold.

Property (8) of \( h \) given in Lemma 3.2.2 guarantees that
\[
S - h_j(h_j + h_{j-1}) \geq 0 \text{ for } j < k, \text{ which is a sufficient condition that }
\]
\[
\sum_{j=k-1}^{k-1} A_j \leq 0. \text{ However, it is not a necessary condition, as the following theorem shows.}
\]
Theorem 3.3.12:

Let $h : J \to \mathbb{R}$ be as described in Lemma 3.2.2 except that properties (3) and (8) are replaced by

$$(3)' \quad h_j > h_{j-1} \text{ for } j \geq k+1$$

$$(8)' \quad h_j = -b < 0 \text{ for } j = 0, \ldots, k-1.$$  

Then Theorem 3.3.8 still holds.

Proof: The change still gives $\sum_{j=1}^{k-1} A_j \leq 0$.

The following theorem is a slight variation of Theorem 3.3.12.

Theorem 3.3.13.

Let $h$ be as described in Theorem 3.3.12. Suppose

$$0 \leq \phi(x) \leq \min\{2B, 1\} \text{ if } x_i < k, \ i = 1, \ldots, p.$$  

Define

$$f_i(x) = \begin{cases} \frac{r_k \phi(x) h(x_i)}{S} & \text{if } x_i > k \\ 0 & \text{if } x_i = k \\ \frac{br_k}{S} \phi(x) \min\{\frac{br_k}{S}, 1\} & \text{if } x_i < k \end{cases}$$

for $i = 1, \ldots, p$.

Then the estimator $\hat{\lambda}(k) = X + f(X)$ satisfies $\Delta(x) \leq 0$ for all $x \in J^p$ (i.e. $\hat{\lambda}(k)$ dominates $X$ uniformly under the squared error loss function).

Proof:

It can be checked that $\sum_{j=1}^{k-1} A_j \leq 0$ for $k > 0$ and that

$$\sum_{j=0}^{l} \sum_{j=0}^{j} \frac{r_k^2 \phi^2(x)}{S}.$$  

Hence $\Delta \leq 0$ as shown in Theorem 3.3.8.

Q.E.D.

Note that estimators $\hat{\lambda}(k)$ of $\lambda$ given in Theorems 3.3.8 and 3.3.12 have the property that they pull the $x_i$'s that are farther away from $k$
more than those \( x_i \)'s that are closer to \( k \). This means that the extreme observations experience a great deal of shifting, while the observations close to \( k \) are shifted to a lesser extent.

There is still another choice of \( h \) that will guarantee \( \Delta \leq 0 \).

**Theorem 3.3.14.**

Let \( h \) be as described in Lemma 3.2.2 except that property (8) is replaced by

\[
(8)'' \quad 3h_{k+1}^2 > h_1 h_0.
\]

Then Theorem 3.3.8 still holds. (In this case, \( |h_0| \) has a larger upper bound).

**Proof:** Note that \( \sum_{j=1}^{k-1} A_j \) is still less than zero in this case.

An example of a function \( h \) described in Theorem 3.3.14 is given below.

**Example 3.3.15.** \((k \geq 2)\)

Let

\[
h_j = \begin{cases} 
\sum_{n=1}^{j} \frac{1}{n} & \text{if } j > k+1 \\
0 & \text{if } j = k \\
-\sum_{n=1}^{k-j} \frac{1}{k+1-n} & \text{if } 0 \leq j < k.
\end{cases}
\]

The estimators \( \hat{\lambda}^{(k)} \) derived thus far have the property that if the \( i \text{th} \) observation is equal to \( k \), then \( \hat{\lambda}_i^{(k)} = k \). That is, there is no shifting of the observations having the value \( k \). The next theorem provides an estimator of \( \lambda \) which improves on the MLE but whose estimate of \( \lambda_i \) is not necessarily equal to \( k \) if the \( i \text{th} \) observation is equal to \( k \).

The theorem unifies and generalizes Theorems 3.1 and 5.1 of Peng [1975].
Theorem 3.3.16.

Let \( h \) be as described in Theorem 3.3.12 except that properties (4) and (7) of \( h \) are replaced by

\[(4)' \quad h_k = -b \]
\[(7)' \quad h_{k+1} \geq \text{Max}\{1, B\} \]

Define

\[ f_i(x) = -\frac{r_k \phi(x)h(x_i)}{S} \quad \text{if } x_i > k \]
\[ = \phi(x) \text{Min}\left\{ \frac{br_k}{S}, 1 - \frac{r_kh_{k+1}}{S} \right\} \quad \text{if } x_i < k \quad \text{for } i = 1, \ldots, p. \]

Suppose \( 0 < \phi(x) \leq \text{Min}\{1, 2B\} \) if \( x_i < k, i = 1, \ldots, p \), and let

\[ \lambda(k)' = X + f(X). \]

Then \( \Delta(x) \leq 0 \) for all \( x \in J^+ \).

The proof of the theorem is similar to that of Theorem 3.3.8. Recall that we denote \( \psi_j(N) = f_i(x) \) if \( x_i = j \).

Proof:

First, note that \( 1 - \frac{r_kh_{k+1}}{S} \geq 0 \) since (7)' holds and

\[ S \geq (p - \sum\limits_{n=0}^{k} N_n)h_k^2 \geq r_kh_{k+1}^2. \]

Case 1. \( r_k \leq 0 \). Then \( \Delta = 0 \leq 0 \).

Case 2. \( r_k > 0 \).

\[(i) \quad \sum\limits_{j=0}^{k} N_j \psi_j^2(N) \leq \frac{r_k^2 \phi^2(x)}{S} \]

\[(ii) \quad \sum\limits_{j=1}^{k} A_j \phi(x) \sum\limits_{j=1}^{k} jN_j[V(N) - V(N-\delta_j+\delta_j-1)] \leq 0 \]

where \( V(N) = \text{Min}\left\{ \frac{br_k}{S}, 1 - \frac{r_kh_{k+1}}{S} \right\} \). Note that \( V(N) = V(N-\delta_j+\delta_j-1) \) for \( 1 \leq j \leq k \).
(iii) \( 2 \sum_{j=k+1}^{l} A_j = 2(k+1)N_{k+1} \left[ \psi_{k+1}(N) - \psi_k(N-\delta_{k+1} + \delta_k) \right] + 2 \sum_{j=k+2}^{l} A_j \)

\[ \leq 2kN_{k+1} \psi_{k+1}(N) - 2(k+1)N_{k+1} \psi_k(N-\delta_{k+1} + \delta_k) - \frac{2r_k^2 \phi(x)B}{S} \]

(the reasoning is similar to that of Theorem 3.3.8).

(i), (ii), and (iii) imply that

\[ \Delta \leq - \frac{2kN_{k+1} r_k \phi(x) h_k}{S} + 2(k+1)N_{k+1} \psi_k(N-\delta_{k+1} + \delta_k) - \frac{r_k^2 \phi(x)}{S} [2B - \phi(x)] \leq 0. \]

(3.3.17)

Remarks:

(1) The bound for the unbiased estimate of the risk deterioration \( \Delta \) given by (3.3.17) depends on \( k \) and \( N_{k+1} \). If \( N_{k+1} \) is likely to be large, then great improvement in risk will result if \( \hat{\lambda}^{(k)} \) is used instead of \( X \). Moreover, the dependence of (3.3.17) on \( k \) implies that the estimators \( \hat{\lambda}^{(k)} \) for different values of \( k \in J^+ \) are competitive.

(2) The special case when \( k = 0, \phi(x) \equiv 1, b = 1 \), and \( h \) is as given in Example 3.2.1 (c) is Peng's [1975] Theorem 5.1, which shrinks all non-zero observations towards zero while a possibly non-zero estimate of \( \lambda_1 \) is given for \( x_i = 0 \).

(3) The case when \( k = 0, \phi(x) \equiv 1, b = 0, \) and \( h \) is as given in Example 3.2.1 (c) is Theorem 3.1 of Peng [1975], which we stated as Theorem 2.2.3 in Section 2.

(4) If \( b = 0 \) in Theorem 3.3.16, \( \lambda_1 \) will be estimated as zero if \( x_1 = 0 \). However, if \( b > 0 \), the estimate for \( \lambda_1 \) will be possibly non-zero if \( x_1 = 0 \). The choice of relatively large value of \( b \) can be interpreted as reflecting the belief that the \( \lambda_1 \)'s are non-zero.
3.4 Adaptive Estimators

The estimators $\hat{\lambda}^{(k)}$ of $\lambda$ suggested in 3.3 pull the MLE towards a prechosen non-negative integer $k$, and the choice of $k$ is guided by the prior knowledge of the $\lambda_i$'s. A natural question which arises is: Is there an estimator $\hat{\lambda}$ of $\lambda$ which shifts the observations towards a point determined by the data itself? We shall show that the answer is affirmative.

Recall that $m = \text{Min}\{x_i\}$. We define new functions $H_i : \mathbb{R}^p \rightarrow \mathbb{R}$, $i = 1, \ldots, p$ as follows:

$$H_i(x) = 1 + \frac{x_i - m}{\sum_{n=2}^{\infty} \frac{1}{m+n}}$$

if $x_i > m+1$ and $m \geq 0$

$$= 1$$

if $x_i = m+1$ and $m > 0$

$$= 0$$

if $x_i = m$ and $m > 0$

$$= 0$$

if $m < 0$ for $i = 1, \ldots, p$.

The functions $H_i$ have similar properties as those of $h$ described in Lemma 3.2.2. We state the properties of the $H_i$'s in the following lemma.

Lemma 3.4.2.

Let $B = 1$. The $H_i$'s, $i = 1, \ldots, p$, have the following properties:

1. $H_i^2(x) - H_i^2(x - e_i)$ is nonincreasing in $x_i$ for $x_i > m+1$ with $m \geq 0$.

2. $x_i [H_i(x) - H_i(x - e_i)]$ is nonincreasing in $x_i$ for $x_i > m$ with $m \geq 0$ and $\lim_{x_i \rightarrow \infty} x_i [H_i(x) - H_i(x - e_i)] = B$.

3. $H_i(x) > H_i(x - e_i)$ if $x_i > m$ and $m > 0$.

4. $H_i(x) = 0$ if $x_i = m$.

5. $H_i(x) \geq 0$ if $m \geq 0$ and $x_i = m + 1$. 
(6) \( H_i(x) \geq (j + x_i)[H_i(x + je_i) - H_i(x + (j-1)e_i)] \) for \( j \geq 1 \), \\
x_i = m + 1, and \( m \geq 0 \).

**Proof:** The proof is straightforward and hence is omitted.

The following theorem provides us with a class of estimators \( \hat{\lambda}^{[m]} \) of \( \lambda \) which pull the MLE towards a point determined by the data, namely the minimum of the \( x_i \)'s.

**Theorem 3.4.3.**

Let \( \hat{\lambda}^{[m]} = (\hat{\lambda}_1^{[m]}, \ldots, \hat{\lambda}_p^{[m]}) \) be such that

\[
\hat{\lambda}_i^{[m]} = X_i - \frac{(p-N_m - 2) + \phi(x)H_i(X)}{p \sum_{i=1}^{p} H_i^2(X)}, \quad i = 1, \ldots, p,
\]

where

1. The \( H_i \)'s are as described in Lemma 3.4.2.
2. \( N_m = \#\{i : X_i = m\} \) and \( p \geq 4 \).
3. \( \phi(x) \) is non-negative and \( \phi(x) \leq 2B \) for some \( B > 0 \).
4. \( \phi(x) \) is nondecreasing in each argument \( x_i \).

Then for all \( \lambda = (\lambda_1, \ldots, \lambda_p) \), \( \hat{\lambda}^{[m]} \) dominates \( X \) under squared error loss.

**Proof:**

Define \( \psi_{i}^{(N)}(x) = f_i(x) \) if \( x_i = j \geq 0 \) and let \( S = \sum_{i=1}^{p} H_i^2(x) \). The proof is very similar to that of Theorem 3.3.8.

**Case 1.** \( (p-N_m - 2)_+ = 0 \). Then \( \Delta \equiv 0 \leq 0 \).

**Case 2.** \( (p-N_m - 2)_+ > 0 \).

(i) \[
\frac{\sum_{j=m}^{x} \sum_{j=m}^{x} \frac{N_{j} \psi_{j}^{2}(N)}{S}}{(p-N_m - 2)^2 \phi^2(x)} = \frac{(p-N_m - 2)^2 \phi^2(x)}{S}
\]
where $h(j)$ is defined to be $H_i(x)$ for $x_i = j$. The summation $\sum_{j=m+2}^{\xi} A_j$ can be shown (cf. Lemma 3.3.5 (3)) to be less than or equal to

$$\frac{2(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+2}^{\xi} h(j) - h(j-1) \sum_{j=m+2}^{\xi} \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}.$$

Hence we have

$$\sum_{j=m+1}^{\xi} A_j \leq \frac{2(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+1}^{\xi} h(j) \sum_{j=m+2}^{\xi} \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}$$

and $\Delta \leq \frac{(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+1}^{\xi} h(j) \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}$ (cf. Lemma 3.3.7)

where $h(j)$ is defined to be $H_i(x)$ for $x_i = j$. The summation $\sum_{j=m+2}^{\xi} A_j$ can be shown (cf. Lemma 3.3.5 (3)) to be less than or equal to

$$\frac{2(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+2}^{\xi} h(j) - h(j-1) \sum_{j=m+2}^{\xi} \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}.$$

Hence we have

$$\sum_{j=m+1}^{\xi} A_j \leq \frac{2(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+1}^{\xi} h(j) \sum_{j=m+2}^{\xi} \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}$$

and $\Delta \leq \frac{(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+1}^{\xi} h(j) \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}$ (cf. Lemma 3.3.7)

where $h(j)$ is defined to be $H_i(x)$ for $x_i = j$. The summation $\sum_{j=m+2}^{\xi} A_j$ can be shown (cf. Lemma 3.3.5 (3)) to be less than or equal to

$$\frac{2(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+2}^{\xi} h(j) - h(j-1) \sum_{j=m+2}^{\xi} \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}.$$

Hence we have

$$\sum_{j=m+1}^{\xi} A_j \leq \frac{2(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+1}^{\xi} h(j) \sum_{j=m+2}^{\xi} \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}$$

and $\Delta \leq \frac{(p-N_m - 2) + \phi(x)}{S} \sum_{j=m+1}^{\xi} h(j) \frac{S - 2h^2(j)}{S - h^2(j) + h^2(j-1)}$ (cf. Lemma 3.3.7)
should be shifted, by the data.

(2) When all the observations $x_1$ of the $p$ Poisson random variables are equal to the same value, say $x_0$, one would conjecture that the parameters $\lambda_1$ of the $p$ random variables are close to one another or even identical. In this case, one would estimate the $\lambda_i$'s by the grand mean $p(\sum_{i=1}^{p} x_i)/p$, which is equal to $x_0$. Our Bayesian point estimators proposed in Section 5 give such an estimate for the $\lambda_i$'s in this situation, as do the estimators $\hat{\lambda}_i[m]$ described in Theorem 3.4.3 above. That is, $\hat{\lambda}_i[m] = x_0$ for all $i$ if $x_i = x_0$ for all $i$, an intuitively appealing result.

As Peng [1975] has noticed, if some of the observations $x_1, \ldots, x_p$ are large, the unbiased estimate of the improvement in risk, i.e. $-\Delta$, will be small. This makes the savings of the proposed estimators small. In order to tackle this problem, Peng used Stein's method [1974] to modify his proposed estimator for Poisson parameters to guard against extreme observations. However, his estimators still give small savings when all the observations are large. For example, when the minimum of the $\lambda_i$'s is greater than or equal to 12, say, the savings from using Peng's modified estimator will still be small. Our proposed estimator $\hat{\lambda}_i[m]$ is useful in this situation. The savings of $\hat{\lambda}_i[m]$ will be much greater than Peng's estimators provided that the observations fall into a relatively narrow interval, i.e. when the observations do not differ very much.

In general, the estimators $\hat{\lambda}_i[m]$ are expected to perform well in a very wide range of parameter values, including the cases when the $\lambda_i$'s are relatively small as well as when they are relatively large. The improvement in risk of $\hat{\lambda}_i[m]$ over the MLE should be considerable, especially when $p$ is large and the $\lambda_i$'s are close to one another. A simulation
result of the behaviour of $\hat{\lambda}[m]$ is reported in Section 7, which supports our conjecture. Moreover, further application of Stein's method [1974] to our estimators $\hat{\lambda}[m]$ will yield a more versatile estimator $\hat{\lambda}[m]'$ of $\lambda$ in that, unlike Peng's estimator, it guards against extreme observations and cases in which all the parameters are large or small. The application is straightforward (cf. Peng [1975]).

We now digress to discuss an obvious application of the techniques in this section to extend Hudson's result ([1977], p. 18) on general discrete exponential families. Though we will not carry out the detailed analysis here (since it is outside the scope of this study), we briefly give an example below. The interested reader should read Hudson [1977] before proceeding to the following example because we will use the notation employed by Hudson. Using his notation, we define

\begin{align*}
(1) \quad m = \min_{i=1}^{p} \{x_i\} \text{ and recall } t(x_i) &= t_j \text{ if } x_i = j \\
(2) \quad b_{x_i} &= \frac{1}{t_1} + \sum_{k=2}^{\infty} \frac{1}{t_k+m} \quad \text{ if } x_i = m+2, \ldots \\
&= \frac{1}{t_1} \quad \text{ if } x_i = m+1 \\
&= 0 \quad \text{ if } x_i = m \\
(3) \quad S = \sum_{i=1}^{p} b_{x_i}^2 \quad p_m = (p-N-m-3)_+ \\
(4) \quad g_i(x) &= -\frac{p}{S} \cdot b_{x_i} \quad g(x) = (g_1(x), \ldots, g_p(x)).
\end{align*}

Then if $p \geq 5$, the estimator $T + g(x)$ dominates $T = (t(X_1), \ldots, t(X_p))$ under squared error loss. The improvement in risk exceeds

$$E_\phi \left[ \frac{2pN}{S} \left( \frac{t_{m+1}}{t_1} - 1 \right) + \frac{p_m^2}{S} \right].$$
SECTION 4. ESTIMATION UNDER K-NSEL

4.1 Introduction

Theorem 2.2.9 provides us with a class of estimators dominating the MLE, \( \hat{X} \), under normalized squared error loss \( L_1 = \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i \). Since \( \hat{X} \) is minimax under \( L_1 \), the estimators given in Theorem 2.2.9 are actually minimax estimators of \( \lambda \). One way to show that an estimator of the form \( \hat{X} + f(\hat{X}) \) is minimax is to show that \( f \) satisfies the conditions in Lemma 2.2.6 and that

\[
\Delta_1 = \sum_{i=1}^{p} \frac{f^2_i(\hat{x} + e_i)}{\hat{x}_i + 1} + 2 \sum_{i=1}^{p} \left[ f_i(\hat{x} + e_i) - f_i(\hat{x}) \right] \leq 0 \text{ for all } x \in J^+ \text{ (Corollary 2.2.8).}
\]

In the first part of this section, we shall use Corollary 2.2.8 to show that the class of minimax estimators given in Theorem 2.2.9 can be enlarged in two ways. The remainder of the section is devoted to attempts to find estimators better than the MLE under the \( k \)-normalized squared error loss (\( k \)-NSEL) function \( L_k(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i^k \), with \( k \geq 2 \). We shall make use of Lemma 2.2.7 to prove the main results and shall also discuss the motivation for using \( k \)-NSEL.

4.2 Minimax Estimators

We shall show in this subsection that a large class of minimax estimators can be obtained which includes the class of estimators proposed by Clevenson and Zidek [1975] as a subclass. The estimators we consider are of the form \( \hat{X} + f(\hat{X}) \) and loss function \( L_1 \) is used. As remarked in 4.1, we need only show that \( \Delta_1(x) \leq 0 \) for all \( x \in J^+ \).
Let \( \psi(z) \) be a nonincreasing real-valued function such that \( z + \psi(z) \) is nondecreasing for \( z \in J^+ \). The following theorem gives a class of estimators which dominate the MLE \( \hat{X} \) and hence are minimax.

