# The $c$-Loss Function: Balancing Total and Individual Risk in the Simultaneous Estimation of Poisson Means

**Emil Aas Stoltenberg**

**THESIS**
**for the degree of**
**MASTER OF SCIENCE**

(Master i Modellering og dataanalyse)

*Department of Mathematics*
*The Faculty of Mathematics and Natural Sciences*
*University of Oslo*
*May 2015*

# The $c$-Loss Function: Balancing Total and Individual Risk in the Simultaneous Estimation of Poisson Means

2015

The $c$-Loss Function: Balancing Total and Individual Risk in the Simultaneous Estimation of Poisson Means

Emil Aas Stoltenberg

http://www.duo.uio.no/

Print: Reprosentralen, Universitetet i Oslo

# Abstract

This thesis is devoted to the simultaneous estimation of the means of $p \geq 2$ independent Poisson distributions. A novel loss function that penalizes bad estimates of each of the means and the sum of the means is introduced. Under this loss function, a class of minimax estimators that uniformly dominate the MLE, is derived. This class is shown to also be minimax and uniformly dominating under the commonly used weighted squared error loss function. Estimators in this class can be fine-tuned to limit shrinkage away from the MLE, thereby avoiding implausible estimates of means anticipated to be bigger than the others. Further light is shed on this new class of estimators by showing that it can be derived by Bayesian and empirical Bayesian methods. Moreover, a class of prior distributions for which the Bayes estimators are minimax and dominate the MLE under the new loss function, is derived. Estimators that shrink the observations towards other points in the parameter space are derived and their performance is compared to similar estimators previously studied in the literature. The most important finding of the thesis is the aforementioned class of estimators that provides the statistician with a convenient way of compromising between two conflicting desiderata (good total *and* individual risk) when estimating an ensemble of Poisson means.

# Preface

I am grateful to my supervisor Nils Lid Hjort for making it so enjoyable to work with this master thesis. The theme of this thesis grew out of the exam project in the course STK4021 on Bayesian statistics that Lid Hjort gave in the autumn of 2013. Decision theory, and particularly its Bayesian version, appealed to me because it provides a coherent basis for statistics and rational decision making. The move from the way statistics is often taught in introductory courses (maximum likelihood, unbiasedness, $p$-values etc.), to the manner in which Lid Hjort lectured over the Bayesian and decision theoretic approach to statistics in STK4021, was my first encounter with the creative and truly fun side of statistics. During the work with this thesis, statistics has at no moment ceased being fun. Quite the contrary. For that I owe many thanks to my supervisor.

Two years ago I submitted my master's thesis at the Faculty of Social Sciences. In that thesis I attempted to predict the outcome of parliamentary elections in Norway, using a Dirichlet-Multinomial model. I was - and still am - proud of that thesis, but did not want to leave the University of Oslo before gaining a more profound understanding of what I had really attempted to do. I am truly glad that I chose to stay at Blindern for two more years, and got at little closer.

I owe many thanks to Calina Haslum Langguth, Adrien Henri Vigier, Tore Wig, Atle Aas, Camilla Stoltenberg and Vilde Sagstad Imeland. All errors are my own.

Blindern, May 2015<br>
Emil Aas Stoltenberg

# Contents

# Figures

# Tables

# 1

# Introduction

This thesis deals with the simultaneous estimation of the parameters of several independent Poisson random variables. Assume that $Y_1, \ldots, Y_p$ are independent Poisson with means $\theta_1, \ldots, \theta_p$, and let $Y$ and $\theta$ be the $p \times 1$ vectors of observations and means. We wish to estimate $\theta$ using an estimator $\delta = (\delta_1, \ldots, \delta_p)$. The two most common loss functions for this problem are the squared- and weighted squared error loss functions, defined by

$$L_m(\delta, \theta) = \sum_{i=1}^{p} \frac{1}{\theta_i^m} (\delta_i - \theta_i)^2$$

with $m = 0$ and $m = 1$ respectively. The usual estimator of $\theta$ is $\delta^o(Y) = Y$, which is the maximum likelihood estimator (MLE), the minimum variance unbiased estimator and minimax with respect to $L_1$ (see Appendix A.2). Peng (1975) showed that the MLE is inadmissible under $L_0$ when $p \geq 3$ and derived an estimator that performs uniformly better than the MLE in terms of risk. Working with the $L_1$ loss function Clevenson and Zidek (1975) derived an estimator shown to possess uniformly smaller risk than the MLE for $p \geq 2$.

As we will see, both Peng's estimator and the estimator of Clevenson and Zidek shrink $Y$ towards the zero boundary of the parameter space. Consequently, both estimators show most of their risk improvement for values of $\theta$ close to zero. Since zero is the boundary of the parameter space for the Poisson distribution, small parameters can only be badly overestimated. Bigger parameter values, on the other hand, can be badly underestimated and shrinkage estimators such as Peng's and that of Clevenson and Zidek might shrink large counts by an amount resulting in implausible estimates of large means. In particular, the shrinkage can be thought to be too large if one in addition to estimating each individual Poisson mean whishes to make sure that the estimate of the sum of the means is not corrupted. These two observations constitute parts of the rationale for the loss function that is the primary focus of this thesis, namely the $c$-Loss function. The $c$-Loss function is a generalization of

$L_1$ and is defined by

$$L_c(\delta,\theta) = \sum_{i=1}^{p} \frac{1}{\theta_i} (\delta_i - \theta_i)^2 + \frac{c}{\sum_{i=1}^{p} \theta_i} \left( \sum_{i=1}^{p} \delta_i - \sum_{i=1}^{p} \theta_i \right)^2 ,$$

where $c$ is a non-negative parameter. In this thesis I show that the MLE is inadmissible under $c$-Loss when $p \geq 2$, and derive a class of minimax estimators that uniformly dominate the MLE. In effect, it is shown that the class of estimators derived in this thesis is also uniformly dominating under the weighted squared error loss function. As such, this class of estimators provides a compromise between the MLE and the shrinkage estimators previously studied in the literature (e.g. Clevenson and Zidek (1975) and Ghosh et al. (1983)), and makes it possible for the statistician to control the amount of shrinkage away from the MLE without sacrificing the risk function optimality.

In the next section I provide an overview of the background for this thesis, namely *Stein's paradox* (see e.g. Efron and Morris (1977)). I introduce the James-Stein estimator and uniformly dominant shrinkage estimators associated with the Poisson distribution. In Section 1.3 I discuss some of the limitations related to estimators derived with the decision theoretic aim of uniform dominance relative to the MLE, and mention some of the interactions between the Bayesian and the decision theoretic approaches. A complete outline of the full thesis is given in Section 1.4.

## 1.1 The risk function

Most statistical studies lead to some form of decision.[1] These decisions range from an inference concerning the probability distribution underlying some phenomenon, to whether one should undertake some action or another. Whatever the objective of a statistical study, it is vital to have some criterion by which to evaluate the consequences of each decision depending on the true state of nature (see e.g. Robert (2001, 51)). In statistical decision theory, this criterion is the loss function $L(\delta, \theta)$. The loss function is a function from $\mathcal{D} \times \Theta$, where $\mathcal{D}$ is the set of all possible decisions and $\Theta$ is the parameter space (the true state of nature), to the positive part of the real line $\mathbb{R}_+$. In this thesis, the set of decisions is the parameter space. In other words, the loss function is used as the criterion for evaluating the performance of an estimator $\delta \in \mathcal{D} = \Theta$ in estimating an unknown parameter $\theta \in \Theta$. An effective way to compare different estimators is the *risk function*

$$R(\delta,\theta) = E_\theta L(\delta,\theta) = \int_{\mathcal{Y}} L(\delta,\theta) f(y|\theta)\, dy, \tag{1.1}$$

[1] Efron (1982, 1986) distinguishes between four basic operations of statistics: enumeration, summary, comparison and inference. They do not all lead to a decision.

which provides the average performance of the estimator in terms of loss. The risk function, since it integrates over the sample space $\mathcal{Y}$, is called the *frequentist* risk function. Throughout this thesis the frequentist risk function is the main tool used for evaluating and comparing the performance of different estimators.

One principle by which to decide whether or not to use an estimator is the principle of *admissibility.* According to this principle an estimator $\delta_0$ is *inadmissible* if there exists an estimator $\delta_1$ such that $R(\delta_0, \theta) \geq R(\delta_1, \theta)$ for all $\theta \in \Theta$, with strict inequality for at least one $\theta_0$. If no such estimator exists, $\delta_0$ is admissible. In this thesis all the risk functions that I consider are continuous, which means that the strict inequality $R(\delta_0, \theta) > R(\delta_0, \theta)$ must hold for an interval $(\underline{\theta}_0, \bar{\theta}_0)$ in $\Theta$ for $\delta_0$ to be inadmissible.

Obviously, it is hard to advocate the use of an inadmissible estimator because there exist other estimators that always perform better. Yet for a given decision problem, there are usually many different admissible estimators. These estimators have risk functions that may cross so that they will be better in different regions of the parameter space (Berger, 1985, 10). In situations where $R(\delta_0, \theta) \geq R(\delta_1, \theta)$ for all $\theta \in \Theta$, with strict inequality for at least one $\theta_0$, I will say that $\delta_1$ *dominates* the estimator $\delta_0$. If the inequality is strict for all $\theta \in \Theta$ the estimator $\delta_0$ will be said to *uniformly dominate* $\delta_0$. Throughout, when I speak of dominance or uniform dominance, the dominance alluded to will always be relative to the MLE if not indicated otherwise.

## 1.2 Shrinkage estimators

Stein's paradox or Stein's phenomenon (Berger, 1985, 360) is in its simplest form the following: Let $X_1, \ldots, X_p$ be independent normal random variables with means $\xi_1, \ldots, \xi_p$ and unit variance. It is desired to estimate these means under the squared error loss function $\sum_{i=1}^{p} (\delta_i - \xi_i)^2$. The MLE of $\xi$ is $\delta^o(X) = X$. Stein (1956) showed that the MLE is inadmissible when $p \geq 3$, and shortly after James and Stein (1961) provided a constructive result by proving that for $p \geq 3$ any estimator of the form

$$\delta^{JS}(X) = \left(1 - \frac{b}{\sum_{i=1}^{p} X_i^2}\right) X, \tag{1.2}$$

where $b$ is a constant with $0 < b < 2(p-2)$, uniformly dominates the MLE.[2] The best choice of $b$ in terms of minimizing risk is $p-2$ (Stigler, 1990, 147). The reason for this result meriting (for some time) the appelation "paradox" is that the observations are independent. At first sight it is hard to understand why

[2] The positive part version of the James-Stein estimator $\left(1 - (p-2)/||X||^2\right)^+ X$, where $(a)^+ = \max\{a, o\}$, improves on $\delta^{JS}$ and stops the estimator from shrinking past zero when $||X||^2 < p - 2$ (Robert, 2001, 98).

information about the price of beer in Bergen and about the unemployment rate in Belgrade might improve the estimate of the height of women in Berlin.

Now, the results of Peng (1975) and Clevenson and Zidek (1975) show that the same quasi-paradoxical conclusion applies to the simultaneous estimation of the means of independent Poisson distributions. Under the squared error loss function $L_0$ Peng proved that the estimator

$$\delta_i^P(y) = Y_i - \frac{(N_0(Y)-2)^+}{D(Y)} h_i(Y_i), \tag{1.3}$$

where $N_\nu(Y) = \#\{i : Y_i > \nu\}$, $h_i(Y_i) = \sum_{k=1}^{Y_i} 1/k$, $D(Y) = \sum_{i=1}^{p} h_i^2(Y_i)$ and $(a)^+ = \max\{a, 0\}$, uniformly dominates the MLE when $p \geq 3$.

The estimator of Clevenson and Zidek for $L_1$ is given by

$$\delta^{CZ}(Y) = \left(1 - \frac{\psi(Z)}{p-1+Z}\right) Y, \tag{1.4}$$

where $Z = \sum_{i=1}^{p} Y_i$ and $\psi$ is a non-decreasing function with $0 < \psi(z) \leq 2(p-1)$. This estimator uniformly dominates the MLE under the weighted squared error loss function when $p \geq 2$. As is evident from (1.2), (1.3) and (1.4) all three estimators shrink the MLE towards zero and the amount of shrinkage decreases as the size of the observations increases. Consequently, for large values of the unknown means, when large observations are to be expected, these three estimators will not differ much from the MLE and the savings in risk will be very modest. The obvious way to fix this deficiency is to modify the estimators so that they shrink the MLE towards some other point in the parameter space than the origin. Provided that hypotheses about the means exist, say they are all close to a common value $\xi_0$, the James-Stein estimator is particularly straightforward to modify for it to yield substantial savings in risk in regions about this point. Using $\delta^{JS}$ for the deviations $x_i - \xi_0$ rather than $x_i$ one obtains an estimator that shrinks the MLE towards $\xi_0$. By way of an empirical Bayesian argument, Lindley (1962, 286) proposed an estimator that uses the empirical mean $\bar{x} = p^{-1} \sum_{i=1}^{p} x_i$ as an estimator of $\xi_0$, leading to

$$\delta^L(x) = \bar{x} + \left(1 - \frac{p-3}{||x-\bar{x}||^2}\right)(x - \bar{x}), \tag{1.5}$$

with $p-3$ as the constant since the parameter $\xi_0$ is estimated (Efron and Morris, 1975, 312). The Lindley-estimator dominates the MLE under $L_0$ in dimension 4 and higher.

For the Poisson distribution uniformly dominating estimators that shrink towards some non-zero point have more complicated forms. This is due to the non-symmetry of the Poisson distribution and zero being the boundary of the parameter space. Moreover, under $L_1$ the weighting implies that overestimation of very small $\theta_i$ incurs heavy penalizations, which limits the possibility of smoothing the observations towards some common point (Hudson, 1985, 249). Nevertheless, simultaneous estimators have been obtained that shrink the Poisson counts towards a non-zero point. The natural place to start in order to find such estimators is to modify $\delta^P$ and $\delta^{CZ}$ in a manner mimicking that of Lindley (1962). That is by subtracting an appropriate quantity from $h$ and $Y$ in (1.3) and (1.4) respectively. Tsui (1981), Hudson and Tsui (1981) and Ghosh et al. (1983) have provided constructive results in this direction where the estimators shrink the MLE towards some pre-chosen or data-generated point in the parameter space. Drawing on the ideas of Stein (e.g. Stein (1981)) these authors considered competitors of the MLE of the form $\delta^o(Y) + F(Y)$ where $F(Y) = (F_1(Y), \ldots, F_p(Y))$, and provided conditions on $F$ for the new estimators to have uniformly smaller risk than the MLE. The most general result concerning such estimators is that of Ghosh et al. (1983), who proved theorems that apply to general loss functions $L_m$ for discrete exponential family distributions (Theorem 3.1 and 4.1 in Ghosh et al. (1983)). By way of these two theorems it is possible to construct functions $F$ for estimators that shrink towards any point in the parameter space. If correctly constructed, the theorems of Ghosh et al. (1983) guarantee that the resulting estimator $Y + F(Y)$ dominates the MLE. In Chapter 3 I provide a more thorough presentation of these estimators and compare their performance to estimators derived in this thesis.

Finally, in many situations hypotheses will exist about the means having some common structure. In the terminology employed so far, this means that one anticpates improvements in risk by smoothing the observations towards several *different* points in the parameter space. Once again, in the case of the normal distribution the James-Stein estimator can be modified to smooth the observations towards any linear model $\xi_i = z_i^t\beta$, $1 \leq i \leq p$. Here $z_i$ is a $k \times 1$ vector of covariates and $\beta$ a $k \times 1$ coefficient vector. The estimator is obtained by replacing the deviances $x_i - \bar{x}$ in (1.5) by $x_i - z_i^t\beta$. Suppose that $\beta$ is estimated from the data by $Z(Z^tZ)^{-1}Z^tx$ where $Z$ is the $p \times k$ design matrix with $z_i^t$ as its rows and $x = (x_1, \ldots, x_p)^t$ (see e.g. Morris (1983b)). This gives the estimator

$$\delta^{EB}(x) = Z\hat{\beta} + \left(1 - \frac{p-k-2}{||x - Z\hat{\beta}||^2}\right)(x - Z\hat{\beta}). \tag{1.6}$$

For the reasons touched upon above, such smoothing estimators are more complicated in the case of the Poisson distribution. The theorems of Ghosh et al. (1983) can be used to derive estimators that shrink towards different points, and Hudson and Tsui (1981, 183) proposed an estimator that shrinks towards different a priori hypotheses about the means. These estimators do not, however, smooth the observations since observations below its hypothesized or data generated point are estimated by the MLE (cf. $\delta_i^{G1}$ in (3.1)). Hudson (1985, 248-249) proposes an estimator that shrinks the observations towards a log-linear model, relying on an approximate log transform of the Poisson data. Due to the transformation, the risk calculations involving this estimator are not exact, and dominance relative to the MLE cannot be established. In Section 4.3 I compare the performance of such smooth-to-structure estimators with similar estimators derived in this thesis.

Constructive results, that is those actually proposing improved estimators, have been obtained for other distributions than the normal and Poisson. Berger (1980, 557-560) obtained estimators that improve on the MLE in estimating the scale parameters of several independent Gamma distributions assuming that the shape parameters are known. Ghosh et al. (1983) derived uniformly dominating estimators (under $L_0$ and $L_1$) for the $w_1, \ldots, w_p$ when the observations come from $p$ independent Negative binomial distributions with parameters $(a_i, w_i)$, assuming that $a_1, \ldots, a_p$ are known. Under $L_1$, Tsui (1984, 155) derived dominating estimators of the means $a_i w_i/(1-w_i)$ of $p$ independent Negative binomial distributions when the $a_i$ are either known or unknown.

What the simulation studies in Chapter 3 and Chapter 4 make clear is that estimators developed with the decision theoretic aim of uniform dominance often yield rather limited improvements in risk compared to the MLE. Berger (1983, 368) suggests that risk domination might be too severe a restriction when estimating multiple Poisson means, and that we might be better off abandoning it in favour of Bayes and empirical Bayes methods producing estimators with superior performance in certain regions of the parameter space. In the next section I discuss such (possibly non-dominating) estimators with a particular focus on the interaction between the decision theoretic and the Bayesian approaches.

## 1.3 Challenges and Bayesian techniques

There are primarily three challenges associated with simultaneous estimators of Poisson means derived under the decision theoretic aim of uniform dominance. The first has to do with the sensitivity of the dominant estimators to the form of the loss function. The second has already been touched upon, namely that the dominant estimators in many cases yield very modest improvements in risk compared to the MLE. Associated with this challenge, a third challenge concerns the balance between good risk performance and sensible estimates of the individual parameters (Efron and Morris, 1972, 130).

In the problem of estimating normal means the choice of squared error loss or weighted squared error loss is unimportant when it comes to dominance (Berger, 1983, 369). Under both loss functions the dominating estimators are on the form (1.2) (provided that the observations have equal variance). As is seen from the estimators $\delta^{P}$ and $\delta^{CZ}$, this is not the case for the Poisson distribution. The form of the loss function does matter. Since a decision maker in many instances is likely to be uncertain about the appropriate loss function, it is preferable that the estimators work well under a variety of different loss functions (Berger, 1983; Robert, 2001, 83). A seemingly innocuous exigence is that the estimators work well under the two most common loss functions, $L_0$ and $L_1$. So far no estimator has been found that uniformly improves on the MLE under both these loss functions.

Since the MLE is minimax under $L_1$ one cannot expect to find estimators that perform substantially better than the MLE in all regions of the parameter space.[3] This means that if one wants to achieve substantial improvements in risk, one has to specify a region where this improvement should occur and accept that the estimator might perform worse than the MLE outside this region (Albert, 1981, 401). In other words, prior information must be brought into the estimation problem. In the development of such estimators a great success in the case of the normal distribution has been the interaction between Bayesian (and empirical Bayesian) methods and the frequentist decision theoretic approach. Bayesian methods have been used in the derivation of estimators, while decision theoretic ideas have been used to fine-tune these estimators (Berger, 1983, 368). Prime examples are the James-Stein type estimators in (1.2) and (1.5) that can be derived by Bayesian methods and then fine-tuned to minimize risk. As we will see in Chapter 3, Bayes and empirical Bayes estimators outperform the uniformly dominating estimators of Poisson means in certain regions of the parameter space. A pertinent question (connected to what is known as *Bayesian robustness* (Ghosh et al., 2006, 72)) is how well Bayesian estimators perform when the prior information is misspecified. In order to make the estimators robust to misspecifications of the prior distribution decision theoretic techniques come into play. A case in point is Albert (1981) who starts out with $Y_i$, $1 \leq i \leq p$ independent Poisson and the conjugate Gamma prior on the means. The posterior distribution is then $\pi(\theta_i \,|\, y_i) \propto \theta_i^{a+y_i-1} \exp\{-(b+1)\theta_i\}$ and the Bayes estimator (under $L_0$) is

$$E[\theta_i \,|\, \text{data}] = \frac{a}{b}\frac{b}{b+1} + \left(1 - \frac{b}{b+1}\right) y_i.$$

Albert then replaces the weight $b/(b+1)$ by $B(y_i)$ and uses decision theoretic techniques to derive a function $B$ that limits shrinkage for observations far from the prior mean. By way of simulations Albert's estimator is shown to perform better than the MLE in a region of the parameter space and to possess smaller risk than $E[\theta_i \,|\, \text{data}]$ in non-favourable regions of the parameter space (i.e. regions of $\Theta$ far from the prior mean).

A slightly different use of Bayesian methods in the frequentist decision theoretic setting is to look at what kind of prior information (if any) that results in dominating estimators. This can illuminate the nature of the dominating estimator at hand. For example, as I show in Section 2.7, the estimator of

[3] Under the squared error loss function $L_0$, there exist no estimator $\delta$ for which $\sup_\theta R(\delta, \theta)$ is finite (see Appendix A.2).

Clevenson and Zidek (1975) can be derived as an empirical Bayes estimator when the Poisson means are independent and come from an exponential distribution with mean $1/b$, where $b$ is estimated from the data.

Lastly, uniformly dominating estimators are constructed to guarantee a reduction in the total risk $R(\delta, \theta)$, but may perform poorly in terms of estimating the individual components $\theta_i$ (Efron and Morris, 1971, 1972). Both the weighted squared error loss function and the squared error loss function can be decomposed into $p$ individual loss functions. Illustratively, we can express the risk as

$$R(\delta, \theta) = \sum_{i=1}^{p} E_\theta L_i(\delta_i, \theta_i) = \sum_{i=1}^{p} R_i(\delta_i, \theta_i).$$

Such a decomposition does not work for the $c$-Loss function. Uniformly dominating estimators guarantee a risk $R$ that is smaller than the risk of the MLE, but may give non-plausible estimates of $\theta_i$ that are different from the others. So even though the risk is uniformly smaller than that of the MLE, this does not prevent one or more individual risk components $R_i$ from being bigger than the corresponding risk components of the MLE (the extent to which depends on the weighting scheme). I look further into this when I motivate the $c$-Loss function in Chapter 2. For now the point is that in using a shrinkage estimator that "has good ensemble properties, the statistician must be aware of the possibility of grossly misestimating the individual components" (Efron and Morris, 1972, 130). An important part of the rationale behind the $c$-Loss function is to develop a compromise between the Clevenson and Zidek estimator and the MLE, which can compete with $\delta^{CZ}$ under the weighted squared error loss function and has decent individual properties.

## 1.4 Outline of the thesis

In Chapter 2 I introduce and give reasons for the $c$-Loss function. Then in Section 2.1 I derive the Bayes solution to the $c$-Loss function and use this solution to prove that the MLE is minimax. Subsequently, in Section 2.3 I derive an explicit estimator that uniformly dominates the MLE under the $c$-Loss function, and show that this estimator belongs to a larger class of minimax estimators. Section 2.4 consists of a comparison of the new estimator derived in 2.3 and the estimator of Clevenson and Zidek (1975). The $c$-parameter is the focus of Section 2.5, where I propose a method for determining the value of this parameter.

In Section 2.6 I extend the $c$-Loss function to situations with different exposure times or multiple independent observations from each of the $p$ distributions, and derive a class of minimax estimators that uniformly dominate the MLE. It is shown that this class of estimators contains the class of estimators derived for the equal exposures/ single-observation case. In Section 2.7 I show that estimators uniformly dominating the MLE under the $c$-Loss function can be constructed from Bayesian arguments in three different ways: as proper Bayes estimators, as an empirical Bayes estimator, and finally as a generalized Bayes estimator where the assumption of independence of the means is relaxed.

Chapter 3 consists of two parts. In the first part in Section 3.1 I study estimators that shrink the MLE towards non-zero points in the parameter space under $L_0$. I derive an estimator by frequentist methods and study its relation to empirical Bayes estimators. Subsequently, the performance of this estimator is compared to those of Ghosh et al. (1983). Secondly, Section 3.2 consists of the same type of analysis, but now under $L_1$. Importantly, I derive a new uniformly dominating estimator that shrinks a subset of the observations to a pre-specified point in the parameter space.

Situations where the Poisson means are thought to have some common structure are studied in Chapter 4. I study a Bayesian hierarchical regression model where the prior expectation of the Poisson mean is log-linear in the covariates. In addition, the model allows for inference on the variance of the unknown means. Via a simulation study this Bayesian regression model is compared to other standard and non-standard Poisson regression models.

Finally, in Chapter 5 I conclude and discuss themes for further research.

# 2

# The $c$-Loss function

In many applications the weighted squared error loss function $L_1$ is the natural loss function to use when estimating an ensemble of Poisson means. The primary reason for this is that in situations where the Poisson parameters are expected to be small, good estimates of $\theta_i$ close to zero are desired. Since the weighted squared error loss function incurs a heavy penalization for bad estimates of small parameter values it is therefore a natural choice. Moreover, contrary to the squared error loss function the *weighted* squared error loss function takes into account that it is not possible to badly underestimate small parameter values (Clevenson and Zidek, 1975, 698).

A slightly different way to view $L_1$ is as an *information* weighted loss function. The Fisher information of a Poisson observation is $\theta_i^{-1}$, which means that the information that an observation $Y_i$ contains about $\theta_i$ is large for small $\theta_i$, and decreasing as the parameter values increase (Lehmann, 1983, 120). Thus, the loss function $L_1$ incurs a heavy penalization for bad estimates of $\theta_i$ when $Y_i$ contains much information about $\theta_i$.

As touched upon in Section 1.1 there are some potential deficiencies with the the estimator $\delta^{CZ}$ derived under $L_1$. The two that provide the motivation for the $c$-Loss function relates to (i) the issue of striking the right balance between good (total) risk performance and sensible estimates of the individual parameters; and (ii) a good estimate of the sum of the individual parameters.

The estimator of Clevenson and Zidek shrinks all the observations towards zero, and in some situations $\delta^{CZ}$ might shrink the observations by an amount deemed to be too large. Particularly, this would be the case in situations with many small or zero counts and a few large counts. As an example, consider a Poisson estimation problem with $p = 7$ observed counts $y = (0, 1, 0, 1, 0, 1, 5)$. The Clevenson and Zidek estimator will here shrink the MLE of $\theta_i$ by 43%, giving 2.86 as our estimate of $\theta_7$. This means that if we trust our estimate of $\theta_7$, the probability of observing what we actually observed or something more extreme is only $1 - P(5 \leq 2.86) = 0.07$. This appears audacious.

More generally, in situations where we in addition to good risk performance, are interested in individual parameters or subpopulations of the parameters it might be advantageous to find a compromise between the Clevenson and Zidek estimator and the MLE. In the terminology of Efron and Morris (1972) estimators that achieve such compromises "limits translation" away from the MLE. Importantly, by shrinking the MLE in a Poisson estimation problem the "gross misestimation of individual components" referred to above, can only occur for parameters a certain distance from zero, simply because zero is the boundary of the parameter space. This means that what we gain by limiting shrinkage of large observations might outweigh what we lose in limiting shrinkage of the small observations.

In many decision problems where one wants to estimate several Poisson means, the decison maker might also be interested in a good estimate of the sum of the Poisson means, or equivalently the mean of the means $\bar{\theta} = p^{-1}\sum_{i=1}^{p}\theta_i$. For example, a decision maker having to make budgetary decisions concerning each of the boroughs of a city *and* the city as a whole, will find herself in such a situation. Recall that the sum of $p$ independent Poisson random variables is itself Poisson with mean equal to the sum of the $p$ Poisson means. I will denote this sum by $\gamma$. A good estimate of $\gamma$ can be of interest in its own right, but it can also be seen as a compromise in situations where the decision maker is uncertain as to whether the $p$ observations come from $p$ different Poisson distributions or the same, i.e. $\theta_1 = \cdots = \theta_p$.

Even though the two considerations discussed above appear to be at odds since they concern individual parameters contra the sum of the parameters, they amount to the same thing (this will be made precise below). As such they provide the motivation for the $c$-Loss function. As mentioned, let $\gamma = \sum_{i=1}^{p}\theta_i$. Then the $c$-Loss function is defined by

$$L_c(\delta, \theta) = \sum_{i=1}^{p} \frac{1}{\theta_i} (\delta_i - \theta_i)^2 + \frac{c}{\gamma}\left(\sum_{i=1}^{p}\delta_i - \gamma\right)^2 . \tag{2.1}$$

This loss function is equal to the weighted squared error loss function $L_1$ plus an extra term, where the weight accorded to this extra term is a function of the user-defined constant $c$. Now I will make precise how the two consideration (i) and (ii) amount to the same thing. Since $\gamma$ is the mean of the Poisson random variable $Z = \sum_{i=1}^{p} Y_i$, the second term in (2.1) is the loss function

$$\frac{c}{\gamma}\left(\sum_{i=1}^{p}\delta_i - \gamma\right)^2 = \frac{cp}{\bar{\theta}}\left(\bar{\delta} - \bar{\theta}\right)^2 ,$$

where $\bar{\delta} = p^{-1}\sum_{i=1}^{p}\delta_i$. The MLE of $\gamma$ is $Z$, and equivalently the MLE of $\bar{\theta}$ is $\bar{Y} = p^{-1}\sum_{i=1}^{p} Y_i$. In this one-dimensional case the MLE is admissible and the unique minimax solution, which means that the MLE cannot be uniformly improved upon (Lehmann, 1983, 277). Stated differently, since $Z$ is the sum of the individual MLE's of $\theta_i$, namely $Y_i$, shrinkage away from $Y_i$ must deteriorate the risk component that penalizes for bad estimates of $\gamma$. This implies that larger values of $c$ must limit shrinkage away from the individual MLE's of $\theta_i$, and consequently one avoids situations as the one above where one of the estimates (that of $\theta_7$) appeared non-plausible. On the other hand, if $c = 0$ we have $\delta^{CZ}$. Hence the compromise between the Clevenson and Zidek estimator and the MLE.