**Theorem 4.2.1.**

Suppose \( X \sim \text{Poisson}(\lambda_i), i = 1, \ldots, p, p \geq 2 \).

Define \( \hat{\lambda}(X) = [1 - \phi(Z)/(Z+\psi(Z))]X \) where

1. \( \psi(z) \geq b > 0 \) for some \( b \)
2. \( \phi(z) \) is nondecreasing
3. \( 0 \leq \phi(z+1) \leq 2\min\{(p-1), \psi(z)\} \) for all \( z \in J^+ \).

Then for all \( \lambda = (\lambda_1, \ldots, \lambda_p) \), \( \hat{\lambda} \) dominates the MLE \( \hat{X} \) under the normalized squared error loss function \( L_1 \).

**Proof:**

Let \( f_i(x) = \frac{-\phi(x)x_i}{z+\psi(z)} \) if \( x \in J^p \)

\[ = 0 \] otherwise.

We need to show that

\[ \Delta_1 = \sum_{i=1}^{p} \frac{f_i^2(x+e_i)}{x_i+1} + 2 \sum_{i=1}^{p} [f_i(x+e_i) - f_i(x)] \leq 0. \]

In fact,

\[ \Delta_1 = \frac{\phi^2(z+1)(z+p)}{[z+1+\psi(z+1)]^2} - 2 \frac{\phi(z+1)(z+p)}{z+1+\psi(z+1)} \frac{\phi(z)z}{z+\psi(z)} \]

\[ \leq \frac{\phi(z+1)}{[z+1+\psi(z+1)]} \frac{(z+p)\phi(z+1)}{z+1+\psi(z+1)} - 2 \frac{z(z+1+\psi(z+1))}{z+\psi(z)} \]

(since \( z+\psi(z) \) is non-decreasing and \( \phi(z) > 0 \))
Note that $\psi(z) - \psi(z + 1) \geq 0$ by assumption and that $z + \psi(z)$ is non-decreasing, so we have

$$\Delta_1 \leq \frac{\phi(z+1)}{[z+1+\psi(z+1)]^2} \{(z+p)\phi(z+1) - 2 ((p-1)z + p\psi(z))\}$$

$$\leq 0 \text{ by condition } (3) \text{ of the theorem.}$$

Therefore $\hat{\lambda} = (1 - \frac{\phi(z)}{z+\psi(z)})X$ dominates $X$.

Q.E.D.

The special case where $\psi(z) = b > 0$ is stated as a corollary below.

**Corollary 4.2.2.**

Suppose $X_i \sim \text{Poisson} (\lambda_i^*), i = 1; \ldots , p$, $p \geq 2$.

Then the estimator $\hat{\lambda} = X - \frac{\phi(z)}{z+b}X$ of $\lambda$ dominates the MLE $X$ under the loss function $L_1$ where

1. $b > 0$
2. $\phi(z)$ is nondecreasing
3. $0 \leq \phi(z) \leq \min\{2(p-1), 2b\}$ and $\phi(z) \neq 0$.

Note that the constant $b$ given in the corollary is an arbitrary positive real number, while the estimators given in Theorem 2.2.9 require $b \geq p-1$. Moreover, the theorem requires that $0 \leq \phi(z) \leq 2(p-1)$, and hence the class of estimators given in Theorem 2.2.9 is in fact a subclass of ours.
The estimator $\hat{\lambda}$ shrinks the MLE towards the origin by the amount $\frac{\phi(z)x}{z+b}$. For every $b$, the maximum shrinkage allowable if $\hat{\lambda}$ is to dominate the MLE is $\frac{\text{Min}\{2(p-1),2b\}x}{z+b}$, which is an increasing function of $b$ (coordinate-wise) whenever $0 < b \leq p-1$ and a decreasing function whenever $b > p-1$. Therefore the maximum shrinkage is obtained when $b = p-1$.

An application of Corollary 4.2.2 above gives us still more interesting estimators of $\lambda$. The result is stated in Corollary 4.2.3 below.

**Corollary 4.2.3.**

Suppose $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, p$, $p \geq 2$. Then the estimator

$$\hat{\lambda} = (1 - \frac{a}{Z+c})^t X$$

of $\lambda$

dominates the MLE $X$ under the loss function $L_1$ where

1. $t \geq 1$, $c > 0$

2. $0 < a < \text{Min}\{\frac{2(p-1)}{t}, c, \frac{2c}{t}\}$.

**Proof:**

Rewrite $\hat{\lambda}$ as follows:

$$\hat{\lambda} = (1 - [1 - (\frac{Z+c-a}{Z+c})^t])X$$

$$= (1 - \frac{\theta(z)}{Z+c})X$$

where $\theta(z) = \frac{(z+c)^t - (z+c-a)^t}{(z+c)^{t-1}}$.

We shall show that (i) $\theta(z)$ is non-decreasing and

(ii) $0 \leq \theta(z) \leq \text{Min}\{2(p-1), 2c\}$. Now, the derivative of $\theta(z)$ is

$$\theta'(z) = 1 - \frac{(z+c-a)^{t-1}(z+c-a+ta)}{(z+c)^t}.$$
Since \((z+c)^t \geq (z+c-a)^t + ta(z+c-a)^{t-1}\) for all \(z \geq 0\), we have \(\theta'(z) \geq 0\). Hence \(\theta(z)\) is non-decreasing. Next, it is clear that \(0 \leq \theta(z)\) for all \(z \geq 0\). Now

\[
\theta(z) = \frac{ta(z+c-a)^{t-1}}{(z+c)^{t-1}} + \psi(z)
\]

where \(\psi(z)\) is such that \(\lim_{z \to \infty} \psi(z) = 0\). As a result, \(\theta(z) \leq \lim_{z \to \infty} \theta(z) = ta\). From condition (2), we have \(\theta(z) \leq \min\{2(p-1), tc, 2c\}\). By Corollary 4.2.2 we see that the estimator

\[
\hat{\lambda} = (1 - \frac{\theta(z)}{z+c})X = (1 - \frac{a}{z+c})^t X
\]

dominates the MLE under the loss function \(L_1\).

Q.E.D.

The estimator \(\hat{\lambda}(X) = \left(1 - \frac{(p-1)}{(Z+p-1)}\right)^2 X\), which is an estimator described in the previous corollary with \(t = 2\) and \(a = c = p-1\), shrinks more towards the origin than does the estimator \(\hat{\lambda}^* = \left(1 - \frac{(p-1)}{(Z+p-1)}\right)X\). Thus, \(\hat{\lambda}\) should give a better estimate of \(\lambda\) than \(\hat{\lambda}^*\) if the parameters \(\lambda_i\), \(i = 1, \ldots, p\) are close enough to zero. The following argument gives us an interesting insight as to why we might arrive at estimators of the form \(\hat{\lambda} = \left(1 - \frac{c}{Z+a}\right)^2 X\).

Let \(X_i \sim \text{Poisson} \left(\lambda_i\right), i = 1, \ldots, p\), be mutually independent and let \(Y_i = 2\sqrt{X_i}, \theta_i = 2\sqrt{\lambda_i}, i = 1, \ldots, p\). It is approximately true that \(Y_i \sim N(\theta_i, 1), i = 1, \ldots, p\), and that the \(Y_i\)'s are mutually independent. That is, approximately, \(Y = (Y_1, \ldots, Y_p) \sim N(\theta, I_p)\), where \(I_p\) is the \(p \times p\) identity matrix and \(\theta = (\theta_1, \ldots, \theta_p)\). The James-Stein estimator

\[
\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)
\]

of \(\theta\) under squared error loss is \(\hat{\theta}_i = (1 - r/Y_i Y_i)Y_i\), \(i = 1, \ldots, p\). Or, in terms of \(X\) and \(\lambda\), \(\hat{\lambda}_{1/2} = (1 - \frac{c}{Z})\sqrt{X_i}, i = 1, \ldots, p\).
where \( Z = \sum_{i=1}^{p} X_i \), \( c = r/4 \), and \( \lambda = (\lambda_1, \ldots, \lambda_p) \) is an estimator of \( \lambda \).

We thus have the estimator \( \hat{\lambda} = (1 - c/Z)^2 X \) of \( \lambda \). Since \( Z \) has a positive probability of being zero, we are prompted to replace \( Z \) by \( Z + a \), where \( a \) is a positive real number. We thus arrive at the estimator

\[
\hat{\lambda} = (1 - \frac{c}{Z+a})^2 X.
\]

One might argue that the above result is obtained through the use of squared error loss instead of normalized squared error loss. However, notice that the squared error loss function is applied only in estimating \( \theta_i \), and asymptotically, \( (\theta_i - \hat{\theta}_i)^2 \) is \( O(\lambda_i^2) \), \( i = 1, \ldots, p \) (i.e. \( \lim_{\lambda_i \to \infty} \frac{(\theta_i - \hat{\theta}_i)^2}{\lambda_i^2} < \infty \)). The normalized squared error loss function has the same asymptotic property, i.e. \( \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i} = O(\lambda_i) \), \( i = 1, \ldots, p \).

Hence, our result can be thought of as if it is obtained under \( L_1 \).

In order that estimators of the form \( \hat{\lambda} = X - \frac{\phi(Z)}{Z+b}X \) dominate the MLE \( X \) under normalized squared loss, the requirement that \( \phi(z) \) be nondecreasing is not a necessary condition. In fact, \( \phi(z) \) can be decreasing for all \( z \) and \( \hat{\lambda} \) will still dominate the MLE \( X \). The following theorem states this fact.

**Theorem 4.2.4**

Suppose \( X_i \sim \text{Poisson}(\lambda_i) \), \( i = 1, \ldots, p \) (\( p \geq 2 \)).

Let \( \hat{\lambda} = (1 - \frac{\phi_t(Z)}{Z+b})X \) be an estimator of \( \lambda \) where

\[
(1) \quad t > 0, \quad b > t+1
\]

\[
(2) \quad \phi_t(z) \text{ is nondecreasing in } z
\]

\[
(3) \quad \phi_t(z) > 0 \text{ and } \phi_t(z) \neq 0
\]

\[
(4) \quad \frac{\phi_t(z)}{(z+b)^t} \leq \operatorname{Min}\{2(p-t-1),2(b-t-1)\}.
\]
Then for all $\lambda = (\lambda_1, \ldots, \lambda_p)$, $\hat{\lambda}$ has smaller risk than $\lambda^o = X$ when the loss function is given by $L_1$.

Proof:

Let $f_i(x) = \frac{-\phi_t(z)x_i}{(z+b)^{t+1}}$ if $x \in J^p$

$$= 0 \quad \text{otherwise, } i = 1, \ldots, p.$$  

Again, we need to show

$$\Delta_1 = \sum_{i=1}^{p} f_i^2(x_i) + \sum_{i=1}^{p} [f_i(x_i) - f_i(x)] \leq 0.$$  

Now

$$\Delta_1 \leq \frac{\phi_t^2(z+1)(z+p)}{(z+b+1)^{2t+2}} + 2 \frac{-\phi_t(z+1)(z+p)}{(z+b+1)^{t+1}} - \frac{\phi_t(z)}{(z+b)^{t+1}} - \frac{2p}{(z+b)^{t+1}}.$$  

(since $\phi_t$ is non-decreasing)

$$\Delta_1 = \frac{\phi_t^2(z+1)(z+p)}{(z+b+1)^{2t+2}} + 2 (z+p) \left[ \frac{1}{(z+b)^{t+1}} - \frac{1}{(z+b+1)^{t+1}} + \frac{2p}{(z+b)^{t+1}} \right].$$  

Observe that the function $f$ defined by $f(z) = \frac{1}{(z+b)^{t+1}}$ is strictly convex for $z \geq 0$. Hence $f(z+1) - f(z) > f'(z)$. Consequently

$$\Delta_1 \leq \frac{\phi_t(z+1)}{(z+b+1)^{t}} \left[ \frac{\phi_t(z+1)}{(z+b+1)^{t}} + \frac{(z+p)}{(z+b)^{t+2}} - \frac{2p}{(z+b)^{t+1}} \right]$$  

$$= \frac{\phi_t(z+1)}{(z+b)^{t+2}} \left[ z \frac{\phi_t(z+1)}{(z+b+1)^{t}} + 2(t+1) - 2p \right]$$  

$$+ \frac{\phi_t(z+1)}{(z+b+1)^{t}} \left[ z \frac{\phi_t(z+1)}{(z+b+1)^{t}} + 2(t+1) - 2b \right].$$  

\[ \phi_t(z+1) \leq 2\text{Min}\{p-t-1,b-t-1\}, \text{i.e. condition (4) of our theorem.} \]

Q.E.D.

If \( b=p-1 \), the theorem holds even when condition (4) in the theorem is replaced by

\[ (4)' \ (z+b)^{-t} \phi_t(z) \leq 2(p-t-1). \]

That is, \( (z+b)^{-t} \phi_t(z) \) has a larger upper bound. We state this result in the theorem below.

**Theorem 4.2.5.**

Let \( X, \lambda \) and \( \hat{\lambda} \) be as given in Theorem 4 with \( b=p-1 \).

Suppose \( \phi_t(z) \) satisfies the following conditions:

\begin{enumerate}
  \item \( t \geq 0, \ p-1 > t \)
  \item \( \phi_t(z) \) is non-decreasing in \( z \)
  \item \( \phi_t(z) \geq 0 \) and \( \phi_t(z) \neq 0 \)
  \item \( \frac{\phi_t(z)}{z+p-1} \leq 2(p-t-1) \)
\end{enumerate}

Then for all \( \lambda \),

\( \hat{\lambda} \) has smaller risk than \( X \) when the error loss function is given by \( L_1 \).

**Proof:**

The proof here is essentially the same as that of Theorem 4.2.4. It can be shown that
\[ \Delta_1 \leq \frac{\phi^2_t(z+1)}{(z+p)^{2t+1}} + 2\phi_t(z+1) \left[ \frac{1}{(z+p-1)^t} - \frac{1}{(z+p)^t} \right] - \frac{(p-1)}{(z+p)^{t+1}} \]

\[ \leq \frac{\phi^2_t(z+1)}{(z+p)^t(z+p-1)^{t+1}} + 2\phi_t(z+1) \left[ \frac{t}{(z+p-1)^{t+1}} - \frac{(p-1)}{(z+p)^{t+1}} \right] \]

\[ \leq \frac{\phi_t(z+1)}{(z+p-1)^{t+1}} \left[ \frac{\phi_t(z+1)}{(z+p)^t} - 2(p-t-1) \right] \]

\[ \leq 0 \text{ (by condition (4)'').} \]

The second inequality holds since the function \( f(z) = \frac{1}{(z+p-1)^t} \) is strictly convex for \( z \geq 0 \).

Q.E.D.

When \( t = 0 \), the above theorem is Theorem 2.2.9 of Section 2. The theorem, due to Clevenson and Zidek [1975], suggests that estimators of the form \( \lambda = (1 - \frac{\theta(Z)}{Z+p-1})X \) will dominate the MLE \( X \) provided that \( 0 \leq \theta(z) \leq 2(p-1) \) and that \( \theta(z) \) is nondecreasing. However, as we see from the theorem above, the requirement that \( \theta(z) \) be nondecreasing is not necessary. A simple example is given below.

Example 4.2.6.

Let \( \hat{\lambda} = (1 - \frac{c}{(Z+p-1)^{t+1}})X \) with \( 0 \leq t < p-1 \) and \( 0 < c < 2(p-t-1) \).

Then by our theorem, \( \hat{\lambda} \) is better than \( X \) under the normalized squared error loss function. We can write \( \hat{\lambda} \) in the form \( (1 - \frac{\theta(Z)}{Z+p-1})X \), where \( \theta(z) = \frac{c}{(z+p-1)^t} \), and note that \( \theta(z) \) is strictly decreasing for all \( z \geq 0 \).
4.3 Better Estimators Under $L^k$

The usage of unbounded loss functions has been a controversial topic because of the famous St. Petersburg's paradox; hence it is natural to consider bounded loss functions. Noting that the loss function $L_1$ is unbounded both when the $\lambda_i$'s are small and when they are large, we are prompted to consider the loss functions $L_k(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i^k}$ (for $k \geq 2$) which are bounded when the $\lambda_i$'s are large. This is desirable if we are primarily interested in the problem of estimating $\lambda$ when the $\lambda_i$'s are relatively large. Furthermore, when the $\lambda_i$'s are large, the variances of the distributions are large, and the estimation problem is more difficult. Intuitively we might expect the MLE to perform fairly well when the $\lambda_i$'s are relatively small, since in this case, the variances of the distributions are small, and the observations are exceedingly likely to fall close to the mode (hence the range of the observations is very likely small). However, when the $\lambda_i$'s are large, the variances are large and the observations are scattered across a wide range of values. This leads us to conjecture that it is possible to improve on the MLE when the $\lambda_i$'s are large.

There is another reason why we might think of using $L_2$ as our loss function in the estimation problem. The parameter of a Poisson distribution can be thought of not only as the mean of the distribution, but also as the variance. With this in mind, a natural loss function to use would be $\sum_{i=1}^{p} \frac{(1 - \hat{\lambda}_i / \lambda_i)^2}{\lambda_i}$, which is $L_2(\lambda, \hat{\lambda})$ (see Brown [1968]).

Before stating the next theorem, we introduce some notation to be used in the theorem.
Definitions:

(1) \( \overline{S} = \sum_{i=1}^{p} (X_i + k)^{(k)} = \sum_{i=1}^{p} (X_i + 1) \cdots (X_i + k) \)

(2) \( S = \sum_{i=1}^{p} (x_i + k)^{(k)} = \sum_{i=1}^{p} (x_i + 1) \cdots (x_i + k) \)

(3) \( \overline{s}^i = \overline{S} - (X_i + k)^{(k)} \)

(4) \( s^i = S - (x_i + k)^{(k)} \)

Theorem 4.3.1 below gives us a class of estimators \( \hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p) \) uniformly dominating the MLE under the loss function \( L_k (k \geq 2) \). These estimators have the following properties:

(1) If the observation from the \( i^{th} \) population is small (less than \( k \)), then the estimator \( \hat{\lambda}_i \) of \( \lambda_i \) is the same as the MLE.

(2) If the observation is large (greater than or equal to \( k \)), then the estimator \( \hat{\lambda}_i \) of \( \lambda_i \) shrinks the MLE towards zero.

Using the definitions given above, the theorem and its proof are stated as follows:

**Theorem 4.3.1.**

Suppose \( X_i \sim \text{Poisson}(\lambda_i), i = 1, \ldots, p, p \geq 2, \) and the loss function is \( L_k (\lambda, \hat{\lambda}) \). Then the estimator \( \hat{\lambda} \) given below dominates the MLE \( X \) uniformly in \( \lambda = (\lambda_1, \ldots, \lambda_p) \):

\[
\hat{\lambda}_i = \text{ith coordinate of } \hat{\lambda} \\
= X_i - \frac{\phi(Z) X_i (X_i-1) \cdots (X_i-k+1)}{\overline{s}^i - X_i (X_i-1) \cdots (X_i-k+1)}
\]
where (1) \[ z = \sum_{i=1}^{p} x_i \]

(2) \( \phi(z) \) is a real valued function increasing in \( z \)

(3) \( 0 \leq \phi(z) \leq 2k(p-1) \) and \( \phi(z) \neq 0 \).

**Proof:**

Let \( f_i(x) = \frac{-\phi(z) x_i (x_i-1) \cdots (x_i-k+1)}{S^i + x_i (x_i-1) \cdots (x_i-k+1)} \) if \( x_i \geq 0, i = 1, \ldots, p \)

\[ = 0 \quad \text{if } x_i < 0. \]

We see that the \( f_i \)'s satisfy the conditions given in Lemma 2.2.6. Hence Lemma 2.2.7 gives us an unbiased estimate \( \Delta_k \) of the deterioration in risk of \( \lambda \).

We have \( D_k = E_{\lambda} \Delta_k = R(\lambda, \lambda) - R(\lambda, X) \)

and

\[ \Delta_k = \sum_{i=1}^{p} \frac{f_i^2(x+k e_i)}{(x_i+k)^{\alpha}} + 2 \sum_{i=1}^{p} \frac{f_i(x+k e_i) - f_i(x+(k-1) e_i)}{(x_i+k)^{\alpha}} \]

Substituting the \( f_i \)'s in the formula, we have

\[ \Delta_k = \frac{\phi^2(z+k)}{-S} + 2 \sum_{i=1}^{p} \frac{-\phi(z+k) \cdot (x_i+k)}{S} + \frac{\phi(z+k) \cdot x_i}{S^i + (x_i+k-1)^{\alpha}} \]

\[ \leq \frac{\phi^2(z+k)}{S} + \frac{2\phi(z+k)}{S} \sum_{i=1}^{p} \frac{-(x_i+k)[S - k(x_i+k+1)(k-1)] + x_i S}{S^i + (x_i+k-1)^{\alpha}} \]

(since \( \phi(z) \) is increasing and \( \phi(z) > 0 \))

\[ = \frac{\phi^2(z+k)}{-S} + \frac{2\phi(z+k)}{S} \sum_{i=1}^{p} \frac{-kS + k(x_i+k)(k)}{S^i + (x_i+k-1)^{\alpha}}, \]

Now observe that \( S > S^i + (x_i+k-1)^{\alpha} \)

and \( S > (x_i+k)^{\alpha} \).
Hence, we have,

\[ \Delta_k \leq \frac{\phi(z+k)}{S} [\phi(z+k) - 2k(p-1)] \]

\[ \leq 0 \]

(since \( 0 \leq \phi(z+k) \leq 2k(p-1) \) and \( \phi(z+k) \neq 0 \)).

\[ \text{Q.E.D.} \]

Remarks:

(1) When \( k = 1 \), Theorem 4.3.1 is the same as Theorem 2.2.9, a result in Clevenson and Zidek [1975].

(2) If \( x_i \leq k-1 \), then \( \hat{\lambda}_i \) gives the same estimate as the MLE, i.e. no shrinkage of the MLE takes place. We see from this that as \( k \to \infty \), the likelihood that shrinkage will be indicated becomes smaller and smaller.

It is intuitively clear that if we shrink the MLE, we should shrink small observations less than we shrink large observations. This is partly because when the observations are small, the underlying parameters are likely to be small, which means that the variances are small and hence each observation is likely to be close to the value of its respective parameter. On the other hand, larger values of the parameters correspond to larger variances of the random variables, and the higher probability of scattering of the observations leads one to suspect that in this case the MLE is less reliable than when the observations are small; one might therefore be more inclined to adjust the MLE. The integer \( k \) in the loss function \( L_k \) reflects the degree of concern about misestimation of relatively small \( \lambda_i \)'s. Large values of \( k \) result in loss functions which are very sensitive to changes in the \( \lambda_i \)'s. Since our estimators shrink those observations that are greater than or equal to \( k \) but leave the
others untouched, the integer \( k \) can be interpreted as an indicator of a person's willingness to move the MLE for better estimation results. Choice of a small value of \( k \) would probably result from a prior belief that the \( \lambda_i \)'s are small, and hence the person is willing to move even small observations which, as pointed out above, are supposed to be more reliable. Choice of a large value of \( k \) would probably result from a prior belief that the \( \lambda_i \)'s are not small, and thus the person is inclined not to move the observations if they are not large enough.

(3) Theorem 3.1 of Clevenson and Zidek [1975] suggests that estimators \( \hat{\lambda} = (1 - \phi(Z)/(Z+p-1))X \) of \( \lambda \) still dominate the MLE under a general loss function \( L_K(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} K(\lambda_i) (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i \) where \( K > 0 \) is some nonincreasing function. When \( K(y) = 1/y^{k-1} \), \( L_K \) is the \( k \)-NSEL \( L_k \).

However, our estimators do not shrink observations that are less than \( k \); only those observations greater than or equal to \( k \) are moved. Therefore if \( \lambda_i \geq k \), our estimators guard against unnecessary shrinkage if the observations happen to be small (i.e. \( < k \)). Since the Clevenson-Zidek estimator shrinks all non-zero observations, we are led to conjecture that our estimators are better than theirs in terms of the percentage in savings compared to the MLE when the \( \lambda_i \)'s are relatively large (i.e. when \( \min(\lambda_i) \geq k \geq 2 \)). Some simulation results which support this conjecture are reported in Section 7.

The next theorem is a generalization of Corollary 4.2.2 to the case when \( k \)-NSEL \( L_k \) is used, where \( k \) is any positive integer.
Theorem 4.3.2.

Suppose $X = (X_1, \ldots, X_p)$ is as given in Theorem 4.3.1 and
\( \hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p) \) is an estimator of \( \lambda \).

Let \( \hat{\lambda}_i = X_i - \frac{\phi(Z)X_i^{(k)}}{S^i + X_i^{(k)} + b} \), \( i = 1, \ldots, p \),

where (1) \( k \) is a positive integer
(2) \( \phi \) is non-decreasing
(3) \( 0 \leq \phi(z) \leq \min\{2 \frac{(b+(p-1)k!}{k!}, 2k(p-1)\}\}
(4) \( b > - (p-1)[k!] \).

Then for all \( \lambda \), \( \hat{\lambda} \) dominates \( X \) under the error loss function \( L_k \).

Proof:

Define \( f_i(x) = \frac{-\phi(z)x_i^{(k)}}{S^i + x_i^{(k)} + b} \) if \( x \in J^i \)
\( = 0 \) otherwise,
\( i = 1, \ldots, p \).

We need to show that
\[
\Delta_k = \sum_{i=1}^{p} \frac{f_i(x+ke_i)}{x_i^{(k)} + (x_i + k)(k)} + 2 \sum_{i=1}^{p} \frac{f_i(x+ke_i) - f_i(x+(k-1)e_i)}{(x_i + k)(k)} 
\leq 0.
\]
Indeed \( \Delta_k \leq \frac{\phi^2(z+k)S}{(S+b)^2} - 2 \frac{\phi(z+k)}{S+b} \sum_{i=1}^{p} \frac{kS - k(x_i + k)(k) + kb}{S^i + (x_i + k - 1)(x_i) + b} \).
\[
\frac{1}{2} \left( \frac{(z+k)^2}{[S+b]^2} - \frac{2(z+k)}{[S+b]^2} (k(p-1)S+pkb) \right)
\]

(since \( kS - k(x_1+k)(k)+kb > 0 \))

\[
\leq \frac{1}{2} \frac{(z+k)^2}{[S+b]^2} \left[ S(\phi(z+k)-2k(p-1)) - 2pkb \right].
\]

Now observe that

\[
S(\phi(z+k)-2k(p-1)) = 2pkb
\]

\[
\leq pk! (\phi(z+k)-2k(p-1)) - 2pkb
\]

(since condition (3) gives \( \phi(z) \leq 2k(p-1) \))

\[
= pk! \left[ \phi(z+k) - 2 \frac{(b+(p-1)k)!}{(k-1)!} \right]
\]

\[
\leq 0 \text{ by condition (3)}.
\]

Hence \( \Delta_k \leq 0 \).

Q.E.D.
SECTION 5. BAYESIAN ANALYSIS

5.1 Introduction

In this section, we explore the problem of estimating the Poisson parameters from another perspective. Recall that \( X_i \sim \text{Poisson}(\lambda_i) \), \( i = 1, \ldots, p \), and that the \( X_i \)'s are mutually independent. Moreover, only one observation is taken from each of the \( p \) Poisson random variables. In contrast to the frequentist approach taken in the last section, we shall take a Bayesian approach and show how to find estimators better than the MLE \( X = (X_1, \ldots, X_p) \). Observe that \( X \) is a Bayes estimator if prior knowledge of the parameters \( \lambda_i \) is non-informative (vague) and the \( \lambda_i \)'s are independent. When substantial prior knowledge is available, significant improvement on the usual estimation procedure would be expected by means of Bayesian methods. That is, in certain situations, we can incorporate the information at hand about the prior distribution in a Bayesian manner and obtain estimators of \( \lambda = (\lambda_1, \ldots, \lambda_p) \) better than \( X \).

Recently, some attempts have been made to study the problem of simultaneously estimating the parameters of several independent Poisson random variables from the Bayesian point of view. As mentioned previously, Clevenson and Zidek [1975] propose a class of estimators that dominate the MLE uniformly under normalized squared error loss. They also provide a Bayesian interpretation of their results. Leonard [1972] assumes that the \( \lambda_i \)'s are independent and identically distributed, with \( \ln \lambda_i \sim N(\mu, \sigma^2) \), \( i = 1, \ldots, p \) for given \( \mu \) and \( \sigma^2 \), that \( \mu \) is uniformly distributed over the real line, and that \( \nu \sigma^2 \) is independent of \( \mu \) and has a chi-square distribution with \( \nu \) degrees of freedom. Modal estimates
of $\ln \lambda_i$, $i = 1, \ldots, p$ are proposed. Later, in another paper, Leonard [1976] briefly discusses the problem again. He assumes that the parameters $\lambda_i$ are exchangeable in the sense of de Finetti [1964]. That is, the prior distributions (two at a time, three at a time, etc.) of the $\lambda_i$'s are invariant under permutation of the suffixes. Given $\alpha > 0$ and $\beta > 0$, the $\lambda_i$'s are assumed to be independent and each $\lambda_i$ has gamma density

$$f(\lambda_i | \alpha, \beta) = \frac{\beta^\alpha \lambda_i^{-1} e^{-\beta \lambda_i}}{\Gamma(\alpha)} \quad \text{for } \lambda_i > 0$$

$$= 0 \quad \text{otherwise.}$$

In the second stage of the distribution, $\ln \beta$ is assumed to be uniformly distributed over the real line and no prior distribution of $\alpha$ is suggested.

In this section, we generalize the above results by adopting various prior distributions on $\alpha$, all included within the broad family of generalized hypergeometric functions, and develop the joint posterior distribution of $\lambda$ as well as the marginal posterior distributions of the $\ln \lambda_i$, $i = 1, \ldots, p$. Point estimators of the $\lambda_i$'s are proposed and compared with the MLE. By means of a computer simulation which will be reported in Section 7, it is found that in certain situations, especially when the parameters $\lambda_i$ are close to each other and the loss is squared error, a substantial savings over the MLE will result. We shall first derive a Bayesian solution for our problem, and then develop the marginal posterior density of $\lambda_i$. 
5.2 Estimates of the Parameters

The following distributions are relevant to a Bayesian solution of our problem:

(1) Given $\lambda_i$, $i=1,...,p$, the observations $x_i,...,x_p$ are independent and have Poisson distributions with parameters $\lambda_i,...,\lambda_p$ respectively.

The joint probability mass function of $x$ given $\lambda$ is

$$f(x|\lambda) = \prod_{i=1}^{p} \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!}.$$

(2) The $\lambda_i$'s are exchangeable a priori, and the prior distribution is described in two stages as follows:

(i) Given $\alpha$ and $\beta$, the $\lambda_i$ are independent and have a gamma density, i.e. for $\alpha > 0$, $\beta > 0$,

$$\pi(\lambda_i | \alpha, \beta) = \beta^\alpha \lambda_i^{\alpha-1} e^{-\beta \lambda_i} / \Gamma(\alpha)$$

if $\lambda_i > 0$

$$= 0$$

otherwise.

(ii) $\alpha$ and $\beta$ are independent, and $\beta$ has a non-informative prior density, i.e., the density of $\beta$ is proportional to $1/\beta$. The parameter $\alpha$ is assumed to have the translated geometric distribution with probability mass function

$$(1-\mu) \quad \text{if } \alpha = 1$$

$$(1-\mu) \mu^{\alpha-1} \quad \text{if } \alpha = 2, 3, ..., \text{ where } 0 < \mu < 1.$$
The joint density of $x, \lambda, \alpha, \beta$ given $\mu$ is proportional to

$$\prod_{i=1}^{p} \frac{e^{-\lambda_i x_i}}{\lambda_i^{x_i!}} \cdot \prod_{i=1}^{p} \frac{\beta \lambda_i^{\alpha-1} e^{-\lambda_i}}{\Gamma(\alpha)} \cdot \beta \cdot \mu^{\alpha-1}.$$ 

Conditioning on $x$ and integrating with respect to $\beta$ gives the joint posterior density of $\lambda$ and $\alpha$, given $x$ and $\mu$,

$$f(\lambda, \alpha | x, \mu) \propto e^{-\Lambda} \prod_{i=1}^{p} \frac{x_i!}{\lambda_i^{1}} \prod_{i=1}^{p} \frac{\lambda_i^{\alpha-1} \Gamma(p\alpha)}{\Gamma(\alpha) \cdot \lambda_i^{p\alpha}}$$

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

Now introduce the common factorial function notation $(y)_a = \frac{\Gamma(y+a)}{\Gamma(y)}$, and let $\phi = (p_p \prod_{i=1}^{p} \lambda_i) / \lambda^p$. Using the Gauss multiplication theorem identity (see, e.g., p. 26 of Rainville [1960])

$$\Gamma(pa) \equiv (2\pi)^{\frac{1}{2}} (1-p)^{pa} \prod_{k=0}^{p-1} \Gamma(\alpha + \frac{k}{p}),$$

we have

$$f(\lambda, \alpha | x, \mu) = C(x, \mu) \prod_{i=1}^{p} \frac{e^{-\Lambda} x_i! \lambda_i^{\alpha-1} \Gamma(p\alpha)}{\lambda_i^{\alpha-1} \lambda^p \Gamma(\alpha) \cdot \lambda_i^{p\alpha}} \phi^{\alpha-1},$$

where $C(x, \mu)$ depends only on $x$ and $\mu$, and is such that:

$$\sum_{\alpha=1}^{\infty} \int f(\lambda, \alpha | x, \mu) d\lambda = 1.$$ 

Summing with respect to $\alpha$ gives the joint posterior density of the $\lambda_i$'s, given $\mu$ and the data $x$. 
\[ f(\lambda|x,\mu) = C(x,\mu) \frac{e^{-\lambda} \prod_{i=1}^{p} \lambda_i^{x_i}}{\Lambda^p} \frac{1}{\prod_{\alpha=1}^{p-1} (\alpha-1)!} \sum_{k=1}^{p-1} \frac{(1 + \frac{k}{p})^{x_i}}{(1)_{\alpha-1} \phi^{\alpha-1}} \]

Notice that the infinite sum here is a generalized hypergeometric function

\[ \sum_{j=0}^{p-2} \frac{(1)_{j}}{(1)_{p-2-j}} \phi^j \]

which converges for \(|\phi| < 1\), i.e. for \(p \mu \sum_{i=1}^{p} \frac{\lambda_i}{\Lambda} < 1\). Since the geometric mean of the \(\lambda_i\)'s cannot exceed their arithmetic mean, it is sufficient for the function to converge if \(\mu < 1\). The joint posterior density of \(\lambda\) is proper if the sum of the observations is positive, which is a reasonable assumption. Now we find the explicit expression of the normalizing constant \(C(x,\mu)\). We make the following transformation of variables:

\[ (\lambda_1, \ldots, \lambda_p) \rightarrow (\Lambda, \theta_1, \ldots, \theta_{p-1}) \]

where \(\lambda_i = \Lambda \theta_i\) for \(i = 1, \ldots, p\), and

\[ \sum_{i=1}^{p} \theta_i = 1, \theta_i > 0. \]

The Jacobian of the transformation is \(\Lambda^{p-1}\).

\[ Jf(\lambda|x,\mu)d\lambda_1d\lambda_2 \cdots d\lambda_p = 1 \text{ implies that} \]
\[ [C(x, \nu)]^{-1} = \int \int e^{-\Lambda} z^{l-1} \frac{\Gamma(j + x_1 + 1)}{\prod_{i=1}^{p-1} j!} \left( 1 + \frac{k}{p} \right)^j j! \prod_{i=1}^{p-1} \frac{\Gamma(j + x_1 + 1)}{j!} d\Lambda d\theta \]

\[ = (z-1)! \sum \frac{\Gamma(j + x_1 + 1)}{\prod_{i=1}^{p-1} j!} \frac{(j + x_1 + 1)}{\prod_{i=1}^{p} i!} \frac{\Gamma(j + x_1 + 1)}{j!} \]

where \( z = \sum_{i=1}^{p} x_1 \).

Now observe that \((x_1 + 1) \) and that

\[ \Gamma(pj + z + p) = \Gamma(z + p) \Gamma(z + p) \]

(cf. p. 22, Lemma 6 of Rainville [1960]).

Consequently, we see that

\[ [C(x, \nu)]^{-1} = \frac{(z-1)! \prod_{i=1}^{p} x_i}{(z+p-1)!} \sum_{j=0}^{p-1} \frac{(z+p-1)!}{\prod_{i=1}^{p} i!} \frac{\Gamma(j + x_1 + 1)}{j!} \frac{\Gamma(j + x_1 + 1)}{j!} \]

\[ = \frac{(z-1)! \prod_{i=1}^{p} x_i}{(z+p-1)!} 2p-1 \Gamma^{2p-2} \left[ \begin{array}{c} 1 + \frac{1}{p}, \ldots, 1 + \frac{p-1}{p}, x_1 + 1, \ldots, x_p + 1; \\ \mu \end{array} \right] \]

That is,

\[ C(x, \nu) = \frac{(z+p-1)!}{(z-1)! \prod_{i=1}^{p} x_i} \frac{\Gamma^{2p-2}}{2p-1} \left[ \begin{array}{c} 1 + \frac{1}{p}, \ldots, 1 + \frac{p-1}{p}, x_1 + 1, \ldots, x_p + 1; \\ \mu \end{array} \right]^{-1} \]
For simplicity, we use the symbol $p^*$ to represent the finite sequence $1 + \frac{1}{p}, \ldots, 1 + \frac{p-1}{p}$, and use the symbol $1^q$ to represent the finite sequence of $q$ 1's. With this notation, $C(x, \mu)$ becomes

\[
\frac{(z+p-1)!}{(z-1)! \prod \limits_{i=1}^{p} x_i!} \left\{ \begin{array}{c}
\binom{p^*, x_1+1, \ldots, x_p+1; \mu}{2p-1} \\
2p-1 \binom{2p-2}{p-2, \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p}}
\end{array} \right\}^{-1}
\]

The posterior means of the components of $\lambda$, given $x$ and $\mu$, are now easily shown to be

\[
\hat{\lambda_i} = E_{x, \mu \mid \lambda_i} = \frac{C(x, \mu)}{C(x+e_i, \mu)} , i=1, \ldots, p,
\]

where $e_i$ is a $p$-vector which has $i^{th}$ coordinate one and the other coordinates zero, and $E_{x, \mu}$ means that the expectation is taken holding $x$ and $\mu$ fixed. This expression can also be written in terms of hypergeometric functions. Accordingly, let $g(x, \mu)$ denote the generalized hypergeometric function,

\[
g(x, \mu) := \binom{p^*, x_1+1, \ldots, x_p+1; \mu}{2p-1} \binom{2p-2}{p-2, \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p}}
\]

In terms of $g$, the posterior mean of $\lambda$ becomes

\[
\hat{\lambda_i} = \frac{z(x_i+1)}{z+p} \frac{g(x+e_i, \mu)}{g(x, \mu)} , i=1, \ldots, p, \text{ where } z = \sum_{i=1}^{p} x_i.
\]

The posterior variances and covariances among the components of $\lambda$ can also be obtained. In terms of $g$, they are given as follows:
\[ \text{Var}_{x, \mu}(\lambda_i) = \mathbb{E}_{x, \mu} \lambda_i^2 - [\mathbb{E}_{x, \mu} \lambda_i]^2 \]

\[ = \frac{z(z+1)(x_i+1)(x_i+2)}{(z+p)(z+p+1)} \cdot \frac{g(x+2e_i, \mu)}{g(x, \mu)} - [\mathbb{E}_{x, \mu} \lambda_i]^2 \]

\[ = \lambda_i \left[ \frac{(z+1)(x_i+2)}{(z+p+1)} \cdot \frac{g(x+2e_i, \mu)}{g(x, \mu)} - \frac{z(x_i+1)}{z+p} \cdot \frac{g(x+e_i, \mu)}{g(x, \mu)} \right] , \]

\[ \text{Cov}_{x, \mu}(\lambda_i, \lambda_j) \]

\[ = \mathbb{E}_{x, \mu} \lambda_i \lambda_j - (\mathbb{E}_{x, \mu} \lambda_i)(\mathbb{E}_{x, \mu} \lambda_j) \]

\[ = \frac{z(z+1)(x_i+1)(x_j+1)}{(z+p)(z+p+1)} \cdot \frac{g(x+e_i+e_j, \mu)}{g(x, \mu)} - \lambda_i \lambda_j \]

\[ i=1, \ldots, p, \ j=1, \ldots, p, \ i \neq j. \]

We see that the \( \lambda_i \)'s are correlated if at least one of the \( x_i \)'s is non-zero. Moreover, since the variances are always non-negative, we have

\[ \frac{(z+1)(x_i+2)}{(z+p+1)} \cdot \frac{g(x+2e_i, \mu)}{g(x, \mu)} \geq \frac{z(x_i+1)}{z+p} \cdot \frac{g(x+e_i, \mu)}{g(x, \mu)} \quad i=1, \ldots, p. \]

That is, for fixed \( i \) the marginal posterior mean \( \mathbb{E}_{x, \mu} \lambda_i \) is a non-decreasing function of \( x_i \) (an intuitively reasonable result).