If the decision maker is uncertain about the number of Poisson distributions being $p > 1$ or one, the $c$-Loss function penalizes the decision maker for mistakenly assuming that she is dealing with $p$ different distributions when she is in fact dealing with one and the same. This is because the individual MLE's cannot be uniformly improved upon if the observations come from the same Poisson distribution. To see this, consider the risk of the first term in the $c$-Loss function when $\theta_1 = \cdots = \theta_p = \theta_0$. The risk is then $E_\theta L_1(\delta, \theta_0) = \theta_0^{-1} \sum_{i=1}^{p} E_\theta \left(\delta_i - \theta_0\right)^2 = p\theta_0^{-1} \left(\delta_i - \theta_0\right)^2$. The point is that if this is in fact the loss function, the MLE (which is $\delta_i = \bar{Y} = p^{-1}Z$ for all $i$) cannot be uniformly improved upon (Lehmann, 1983, 277), and at the same time $Z$ cannot be improved upon in estimating $\gamma$ irrespective of the number of Poisson distributions being $p$ or one. Hence, larger values of $c$ pulls the estimates towards what would have been best if the observations came from the same distribution.

Before I move ahead and derive an estimator that uniformly dominates the MLE under the $c$-Loss function, I consider the Bayes solution and use this to establish that the MLE is minimax. The minimax result will be of importance in later sections.

## 2.1 Bayes and minimax

The risk function $R(\delta, \theta)$ in (1.1) is called the frequentist risk function because it intergrates over the sample space $\mathcal{Y}$. In the Bayesian approach to decision theory the sample is taken as given and one instead integrates over the parameter space $\Theta$ to obtain the *posterior expected loss*

$$\rho(\pi(\theta \,|\, y), \delta) = E[L(\delta, \theta) \,|\, y] = \int_\Theta L(\delta, \theta) \pi(\theta \,|\, y) \, d\theta.$$

Associated with the posterior expected loss is the *Bayes risk* $\mathrm{BR}(\pi, \delta)$ which is the expectation of the frequentist risk with respect to the prior distribution of $\theta$,

$$\mathrm{BR}(\pi, \delta) = ER(\delta, \theta) = \int_\Theta \left\{ \int_\mathcal{Y} L(\delta, \theta) f(y|\theta) \, dy \right\} \pi(\theta) \, d\theta.$$

An important relation between the posterior expected loss and the Bayes risk is that the estimator that minimizes $\rho(\pi(\theta \,|\, y), \delta)$ is also the estimator that minimizes $\mathrm{BR}(\pi,\delta)$ (Berger, 1985, 159). This follows from Fubini's theorem since

$$\begin{aligned}\mathrm{BR}(\pi,\delta) &= \int_\Theta \int_{\mathcal{Y}} L(\delta,\theta) f(y|\theta)\, dy\, \pi(\theta)\, d\theta \\ &= \int_{\mathcal{Y}} \int_\Theta L(\delta,\theta)\pi(\theta|y)\, d\theta\, m(y)\, dy = \int_{\mathcal{Y}} \rho(\pi(\theta \mid y),\delta)\, m(y)\, dy,\end{aligned}$$

where $m(y) = \int_\Theta f(y|\theta)\pi(\theta)\, d\theta$ is the marginal distribution of the data. The estimator that minimizes the posterior expected loss (equivalently the Bayes risk) is called the *Bayes solution* (or Bayes estimator) and will be denoted $\delta^B$. The quantity $\mathrm{BR}(\pi,\delta^B)$ will be called the *minimum Bayes risk* and be denoted $\mathrm{MBR}(\pi)$.

The Bayes solution is of interest either because one is a Bayesian trusting the prior used, or as a tool to develop improved estimators in frequentist settings. In a first part I derive the Bayes solution to the $c$-Loss function and study its properties. Thereafter, I use the Bayes estimator to prove that the MLE is minimax under the $c$-Loss function. In Section 2.7 I continue the Bayesian analysis and derive uniformly dominating estimators by Bayesian methods.

### 2.1.1 The Bayes solution

Let $Y_1,\ldots,Y_p$ be $p \geq 2$ independent Poisson random variables with means $\theta_1,\ldots,\theta_p$. Assume that the Poisson means are independent and come from a prior distribution $\pi(\theta)$ on $\Theta$. Consider the posterior expected loss and change the order of integration and derivation, take the partial derivative with respect to $\delta_j$ and set this equal to zero. This gives,

$$\begin{aligned}E\left[\frac{\partial}{\partial\delta_j} L_c(\delta,\theta) \mid Y\right] &= E\left[2\frac{1}{\theta_j}(\delta_j-\theta_j) + \frac{2c}{\gamma}\left(\sum_{i=1}^p \delta_i - \gamma\right)\right] \\ &= 2\delta_j E[\theta_j^{-1} \mid Y] - 2 + 2c\left\{\sum_{i=1}^p \delta_i\right\} E[\gamma^{-1} \mid Y] - 2c = 0.\end{aligned}$$

Since this equation must hold for all $j = 1,\ldots,p$ we get the system of equations given by,

$$\begin{aligned}\delta_1(E[\theta_1^{-1}|Y] + cE[\gamma^{-1}|Y]) + cE[\gamma^{-1}|Y] \sum_{\{i:i\neq 1\}} \delta_i &= 1+c \\ \vdots \qquad\qquad\qquad & \qquad \vdots \\ \delta_p(E[\theta_p^{-1}|Y] + cE[\gamma^{-1}|Y]) + cE[\gamma^{-1}|Y] \sum_{\{i:i\neq p\}} \delta_i &= 1+c.\end{aligned}$$

In Appendix A.3 I solve this system of equations and obtain a general expression for the $j$'th coordinate of the Bayes estimator, namely

$$\delta_j^B(Y) = \frac{1+c}{E[\theta_j^{-1}|Y]\left(1 + cE[\gamma^{-1}|Y]\sum_{i=1}^p E[\theta_i^{-1}|Y]^{-1}\right)}. \tag{2.2}$$

In principle, as long as it is possible to explicitly compute the posterior expectations in (2.2), analytic expressions for the Bayes estimator can be obtained. Here, I will rely on the conjugate Gamma prior distribution. Let the Poisson means $\theta_1, \dots, \theta_p$ be independent Gamma random variables with means $a/b$ and variances $a/b^2$, denoted $\mathcal{G}(a,b)$. If not explicitly stated otherwise, I assume that $a > 1$. Given the data the distribution of $\theta_i$ is $\mathcal{G}(a + y_i, b+1)$ for $i = 1, \dots, p$. Some integration included in Appendix A.1 shows that generally if $G \sim \mathcal{G}(\alpha, \beta)$ and $\alpha > 1$, then $E[G^{-1}] = \beta/(\alpha - 1)$. Thus, with a $\mathcal{G}(a,b)$ prior the posterior expectations in (2.2) are $E[\theta_j^{-1}|Y] = (b+1)/(a+y_j-1)$ and $E[\gamma^{-1}|Y] = (b+1)/(pa + Z - 1)$ where $Z = \sum_{i=1}^p Y_i$. Since

$$\sum_{i=1}^p \{E[\theta_i^{-1}|Y]\}^{-1} = \sum_{i=1}^p \left(\frac{b+1}{a+y_i-1}\right)^{-1} = \frac{p(a-1)+Z}{b+1},$$

the latter term in the denominator in (2.2) is $cg(z)$ where $g$ is defined as $g(z) = (p(a-1)+z)/(pa-1+z)$. In summary, the Bayes estimator with the conjugate Gamma prior is given by

$$\delta_j^{\mathrm{B}}(y) = \frac{1+c}{\frac{b+1}{a+y_j-1}\left(1 + c\frac{p(a-1)+z}{pa-1+z}\right)} = \frac{1+c}{1+cg(z)}\frac{a+y_j-1}{b+1}. \tag{2.3}$$

Recall that $p \geq 2$, and assume that $a \geq 1$. Then we see that $0 < g < 1$, and that the first factor in (2.3) is always bigger than one. Note also that we might write $\delta_j^{\mathrm{B}} = (1+c)/(1+cg(z))\{E[\theta_j^{-1}|y]\}^{-1}$ where $\{E[\theta_j^{-1}|y]\}^{-1}$ is the Bayes solution under weighted squared error loss $L_1$. Hence

$$\{E[\theta_j^{-1} \,|\, y]\}^{-1} = \frac{a+y_j-1}{b+1} < \frac{a+y_j}{b+1} = E[\theta_j \,|\, y],$$

where $E[\theta_j \,|\, y]$ is the Bayes estimator under the squared error loss function $L_0$. This shows that the effect of weighting (i.e. $L_1$ vs. $L_0$) is to shrink the estimates relative to the Bayes estimator under $L_0$. Under the $c$-Loss function, this shrinkage effect is counteracted by the term $(1+c)/(1+cg(z))$ which is increasing in $c$. This means that as more weight is put on a good estimate of $\gamma = \sum_{i=1}^p \theta_i$, the Bayes estimator will move closer to $E[\theta_j \,|\, y]$. More succinctly, when $a \geq 1$ we will always have that

$$\begin{array}{ccccc} \{E[\theta_j^{-1}|y]\}^{-1} & \leq & \delta_j^B & \leq & E[\theta_j|y] \\ \uparrow & & \uparrow & & \uparrow \\ L_1 & & L_c & & L_0, \end{array}$$

where the arrows indicate the Bayes solutions associated with the given loss function. Where in the interval between $\{E[\theta_j^{-1}\,|\,y]\}^{-1}$ and $E[\theta_j\,|\,y]$ the Bayes solution under the $c$-Loss function is to be found is a function of $Z$ and the user-defined constant $c$. We see that $g$ is strictly increasing in $z$ and that as $z$ grows $g$ converges to one from below. Thus, the factor $(1+c)/(1+cg(z))$ is decreasing towards one, which means that

$$\delta_j^B \longrightarrow \{E[\theta_j^{-1}|y]\}^{-1},$$

as $z$ goes to infinity. This is natural, because a very large $Z$ is unlikely to occur unless $\gamma$ is very large. A very large $\gamma$ means that the Fisher information $\gamma^{-1}$ contained in the observation $Z$ is very small, and the second term in the $c$-Loss function can be disregarded.

### 2.1.2 Minimaxity

In this section I prove that the MLE $\delta^o(Y) = Y$ is minimax under the $c$-Loss function. An estimator is minimax if it minimizes the expected loss in the worst possible scenario. More precisely, the estimator $\delta^m$ is minimax if $\sup_\theta R(\delta^m,\theta) \leq \sup_\theta R(\delta,\theta)$ for all estimators $\delta$. From the definition of the minimum Bayes risk above it follows that

$$\mathrm{MBR}(\pi) = \int_\Theta R(\delta^B,\theta)\pi(\theta)\,d\theta \leq \int_\Theta R(\delta,\theta)\pi(\theta)\,d\theta \leq \sup_{\theta\in\Theta} R(\delta,\theta),$$

where $\delta$ can be any estimator, hence also the minimax estimator. Therefore, we have the relation $\mathrm{MBR}(\pi) \leq \sup_\theta R(\delta^m,\theta)$, where $\delta^m$ is minimax. In order to prove that the MLE is minimax under the $c$-Loss function I will use the following lemma (well known in the literature).

**Lemma 2.1.1.** *Let $\{\pi_n\}_{n=1}^\infty$ be a sequence of prior distributions for which the minimum Bayes risk satisfies*

$$MBR(\pi_n) \to \sup_\theta R(\delta,\theta)$$

*as $n\to\infty$. Then $\delta$ is minimax.*

*Proof.* Assume that $\mathrm{MBR}(\pi_n) \to \sup_\theta R(\delta,\theta)$ as $n\to\infty$. Let $\delta^*$ be any other estimator. Then

$$\sup_\theta R(\delta^*,\theta) \geq \int_\Theta R(\delta^*,\theta)\pi_n(\theta)\,d\theta \geq \mathrm{MBR}(\pi_n)$$

for all $n \geq 1$. Since this holds for all $n$ we must have that $\sup_\theta R(\delta^*,\theta) \geq \sup_\theta R(\delta,\theta)$. Since $\delta^*$ is any estimator this implies that $\delta$ must be minimax. □

Under the $c$-Loss function the MLE has constant risk

$$\begin{aligned} R(Y,\theta) &= E_\theta L_c(Y,\theta) \\ &= \sum_{i=1}^{p} \theta_i^{-1} E_\theta (Y_i - \theta_i)^2 + c\gamma^{-1} E_\gamma (Z-\gamma)^2 = p + c. \end{aligned}$$

In view of Lemma 2.1.1, in order to show that the MLE is minimax I must show that the MBR($\pi_n$) converges to $p+c$ as $n$ goes to infinity for a given sequence of priors. Such a sequence of priors does, indeed, exist.

**Theorem 2.1.2.** *The MLE $\delta^o(Y) = Y$ is minimax under the $c$-Loss function.*

*Proof.* (See Appendix A.4 for a more thorough version of this proof). Consider the prior sequence $\{\pi_n\}_{n=1}^{\infty} = \{\mathcal{G}(1, b_n)\}_{n=1}^{\infty}$ where $b_n = b/n$. With this prior sequence the Bayes estimator in (2.3) is given by

$$\delta_j^B = \frac{(1+c)(p-1+z)}{p-1+(1+c)z} \frac{y_j}{b_n+1}.$$

Using that $Y_i \,|\, Z$ is binomial with mean $Z\eta_i$ and variance $Z\eta_i(1-\eta_i)$ where $\eta_i = \theta_i/\gamma$ (see Lemma A.1.1), the risk of $\delta^B$ can be written

$$\begin{aligned} R(\delta^B,\theta) &= E_\theta L(\delta^B,\theta) = E_\theta E[L(\delta^B,\theta)|Z] \\ &= E_\theta \left\{ \frac{(1+c)^2(p-1+Z)^2}{p-1+(1+c)Z} \frac{Z}{\gamma(b+1)^2} \right. \\ &\qquad \left. - \{2\frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z} \frac{Z}{b+1} + (1+c)\gamma \right\}. \end{aligned}$$

Because

$$\mathrm{MBR}(\pi) = E^\pi[E_\gamma[L(\delta^B,\theta)]] = \int_{\mathcal{Y}} E^{\pi^*}[L(\delta^B,\theta) \,|\, Z]\, m(z)\, dz,$$

where $m(z)$ is Negative binomial with parameters $p$ and $(b+1)^{-1}$, and

$$\gamma \,|\, Z \sim \pi^*(\gamma \,|\, z) = \mathcal{G}(p+z, b_n+1),$$

the minimum Bayes risk can be expressed as

$$\begin{aligned} \mathrm{MBR}(\pi_n) &= E^\pi R(\delta^B,\theta) = E^m[E^{\pi^*}[L(\delta^B,\theta) \,|\, Z]] \\ &= E^m \left[ \frac{1+c}{b+1} \left\{ p - \frac{c(p-1)Z}{p-1+(1+c)Z} \right\} \right]. \end{aligned} \tag{2.4}$$

Define the integrand in the last line in Equation (2.4) as $h(z)$, so that MBR($\pi_n$) $= E^m[h(Z)]$ and use that $h(z)$ is a convex function. Then by Jensen's inequality we have that

$$\begin{aligned} \mathrm{MBR}(\pi_n) &= E^m\,[\,h(Z)] \geq h(E^m[\,Z]\,) \\ &\geq \frac{(1+c)p}{b_n+1} - \frac{1+c}{b_n+1} \frac{cp(p-1)}{b_n(p-1)+(1+c)p}, \end{aligned} \tag{2.5}$$

where I have used that the marginal expectation of $Z$ is $E^m[Z] = p/b_n$. The expression in (2.5) converges to $p+c$ as $n \to \infty$. Hence, $\text{MBR}(\pi_n) \geq p+c$ for all $n$. But since $\text{MBR}(\pi_n) \leq \sup_\theta R(\delta, \theta)$ the minimum Bayes risk must converge to $p+c$ as $n \to \infty$. By Lemma 2.1.1 we then have that the MLE $\delta^o(Y) = Y$ is minimax under $c$-Loss. □

This theorem will be important in what follows because it implies that estimators that uniformly dominate the MLE are minimax. Moreover, since $Y$ is a constant risk minimax estimator and inadmissible (see Theorem 2.3.2), it follows that $Y$ is the worst minimax estimator (Robert, 2001, 97).

## 2.2 Two crucial lemmata

A crucial identity in proving risk dominance of the James-Stein estimator is what is known as Stein's identity. If $X$ is normally distributed with mean $\xi$ and unit variance, and $g$ is a function that satisfies some very mild conditions (Stein, 1981, 1136), then $E_\xi(X-\xi)g(X) = E_\xi g'(X)$ where $g'$ is the derivative of $g$. The importance of this identity derives from the fact that it enables us to express the improvement in risk of an estimator over the MLE as a function that is independent of the unknown parameters. In the Poisson setting, an analogous result holds.

**Lemma 2.2.1.** *If $Y$ is Poisson with mean $\theta$ and $f$ is a function such that $f(y) = 0$ for all $y \leq 0$ and $E_\theta|f(Y)| < \infty$, then*

$$E_\theta f(Y)/\theta = E_\theta f(Y+1)/(Y+1) \tag{2.6}$$

*and*

$$E_\theta \theta f(Y) = E_\theta f(Y-1)Y. \tag{2.7}$$

*Let $Y = (Y_1, \dots, Y_p)$ be a $p$ dimensional vector of independent Poisson random variables with means $\theta_1, \dots, \theta_p$, and let $F(Y) = (F_1(Y), \dots, F_p(Y))$ be a function where $F_i : \mathbb{N}^p \to \mathbb{R}$ is such that $F_i(y) = 0$ if $y_i \leq 0$ and $E_\theta|F_i(Y)| < \infty$. Then*

$$E_\theta F_i(Y)/\theta_i = E_\theta F_i(Y + e_i)/(Y_i + 1)$$

*and*

$$E_\theta \theta_i F_i(Y) = E_\theta F_i(Y - e_i) Y_i,$$

*where $e_i$ is the $p \times 1$ vector whose i'th component is one and the rest are zero.*

Versions of this lemma are used and proved in all contributions to the literature on the simultaneous estimation of Poisson means (see e.g. Tsui and Press (1982, 94)). It will be used throughout this thesis. For completeness, I include the proof.

*Proof.* The proof of (2.6) is

$$E_\theta[f(Y)/\theta] = \sum_{y=0}^{\infty} \frac{f(y)}{\theta}\frac{1}{y!}\theta^y e^{-\theta} = \frac{f(0)}{\theta} + \sum_{y=1}^{\infty} f(y)\frac{1}{y!}\theta^{y-1}e^{-\theta}$$
$$= 0 + \sum_{y=0}^{\infty} f(y+1)\frac{1}{(y+1)y!}\theta^y e^{-\theta} = E_\theta[f(Y+1)/(Y+1)].$$

To prove the equivalent identity for the multivariate case condition on $\{Y_j \,|\, j \neq i\}$,

$$E_\theta F_i(Y)/\theta_i = E_\theta E[F_i(Y)/\theta_i \,|\{Y_j \,|\, j \neq i\}] = E_\theta F_i(Y + e_i)/(Y_i + 1).$$

The proofs of (2.7) and its multivariate extension are similar and are included in Appendix A.5. □

The idea, originally due to Stein (see e.g. Stein (1981)), for finding improved estimators is to consider competitors that are equal to the MLE plus an extra term, $\delta^*(Y) = \delta^o(Y) + f(Y)$. Then, the goal is to express the *difference in risk* $R(\delta^*, \theta) - R(\delta^o, \theta)$ as a function independent of the unknown parameters.

**Lemma 2.2.2.** *If $R(\delta^*, \theta) - R(\delta^o, \theta) = E_\theta D(Y)$ and $D(Y) \leq 0$ for all $Y$ with strict inequality for at least one datum $Y$, then $\delta^*$ dominates $\delta^o$ in terms of risk. In other words, $\delta^o$ is inadmissible.*

This lemma follows from the fact that if $X$ and $Y$ are two random variables such that $X \leq Y$, then $EX \leq EY$. When $D$ is independent of the unknown parameters I write $D = D(Y)$, when the difference in loss is not independent of the unknown parameters it will be denoted $D(Y \,|\, \theta)$.

## 2.3 Finding improved estimators

In order to find a class of estimators that improve on the MLE under the $c$-Loss function I consider estimators $\delta^* = (\delta_1^*, \ldots, \delta_p^*)$ of the form

$$\delta^* = (1 - \phi(Z))Y, \tag{2.8}$$

where $Z = \sum_{i=1}^p Y_i$ and $E_\theta|\phi(Z)| < \infty$. To find an expression for the difference in risk between $\delta^*$ in (2.8) and the MLE $\delta^o(Y) = Y$ I first look at the difference of the loss functions. By expanding the squares we get the following expression

$$D_c(Y \,|\, \gamma) = \sum_{i=1}^{p} \frac{1}{\theta_i}(\phi^2(Z) - 2\phi(Z))Y_i^2 - c\frac{1}{\gamma}(\phi^2(Z) - 2\phi(Z))Z^2 + 2(1+c)\phi(Z)Z.$$

Then use the fact that conditional on $Z$, the random variable $Y = (Y_1, \ldots, Y_p)$ is multinomial with $Z$ trials and cell probabilites $\eta_i = \theta_i/\gamma$, and apply Lemma 2.2.1. This gives the following lemma.

**Lemma 2.3.1.** *Under the c-Loss function the difference in risk is given by*

$$\begin{aligned}E_\theta[D_c(Y\,|\,\gamma)] &= E_\theta E[D_c(Y\,|\,\gamma)|Z]\\ &= E_\theta\left\{(\phi^2(Z)-2\phi(Z))\frac{Z[(p-1)+(1+c)Z]}{\gamma}+2(1+c)\phi(Z)Z\right\}\\ &= E_\theta\left\{(\phi^2(Z+1)-2\phi(Z+1))[(p-1)+(1+c)(Z+1)]\right.\\ &\qquad\qquad\left.+\,2(1+c)\phi(Z)Z\right\}\\ &= E_\theta D_c(Z).\end{aligned}$$

Here $D_c$ is independent of the parameters, so a function $\phi(Z)$ that ensures that $D_c(Z)\le 0$ for all $Z$ with strict inequality for at least one $Z$ yields an estimator that uniformly dominates the MLE. Assume that $\phi(z)z \le \phi(z+1)(z+1)$ for all $z\ge 0$, and define

$$D_c^*(Z)=(\phi^2(Z+1)-2\phi(Z+1))[(p-1)+(1+c)(Z+1)]+2(1+c)\phi(Z+1)(Z+1).$$

Then $R(\delta^*,\theta)-R(Y,\theta)=E_\theta D_c(Y)\le E_\theta D_c^*(Y)$ for all $Y\in\mathcal{Y}$. Treating $\phi$ as a constant, then taking the partial derivative with respect to $\phi$ and setting this equal to zero, we get

$$\frac{1}{2}\frac{\partial}{\partial\phi}D_c^*(Z)=[\phi(Z+1)-1](p-1+(1+c)(Z+1))+(1+c)(Z+1)=0,$$

which is solved for $\phi_c(Z+1)=(p-1)/(p-1+(1+c)(Z+1))$. We thereby obtain the estimator

$$\delta_1^c(Y)=\left(1-\frac{p-1}{p-1+(1+c)Z}\right)Y, \tag{2.9}$$

which is, as expected, equal to $\delta^{CZ}(Y)$ for $c=0$. In order to show that this estimator dominates the MLE we must show that its risk is smaller than that of the MLE. The risk of the MLE is constant and equal to $p+c$.

**Theorem 2.3.2.** *For all $\theta\in\Theta$ the estimator $\delta_1^c$ given in* (2.9) *has smaller risk than the MLE $\delta^o(Y)=Y$ when loss is given by $L_c$ in* (2.1).

*Proof.* Use the expression for $D_c$ above, then

$$\begin{aligned} R(\delta_1^c, \theta) &= p + c + D_c \\ &= p + c - \frac{1}{\gamma} E_\gamma \{\phi(Z) Z [2(p - 1 + (1+c)(Z-\gamma)) \\ &\qquad\qquad\qquad\qquad - \phi(Z)[p-1+(1+c)Z]]\} \\ &= p + c - \frac{1}{\gamma} E_\gamma \{\phi(Z) Z [2(p - 1 + (1+c)(Z-\gamma)) - (p-1)]\} \\ &\leq p + c - \frac{2}{\gamma} E_\gamma \{\phi(Z) Z (1+c)(Z-\gamma)\} \\ &= p + c - 2(1+c) E_\gamma \left[\frac{\phi(Z)Z^2}{\gamma} - \phi(Z)Z\right] \\ &= p + c - 2(1+c) E_\gamma [\phi(Z+1)(Z+1) - \phi(Z)Z] \\ &< p + c - 2(1+c) E_\gamma [\phi(Z+1)(Z+1) - \phi(Z+1)(Z+1)] \\ &= p + c = R(Y, \theta), \end{aligned}$$

for all $\theta$ because $\phi(Z)Z$ is a strictly increasing function of $Z$ (this will be proved in a more general setting in Corollary 2.3.3 below). □

Since the MLE is minimax, this means that we have found a minimax estimator with uniformly smaller risk than the MLE. In effect, from the the first inequality of the proof of Theorem 2.3.2 it is easy to see that $\delta_1^c$ is a member of a larger class of minimax estimators, all with uniformly smaller risk than the MLE. That is $\delta_1^c \in \mathcal{D}_c$ where $\mathcal{D}_c$ is the class of all estimators of the form

$$\delta^c(Y) = \left(1 - \frac{\psi(Z)}{p - 1 + (1+c)Z}\right) Y,$$

where the function $\psi$ is such that $0 < \psi(z) < 2(p-1)$ is non-decreasing for all $z \geq 0$.

**Corollary 2.3.3.** *All estimators $\delta^c \in \mathcal{D}_c$ have uniformly smaller risk than the MLE.*

*Proof.* This is basically the same proof as for $\delta_1^c$.

$$\begin{aligned} R(\delta^c, \theta) &= p + c - \frac{1}{\gamma} E_\gamma \{\phi(Z) Z [2(p - 1 + (1+c)(Z-\gamma)) - \psi(Z)]\} \\ &\leq p + c - \frac{2}{\gamma} E_\gamma \{\phi(Z) Z (1+c)(Z-\gamma)\} \\ &= p + c - 2(1+c) E_\gamma \left[\frac{\phi(Z)Z^2}{\gamma} - \phi(Z)Z\right] < p + c = R(Y, \theta), \end{aligned}$$

since $0 < \psi(z) \leq 2(p-1)$, hence $2(p-1) - \psi(z) > 0$ and we get the first inequality. The second inequality is obtained by using Lemma 2.2.1 to get rid of $\gamma$, then using that $\phi(Z)Z$ is strictly increasing. This function is strictly increasing because $\psi$ is non-decreasing, hence $\psi'(z) \geq 0$ for all $z \geq 0$. Denote $G(z) = p - 1 + (1+c)z$, then $G'(z) = (1+c)$. The first derivative of $\phi(z)z$ with respect to $z$ is then

$$\frac{d}{d\,z}\,\phi(z)z = \frac{(\psi'(z)z + \psi(z))G(z) - (1+c)\psi(z)z}{G(z)^2}.$$

The denominator in this expression is always positive, and the nominator is

$$\begin{aligned}(\psi'(z)z + \psi(z))G(z) - (1+c)\psi(z)z& \\ &\geq \psi(z)G(z) - (1+c)\psi(z)z = (p-1)\psi(z) > 0,\end{aligned}$$

since $0 < \psi(z) \leq 2(p-1)$ and $p \geq 2$. Thus $(\phi(z)z)' > 0$, which proves the assertion. □

As with $\delta^{CZ}$ the new estimators $\delta^c \in \mathcal{D}_c$ shrink the MLE towards the origin, but the amount of shrinkage is less than for $\delta^{CZ}$, and can be controlled by the statistician. In Figure 2.1 I plot the risk of $\delta_1^c$ for different values of $\gamma$. This plot is obtained by adding $p + c$ to the expression for the difference in risk in Lemma 2.3.1, which yields an expression the loss of $\delta_1^c$ that is solely a function of $Z$, hence $R(\delta_1^c, \theta)$ is a function of $\gamma$ only,

$$R(\delta^c, \theta) = E_\gamma \left\{ \frac{(p-1)^2}{p-1+(1+c)(Z+1)} - \frac{2(p-1)^2}{p-1+(1+c)Z} \right\} + p + c.$$

This risk function can be computed numerically for increasing values of $\gamma$ in order to compare the loss of $\delta_1^c$ to that of the MLE. In Figure 2.1 one clearly sees how the savings in risk are substantial for small values of $\gamma$, while for larger values of $\gamma$ the risk of $\delta_1^c$ becomes almost indistinguishable from that of the MLE. In the next section I undertake a more thorough comparison of $\delta_1^c$ and the estimator of Clevenson and Zidek (1975).

## 2.4 A comparison of the estimators

Further insight into the difference between the two estimators $\delta^{CZ}$ and $\delta_1^c$ is gained by studying how $\delta^{CZ}$ performs under the $c$-Loss function and how $\delta_1^c$ performs under the weighted squared error loss function $L_1$. Recall that an important part of the motivation for the $c$-Loss function was, in the terminology of Efron and Morris (1971, 1972), to “limit translation” away from the MLE in order to achieve more plausible estimates of large $\theta_i$. The question is whether this strategy of limiting shrinkage away from the MLE works to the detriment

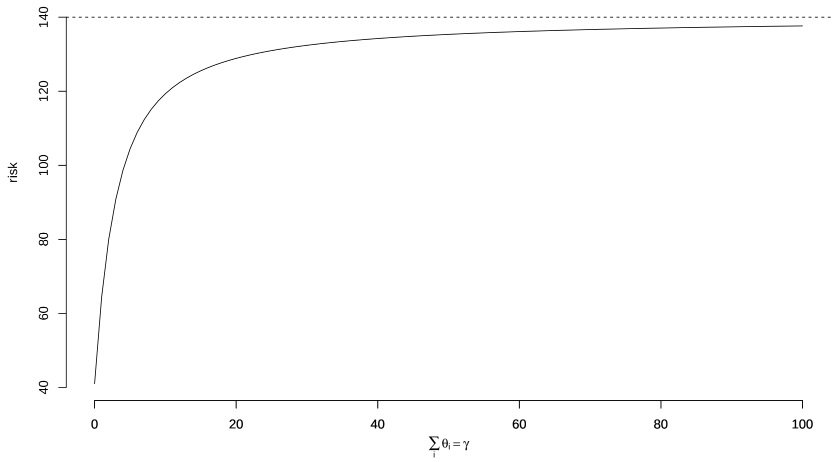

Figure 2.1: A plot of $R(\delta_1^c, \theta)$ with sample size $p = 100$ and $c = 40$ for increasing values of $\gamma = \sum_{i=1}^p \theta_i$. The horizontal line is the constant risk $p + c = 140$ of the MLE.

of the uniform dominance under weighted squared error loss, or not. Intuitively, since in a weighted squared error loss sense

$$\text{dist}(Y, \theta) \geq \text{dist}(\delta^{CZ}, \theta),$$

any estimator on the line segment between $Y$ and $\delta^{CZ}$ should have smaller risk than the MLE. Continuing this heuristic argument, since $\delta^{CZ}$ and $Y$ are the two limiting cases of $\delta^c$ when $c = 0$ and $c \to \infty$ respectively, $\delta^c$ should belong to the class of minimax estimators dominating the MLE when loss is $L_1$. This is indeed the case.