Recall that the foregoing result is based on the assumption that \( \alpha \) follows a translated geometric distribution. Similar results can be derived with different discrete prior distributions of \( \alpha \). Two examples are given below.

**Example 5.2.1.**

Suppose the prior distribution of \( \alpha \) is Poisson with counting density proportional to \( \exp(-\mu)\mu^{\alpha-1}, \ \alpha=1,2, \ldots, 0 < \mu < \infty \) and \( \mu \) is
known. The normalizing constant in this case is

\[ C(x, \mu) = \frac{(z+p-1)!}{(z-1)!} \sum_{i=1}^{p} x_i \left\{ 2p-1 \text{F}_{2p-1} \left[ \begin{array}{c} \mu \x+y+1, \ldots, x+y+1; \\ \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p} \end{array} \right] \right\}^{-1} \]

The posterior mean and the covariance matrix of \( \lambda \) can then be calculated.

\[ E(\lambda | x, \mu) = \frac{z(x_i+1)}{z+p} \frac{g(x+e_i, \mu)}{g(x, \mu)}, \quad i=1, \ldots, p, \]

where \( g(x, \mu) = 2p-1 \text{F}_{2p-1} \left[ \begin{array}{c} \mu \x+y+1, \ldots, x+y+1; \\ \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p} \end{array} \right] \).

Example 5.2.2.

Suppose the prior distribution of \( \alpha \) is negative binomial with counting density proportional to \( m+\alpha-2 \choose \alpha-1 \mu^{\alpha-1}(1-\mu)^m, \alpha=1,2, \ldots \), for some known \( m \) and \( \mu \) such that \( m > 1 \) and \( 0 < \mu < 1 \). The joint posterior density of \( \lambda \) given \( m, \mu \) and \( x \) is

\[ f(\lambda | m, \mu, x) = C(x, m, \mu) \frac{\phi}{\Lambda_p} \prod_{i=1}^{p} x_i \left\{ \begin{array}{c} p^*; \mu; \\ \frac{z+p}{p}; \ldots, \frac{z+2p-1}{p} \end{array} \right\} \]

where \( \phi = p^p \left[ \prod_{i=1}^{p} \lambda_i \right]^{\mu-1} \).

The normalizing constant, \( C(x, m, \mu) \), is given by

\[ C(x, m, \mu) = \frac{(z+p-1)!}{(z-1)!} \sum_{i=1}^{p} x_i \left\{ 2p-1 \text{F}_{2p-1} \left[ \begin{array}{c} \mu \x+y+1, \ldots, x+y+1, m; \\ \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p} \end{array} \right] \right\}^{-1} \]
The posterior mean given $x$, $m$ and $\mu$, is

$$
\mathbb{E}(\lambda | x, m, \mu) = \frac{z(x+1)}{z+p} \cdot \frac{g(x+e_i, m, \mu)}{g(x, m, \mu)}, \quad i = 1, \ldots, p,
$$

where $g(x, m, \mu) = 2p^x 2p^{-1}^p, x+1, \ldots, x+1, m$;

$$
\begin{bmatrix}
p^x, x+1, \ldots, x+1, m; \\
1, z+p, z+2p-1
\end{bmatrix}. \quad \frac{\mu}{p}, \ldots, \frac{\mu}{p}
$$

In many situations it is desirable to have greater flexibility in translating one's prior beliefs into a parametric family of prior distributions. A family which is richer in parameters will usually be sufficient. Accordingly, richer results can be obtained by choosing the prior distribution of $\alpha$ to have the following (hypergeometric) form:

$$
P(\alpha = k) = \frac{1}{w} \prod_{j=1}^{u} \left( a_j \right)^{k-1} \frac{\mu^{k-1}}{(k-1)!}, \quad k = 1, \ldots, (5.2.3)
$$

for some known $(a_1, \ldots, a_u, b_1, \ldots, b_v, \mu)$,

where:

1. $w = \prod_{j=1}^{u} v_j \left( a_j \right)^{k-1} \frac{\mu^{k-1}}{(k-1)!}$ is well defined,

$$
\begin{bmatrix}
a_1, \ldots, a_u; \\
b_1, \ldots, b_v; \\
\mu
\end{bmatrix}
$$

2. $\prod_{j=1}^{u} d_j = 1$, if $u = 0$,
\( (3) \) \( u \leq v+1 \) and

\[
\mu \geq 0, \text{ if } u < v+1,
\]

\[
0 \leq \mu < 1, \text{ if } u = v+1.
\]

Let \( a \) denote \((a_1, \ldots, a_u)\) and \( b \) denote \((b_1, \ldots, b_v)\). The resulting joint posterior density of \( \lambda \) given \( a, b, \mu \) and \( x \) is

\[
f(\lambda \mid a, b, \mu, x) = C(x, a, b, \mu) \frac{p^p \prod_{i=1}^{p} x_i^i}{\prod_{i=1}^{p+u} x_i^i Y}^\mu \begin{pmatrix} p^*, a; \\ 1_{p-1}, b; \end{pmatrix},
\]

where \( \phi = p^p \prod_{i=1}^{p} \frac{X_i}{x_i^i} \mu^\mu \).

and \([C(x, a, b, \mu)]^{-1}\)

\[
\frac{(z-1)! \prod_{i=1}^{p} x_i^i}{(z+p-1)!} = \frac{p^p \prod_{i=1}^{p} x_i^i}{\prod_{i=1}^{p+u} x_i^i Y}^\mu \begin{pmatrix} p^*, x_1+1, \ldots, x_p+1, a; \\ 1_{p-1}, \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p}, b; \end{pmatrix}.
\]

In Example 5.2.1, we had \( u=v=0 \) and \( 0 \leq \mu < \infty \). In Example 5.2.2, we had \( v=0, u=1, a_1 = m, 0 \leq \mu < 1 \). As before, to simplify the representation for the posterior density, and to aid in understanding its behavior, we define the generalized hypergeometric function \( g \) as follows:

\[
g(x, a, b, \mu) = 2p^p \prod_{i=1}^{p} x_i^i Y^\mu \begin{pmatrix} p^*, x_1+1, \ldots, x_p+1, a; \\ 1_{p-1}, \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p}, b; \end{pmatrix}.
\]
In the case of the hypergeometric family of prior distributions, we can express the posterior mean vector and the posterior covariance matrix \([\text{Cov}_{x,a,b,\mu}(\lambda_i, \lambda_j)]_{pxp}\) in terms of \(g\). Indeed,

\[
1. E_{x,a,b,\mu,\lambda_i} = \frac{z(x_i+1)}{z+p} \frac{g(x+e_i,a,b,\mu)}{g(x,a,b,\mu)}, \quad i=1, \ldots, p, \tag{5.2.5}
\]

\[
2. E_{x,a,b,\mu,\lambda_i}^2 = \frac{z(z+1)(x_i+1)(x_i+2)}{(z+p)(z+p+1)} \frac{g(x+2e_i,a,b,\mu)}{g(x,a,b,\mu)},
\]

\[
3. \text{Var}_{x,a,b,\mu,\lambda_i}(\lambda_i) = E_{x,a,b,\mu,\lambda_i^2} - [E_{x,a,b,\mu,\lambda_i}]^2
\]

\[
= \left[ E_{x,a,b,\mu,\lambda_i} \right] \left[ \frac{(z+1)(x_i+2)}{(z+p+1)} \frac{g(x+2e_i,a,b,\mu)}{g(x,e_i,a,b,\mu)} - \frac{z(x_i+1)}{z+p+1} \frac{g(x+e_i,a,b,\mu)}{g(x,a,b,\mu)} \right],
\]

\[
4. \text{Cov}_{x,a,b,\mu}(\lambda_i, \lambda_j) = \frac{z(z+1)(x_i+1)(x_j+1)}{(z+p)(z+p+1)} \frac{g(x+e_i+e_j,a,b,\mu)}{g(x,a,b,\mu)} - [E_{x,a,b,\mu,\lambda_i}][E_{x,a,b,\mu,\lambda_j}],
\]

\(i=1, \ldots, p, \quad j=1, \ldots, p, \quad i\neq j.\)

The marginal posterior mean, \(E_{x,a,b,\mu,\lambda_i}\), is seen to be a non-decreasing function of \(x_i\). This follows directly since \(\text{Var}_{x,a,b,\mu,\lambda_i}(\lambda_i) \geq 0\). The MLE of \(\lambda_i, x_i\), is obviously an increasing function of \(x_i\).

Since the posterior moments involve the function \(g(x,a,b,\mu)\), we would like to investigate the properties of \(g\). In particular, we would like to compare the value \(g(x+e_i,a,b,\mu)\) and the value \(g(x,a,b,\mu)\). A sufficient condition that \(g(x+e_i,a,b,\mu)\) is greater than or equal to
\( g(x, a, b, \mu) \) is that each individual term of the infinite sum of 
\( g(x+e_i, a, b, \mu) \) is greater than or equal to the corresponding term of the 
infinite sum of \( g(x, a, b, \mu) \). In other words, 
\[
g(x+e_i, a, b, \mu) \geq g(x, a, b, \mu)
\]

\[
p^{-1} \prod_{j=1, j \neq i}^p \left( \frac{x_j + 1}{\alpha} \right) u 
\]

\[
\text{if } \frac{1}{(x_i + 1)^{\alpha}} \prod_{j=1}^p (x_j + 2)^{\alpha} \prod_{j=1}^p (a_j)^{\alpha}
\]

\[
\frac{(1)^{p-1} \prod_{k=1}^p \left( \frac{z+p+k}{\alpha} \right) \prod_{j=1}^p (b_j)^{\alpha}}{(1)^{p-1} \prod_{k=0}^p \left( \frac{z+p+k}{\alpha} \right) \prod_{j=1}^p (b_j)^{\alpha}}
\]

for \( \alpha = 0, 1, \ldots \).

The condition can be simplified to

\[
\frac{(x_i + 2 + \alpha - 1)}{z + 2 p + (\alpha - 1)} \geq \frac{(x_i + 1)}{z + p}, \text{ for } \alpha = 1, 2, \ldots
\]

which is equivalent to

\[
\frac{z}{p} \geq x_i.
\]

Consequently, if the grand mean, \( \sum_{i=1}^p \frac{x_i}{p} \), is strictly greater than \( x_i \),
then \( g(x+e_i, a, b, \mu) \geq g(x, a, b, \mu) \), and \( g(x+e_i, a, b, \mu) = g(x, a, b, \mu) \) if
\[
\sum_{j=1}^p x_j / p = x_i.
\]
On the other hand, if \( \sum_{j=1}^p x_j / p < x_i \), then
\( g(x+e_i, a, b, \mu) < g(x, a, b, \mu) \). In fact, if the grand mean is less than \( x_i \),
\( g(x, a, b, \mu) \) is a strictly decreasing function of \( x_i \).

Now suppose that all the \( x_i \)'s are identical and equal to \( x_0 \), say.

Then the marginal posterior means are
\[
E_{x,a,b,\mu^i} \frac{z(x_i+1)}{z+p} \cdot \frac{g(x+e_i,a,b,\mu)}{g(x,a,b,\mu)} = x_0', \ i=1, \ldots, p.
\]

That is to say, \( E_{x,a,b,\mu^i} \) equals the MLE in this case. This is a natural result. For, the assumption of having an exchangeable prior distribution of \( \lambda \) implies that our prior beliefs about the \( \lambda_i \)'s is that they are "probabilistically close" to one another. Then with identical observations of the \( x_i \)'s, we have supporting evidence that the \( \lambda_i \)'s are very close, or even identical to one another. Hence, the grand mean \( x_0 \) is an appropriate estimate for each of the \( \lambda_i \)'s.

5.3 Marginal Posterior Density

In this section, we will derive the marginal posterior density of \( \lambda_i \), for \( i=1, \ldots, p \). Recall that the joint posterior density of the \( \lambda_i \)'s is given by

\[
f(\lambda|x,a,b,\mu) = C(x,a,b,\mu) \frac{\prod_{i=1}^{p} \lambda_i^{x_i}}{\Lambda^p} p+u-1 \cdot p+v-1 \cdot \phi,
\]

where \( p^* = (1 + \frac{1}{p}, 1 + \frac{2}{p}, \ldots, 1 + \frac{p-1}{p}) \),

\[
\phi = \mu^p \left[ \prod_{i=1}^{p} \lambda_i \right]^{-p} \quad \text{and} \quad \Lambda = \sum_{i=1}^{p} \lambda_i.
\]
The joint posterior density of \( \lambda \) can be thought of as a product of three functions of \( \lambda \). The first function, \( e^{-\lambda} \prod_{i=1}^{p} \lambda_i \), is maximized when \( \lambda_i = x_i, \ i=1, \ldots, p \). The second function, \( \frac{1}{\lambda^p} \), increases as the \( \lambda_i \)'s decrease. The third function,

\[
\begin{bmatrix}
    p^*, a; \\
p+u-1 \\
1-p! b;
\end{bmatrix}
\]

is maximized when the \( \lambda_i \)'s are equal to one another. Hence the posterior mode is shifted from \( x \) to a position where the \( \lambda_i \)'s are more nearly equal and towards the origin.

The posterior mean can be used as an estimator of \( \lambda \); it lies between the MLE and the grand mean \( \bar{x} \). To see this, we utilize another representation of the posterior mean. Conditioning on \( \alpha \), the data, and the prior information, the posterior mean of \( \lambda_i \) is:

\[
(\frac{x}{x+\alpha}) x_i + (\frac{\alpha}{x+\alpha}) \bar{x},
\]

which is a convex combination of \( x_i \) and \( \bar{x} \). The posterior mean of \( \lambda_i \) is therefore equal to

\[
E_{x,a,b,\mu} \lambda_i = E_{x,a,b,\mu} \left[ (\frac{x}{x+\alpha}) x_i + (\frac{\alpha}{x+\alpha}) \bar{x} \right]
\]

where \( E_{x,a,b,\mu} \) is the expectation taken with respect to the posterior distribution of \( \alpha \). Now, for all \( \alpha \), the expression inside the brackets lies between the MLE \( x_i \), and the grand mean \( \bar{x} \). Hence the posterior mean of \( \lambda_i \) also lies between \( x_i \) and \( \bar{x} \). This does not mean that the posterior mean of \( \lambda_i \) an estimate of \( \lambda_i \) is necessarily better than the MLE \( x_i \) for a
particular $i$. We only expect that most of the posterior means of
$\lambda_i$, $i=1,\ldots,p$, are closer to the true parameters than the MLE. For a
particular $\lambda_i$, the squared error loss in using the Bayes estimator may
be greater than that of the MLE. However, the total squared error loss
is expected to be much smaller than that of the MLE when the $\lambda_i$'s are
"probabilistically close" to one another. On the other hand, when the
$\lambda_i$'s are spread across a wide range, the Bayes estimator is not expected
to perform much better than the MLE. In fact, there may be cases where
the MLE is better. Such a case is discussed in Section 7.

Next, we find the marginal posterior density for $\lambda_i$ in several
instances, first, when the distribution of $\alpha$ is degenerate, and then,
in general. Results for various specific distributions of $\alpha$ will be
studied numerically.

Consider first, the case when the distribution of $\alpha$ is degenerate,
i.e. $p(\alpha=s) = 1, s > 0$. The posterior density of $\lambda$ given $x$ and $s$ is

$$f(\lambda|x,s) = C(x,s) \frac{e^{-A} \prod_{i=1}^{p} x_i^{s-1}}{\Gamma(ps) \lambda^{ps}} .$$

In order to find the marginal posterior density of $\lambda_i$, we introduce the
following notation:

**Definitions**

(1) $z(i) = \sum_{j \neq i} x_j$

(2) $A(i) = \sum_{j \neq i} \lambda_j$

(3) $\theta(i) = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$.
Now transform $\lambda$ according to:

1. $\lambda_i \rightarrow \lambda'_{i}$

2. $\lambda_j \rightarrow \Lambda(i) \theta_j$ where $\Sigma_{k \neq i} \theta_k = 1$ and $j \neq i$.

The Jacobian of this transformation is $\Lambda^{p-2}(i)$. After the transformation, $f(\lambda|x,s)$ becomes $f(\lambda', \Lambda(i), \theta(i)|x,s)$.

$$f(\lambda|x,s) \propto e^{-\lambda_i x_i + s - 1} \prod_{j \neq i} \Theta_j \left( \begin{array}{c} x_j + s - 1 \\ \Lambda(i) \end{array} \right) \frac{\Lambda(i)^{-\lambda_i x_i + s - 1}}{(\lambda_i + \Lambda(i))^{p-1}}.$$

(Note that $f(\cdot)$ is being used generically to denote the density of the indicated arguments).

Integrating out $\Lambda(i)$ and $\theta(i)$, we have the marginal posterior density of $\Lambda_i$ given by

$$f(\lambda_i|x,s) \propto e^{-\lambda_i x_i + s - 1} II(\lambda_i),$$

where

$$II(\lambda_i) = \frac{\int_{0}^{\infty} e^{-y} y^{z(i)+(p-1)s-1}}{(\lambda_i + y)^{ps}} dy.$$

Let $t = (y/\lambda_i)$. Then $II(\lambda_i)$ becomes

$$\frac{z(i)^{-s}}{\lambda_i^{s}} \int_{0}^{\infty} e^{-\frac{t}{\lambda_i} + \frac{z(i)+(p-1)s-1}{(1+t)^{ps}}} dt.$$
Hence the marginal posterior density of $\lambda_1$ may be written

$$f(\lambda_1|x,s) = e^{-\lambda_1 x_i} \int_0^\infty e^{\frac{-\lambda_1 t}{t^{p-1}} (p-1)s-1} \frac{t^{p-1}}{(1+t)^{ps}} dt.$$ 

It is reasonable to assume that at least one of the $x_i$'s is non-zero and hence, $z > 0$ is a reasonable assumption. In this case, $f(\lambda_1|x,s)$ is a proper density.

Let us examine the property of the marginal posterior density $f(\lambda_1|x,s)$ of $\lambda_1$ given $x$ and $s$.

If $z_{(i)} < s$, then the integral

$$I_{s-1}(\lambda_1) = \int_0^\infty e^{\frac{-\lambda_1 t}{t^{p-1}} (p-1)s-1} \frac{t^{p-1}}{(1+t)^{ps}} dt \quad (s > 0)$$

is finite even when $\lambda_1$ is zero. If $s=1=z_{(i)}$, we have

$$f(\lambda_1|x,s) = e^{-\lambda_1 x_i} \int_0^\infty e^{\frac{-\lambda_1 t}{t^{p-1}}} \frac{t^{p-1}}{(1+t)^{p}} dt,$$

and $f(\lambda_1|x,s)$ is strictly decreasing if $x_1 = 0$. It is straightforward to check that the integral is a monotone decreasing, convex function of $\lambda_1$. Moreover, $f(\lambda_1|x,s)$ tends to infinity as $\lambda_1$ tends to zero. This is not surprising because $x_1 = 0$ and $z = 1$ would indicate that $\lambda_1$ is very likely to be near zero. In general, we would expect the shape of the marginal posterior density to be gamma-like. If $x_1 > 0$ and $s > 1$, the curve of $f(\lambda_1|x,s)$ is again unimodal and skewed to the right.

Now, we consider the cases in which $\alpha$ has the discrete distributions suggested in this paper. The derivation of the marginal posterior
density of $\lambda_1$ is straightforward but tedious. In fact,

$$f(\lambda_1 | x, a, b, u) = e^{-\lambda_1} \sum_{s=0}^{\infty} \frac{p^s \prod_{j \neq i} (x_j + 1) \cdot \prod_{k=1}^{p-1} (1 + \frac{k}{p}) \cdot \prod_{j=1}^{u} (a_j) \cdot \prod_{j=1}^{v} (b_j) \cdot (p-1)^{s+1} \cdot \prod_{s=0}^{p-2} (1 - \frac{(i+p-1+k)}{p-1}) \cdot \prod_{j=1}^{v} (p-1)^{(p-1)s} \cdot \prod_{j=1}^{v} (b_j)^{s}}{s! \prod_{k=0}^{p-2} (\frac{(i+p-1+k)}{p-1})}$$

Define the hypergeometric function $h_{(i)}(x, \mu, t)$

$$h_{(i)}(x, \mu, t) = \sum_{s=0}^{\infty} \frac{p^s \prod_{j \neq i} (x_j + 1) \cdot \prod_{k=1}^{p-1} (1 + \frac{k}{p}) \cdot \prod_{j=1}^{u} (a_j) \cdot \prod_{j=1}^{v} (b_j) \cdot (p-1)^{s+1} \cdot \prod_{s=0}^{p-2} (1 - \frac{(i+p-1+k)}{p-1}) \cdot \prod_{j=1}^{v} (p-1)^{(p-1)s} \cdot \prod_{j=1}^{v} (b_j)^{s}}{s! \prod_{k=0}^{p-2} (\frac{(i+p-1+k)}{p-1})} \cdot \left[ \frac{\mu^p t^{p-1}}{(p-1)(p-1)!(1+t)^p} \right]$$

$$= 2p-2+u \cdot 2p-2+v \cdot p^{p-1} \cdot x_{i+1}, \ldots, x_{i-1+1}, x_{i+1+1}, \ldots, x^{+1}, a; \right\}
\begin{bmatrix}
1_{p-1}, \frac{z(i)+p-1}{p-1}, \ldots, \frac{z(i)+2p-3}{p-1}, b;
\end{bmatrix}
\right)$$

where $w = \frac{\mu^p t^{p-1}}{(p-1)(p-1)!(1+t)^p}$.

We then have for the general (non-degenerate $a$) marginal posterior density,

$$f(\lambda_1 | x, a, b, u) = e^{-\lambda_1} \sum_{z=1}^{\infty} f_{\lambda_1 z} \int_0^\infty e^{-\frac{\lambda_1 z}{t}} \cdot h_{(i)}(x, \mu, t) dt.$$
It can be checked that the normalizing constant is

\[
C(x,a,b,\mu) = \frac{p}{\prod_{j \neq 1} \Gamma(x_j + 1)/\Gamma(z_{(i)} + p - 1)}
\]

\[
= \frac{(z + p - 1)!}{(z - 1)! \prod_{j=1}^p x_j!} \left\{ \begin{array}{c}
\frac{p^x_{z+1, \ldots, z+p-1}}{2p+u-1^z+2p-v-1} \\
1_{p-1}, \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p}, b;
\end{array} \right\}^{-1} 
\]

\[
\cdot \frac{\prod_{j \neq 1} x_j!}{(z_{(1)} + p - 2)!}
\]

\[
= \frac{(z + p - 1)!}{(z - 1)! (z_{(1)} + p - 2)! x_1!} \left\{ \begin{array}{c}
\frac{p^x_{z+1, \ldots, z+p-1}}{2p+u-1^z+2p-v-1} \\
1_{p-1}, \frac{z+p}{p}, \ldots, \frac{z+2p-1}{p}, b;
\end{array} \right\}^{-1} 
\]

The shape of this density function is also gamma-like, by the same argument as before.

5.4 Summary

In this section we proposed Bayes estimators for simultaneously estimating the parameters of \( p \) distributions of independent Poisson random variables when the prior distribution of the parameters is assumed to be exchangeable. Substantial improvement over the usual procedure (which is the MLE) is expected when the parameters are close to one another, especially when \( p \) is large, because the assumption of exchangeability implies that the larger \( p \) is, the more information we have about the \( \lambda_i \)'s. Our claim is supported in Section 7 by the results
of a computer simulation designed to compare the estimation efficiency of the Bayes estimators with the MLE. The measure used to assess the performance of the estimators is mean squared error, which is often a reasonable measure of the overall adequacy of an estimator.
SECTION 6. EMPIRICAL BAYES ESTIMATION

6.1 Background

In estimating the mean of a multivariate normal random vector, Efron and Morris [1973] give an interpretation of the James-Stein estimator from an empirical Bayes point of view. If we let

\[ V = (V_1, \ldots, V_p) \sim N(\theta, I_p), \quad S = \sum_{i=1}^{p} V_i^2, \quad \text{and} \quad p \geq 3, \]

where \( I_p \) is the \( p \times p \) identity matrix, then the James-Stein estimator of \( \theta \) is

\[ (1 - (p-2)/S)V. \]

Efron and Morris assume that the coordinates \( \theta_i \) of \( \theta \) are independently and identically \( \sim N(0, \sigma^2) \). In this situation, the Bayes rule is \( \hat{\theta}^* = (1 - B)V \), with \( B = 1/(1 + \sigma^2) \). The marginal distribution of \( V \) is \( N_p(0, (1+\sigma^2)I_p) \) and \( S \) is distributed as \( (1+\sigma^2)\chi^2_p \), a multiple of a chi-square distribution with \( p \) degrees of freedom. Replacing \( B \) by \( B(S) = (p-2)/S \), which is an unbiased estimate of \( B \), yields the empirical Bayes estimator \( \hat{\theta} = (1 - \hat{B}(S))V \) which is the James-Stein estimator. Efron and Morris perform much of their analysis based on what they define as the "relative savings loss", i.e. the regret from using the empirical Bayes rule \( \hat{\theta} \) instead of the actual Bayes rule divided by the corresponding regret if the MLE, \( V \), is used instead of the Bayes rule. If the risk of the estimator \( \hat{\theta} \) is denoted by

\[ R(B, \hat{\theta}) = E_B R(\theta, \hat{\theta}), \]

where \( E_B \) indicates that expectation is taken with respect to the above prior distribution, the relative savings loss is

\[ \text{RSL}(B, \hat{\theta}) = \frac{R(B, \hat{\theta}^*) - R(B, \hat{\theta})}{R(B, V) - R(B, \hat{\theta}^*)}. \]

Under the above assumptions, straightforward calculations show that \( \text{RSL}(B, \hat{\theta}) \) can be written as \( \text{RSL}(B, \hat{\theta}) = E_B \{ (\hat{B}(S) - B) / B \}^2 \). Here \( E_B \) indicates expectation with respect to \( S \sim (1/B)\chi^2_{p+2} \). The empirical
Bayes approach thus reduces the p-dimensional problem of estimating 
\( \theta = (\theta_1, \ldots, \theta_p) \) from \( V = (V_1, \ldots, V_p) \) to the one-dimensional problem of estimating \( B \), or more precisely, \( 1/(1+A) \).

6.2 Relative Savings Loss in the Poisson Case

In this section, we use the normalized squared error loss

\[
L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i}
\]

as our measure for the loss in estimating \( \lambda_i \) by \( \hat{\lambda}_i \), \( i = 1, \ldots, p \). This is the loss function employed by Clevenson and Zidek [1975]. As they suggested, the above approach used on the multivariate normal estimation problem seems applicable to our problem of simultaneously estimating the parameters \( \lambda_1, \ldots, \lambda_p \) of the independent Poisson variables \( X_1, \ldots, X_p \). Our developments here parallel those in the normal case. Although no new estimators are found, the empirical Bayes approach provides an alternate way to view our problem and derive the estimators given in Theorem 2.2.9 so that we have a better understanding of those estimators.

We will next derive the "relative savings loss" for the Poisson case. Let \( \lambda_1, \ldots, \lambda_p \) be independently and identically distributed with prior distribution \( a \chi^2_2 \), a scalar multiple of a chi-square distribution with two degrees of freedom. The joint prior density of \( \lambda_1, \ldots, \lambda_p \) is proportional to \( \exp(-\sum \lambda_i/a) d\lambda_1 \cdots d\lambda_p \) for \( \lambda_i > 0 \), and zero otherwise. The Bayes estimator of \( \lambda \) when such a prior is used is \( (1-b)X \), where \( b = 1/(1 + a) \). This estimator has Bayes risk \( p(1 - b) \), while the MLE, \( X \), has risk \( p \). Thus, the regret at using the MLE instead of the Bayes estimator is \( pb \). The marginal distribution of \( Z = \sum_{i=1}^{p} X_i \), given \( b \), is
the negative binomial distribution $NB(b,p)$ with probability mass function $Pr(z|b) = \binom{z+p-1}{z} b^p (1-b)^z$, for $z \in J^+$. Suppose $b$ is to be estimated by $\hat{b}(Z)$. Then the regret at using the empirical Bayes estimator $\hat{\lambda} = (1-\hat{b}(Z))X$ instead of the Bayes rule $(1-b)X$ is

$$E_b E_X \left[ \sum_{i=1}^{p} \frac{(1 - \hat{b}(Z))X_i - \lambda_i^2}{\lambda_i^2} - p(1 - b) \right] = E_b \left( \frac{1}{1 - \hat{b}(Z)} \right) \left[ Z\hat{b}^2(Z) - 2\hat{b}(Z)Z + 2\hat{b}(Z)Z(1-b) \right] + p - p(1-b)$$

$$= E_b \left[ \frac{Z(\hat{b}(Z) - b)^2}{(1 - b)} \right] / (1 - b),$$

where $E_b$ denotes expectation with respect to the marginal distribution of $Z$ as given above. The relative savings loss in this case is

$$RSL(b, \hat{\lambda}) = E_b \left[ Z(\hat{b}(Z) - b)^2 / pb(1 - b) \right]$$

$$= \sum_{z=0}^{\infty} \frac{(\hat{b}(z) - b)^2}{(1-b)pb} \cdot \frac{(z+p-1)!}{z!(p-1)!} \cdot b^p (1-b)^z$$

$$= \sum_{z=1}^{\infty} \frac{(\hat{b}(z) - b)^2}{(1-b)pb} \cdot \frac{(z+p-1)!}{(z-1)!(p-1)!} \cdot b^p (1-b)^z$$

$$= \sum_{z=0}^{\infty} \frac{(\hat{b}(z+1) - b)^2}{b^2} \cdot \frac{(z+p)!}{z!p!} \cdot b^{p+1} (1-b)^z$$

$$= E_b \left[ \frac{(\hat{b}(Z+1) - b)}{b} \right]^2$$

where $E_b$ denotes expectation with respect to the distribution of $Z$ and $Z \sim NB(b,p+1)$. 
6.3 The Plus Rules

Suppose we have an estimator of the form $\hat{\lambda} = (1 - b(Z))X$ and that there is a positive probability that $b(Z) > 1$. The following proposition shows that we can always improve the relative savings loss RSL of such an estimator by replacing $b(Z)$ with $b^+(Z) = \min\{1, b(Z)\}$.

Intuitively, we would expect that the estimator $\hat{\lambda} = (1 - b(Z))X$ can be improved upon by replacing $b(Z)$ with 1 when $b(Z)$ is greater than 1, because the Bayes estimator is $(1 - b)X$, where $b$ is known to be between 0 and 1. Hence the proposition is a very natural result.

**Proposition 6.3.1.**

Let $\hat{\lambda}^+ = (1 - b^+(Z))X$. Then $RSL(b, \hat{\lambda}) - RSL(b, \hat{\lambda}^+) = (1/b^2) \frac{1}{b} \{ [b(Z+1) - b^+(Z+1)] + (1-b)]^2 - [1 - b]^2 \}$.

The function of $Z$ in the outermost brackets is nonnegative and strictly greater than 0 if $b(Z) > 1$, so that $RSL(b, \hat{\lambda}) \geq RSL(b, \hat{\lambda}^+)$ for all $b$, with strict inequality if $Pr\{b(Z) > 1\}$ is positive for any value of $b$.

The proof of the above proposition is immediate and hence is omitted. The estimator $\hat{\lambda}^+$ may be called the "plus rule" of $\hat{\lambda}$.

6.4 Bayes Rules with Respect to Other Priors

In this subsection, we shall consider some Bayes rules with respect to priors which are members of the family considered by Clevenson and Zidek [1975]. We reparametrize the parameters $\lambda_1, \ldots, \lambda_p$ as $(\theta_i, \Lambda)$, $i = 1, \ldots, p$, where $\Lambda = \sum_{i=1}^{p} \theta_i$ and $\theta_i = \lambda_i / \Lambda$, and suppose the joint prior distribution of $(\theta_i, \Lambda)$, $i = 1, \ldots, p$, is proportional to

$$\exp(\frac{-\Lambda/a}{\Lambda_0} \prod_{i=1}^{p} d\theta_i)$$

when $\Lambda > 0$ and $\sum_{i=1}^{p} \theta_i = 1$, and zero otherwise.

\[ (6.4.1) \]
where \( a \) is a positive integer. The Bayes estimator with respect to this prior distribution is 
\[
\hat{\lambda}_a = (1-b)(Z+a)X / (Z+p-1),
\]
where \( b = 1/(1+a) \).

The marginal distribution of \( Z \) has probability mass function \( \text{NB}(b, a+1) \).

Since \( \mathbb{E}_b \{ Z/(Z+\alpha) \} = (1-b) \), estimating \( (1-b) \) by \( Z/(Z+\alpha) \) leads to the empirical Bayes estimator
\[
\hat{\lambda}^*(X) = \left[ \frac{Z}{Z+p-1} \right] X = \left[ 1 - \frac{(p-1)}{(Z+p-1)} \right] X,
\]
which is independent of \( a \). This estimator belongs to the class of estimators mentioned in Theorem 2.2.9. The risk of \( \hat{\lambda}_a^* \) can be calculated as follows.

\[
R(b, \hat{\lambda}_a) = \mathbb{E}_b \left[ \mathbb{E}_{\Lambda; \theta} \sum_{i=1}^{p} \frac{[(1-b)\frac{Z+\alpha}{Z+p-1} X_i - \Lambda \theta_i]^2}{\Lambda \theta_i} \right]
\]

\[
= \mathbb{E}_b \left[ \mathbb{E}_Z \left( \frac{1}{\Lambda} \right) \cdot \{(1-b)^2 \left( \frac{Z+\alpha}{Z+p-1} \right)^2 \cdot Z(Z+p-1) - 2(1-b) \frac{Z+\alpha}{Z+p-1} \Lambda + \Lambda^2} \right]
\]

where \( \mathbb{E}_b \) indicates expectation with respect to the prior stated above and \( \mathbb{E}_Z \) indicates expectation with respect to Poisson \( (\Lambda) \). Note that

\[
\mathbb{E}_{Z;b} \Lambda = [(Z+a+1)/(1+(1/a))] = (1-b)(Z+a+1) \text{ and}
\]

\[
\mathbb{E}_{Z;b} \left( \frac{1}{\Lambda} \right) = 1 / \{(1-b)(Z+a)\},
\]

where \( \mathbb{E}_{Z,b} \) is the expectation taken with respect to the posterior distribution of \( \Lambda \) given \( Z \) and \( b \). Hence

\[
R(b, \hat{\lambda}_a) = \mathbb{E}_b \left\{ (1-b)(Z+a+1) - (1-b)(Z+a)\alpha/(Z+p-1) \right\}
\]

\[
= p(1-b) + (1-b)(a+1-p) \mathbb{E}_b \left[ (p-1)/(Z+p-1) \right]
\]
where $\mathbb{E}_b$ denotes expectation with respect to the probability mass function $NB(b, \alpha + 1)$. For example, when $\alpha = p$,

$$R(b, \lambda_p) = p(1-b) + (1-b)[b^2 + (p-1)b(1-b)/p]$$

$$= p - b[p - 1 + 1/p + (1-2/p)b + b^2/p]$$

$$\leq p \text{ for all } b \in (0,1].$$

6.5 **Truncated Bayes Rules**

In this subsection, we consider again the class of estimators of the form $(1 - b(Z))X$. Theorem 2.2.9 provides a class of such estimators dominating the MLE. We will attempt to find conditions under which they are also Bayes with respect to some prior distribution. As before, $X_i \sim \text{Poisson } (\lambda_i)$ independently, $i = 1, \ldots, p$. Suppose each

$$\lambda_i \sim (1/\alpha) \exp(\lambda_i/a),$$

and let $b = 1/(1+a)$. Furthermore, we suppose that $b \sim h(b)$, where $h(b)$ is a prior probability density function of $b$ putting all of its probability on the interval $(0,1]$. In this setting, the Bayes rule under normalized squared loss is calculated to be

$$\hat{\lambda}_h^*(x) = \mathbb{E}_h [E(\lambda | x, b)]$$

$$= \int_0^1 (1-b) \lambda h_Z(b) \ db$$

$$= [1 - b^*_h(Z)]X \quad (6.5.1)$$

where $\mathbb{E}_h$ indicates expectation with respect to the situation described above, $h_Z(b)$ is the conditional density of $b$ given $Z$, and $b^*_h(Z)$ is the conditional expectation of $b$ given $Z$. The Bayes rule $\hat{\lambda}_h^*$ is thus of the form that we have been considering and Theorem 2.2.9 is applicable.
The following lemma gives equivalent definitions of the Bayes rule with respect to \( h \). In particular, an alternate definition which proves to be more convenient for our analysis than the usual one is suggested.

**Lemma 6.5.1.**

The following are equivalent:

(i) \( \hat{\lambda}_h^* \) is that \( \hat{\lambda} \) which minimizes \( \int_0^1 R(b, \hat{\lambda}) h(b) \, db \).

(ii) \( \hat{\lambda}_h^* \) is that \( \hat{\lambda} \) which minimizes \( \int_0^1 [R(b, \hat{\lambda}) - p(1-b)]/pb \, g(b) \, db \)
where \( g(b) = pbh(b) \).

(iii) \( \hat{\lambda}_h^* \) is that \( \hat{\lambda} = (1 - b(Z))X \) for which \( b(Z) \) minimizes

\[
\int_0^1 b(pZ + 1) - b \, g(b) \, db \]

where \( E_b \) means expectation with respect to \( Z \sim NB(b, p+1) \).

**Proof:**

Form (i) is the usual definition; form (ii) follows directly because the minimization is over \( \hat{\lambda} \). Form (iii) is equivalent to form (ii) by (6.2.1). Q.E.D.

Finding the Bayes rules having the form \( \hat{\lambda} = (1 - b(Z))X \) is thus equivalent to finding \( \hat{b} \) which minimizes (6.5.2). The \( \hat{b}(z) \) that gives the minimization in (iii) is \( b_h^*(z) \), which is given by

\[
b_h^*(z+1) = \frac{\int_0^1 g(b) b^p (1-b)^Z \, db}{\int_0^1 g(b) b^{p-1} (1-b)^Z \, db} \quad (6.5.3)
\]

or

\[
b_h^*(z+1) = \frac{\int_0^1 h(b) b^{p+1} (1-b)^Z \, db}{\int_0^1 h(b) b^p (1-b)^Z \, db} \quad (6.5.4)
\]
If we take the improper prior \( h(b) = 1/b^2 \) or \( g(b) = p/b \), then the Bayes rule is \((1 - b^\star(Z))X\) with

\[
\frac{1}{b} \int_0^1 b^{-2} b^{p+1} (1-b)^Z \, db \\
\frac{1}{b} \int_0^1 b^{-2} b^p (1-b)^Z \, db \\
= (p-1)/(z+p)
\]

or \( b^\star(h(z)) = (p-1)/(z+p-1) \).