**Corollary 2.4.1.** *All estimators $\delta^c \in \mathcal{D}_c$ have uniformly smaller risk than $\delta^o(Y) = Y$ under weighted squared error loss $L_1$.*

*Proof.* The risk of $\delta^c \in \mathcal{D}_c$ under $L_1$ is obtained by setting $c = 0$ in the second line in the proof of Theorem 2.3.2, that is

$$\begin{aligned} R(\delta^c, \theta) &= p - \frac{1}{\gamma} E_\gamma \left\{ \phi(Z) Z [2(p - 1 + (Z - \gamma)) - \phi(Z)[p - 1 + Z]] \right\} \\ &\leq p - \frac{1}{\gamma} E_\gamma \left\{ \phi(Z) Z [2(p - 1 + (Z - \gamma)) - \psi(Z)] \right\}, \end{aligned}$$

where the inequality follows since $(p - 1 + Z)/(p - 1 + (1 + c)Z) \leq 1$ for all values of $c \geq 0$. Proceed as in the proof of Corollary 2.3.3. $\square$

A similar result for the estimator of Clevenson and Zidek exposed to the $c$-Loss function does not hold in general. To see this, write

$$\phi^{CZ}(z) = \frac{\psi(z)}{p-1+z} = \frac{p-1+(1+c)z}{p-1+z}\,\phi^c(z) = \frac{\psi^*(Z)}{p-1+(1+c)Z},$$

where $\psi^*(Z) = \{(p-1+(1+c)Z)/(p-1+Z)\}\,\psi(Z)$. The Clevenson and Zidek estimator can then be expressed as

$$\delta^{CZ} = \left(1 - \frac{\psi^*(Z)}{p-1+(1+c)Z}\right) Y_i.$$

Here, the nominator in the shrinkage term has

$$\sup_{z\geq 0} \psi^*(Z) = (1+c)\,2(p-1) \geq 2(p-1),$$

with equality for $c = 0$ only. This shows that the function $\psi^*$ does not in general satisfy the conditions of Corollary 2.3.3. The optimal Clevenson and Zidek estimator in terms of minimizing risk is obtained by setting $\psi(Z)$ in (1.4) equal to $p-1$. With this choice of $\psi$, we see that the estimator satisfies the conditions of Corollary 2.3.3 provided that $c \leq 1$. In conclusion, the estimator of Clevenson and Zidek does not in general dominate the MLE under $c$-Loss. This fact is amply illustred by the simulations summarized in Table 2.2. Table 2.1 summarizes simulation results for $p = 5$, $p = 10$ and $p = 15$ under the weighted squared error loss function $L_1$. In these simulations I follow the approach of Hwang (1982, 97-98) and Ghosh et al. (1983), which consists of drawing a sample $\theta_i$, $1 \leq i \leq p$ from a uniform distribution on the interval $(a, b)$. Then one observation from each of the $p$ distributions are generated. This second step is repeated $10^5$ times and the estimated risks of $\delta^{CZ}$, $\delta^c$ and the MLE are calculated under $L_1$. Finally, the percentage savings in risk from using an alternative estimator $\delta$ as compared to the MLE, $[R(Y,\theta) - R(\delta,\theta)]/R(Y,\theta) \times 100$ are calculated and

| | $p=5$ | | $p=10$ | | $p=15$ | |
|---|---|---|---|---|---|---|
| range of $\theta_i$ | $\delta^{CZ}$ | $\delta^c$ | $\delta^{CZ}$ | $\delta^c$ | $\delta^{CZ}$ | $\delta^c$ |
| (0,4) | 34.59 | 9.84 | 21.09 | 6.65 | 18.71 | 5.99 |
| (0,8) | 25.54 | 7.45 | 20.94 | 6.55 | 19.12 | 6.08 |
| (8,12) | 18.36 | 5.51 | 18.65 | 5.86 | 17.82 | 5.68 |
| (12,16) | 14.65 | 4.46 | 16.66 | 5.26 | 16.55 | 5.29 |
| (0,12) | 14.84 | 4.49 | 16.19 | 5.10 | 16.36 | 5.21 |
| (4,16) | 13.75 | 4.16 | 15.24 | 4.80 | 15.76 | 5.01 |

Table 2.1: Percentage savings in risk relative to the MLE $Y$ under weighted squared error loss $L_1$. The parameter $c$ was set to 5 in the simulations.

| | $p = 5$ | | $p = 10$ | | $p = 15$ | |
|---|---|---|---|---|---|---|
| range of $\theta_i$ | $\delta^{CZ}$ | $\delta^c$ | $\delta^{CZ}$ | $\delta^c$ | $\delta^{CZ}$ | $\delta^c$ |
| (0,4) | -15.41 | 3.04 | -15.18 | 2.17 | -18.37 | 2.15 |
| (0,8) | -15.22 | 2.26 | -16.70 | 2.19 | -19.44 | 2.22 |
| (8,12) | -13.06 | 1.70 | -16.42 | 2.01 | -19.21 | 2.11 |
| (12,16) | -11.61 | 1.38 | -15.86 | 1.84 | -18.76 | 1.99 |
| (0,12) | -12.14 | 1.38 | -16.30 | 1.79 | -19.13 | 1.97 |
| (4,16) | -11.91 | 1.27 | -16.15 | 1.70 | -19.07 | 1.91 |

Table 2.2: Percentage savings in risk relative to the MLE under the $c$-Loss function. The parameter $c$ was set to 5 in the simulations.

reported in the table. In Table 2.2 the results of the same simulation procedure with the $c$-Loss function are summarized. Under weighted squared error loss we see that the savings in risk are sizable when using the estimator $\delta^{CZ}$. As expected, the largest reduction in risk 34.59%, is obtained for the interval with the smallest values of $\theta_i$ when $p = 5$. Under $L_1$ the new estimator $\delta^c$ does not perform as impressively as $\delta^{CZ}$, but still improves on the MLE for all intervals and all values of $p$. Under the $c$-Loss function the new estimator $\delta^c$ outperforms the estimator of Clevenson and Zidek. In Table 2.2 we clearly see that $\delta^{CZ}$ is a lousy estimator when one seeks good estimates of $\gamma$ in addition the the individual $\theta_i$. That being said, the new estimator only leads to an estimated reduction in risk of around 2% compared to the MLE in these simulations. Keep in mind, however, that a value of $c$ fine-tuned to the sample size and the expected size of the parameters, would likely give more improvement in risk (cf. Section 2.5).

In the data that gave rise to the study of Clevenson and Zidek (1975), $\theta_i$ represented the expected number of oilwell discoveries in the Canadian province of Alberta obtained from wildcat exploration during month $i$ over a period of thirty years (Clevenson and Zidek, 1975, 703). In order to compare the estimators on the wildcat exploration data I follow the strategy of Clevenson and Zidek and use the observation for every third month of each half year, March and September in the period from 1953 to 1970. Their strategy is to use the average number of discoveries in each half year surrounding the monthly observation to provide the "true" value of the parameter. Table A.1 included in Appendix A.7 summarizes this empirical comparison. In this study $c$ was set to 40 to limit shrinkage for the relatively large observations $y_{13} = y_{16} = 3$ and $y_{25} = 5$. Despite this large value of $c$ the new estimator $\delta^c$ differs markedly from the MLE. Under $L_1$ the total losses are $L_1(Y, \theta) = 39.26$, $L_1(\delta^{CZ}, \theta) = 14.34$ and $L_1(\delta_1^c, \theta) = 19$. As predicted by theory, both $\delta^{CZ}$ and the new estimator perform much better than the MLE, giving reductions in loss of 63.47% and 51.60% respectively. Exposed to the $c$-Loss function $\delta^c$ and $\delta^{CZ}$ are not able to match the performance of the MLE. The loss ratios are $L_c(\delta^c, \theta)/L_c(Y, \theta) = 1.63$ and $L_c(\delta^{CZ}, \theta)/L_c(Y, \theta) = 5.03$, which shows that the estimator of Clevenson and Zidek (1975) performs very poorly when it is penalized for bad estimates of $\gamma$.

Recall that this is an empirical comparison and does not invalidate the theoretical results derived above. To see this I simulated $10^5$ draws from the 36 Poisson distributions in Table A.1 and computed the risk of the three estimators under $L_1$ and $L_c$ with $c = 40$. In accordance with theory, $\delta^c$ now beats the MLE with risk reductions of 6.02% under $L_1$ and 1.58% under the $c$-Loss function. The Clevenson and Zidek estimator improves by 57.16% under $L_1$, but once again performs very poorly under the $c$-Loss function. Its estimated risk ratio is $R(\delta^{CZ}, \theta)/R(Y, \theta) = 4.87$.

## 2.5 A closer look at $c$

One of the primary motivations for the $c$-Loss function was to strike a balance between good risk performance and plausible estimates of the individual parameters. Using the new estimator $\delta^c$, this balancing is controlled by the value of $c$. The risk performance that we consider here is with respect to the weighted squared error loss function (with respect to the $c$-Loss function the question "what $c$" would be tautological). Exposed to $L_1$ the estimator $\delta^c$ minimizes risk when $c$ is set to zero, but $c = 0$ might produce non-plausible estimates of large parameters. The question is therefore how to choose $c$.

I will now propose one way in which the $c$ parameter can be chosen. Consider the weighted squared error loss function in estimating $\gamma$ with $\delta$, namely

$$L(\delta, \gamma) = \frac{1}{\gamma}(\delta - \gamma)^2 . \tag{2.10}$$

Under this loss function $Z = \sum_{i=1}^{p} Y_i$ is the MLE, minimax and cannot be uniformly improved upon (Lehmann, 1983, 277). In addition, the MLE has constant risk equal to one. Using Lemma 2.2.1 we find that the risk of $\delta_1^c$ under (2.10) is

$$R(\delta_1^c, \gamma) = E_\gamma \left[L(\delta_1^c, \gamma)\right] = 1 + E_\gamma \left[(\phi^2(Z+1) - 2\phi(Z+1))(Z+1) + 2\phi(Z)Z\right].$$

A well posed question that will determine $c$ is then: given that the statistician is willing to make a guess at $\gamma$ and only tolerates a deterioration of risk (under (2.10)) in using $\delta_1^c$ instead of $Z$ of $K\%$, what $c$ value should she choose? With a prior guess of $\gamma$ the risk $R(\delta_1^c, \gamma)$ can be computed and compared to the risk of the MLE. Since the risk performance of $\delta_1^c$ under (2.10) deteriorates with lower values of $c$, $K\%$ provides a lower bound on the $c$ value. The upper bound is naturally provided by the fact that the statistician wants to minimize the risk under $L_1$, $R(\delta_1^c, \theta) = E_\theta \sum_{i=1}^{p} \theta_i^{-1}(\delta_i^c - \theta_i)^2$. In other words, we seek the

| $\theta_i$ | $\theta_i^{-1}(\delta^{CZ} - \theta_i)^2$ | $\theta_i^{-1}(\delta_1^c - \theta_i)^2$ |
|---|---|---|
| 1.48 | 33.94 | 14.94 |
| 0.68 | 39.54 | 14.24 |
| 1.84 | 30.41 | 13.05 |
| 1.24 | 35.32 | 14.43 |
| 1.46 | 33.97 | 14.59 |
| 1.38 | 34.02 | 14.09 |
| 1.64 | 31.95 | 13.60 |
| 1.46 | 33.27 | 13.48 |
| 6.98 | -11.38 | 6.85 |
| 6.91 | -9.62 | 7.90 |
| $L_1$ | 25.14 | 12.72 |

Table 2.3: Estimated total and componentwise percentage savings in risk relative to the MLE under $L_1$. 100000 simulations with $p = 10$ and $c = 3$. See Appendix C.1 for details.

smallest value of $c$ that ensures a deterioration of risk of less than $K\%$, hence our optimal $c$, denoted $\hat{c}$, is

$$\hat{c} = \min_c \left\{ c \in [0, \infty) \mid \frac{R(\delta_1^c, \gamma) - R(Z, \gamma)}{R(Z, \gamma)} \times 100 \leq K\% \right\}. \tag{2.11}$$

Lower tolerance levels $K$ will increase the size of $\hat{c}$ (how much depends on the prior guess of $\gamma$ and the sample size $p$), and consequently $\delta^c$ will be pulled closer to the MLE and the possibility of non-plausible estimates of the large parameters is reduced.

As an illustration of (2.11) Table A.2 in Appendix A.7 gives the optimal $\hat{c}$-values for five different tolerance levels $K$ and varying prior guesses of $\gamma$ for sample sizes of $p = 8$ and $p = 40$. (The R-script used to approximate (2.11) is found in Appendix C.2). A feature of the $\hat{c}$-values reported in Table A.2 is that as a function of $\gamma$, $\hat{c}$ is first increasing, reaching its max for medium sized $\gamma$, then decreasing. This reflects the fact that it is in situations with many small and a few large Poisson means, that fine tuning of $c$ is most critical.

In Table 2.3 I report the results of a simulation study where most of the true parameters were small and a few were large. Eight of the $p = 10$ true parameters ($\theta_i$, $1 \leq i \leq 8$) were generated from a uniform distribution on $(0, 2)$, while two remaining larger ones ($\theta_9$ and $\theta_{10}$) came from a uniform on $(5, 8)$. The R-script in Appendix C.2 was used to find $\hat{c}$. A prior guess of $\gamma$ of 28 and a tolerance level of 10% gave $\hat{c} = 3$. This rather low tolerance level is meant to reflect our anticipation of some of the unknown parameters being larger than the others, and that we desire plausible estimates of these. In Table 2.3 we see that both $\delta^{CZ}$ and $\delta_1^c$ improve on the MLE. Considering the parameters separately, $\delta^{CZ}$ performs much better than the MLE in estimating the small parameters with risk savings well above 30 percent. The new estimator $\delta_1^c$ also improves on the MLE, though with a lesser amount than $\delta^{CZ}$. Crucially, when it comes to the two larger parameters the estimator of Clevenson and Zidek loses against the MLE while $\delta_1^c$ improves on the MLE by around seven percent. This improvement for the two large parameters is in part a product of the low tolerance level $K$ used to find $\hat{c}$.

Due to the fact that both $\delta^{CZ}$ and $\delta^c$ are constructed to improve the *total* risk, a somewhat surprising feature of Table 2.3 is their remarkable performance in terms of the *individual* risks when it comes to estimating the (small) parameters. A possible explanation for this is simply that with one observation from each distribution, the MLE lives in $\mathbb{N} \cup \{0\}$ while $\delta^{CZ}$, $\delta^c$ and the parameters live in $\mathbb{R}_+$. Thus, the minimal distance from the MLE to $\theta_i$ is constrained by the fact that the MLE equals $0, 1, 2, \ldots$ and so on.

In this section I have proposed a method for determining $c$ that relies on the specification of two conflicting desiderata. There are surely other methods. A reassuring property of $\delta^c \in \mathcal{D}_c$ is that they are robust with respect to the weighted squared error loss function. That is, they are minimax and dominate the MLE whatever value the parameter $c \geq 0$ is given.

## 2.6 Different exposure levels

So far I have been studying situations with a single observation from each of the $p$ Poisson distributions, and assumed that the exposures are equal. In many applications the Poisson counts will either be generated over different intervals of time, or we might have $n_i$ independent observations from each of the $p$ distributions. In the first case $Y_1, \ldots, Y_p$ are independent Poisson random variables with means $t_1\theta_1, \ldots, t_p\theta_p$, where $t_i > 0$ for $i = 1, \ldots, p$ are known exposures. In the second case, we observe $Y_{i,1}, \ldots, Y_{i,n_i} \sim \mathcal{P}(\theta_i)$ for $i = 1 \ldots, p$. Obviously, $Y_i = \sum_{j=1}^{n_i} Y_{i,j}$, $1 \leq i \leq p$ are Poisson with means $n_i\theta_i$. The two situations are equivalent. In what follows I will stick to the notation where the exposures are denoted $t_i$.

Recall that parts of the rationale behind the weights in $L_1$ is the fact that $\theta_i^{-1}$ is the Fisher information in an observation $Y_i \sim \mathcal{P}(\theta_i)$. In the exposure/multiple observations case, the Fisher information in an observation $Y_i$ is $t_i/\theta_i$. As a consequence, the information weighting argument leads to the loss function

$$L_{1,t}(\delta, \theta) = \sum_{i=1}^{p} \frac{t_i}{\theta_i} (\delta_i - \theta_i)^2 . \tag{2.12}$$

Quite intuitively, the amount of information in an observation increases with the exposure, and the loss function $L_{1,t}$ penalizes bad estimates heavily when the exposure is high. If, however, the objective is precise estimates of small $\theta_i$, it seems arbitrary to penalize bad estimates as a function of the exposure. In other words, $L_1$ might still be a reasonable loss function in the exposure/multiple observations setting. In Corollary 2.6.2 below I show that the same estimator uniformly dominates the MLE under both $L_1$ and $L_{1,t}$. In the exposure setting the MLE is equal to the observed rate, denoted $r_i = Y_i/t_i$. By Lemma 2.1.1 we can prove that the MLE is minimax under both loss functions $L_1$ and $L_{1,t}$.

**Lemma 2.6.1.** *The MLE whose i'th coordinate is given by $r_i = Y_i/t_i$ is minimax under the loss functions $L_1$ and $L_{1,t}$ as given in* (2.12).

*Proof.* Consider the sequence of priors given by $\pi_n = \{\mathcal{G}(1, b_n)\}_{n=1}^{\infty}$ where $b_n = b/n$. The i'th coordinate of the Bayes estimator under $L_{1,t}$ is then $y_i/(b_n + t_i)$. The minimum Bayes risk is then

$$\begin{aligned}
\mathrm{MBR}(\pi_n) &= EE_\theta \sum_{i=1}^{p} \frac{t_i}{\theta_i}\left(\frac{Y_i}{b_n + t_i} - \theta_i\right)^2 = E\sum_{i=1}^{p}\frac{t_i}{\theta_i}\left\{\frac{t_i\theta_i}{(b_n+t_i)^2} + \left(\frac{t_i\theta_i}{b_n+t_i} - \theta_i\right)^2\right\} \\
&= \sum_{i=1}^{p}\left\{\frac{t_i^2}{(b_n+t_i)^2} + \frac{t_i}{b_n}\left(\frac{-b_n}{b_n+t_i}\right)^2\right\} = \sum_{i=1}^{p}\left\{\frac{t_i^2}{(b_n+t_i)^2} + \frac{t_i b_n}{(b_n+t_i)^2}\right\} \\
&= p - \sum_{i=1}^{p}\frac{b_n}{b_n+t_i} \longrightarrow p = R(r,\theta)
\end{aligned}$$

when $n \to \infty$. Under $L_1$ the $\mathrm{MBR}(\pi_n)$ goes to $\sum_{i=1}^{p} 1/t_i = R(r,\theta)$ when $n \to \infty$ with the same prior sequence. □

I will now prove that the natural generalization of the Clevenson and Zidek estimator to the exposure setting, dominates the MLE under both $L_1$ and $L_{1,t}$.

**Corollary 2.6.2.** *Under both $L_1$ and $L_{1,t}$ the estimator given componentwise by $(1 - (p-1)/(p-1+Z))r_i$ dominates $r_i = Y_i/t_i$.*

*Proof.* Write $(1-(p-1)/(p-1+Z))r_i$ as $r_i + f_i(Y)$ where

$$f_i(Y) = -\frac{(p-1)Y_i/t_i}{p-1+Z}.$$

This function satisfies Lemma 2.2.1. The difference in risk compared to the MLE is then

$$\begin{aligned}
R(\delta^*,\theta) - R(r,\theta) &= \sum_{i=1}^{p}\frac{t_i}{\theta_i}\left\{(r_i + f_i(Y) - \theta_i)^2 - (r_i - \theta_i)^2\right\} \\
&= \sum_{i=1}^{p} t_i\left\{\frac{t_i f_i^2(Y)}{t_i\theta_i} + 2\frac{f_i(Y)Y_i}{t_i\theta_i} - 2f_i(Y)\right\} \\
&= \sum_{i=1}^{p} t_i\left\{\frac{t_i f_i^2(Y+e_i)}{Y_i+1} + 2f_i(Y+e_i) - 2f_i(Y)\right\},
\end{aligned}$$

under $L_{1,t}$, and

$$E_\theta \sum_{i=1}^{p} \left\{ t_i f_i^2(Y+e_i)/(Y_i+1) + 2f_i(Y+e_i) - 2f_i(Y) \right\},$$

under $L_1$. Hence, it suffices to show that the expression inside the brackets is smaller than or equal to zero for all $Y$, with strict inequality for at least one $Y$. Let $t_{\min} = \min_{1\le i\le p} t_i$. Then,

$$\begin{aligned}
D(Y) &= \sum_{i=1}^{p} \left\{ \frac{t_i f_i^2(Y+e_i)}{Y_i+1} + 2\Delta_i f_i(Y+e_i) \right\} \\
&= \sum_{i=1}^{p} \frac{1}{t_i} \left\{ \frac{(p-1)^2(Y_i+1)}{(p+Z)^2} - 2\frac{(p-1)(Y_i+1)}{p+Z} + 2\frac{(p-1)Y_i}{p-1+Z} \right\} \\
&\le \frac{1}{t_{\min}} \left\{ \frac{(p-1)^2}{p+Z} - 2(p-1) + 2\frac{(p-1)Z}{p-1+Z} \right\} \\
&= \frac{1}{t_{\min}} \frac{p-1}{(p+Z)(p-1+Z)} \left\{ (p-1)(p-1+Z) - 2(p-1)(p+Z) \right\} < 0
\end{aligned}$$

since $p \ge 2$. This shows that $D(Y) < 0$ for all $Y$, which means that $(1-(p-1)/(p-1+Z))r$ is minimax and uniformly dominates the MLE. □

In summary, Lemma 2.6.1 and Corollary 2.6.2 show that the exposure-version of the Clevenson and Zidek estimator

$$\delta_i^{CZt}(Y) = \left(1 - \frac{p-1}{p-1+Z}\right)\frac{Y_i}{t_i},$$

is minimax and uniformly dominates the MLE under both $L_1$ and $L_{1,t}$.

Next, these results will be extended to a version of the $c$-Loss function suitably adjusted to the exposure setting. A good extension of $L_c$ is the loss function defined by

$$L_{c,t}(\delta,\theta) = \sum_{i=1}^{p} \frac{t_i}{\theta_i}(\delta_i-\theta_i)^2 + c\frac{\left(\sum_{i=1}^p \delta_i - \sum_{i=1}^p \theta_i\right)^2}{\sum_{i=1}^p \theta_i/t_i}. \tag{2.13}$$

The loss function $L_{c,t}$ penalizes for bad estimates of $\gamma$, and in accordance with the information weighting argument this penalization increases as the exposure increases. On the other hand, the penalization is decreasing in the size of $\theta_i$, reflecting the deference paid to small observations. Note also that when $t_i = 1$ for all $i$, $L_{c,t}$ equals $c$-Loss function. Moreover, a quick calculation shows that the MLE has constant risk

$$\begin{aligned}
R(\delta^o, \theta) &= p + \frac{c}{\sum_{i=1}^p \theta_i/t_i} E_\theta \left( \sum_{i=1}^p Y_i/t_i - \sum_{i=1}^p \theta_i \right)^2 = p + cE_\theta \frac{\left(\sum_{i=1}^p \frac{Y_i - t_i\theta_i}{t_i}\right)^2}{\sum_{i=1}^p \theta_i/t_i} \\
&= p + \frac{c}{\sum_{i=1}^p \theta_i/t_i} E_\theta \sum_{i=1}^p \left( \frac{Y_i - t_i\theta_i}{t_i} \right)^2 + \frac{2c}{\sum_{i=1}^p \theta_i/t_i} \sum_{i \neq j} \frac{1}{t_i t_j} \mathrm{Cov}(Y_i, Y_j) \\
&= p + \frac{c}{\sum_{i=1}^p \theta_i/t_i} \sum_{i=1}^p \frac{\theta_i}{t_i} = p + c,
\end{aligned}$$

since $\mathrm{Cov}(Y_i, Y_j) = 0$ because $Y_i$ and $Y_j$ are independent for $i \neq j$.

In order to find an estimator with risk smaller than $p + c$, I consider estimators of the form $\delta_i^*(Y) = (1 - \phi(Z))Y_i/t_i$ for $i = 1, \ldots, p$. The case of equal exposures is straightforward. When the exposures are equal $t_1 = \cdots = t_p = t > 0$ (equivalently, equal sample sizes), the risk with respect to the loss function in (2.13) reduces to

$$R(\delta^*, \theta) = E_\theta \left\{ \sum_{i=1}^p \frac{1}{t\theta_i} \left((1 - \phi(Z))Y_i - t\theta_i\right)^2 + \frac{c}{t\gamma} \left( \sum_{i=1}^p (1 - \phi(Z))Z - t\gamma \right)^2 \right\}.$$

From this expression we see immediately, using the arguments of Section 2.3, that the improved estimator must be

$$\delta_i^c/t = \left(1 - \frac{p-1}{p-1+(1+c)Z}\right) \frac{Y_i}{t}.$$

In the remainder of this section I derive a class of estimators that dominate the MLE when the exposures are possibly unequal. To prove dominance I derive an expression that bounds the difference in risk between estimators of the type $\delta^*$ and $r_i$. The difference in risk is

$$\begin{aligned}
E_\theta D_{c,t}(Y, \theta) &= R(\delta^*, \theta) - R(r, \theta) \\
&= E_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \sum_{i=1}^p \frac{t_i}{\theta_i} \left( \frac{Y_i}{t_i} \right)^2 + 2\phi(Z)Z \right\} \\
&\quad + \frac{c}{W} E_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \left( \sum_{i=1}^p \frac{Y_i}{t_i} \right)^2 + 2\phi(Z)\gamma \sum_{i=1}^p \frac{Y_i}{t_i} \right\},
\end{aligned} \tag{2.14}$$

where $W = \sum_{i=1}^p \theta_i/t_i$. Let $\gamma_T = \sum_{i=1}^p t_i\theta_i$, so that $Z$ is Poisson with mean $\gamma_T$. Using Lemma A.1.1 the first term in (2.14) equals

$$E_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \frac{Z[p-1+Z]}{\gamma_T} + 2\phi(Z)Z \right\}.$$

Since $\theta_i \geq 0$, $1 \leq i \leq p$ we have that $\sum_{i=1}^p \theta_i^2 \leq \left(\sum_{i=1}^p \theta_i\right)^2 = \gamma^2$, which gives the following inequality for the second term

$$\begin{aligned}
&\frac{c}{W} E_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \left( \frac{Z}{\gamma_T} \sum_{i=1}^p \frac{\theta_i}{t_i} \left(1 - \frac{t_i\theta_i}{\gamma_T}\right) + \gamma^2 \frac{Z^2}{\gamma_T^2} \right) + 2\phi(Z)\gamma^2 \frac{Z}{\gamma_T} \right\} \\
&= cE_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \left( \frac{Z}{\gamma_T} - \frac{1}{W} Z \sum_{i=1}^p \frac{\theta_i^2}{\gamma_T^2} + \frac{1}{W}\gamma^2 \frac{Z^2}{\gamma_T^2} \right) + 2\phi(Z) \frac{1}{W}\gamma^2 \frac{Z}{\gamma_T} \right\} \\
&\leq cE_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \frac{1}{\gamma_T} \left( Z - \frac{1}{W}\frac{\gamma^2}{\gamma_T} Z + \frac{1}{W}\frac{\gamma^2}{\gamma_T} Z^2 \right) + 2\phi(Z) \frac{1}{W}\gamma^2 \frac{Z}{\gamma_T} \right\}.
\end{aligned}$$

Let $\tau = t_{\min}/t_{\max}$ and notice that $W = \sum_{i=1}^p \theta_i/t_i \leq \sum_{i=1}^p \theta_i/t_{\min} = \gamma/t_{\min}$ and similarly $W \geq \gamma/t_{\max}$. Thus

$$\frac{1}{W}\frac{\gamma^2}{\gamma_T} = \frac{1}{\sum_i \theta_i/t_i} \frac{\gamma^2}{\sum_i t_i\theta_i} \geq t_{\min} \frac{\gamma}{\sum_i t_i\theta_i} \geq \frac{t_{\min}}{t_{\max}} = \tau$$

and

$$\frac{1}{W}\frac{\gamma^2}{\gamma_T} = \frac{1}{\sum_i \theta_i/t_i} \frac{\gamma^2}{\sum_i t_i\theta_i} \leq t_{\max} \frac{\gamma}{\sum_i t_i\theta_i} \leq \frac{t_{\max}}{t_{\min}} = \frac{1}{\tau}.$$

This means that for all $\theta \in \Theta$ we have that $\tau \leq \gamma^2(W\gamma_T)^{-1} \leq 1/\tau$. We then have that the second term in (2.14) is less than or equal to

$$cE_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \frac{1}{\gamma_T} \left( Z - \tau Z + \frac{1}{\tau} Z^2 \right) + 2\phi(Z) \frac{1}{\tau} Z \right\}.$$

Putting these two terms together, the expression that bounds the difference in risk is

$$\begin{aligned}
E_\theta D_{c,t}(Y, \theta) \leq E_\theta \Bigg\{ &(\phi^2(Z) - 2\phi(Z)) \frac{Z[p - 1 + c(1-\tau) + (1 + c/\tau)Z]}{\gamma_T} \\
&+ 2\left(1 + \frac{c}{\tau}\right) \phi(Z) Z \Bigg\} = E_\theta D_{c,t}^*(Z, \gamma_T).
\end{aligned}$$

Notice that by the assumption $\phi(z)z \leq \phi(z+1)(z+1)$ combined with Lemma 2.2.1 we have that $E_\theta D_{c,t}(Z, \gamma_T) \leq E_\theta D_{c,t}^*(Z)$ where $D_{c,t}^*(Z)$ is independent of $\gamma_T$. This makes it possible to use the techniques of Section 2.3 to derive explicit estimators. The main finding of this section can now be stated.