Thus the empirical Bayes estimator \( \lambda^\star = (1 - (p-1)/(Z+p-1))X \) derived in Section 6.4 is actually a Bayes rule.

We next inquire if the Bayes rules thus obtained will still dominate the MLE. Let \( \phi_s \) be the class of estimators

\[
\{1 - [(p-1) \phi(Z)/(Z+p-1)]\}X
\]

described in Theorem 2.2.9 satisfying \( \lim_{z \to \infty} \phi(z) = s \leq 2 \). The following theorem gives the class of estimators obtained by modifying the Bayes rule obtained in this section so that they are in \( \phi_s \) for some \( 0 < s < 2 \).

**Theorem 6.5.5.**

Suppose that \( h(b) \) is such that

\[
\phi_h^\star(z) = b_h^\star(z)/[(p-1)/(Z+p-1)]
\]

is nonnegative and nondecreasing in \( z \). The estimator in \( \phi_s \) which minimizes the Bayes risk versus \( h \), \( E_h R(b, \lambda) \), is given by

\[
\lambda_h^S(X) = \{1 - [(p-1)/(Z+p-1)]\phi_h^S(z)\}X
\]

where \( \phi_h^S(z) = \min \{s, \phi_h^\star(z)\} \).
Proof:

Condition on Z=z, and let $g_z(b) = pbh_z(b)$. Then $g_z(b) = g(b)b^{p+1}(1-b)^z$
where $g(b) = pbh(b)$ as before. For any estimator $\lambda = (1 - b(Z))X$,

$$
\int_0^1 [(b(z+1)-b)/b]^2 g_z(b) \, db
= \int_0^1 [(b(z+1) - b_h^{(z+1)})/b]^2 g_z(b) \, db + [(b_h^{(z+1)} - b)/b]^2 g_z(b) \, db.
$$

The cross term is zero because of formula (6.5.3) for $b_h^{(z+1)}$. In terms of

$$
\phi(z) = b(z)/[(p-1)/(z+p-1)],
$$

the above equality can be written as

$$
\int_0^1 [(b(z+1) - b)/b]^2 g_z(b) \, db
= \int_0^1 [(\phi(z+1) - \phi_h^{(z+1)})^2 [(p-1)/(z+p-1)]^2 g_z(b) \, db
+ \int_0^1 [(b_h^{(z+1)} - b)/b]^2 g_z(b) \, db.
$$ (6.5.6)

It is seen that $\phi_h^S(z)$ minimizes the right hand side of the above equality
for all z among $\phi(z)$ giving rules in $\phi_h^S$.

Integrating over the marginal distribution of Z shows that $\lambda_h^S$ minimizes the integral $\int_0^1 RSL(b, \lambda) g(b) \, db$ for $\lambda$ in $\phi_h^S$.

Q.E.D.

The estimator thus obtained is called a "truncated Bayes rule", a
term introduced by Efron and Morris [1973].

We now define $b_h^S(z) = [(p-1)/(z+p-1)]\phi_h^S(z)$
with $s \leq \lim_{Z \to \infty} \phi_h^S(z)$, and let $\lambda_h^S(X) = (1 - b_h^S(Z))X$. The following lemma
shows that the risk of $\lambda_h^S$ is a decreasing convex function of $s$. 

Lemma 6.5.7.

\[ E_h R(b, \hat{\lambda}_h^s) \] is a strictly decreasing convex function of s.

Proof:

Since \( \{ \phi_h^s(z) - \phi^*(z) \}^2 = [\text{Max} \{0, \phi^*_h(z) - s\}]^2 \) is a decreasing convex function of s, the result follows from (6.5.6).

Q.E.D.

6.6 The Risk Function of the Estimator \( \hat{\lambda}^* \)

As remarked in the previous subsection, \( \hat{\lambda}^* = [1 - (p-1)/(Z+p-1)]X \) can be viewed as an empirical Bayes estimator under the assumption that the prior distribution of \( \lambda \) is given by (6.4.1). Since \( \hat{\lambda}^* \) is independent of \( a \), this estimator is robust in that any member of a whole family of prior distributions may be chosen and still we arrive at the same empirical Bayes estimator. We therefore proceed to calculate the risk function \( R(\lambda, \hat{\lambda}^*) \) of the estimator \( \hat{\lambda}^* \). We first derive an expression for the risk function as a function of b.

As in subsection 6.2, we assume that \( \lambda_1, \ldots, \lambda_p \) are independently and identically distributed with prior distribution \( a \lambda_2^2 \). Equivalently, the prior is of the form given above in (6.4.1), with \( a = p - 1 \). The RSL can then be calculated as follows.

\[
RSL(b, \hat{\lambda}^*|a=p-1) = \sum_{b} \left[ \frac{(p-1)/(z+p)}{(z+p)!(b^2z!p!)}b^{p+1}(1-b)^z - 2(p-1)/p + 1 \right]
\]
\[= (p-1)^2 b^{p-1} \sum_{z=0}^{\infty} \frac{1}{(z+p)}[(z+p-1)!/(z!p!)](1-b)^z - 2(p-1)/p + 1 \]

\[= [(p-1)^2 b^{p-1}/(p(1-b)^p)] \sum_{z=0}^{\infty} [(z+p-1)!/(z!(p-1)!)] \int_{0}^{1} (1-t)^{z+p-1} \, dt \]

\[- 2(p-1)/p + 1 \]

\[= [(p-1)^2 b^{p-1}/(p(1-b)^p)] \int_{0}^{1} \sum_{z=0}^{\infty} [(z+p-1)!/(z!(p-1)!)] (1-t)^{z+p-1} \, dt \]

\[- 2(p-1)/p + 1 \]

\[= [(p-1)^2 b^{p-1}/(p(1-b)^p)] \int_{0}^{1} (1-t)^{p-1}/T^p \, dt - 2(p-1)/p + 1. \]

Observe that as \(b\) tends to zero, \(RSL(b, \lambda | \alpha = p-1)\) tends to \((p-1)/p - 2(p-1)/p + 1 = 1 - (p-1)/p = 1/p\). As \(b\) tends to one, \(RSL(b, \hat{\lambda} | \alpha = p-1)\) tends to \((p-1)^2/p^2 - 2(p-1)/p + 1 = 1/p^2\).

Using the above result to calculate the risk of the estimator as a function of \(b\), we have

\[R(b, \hat{\lambda} | \alpha = p-1) = RSL(b, \hat{\lambda} | \alpha = p-1)pb + p(1-b) \]

\[= (p-1)^2 b^p/(1-b)^p \int_{0}^{1} (1-t)^{p-1}/t^p \, dt - 2(p-1)b \]

\[+ pb + p(1-b) \]

\[= p - 2(p-1)b + (p-1)^2 b^p/(1-b)^p \int_{0}^{1} (1-t)^{p-1}/t^p \, dt. \]

Recall that this expression depends on the choice of \(\alpha\). Since this is not a simple expression, in the next subsection we explore the possibility of simplifying the expression by varying \(\alpha\).
6.7 **Simplification of** $R(b, \hat{\lambda}^*)$

In our attempt to obtain a simpler expression for $R(b, \hat{\lambda}^*)$, we now adopt another prior distribution. Suppose the joint prior distribution is as given in (6.4.1), with $\alpha = p$. The risk $R(b, \hat{\lambda}^*|\alpha=p)$ when such a prior is used is

$$E_a E_{\Lambda, \theta} \frac{p}{\sum_{i=1}^{P} ((1-b(Z))X_i - \Lambda \theta_i)^2/(\Lambda \theta_i)}$$

$$= E_a E_Z (1/\Lambda) [Z(Z+p-1)\{\hat{b}^2(Z) - b(Z) + 2b(Z)/Z+p-1\}] + p$$

(conditioned on $Z$, the $X_i$'s have a multinomial distribution with parameters $\theta_i$, $i = 1, \ldots, p$)

$$= E_b \frac{1}{1-b} \left[ \frac{Z(Z+p-1)b^2(Z)}{Z+p} - \frac{2b(Z)(Z+p-1)}{Z+p} + 2b(Z)Z(1-b) \right] + p$$

where $E_b$ indicates expectation with respect to the marginal distribution of $Z \sim \text{NB}(b, p+1)$. With $\hat{b}(Z) = (p-1)/(Z+p-1)$, we have

$$E_b \frac{1}{1-b} \frac{Z(Z+p-1)b^2(Z)}{Z+p}$$

$$= \frac{1}{1-b} \sum_{z=0}^{\infty} \frac{z(z+p-1)}{z+p} \frac{(p-1)^2}{(z+p-1)^2} \frac{(z+p)!}{z!p!} b^{p+1} (1-b)^z$$

$$= \frac{(p-1)^2}{1-b} \sum_{z=1}^{\infty} \frac{(z+p-2)!}{(z-1)!p!} b^{p+1} (1-b)^z$$

$$= \frac{(p-1)^2}{1-b} \sum_{z=0}^{\infty} \frac{(z+p-1)!}{z!p!} b^{p+1} (1-b)^{z+1}$$

$$= (p-1)^2 b/p.$$
(ii) \(-2E_b \frac{1}{1-b} b(z) \frac{Z(Z+p-1)}{Z+p}\)

\[
= -2 \sum_{z=0}^{\infty} \frac{p-1}{z+p-1} \frac{(z+p)!}{z!p!} b^{p+1}(1-b)^z \frac{z+p-1}{z+p}
\]

\[
= -2(p-1).
\]

(iii) \(2E_b b(Z)Z\)

\[
= 2 \sum_{z=0}^{\infty} \frac{p-1}{z+p-1} \frac{(z+p)!}{z!p!} b^{p+1}(1-b)^z
\]

\[
= 2(p-1) \sum_{z=1}^{\infty} \frac{(z+p-2)!}{(z-1)!p!} b^{p+1}(1-b)^z
\]

\[
= 2(p-1) \sum_{z=0}^{\infty} \frac{(z+p-1)!}{z!p!} b^{p+1}(1-b)^{z+1}
\]

\[
= \frac{2(p-1)}{p} b(1-b) [p(1-b)/b + p + 1]
\]

\[
= 2(p-1) - 2b(p-1)^2/p - 2b^2(p-1)/p.
\]

Consequently, \(R(b, \lambda^*|a=p)\) becomes

\[
p - b(p-1)^2/p - 2b^2(p-1)/p, \quad 0 < b \leq 1. \quad (6.7.1)
\]

Equivalently, \(R(a, \lambda^*) = p - (p-1)^2/(p(1+a)) - 2(p-1)/(p(1+a)^2), \quad a \geq 0. \quad (6.7.2)\)

\(R(b, \lambda^*|a=p)\) is clearly concave, and decreases in \(b\) from \(R(0, \lambda^*|a=p) = p\) to \(R(1, \lambda^*|a=p) = 1/p\).
6.8 Risk Function of \( \hat{\lambda}^* \) as a Function of \( \Lambda \)

We have now obtained a relatively simple expression for \( R(b,\hat{\lambda}^*) \), and note that \( R(\lambda,\hat{\lambda}^*) \) depends on \( \lambda \) only through the value of \( \Lambda = \sum_{i=1}^{p} \lambda_i \). The risk function \( R(\lambda,\hat{\lambda}^*) \) is therefore actually a function of \( \Lambda \), \( R(\Lambda,\hat{\lambda}^*) \), for which we are now in a position to derive an expression. Usually, knowledge about \( R(\lambda,\hat{\lambda}) \) implies knowledge of \( R(b,\hat{\lambda}) \) when the prior distribution of \( \lambda \) is known. The result below shows that we can reverse the direction and gain information about \( R(\lambda,\hat{\lambda}^*) \) from knowledge about \( R(b,\hat{\lambda}^*) \).

With (6.4.1) as the prior distribution \( (\alpha=p) \), \( \Lambda \) has density function

\[
P_\alpha(\lambda) = (\lambda/a)^p \left[ \exp(-\lambda/a) \right] / [a \Gamma(p+1)] \quad \text{if } \lambda \geq 0
\]

\[= 0 \quad \text{if } \lambda < 0.
\]

By definition, \( R(a,\hat{\lambda}^*) = \int_0^\infty R(\Lambda,\hat{\lambda}^*) P_\alpha(\Lambda) \, d\Lambda \), which can be written as

\[
R(a,\hat{\lambda}^*) = \int_0^\infty R(\Lambda,\hat{\lambda}^*) \Lambda^p \exp(-\Lambda) / \Gamma(p+1) \, d\Lambda. \quad (6.8.1)
\]

Differentiating both sides of (6.8.1) \( j \) times with respect to \( a \) gives

\[
\frac{d^j R(a,\hat{\lambda}^*)}{da^j} \left|_{a=0} \right. = \frac{\Gamma(p+j+1)}{\Gamma(p+1)} \frac{d^j R(\Lambda,\hat{\lambda}^*)}{d\Lambda^j} \left|_{\Lambda=0} \right. . \quad (6.8.2)
\]

However, from (6.7.2), the derivative on the left side of (6.8.2) is

\[-(p-1)^2(-1)^j j! / p - 2(p-1)(-1)^j(j+1)! / p, \ j = 1, \ldots, \infty .
\]

Therefore

\[
\frac{d^j R(\Lambda,\hat{\lambda}^*)}{d\Lambda^j} \left|_{\Lambda=0} \right. = \frac{\Gamma(p+1)}{\Gamma(p+j+1)} \left[ -(p-1)^2 \right]^{j} (-1)^j j! - \frac{2(p-1)}{p} (-1)^j (j+1)! .
\]
Representing $R(\Lambda, \hat{\lambda}^*)$ as a power series expansion about $\Lambda = 0$ gives

$$R(\Lambda, \hat{\lambda}^*) = p - \frac{2(p-1)}{p} \sum_{j=0}^{\infty} \frac{(j+1)\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^j = \frac{(p-1)^2}{p} \sum_{j=0}^{\infty} \frac{\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^j.$$  

(6.8.3)

Furthermore, $\sum_{j=0}^{\infty} \frac{\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^j = \int_0^1 p \exp(-\Lambda t) (1-t)^{p-1} dt$

(see Abramowitz and Stegun [1964], p. 505).

The series $\sum_{j=0}^{\infty} \frac{\Gamma(p)}{\Gamma(p+j)} (-\Lambda)^j$ is uniformly convergent on bounded intervals of the real line. Hence

$$\sum_{j=0}^{\infty} \frac{(j+1)\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^j$$

$$= \sum_{j=0}^{\infty} \frac{\Gamma(p+1)}{\Gamma(p+j+1)} \left[ -\frac{d}{d\Lambda} (-\Lambda)^{j+1} \right]$$

$$= \frac{d}{d\Lambda} \sum_{j=0}^{\infty} \frac{\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^{j+1}$$

$$= \frac{d}{d\Lambda} \Lambda p \int_0^1 \exp(-\Lambda t) (1-t)^{p-1} dt$$

$$= p \int_0^1 \exp(-\Lambda t) (1-t)^{p-1} dt - p \int_0^1 t \exp(-\Lambda t) (1-t)^{p-1} dt.$$  

(6.8.4)

The last term can be rewritten as $p \int_0^1 \exp(-\Lambda t) t(1-t)^{p-1} dt$

$$= p \left[ \exp(-\Lambda t) t(1-t) \right]_0^1 + \int_0^1 \exp(-\Lambda t) \{ (1-t)^{p-1} - (p-1)t(1-t)^{p-2} \} dt$$

(integration by parts)
\[
= -p \left[ \int_0^1 \exp(-At)(1-t)^{P-1} dt - (p-1) \int_0^1 t(1-t)^{P-2} \exp(-At) dt \right].
\]

Equation (6.8.4) thus becomes \( p(p-1) \int_0^1 t(1-t)^{P-2} \exp(-At) dt. \)

Consequently, (6.8.3) has the form

\[
R(\Lambda, \lambda^*) = p - \frac{2(p-1)}{p} \left[ p(p-1) \int_0^1 t(1-t)^{P-2} \exp(-At) dt \right] - \frac{(p-1)^2}{p} \left[ \int_0^1 (1-t)^{P-1} \exp(-At) dt \right]
\]

\[
= p - 2(p-1) \int_0^1 t(1-t)^{P-2} \exp(-At) dt - (p-1)^2 \int_0^1 \exp(-At)(1-t)^{P-1} dt.
\]

Thus \( R(\Lambda, \lambda^*) \) is seen to be an increasing concave function of \( \Lambda \). We summarize the above into the following theorem.

**Theorem 6.8.5.**

Let \( X_1, \ldots, X_p \) be independent Poisson variables with parameters \( \lambda_1, \ldots, \lambda_p, p \geq 2 \). Then the risk \( R(\Lambda, \lambda^*) \) of the estimator

\[
\lambda^* = (1 - (p-1)/(Z+p-1)) X
\]

is an increasing concave function of \( \Lambda = \sum_{i=1}^p \lambda_i \) from \( R(0, \lambda^*) = 1/p \) to \( R(\infty, \lambda^*) = p \). It has power series

\[
R(\Lambda, \lambda^*) = p - \frac{2(p-1)}{p} \sum_{j=0}^\infty \frac{(j+1)\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^j - \frac{(p-1)^2}{p} \sum_{j=0}^\infty \frac{\Gamma(p+1)}{\Gamma(p+j+1)} (-\Lambda)^j.
\]

It also has the integral representation

\[
R(\Lambda, \lambda^*) = p - (p-1) \int_0^1 \exp(-At)(1-t)^{P-1} dt - 2(p-1) \int_0^1 \exp(-At) t(1-t)^{P-2} dt.
\]

This result is very similar to Corollary 2 of Efron and Morris [1973], which deals with the normal case.
6.9 Risk Function of $\hat{\lambda}^*$ as a Function of $\lambda$

The next theorem gives the risk of the estimator $\hat{\lambda}^*$ as a function of $\lambda = (\lambda_1, \ldots, \lambda_p)$.

Theorem 6.9.1.

The risk of the estimator $\hat{\lambda}^* = \left(1 - \frac{(p-1)}{(Z+p)}\right)X$ as a function of $\lambda = (\lambda_1, \ldots, \lambda_p)$ is

$$R(\lambda, \hat{\lambda}^*) = p - (p-1)E^2_{\lambda} \left[\frac{1}{Z+p}\right] - 2(p-1)E^2_{\alpha} \left[\frac{1}{(Z+p)(Z+p-1)}\right]$$

$$= p - (p-1)E_{\lambda} \left[\frac{Z+p+1}{(Z+p)(Z+p-1)}\right]. \quad (6.9.2)$$

Proof:

The left hand side of (6.9.2), as pointed out earlier, is a function of $\Lambda = \sum_{i=1}^{p} \lambda_i$, say $f(\Lambda)$. The right hand side of (6.9.2) is also a function of $\Lambda$, say $g(\Lambda)$. By definition,

$$E_b f(\Lambda) = p - (p-1)b/p - 2(p-1)b^2/p$$

where $E_b$ indicates expectation with respect to the joint prior distribution (6.4.1) with $a = p$. Also,

$$E_b g(\Lambda) = p - (p-1)^2 E_{\nu} E_{\lambda} \left[\frac{1}{Z+p}\right] - 2(p-1)^2 E_{\nu} E_{\lambda} \left[\frac{1}{(Z+p)(Z+p-1)}\right]$$

$$= p - (p-1)^2 E_{\nu} \left[\frac{1}{Z+p}\right] - 2(p-1)^2 E_{\nu} \left[\frac{1}{(Z+p)(Z+p-1)}\right]$$

where $E_{\nu}$ indicates expectation with respect to the marginal distribution of $Z \sim NB(b,p+1)$. As a result,

$$E_b g(\Lambda) = p - (p-1)^2 b/p - 2(p-1)b^2/p.$$
Therefore $E_b g(\Lambda) = E_b f(\Lambda)$ for every value of $b$, $0 < b < 1$, or equivalently, for every value of $a$, $a > 0$. Since the distributions of $\Lambda = \Lambda^p \exp(-\Lambda/a) \, d\Lambda$ are complete as a function of $a$, $f$ and $g$ must be the same function.