**Theorem 2.6.3.** *If $\psi(Z)$ is a non-decreasing function such that $0 \leq \psi(Z) \leq 2[p - 1 + c(1-\tau)]$ then the class of estimators given by*

$$\delta_i^{ct}(Y) = \left(1 - \frac{\psi(Z)}{p - 1 + c(1-\tau) + (1 + c/\tau)Z}\right) \frac{Y_i}{t_i},$$

*is minimax and uniformly dominates the MLE $Y_i/t_i$ under the loss function $L_{c,t}$ as given in* (2.13).

*Proof.* Use the fact that $E_\theta D_{c,t} \leq E_\theta D^*_{c,t}$ , then

$$\begin{aligned}
R(\delta^{ct},\theta) &= p + c + E_\theta D_{c,t}(Y,\theta) \leq p + c + E_{\gamma_T} D^*_{c,t}(Z,\gamma_T)\\
&= p + c - \frac{1}{\gamma_T} E_{\gamma_T}\phi(Z)Z\left\{2\left[p-1+c(1-\tau)+(1+c/\tau)Z+(1+c/\tau)\gamma_T\right]\right.\\
&\qquad \left. -\ \phi(Z)(p-1+c(1-\tau)+(1+c/\tau)Z)\right\}\\
&= p + c - \frac{1}{\gamma_T} E_{\gamma_T}\phi(Z)Z\left\{2\left[p-1+c(1-\tau)+(1+c/\tau)(Z-\gamma_T)\right] - \psi(Z)\right\}\\
&\leq p + c - \frac{1}{\gamma_T} E_{\gamma_T}\phi(Z)Z\left\{(1+c/\tau)(Z-\gamma_T)\right\}\\
&= p + c - 2(1+c/\tau)E_{\gamma_T}\left[\phi(Z)Z^2/\gamma_T - \phi(Z)Z\right]\\
&= p + c - 2(1+c/\tau)E_{\gamma_T}\left[\phi(Z+1)(Z+1) - \phi(Z)Z\right]\\
&< p + c - 2(1+c/\tau)E_{\gamma_T}\left[\phi(Z+1)(Z+1) - \phi(Z+1)(Z+1)\right]\\
&= p + c = R(Y/t,\theta),
\end{aligned}$$

for all $\theta$ because $\phi(Z)Z$ is increasing in $Z$ (cf. Corollary 2.3.3). □

Notice that when $t_1 = \cdots = t_p$ the ratio $\tau = 1$ and the estimators of Theorem 2.6.3 reduce to the natural generalization of $\delta^c$ to the equal exposure case, namely $\delta^c/t$. Lastly, in order to ensure that the class of estimators in Theorem 2.6.3 is minimax we need to establish that the MLE is minimax under the loss function $L_{c,t}$ in (2.13). Let $R_1 + R_c$ denote the two components of the risk of $L_{c,t}$, i.e.

$$E_\theta L_{c,t}(\delta,\theta) = E_\theta L_{1,t}(\delta,\theta) + E_\theta\, c\,\frac{\left(\sum_{i=1}^p \delta_i - \sum_{i=1}^p \theta_i\right)^2}{\sum_{i=1}^p \theta_i/t_i} = R_1(\delta,\theta) + R_c(\delta,\theta).$$

From Lemma 2.6.1 we know that $Y/t$ is minimax under $L_{1,t}$. Moreover, according to Corollary 3.2 in Lehmann (1983, 277) the MLE of $\gamma$ is the unique minimax solution under any loss function of the form $(\delta-\gamma)/\mathrm{Var}_\theta(\delta)$. Since the MLE of $\gamma$ is $\sum_{i=1}^p Y_i/t_i$ and

$$\mathrm{Var}_\theta \sum_{i=1}^p Y_i/t_i = \sum_{i=1}^p \theta_i/t_i = W,$$

we see that the second term in $L_{c,t}$ is on this form. Hence $\sum_{i=1}^p Y_i/t_i$ is the unique minimax solution, and from the risk calculation above we have that $R_c(Y/t,\gamma) = c$. Now, assume that $\delta$ is any estimator and let the set $A \subset \Theta$ be defined by $A = \{\theta \mid R_c(\delta,\theta) \geq c\}$. Then

$$\begin{aligned}\sup_\theta R(\delta,\theta) &\geq \int_\Theta R(\delta,\theta)\,\pi_n(\theta)\,d\theta = \int_\Theta \{R_1(\delta,\theta) + R_c(\delta,\theta)\}\ \pi_n(\theta)\,d\theta \\ &= \int_\Theta R_1(\delta,\theta)\,\pi_n(\theta)\,d\theta + \int_\Theta R_c(\delta,\theta)\,\pi_n(\theta)\,d\theta \\ &\geq \int_\Theta R_1(\delta,\theta)\,\pi_n(\theta)\,d\theta + \int_A c\,\pi_n(\theta)\,d\theta + \int_{\Theta\setminus A} R_c(\delta,\theta)\,\pi_n(\theta)\,d\theta \\ &\geq \int_\Theta R_1(\delta,\theta)\,\pi_n(\theta)\,d\theta + c \geq \int_\Theta R(\delta_n^B,\theta)\,\pi_n(\theta)\,d\theta + c,\end{aligned}$$

where $\delta_n^B = y_i/(b_n+1)$ is the Bayes solution with respect to the $\mathcal{G}(1,b_n)$ prior. From Lemma 2.6.1 we have that the last line above goes to $p+c$ as $n \to \infty$. This shows that

$$\sup_\theta R(\delta,\theta) \geq p + c = \sup_\theta R(Y/t,\theta).$$

Since $\delta$ could be any estimator this means that the MLE is minimax under the loss function $L_{c,t}$ in (2.13). It follows that the class of estimators in Theorem 2.6.3 is also minimax.

## 2.7 More Bayesian analysis

In this section I continue the Bayesian analysis of Section 2.1 and use Bayesian and empirical Bayesian methods to derive estimators in the class $\mathcal{D}_c$. In a first part I draw on the techniques of Ghosh and Parsian (1981) for the $L_1$-setting to derive a class of proper Bayes minimax estimators that uniformly dominate the MLE under $c$-Loss. Second, I derive the explicit estimator $\delta_1^c$ in (2.9) as an empirical Bayes estimator. Finally, I drop the assumption of the Poisson means being independent, and show that $\delta_1^c$ can be derived as a generalized Bayes estimator.

Let the Poisson means be independent with prior distribution $\theta_i \mid b \sim \mathcal{G}(1,b)$ for $i = 1,\dots,p$. Furthermore, let the $b > 0$ parameter in the $\mathcal{G}(1,b)$ have the prior distribution

$$b \sim \pi_2(b) = \frac{1}{B(\alpha,\beta)} b^{\alpha-1}\,(b+1)^{-(\alpha+\beta)}, \tag{2.15}$$

where $\alpha$ and $\beta$ are positive parameters and $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. Given the data and $b$, the posterior distribution of the Poisson means is $\mathcal{G}(y_i+1, b+1)$ for $i = 1,\dots,p$. Using the results of Section 2.1.1 we then have that given $b$, the Bayes solution under the $c$-Loss function with a $\mathcal{G}(1,b)$ prior on $\Theta$ is

$$\delta_j(Y) = \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z}\{E[\theta_i^{-1} \mid Y, b]\}^{-1},$$

where $\{E[\theta_i^{-1} \,|\, Y, b]\}^{-1} = Y_i/(b+1)$. By the tower property of conditional expectation

$$\begin{aligned}\{E[\theta_i^{-1} \,|\, Y]\}^{-1} &= \{E[E[\theta_i^{-1} \,|\, Y, b] \,|\, Y]\}^{-1} \\ &= \{E[Y_i/(b+1) \,|\, Y]\}^{-1} = \{E[(b+1) \,|\, Y]\}^{-1} Y_i.\end{aligned}$$

Moreover, the joint distribution of $Y$ and $b$ is

$$\begin{aligned}f(y,b) &= \prod_{i=1}^{p} \left\{ \int_\Theta \frac{1}{\Gamma(y_i+1)} \theta_i^{y_i}\, e^{-(b+1)\theta_i}\, d\theta_i \right\} \pi_2(b) \\ &= \left\{ b^p \frac{1}{(b+1)^{y_i+1}} \right\} \pi_2(b) = b^p\, (b+1)^{-(z+p)}\, \pi_2(b).\end{aligned}$$

Since this is the joint distribution, Bayes' theorem states that the conditional distribution of $b$ given $Y$ is proportional to $b^p\,(b+1)^{-(z+p)}\,\pi_2(b)$. This shows that $b$ only depends on $Y_i$ through the sum $Z = \sum_{i=1}^p Y_i$ (Ghosh and Parsian, 1981, 283). It follows that $E[(b+1) \,|\, Y] = E[(b+1) \,|\, Z]$, and that the Bayes estimator of $\theta_i$, $1 \leq i \leq p$ is

$$\delta_j(Y) = \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z} \frac{Y_i}{E[b+1 \,|\, Z]}. \tag{2.16}$$

In order to find the conditional expectation in (2.16) it is convenient to first find the normalizing constant $K$ of the conditional distribution $f(b \,|\, y) = K^{-1} b^p\,(b+1)^{-(z+p)}\,\pi_2(b)$.

$$\begin{aligned}K &= \int_0^\infty b^{p+\alpha-1}\,(b+1)^{-(z+p+\alpha+\beta)}\,db = \int_0^\infty \left(\frac{b}{b+1}\right)^{p+\alpha} \frac{1}{b}\,(b+1)^{-(z+\beta)}\,db \\ &= \int_0^1 u^{p+\alpha}\,\frac{1-u}{u}\,(1-u)^{z+\beta-2}\,du = \int_0^1 u^{p+\alpha-1}\,(1-u)^{z+\beta-1}\,du \\ &= B(p+\alpha, z+\beta),\end{aligned}$$

where I have used the change of variable $u = b/(b+1)$. The expectation of $b+1$ given $Z$ is then

$$\begin{aligned}E[b+1 \,|\, Z] &= K^{-1} \int_0^\infty b^{p+\alpha-1}\,(b+1)^{-(z+p+\alpha+\beta-1)}\,db \\ &= K^{-1} \int_0^\infty \left(\frac{b}{b+1}\right)^{p+\alpha-1} (b+1)^{-(z+\beta)}\,db \\ &= K^{-1} \int_0^1 u^{p+\alpha-1}\,(1-u)^{z+\beta-2}\,du = \frac{B(p+\alpha, z+\beta-1)}{B(p+\alpha, z+\beta)} \\ &= \frac{\Gamma(z+\beta-1)}{\Gamma(z+\beta)} \frac{\Gamma(p+\alpha+\beta)}{\Gamma(p+\alpha+\beta-1)} = \frac{p+\alpha+\beta+z-1}{z+\beta-1}.\end{aligned}$$

By this analysis I reach a conclusion that extends the results of Ghosh and Parsian (1981) in a $L_1$-setting, to the $c$-Loss function.

**Proposition 2.7.1.** *Assume that $p > 2 + c$ and consider the family of prior distributions in (2.15) where*

$$0 < \alpha \leq \frac{p-2-c}{1+c}$$

*and $\beta > 0$. Then the Bayes solution under the $c$-Loss function is minimax and uniformly dominates the MLE, hence it is a member of $\mathcal{D}_c$.*

*Proof.* Inserting the expression for $E[b+1 \,|\, Z]$ in (2.16) we obtain

$$\delta_j(Y) = \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z} \frac{z+\beta-1}{p+\alpha+\beta+z-1} Y_i. \tag{2.17}$$

Recall that the estimators in $\mathcal{D}_c$ are of the form $(1-\psi(Z)/(p-1+(1+c)Z))Y_i$ where $\psi$ is non-decreasing and $0 \leq \psi(z) \leq 2(p-1)$ for all $z$ (cf. Corollary 2.3.3). By some algebra we obtain that for the Bayes solution we here consider

$$\begin{aligned}\psi(z) &= p-1+(1+c)z-(1+c)(p-1+z)\frac{z+\beta-1}{p+\alpha+\beta+z-1}\\ &= (1+c)\frac{(p-1+z)(p+\alpha)}{p-1+\alpha+\beta+z} - c(p-1).\end{aligned}$$

This function is non-decreasing for all $z \geq 0$. Moreover, we see that it is bounded above by

$$\sup_{z\geq 0} \psi(z) = (1+c)(p+\alpha) \leq 2(p-1),$$

since $\alpha \leq (p-2-c)/(1+c)$. This means that the class of Bayes solutions in (2.17) where $\alpha$ satisfies the condition of the proposition, is minimax and uniformly dominate the MLE. □

In the following, my focus is shifted from a pure Bayes setting to the (parametric) empirical Bayes approach. In the context of the empirical Bayes approach one uses a family of prior distributions $\pi(\theta_i \,|\, b)$ and estimates $b$ via the marginal distribution of all the data $m(y_1, \ldots, y_p \,|\, b)$ (Carlin and Louis, 2009, 226). Above we saw that the joint distribution of $Y_1, \ldots, Y_p$ and $b$ was $b^p\,(b+1)^{-(z+p)}\,\pi_2(b)$. It follows that with no prior distribution on the parameter $b$, the marginal distribution of all the data is proportional to $b^p\,(b+1)^{-(z+p)}$. It is easily seen that this marginal distribution is a Negative binomial with parameters $p$ and $q = (b+1)^{-1}$ (see Appendix A.1), that is

$$m(z \,|\, b) = \frac{\Gamma(z+p)}{z!\,\Gamma(p)} \left(1 - \frac{1}{b+1}\right)^p \left(\frac{1}{b+1}\right)^z = \frac{\Gamma(z+p)}{z!\,\Gamma(p)} (1-q)^p\, q^z.$$

An alternative to estimating $b$ directly is to estimate $q = (b+1)^{-1}$. This choice is natural since $q$ appears in the Bayes solution in (2.16). The MLE of $q$ is

$z/(z+p)$, which is slightly biased towards underestimating $q$. An unbiased estimator of $q$ is $z/(z+p-1)$,

$$\begin{aligned} E\,\frac{Z}{Z+p-1} &= \sum_{z=0}^{\infty} \frac{z}{z+p-1}\frac{\Gamma(z+p)}{\Gamma(z+1)\Gamma(p)}\,(1-q)^p\,q^z \\ &= q\sum_{z=0}^{\infty} \frac{\Gamma(z+p-1)}{\Gamma(z)\Gamma(p)}\,(1-q)^p\,q^{z-1} = q. \end{aligned}$$

Inserting the estimator $\hat{q} = z/(p-1+z)$ of $q = (b+1)^{-1}$ in the Bayes solution in (2.16) we get

$$\begin{aligned} \delta_i^{eb}(Y) &= \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z}\,\hat{q}\,Y_i = \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z}\frac{Z}{p-1+Z}\,Y_i \\ &= \left(1-\frac{p-1}{p-1+(1+c)Z}\right)Y_i = \delta_1^c(Y). \end{aligned}$$

These calculations show that the estimator $\delta_1^c$ of (2.9) that I derived in Section 2.3, is an empirical Bayes estimator. As a side note, it is interesting that the argument for and the result of estimating $(b+1)^{-1}$ rather than $b$, parallels that in a normal setting. In a normal-normal model, i.e. $X_i \,|\, \xi_i \sim N(\xi, 1)$ and $\xi_i \,|\, \tau^2 \sim N(0, \tau^2)$, the Bayes estimator is $(1 - 1/(\tau^2+1))x_i$. Here the best unbiased estimator of $(\tau^2+1)^{-1}$ is $(p-2)/||x||^2$, which inserted in the Bayes estimator gives the James-Stein estimator (Robert, 2001, 485). Estimating $\tau^2$ directly by the MLE does not yield the optimal James-Stein estimator.

Finally, I show that the estimator $\delta_1^c$ can be derived as a generalized Bayes estimator. Reparametrize the Poisson means $(\theta_1, \ldots, \theta_p) = (\alpha_1\lambda, \ldots, \alpha_p\lambda)$, and assume that

$$(\alpha_1, \ldots, \alpha_p) \sim \frac{\Gamma(\sum_{i=1}^p a_i)}{\prod_{i=1}^p \Gamma(a_i)}\,\alpha_1^{a_1-1}\cdots\alpha_p^{a_p-1},$$

where $\sum_{i=1}^p \alpha_i = 1$ and $a_i > 0$ for all $i$. That is, $(\alpha_1, \ldots, \alpha_p)$ is Dirichlet distributed with parameters $(a_1, \ldots, a_p)$. In the remainder I define $a_0 = \sum_{i=1}^p a_i$. In addition, let the parameter $\lambda$ have the prior distribution $\pi(\lambda)$. From Lemma A.1.1 concerning the relation between the Poisson and the Multinomial distributions, we have that

$$P(Y_1, \ldots, Y_p \,|\, Z) = \frac{P(\{Y_1, \ldots, Y_p\} \cap \{Z\})}{P(Z)} = \frac{P(Y_1, \ldots, Y_p)}{P(Z)},$$

where $P(Y_1, \ldots, Y_p \,|\, Z)$ is shown to be the Multinomial distribution with cell probabilities $\theta_1/\gamma, \ldots, \theta_p/\gamma$. Note that with the parametrization I work with here, namely $\theta_i = \alpha_i\lambda$, the cell probabilities are equal to $\alpha_i$. Moreover, the sum $Z = \sum_{i=1}^p Y_i$ is distributed Poisson with mean $\sum_{i=1}^p \alpha_i\lambda = \lambda$. Using the factorization of the likelihood found above we get that the posterior distribution of $\theta_1,\ldots,\theta_p$ is

$$\begin{aligned}\pi(\theta_1,\ldots,\theta_p\,|\,Y) &\propto P(Y_1,\ldots,Y_p\,|\,Z)P(Z)\,\mathrm{Dirichlet}(a_1\ldots,a_p)\,\pi(\lambda)\\ &= \frac{z!}{y_1!\cdots y_p!}\alpha_1^{a_1+y_1-1}\cdots\,\alpha_p^{a_p+y_p-1}\,\lambda^z e^{-\lambda}\,\pi(\lambda)\\ &\propto \mathrm{Dirichlet}(a_1+y_1,\ldots,a_p+y_p)\,\mathcal{G}(z+1,1)\,\pi(\lambda),\end{aligned}\tag{2.18}$$

which also shows that $(\alpha_1,\ldots,\alpha_p)$ and $\lambda$ are independent. With this parametrization the Bayes solution under the $c$-Loss function is

$$\delta_j^B(Y)=\frac{1+c}{1+cE[\lambda^{-1}\,|\,Y]\sum_{i=1}^p\{E[\theta_i^{-1}\,|\,Y]\}^{-1}}\,\{E[\theta_j^{-1}\,|\,Y]\}^{-1}.\tag{2.19}$$

With respect to the posterior distribution in (2.18), the expectation $E[\theta_j^{-1}\,|\,Y]$ in this expression is given by

$$\begin{aligned}E[\theta_j^{-1}\,|\,Y] &= \int_0^\infty\int_S \frac{1}{\alpha_j\lambda}\,\pi(\theta_1,\ldots,\theta_p\,|\,Y)\,d\alpha\,d\lambda\\ &= \int_0^\infty\int_S\frac{1}{\alpha_j\lambda}\left\{\frac{\Gamma(a_0+z)}{\prod_{i=1}^p\Gamma(a_i+y_i)}\prod_{i=1}^p\alpha_i^{a_i+y_i-1}\right\}\mathcal{G}(z+1,1)\pi(\lambda)\,d\alpha\,d\lambda\\ &= \int_0^1\mathcal{G}(z+1,1)\pi(\lambda)\,d\lambda\int_S\alpha_j\frac{\Gamma(a_0+z)}{\prod_{i=1}^p\Gamma(a_i+y_i)}\prod_{i=1}^p\alpha_i^{a_i+y_i-1}\,d\alpha.\end{aligned}$$

Here the expectation of $\alpha_j$ over the simplex $S$ can be computed explicitly,

$$\begin{aligned}E[\alpha_j\,|\,Y] &= \int_S\frac{\Gamma(a_0+z)}{\prod_{i=1}^p\Gamma(a_i+y_i)}\alpha_j^{a_j+y_j-2}\prod_{i\neq j}\alpha_i^{a_i+y_i-1}\,d\alpha\\ &= \frac{\Gamma(a_j+y_j-1)}{\Gamma(a_j+y_j)}\int_S\frac{\Gamma(a_0+z)}{\Gamma(a_j+y_j-1)\prod_{i\neq j}\Gamma(a_i+y_i)}\alpha_j^{a_j+y_j-2}\prod_{i\neq j}\alpha_i^{a_i+y_i-1}\,d\alpha\\ &= \frac{\Gamma(a_j+y_j-1)\Gamma(a_0+z)}{\Gamma(a_j+y_j)\Gamma(a_0+z-1)}\\ &\qquad\times\int_S\frac{\Gamma(a_0+z-1)}{\Gamma(a_j+y_j-1)\prod_{i\neq j}\Gamma(a_i+y_i)}\alpha_j^{a_j+y_j-2}\prod_{i\neq j}\alpha_i^{a_i+y_i-1}\,d\alpha\\ &= \frac{\Gamma(a_j+y_j-1)\Gamma(a_0+z)}{\Gamma(a_j+y_j)\Gamma(a_0+z-1)}=\frac{a_0+z-1}{a_j+y_j-1}.\end{aligned}$$

Inserting this in the posterior expectation of $\theta_j^{-1}$ gives

$$E[\theta_j^{-1}\,|\,Y]=\frac{a_0+z-1}{a_j+y_j-1}\int_0^\infty\frac{1}{\lambda}\,\mathcal{G}(z+1,1)\pi(\lambda)\,d\lambda,$$

for $j = 1, \dots, p$. Let $\lambda$ have the non-informative prior distribution that is uniform on the positive real line, $\pi(\lambda) \propto I(\lambda > 0)$ where $I(\cdot)$ is the indicator function. Despite this being an improper prior, the posterior distribution of $\lambda$ given $Z$ is not improper. We have that

$$E[\lambda^{-1} \,|\, Z] = \int_0^\infty \frac{1}{\lambda}\, \mathcal{G}(z+1,1) I(\lambda > 0)\, d\lambda = \frac{1}{z},$$

which gives

$$E[\theta_j^{-1} \,|\, Y] = \frac{a_0 + z - 1}{a_j + y_j - 1}\,\frac{1}{z}.$$

In addition, the sum in (2.19) equals

$$\sum_{i=1}^{p} \{E[\theta_i^{-1} \,|\, Y]\}^{-1} = z \sum_{i=1}^{p} \frac{a_j + y_j - 1}{a_0 + z - 1} = \frac{(a_0 + z - p)z}{a_0 + z - 1}.$$

Now, let $\alpha_1, \dots, \alpha_p$ be uniformly distributed over the simplex $S$. This is achieved by setting $a_1 = \cdots = a_p = 1$. Then the sum $a_0 = p$. In summary, with $\lambda$ uniform over $\mathbb{R}_+$ and the $(\alpha_1, \dots, \alpha_p)$ uniform on the simplex $S = [0,1]^p$, the Bayes solution under the $c$-Loss function equals

$$\delta_j^B(Y) = \frac{1+c}{1 + c\frac{1}{Z}\frac{Z^2}{p-1+Z}} \frac{Y_j}{p-1+Z} = \left(1 - \frac{p-1}{p-1+(1+c)Z}\right) Y_j = \delta_1^c(Y).$$

This means that in addition to being an empirical Bayes estimator, the new estimator $\delta_1^c$ is also a generalized Bayes estimator.

## 2.8 Remark I: Estimation of Multinomial probabilities

In the preceeding sections I have several times used the result that $(Y_1, \dots, Y_p)$ given $Z = \sum_{i=1}^p Y_i$ is Multinomial with cell probabilities $\theta_1/\gamma, \dots, \theta_p/\gamma$. We write

$$(Y_1, \dots, Y_p) \,|\, Z \sim \text{Multi}(Z, \theta_1/\gamma, \dots, \theta_p/\gamma).$$

The estimator $\delta_1^c$ yields an estimator of $\lambda_i = \theta_i/\gamma$, $1 \le i \le p$, namely

$$\delta_1^c/Z = \left(1 - \frac{p-1}{p-1+(1+c)Z}\right) \frac{Y_i}{Z}.$$

Assume that the loss in estimating $\lambda_i$ is the natural counterpart of the $c$-Loss function, $\sum_{i=1}^p \lambda_i^{-1}(\delta_i - \lambda_i)^2 + c\left(\sum_{i=1}^p \delta_i - 1\right)^2$, since $\sum_{i=1}^p \lambda_i = 1$. Moreover, assume that $n$ is the known number of independent trials and $Y_i$ equals the number of outcomes of type $i$, so $\sum_{i=1}^{p} Y_i = n$. We are then in a Multinomial situation with $p$ categories. From the heuristic argument above the new estimator should be

$$\delta^{cM} = \left(1 - \frac{p-1}{p-1+(1+c)n}\right)\frac{Y}{n}.$$

Using the function $\phi(n) = (p-1)/(p-1+(1+c)n)$ the difference in risk between $\delta^{cM}$ and the usual estimator of Multinomial cell probabilities is

$$\begin{aligned} D &= (\phi^2(n) - 2\phi(n))\sum_{i=1}^{p}\frac{1}{n^2\lambda_i}E_\lambda[Y_i^2] + 2\phi(n) + c\phi^2(n) \\ &= (\phi^2(n) - 2\phi(n))\frac{p-1+n}{n} + 2\phi(n) + c\phi^2(n) \\ &= \phi^2(n)\frac{p-1+(1+c)n}{n} - 2\phi(n)\frac{p-1}{n} = \frac{(p-1)^2 - 2(p-1)^2}{\{p-1+(1+c)n\}n} \leq 0 \end{aligned}$$

for all $p \geq 2$. Interestingly, this constant risk estimator dominates the MLE $Y_i/n$, $1 \leq i \leq p$ even though it underestimates the total probability, that is

$$\delta_1^{cM} + \cdots + \delta_p^{cM} < 1.$$

If we set $c = 0$ in $\delta^{cM}$ we get an estimator that uniformly dominates the MLE under $L_1$ (Clevenson and Zidek, 1975, 703). Under the squared error loss function there exists no estimator that uniformly improves on the MLE (Fienberg and Holland, 1973, 684).

## 2.9 Remark II: A squared error $c$-Loss function

In the case of the squared error loss function $L_0$, the analogous extension of the $c$-Loss function is

$$L_{c^*}(\delta,\theta) = \sum_{i=1}^{p}(\delta_i - \theta_i)^2 + c^*\left(\sum_{i=1}^{p}\delta_i - \sum_{i=1}^{p}\theta_i\right)^2. \tag{2.20}$$

This loss function can be expressed as the quadratic form

$$L_{c^*}(\delta,\theta) = (\delta-\theta)^t A(\delta-\theta),$$

where $\theta = (\theta_1,\ldots,\theta_p)^t$ and $\delta = (\delta_1,\ldots,\delta_p)^t$ are $p \times 1$ vectors of parameters and estimators respectively, while $A$ is a $p \times p$ matrix of the form

$$A = \begin{bmatrix} 1+c^* & \cdots & c^* \\ \vdots & \ddots & \vdots \\ c^* & \cdots & 1+c^* \end{bmatrix}.$$

The matrix $A$ is symmetric $A = A^t$, from which it follows that $A$ is orthogonally diagonalizable. That is, there exists an orthogonal matrix $Q$ (which can be normalized and made into an orthonormal matrix) and a diagonal matrix $\Lambda$ such that $A = Q\Lambda Q^t$ (see e.g. Theorem 2 in Lay (2012, 396)). The rows of $Q$ are the orthonormal eigenvectors $u_i$, $1 \leq i \leq p$ of $A$ and $\Lambda$ has the eigenvalues of $A$ on its diagonal and zero elsewhere. From this decomposition of $A$ we get the following well known inequality that I will use below

$$\begin{aligned} x^t A x &= x^t Q\Lambda Q^t x = (Q^t x)^t \Lambda (Q^t x) \\ &= \sum_i \lambda_i (u_i^t x_i)^2 \leq \lambda_{\max} \sum_i (u_i^t x_i)^2 = \lambda_{\max} ||x||^2, \end{aligned} \tag{2.21}$$

where $\lambda_i$ are the eigenvalues of $A$ and $\lambda_{\max} = \max_{1 \leq i \leq p} \lambda_i$. We will consider estimators of the form

$$\delta^*(Y) = Y + f(Y),$$

where $f_i(Y) = (f_1(Y), \ldots, f_p(Y))$ satisfies the conditions in Lemma 2.2.1. The difference in risk between $Y + f(Y)$ and the MLE under $L_0$ is

$$\begin{aligned} &E_\theta \left[ L_0(\delta^*, \theta) - L_0(Y, \theta) \right] \\ &= E_\theta \sum_{i=1}^p \left\{ f_i^2(Y) + f_i(Y) Y_i - f_i(Y) \theta_i \right\} \\ &= E_\theta \sum_{i=1}^p \left\{ f_i^2(Y) + Y_i \left( f_i(Y) - f_i(Y - e_i) \right) \right\} = E_\theta D(f(Y)), \end{aligned}$$

where $e_i$ is the $p \times 1$ vector with i'th component equal to one a zero elsewhere. The function $f$ that ensures that $D(Y) \leq 0$ for all $Y$ is the function of Peng (1975). This function was defined in (1.3) as $f_i(Y) = (N_0(Y) - 2)^+ h_i(Y_i)/D(Y)$. The difference in risk between $\delta^*(Y) = Y + f(Y)$ and $\delta^o(Y) = Y$ under the loss function $L_{c^*}$ in (2.20) is then

$$\begin{aligned} E_\theta D_{c^*}(f(Y)) &= E_\theta \left[ (Y + f(Y) - \theta)^t A (Y + f(Y) - \theta) - (Y - \theta)^t A (Y - \theta) \right] \\ &\leq \lambda_{\max} E_\theta \left[ ||Y + f(Y) - \theta||^2 - ||Y - \theta||^2 \right] \\ &= \lambda_{\max} E_\theta D(f(Y)), \end{aligned}$$

where I have used the inequality in (2.21). The function $f$ that ensures that $D_{c^*}(f(y)) \leq 0$ for all $y$ must then be that of Peng (1975) divided by the largest eigenvalue $\lambda_{\max}$, that is $(N(Y) - 2)^+ h_i(Y_i)/(\lambda_{\max} D(Y))$.