Q.E.D.

Both Theorem 6.8.5 and Theorem 6.9.1 show that $R(\lambda, \hat{\lambda}^*) < p$ for all $\lambda$, so $\hat{\lambda}^*$ is better than the MLE (whose risk is $p$) for all $\lambda$. Furthermore, $[1/(z+p)]$ and $[1/((z+p)(z+p-1))]$ are convex functions in $z$. Hence by Jensen's inequality, the upper bound $UB(\Lambda)$ of $R(\lambda, \hat{\lambda}^*) = R(\Lambda, \hat{\lambda}^*)$ is of the form

$$UB(\Lambda) = p - \frac{(p-1)^2}{\Lambda + p} - \frac{2(p-1)^2}{\Lambda^2 + 2p\Lambda + p^2 - p}$$

$$= p - \frac{(p-1)^2}{\Lambda + p} - \frac{2(p-1)^2}{(\Lambda + p)^2 - p}.$$ 

In particular, when $\Lambda = 0$, then $UB(0) = 1/p$. Since $R(0, \hat{\lambda}^*) = 1/p$, the upper bound is attained when $\Lambda = 0$.

6.10 Risk Function of the Clevenson-Zidek Estimators

By Theorem 2.2.9, the estimators

$$\hat{\lambda}^s = (1 - [s(p-1)/(Z+p-1)])X, \ 0 \leq s \leq 2,$$

are uniformly better than the MLE. The risk functions of $\hat{\lambda}^s$ are similar to those given in Theorem 6.8.5 and Theorem 6.9.1. They are given in the following theorems (proofs are similar to those in Theorems 6.8.5 and 6.9.1).
Theorem 6.10.1.

The risk function of $\hat{\lambda}^S$ ($0 \leq s \leq 2$) as a function of $b$ is
\[
R(b, \hat{\lambda}^S) = p - (p-1)^2 b[1 - (s-1)^2]/p - 2b^2 s(p-1)/p.
\]

Theorem 6.10.2.

The risk function of $\hat{\lambda}^S$ ($0 \leq s \leq 2$) as a function of $\Lambda = \sum_{i=1}^{P} \lambda_i$ is
\[
R(\Lambda, \hat{\lambda}^S) = \int_0^1 \exp(-\Lambda t) (1-t)^{P-1} dt - 2(p-1)^2 \int_0^1 \exp(-\Lambda t) t(1-t)^{P-2} dt.
\]

This is an increasing concave function of $\Lambda$ from
\[
R(0, \hat{\lambda}^S) = 1/p + (p-1)^2 (s-1)^2/p + 2(p-1)(1-s)/p
\]
to $R(\infty, \hat{\lambda}^S) = p$.

Theorem 6.10.3.

The risk function of $\hat{\lambda}^S$ ($0 \leq s \leq 2$) as a function of $\lambda = (\lambda_1, \ldots, \lambda_p)$ is
\[
R(\lambda, \hat{\lambda}^S) = p - (p-1)^2 \sum_{i=1}^{P} \lambda_i [1/(Z+p)] - 2(p-1)^2 \sum_{i=1}^{P} \lambda_i [1/(Z+p)(Z+p-1)].
\]

6.11 Summary

In this section, we confine our attention to estimators of $\lambda$ of the form $\hat{\lambda} = (1 - b(Z))X$ and carry out our analysis by means of empirical Bayes methods. We begin with a description of the method used in the normal case and then proceed to calculate the "relative savings loss" (a term coined by Efron and Morris [1973]) in the Poisson case when normalized squared error loss is the criterion. The relative savings loss plays a fundamental role in the analysis performed by Efron and Morris [1973] in the normal case. We obtain "plus rules ($\hat{\lambda}^+ = (1 - b(Z))^+X$) and "truncated Bayes rules", results similar to those in the normal case.

The remainder of the section is devoted to the calculation of the risk $R(\lambda, \hat{\lambda}^S)$ of the Clevenson-Zidek estimators.
\lambda^s = [1 - s(p-1)/(Z+p-1)]X

as a function of \lambda, where 0 \leq s \leq 2. We find that R(\lambda, \lambda^s) is actually a function of \lambda only through the value of \Lambda = \sum_{i=1}^{p} \lambda_i.

Moreover, R(\lambda, \lambda^s) is an increasing and concave function of \Lambda.
SECTION 7. COMPUTER SIMULATION

7.1 Introduction

In this section we describe the results of a computer simulation used to compare some of our proposed estimators with the MLE. We also compare a Clevenson-Zidek estimator and one of Peng's estimators (Theorem 2.2.3) with the MLE. Finally, we discuss the performance of the estimators.

The computations reported here were performed both on the IBM 370/168 computer at the University of British Columbia and the Data General NOVA 840 computer at the University of California, Riverside. A FORTRAN program was used in the IBM computer and a BASIC program was used in the NOVA computer. First, the number \( p \) of independent Poisson random variables is chosen. Second, \( p \) parameters \( \lambda_i \) are generated randomly within a certain range \((c,d)\). Third, one observation of each of the \( p \) distributions with the parameters obtained in the second step is generated. Estimates of the parameters are then calculated according to the estimator \( \hat{\lambda} \) we want to test. The third step is repeated 2000 times and the risks under the relevant loss functions for both the estimator and the MLE are calculated. The percentage of the savings in using \( \hat{\lambda} \) as compared to the MLE, \( \frac{R(\lambda,X) - R(\lambda,\hat{\lambda})}{R(\lambda,X)} \times 100 \)\% is calculated. The whole process is then repeated at least three times and the average percentage of the savings is calculated.

In the case when our Bayesian estimators (Section 5) are used, the hyperparameters \((u,v,\mu,a,b)\) which specify the prior information are chosen beforehand, i.e. before the third step takes place. We chose the range of the parameters \( \lambda_i \) in such a way that we might check the performance of
the estimators when the parameters $\lambda_i$ are relatively close to one another. This is especially relevant when the performance of our Bayes estimators are checked, since this is the case where we expect that a substantial improvement over the usual procedure will take place. For purposes of comparison, we also include a simulation study of the performance of the estimators when the range of the parameters $\lambda_i$ is wide. Our expectation that the Bayes estimator might not always be better than the MLE when the range of the $\lambda_i$'s is wide was substantiated when negative savings resulted from a choice of $p = 3$, $\mu = 9.0$, and $(c,d) = (0,20)$. The MLE performed well in this case because $p$ is small, the range of the $\lambda_i$'s is wide, and the prior was purposely selected to be an improper choice for the problem.

In calculating the percentage of improvement of the various estimators over the MLE, the appropriate loss functions must be used. Recall that $L_k(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i^k$. In Tables I and II, the loss function $L_2$ is the criterion, while in Tables III and IV, $L_1$ is used. The loss function $L_4$ is the criterion in Table V, and Tables VI through XIV use squared error loss.

In most of the cases, the improvement percentage is seen to be an increasing function of $p$, the number of independent Poisson distributions. For the non-Bayes estimators, we see that in general, the improvement percentage decreases as the magnitude of the $\lambda_i$'s increases. The improvement percentage of the Bayes estimators depends on the choice of the prior hyperparameters; proper choice leads to substantial improvement over the MLE.
7.2 Estimators Under k-NSEL

In Section 4 we derived, for each non-negative integer k, a class of estimators \( \hat{\lambda}^k \) dominating the MLE under k-NSEL \( L_k(\lambda, \hat{\lambda}) = \frac{p}{\sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2/\lambda_i^k} \).

These estimators are of the form

\[
\hat{\lambda}_i^k = X_i - \frac{\phi(z) X_i (X_i - 1) \cdots (X_i - k + 1)}{\sum_{j \neq i} (X_j + 1) \cdots (X_j + k) + X_i (X_i - 1) \cdots (X_i - k + 1)}
\]

where \( \phi(z) \in [0, 2k(p-1)] \) and is nondecreasing. The estimator \( \hat{\lambda}_2 \) (i.e. k=2) is of considerable appeal because this is the case where a natural loss function \( L_2(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} (1 - \frac{\hat{\lambda}_i}{\lambda_i})^2 \) is used in estimating scale parameters. Our computer simulation results in this subsection are mainly devoted to the performance of an estimator \( \hat{\lambda}_2 \) and a Clevenson-Zidek estimator \( \hat{\lambda}_1 \). More specifically, we choose \( \phi(z) = 2[p-1] \) (i.e. k[p-1]) for \( \hat{\lambda}_2 \), and \( \hat{\lambda}_1 = (1 - \frac{p}{Z+p})Z \).

In Table I, we see that for the ranges considered, the percentage of improvement in risk of \( \hat{\lambda}_2 \) over the MLE is considerable when the parameters fall into a narrow interval. For each value of p, \( \hat{\lambda}_2 \) performs best when the parameters are in the intervals (0, 4) and (4, 8). The improvement decreases gradually as the magnitude of the \( \lambda_i \)'s increases. In contrast, the Clevenson-Zidek estimator \( \hat{\lambda}_1 \) performs very well only when the parameters are relatively small, with the improvement percentage decreasing dramatically as the magnitude of the \( \lambda_i \)'s increases (Table III). This is as conjectured in Section 4. Tables II and IV show results for wider ranges of the \( \lambda_i \)'s. Although the improvement percentages of \( \hat{\lambda}_2 \) over the MLE are by no means substantial, they are nevertheless greater
Table I. Improvement Percentage of $\hat{\lambda}^2$ over the MLE

Narrow Range for $\lambda_i$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=3</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>24.97</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>23.08</td>
</tr>
<tr>
<td>(8, 12)</td>
<td>14.89</td>
</tr>
<tr>
<td>(12, 16)</td>
<td>11.04</td>
</tr>
</tbody>
</table>

Table II. Improvement Percentage of $\hat{\lambda}^2$ over the MLE

Wide Range for $\lambda_i$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>11.62</td>
</tr>
<tr>
<td>(10, 30)</td>
<td>11.99</td>
</tr>
</tbody>
</table>
Table III. Improvement Percentage of $\hat{\lambda}_i^1$ over the MLE

Narrow Range for $\lambda_i$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=3</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>24.77</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>7.34</td>
</tr>
<tr>
<td>(8, 12)</td>
<td>4.02</td>
</tr>
<tr>
<td>(12, 16)</td>
<td>2.44</td>
</tr>
</tbody>
</table>

Table IV. Improvement Percentage of $\hat{\lambda}_i^1$ over the MLE

Wide Range for $\lambda_i$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>8.23</td>
</tr>
<tr>
<td>(10, 30)</td>
<td>4.08</td>
</tr>
</tbody>
</table>
than those of \( \hat{\lambda}^1 \). Of course, the different loss functions employed for \( \hat{\lambda}^2 \) and \( \hat{\lambda}^1 \) might contribute to such a difference.

Table V reveals the performance of the estimator \( \hat{\lambda}^4 \) with 
\[ \phi(z) \equiv 4[p-1] \] (i.e. \( k[p-1] \)). When the parameters are confined to the interval \((4, 8)\), the percentages of improvement are seen to be above 30%, which is rather substantial. However, when the parameters are in the interval \((0, 4)\), the improvements are unimpressive. This is as expected because \( \hat{\lambda}^4 \) gives identical estimates as the MLE when the observations are less than four, and when the \( \lambda_i \)'s are in the interval \((0, 4)\), such observations are likely to occur (cf. the remarks following the proof of Theorem 4.3.1). The slight improvements shown in the \((0, 4)\) row of Table V are due to the fact that some observations are greater than four. While the MLE estimates \( \lambda_i \) by \( x_i \), the estimator \( \hat{\lambda}^4 \) shrinks the observations greater than four towards zero, resulting in some improvement.

7.3 Bayes Estimators

The Bayes estimators \( \hat{\lambda} \) developed in Section 5 are of the form

\[
\hat{\lambda}_i = E_{x, a, b, \mu} \lambda_i \]

\[
= \frac{Z(X_i+1)}{(Z+p)} \cdot \frac{g(X+e_i, a, b, \mu)}{g(X, a, b, \mu)} , \quad i = 1, \ldots, p \tag{7.3.1}
\]

where \( g(x, a, b, \mu) \) is the generalized hypergeometric function given in equation (5.2.4). Basically, (7.3.1) depends on the choice of the prior distribution of \( \alpha \), which is given in equation (5.2.3). A particular prior distribution of \( \alpha \) is determined by the specification of the
Table V. Improvement Percentage of $\hat{\lambda}$ over the MLE

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_1$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=4</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>7.51</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>30.03</td>
</tr>
</tbody>
</table>

Table VI. Poisson-Distributed $\alpha$ ($u=0, v=0$)

Narrow Range for $\lambda_1$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_1$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
</tr>
<tr>
<td>(0, 2)</td>
<td>0.5</td>
</tr>
<tr>
<td>(2, 4)</td>
<td>2.0</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>5.0</td>
</tr>
<tr>
<td>(8, 12)</td>
<td>9.0</td>
</tr>
<tr>
<td>(12, 16)</td>
<td>13.0</td>
</tr>
<tr>
<td>(16, 20)</td>
<td>17.0</td>
</tr>
</tbody>
</table>
hyperparameters \((u,v,a,b,\mu)\). In our simulation study, we shall restrict ourselves to the following cases:

1. \(u=0, v=0, 0 \leq \mu < \infty\) (Poisson-distributed \(a\))
2. \(u=1, v=0, a_1=1.0, 0 \leq \mu < 1\) (geometric-distributed \(a\))
3. \(u=1, v=0, a_1=3.0, 0 \leq \mu < 1\) (negative binomial-distributed \(a\)).

The choice of \(\mu\) reflects one's belief about the magnitude of the parameters \(\lambda_i\). A relatively large value of \(\mu\) indicates that the \(\lambda_i\)'s are believed to be large, while a relatively small value means that the \(\lambda_i\)'s are believed to be small.

Recall that \(\lambda_i \sim \text{gamma}(\alpha, \beta)\). Observe that for given \(\mu\) and \(\beta\), the marginal expectation of \(X_{i\beta}\) is

\[
E(X_{i\beta}) = E(E(X_{i\beta}|\lambda)) = E(E(\lambda_{i|\beta})|\alpha, \beta) = E(\frac{\alpha}{\beta} | \beta) = \frac{1}{\beta} E\alpha
\]

where the expectations are taken appropriately. Note that \(E\alpha\) is a function of \(\mu\).

Suppose that the scale parameter \(\beta\) is equal to one, so that our problem is cast in terms of units of \(\beta\). In order that \(E(X_{i\beta}|\beta=1)\) might be close to the parameter \(\lambda_i\), \(\mu\) is chosen so that \(E\mu = E(X_{i\beta}|\beta=1) = \frac{c+d}{2}\) if the \(\lambda_i\)'s are believed to be in the interval \((c, d)\). For example, if \((c, d) = (4, 8)\), \(\mu=5\) is chosen. However, this is only one way to choose \(\mu\); other choices are certainly possible. In Table VII, three values of \(\mu\) were chosen. The choice of \(\mu=12\) seemed to result in the greatest savings.

This leads us to suspect that there are in fact better choices for \(\mu\) than that described above. In retrospect, it seems that we should consider not only the length of the interval as a measure of closeness, but also, and possibly more importantly, the relative magnitude of the
$\lambda_i$'s within an interval. We might use the ratio $c/d$, where $c > 0$, as an indicator of the closeness of the $\lambda_i$'s. When the ratio is close to zero, the $\lambda_i$'s may be said to be relatively spread out in that interval, and when the ratio is close to one, they may be said to be relatively close. Hence for intervals of fixed length, the parameters $\lambda_i$ are considered to be closer if the lower bound of the interval is further away from zero. In other words, the closeness of the $\lambda_i$'s does not depend solely on the Euclidean distance between the $\lambda_i$'s, but also on their location.

Recall that the Bayes estimate is a weighted average of $x_i$ and $\bar{x}$. If the $\lambda_i$'s are very close to one another, more weight should be given to $\bar{x}$, which means that $\mu$ should be given a relatively large value. If the $\lambda_i$'s are spread out, a small value of $\mu$ should be chosen. In Table VIII, we see that a choice of $\mu=0.1$ is better than 1.0 and 9.0 when the interval is (0, 20). Since the parameters are quite spread out, the Bayes estimate should give more weight to the MLE, i.e. a small value of $\mu$ is suitable. However, in the interval (10, 30), the $\lambda_i$'s are relatively close (although the length of the interval is still 20) and we see that a choice of $\mu=19$ is better than a choice of $\mu=14$, a phenomenon similar to that encountered in Table VII. Thus, the Bayes estimators remain superior to the MLE even though the range of the parameters $\lambda_i$ is wide, if we are acute enough to select the hyperparameters appropriately.

From Table VI, we conclude that the percentage of improvement over the MLE is substantial when the parameters fall into a narrow interval, which is as expected. Tables IX and X both show that the Bayes estimators using other prior distributions (negative binomial- and geometric-distributed $\alpha$) are still superior to the MLE and the savings in mean squared error is substantial.
Table VII. Poisson-Distributed $\alpha$ ($u=0$, $v=0$)

**Effect of $\mu$**

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>1.0</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>5.0</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>12.0</td>
</tr>
</tbody>
</table>

Table VIII. Poisson-Distributed $\alpha$ ($u=0$, $v=0$)

**Wide Range for $\lambda_i$'s**

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
</tr>
<tr>
<td>(0, 10)</td>
<td>2.0</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>0.1</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>1.0</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>9.0</td>
</tr>
<tr>
<td>(10, 20)</td>
<td>14.0</td>
</tr>
<tr>
<td>(10, 30)</td>
<td>14.0</td>
</tr>
<tr>
<td>(10, 30)</td>
<td>19.0</td>
</tr>
</tbody>
</table>
Table IX. Negative Binomial-Distributed $\alpha$ ($u=1$, $v=0$, $a_1=3.0$)

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_1$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>0.25</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>0.5</td>
</tr>
<tr>
<td>(8, 12)</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table X. Geometric-Distributed $\alpha$ ($u=1$, $v=0$, $a_1=1.0$)

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_1$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>0.5</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>0.833</td>
</tr>
</tbody>
</table>
7.4 Estimators Under Squared Error Loss

We shall next compare the performance of Peng's estimator $\hat{\lambda}^{(0)}$ (a special case of $\hat{\lambda}^{(k)}$ described in Section 3) and our new estimator $\hat{\lambda}^{[m]}$, which shrinks the MLE towards the minimum of the observations, where

$$\hat{\lambda}^{[m]}_i = x_i - \frac{(p-N_m - 2) + H_i(X)}{\sum_{i=1}^{p} H_i^2(X)}, \quad i = 1, \ldots, p,$$

and the $H_i$'s are as defined in equation (3.4.1). We have argued that this adaptive estimator $\hat{\lambda}^{[m]}$ should perform better than $\hat{\lambda}^{(0)}$ across a broad spectrum of values for the $\lambda_i$'s.

Table XI shows that although $\hat{\lambda}^{(0)}$ provides a noticeable improvement over the MLE, the improvement percentage decreases rapidly as the $\lambda_i$'s move away from zero. In contrast, the improvement percentages in Table XIII remain noticeable even when the $\lambda_i$'s are in the interval (12, 16). This supports our conjecture that the estimator $\hat{\lambda}^{[m]}$ is superior to $\hat{\lambda}^{(0)}$ when $p \geq 4$. The improvement percentages for $p=3$ in Table XIII are all zero because $\lambda^{[m]}$ is identical with the MLE when $p \leq 3$. Thus Peng's estimator has the merit that it provides a better estimator than the MLE under squared error loss when $p \geq 3$. Note, however, that use of $\hat{\lambda}^{(0)}$ implicitly involves a choice of $k=0$, towards which the observations are shrunk. If this kind of subjectivity is to be avoided and the shrinkage determined only by the data (using $\hat{\lambda}^{[m]}$), one degree of freedom is lost and improvement over the MLE results only when $p \geq 4$.