Notice that we might write $A$ as $I_p + c^* 1_p 1_p^t$ where $1_p$ is the $p \times 1$ vector with only ones. If $v = (v_1, \ldots, v_p)^t$ is an eigenvector of the matrix $A$ then

$$A\,v - \lambda\,v = (I_p + c^* 1_p 1_p^t) v - \lambda\,v = 0,$$

which on matrix form is

$$\begin{bmatrix}(1-\lambda)v_1\\ \vdots \\ (1-\lambda)v_p\end{bmatrix} + c^* \begin{bmatrix}v_1+v_2\cdots+v_p\\ \vdots \\ v_1+v_2\cdots+v_p\end{bmatrix} = 0.$$

From this we see that either the eigenvectors are such that $v_1 + \cdots + v_p = 0$ in which case the eigenvalue must be one. The other possibility is that $v_1 = v_2 = \cdots = v_{p-1} = v_p$ in which case the eigenvalue is $1 + p\,c^*$. This shows that the largest eigenvalue of $A$ is $1 + pc^*$. In conclusion, the estimator

$$\delta_i^{P^*}(Y) = Y_i - \frac{(N(Y)-2)^+}{(1+pc^*)D(Y)} h_i(Y_i),$$

uniformly dominates the MLE when loss is given by $L_{c^*}$ in (2.20).

# 3

# Shrinking towards a non-zero point

Except for the Bayes estimators in Section 2.1 all the estimators considered so far shrink the observations towards zero, the boundary of the parameter space. Consequently, substantial savings in risk are only obtained when the $\theta_i$, $1 \leq i \leq p$ are all close to zero. As mentioned in the introduction, an important question is whether there are estimators that improve on the MLE that shrink the observations towards some other point than zero. This question has been answered by the affirmative by Tsui (1981), Hudson and Tsui (1981) and Ghosh et al. (1983). The impetus for developing such estimators is that they should give larger savings in risk when the $\theta_i$ are large, and at the same time maintaining risk dominance relative to the MLE. In Section 3.1 and 3.2 I present some of these estimators (with respect to $L_0$ and $L_1$ respectively) and compare them to estimators where the requirement of uniform dominance is relaxed.

## 3.1 Squared error loss

Tsui (1981), Hudson and Tsui (1981) and Ghosh et al. (1983) derive estimators that shrink the observations towards a pre-specified point or some point generated by the data, while maintaining uniform risk dominance compared to the MLE. These estimators build on the ideas of Peng (1975) and can be viewed as extensions of the estimator $\delta^P$ in (1.3). The proofs of risk dominance of these estimators are similar to that of $\delta^P$ albeit slightly more involved (see e.g. Ghosh et al. (1983)). In Appendix A.6 I prove the risk dominance of Peng's estimator.

The first estimator I present is due to Ghosh et al. (1983). It shrinks the MLE towards a prior guess $\nu = (\nu_1, \dots, \nu_p)$ in $\Theta$. Let $N(Y) = \#\{i : Y_i > \nu_i)$ count the number of $Y_i$ bigger than $\nu_i$, $h_j = \sum_{k=1}^{j} k^{-1}$, and $D(Y) = \sum_{j=1}^{p} d_i(Y_i)$

where

$$d_i(Y_i) = \begin{cases} \{h(Y_i) - h(\nu_i)\}^2 + \frac{1}{2}\{3h(\nu_i) - 2\}^+ & Y_i < \nu_i \\ \{h(Y_i) - h(\nu_i)\}\{h(Y_i + 1) - h(\nu_i)\} & Y_i \geq \nu_i. \end{cases}$$

Then the estimator $\delta^{G1}$ whose i'th component is given by

$$\delta_i^{G1}(Y) = Y_i - \frac{(N(Y) - 2)^+}{D(Y)}(h_i(Y_i) - h(\nu_i)), \tag{3.1}$$

shrinks $Y_i$ towards the prior guess $\nu_i$ of $\theta_i$ for each $i = 1, \ldots, p$. $\delta_i^{G1}$ dominates the MLE under squared error loss when $p \geq 3$.

Now, define $N(Y) = \#\{i : Y_i > Y_{(1)})$ and $H_i(Y) = h(Y_i) - h(Y_{(1)})$ with $h$ as above and $Y_{(1)} = \min_{1 \leq i \leq p} Y_i$. Let $D(Y) = \sum_{i=1}^p H_i(Y)H(Y + e_i)$. Then the estimator given

$$\delta_i^{G2}(Y) = Y_i - \frac{(N(Y) - 2)^+}{D(Y)} H_i(Y) \tag{3.2}$$

dominates $\delta^o$ under $L_0$ when $p \geq 4$. Finally, the estimator

$$\delta_i^{G3}(Y) = Y_i - \frac{(N(Y) - 2)^+}{D(Y)}\{h(Y_i) - h(\text{median}(Y)\}, \tag{3.3}$$

where the functions are as in $\delta^{G1}$ with the $\nu_i$ replaced by median$(Y)$, dominates the MLE when $p \geq 6$. The estimators $\delta^{G1}$, $\delta^{G2}$ and $\delta^{G3}$ all dominate the MLE under $L_0$ (given that $p$ is sufficiently large), but as we will see below (cf. Figure 3.1), the reductions in risk of these estimators are not particularly impressive. These rather modest reductions in risk makes it tempting to instead consider estimators that yield substantial savings in risk in plausible regions of the parameter space, but that fail to be uniformly dominating. As we saw in the introduction, Berger (1983, 368) proposed that the aim of uniform risk domination might be too strict. In the same vein, Morris (1983a) thinks that "statisticians need simple rules that shift towards a good center (near the mean of the data)". Following this advice, I therefore develop an estimator that shrinks the observations towards the mean of the data, while at the same time insuring against overly bad performance in situations where the $\theta_i$ have little or nothing in common.

To gain insight into the construction of this estimator, consider the Bayes estimator of $\theta_i$, $1 \leq i \leq p$ with a Gamma prior with mean $\mu = a/b$ and variance $a/b^2$,

$$\delta_i^{\mathrm{B}} = \mu\frac{b}{b+1} + \left(1 - \frac{b}{b+1}\right) y_i. \tag{3.4}$$

Recall that what we are competing against is the risk of the MLE, which is $R(Y, \theta) = p\bar{\theta}$. Defining $w = b/(b+1)$ the risk of the Bayes estimator can be written

$$R(\delta^{\mathrm{B}}, \theta) = p\bar{\theta}(1 - w)^2 + w^2 \sum_{i=1}^{p} (\mu - \theta_i)^2. \tag{3.5}$$

The first term in this expression is always less than $p\bar{\theta}$, and as $w$ increases, expressing more confidence in the prior, the term diminishes. In addition, the closer the prior mean $\mu$ is to $\theta_i$, $1 \leq i \leq p$ the smaller is the second term. From this second term it is also clear that there is little to gain from using the Bayes estimator if we *a priori* believe that the Poisson means are very heterogenous, simply because one guess $\mu$ at $p$ different $\theta_i$ is doomed to be unsatisfactory. I desire to find an estimator that acts like (3.4) (shrinks to some common value), but that does not rely on prior information and does not lead to disastrous results if the $\theta_i$ are too spread out. The natural place to start is by replacing $\mu$ in the Bayes estimator by the empirical mean. Consider the estimator where the $i$'th component takes the form

$$\delta_i = B\bar{y} + (1 - B)y_i. \tag{3.6}$$

The task is then to find a weight function $B$ that is optimal in terms of risk. Using that $E_\theta Y_i \bar{Y} = p^{-1} E_\theta[E[Y_i Z \mid Z]] = p^{-1}\theta_i(1+\gamma)$, then

$$E_\theta(Y_i - \bar{Y})(Y_i - \theta_i) = \theta_i - p^{-1}\theta_i$$

and

$$E_\theta(Y_i - \bar{Y})^2 = \theta_i + \theta_i^2 - 2p^{-1}\theta_i(1+\gamma) + p^{-2}(\gamma + \gamma^2).$$

Under squared error loss the risk of this estimator is

$$\begin{aligned}
R(\delta, \theta) &= E_\theta \sum_{i=1}^{p} ((Y_i - \theta_i) - B(Y_i - \bar{Y}))^2 \\
&= \sum_{i=1}^{p} E_\theta \left\{ (Y_i - \theta_i)^2 - 2B(Y_i - \bar{Y})(Y_i - \theta_i) + B^2(Y_i - \bar{Y})^2 \right\} \\
&= \gamma - 2B \sum_{i=1}^{p} E_\theta(Y_i - \bar{Y})(Y_i - \theta_i) + B^2 \sum_{i=1}^{p} E_\theta(Y_i - \bar{Y})^2 \\
&= p\bar{\theta} - 2B(p-1)\bar{\theta} + B^2(p-1)\bar{\theta} + B^2(p-1)S_\theta^2,
\end{aligned}$$

where $S_\theta^2 = (p-1)^{-1}\sum_{i=1}^p(\theta_i - \bar{\theta})^2$ and $\bar{\theta} = p^{-1}\sum_{i=1}^p \theta_i$. In terms of minimizing the risk the optimal $B$ is then

$$B = \frac{\bar{\theta}}{S_\theta^2 + \bar{\theta}}. \tag{3.7}$$

In this expression for $B$ we have the unknown quantities $\bar{\theta}$ and $S_\theta^2$. An unbiased estimator of $\bar{\theta}$ is $\bar{Y}$. More care must be taken when estimating $S_\theta^2$. Define $S_{yy}^2 = (p-1)^{-1}\sum_i (y_i - \bar{y})^2$, then $E_\theta[S_{yy}^2 - \bar{Y}] = S_\theta^2$, which suggests estimating the unknown variance of the parameters by $(S_{yy}^2 - \bar{y})^+ = \max\{S_{yy}^2 - \bar{y}, 0\}$ and replacing the unknown variance in (3.7) by this unbiased estimate. This yields

$\hat{B} = \bar{y}/((S_{yy}^2 - \bar{y})^+ + \bar{y})$ as an estimator of the optimal $B$. Thereby, we obtain the estimator

$$\delta_i^m(Y) = \frac{\bar{y}}{(S_{yy}^2 - \bar{y})^+ + \bar{y}}\,\bar{y} + \left(1 - \frac{\bar{y}}{(S_{yy}^2 - \bar{y})^+ + \bar{y}}\right) y_i. \tag{3.8}$$

This estimator has some interesting properties. First, it is almost equal to the empirical Bayes estimator. Exploiting the Negative binomial marginal distribution of the data the parameters $a$ and $b$ can be estimated by the method of moments. The method of moments estimators of $a$ and $b$ are $b\bar{y}$ and $\bar{y}/((p-1)/pS_{yy}^2 - \bar{y})$ respectively. Since $b$ cannot be negative we consider the positive part version of this estimator and set $\hat{b}_{\text{mom}} = \bar{y}/((p-1)/pS_{yy}^2 - \bar{y})^+$ Consequently, the empirical Bayes estimator is of the form (3.6) where $B$ is estimated by

$$\hat{B}_{\text{eb}} = \frac{\bar{y}}{(\frac{p-1}{p}S_{yy}^2 - \bar{y})^+ + \bar{y}}\,.$$

Thus, we see that the two estimators of $B$ are indistinguishable for sufficiently large $p$. Second, $\delta^m$ seems to be closely related to the empirical Bayes estimator in a normal-normal model (see e.g. Morris (1983b)). Consider $X_i \mid \xi_i \sim N(\xi_i, \sigma^2)$ independent with $\sigma$ known and $\xi_i \sim N(\mu, \tau^2)$ independent. Then the posterior distribution $\xi_i \mid x$ is $N(\xi^*, \sigma^2(1-A))$ where $A = \sigma^2/(\sigma^2 + \tau^2)$ and

$$\xi^* = A\mu + (1-A)x.$$

Via the marginal distribution of the data we find the ML-estimates of $(\mu, \tau)$, which are $\hat{\mu} = \bar{x} = p^{-1}\sum_{i=1}^p x_i$ and $\hat{\tau}^2 = (s^2 - \sigma^2)^+ = \max\{0, s^2 - \sigma^2\}$. Then the empirical Bayes estimator in the normal-normal model with $\sigma$ known is

$$\delta_i^{\text{eb}} = \frac{\sigma^2}{\sigma^2 + (S_{xx}^2 - \sigma^2)^+}\bar{x} + \left(1 - \frac{\sigma^2}{\sigma^2 + (S_{xx}^2 - \sigma^2)^+}\right) x_i.$$

As with the simultaneous estimator of Poisson means $\delta^m$ in (3.8), the target of shrinkage for $\delta_i^{eb}$ is the mean of the observations. The amount of shrinkage is a function of how much the empirical variance $S_{xx}^2$ exceeds $\sigma^2$ (Carlin and Louis, 2009, 228). If $S_{xx}^2 \leq \sigma^2$ the estimator $\delta_i^{\text{eb}}$ is equal to the MLE of $\xi$ in a model where the observations are independent and identically distributed from a normal distribution with mean $\xi$. The simultaneous estimator of Poisson means $\delta^m$ shares this property, because if $Y_1, \ldots, Y_p$ are independent and identically distributed Poisson with mean $\theta$, then $E_\theta[S_{yy}^2 - \bar{y}] = 0$ and the estimator $\delta^m$ is equal to the MLE in the one-dimensional case $p = 1$. Therefore, the term $(S_{yy}^2 - \bar{y})^+$ can be viewed as a measure of the amount of heterogeneity in the data, that is how far the Poisson means $\theta_1, \ldots, \theta_p$ are from being equal. As the heterogeneity of the data increases the shrinkage towards the mean decreases and the estimates are pushed towards the observations.

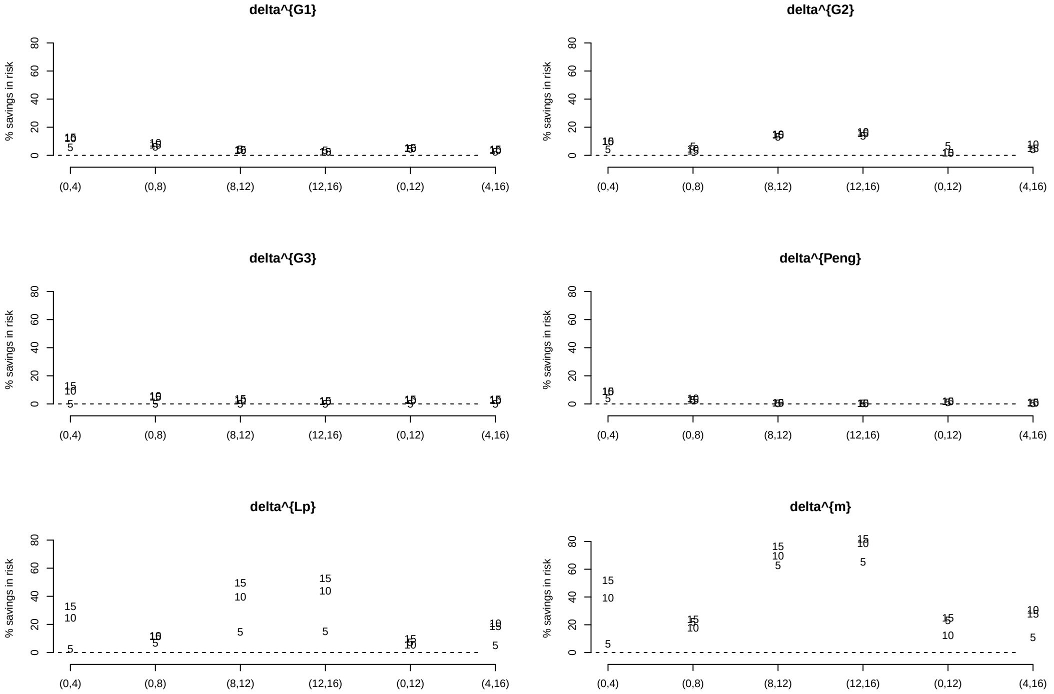

Figure 3.1: Savings in risk $(R(Y,\theta) - R(\delta,\theta))/R(Y,\theta) \times 100$ under $L_0$ for the estimators presented in Section 3.1. The savings in risk are simulated for six different intervals with values of $p$ equal to 5, 10 and 15. Estimates based on 10000 simulations.

Hudson (1985) and Ghosh et al. (1983) develop an estimator that instead of shrinking towards the mean of the data, shrinks towards the geometric mean of the data. They observe that the function $h(Y_i) = \sum_{k=1}^{Y_i} k^{-1}$ used in (1.3) is close to $\log Y_i$ when $Y_i$ is sufficiently large, and that the log transform of Poisson data is often seen as approximately normally distributed (Ghosh et al., 1983, 355). They therefore propose a Lindley type estimator (cf. Equation (1.5)) for the Poisson case

$$\delta_i^{Lp}(Y) = Y_i - \frac{(N(Y)-2)^+}{\sum_{j=1}^{p}(h(Y_j) - \bar{h})^2}(h(Y_i) - \bar{h}),$$

where $N(Y) = \#\{i : h(X_i) > \bar{h}\}$ and $\bar{h} = p^{-1}\sum_{i=1}^{p} h(Y_i)$. Since $h(Y_i) \approx \log Y_i$, we have that $\bar{h} \approx \log\left(\prod_{i=1}^{p} Y_i\right)^{1/p}$, which is the geometric mean of the observations.

Figure 3.1 provides a graphical summary of the simulation results for the six estimators presented in this section for three different values of $p$ and six different intervals for the unknown parameters.

The estimators $\delta^{Lp}$ and $\delta^{m}$ perform better than the four other estimators for all the intervals. Not surprisingly, when $\theta_i \in (8, 12)$ and $\theta_i \in (12, 16)$ these estimators outperform the three others, with savings in risk of about 50%. Interestingly, it seems that the new estimator $\delta^m$ is much less sensible to the sample size $p$ than the Lindley type estimator $\delta^{Lp}$. For the intervals $(8, 12)$ and $(12, 16)$ we see that the two estimators perform about equally well for $p = 10$ and $p = 15$, but that for $p = 5$ the performance of $\delta^m$ is superior to that of $\delta^{Lp}$. To gain insight into this phenomenon, I have compared the variance of the two estimators when $p = 5$. In the box plots in Figure 3.2 we see that the means of the estimates of the Lindley type estimator is closer to the true $\theta_1, \dots, \theta_5$, but that $\delta^{Lp}$ shows somewhat higher variability than $\delta^m$. As a consequence, the estimated savings in risk from using $\delta^{Lp}$ compared to the MLE is 14.74%, while that of $\delta^m$ is much higher at 62.68%. The findings of these simulations give credence to the statement of Berger (1983) concerning uniform risk dominance being a perhaps too restrictive criterion. Since the non-dominating estimators $\delta^{Lp}$ and $\delta^m$ perform better or equally well for all the intervals, these simulations indicate that the $\theta_i$ must have very little in common for it to be dangerous to use non-dominating estimators. Moreover, as we see from Figure 3.1, the potential gains from using non-dominating estimators are huge.

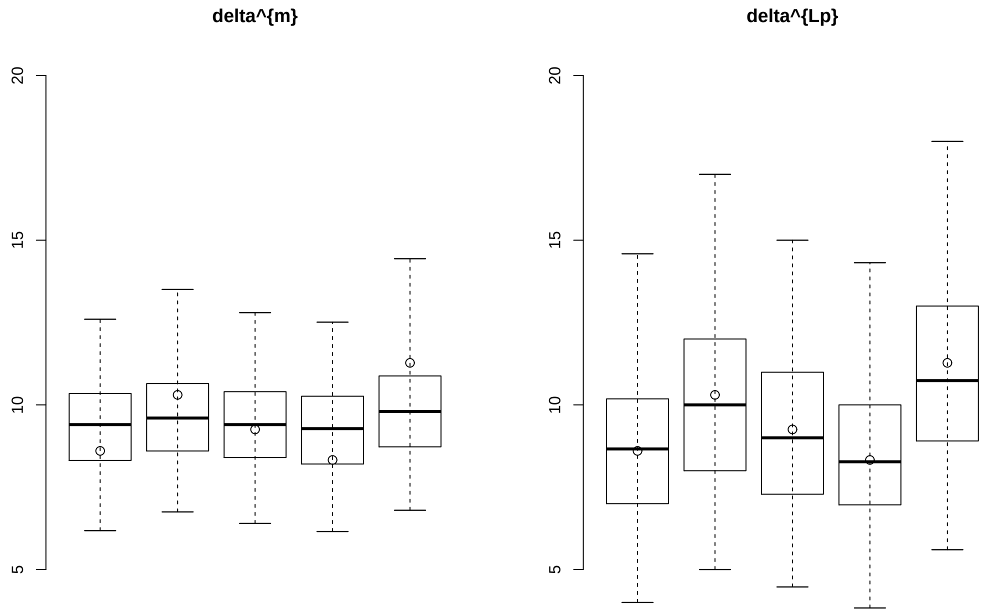

Figure 3.2: True values of the parameters are $\theta \in (8, 12)$ with $p = 5$. The box plots indicate the variance of $\delta^{Lp}$ and $\delta^m$ based on $10^4$ simulations. Estimates based on 10000 simulations (the same simulations as in Figure 3.1).

## 3.2 Weighted squared error loss

Estimators that shrink the MLE towards some non-zero point in the parameter space under the weighted squared error loss function $L_1$ are not as intensively studied as their $L_0$-counterparts presented in Section 3.1. One reason for this is that due to the attention paid to small values of $\theta_i$ under $L_1$, such estimators are harder to find than under $L_0$. Ghosh et al. (1983, 357) propose an estimator that shrinks the observations towards $Y_{(1)} = \min_{1\leq i\leq p} Y_i$ instead of zero. This shift in point of attraction from zero to a possibly non-zero point is not detrimental to the risk function optimality. Their estimator is defined by

$$\delta_i^{Gm}(Y) = Y_i - \frac{(N(Y)-1)^+}{\sum_{j=1}^p (Y_j - Y_{(1)})}(Y_i - Y_{(1)}),$$

where $N(Y)$ counts the number of observations strictly bigger than $Y_{(1)}$. As seen from the simulations in Table 3.1 much can be gained from shrinking towards $Y_{(1)} \geq 0$ instead of automatically shrinking towards zero when the Poisson means are of a certain size. With the possibility of small or zero counts, however, the gains from using $\delta^{Gm}$ compared to $\delta^{CZ}$ are negligible.

Under $L_1$ it seems to be difficult to find improved estimators that shrink down towards *and* up towards a common point $\nu$, i.e. estimators $\delta$ that have the property that if $Y_i \leq \nu$ then $\delta_i \geq Y_i$ and if $Y_i \geq \nu$ then $\delta_i \leq Y_i$. Even the Bayes estimator under $L_1$ takes this into account. Consider $\theta_i$, $1 \leq i \leq p$ independent $\mathcal{G}(a,b)$. Then the Bayes estimator under $L_1$ is

$$E[\theta_i^{-1}\,|\,\text{data}] = \frac{a-1}{b}\,\frac{b}{b+1} + \left(1 - \frac{b}{b+1}\right) y_i. \tag{3.9}$$

This estimator shrinks towards $(a-1)/b$ instead of the mean $a/b$ of the prior, reflecting the great deference paid to small counts under the weighted squared error loss function. The form of (3.9) proposes a slight modification of $\delta^m$ for the weighted squared error loss function. Since the weight $b/(b+1)$ is in a sense estimated by $\hat{B}$ as given in (3.8), $(\hat{B}-1)/\hat{B}$ is an estimator of $1/b$. This gives a shrink-to-mean estimator suitably adjusted to the $L_1$ loss function, namely

$$\delta^{m1}(Y) = \begin{cases} \hat{B}\,\bar{y} + (1-\hat{B})\,(y_i - 1), & \text{if } y_i \geq 1 \\ 0 & \text{if } y_i = 0\,. \end{cases} \tag{3.10}$$

As with $\delta^m$, this estimator shrinks towards the mean of the observations, but the observations have to be somewhat bigger in order to be pulled upwards to the empirical mean. In other words, fewer observations are pulled upwards.

The estimator I now develop acknowledges the difficulty of pulling the MLE *up* to a common point, and therefore only shrinks observations above a certain point. Such a scheme is relevant in situations where there exist hypotheses about

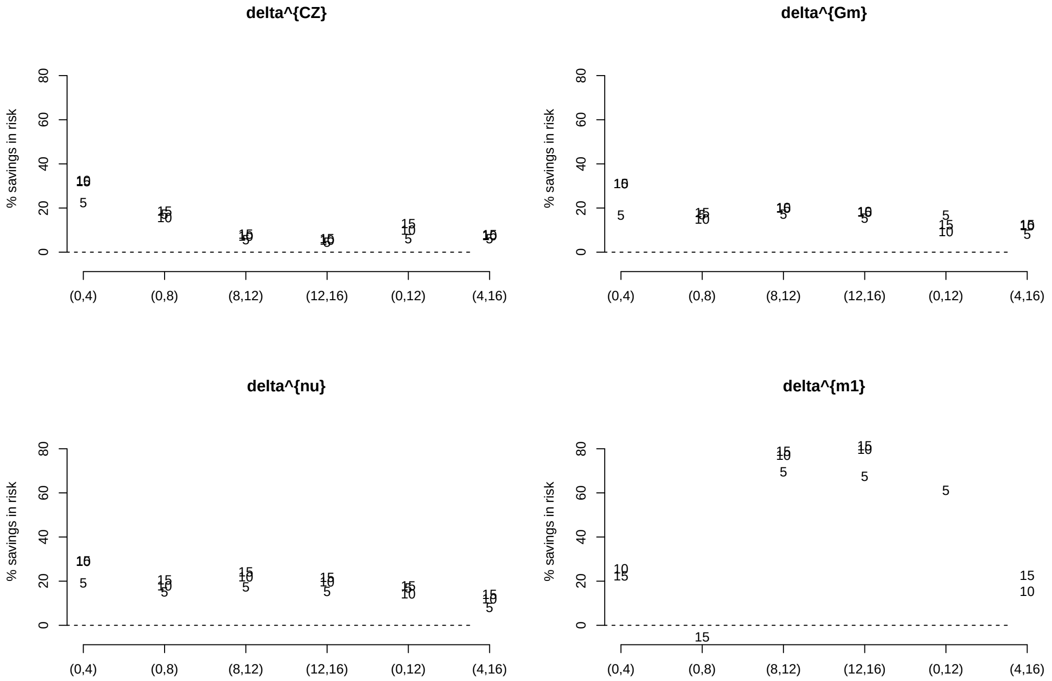

Figure 3.3: Savings in risk $(R(Y,\theta) - R(\delta,\theta))/R(Y,\theta) \times 100$ (under $L_1$) for the four estimators presented in Section 3.2. The simulation procedure is as described in Section 2.4, with estimates based on 200000 simulations (2000 simulation of $y$ times 100 simulations of $\theta$).

subsets of the Poisson parameters, where this subset is hypothesized to consist of parameters that are larger than the remaining parameteres. In effect, the new estimator shrinks a subset of the observations towards some predetermined point $\nu$, while the remaining are estimated by the MLE.

**Proposition 3.2.1.** *Define the set $Y_\nu = \{Y_i : Y_i \geq \nu\}$ with $\nu$ a non-negative integer, $p_\nu = \sum_{i=1}^{p} I(Y_i \in Y_\nu)$ where $I(\cdot)$ is the indicator function, and let $Z_\nu = \sum_{i=1}^{p} Y_i\, I(Y_i \in Y_\nu)$. The estimator $\delta_i^\nu = Y_i + f_i(Y)$ where $f$ satisfies the conditions of Lemma 2.2.1 is defined by*

$$f_i(Y) = -I(Y_i > \nu)\frac{\psi(Z)(Y_i - \nu)}{p_\nu - 1 + Z_\nu - p_\nu \nu}, \tag{3.11}$$

*where $\psi$ is a non-decreasing function such that $0 \leq \psi(z) \leq 2(p_\nu - 1)$ for all $z \geq 0$. If $p_\nu \geq 2$ and at least one $Y_i$ is strictly bigger than $\nu$, then $\delta^\nu$ dominates $\delta^o(Y) = Y$ under $L_0$.*

*Proof.* Let $\Delta_i f(Y) = f(Y) - f(Y - e_i)$, then using Lemma 2.2.1 we have that the difference in risk is

$$R(\delta^\nu, \theta) - R(Y, \theta) = E_\theta D(Y),$$

where $D(Y)$ is independent of the parameters. Moreover, $D(Y)$ satisfies

$$\begin{aligned}
D(Y) &= \sum_{i=1}^{p} \left\{ \frac{f_i^2(Y+e_i)}{Y_i+1} + 2\Delta_i f_i(Y+e_i) \right\} \\
&= \sum_{i=1}^{p} \mathrm{I}(Y_i+1>\nu) \frac{\psi^2(Z+1)(Y_i+1-\nu)^2}{(p_\nu+Z_\nu-p_\nu\nu)^2} \frac{1}{Y_i+1} \\
&\qquad\qquad - 2\sum_{i=1}^{p} \Delta_i \mathrm{I}(Y_i+1>\nu) \frac{\psi(Z+1)(Y_i+1-\nu)}{p_\nu-1+Z_\nu+1-p_\nu\nu} \\
&\le \sum_{i=1}^{p} \mathrm{I}(Y_i+1>\nu) \left\{ \frac{\psi^2(Z+1)(Y_i+1-\nu)^2}{(p_\nu+Z_\nu-p_\nu\nu)^2} \frac{1}{Y_i+1} \right. \\
&\qquad\qquad \left. - 2\Delta_i \frac{\psi(Z+1)(Y_i+1-\nu)}{p_\nu-1+Z_\nu+1-p_\nu\nu} \right\} \\
&= \sum_{j=1}^{p_\nu} \left\{ \frac{\psi^2(Z+1)(Y_j+1-\nu)^2}{(p_\nu+Z_\nu-p_\nu\nu)^2} \frac{1}{Y_j+1} - 2\Delta_j \frac{\psi(Z+1)(Y_j+1-\nu)}{p_\nu-1+Z_\nu+1-p_\nu\nu} \right\} \\
&\le \sum_{j=1}^{p_\nu} \left\{ \frac{\psi^2(Z+1)(Y_j+1-\nu)}{(p_\nu+Z_\nu-p_\nu\nu)^2} - 2\Delta_j \frac{\psi(Z+1)(Y_j+1-\nu)}{p_\nu-1+Z_\nu+1-p_\nu\nu} \right\} \\
&= \frac{\psi^2(Z+1)}{p_\nu+Z_\nu-p_\nu\nu} - 2\psi(Z+1) + 2\frac{\psi(Z)(Z_\nu-p_\nu\nu)}{p_\nu-1+Z_\nu-p_\nu\nu} \\
&\le \frac{\psi(Z+1)}{p_\nu+Z_\nu-p_\nu\nu} \left\{ \psi(Z+1) - 2(p_\nu+Z_\nu-p_\nu\nu) \right. \\
&\qquad\qquad \left. +2\frac{(Z_\nu-p_\nu\nu)(p_\nu+Z_\nu-p_\nu\nu)}{p_\nu-1+Z_\nu-p_\nu\nu} \right\} \\
&= K(Z)\left\{ (Z_\nu-p_\nu)[\psi(Z+1)-2(p_\nu-1)] \right. \\
&\qquad\qquad \left. +(p_\nu-1)[\psi(Z+1)-2p_\nu] \right\} \le 0,
\end{aligned}$$

where

$$K(Z) = \frac{\psi(Z+1)}{(p_\nu+Z_\nu-p_\nu\nu)(p_\nu-1+Z_\nu-p_\nu\nu)}.$$

The second inequality is obtained because for all $Y_j \in Y_\nu$ we have that $Y_j+1 > Y_j+1-\nu > 0$. For the third inequality I have used that the function $\psi$ is non-decreasing. □

So for any point $\nu$ for which $p_\nu \ge 2$ we have an estimator that uniformly dominates the MLE. Figure 3.3 gives a visual summary of the simulations comparing the estimated risk performances of $\delta^{CZ}$ and $\delta^{Gm}$ with that of the new estimators $\delta^{m1}$ and $\delta^{\nu}$. In these simulations $\nu$ was set to one unit below the empirical median and $\psi$ to $p_\nu - 1$. On a side note, it is likely that with a slight

| | | $(0,4)$ | | | $(0,8)$ | | | $(8,12)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | $p=5$ | $p=10$ | $p=15$ | $p=5$ | $p=10$ | $p=15$ | $p=5$ | $p=10$ | $p=15$ |
| $\delta^{CZ}$ | 22.49 | 32.37 | 32.06 | 17.28 | 15.67 | 18.53 | 5.53 | 7.32 | 8.22 |
| $\delta^{Gm}$ | 16.67 | 30.97 | 30.99 | 16.96 | 14.97 | 17.87 | 17.12 | 20.18 | 20.07 |
| $\delta^{\nu}$ | 19.08 | 28.95 | 29.29 | 15.17 | 17.77 | 20.68 | 17.48 | 21.86 | 24.30 |
| $\delta^{m1}$ | -32.20 | 25.50 | 22.44 | -11.34 | -31.3 | -5.22 | 69.58 | 77.00 | 78.87 |
| | | $(12,16)$ | | | $(0,12)$ | | | $(4,16)$ | |
| | $p=5$ | $p=10$ | $p=15$ | $p=5$ | $p=10$ | $p=15$ | $p=5$ | $p=10$ | $p=15$ |
| $\delta^{CZ}$ | 4.44 | 5.55 | 6.14 | 6.09 | 10.06 | 13.02 | 6.00 | 7.46 | 8.05 |
| $\delta^{Gm}$ | 15.30 | 18.27 | 18.23 | 16.80 | 9.33 | 12.56 | 8.07 | 12.08 | 12.43 |
| $\delta^{\nu}$ | 15.42 | 19.77 | 21.81 | 16.97 | 14.22 | 17.94 | 7.97 | 11.99 | 14.02 |
| $\delta^{m1}$ | 67.38 | 79.71 | 81.25 | 61.20 | -45.78 | -26.74 | -14.08 | 15.28 | 22.62 |

Table 3.1: Percentage improvement in risk compared to the MLE under $L_1$ of the four estimators presented in Section 3.2. These are the same simulations as in Figure 3.3.

modification, the estimator $\delta^{\nu}$ is a good estimator under $c$-Loss. A conjecture is that the estimator $Y_i + f_i(Y)$, with

$$f_i(Y) = -\mathrm{I}(Y_i > \nu)\frac{(p_\nu - 1)(Y_i - \nu)}{p_\nu - 1 + (1+c)(Z_\nu - p_\nu \nu)},$$

dominates the MLE under the $c$-Loss function.

Table 3.1 provides a precise summary of the percentage savings in risk compared to the MLE. Two striking features of Table 3.1 are the solid performance of the estimator $\delta^{\nu}$ of Proposition 3.2.1, and the erratic performance of $\delta^{m1}$ of (3.10). The estimator $\delta^{\nu}$ appears insensitive to differing sizes of the samples $p$ and to differing sizes of the parameters. In view of these simulations it appears hazardous to use $\delta^{m1}$ if some of the parameters are thought to be small, while for the intervals $\theta_i \in (8,12)$ and $\theta_i \in (12,16)$, the risk performance of $\delta^{m1}$ is impressive.

## 3.3 Treating it as normal

As we have seen, the issue of shrinking towards some other point than the origin only arises when the Poisson parameters are thought to be of a certain size. If the Poisson means are thought to be small, there is no reason to shrink anywhere else than zero. By the central limit theorem the Poisson random variable $Y_i$ with mean $\theta_i$ is approximately normal with mean and variance $\theta_i$. One way to see this is to assume that $\theta_i$ is a natural number and think of $Y_i$ as the sum of

| | $\theta_i \in (0,8)$ | | | | $\theta_i \in (4,8)$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Poisson | | Normal | | Poisson | | Normal | |
| | $\delta^{CZ}$ | $\delta^{m1}$ | $\delta^{JS}$ | $\delta^{L}$ | $\delta^{CZ}$ | $\delta^{m1}$ | $\delta^{JS}$ | $\delta^{L}$ |
| $L_1$ | 19.44 | 3.69 | 15.86 | -1105.35 | 13.31 | 68.55 | 10.51 | 44.13 |
| | $\delta^{P}$ | $\delta^{m}$ | $\delta^{JS}$ | $\delta^{L}$ | $\delta^{P}$ | $\delta^{m}$ | $\delta^{JS}$ | $\delta^{L}$ |
| $L_0$ | 4.14 | 35.31 | 13.74 | -2.67 | 2.48 | 72.70 | 10.35 | 44.54 |

Table 3.2: Normal- and Poisson theory estimators on transformed and non-transformed Poisson data with $p = 31$. Estimated percentage savings in risk under $L_0$ and $L_1$ based on $10^4$ simulations.

$\theta_i$ Poisson random variables with mean one. Then,

$$\frac{\left(\sum_{j=1}^{\theta_i} Y_j^* - \theta_i\right)}{\sqrt{\theta_i}} \xrightarrow{d} N(0,1),$$

where $Y_j^* \sim \mathcal{P}(1)$ (Casella and Berger, 2002, 237). In this section I concentrate on uses of James-Stein type estimators in the Poisson setting. Recall that the prototypical example of the James-Stein estimator pertains to situations where $X_1, \ldots, X_p$ are normal with means $\xi_i$, $1 \leq i \leq p$ and equal variance (Efron and Morris, 1975). Using the Delta method we have that since $Y_i \approx N(\theta_i, \theta_i)$, the random variable $X_i = 2\sqrt{Y}_i$ is approximately normal with mean $\xi_i = 2\sqrt{\theta}_i$ and unit variance. For this reason it is tempting to apply the variance stabilizing transformation to the Poisson observations, treat $X_i = 2\sqrt{Y}_i$ as normal, use a James-Stein type estimator and transform back.

In doing so it is illuminating to note the resemblance of $\delta^{CZ}$ to the James-Stein estimator in (1.2). If the variance stabilizing transformation $2\sqrt{y}$ has been applied, then transforming back to the Poisson world we have that

$$\frac{1}{4}\left(\delta_i^{JS}\right)^2 = \frac{1}{4}\left(1 - \frac{b}{\sum_{i=1}^{p} X_i^2}\right)^2 X_i^2 = \left(1 - \frac{4b}{Z}\right)^2 Y_i.$$

Since $Z$ has a non-zero probability of being zero, it is reasonable to add some constant to the denominator above (Tsui and Press, 1982, 96). Hence, we have an estimator that looks very much like that of Clevenson and Zidek (1975). Despite this being an approximative argument, the similarity of form between normal and Poisson shrinkage estimators indicates that James-Stein type estimators might be very effective after a variance stabilizing transformation of the Poisson counts (Hudson, 1985, 248). A quick simulation study corroborates this speculation. In the simulation study reported in Table 3.2, I compare Poisson theory estimators with two James-Stein type estimators, $\delta^{JS}$ in (1.2) with $b = p - 2$ and $\delta^{L}$ in (1.5), under both $L_0$ and $L_1$.

Under the weighted squared error loss function $L_1$ we see that in the first interval the performance of the James-Stein estimator almost matches that of

$\delta^{CZ}$. This is probably due to the fact that for the largest $\theta_i \in (0,8)$ the normal approximation works reasonably well. The new estimator $\delta^{m1}$ of (3.10) barely beats the MLE when the $\theta_i$ are in $(0,8)$. In this interval the performance of the Lindley-estimator is extremely bad. Most likely, $\delta^L$ is penalized heavily for pulling too many observations up towards the empirical mean and thereby overestimating small means.

When the parameter interval is $(4,8)$, on the other hand, both $\delta^{m1}$ and the Lindley-estimator outperform the two shrink-to-zero estimators. In this interval overestimation is not as severly penalized simply because the means are not that small. The performance of the new estimator $\delta^{m1}$ is outstanding, resulting in a risk reduction of 68.55% compared to the MLE. But, as seen above (cf. Table 3.1), it is very risky to use this estimator if one suspects some of the $\theta_i$ to be small.

Under the squared error loss function $L_0$ the James-Stein estimator shows better performance than the estimator of Peng (1975) when $\theta_i \in (0,8)$. As a consequence of not severly penalizing overestimation of small parameters (using $L_0$), the best risk performance for $\theta_i \in (0,8)$ is achieved by the shrink-to-mean estimator $\delta^m$ of (3.8), with a reduction in risk of 35.31% relative to the MLE. For the second interval $\theta_i \in (4,8)$, $\delta^{JS}$ once again beats the estimator of Peng. The two estimators $\delta^m$ and $\delta^L$ show very solid risk performance, beating the two shrink-to-zero estimators by large amounts. For $\theta_i \in (4,8)$, the performance of $\delta^m$ is truly remarkable. The reason for $\delta^m$ performing better than $\delta^L$ is probably that $\delta^m$ reduces to the one-dimensional MLE, namely $\bar{Y}$, when the sample variance is low.

Two findings of this simulation study are worth emphasizing. First, the James-Stein estimator seems very robust confronted with deviances from the normal distribution under both $L_0$ and $L_1$. Altough $\delta^{JS}$ does not beat $\delta^{CZ}$ under $L_1$, its performance is rather solid for both intervals. Second, under $L_0$ the (normal theory) James-Stein estimator actually performs much better than the (Poisson theory) estimator of Peng (1975). I suspect that for intervals with a lower upper bound than 8, i.e. $\theta_i \in (0,4)$, the estimator of Peng will perform better than the James-Stein estimator under $L_0$. Whether this is the case, and how small the $\theta_i$ must be for $\delta^P$ to be the superior estimator (compared with $\delta^{JS}$), is an interesting theme for further study.

# 4

# Poisson regression

So far I have been looking at estimators that shrink the observations towards some point in the parameter space. This point has either been the origin, some pre-specified point or a point generated by the data. I will now consider estimators that shrink the observations towards different points in the parameter space, where these points are functions of covariates. In a first part, I consider a hierarchical Bayesian regression model. Second, I look at an empirical Bayes version of this model. These two models are compared to the usual Poisson regression model and three other regression models. Throughout this chapter I assume that loss is given by the squared error loss function $L_0$.

To gain intuition into the development of the Bayes model, I start by comparing the risks of two Bayes estimators to that of the MLE. In Section 3.1 we saw that the risk $p\bar{\theta}$ of the MLE could be improved upon provided that we have prior information about the parameters. The risk of the Bayes estimator $\mu w + (1-w)y_i$ (given in (3.5)) is restated here

$$R(\delta^B, \theta) = p\bar{\theta}(1-w)^2 + w^2 \sum_{i=1}^{p} (\mu - \theta_i)^2 .$$

From this risk function, we see how a good prior guess $\mu$ of $\theta_i$ decreases the risk. Importantly, since we make one guess $\mu$ for $p$ parameters it is crucial that the parameters $\theta_1, \ldots, \theta_p$ are not too different. An intuitive way to further improve the savings in risk is to guess not once, but $p$ times. Thereby replacing $\mu$ in (3.5) (restated above) with $\mu_1 \ldots, \mu_p$, resulting in the risk function

$$R = p\bar{\theta}(1-w)^2 + w^2 \sum_{i=1}^{p} (\mu_i - \theta_i)^2 . \tag{4.1}$$

This is the idea behind the pure and empirical Bayes regression models that I study in Section 4.2. Before I present these two models, three other potential models are introduced. In Section 4.3 the differences between these estimators are illustrated and their risk performances are compared by way of a simulation study.

## 4.1 Standard and non-standard models

It is worth mentioning that if one thinks that the Poisson means are structured in some way, but have no information about this structure the estimator $\delta^o(Y) = Y$ is a good candidate. This estimator will still be referred to as the MLE, and serves as the baseline model in the simulations in Section 4.3.

The standard Poisson regression model takes the observations $Y_i \mid z_i, 1 \leq i \leq p$ to be independent $\mathcal{P}(\exp(z_i^t\beta))$. The $z_i$ are $k \times 1$ vectors of covariates, and the $k \times 1$ coefficient vector $\beta$ is estimated by the maximum likelihood method. In most applications of the Poisson regression model, the primary interest is in inference on $\beta$. In this thesis, focus is on the Poisson means for which the Poisson regression model estimates are

$$\hat{\theta}_i = \exp(z_i^t\hat{\beta}),\ 1 \leq i \leq p.$$

Contrary to the estimators that are presented below, the estimated means of the Poisson regression model all lie on the curve $(z_i, \exp(z_i^t\hat{\beta}))$, $1 \leq i \leq p$.

In settings where it is thought that there is something to gain from structuring the prior guesses of the Poisson means, some of the means must be anticipated to be rather large. If this was not the case, one might as well use $\delta^{CZ}$ or one of the estimators of Section 3.1. Consequently, in settings where it is deemed advantageous to structure the prior guesses, the variance stabilizing transformation $X_i = 2\sqrt{Y_i}$ is likely to work well. This means that one might use the James-Stein type estimator $\delta^{EB}$ in (1.6) on the transformed data $X_1, \ldots, X_p$, then transform back $1/4(\delta^{EB})^2$. $\delta^{EB}$ shrinks the observations towards $Z\hat{\beta}$ where $\hat{\beta} = (Z^tZ)^{-1}Z^tX$. $Z$ is the $p \times k$ design matrix with $z_i^t$ as its rows.

Hudson (1985) devised a regression method inspired by the estimator $\delta^{EB}$ for the Poisson setting. His method consists of transforming each observation $H_i(Y_i) = \sum_{k=1}^{Y_i} k^{-1}$ and use that $\log(x + 0.56)/0.56)$ is a very good approximation to $H_i(x)$ (Hudson, 1985, 248). Then the fitted values $\hat{H}_i = Z(Z^tZ)^{-1}Z^tH$ are calculated, and finally one "transforms" back to the Poisson world $\hat{Y}_i = 0.56(\exp(\hat{H}) - 1)$. The estimator of Hudson (1985) is given componentwise as

$$\delta_i^H(Y) = Y_i - \frac{(N_0(Y) - k - 2)^+}{||H - \hat{H}||^2}(H_i - \hat{H}_i), \tag{4.2}$$

if $Y_i + 0.56$ is bigger than the shrinkage factor $(N_0(Y) - k - 2)^+/||H - \hat{H}||^2$ and is equal to $\hat{Y}_i$ otherwise. $N_0(Y)$ counts the number of observations bigger than zero.

The three models briefly introduced in this section will be compared with the pure and the empirical Bayes regression models that I study next.

## 4.2 Pure and empirical Bayes regression

Assume that $Y_i \,|\, z_i \sim \mathcal{P}(\theta_i)$ for $i = 1, \ldots, p$ are independent, where $z_i$ are $k \times 1$ vectors of fixed covariates. As discussed above, in an attempt to further improve on the risk $p\bar{\theta}(1-w)^2 + w^2 \sum_{i=1}^{p} (\mu - \theta_i)^2$ I will put some structure on the prior mean $\mu$, and consider $\mu_1, \ldots, \mu_p$. Let the Poisson means be

$$\theta_i \sim \mathcal{G}(b\mu_i, b), \; 1 \leq i \leq p$$

independent, and model the prior means by $\mu_i = \exp(z_i^t \beta)$. If sufficient prior information is available to determine $\beta$ and $b$ we have the Bayes estimator $\mu_i w + (1-w) y_i$, $w = b/(b+1)$ whose risk function is given in (4.1). In most applications the statistician is likely to be uncertain about $(\beta_1, \ldots, \beta_k, b)$, and it is natural to express this uncertainty through probability distributions over these $k+1$ parameters.[1] I put a multivariate normal distribution over the regression parameters $\beta$, i.e. $\beta \sim N_k(\xi, \Sigma)$. To describe the uncertainty associated with the parameter $b$ I use a Gamma distribution, $b \sim \mathcal{G}(\zeta, \eta)$. We will assume that $\beta$ and $b$ are independent, which means that the distribution of the so-called hyperparameters (the parameters in the prior on the prior) is $\pi_2(\beta, b) = \mathcal{G}(\zeta, \eta) N_k(\xi, \Sigma)$. In a hierarchical setup like this it is common to call the $\theta_i$ the *individual* parameters and $(\beta, b)$ the *structural* parameters (Christiansen and Morris, 1997). In a medical study with $p$ patients for example, $\theta_i$ describes a characteristic of patient $i$, while $(\beta, b)$ describes how the differing characteristics of patient $i = 1, \ldots, p$ are related.

The parameters of primary interest for inference are the Poisson means $\theta_1, \ldots, \theta_p$. Since the posterior distribution of these cannot be derived analytically, I rely on Gibbs sampling in order to draw samples from the joint posterior disitribution $\theta_1, \ldots, \theta_p, \beta, b \,|\, \text{data}$. This distribution is

$$\begin{aligned}
\pi(\theta, \beta, b \,|\, y) &\propto \left\{ \prod_{i=1}^{p} \mathcal{P}(\theta_i) \mathcal{G}(b\mu_i, b) \right\} \mathcal{G}(\zeta, \eta) N_k(\xi, \Sigma) \\
&\propto \left\{ \prod_{i=1}^{p} \frac{b^{b\mu_i}}{\Gamma(b\mu_i)} \theta_i^{y_i + b\mu_i - 1} e^{-(b+1)\theta_i} \right\} b^{\zeta - 1} e^{-\eta b} N_k(\xi, \Sigma) \qquad (4.3) \\
&= \frac{b^{b \sum_{i=1}^{p} \mu_i + \zeta - 1}}{\prod_{i=1}^{p} \Gamma(b\mu_i)} \left\{ \prod_{i=1}^{p} \theta_i^{y_i + b\mu_i - 1} \right\} e^{-(b+1)\sum_{i=1}^{p} \theta_i - \eta b} N_k(\xi, \Sigma).
\end{aligned}$$

[1] To quote Gelman and Robert (2013, 3): "(...) priors are not reflections of a hidden "truth" but rather evaluations of the modeler's uncertainty about the parameter."

The point of the Gibbs sampler is to break a complex problem into a sequence of smaller problems (Robert and Casella, 2010, 200). So instead of attacking the joint posterior density above directly, we approach it by way of the full conditional distributions. These full conditional distributions are given by

$$\theta_i \,|\, \beta, b, y \sim \mathcal{G}(y_i + b\mu_i, b + 1), \tag{4.4}$$

for the $i = 1, \ldots, p$ Poisson means. For the structural parameters $(\beta, b)$ we have

$$\beta \,|\, \theta, b, y \propto \left\{ \prod_{i=1}^{p} \frac{b^{b\mu_i}}{\Gamma(b\mu_i)} \theta_i^{b\mu_i} \right\} N_k(\xi, \Sigma), \tag{4.5}$$

and

$$b \,|\, \theta, \beta, y \propto \frac{b^{b\sum_{i=1}^{p} \mu_i + \zeta - 1}}{\prod_{i=1}^{p} \Gamma(b\mu_i)} \left\{ \prod_{i=1}^{p} \theta_i^{b\mu_i} \right\} e^{-(p\bar{\theta} + \eta)b}. \tag{4.6}$$

More details on the Gibbs sampler and its implementation are included in Appendix B. Notice that the posterior expectation of $\theta_i$, $1 \leq i \leq p$ is

$$E[\theta_i \,|\, Y] = E[E[\theta_i \,|\, Y, \beta, b] \,|\, Y] = E\left[ \mu_i \frac{b}{b+1} + \left(1 - \frac{b}{b+1}\right) y_i \,|\, Y \right],$$

which shows that the posterior mean of $\theta_i$ is still some weighted average of the prior mean $\mu_i$ and the observation $y_i$. In Figure 4.1 we clearly see how this weighting scheme works. In addition to the observations (circles) and the Bayes estimates (triangles), the plot contains the true (solid) and the estimated (dashed) regression curves. For this example the true parameters were $\beta = (0.20, 0.50)^t$ and $b = 0.20$, then $\theta_i$, $1 \leq i \leq p$ were drawn from Gamma distributions with means $\mu_i$ and variances $\mu_i / b$. The prior on $\beta$ was $N_k(0, I_k)$ and $b$ was given a $\mathcal{G}(2, 3)$ prior. In order to draw samples from the posterior distribution the Gibbs sampler was run for 10000 iterations, of which the last 8000 were retained. This gave the posterior estimates reported in Table 4.1.

| $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{b}$ |
|---|---|---|
| 0.045 | 0.51 | 0.17 |
| [−0.478, 0.547] | [0.430, 0.591] | [0.129, 0.223] |

Table 4.1: Posterior estimates of the structural parameters $(\beta, b)$. The second row contains the highest posterior density regions with 90% probability mass between the lower and upper limit.

In the plot we see how the model produces estimates of $\theta_i$ that are compromises between the observations and the estimated regression line $(z_i, \hat{\mu}_i)$.

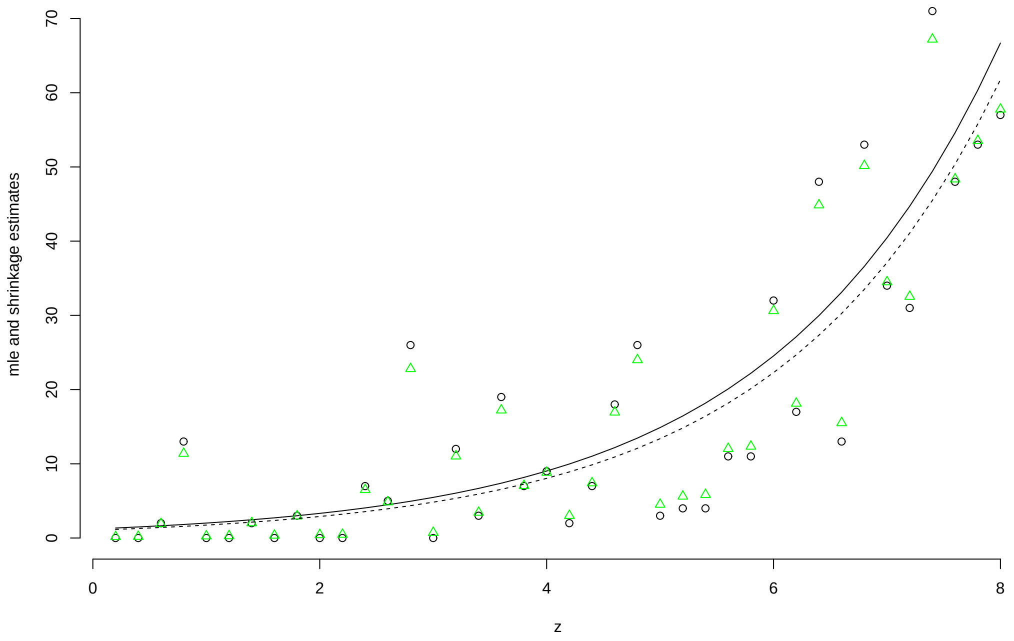

Figure 4.1: The regression model in Section 4.2. The circles are the ML-estimates (the observations), the triangles are the Bayes estimates. The solid curve is the unknown regression line and the dotted line is the one based on $\hat{\beta}$.

In the context of the regression model above, the empirical Bayes approach consists of estimating the hyperparameters $(\beta, b)$ via the marginal likelihood of all the data $m(y_1, \ldots, y_p \,|\, \beta, b)$. Assuming that the Poisson counts are mutually independent this marginal distribution is given by

$$\begin{aligned} m(y_1, \ldots, y_p \,|\, \beta, b) &= \prod_{i=1}^{p} \int_0^\infty \mathcal{P}(\theta_i)\mathcal{G}(b\mu_i, b)\, d\theta_i \\ &= \prod_{i=1}^{p} \frac{\Gamma(y_i + b\mu_i)}{y_i!\Gamma(b\mu_i)} \left(1 - \frac{1}{b+1}\right)^{b\mu_i} \left(\frac{1}{b+1}\right)^{y_i}, \end{aligned}$$

which is the product of $p$ Negative binomial distributions with parameters $b\mu_i$ and $(1+b)^{-1}$. From this distribution the parameters $\beta$ and $b$ can be estimated by the maximum likelihood method and then plugged into the prior distribution, resulting in the posterior estimator

$$E[\theta_i \,|\, Y] = \exp(z_i^t \hat{\beta})\, \frac{\hat{b}}{\hat{b}+1} + \left(1 - \frac{\hat{b}}{\hat{b}+1}\right) y_i.$$

In the next section I compare the performance of the pure and empirical Bayes estimators with those presented in Section 4.1.

## 4.3 A simulation study

Table 4.2 summarizes the simulated risks and percentagewise reductions in risk relative to the MLE of the estimators presented above. Except for the standard Poisson regression model, all the models improve on the MLE. This means that some smoothing towards an estimated regression line improves performance, while the Poisson regression model obviously smooths too much.

| Estimator | $R(\,\cdot\,,\theta)$ | % risk reduction |
|---|---|---|
| Pure Bayes | 673.68 | 13.54 |
| Emp. Bayes | 670.30 | 13.97 |
| $\delta^H$ | 762.74 | 2.11 |
| Poisson reg. | 3327.35 | -327.05 |
| $1/4(\delta^{EB})^2$ | 671.24 | 13.85 |
| MLE | 779.15 | 0.00 |

Table 4.2: Simulated risks and percentage reductions in risk relative to the MLE $(R(Y,\theta) - R(\,\cdot\,,\theta))/R(Y,\theta) \times 100$. Estimates based on 500 simulations. A negative "risk reduction" indicates that the estimator performed worse than the MLE. More details are found in Appendix C.7.

The most surprising feature of Table 4.2 is the performance of $1/4\left(\delta^{EB}\right)^2$, the transformation of an estimator constructed for a normal setting. In view of the parameter values many of the simulated observations are very small, which should limit the efficiency of the normal approximation to the Poisson. Despite this fact, the risk performance of $1/4\left(\delta^{EB}\right)^2$ is the same as for the Bayes and empirical Bayes models, with a reduction in risk of 13.85% compared to the MLE. This is most likely a consequence of using the squared error loss function. With the squared error loss function the risk-competition is won and lost in the estimation of large parameters, for which the normal approximation works well. It is, in other words, unlikely that $1/4\left(\delta^{EB}\right)^2$ would have done equally well under $L_1$.

At the same time, the estimator of Hudson (1985), which in a sense is $\delta^{EB}$ adjusted to the Poisson distribution, only improves with 2.20% on the MLE. The unexciting performance of Hudson's estimator is probably a consequence of this estimator being constructed with an aim of uniform dominance (this was the aim, but Hudson (1985, 250-251) only proves dominance approximately). As a consequence, the estimator $\delta^H$ smooths the observations towards a curve that is too low in the plane (cf. the estimates in Figure 4.2).

The Bayes estimator and the empirical Bayes estimator show solid performance. Deliberately the parameters of the prior distributions were somewhat misspecified in order to "level the playing field" with the other estimators. As a result, the empirical Bayes estimator performs slightly better than the pure Bayesian model.

One advantage of the two Bayesian models compared to the three others, is that they provide estimates of the between-individual variability $b$. This is the variability of the individual parameters $\theta_1, \dots, \theta_p$. Therefore, contrary to the other models, the hierarchical nature of the Bayesian model and the empirical Bayesian model yields a nice summary of the two levels of variation, the sampling variation and that between the distributions generating the samples (Christiansen and Morris, 1997, 619).

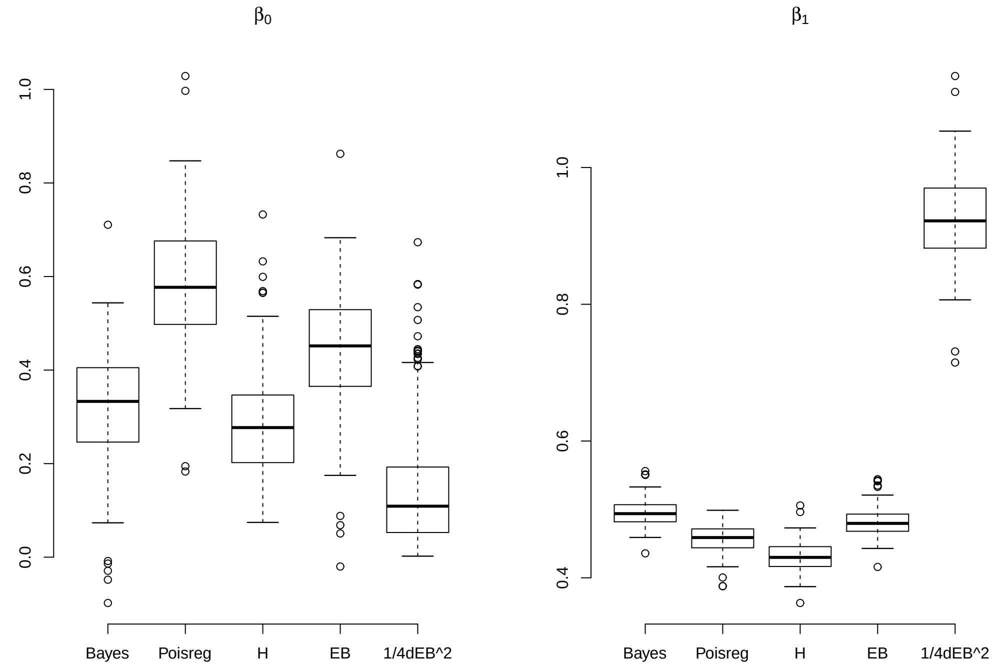

Figure 4.2: The box plots summarize the estimates of the intercept and the slope for the five different models. They are based on 500 simulations. The boxes indicate the interquartile range, the solid line is the average and the dashed line connects the 0.025 and 0.975 quantiles.