Finally, Tables XII and XIV show our results for wide ranges of the parameters $\lambda_i$. In both cases, the improvement percentages are seen to be minimal, which is not surprising because the risk of the MLE is large in this case, and hence the relative savings are bound to be small.
Table XI. Improvement Percentage of $\hat{\lambda}^{(0)}$ over the MLE

Narrow Range for $\lambda_1$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_1$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=3</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>2.23</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>0.58</td>
</tr>
<tr>
<td>(8, 12)</td>
<td>0.17</td>
</tr>
<tr>
<td>(12, 16)</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table XII. Improvement Percentage of $\hat{\lambda}^{(0)}$ over the MLE

Wide Range for $\lambda_1$'s

<table>
<thead>
<tr>
<th>Range of the Parameters $\lambda_1$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>0.69</td>
</tr>
<tr>
<td>(10, 30)</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Table XIII. Improvement Percentage of $\hat{\lambda}_{[m]}$ over the MLE
Narrow Range for $\lambda_i$'s

<table>
<thead>
<tr>
<th>Range of Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=3</td>
</tr>
<tr>
<td>(0, 4)</td>
<td>0.00</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>0.00</td>
</tr>
<tr>
<td>(8, 12)</td>
<td>0.00</td>
</tr>
<tr>
<td>(12, 16)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table XIV. Improvement Percentage of $\hat{\lambda}_{[m]}$ over the MLE
Wide Range for $\lambda_i$'s

<table>
<thead>
<tr>
<th>Range of Parameters $\lambda_i$</th>
<th>Percentage of Improvement over the MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=5</td>
</tr>
<tr>
<td>(0, 20)</td>
<td>0.79</td>
</tr>
<tr>
<td>(10, .30)</td>
<td>2.30</td>
</tr>
</tbody>
</table>
7.5 Comparison of the Estimators

Based on our computer simulation results, it appears that among the estimators considered, the Bayes estimators provide by far the most improvement over the MLE when the parameters are in a relatively narrow interval. When prior knowledge indicates that the $\lambda_i$'s are exchangeable and are likely to fall into a certain relatively narrow interval $(c, d)$, there is no doubt that the Bayes estimator should be used. Recall also that the Bayes estimators remain superior to the MLE even though the range of the parameters $\lambda_i$ is wide, if we choose the hyperparameters suitably. It should be noted, however, that the Bayes estimators do not dominate the MLE uniformly in $\lambda$, and hence should not be used indiscriminately.

On the other hand, the estimators $\hat{\lambda}^k$, $\hat{\lambda}^{(k)}$, and $\hat{\lambda}^{[m]}$ are guaranteed to have lower risk than the MLE uniformly in $\lambda$ under appropriate loss functions. The estimators $\hat{\lambda}^k$ and $\hat{\lambda}^{(k)}$ can be used advantageously when the integer $k$ is chosen appropriately. The estimator $\hat{\lambda}^{[m]}$ is useful if $p \geq 4$ and prior knowledge of $k$ used in the estimator $\hat{\lambda}^{(k)}$ is vague.
8.1 Motivation

Up to this point, we have confined our attention to the situation in which only one observation is taken from each of \( p \) independent Poisson populations. We now suppose that \( X_{i1}, \ldots, X_{in_i} \sim \text{Poisson}(\lambda_i) \), where \( n_i > 1, \ i = 1, \ldots, p \), and that all the \( X_{ij} \)'s are independent. Letting \( X_i = \sum_{j=1}^{n_i} X_{ij}, \ i = 1, \ldots, p \), the MLE of \( \lambda \) is \( \left( \frac{X_1}{n_1}, \ldots, \frac{X_p}{n_p} \right) \). We pose the following question:

**Question 8.1.1.**

Given the situation described above, are there estimators \( \hat{\lambda} \) of \( \lambda \) which dominate the MLE under an appropriate loss function?

In the case of simultaneously estimating \( p \) normal means \( \theta_i \), we suppose that \( X_{ij} \sim \mathcal{N}(\theta_i, 1) \) for \( i = 1, \ldots, p \) and \( j = 1, \ldots, n_i \). The variables \( X_i/n_i \) are still normal and we are essentially dealing with the problem in which one observation is taken from each population. In our case, unfortunately, the distribution of \( X_i/n_i \) is no longer Poisson if \( n_i > 1 \). We therefore cannot apply our previous results directly to this problem. However, if our interest is to estimate \( n_i \theta_i, \ i = 1, \ldots, p \), then the foregoing theory can be applied because \( X_i \sim \text{Poisson}(n_i \lambda_i) \) in this case.

We first consider the squared error loss function. The goal is to find an estimator \( \hat{\lambda} \) of \( \lambda \) such that

\[
E_\lambda \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2 \leq E_\lambda \sum_{i=1}^{p} (\lambda_i - \frac{X_i}{n_i})^2 \quad \text{for all } \lambda
\]
with strict inequality for some $\lambda$. Observe that

$$E_\lambda \sum_{i=1}^{p} \frac{X_i}{n_i} (\frac{1}{n_i} - \lambda_1)^2 = E_\lambda \sum_{i=1}^{p} \frac{1}{n_i^2} (X_i - n_i \lambda_1)^2$$

and that $X_i \sim \text{Poisson}(n_i \lambda_1)$, $i = 1, \ldots, p$. We are thus led to consideration of the following problem.

**Problem 8.1.2.**

Suppose $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, p$, and that $x_i$ is an observation of $X_i$, $i = 1, \ldots, p$. If possible, find an estimator $\hat{\lambda}$ of $\lambda_i$ such that $\hat{\lambda}$ dominates the MLE $X = (X_1, \ldots, X_p)$ uniformly under the generalized squared error loss function $L^c = \sum_{i=1}^{p} c_i (\lambda_i - \hat{\lambda}_i)^2$, where $c_i > 0$, $i = 1, \ldots, p$.

Solution of Problem 8.1.2 will automatically provide an answer to Question 8.1.1. In the next subsection we shall show that a solution to Problem 8.1.2 exists provided certain conditions on the $c_i$'s hold.

### 8.2 Estimators Under Generalized Squared Error Loss

Let $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, p$ and $X = (X_1, \ldots, X_p)$. The estimators $\hat{\lambda}$ of $\lambda$ we consider here are again of the form $X + f(X)$ where $f(X)$ is as described in Section 2.1 and satisfies the following conditions:

1. $f_i(x) = 0$ if $x$ has a negative coordinate
2. $E_\lambda |f_i(x + je_1)| < \infty$ for $j = 0, 1$.

The results in this subsection are very similar to those in Section 6, as are the derivations. The next lemma, which is similar to Lemma 2.2.2, gives the deterioration in risk of $\hat{\lambda} = X + f(X)$ as compared to the risk of the MLE $X$. 
Lemma 8.2.1. ...

Under the loss function $L^c(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} c_i (\lambda_i - \hat{\lambda}_i)^2$, the deterioration in risk $D$ of $\hat{\lambda} = X + f(X)$ as compared to the risk of $X$ is

$$D = R(\lambda, \hat{\lambda}) - R(\lambda, X) = E_{\lambda} \Delta$$

where

$$\Delta = \sum_{i=1}^{p} c_i f_i^2(X) + 2 \sum_{i=1}^{p} c_i X_i [f_i(X) - f_i(X_i)]$$

The following theorem gives estimators $\hat{\lambda}$ of $\lambda$ dominating $X$ under $L^c$.

Theorem 8.2.2.

Let $X_i \sim \text{Poisson}(\lambda_i), i = 1, \ldots, p$ and $X = (X_1, \ldots, X_p)$ (p > 3). Define

$$f_i(x) = \begin{cases} \frac{1}{\sqrt{c_i}} \left( \sum_{j: x_j \neq 0} \sqrt{c_j} - 2\sqrt{c^*} \right) h(x_i)/S & \text{if } x_i > 0 \\ 0 & \text{if } x_i < 0 \end{cases}$$

for $i = 1, \ldots, p$, where

1. $c^* = \max\{c_i\}_{i=1}^{p}$
2. $h(j) = \frac{1}{n} \sum_{n=1}^{p} 1/n$ if $j > 1$
   \[
   \frac{1}{n} \quad \text{if } j < 0
   \]
3. $S = \sum_{i=1}^{p} h^2(x_i)$.

Let $f(X) = (f_1(X), \ldots, f_p(X))$. Then the estimator $\hat{\lambda} = X + f(X)$ dominates $X$ uniformly in $\lambda$ under the loss function $L^c$ provided

$$\sum_{i=1}^{p} \sqrt{c_i} > 2\sqrt{c^*}.$$
Proof:

The proof is very similar to that of Theorem 6.3.8. It can be shown that the \( A \) given in Lemma 8.2.1 is less than or equal to 

\[-(\sum_{j:x_j \neq 0} \sqrt{c_j} - 2\sqrt{c^*})^2/S \]

and hence \( R(\lambda, \hat{\lambda}) \leq R(\lambda, X) \) for all \( \lambda \).

Q.E.D.

If we interpret \( c_i \) as the cost of misestimation of \( \lambda_i \) per squared unit of length, then the condition \( \sum_{i=1}^{p} \sqrt{c_i} > 2\sqrt{c^*} \) can be interpreted as meaning that no cost per unit length of a particular \( \lambda_i \) can be greater than the total cost per unit length of the remaining \( \lambda_i \)'s.

Intuitively, we would not expect that an improvement over \( X \) can be made if there is a particular cost \( \sqrt{c_i} \) (per unit length) dominating the rest of the costs, since we know that in the one-dimensional case \( (p=1) \), \( X \) is admissible under squared error loss. Hence the condition \( \sum_{i=1}^{p} \sqrt{c_i} > 2\sqrt{c^*} \) is intuitively reasonable.

The estimator \( \hat{\lambda} \) given in Theorem 8.2.2 is not the only solution to problem 8.1.2. In fact, solutions similar to those given in Section 6 can be obtained. For example, we may suppose the function \( h \) satisfies Lemma 6.2.2 and obtain theorems similar to Theorems 6.3.8, 6.3.12, 6.3.13, etc. However, we will not carry out the details here.

Let \( c_i = 1/n_i^2 \). The following corollary suggests an answer to Question 8.1.1.

**Corollary 8.2.3.**

Let \( X_{i,j} \sim \text{Poisson} (\lambda_i) \) where \( n_i \geq 1, i = 1, \ldots, p \), and \( p \geq 3 \).

Let \( X_i = \sum_{j=1}^{n_i} X_{i,j}, i = 1, \ldots, p \). Define \( \lambda = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p) \) as
117

\[ \lambda_i = \frac{X_i}{n_i} - \frac{\sum_{j:X_j \neq 0} \frac{1}{n_j}}{n_i} + \frac{X_i}{n_i} = \frac{\sum_{j=1}^{p} \frac{X_j}{n_i} (\sum_{j=1}^{p} \frac{1}{n_j})^2}{n_i}, \quad i = 1, \ldots, p \]

where

1. \( n_* = \min \{ n_i \} \)
2. \( \sum_{i=1}^{p} \frac{1}{n_i} = 0. \)

Suppose \( \sum_{i=1}^{p} \frac{1}{n_i} > \frac{2}{n_*}. \) \( (8.2.4) \)

Then the estimator \( \hat{\lambda} \) dominates the MLE \( \left( \frac{X_1}{n_1}, \ldots, \frac{X_p}{n_p} \right) \) under the squared error loss function \( L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2. \)

Remarks:

1. Condition \( (8.2.4) \) guarantees that the proposed estimator \( \hat{\lambda} \) is different from the MLE with positive probability.
2. When \( n_i = 1 \) for all \( i, \) the estimator \( \hat{\lambda} \) is Peng's estimator \( \hat{\lambda}^{(0)} \) (cf. Theorem 2.2.3). Moreover, condition \( (8.2.4) \) holds automatically if \( p \geq 3. \)
3. Suppose the \( i_{th} \) population has sample size \( n_* \). Condition \( (8.2.4) \) says that improvement over the MLE is possible under squared error loss if the sample sizes \( n_j \) of the other populations are not too large.

8.3 Estimators Under Generalized k-NSEL

Answers to Question 8.1.1 are possible under other loss functions as well. The techniques are much the same as those used above, i.e.
Question 8.1.1 is transformed into a problem similar to 8.1.2, with a different loss function, and analysis similar to that in a previous section is employed to derive our results. We will therefore for the most part merely state the results.

The following lemma is similar to Lemma 2.2.7.

Lemma 8.3.1.

Let $X_i \sim \text{Poisson} (\lambda_i)$, $i = 1, \ldots, p$ and let $f_i : \mathbb{R}^p \to \mathbb{R}$, $i = 1, \ldots, p$ satisfy the conditions given in Lemma 2.2.6. Define

$$\lambda = X + f(X).$$

Then, under the loss function

$$L_k^C(\lambda, \lambda) = \sum_{i=1}^{p} c_i (\lambda_i - \lambda_i)^2 / \lambda_i,$$

(8.3.2)

the deterioration in risk of $\lambda$ as compared to $X$ is $R(\lambda, \lambda) - R(\lambda, X)$

$$= E_{\lambda} \Delta_k$$

where

$$\Delta_k = \sum_{i=1}^{p} c_i f_i^2(X + ke_i) / (X_i + k)^{k} + 2 \sum_{i=1}^{p} c_i (X_i + k) \frac{f_i(X + ke_i) - f_i(X + (k-1)e_i)}{(X_i + k)^k} .$$

Proof: The proof is virtually the same as that of Lemma 2.2.7.

The next theorem supplies estimators that dominate the MLE under the loss function (8.3.2).

Theorem 8.3.3.

Let $X_i \sim \text{Poisson} (\lambda_i)$, $i = 1, \ldots, p$, and the loss function be as given in (8.3.2). Define

$$f_i(x) = -k(p-1) / \sqrt{c_i} x_i^{(k)} / (s_i + x_i^{(k)})$$

if $x_i > 0$

$$= 0$$

if $x_i < 0$
$i = 1, \ldots, p$, where

(1) $c^*_i = \min\{c_i\}$

(2) $S^i = \sum_{j \neq i} (x_{j+k})^2$

(3) $x_{i}^{(k)} = x_{i}^{(k-1)}(x_{i}^{(k-2)} \cdots (x_{i}^{(k-k)})).$

Let $f(X) = (f_1(X), \ldots, f_p(X))$ and $\hat{\lambda} = X + f(X)$. Then the estimator $\hat{\lambda}$ of $\lambda$ dominates the MLE $X$ uniformly in $\lambda$ under the loss function (8.3.2).

**Proof:**

Using similar techniques as in the proof of Theorem 4.3.1, we can show that $\Delta_k$ given in Lemma 8.3.1 above satisfies

$$\Delta_k \leq -\frac{c^*_k^2(p-1)^2}{\sum_{i=1}^{p} (x_{i+k})^2} \leq 0.$$  

Q.E.D.

The following corollaries provide other answers to Question 8.1.1.

**Corollary 8.3.4.**

Let $X_{ij} \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, p$, $j = 1, \ldots, n_i$. Let

$$X_i = \sum_{j=1}^{n_i} X_{ij}, \quad i = 1, \ldots, p.$$  

Define $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ by

$$\hat{\lambda}_i = \frac{X_i}{n_i} - \frac{k(p-1)}{n_i} \left[ \frac{n_i}{x_i} \right] \frac{((k-2)/2)}{x_i^{(k)} / (s_i^2 + x_i^{(k)})}.$$
i = 1, \ldots, p, \text{ where } n^* = \max_i n_i. \text{ Then } \hat{\lambda} \text{ dominates the MLE}

\sum_{i=1}^{p} X_i \text{ under } k\text{-NSEL } L_k(\lambda, \hat{\lambda}) = \frac{\sum_{i=1}^{p} (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i^k}{\lambda_i^k} \text{ with } k = 1 \text{ or } 2.

\textbf{Proof:} Use Theorem 8.3.3 with } c_i = n_i^{k-2} \text{ and note that } c^* = (n^*)^{k-2} \text{ if } k = 1 \text{ or } 2.

\textbf{Corollary 8.3.5.}

\text{independent}

Let \( X_{ij} \sim \text{Poisson } (\lambda_i), i = 1, \ldots, p, j = 1, \ldots, n_i. \text{ Let}

\sum_{j=1}^{n_i} X_i = \sum_{j=1}^{n_i} X_{ij}, \ i = 1, \ldots, p. \text{ Define } \hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p) \text{ by}

\hat{\lambda}_i = \frac{X_i}{n_i} - \frac{k(p-1)}{n_i} \left( \frac{n^*}{n_i} \right)^{(k-2)/2} \cdot \frac{X_i^{(k)} / (\sum_i X_i^{(k)})}{n_i},

i = 1, \ldots, p, \text{ where } n^* = \min_i n_i. \text{ Then } \hat{\lambda} \text{ dominates the MLE } \left( \frac{X_1}{n_1}, \ldots, \frac{X_p}{n_p} \right)

\text{under } k\text{-NSEL with } k \geq 3.

\textbf{Proof:} Observe that } c_i = (n_i)^{k-2} \text{ and } c^* = (n^*)^{k-2} \text{ if } k \geq 3.

There are, of course, other estimators dominating the MLE under

the loss function \( L(\lambda, \hat{\lambda}) = \sum_{i=1}^{p} c_i (\lambda_i - \hat{\lambda}_i)^2 / \lambda_i^k. \text{ The results will be}

\text{similar to those derived in Section 4, and we shall therefore not set down the details here.}

8.4 \textbf{An Application}

There are other situations in which the results of Theorem 8.2.2 and

8.3.3 might be useful. One such situation is described as follows:

Let \( P_i(\lambda_i), i = 1, \ldots, p, \text{ be p independent Poisson processes with} \)
intensity parameters $\lambda_i$. Let $X_i$ be the number of counts of the process observed during the period of time $(0,t_i)$, $i = 1, \ldots, p$. The MLE of $\lambda = (\lambda_1, \ldots, \lambda_p)$ is $\hat{\lambda} = \left( \frac{X_1}{t_1}, \ldots, \frac{X_p}{t_p} \right)$. The results in subsection 8.2 and 8.3 show that there are other estimators $\hat{\lambda}$ of $\lambda$ dominating the MLE under squared error loss or under $k$-NSEL, provided the $t_i$'s satisfy certain conditions. In fact, the estimators described in Corollaries 8.2.3, 8.3.4, and 8.3.5 can be employed here with $n_i$ replaced by $t_i$ and $n_\ast = t_\ast = \text{Min}(t_i)$ and $n_\ast = t_\ast = \text{Max}(t_i)$. 
SECTION 9. SUGGESTIONS FOR FURTHER RESEARCH

In Section 6, we gave an empirical Bayes interpretation for the estimators derived under the normalized squared error loss $L_1$. It is therefore natural to seek a similar interpretation for our new estimators $\hat{\lambda}^k$ in Section 4 which are derived under $k$-NSEL $L_k$, with $k > 2$. Moreover, since some of the estimators under $L_1$ are actually Bayes estimators, one might attempt to check if some of the estimators of $\hat{\lambda}^k$ are Bayes estimators under some appropriate prior distributions. If such prior distributions can be found, then we have a clear understanding of when we should use those estimators $\hat{\lambda}^k$.

In the squared error loss case, Peng [1975] has already suggested that some insights about his estimators $\hat{\lambda}^{(0)}$ (see Theorem 2.2.3) might be obtained from an empirical Bayes interpretation for the estimators. Up to this point, no successful attempt has been made.

Though the estimators $\hat{\lambda}^{(k)}$ suggested in Section 3 dominate the MLE under squared error loss when $p > 3$, they are not admissible. It might be of interest to see if there are admissible estimators which also dominate the MLE.
BIBLIOGRAPHY