In interpreting these results it is important to keep in mind that in the simulations $\theta_1, \dots, \theta_p$ were held fixed. First they were generated from $\mathcal{G}(b\mu_i, b)$, $1 \leq i \leq p$ distributions with $(\beta_1, \beta_2) = (0.2, 0.5)$ and $b = 0.20$, then held constant for the 500 simulations. Another sample of $\theta_1, \dots, \theta_p$ is likely to give somewhat different results. The differing performances of the estimators are functions of two things: what the MLE is smoothed towards and by how much. Figure 4.2 therefore provides a graphical summary of the estimates of the intercept and the slope of the curves that the five models smooth the observations towards. The average estimate of $b$ for the Bayesian and empirical Bayesian models were 0.23 [0.16, 0.32] and 0.20 [0.14, 0.29] respectively (0.025 and 0.975 quantiles in the brackets), meaning that on average the pure Bayesian model smoothed the observations a little more than the empirical Bayesian one.

# 5

# Concluding remarks

In this thesis I have derived a class of minimax estimators that uniformly dominates the MLE under the apparently novel $c$-Loss function. As we have seen the $c$-Loss function serves a dual purpose. On one hand, it is in many instances a reasonable loss function in itself, because we seek precise estimates of $\theta_i$, $1 \leq i \leq p$ and a precise estimate of $\gamma = \sum_{i=1}^{p} \theta_i$. On the other hand, it enables the derivation of a class of minimax estimators that uniformly dominate the MLE under the commonly used weighted squared error loss function. Clevenson and Zidek's class of estimators is a subclass of the class $\mathcal{D}_c$ derived in this thesis. The introduction of the $c$-parameter allows for easy control of the amount of shrinkage away from the MLE. Situations where such limited shrinkage is of interest were discussed in Section 2.5 and a method for finding $c$ was proposed. A direct continuation of this thesis would be to look at other methods for determining the optimal $c$-value and, more generally, other methods for deriving estimators that are optimal under conflicting desiderata when estimating several Poisson means.

In the following four sections I outline four themes related to this thesis that ought to be further explored.

## 5.1 Dependent Poisson means

In the Bayesian models of this thesis the Poisson means $\theta_1, \ldots, \theta_p$ have been assumed independent (the Dirichlet model being the only exception). In this section I sketch one way in which the assumption of the $\theta_i$ being independent might be relaxed. If $X$ is a random variable with cumulative density $F_X$, then the random variable $U = F_X(X)$ is uniformly distributed on the unit interval. Reverting the argument we have that $X = F_X^{-1}(U)$ have cumulative density $F_X$. I will now consider a model were the $\theta_i$ come from the conjugate Gamma distribution, but are dependent. Let the Poisson counts $Y_1, \ldots, Y_p$ be independent given $\theta_i$, $1 \leq i \leq p$, and assume that the Poisson means are given by

$$\theta_i = G^{-1}(\Phi(V_i); a, b), \quad \text{for } i = 1, \dots, p,$$

where $\Phi(\cdot)$ is the standard normal cumulative density function, and $G^{-1}$ is the inverse cumulative density function of a Gamma distribution. The $p \times 1$ vector $V = (V_1, \dots, V_p)^t$ is distributed $V \sim N_p(0, \Sigma)$ where $\Sigma$ is a $p \times p$ covariance matrix. The covariance structure can for example be hypotesized to be decreasing in the temporal or spatial distance between the means. A straightforward way to model this is by the covariance of the normals $V_i$, $1 \leq i \leq p$, setting $\text{Cov}(V_i, V_j) = \rho^{|i-j|}$, $\rho \leq 1$, i.e. if $p = 4$ the covariance matrix $\Sigma$ takes the form

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

This means that the covariance structure of the Poisson means is modelled via the covariance structure of a multivariate normal distribution. To get a sense of how the covariance of the normals translates to the covariance of the Poisson means, the three panels in Figure 5.1 plot $\rho^{|i-j|}$ and the simulated estimates of the correlation $(b^2/a)\text{Cov}(\theta_i, \theta_j)$ for $|i - j| = 0, 1, \dots, p - 1$. The three panels in Figure 5.1 suggest that the covariance structure of the multivariate normal random variables $V_i$, $1 \leq i \leq p$ translates more or less directly to the covariance structure of $\theta_i$, $1 \leq i \leq p$. If this is in fact the case, it is a very appealing feature of the model outlined above.

## 5.2 Constrained parameter space

In many settings with Poisson data the statistician will have information about constraints on the parameter space (see e.g. (Johnstone and MacGibbon, 1992, 808)). For example, suppose that it is known that $\theta_i \leq \theta_0$ for all $i$. An interesting scenario is when the upper bound is small, for example $\theta_0 = 1$, so that the parameter space is the simplex $S = [0, 1]^p$. Under the squared error loss function $L_0$ the risk of the MLE is $R(Y, \theta) = p\bar{\theta}$, and $\sup_\theta R(Y, \theta) = p$ when the parameter space is $S$. Notice that any constant estimator not equal to one, that is $\delta(Y) = d = [0, 1)^p$, has

$$\sup_{\theta \in [0,1)} R(d, \theta) \leq p.$$

Intuitively, all estimators $\delta \in [0, 1)$ (which excludes the MLE) should be minimax and have smaller max risk than the MLE. An intriguing question is whether

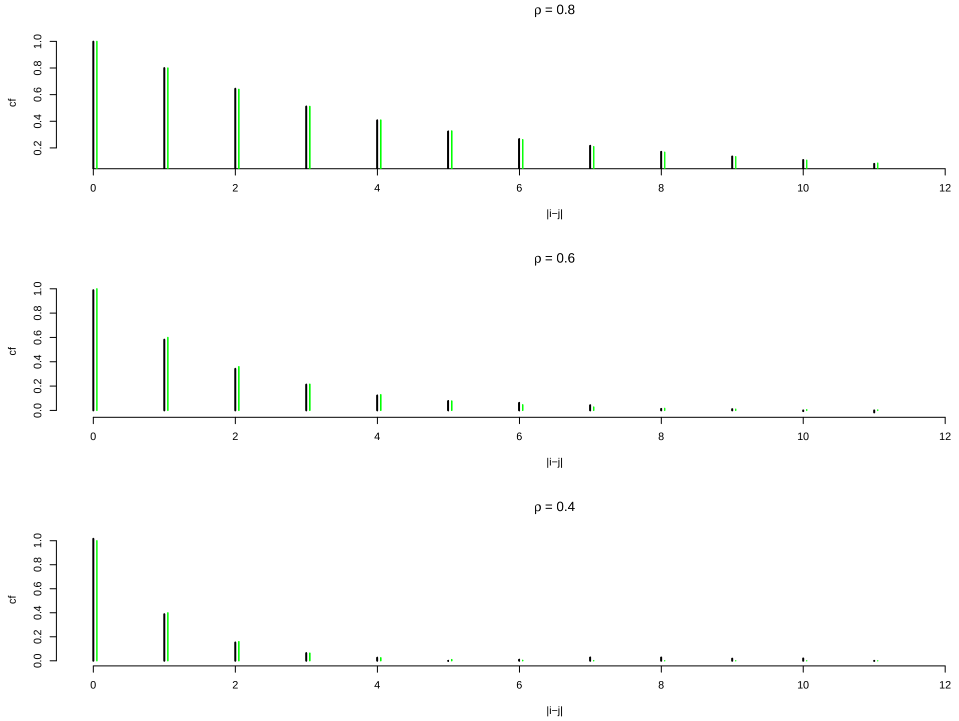

Figure 5.1: Estimates of $(b^2/a)\text{Cov}(\theta_i, \theta_j)$ (solid lines) and $\rho^{|i-j|}$ (green lines) for $|i-j| = 0, 1, \ldots, p-1$ with $p = 12$.

this is in fact the case, and whether we can find estimators with other nice properties in this setting, for example estimators that uniformly improve on the constant estimator in terms of risk. This relates to how to utilize additional prior information about the Poisson means. A Bayesian "baseline" model is assuming $\theta_1, \ldots, \theta_p$ independent and uniform over the simplex, i.e. $\pi(\theta_i) = \text{I}(\theta_i \in S)$. The Bayes solution under $L_0$ is then

$$E[\theta_i \,|\, \text{data}\,] = \frac{\int_{\Theta_0} \theta_i^{y_i+1}\, e^{\theta_i}\, d\theta_i}{\int_{\Theta_0} \theta_i^{y_i}\, e^{\theta_i}\, d\theta_i} = \frac{\Gamma(y_i+2)}{\Gamma(y_i+1)} \frac{G(1,\, y_i+2, 1)}{G(1,\, y_i+1, 1)},$$

where $G$ is the cumulative density function of the Gamma distribution. Knowing that the Poisson means are constrained to the simplex $[0,1]^p$ also invites an analysis similar to that in Section 2.7 using a Dirichlet prior directly on $(\theta_1, \ldots, \theta_p)$. The properties of these Bayes solutions, as well as differing Bayes solutions under other plausible loss functions, is a theme I would like to investigate further.

## 5.3 Estimation of Poisson means in real time

Assume that we observe $Y_1$ to $Y_p$ sequentially, where $Y_i$ comes from $\mathcal{P}(\theta_i(t_i - t_{i-1}))$ for $i = 1, \dots, p$. Assume that all the intervals $[t_{i-1}, t_i)$ are of equal length (this is not a vital assumption for the problem presented). The study starts at $t_0$ and at $t_1$ the statistician is supposed to provide an estimate of $\theta_1$, at $t_2$ an estimate of $\theta_2$ and so on. Whilst the estimates are provided sequentially, the loss is incurred at the end of the study period. Natural loss functions are any of those studied in this thesis, i.e. $L_1$, $L_{1,t}$, $L_c$ or $L_{c,t}$.

At *the end* of the period the optimal estimators in terms of improved risk relative to the MLE are the estimators I have been studying in this thesis. The point is that these estimators cannot be used when the estimates are supposed to be delivered sequentially in real time. They cannot be used simply because they all involve $Z = \sum_{i=1}^{p} Y_i$, which is unknown to the statistician at the point in time when the estimates are supposed to be provided.

Intuitively, it should be possible to improve on the MLE in such a situation. It would be interesting to investigate how a, say $\hat{Z}$, ought to be constructed in order to achieve the maximum savings in risk relative to the MLE, and if such an estimate of $Z$ can be constructed without losing the minimax property of the estimators studied in this thesis.

## 5.4 A final note on admissibility

I close off with a note on admissibility and the estimators in the class $\mathcal{D}_c$ of Corollary 2.3.3. In particular, I discuss the question of admissibility of the estimator that minimizes the $c$-Loss function, namely $\delta_1^c$ in (2.9). It turns out that we do not know whether this estimator is admissible or not. In a first part, I show how the admissibility of Clevenson and Zidek type estimators can be proven. Thereafter, I substantiate why we are not able to prove admissibility of the optimal estimator (in terms of minimizing risk) under the $c$-Loss function.

By way of a Bayesian analysis Clevenson and Zidek derive the estimator

$$\delta_m^{CZ}(Y) = \left(1 - \frac{m+p-1}{m+p-1+Z}\right) Y. \tag{5.1}$$

When $m = 0$ this estimator is equal to the one in (1.4) with $\psi(Z) = p - 1$. Importantly, the resulting estimator

$$\delta_0^{CZ}(Y) = \left(1 - \frac{p-1}{p-1+Z}\right) Y,$$

is the best version of their estimator in terms of minimizing risk under $L_1$. Clevenson and Zidek show that $\delta_m^{CZ}$ in (5.1) is admissible for $m > 1$, since it is proper Bayes. Drawing on the heuristic method for determining admissibility developed by Brown (1979), Clevenson and Zidek suspected that their estimator is admissible for $m \geq 0$. If this is indeed the case, then the best version of their estimator in terms of minimizing risk, which is obtained by setting $m = 0$, is admissible. Admissibility of $\delta_0^{CZ}$ is, however, yet to be proven.

The work of Johnstone (1984, 1986) on the connection between admissibility of estimators and the recurrence of associated Markov chains leads to a nice theorem for determining the admissibility or inadmissibility of simultaneous estimators of Poisson means under weighted squared error loss $L_1$ (see e.g. Robert (2001, 399)). Johnstone (1984, 1986) provides sufficient conditions for proving the admissibility of generalized Bayes estimators under $L_1$. According to these conditions (Theorem 8.2.18 in Robert (2001, 399)) a generalized Bayes estimator of the form

$$\delta(Y) = (1 - \phi(Z))\, Y,$$

is admissible under $L_1$ if there exist finite $K_1$ and $K_2$ such that $\sqrt{z}\phi(z) \leq K_1$ for every $z \geq 0$, and for $z > K_2$,

$$z\, \phi(z) \geq p - 1.$$

The Clevenson and Zidek estimator in (5.1) is a generalized Bayes estimator. To see this, consider the two-stage prior setup of Section 2.7 where $\theta_i$, $1 \leq i \leq p$ are independent $\mathcal{G}(1, b)$ and $b \sim \pi_2(b)$. Let now $\pi_2(b) \propto b^{m-2}(b+1)^{-m}$, $b > 0$. Then $\pi_2$ is a proper prior for $m > 1$, and an improper prior when $0 \leq m \leq 1$. The generalized Bayes solution is then (5.1). The function

$$\sqrt{z}\phi(z) = \sqrt{z}\,\frac{m+p-1}{m+p-1+z},$$

is first increasing, reaching its maximum at $z = m + p - 1$, then decreasing. This means that (5.1) satisfies

$$\sqrt{z}\phi(z) = \sqrt{z}\,\frac{m+p-1}{m+p-1+z} \leq \frac{1}{2}\sqrt{m+p-1} = K_1,$$

for all $z$, and for all $m \geq 0$. The second of the conditions of Johnstone (1984, 1986) does, however, only hold for $m$ strictly bigger than zero. With $\phi(z) = m + p - 1$, the second condition implies that there must exist a $K_2$ such that for all $z > K_2$ the inequality

$$z\, m \geq (p-1)(m+p-1),$$

holds. Clearly, such a finite value $K_2$ only exists when $m > 0$. In conclusion, the work of Johnstone (1984, 1986) has provided the tools for determining the admissibility of Clevenson and Zidek type estimators. The question concerning the admissibilty of $\delta_0^{CZ}$ with $m = 0$ remains open, but with $m > 0$ very small, one can get arbitrarily close to $\delta_0^{CZ}$ with admissible estimators.

When it comes to the $c$-Loss function we are further from proving that the optimal estimator in terms of minimizing risk $\delta_1^c$, is admissible. What we do know about the estimators in the class $\mathcal{D}_c$ is that the class of Bayes estimators derived in Proposition 2.7.1 are admissible because they are proper Bayes. In addition, we know that the optimal estimator under the $c$-Loss function $\delta_1^c$ of (2.9), restated here

$$\delta_1^c(Y) = \left(1 - \frac{p-1}{p-1+(1+c)Z}\right) Y,$$

is very close to a proper Bayes solution. The same two-stage prior setup as for deriving $\delta_m^{CZ}$ in (5.1), yields the Bayes solution under $c$-Loss

$$\delta^B(Y) = \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z} \frac{Z}{m+p-1+Z} Y. \tag{5.2}$$

This estimator is proper Bayes for $m > 1$ and generalized Bayes for $0 \leq m \leq 1$. With $m$ close to one (i.e. $m = 1.0001$) we see that the proper Bayes solution $\delta^B$ is very close to $\delta_1^c$, particularly for large $p$ and/or $Z$.

To summarize, although it is not known whether $\delta_1^c$ is admissible or not, I have shown that it is generalized Bayes (see Section 2.7) and that it is indistinguishable from a proper Bayes estimator for $p$ and $Z$ of a certain size. In addition, from the derivation of $\delta_1^c$ in Section 2.3 it follows that $\delta_1^c$ is not uniformly dominated by any estimator on the form $(1 - \phi(Z))Y$.

The theorem of Johnstone (1984, 1986) cannot be used for $\delta^B$ since it only applies to the weighted squared error loss function $L_1$. A truly interesting continuation of this thesis would be to look for admissibility conditions for estimators under the $c$-Loss function, similar to those Johnstone (1984, 1986) derive for $L_1$. If such an endeavour is successful, I suspect that it should be possible to prove that $\delta^B$ in (5.2) is admissible for $m > 0$, not only for $m > 1$. More generally, this line of work is concerned with finding possible admissibility conditions for optimal simultaneous estimators (in terms of risk) under loss functions that aim to balance the total and the individual risk. The $c$-Loss function is one such loss function, and $\delta_1^c$ is one such compromising estimator.

# Appendix A

# Some calculations and two tables

This appendix consists of well known results used throughout the text as well as calculations and proofs that are omitted in the text.

## A.1 The Poisson and the Gamma distribution

A Poisson random variable $Y$ with mean $\theta > 0$ has probability mass function

$$P(Y = y \,|\, \theta) = \frac{1}{y!}\, \theta^y \, e^{-\theta}.$$

The moment generating function is

$$M_Y(t) = E[e^{tY}] = \sum_{y=0}^{\infty} e^{tY} \frac{1}{y!} \theta^y e^{-1} = e^{-\theta} \sum_{y=0}^{\infty} \frac{(e^t\theta)^y}{y!} = e^{\theta(e^t-1)}.$$

If $Y_1, \ldots, Y_p$ are independent Poisson with means $\theta_i$, $1 \le i \le p$ and $Z = \sum_{i=1}^{p} Y_i$, then the moment generating function of $Z$ is $M_Z(t) = \prod_{i=1}^{p} e^{\theta_i(e^t-1)} = e^{\gamma(e^t-1)}$, where $\gamma = \sum_{i=1}^{p} \theta_i$. This shows that $Z$ is a Poisson random variable with mean $\gamma$.

If $\theta$ is a Gamma random variable, denoted $\theta \sim \mathcal{G}(a, b)$, the probability density function of $\theta$ is $f(\theta \,|\, a, b) = (b^a/\Gamma(a))x^{a-1}e^{-bx}$, $\theta > 0$ and $a, b > 0$. $E[\theta] = a/b$ and $\mathrm{Var}(\theta) = a/b^2$. Provided that $a > 1$, the expectation of $\theta^{-1}$ is

$$\begin{aligned} E[\theta^{-1}] &= \frac{b^a}{\Gamma(a)} \int_0^\infty \theta^{-1} x^{a-1} e^{-b\theta}\, d\theta = \frac{b^a}{\Gamma(a)} \int_0^\infty \theta^{a-2} e^{-b\theta}\, d\theta \\ &= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a-1)}{b^{a-1}} = \frac{b^a}{(a-1)\Gamma(a-1)} \frac{\Gamma(a-1)}{b^{a-1}} \\ &= \frac{b}{a-1}. \end{aligned}$$

The moment generating function of the Gamma distribution is $M_\theta(t) = E[e^{t\theta}] = (1-t/b)^{-a}$. If $\theta_1, \ldots, \theta_p$ are independent $\mathcal{G}(a,b)$, then $M_\gamma(t) = \prod_{i=1}^p (1-b/t)^{-a} = (1-b/t)^{-\sum_{i=1}^p a} = (1-b/t)^{-pa}$, which shows that $\gamma \sim \mathcal{G}(pa, b)$.

**Lemma A.1.1.** *Define $\eta_i = \theta_i/\gamma$, then $(Y_1, \ldots, Y_p) \,|\, Z$ follows the multinomial distribution for $Z$ trials with probabilities $(\eta_1, \ldots, \eta_p)$.*

*Proof.* Since the event $\{Y_1 = y_1, \ldots, Y_p = y_p\}$ is a subset of the event $\{Z = z\}$, $\{Y_1 = y_1, \ldots, Y_p = y_p\} \cap \{Z = z\} = \{Y_1 = y_1, \ldots, Y_p = y_p\}$. Thus

$$\begin{aligned} P(Y_1, \ldots, Y_p \,|\, Z) &= \frac{P(\{Y_1, \ldots, Y_p\} \cap \{Z\})}{P(Z)} = \frac{P(Y_1, \ldots, Y_p)}{P(Z)} \\ &= \frac{P(Y_1) \cdots P(Y_p)}{P(Z)} = \frac{z!}{y_1! \cdots y_p!} \frac{\theta_1^{y_1} \cdots \theta_p^{y_p} e^{\theta_1} \cdots e^{\theta_p}}{\gamma^Z e^{-\gamma}} \\ &= \frac{z!}{y_1! \cdots y_p!} \frac{\theta_1^{y_1} \cdots \theta_p^{y_p}}{\gamma^{\sum_i y_i}} = \frac{z!}{y_1! \cdots y_p!} \eta_1^{y_1} \cdots \eta_p^{y_p}, \end{aligned}$$

which is the probability mass function of the Multinomial$(Z, \eta_1, \ldots, \eta_p)$. □

When $(Y_1, \ldots, Y_p) \,|\, Z$ is Multinomial$(Z, \eta_1, \ldots, \eta_p)$ the expectation is $E[Y_i \,|\, Z] = Z\eta_i$ and the variance is $\mathrm{Var}(Y_i \,|\, Z) = Z\eta_i(1-\eta_i)$.

A result used many times throughout the thesis is the following: Let the Poisson random variable $Y$ have mean $\theta$ and assume that $\theta \sim \mathcal{G}(a,b)$. Then the marginal distribution of $Y$ is

$$\begin{aligned} m(y \,|\, a, b) &= \int_0^\infty \frac{b^a}{y!\Gamma(a)} \theta^{a+y-1} e^{-(b+1)} \, d\theta = \frac{b^a}{y!\Gamma(a)} \frac{\Gamma(a+y)}{(b+1)^{a+y}} \\ &= \frac{\Gamma(a+y)}{y!\Gamma(a)} \left(1 - \frac{1}{b+1}\right)^a \left(\frac{1}{b+1}\right)^y. \end{aligned}$$

This is a Negative binomial distribution with parameters $a$ and $(b+1)^{-1}$. Its expectation is $E^m[Y] = a/b$ and its variance is $\mathrm{Var}_m(Y) = a/b(1+1/b)$.

## A.2 Minimaxity of the MLE under $L_1$

In this section I prove that $Y$ is minimax under $L_1$ by showing that the MBR$(\pi)$ converges to $R(Y, \theta)$ for a given sequence of priors when loss is $L_1$, cf. Lemma 2.1.1.

*Proof.* (Minimaxity of MLE under $L_1$). Define $\mu_{-1} = (a-1)/b$. Then the Bayes solution under $L_1$ is $w\mu_{-1} + (1-w)y_i = y_i - w(y_i - \mu_{-1})$. The minimum Bayes risk is then

$$\begin{aligned}\mathrm{MBR}(\pi) = E^{\pi}E_{\theta}\sum_{i=1}^{p}\frac{1}{\theta_i}(y_i - w(y_i-\mu_{-1})-\theta)^2 &= p(1-w)^2 + \sum_{i=1}^{p}E^{\pi}\frac{1}{\theta_i}(\mu_{-1}-\theta_i)^2 \\ &= p\frac{1}{(b+1)^2} + pE^{\pi}\left\{\frac{a-1}{b} - 2\frac{a-1}{b} + \frac{a}{b}\right\} = \frac{p}{b+1}.\end{aligned}$$

With the sequence of priors $\pi_n = \mathcal{G}(a, b_n)$, $b_n = 1/n$ the $\mathrm{MBR}(\pi)$ converges to $p$ as $n \to \infty$. □

Under the squared error loss function the MLE has risk $R(Y,\theta) = \sum_{i=1}^{p}\theta_i$. Since $\Theta = [0,\infty)$ the supremum over this risk is not finite. In effect, there exists no estimator $\delta^c$ for which $\sup_\theta R(\delta,\theta)$ is finite. Imagine (a game theoretic situation) where the statistician plays against an intelligent opponent that controls the state of nature and desires to maximize the loss of the statistician (see e.g. Berger (1985, 308-309)). It is then easy to see that there exists no strategy (estimator) $\delta$ for which $E_\theta \sum_{i=1}^{p}(\delta_i-\theta_i)^2$ is finite.

In Section 1.3 I wrote that since the MLE is minimax under $L_1$, it is hard to find estimators that substantially improve on it over the entire parameter space. This comment does to a certain extent apply to the MLE under $L_0$ also, because it is the minimum variance unbiased estimator. The risk of any biased competitor $\delta$ is

$$R(\delta,\theta) = \sum_{i=1}^{p}\mathrm{Var}_\theta(\delta_i) + \sum_{i=1}^{p}\mathrm{bias}_\theta^2(\delta_i).$$

It is hard to get a small $\sum_{i=1}^{p}\mathrm{bias}_\theta^2(\delta_i)$ if the $\theta_i$ are very spread out, while the MLE always has bias equal to zero. This is in some sense parallel to the minimaxity of the MLE under $L_1$.

## A.3 Bayes estimator under $c$-Loss

In Section 2.1 I derived the equations

$$\delta_j\left(E[\theta_j^{-1}\,|\,Y] + cE[\gamma^{-1}\,|\,Y]\right) + cE[\gamma^{-1}\,|\,Y]\sum_{\{i:i\neq j\}}\delta_i = 1 + c, \qquad \text{(A.1)}$$

which must hold for all $j = 1, \ldots, p$. Writing this system of equations on matrix form and row reduce gives

$$\begin{bmatrix} E[\theta_1^{-1}] + cE[\gamma^{-1}] & cE[\gamma^{-1}] & \cdots & cE[\gamma^{-1}] & 1+c \\ cE[\gamma^{-1}] & E[\theta_2^{-1}] + cE[\gamma^{-1}] & \cdots & cE[\gamma^{-1}] & 1+c \\ \vdots & cE[\gamma^{-1}] & \ddots & \vdots & \vdots \\ cE[\gamma^{-1}] & cE[\gamma^{-1}] & \cdots & E[\theta_p^{-1}] + cE[\gamma^{-1}] & 1+c \end{bmatrix}$$
$$\sim \begin{bmatrix} E[\theta_1^{-1}] + cE[\gamma^{-1}] & cE[\gamma^{-1}] & \cdots & cE[\gamma^{-1}] & 1+c \\ -E[\theta_1^{-1}] & E[\theta_2^{-1}] & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ -E[\theta_1^{-1}] & 0 & \cdots & E[\theta_p^{-1}] & 0 \end{bmatrix}.$$

From this row reduction we see that all the estimators are proportional, and are on the form

$$\delta_i = E[\theta_j^{-1}]/E[\theta_i^{-1}]\delta_j, \tag{A.2}$$

for all $i = 1, \ldots, p$, $j = 1, \ldots, p$. Inserting this in (A.1) yields

$$\begin{aligned} &\delta_j \left(E[\theta_j^{-1} \,|\, Y] + cE[\gamma^{-1} \,|\, Y]\right) + cE[\gamma^{-1} \,|\, Y] \sum_{\{i: i \neq j\}} \delta_i = \delta_j E[\theta_j^{-1} \,|\, Y] + cE[\gamma^{-1} \,|\, Y] \sum_{i=1}^{p} \delta_i \\ &= \delta_j E[\theta_j^{-1} \,|\, Y] + cE[\gamma^{-1} \,|\, Y] \delta_j E[\theta_j^{-1} \,|\, Y] \sum_{i=1}^{p} \{E[\theta_i^{-1} \,|\, Y]\}^{-1} \\ &= \delta_j E[\theta_j^{-1} \,|\, Y] \left(1 + cE[\gamma^{-1} \,|\, Y] \sum_{i=1}^{p} \{E[\theta_i^{-1} \,|\, Y]\}^{-1}\right) = 1 + c. \end{aligned}$$

Solve for $\delta_j$ to obtain the Bayes solution in (2.2).

## A.4 Minimaxity of the MLE under $c$-Loss

Here are the calculations involved in the proof of Theorem 2.1.2. The Bayes estimator we are considering is

$$\delta_j^B(Y) = \frac{(1+c)(p-1+z)}{p-1+(1+c)z} \frac{y_j}{b_n+1} = \psi(z) \frac{y_j}{b_n+1},$$

which defines the function $\psi$. In the following I drop the subscript on $b_n$ to ease notation. To obtain the expression for $\mathrm{MBR}(\pi_n)$ in (2.4) I write the risk of $\delta_j^B$ under $L_c$ as a sum $R = E_\theta L_c = E_\theta S_1 + cE_\theta S_2$. Look at the expectation of $S_1$

and use Lemma A.1.1, then

$$\begin{aligned}
E_\theta S_1 &= E_\theta \sum_{i=1}^{p} \theta_i^{-1} \left( \psi(Z) \frac{Y_i}{b+1} - \theta_i \right)^2 \\
&= E_\theta E \left\{ \sum_{i=1}^{p} \theta_i^{-1} \left( \psi^2(Z) \frac{Y_i^2}{(b+1)^2} - 2\psi(Z) \frac{Y_i}{b+1} \theta_i + \theta_i^2 \right) \mid Z \right\} \\
&= E_\theta \sum_{i=1}^{p} \theta_i^{-1} \left( \psi^2(Z) \frac{E[Y_i^2|Z]}{(b+1)^2} - 2\psi(Z) \frac{E[Y_i|Z]}{b+1} \theta_i + \theta_i^2 \right) \\
&= E_\theta \sum_{i} (\eta_i \gamma)^{-1} \left( \psi^2(Z) \frac{[Z\eta_i(1-\eta_i) + Z^2\eta_i^2]}{(b+1)^2} - 2\psi(Z) \frac{Z\eta_i}{b+1} \theta_i + \theta_i^2 \right) \\
&= E_\theta \left\{ \psi^2(Z) \frac{Z(p-1+Z)}{\gamma\,(b+1)^2} - 2\psi(Z) \frac{Z}{b+1} + \gamma \right\}.
\end{aligned}$$

The expectation of the second term $cS_2$ is

$$\begin{aligned}
E_\theta c S_2 = E_\theta \frac{c}{\gamma} \left( \sum_{i=1}^{p} \psi(Z) \frac{Y_i}{b+1} - \gamma \right)^2 &= E_\theta \frac{c}{\gamma} \left( \psi(Z) \frac{Z}{b+1} - \gamma \right)^2 \\
&= E_\theta \left\{ c\psi^2(Z) \frac{Z^2}{\gamma\,(b+1)^2} - 2c\,\psi(Z) \frac{Z}{b+1} + c\gamma \right).
\end{aligned}$$

Putting the two terms together and inserting the expression for $\psi$ we obtain

$$\begin{aligned}
R(\delta^B, \theta) &= E_\theta S_1 + cS_2 \\
&= E_\theta \left\{ \psi^2(Z) \frac{Z(p-1+(1+c)Z)}{\gamma\,(b+1)^2} - 2(1+c)\psi(Z) \frac{Z}{b+1} + (1+c)\gamma \right\} \\
&= E_\theta \left\{ \frac{(1+c)^2(p-1+Z)^2}{p-1+(1+c)Z} \frac{Z}{\gamma\,(b+1)^2} \right. \\
&\qquad \left. -2 \frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z} \frac{Z}{b+1} + (1+c)\gamma \right\}.
\end{aligned}$$

Now, use that the MBR($\pi$) can be expressed as

$$E^\pi R(\delta^B, \theta) = E^m \left[ E^{\pi^*} \left[ L(\delta^B, \theta) \mid Z \right] \right]$$

where $E^\pi$ is the expectation with respect to the prior distribution on $\theta_i$, $E^m$ is the expectation taken over the marginal distribution of all the data, and $E^{\pi^*}$ is the expectation over the posterior distribution of $\gamma$ given all the data. Since

$\gamma \mid Z$ is distributed $\mathcal{G}(p+z, b+1)$ we get that

$$\begin{aligned}
\mathrm{MBR}(\pi) &= E^m\left[E^{\pi^*}\left[L(\delta^B,\theta)\mid Z\right]\right] \\
&= E^m\left\{\frac{(1+c)^2(p-1+Z)^2}{p-1+(1+c)Z}\frac{Z}{(b+1)^2}E^{\pi^*}[\gamma^{-1}\mid Z]\right. \\
&\qquad\left. -\,2\frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z}\frac{Z}{b+1}+(1+c)E^{\pi^*}[\gamma\mid Z]\right\} \\
&= E^m\left\{\frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z}\frac{Z}{b+1}\right. \\
&\qquad\left. -2\frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z}\frac{Z}{b+1}+(1+c)\frac{p+Z}{b+1}\right\} \\
&= E^m\left\{(1+c)\frac{p+Z}{b+1}-\frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z}\frac{Z}{b+1}\right\}.
\end{aligned}$$

Rearranging the expression we are taking the expectation over gives

$$\begin{aligned}
\mathrm{MBR}(\pi) &= E^m\frac{1+c}{b+1}\left\{p+Z-\frac{(1+c)(p-1+Z)}{p-1+(1+c)Z}Z\right\} \\
&= E^m\frac{1+c}{b+1}\left\{p+Z\left(1-\frac{(1+c)(p-1+Z)}{p-1+(1+c)Z}\right)\right\} \\
&= E^m\frac{1+c}{b+1}\left\{p-\frac{c(p-1)Z}{p-1+(1+c)Z}\right\}.
\end{aligned}$$

Define the intergrand in this expression as $h(z)$. Its second derivative is

$$\frac{d^2}{d\,z^2}h(z) = \frac{d}{d\,z}\left\{\frac{d}{d\,z}\frac{c(p-1)^2}{(p-1+(1+c)z)^2}\right\} = 2\frac{(1+c)^2}{b+1}\frac{c(p-1)^2}{(p-1+(1+c)z)^3} > 0$$

for all $z \geq 0$. This shows that $h$ is convex. The marginal distribution of $Z$ is Negative binomial with parameters $p$ and $(b+1)^{-1}$, so $E^m[Z] = p/b$. Using the convexity of $h$ we have by Jensen's inequality that

$$\begin{aligned}
\mathrm{MBR}(\pi) = E^m\left[h(Z)\right] \geq h(E^m[Z]) &= \frac{(1+c)p}{b+1}-\frac{1+c}{b+1}\frac{c(p-1)E^m[Z]}{p-1+(1+c)E^m[Z]} \\
&= \frac{(1+c)p}{b+1}-\frac{1+c}{b+1}\frac{c(p-1)p/b}{p-1+(1+c)p/b} \\
&= \frac{(1+c)p}{b+1}-\frac{1+c}{b+1}\frac{cp(p-1)}{b(p-1)+(1+c)p}
\end{aligned}$$

which is the expression in (2.5) for the MBR in Theorem 2.1.2.

## A.5 Proof of crucial lemmata cont.

The proof of the identity in (2.7) is

$$\begin{aligned} E_\theta \theta\, f(Y) &= \sum_{y=0}^{\infty} f(y) \frac{1}{y!} \theta^{y+1} e^{-\theta} = 0 + \sum_{y=1}^{\infty} f(y) \frac{1}{y!} \theta^{y+1} e^{-\theta} \\ &= \sum_{y=1}^{\infty} f(y-1) \frac{y}{y!} \theta^{y} e^{-\theta} = \sum_{y=0}^{\infty} f(y-1) \frac{y}{y!} \theta^{y} e^{-\theta} = E_\theta f(Y-1) Y \end{aligned}$$

where I have used the assumption that $f(y) = 0$ for all $y \geq 0$. To prove the multivariate equivalent, condition on $\{Y_j\,;\, j \neq i\}$, then

$$\begin{aligned} E_\theta \theta_i\, F_i(Y) &= E_\theta E\left[\theta_i\, F_i(Y) \,|\, \{Y_j\,;\, j \neq i\}\right] \\ &= E_\theta\, F_i(Y - e_i) Y_i. \end{aligned}$$

## A.6 Estimation of Poisson means under $L_0$

In this section I prove that the estimator of Peng (1975) given in (1.3) dominates the MLE under $L_0$. The proofs of dominance of the generalizations of $\delta^P$ in (3.1), (3.2) and (3.3) are similar and can be found in Ghosh et al. (1983). The competitor of the MLE is given componentwise by

$$\delta_i^*(Y) = Y_i + f_i(Y),$$

where $f$ is a function that satisfies Lemma 2.2.1. Let $\Delta_i F(Y) = F(Y) - F(Y - e_i)$. Applying Lemma 2.2.1 the difference in risk between $\delta^*$ and $\delta^o$ can be expressed as

$$\begin{aligned} R(\delta^*, \theta) - R(Y, \theta) &= E_\theta \left\{ \sum_{i=1}^{p} (Y_i - f_i(Y) - \theta_i)^2 - (Y_i - \theta_i)^2 \right\} \\ &= E_\theta \sum_{i=1}^{p} \left\{ f_i^2(Y) + 2 f_i(Y)(Y_i - \theta_i) \right\} = E_\theta \sum_{i=1}^{p} \left\{ 2 Y_i \Delta_i f_i(Y) + f_i^2(Y) \right\}. \end{aligned}$$

So $R(\delta^*, \theta) - R(Y, \theta) = 2 E_\theta D_0(Y)$ where

$$D_0(Y) = \sum_{i=1}^{p} \left\{ Y_i \Delta_i f_i(Y) + \frac{1}{2} f_i^2(Y) \right\}.$$

Thus, if $D_0(Y) \leq 0$ with strict inequality for at least one datum, then $\delta^*$ dominates the ML-estimator.

*Proof.* (Peng (1975)) Let $D_i = D(Y - e_i)$. The estimator $\delta^P(Y) = Y + f(Y)$ is defined by

$$f_i(Y) = -\frac{(N_0(Y) - 2)^+}{D(Y)} h_i(Y_i),$$

for $i = 1, \dots, p$. The case $N_0(Y) \leq 2$ is trivial because then $D_0(Y) = 0$. Assume that $N(Y) > 2$. Then

$$\frac{1}{2}\sum_{i=1}^{p} f_i^2(Y) = \frac{1}{2}\sum_{i=1}^{p} \frac{(N_0(Y) - 2)^2}{D^2} h_i^2(Y_i) = \frac{1}{2}\frac{(N_0(Y) - 2)^2}{D} \leq \frac{(N_0(Y) - 2)^2}{D},$$

with a strict inequality when $N_0(Y) > 2$. Furthermore,

$$\begin{aligned}\Delta_i f_i(Y) &= -\frac{(N_0(Y) - 2)^+}{D(Y)} h_i(Y_i) + \frac{(N_0(Y - e_i) - 2)^+}{D(Y - e_i)} h_i(Y_i - 1) \\ &\leq (N_0(Y - e_i) - 2)^+ \Delta_i \frac{h_i(Y_i)}{D(Y)},\end{aligned}$$

since $(N_0(Y - e_i) - 2)^+ \leq (N_0(Y) - 2)^+$. And

$$\begin{aligned}-\Delta_i \frac{h_i(Y_i)}{D(Y)} &= -\frac{h_i(Y_i)}{D} + \frac{h_i(Y_i - 1)}{D_i} \\ &= -\frac{h_i(Y_i)}{D} + \frac{h_i(Y_i - 1)}{D} + \frac{h_i(Y_i - 1)D}{D_i D} - \frac{h_i(Y_i - 1)D_i}{DD_i} \\ &= -\frac{\Delta_i h_i(Y_i)}{D} + \frac{h_i(Y_i - 1)\Delta_i D}{DD_i}.\end{aligned}$$

Then

$$\begin{aligned}\sum_{i=1}^{p} Y_i \Delta_i f_i(Y) &= \sum_{i=1}^{p} \left\{ Y_i (N_0(Y) - 2)^+ \left( -\frac{\Delta_i h_i(Y_i)}{D} + \frac{h_i(Y_i - 1)\Delta_i D}{DD_i} \right) \right\} \\ &= \frac{(N_0(Y) - 2)^+}{D} \sum_{i=1}^{p} \left\{ -Y_i \Delta_i h_i(Y) + Y_i \frac{h_i(Y_i - 1)\Delta_i D}{D_i} \right\} \\ &\leq \frac{(N_0(Y) - 2)^+}{D} \sum_{i=1}^{p} \left\{ -N_0(Y) + Y_i \frac{h_i(Y_i - 1)\Delta_i h_i^2(Y_i)}{D_i} \right\},\end{aligned}$$

since $\sum_{i=1}^{p} Y_i \Delta_i h_i(Y) = p \leq N_0(Y)$ and $\Delta_i D(Y) = D(Y) - D(Y - e_i) = \sum_{j=1}^{p} [h_j^2(Y_j) - h_j^2(Y_j - I(j = i))] = \Delta_i h_i(Y_i)$. Since

$$\begin{aligned}\Delta_i h_i^2(Y_i) &= \left( \sum_{k=1}^{Y_i} \frac{1}{k} \right)^2 - \left( \sum_{k=1}^{Y_i - 1} \frac{1}{k} \right)^2 = \left( \sum_{k=1}^{Y_i} \frac{1}{k} \right)^2 - \left( \sum_{k=1}^{Y_i} \frac{1}{k} - \frac{1}{Y_i} \right)^2 \\ &= 2\frac{1}{Y_i} \sum_{k=1}^{Y_i} \frac{1}{k} - \frac{1}{Y_i^2},\end{aligned}$$

we get that

$$\begin{aligned}Y_i h_i(Y_i-1)\Delta_i(Y_i) &= h_i(Y_i-1)\left[2h(Y_i)-1/Y_i\right] = h_i(Y_i-1)\left[2h(Y_i-1)+1/Y_i\right]\\ &\leq h_i(Y_i-1)\left[2h(Y_i-1)+2/Y_i\right] = 2h_i^2(Y_i-1).\end{aligned}$$

Setting all this together gives

$$\begin{aligned}D_0(Y) &\leq \frac{(N_0(Y)-2)^+}{D}\sum_{i=1}^{p}\left\{-N_0(Y)+Y_i\frac{h_i(Y_i-1)\Delta_i h_i^2(Y_i)}{D_i}\right\}+\frac{(N_0(Y)-2)^2}{D}\\ &= -\frac{(N_0(Y)-2)^+}{D}\left\{N_0(Y)-\sum_{i=1}^{p}Y_i\frac{h_i(Y_i-1)\Delta_i h_i^2(Y_i)}{D_i}-(N_0(Y)-2)^+\right\}\\ &\leq -\frac{(N_0(Y)-2)^+}{D}\left\{N_0(Y)-\sum_{i=1}^{p}Y_i\frac{2h_i^2(Y_i-1)}{D_i}-(N_0(Y)-2)^+\right\}\\ &\leq -\frac{(N_0(Y)-2)^+}{D}\left\{N_0(Y)-2-(N_0(Y)-2)^+\right\}\leq 0,\end{aligned}$$

provided that $N_0(Y)\geq 3$. The inequality is strict when $(N_0(Y)-2)^+h_i(Y_i-1)\Delta_i h_i^2(Y_i)>0$ for at least two observations. In other words, $N_0(Y)\geq 3$ and $N_1(Y)\geq 2$. □

## A.7 Comparing $\delta^{CZ}$ and $\delta_1^c$ and finding optimal $c$

Table A.1 is a reproduction of Table 1 in Clevenson and Zidek (1975, 704) extended to include the results of using the new estimator $\delta_1^c$ in (2.9). This table is the basis for the empirical comparison of $\delta^{CZ}$ and $\delta_1^c$ referred to in Section 2.4. Below is Table A.2 referred to in Section 2.5, providing a summary of optimal $c$-values.

| i | $y_i$ | $\theta_i$ | $\delta_i^{cz}$ | $\delta_i^c$ | $(y_i - \theta_i)^2/\theta_i$ | $(\delta_i^{cz} - \theta_i)^2/\theta_i$ | $(\delta_i^c - \theta_i)^2/\theta_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1.17 | 0 | 0 | 1.17 | 1.17 | 1.17 |
| 2 | 0 | 0.83 | 0 | 0 | 0.83 | 0.83 | 0.83 |
| 3 | 0 | 0.5 | 0 | 0 | 0.5 | 0.5 | 0.5 |
| 4 | 1 | 1 | 0.45 | 0.67 | 0 | 0.3 | 0.11 |
| 5 | 2 | 0.83 | 0.91 | 1.33 | 1.65 | 0.01 | 0.31 |
| 6 | 1 | 0.83 | 0.45 | 0.67 | 0.03 | 0.17 | 0.03 |
| 7 | 0 | 1.17 | 0 | 0 | 1.17 | 1.17 | 1.17 |
| 8 | 2 | 0.83 | 0.91 | 1.33 | 1.65 | 0.01 | 0.31 |
| 9 | 0 | 0.67 | 0 | 0 | 0.67 | 0.67 | 0.67 |
| 10 | 0 | 0.17 | 0 | 0 | 0.17 | 0.17 | 0.17 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 1 | 0.33 | 0.45 | 0.67 | 1.36 | 0.05 | 0.34 |
| 13 | 3 | 1.5 | 1.36 | 2 | 1.5 | 0.01 | 0.17 |
| 14 | 0 | 0.5 | 0 | 0 | 0.5 | 0.5 | 0.5 |
| 15 | 0 | 1.17 | 0 | 0 | 1.17 | 1.17 | 1.17 |
| 16 | 3 | 1.33 | 1.36 | 2 | 2.1 | 0 | 0.34 |
| 17 | 0 | 0.5 | 0 | 0 | 0.5 | 0.5 | 0.5 |
| 18 | 2 | 1.17 | 0.91 | 1.33 | 0.59 | 0.06 | 0.02 |
| 19 | 1 | 0.5 | 0.45 | 0.67 | 0.5 | 0 | 0.06 |
| 20 | 2 | 0.5 | 0.91 | 1.33 | 4.5 | 0.33 | 1.39 |
| 21 | 0 | 1.33 | 0 | 0 | 1.33 | 1.33 | 1.33 |
| 22 | 0 | 0.83 | 0 | 0 | 0.83 | 0.83 | 0.83 |
| 23 | 0 | 0.33 | 0 | 0 | 0.33 | 0.33 | 0.33 |
| 24 | 1 | 1.5 | 0.45 | 0.67 | 0.17 | 0.73 | 0.46 |
| 25 | 5 | 1.33 | 2.27 | 3.33 | 10.13 | 0.66 | 3.02 |
| 26 | 0 | 0.67 | 0 | 0 | 0.67 | 0.67 | 0.67 |
| 27 | 1 | 0.67 | 0.45 | 0.67 | 0.16 | 0.07 | 0 |
| 28 | 0 | 0.33 | 0 | 0 | 0.33 | 0.33 | 0.33 |
| 29 | 0 | 0.33 | 0 | 0 | 0.33 | 0.33 | 0.33 |
| 30 | 1 | 0.33 | 0.45 | 0.67 | 1.36 | 0.05 | 0.34 |
| 31 | 0 | 0.5 | 0 | 0 | 0.5 | 0.5 | 0.5 |
| 32 | 1 | 0.83 | 0.45 | 0.67 | 0.03 | 0.17 | 0.03 |
| 33 | 0 | 0.67 | 0 | 0 | 0.67 | 0.67 | 0.67 |
| 34 | 1 | 0.33 | 0.45 | 0.67 | 1.36 | 0.05 | 0.34 |
| 35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 36 | 1 | 0.5 | 0.45 | 0.67 | 0.5 | 0 | 0.06 |
| $L_1$ | | | | | 39.26 | 14.34 | 19 |
| $L_c$ | $Z = 29$ | $\gamma = 25.98$ | | | 53.29 | 268.21 | 87.05 |

Table A.1: An empirical comparison of $\delta^{cz}$ and $\delta_1^c$ on the oil-well exploration data in Clevenson and Zidek (1975, 707). In this study the parameter $c$ was set to 40.

| | Tolerance level $K\%$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p = 40$ | | | | | $p = 8$ | | | | |
| $\gamma$ | 2% | 5% | 10% | 15% | 20% | 2% | 5% | 10% | 15% | 20% |
| 1 | 7 | 5 | 3 | 2 | 1 | 200 | 200 | 200 | 200 | 200 |
| 2 | 23 | 17 | 12 | 9 | 7 | 3 | 2 | 1 | 0 | 0 |
| 3 | 43 | 28 | 19 | 15 | 12 | 6 | 4 | 2 | 1 | 1 |
| 4 | 59 | 37 | 25 | 19 | 16 | 9 | 5 | 3 | 2 | 2 |
| 5 | 68 | 41 | 27 | 21 | 18 | 11 | 6 | 4 | 3 | 2 |
| 6 | 72 | 43 | 29 | 22 | 18 | 12 | 7 | 4 | 3 | 2 |
| 7 | 73 | 44 | 29 | 23 | 19 | 12 | 7 | 4 | 3 | 2 |
| 8 | 72 | 44 | 29 | 23 | 19 | 12 | 7 | 4 | 3 | 2 |
| 9 | 71 | 43 | 29 | 22 | 19 | 12 | 6 | 4 | 3 | 2 |
| 10 | 70 | 42 | 28 | 22 | 18 | 11 | 6 | 4 | 3 | 2 |
| 11 | 68 | 41 | 28 | 22 | 18 | 11 | 6 | 4 | 3 | 2 |
| 12 | 66 | 40 | 27 | 21 | 18 | 11 | 6 | 4 | 3 | 2 |
| 13 | 64 | 39 | 26 | 21 | 17 | 10 | 6 | 3 | 2 | 2 |
| 14 | 63 | 38 | 26 | 20 | 17 | 10 | 6 | 3 | 2 | 2 |
| 15 | 61 | 37 | 25 | 20 | 17 | 10 | 5 | 3 | 2 | 2 |
| 16 | 60 | 36 | 24 | 19 | 16 | 9 | 5 | 3 | 2 | 2 |
| 17 | 58 | 35 | 24 | 19 | 16 | 9 | 5 | 3 | 2 | 2 |
| 18 | 57 | 35 | 23 | 18 | 15 | 9 | 5 | 3 | 2 | 2 |
| 19 | 56 | 34 | 23 | 18 | 15 | 9 | 5 | 3 | 2 | 1 |
| 20 | 55 | 33 | 22 | 18 | 15 | 9 | 5 | 3 | 2 | 1 |
| 21 | 53 | 33 | 22 | 17 | 15 | 8 | 5 | 3 | 2 | 1 |
| 22 | 52 | 32 | 22 | 17 | 14 | 8 | 5 | 3 | 2 | 1 |
| 23 | 51 | 31 | 21 | 17 | 14 | 8 | 4 | 3 | 2 | 1 |
| 24 | 50 | 31 | 21 | 16 | 14 | 8 | 4 | 3 | 2 | 1 |
| 25 | 50 | 30 | 20 | 16 | 14 | 8 | 4 | 2 | 2 | 1 |
| 26 | 49 | 30 | 20 | 16 | 13 | 8 | 4 | 2 | 2 | 1 |
| 27 | 48 | 29 | 20 | 16 | 13 | 7 | 4 | 2 | 2 | 1 |
| 28 | 47 | 29 | 19 | 15 | 13 | 7 | 4 | 2 | 2 | 1 |
| 29 | 46 | 28 | 19 | 15 | 13 | 7 | 4 | 2 | 1 | 1 |
| 30 | 46 | 28 | 19 | 15 | 13 | 7 | 4 | 2 | 1 | 1 |
| 31 | 45 | 27 | 19 | 15 | 12 | 7 | 4 | 2 | 1 | 1 |
| 32 | 44 | 27 | 18 | 14 | 12 | 7 | 4 | 2 | 1 | 1 |
| 33 | 44 | 27 | 18 | 14 | 12 | 7 | 4 | 2 | 1 | 1 |
| 34 | 43 | 26 | 18 | 14 | 12 | 7 | 3 | 2 | 1 | 1 |
| 35 | 43 | 26 | 18 | 14 | 12 | 6 | 3 | 2 | 1 | 1 |
| 36 | 42 | 26 | 17 | 14 | 11 | 6 | 3 | 2 | 1 | 1 |
| 37 | 41 | 25 | 17 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 38 | 41 | 25 | 17 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 39 | 40 | 25 | 17 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 40 | 40 | 24 | 16 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 41 | 39 | 24 | 16 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 42 | 39 | 24 | 16 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 43 | 39 | 24 | 16 | 13 | 11 | 6 | 3 | 2 | 1 | 1 |
| 44 | 38 | 23 | 16 | 12 | 10 | 6 | 3 | 2 | 1 | 1 |
| 45 | 38 | 23 | 16 | 12 | 10 | 6 | 3 | 2 | 1 | 1 |
| 46 | 37 | 23 | 15 | 12 | 10 | 5 | 3 | 2 | 1 | 1 |
| 47 | 37 | 23 | 15 | 12 | 10 | 5 | 3 | 2 | 1 | 1 |
| 48 | 37 | 22 | 15 | 12 | 10 | 5 | 3 | 1 | 1 | 1 |
| 49 | 36 | 22 | 15 | 12 | 10 | 5 | 3 | 1 | 1 | 1 |
| 50 | 36 | 22 | 15 | 12 | 10 | 5 | 3 | 1 | 1 | 1 |

Table A.2: Optimal values of $c$ for $p = 8$ and $p = 40$ for varying prior guesses of $\gamma$. For simplicity, the $c$-values were restricted to $\mathbb{N} \cup \{0\}$. They were computed with the `R`-script in Appendix C.2.

# Appendix B

# MCMC for the Poisson regression model

In order to draw samples from the joint posterior distribution $\{\theta_i\}_{i=1}^p, \beta, c \,|\, Y$ given in (4.3) I rely on Markov Chain Monte Carlo methods, particularly the Gibbs sampler (see e.g. Robert and Casella (2010)). The Gamma distribution of $\{\theta_i\}_{i=1}^p \,|\, \beta, c,$ data in (4.4) is straightforward to sample from since `R` and most other statistical software include routines for sampling from standard distributions. The full conditional distributions given in (4.5) and (4.6), on the other hand, are not standard. Therefore I implement two Metropolis-Hastings (MH) algorithms in order to draw samples from these.

The MH-algorithm for $\beta \,|\, \{\theta_i\}_{i=1}^p, c, y$ in (4.5) is explained in Algorithm (B.1).

```
Given $\beta_{(n)} = (\beta_{1,(n)}, \ldots, \beta_{k,(n)})$
1.  Draw
      $\beta_{(n+1)} \sim N_k(\beta_{(n+1)} - \beta_{(n)}, I_k)$
2.  Take
      $\beta_{n+1} = \begin{cases} \beta_{(n+1)} & \texttt{with probability } A(\beta_{(n+1)}, \beta_{(n)}) \\ \beta_{(n)} & \texttt{with probability } 1 - A(\beta_{(n+1)}, \beta_{(n)}) \end{cases}$        (B.1)
  where,
      $A(\beta_{(n+1)}, \beta_{(n)}) = \frac{\pi(\beta_{(n+1)} \,|\, \{\theta_i\}_{i=1}^p, c, y)}{\pi(\beta_{(n)} \,|\, \{\theta_i\}_{i=1}^p, c, y)}$
```

The sampling from the full conditional distribution $c \,|\, \{\theta_i\}_{i=1}^p, \beta, y$ works in the samme manner.

```
Given $c_{(n)}$
1.  Draw
        $c_{(n+1)} \sim \frac{N(c_{(n+1)} - c_{(n)}, 1)}{\Phi(c_{(n)})}$
2.  Take                                                                        (B.2)
        $c_{(n+1)} = \begin{cases} c_{(n+1)} & \text{with probability } A(c_{(n+1)}, c_{(n)}) \\ c_{(n)} & \text{with probability } 1 - A(c_{(n+1)}, c_{(n)}) \end{cases}$
  where,
        $A(c_{(n+1)}, c_{(n)}) = \frac{\pi(c_{(n+1)} \mid \{\theta_i\}_{i=1}^p, \beta, y)}{\pi(c_{(n)} \mid \{\theta_i\}_{i=1}^p, \beta, y)} \frac{\Phi(c_{(n)})}{\Phi(c_{(n+1)})}$
```

where $\Phi(\cdot)$ is the standard normal cumulative density function. Finally, these two MH-algorithms are used in the Gibbs-sampler described in (B.3).

```
Given $(\beta_{(0)}, c_{(0)})$
For $n = 1, \ldots, N$
1.  Draw
        $\theta_{i,(n+1)} \sim \mathcal{G}(\theta_{i,0}^{(n)}/c_{(n)} + y_i, 1/c_{(n)} + t_i)$, for $i = 1, \ldots, p$
2.  Draw                                                                        (B.3)
        $\beta_{(n+1)} \sim \pi(\beta \mid \{\theta_i\}_{i=1,(n+1)}^p, c_{(n)}, \text{data})$
3.  Draw
        $c_{(n+1)} \sim \pi(c \mid \{\theta_i\}_{i=1,(n+1)}^p, \beta_{(n+1)}, \text{data})$
```

## B.1 The MCMC-algorithm implemented in R

Below are the R-scripts implementing this Gibbs-sampler. The R-function called MH_beta() implements the MH-algorithm in (B.1), while the function MH_b() implements the MH-algorithm in (B.2). Finally, these two algorithms are used in the Gibbs-sampler, given in the last R-script and called gibbs().

```r
require(compiler)
enableJIT(3)
## A Gibbs sampler.
##
MH_beta <- function(sims,Bprior,Sigma,Z,theta,b){
  #--------------------------------
  # Bprior = c(\beta_1,...,\beta_p) is the
  # prior mean of the regression coefficients
  # Sigma is the k/times k prior covariance matrix.
  # Z is the design matrix. Rows = p k\times 1
  # vectors z_i^t
  #--------------------------------
  k <- length(Bprior) # number of regression coeffs.
  out <- matrix(NA,nrow=sims,ncol=k)
  Bold <- Bprior # use prior as start value
  for(i in 1:sims){
    # symmetric proposal
    Bprop <- Bold + rnorm(k,0,1)
    # make A (the probability) on log scale
    mu.prop <- exp(Z%*%Bprop) # \mu_i=\exp(z_i^t\beta)
    mu.old <- exp(Z%*%Bold)
    l.A.nom <- sum(log(dgamma(theta,b*mu.prop,b))) + sum(log(
   dnorm(Bprop,Bprior,diag(Sigma))))
    l.A.denom <- sum(log(dgamma(theta,b*mu.old,b))) + sum(log
   (dnorm(Bold,Bprior,diag(Sigma))))
    #
    A <- exp(l.A.nom-l.A.denom)
    accept <- rbinom(1,1,min(A,1))
    Bold <- accept*Bprop + (1-accept)*Bold
    out[i,] <- Bold
    }
    return(out)
}
##
MH_b <- function(sims,bstart,zeta,eta,B,theta,Z){
  #--------------------------------
  # bstart is the start value of b
  # (zeta,eta) are the prior parameters
  # B = c(\beta_1,...,\beta_p) is the coefficent vector
  # Z is the design matrix. Rows = p k\times 1
  # vectors z_i^t
  #--------------------------------
  out <- numeric(sims)
  p <- length(theta) # number of observations
  b.old <- bstart
  # make A (the probability)
  mu <- exp(Z%*%B)
  mu.bar <- mean(mu)
  theta.bar <- mean(theta)
  W <- sum(mu*log(theta)) # \sum_i \mu_i\log\theta_i
  for(i in 1:sims){
    # non-symmetric proposal (b.prop > 0)
    repeat{
      b.prop <- b.old + rnorm(1,0,1)
      if(b.prop>0){
        break}
    }
    # on log-scale
    l.A.nom <- sum(log(dgamma(theta,b.prop*mu,b.prop))) + log(dgamma(b.prop,zeta,eta))
    l.A.denom <- sum(log(dgamma(theta,b.old*mu,b.old))) + log(dgamma(b.old,zeta,eta))
    # correct for non-symmetric proposal
    A <- exp(l.A.nom-l.A.denom)*pnorm(b.old,0,1)/pnorm(b.prop,0,1)
    #
    accept <- rbinom(1,1,min(A,1))
    b.old <- accept*b.prop + (1-accept)*b.old
    out[i] <- b.old
  }
  return(out)
}
##
gibbs <- function(sims,y,Z,Bprior,Sigma,bstart,zeta,eta,subsims=5*10^2,subburnin=2*10^2){
  #---------------------------------
  # (i) y = (y_1,...,y_p) are the observations.
  # (ii) Z is the design matrix. Rows = p k\times 1
  # vectors z_i^t.
  # (iii) Bprior = c(\beta_1,...,\beta_p) are the prior means of the
  # regression coefficients.
  # (iv) Sigma is the k/times k prior covariance matrix.
  # (v) (zeta,eta) are the prior parameters for \pi(b) = Gamma(zeta,eta)
  #---------------------------------
  k <- length(Bprior) # number of regression coefficients/covariates
  p <- length(y) # number of observations
  out <- matrix(NA,nrow=sims,ncol=k+p+1)
  colnames(out) <- c(paste("beta",1:k,sep=""),"b",paste("theta",1:p,sep=""))
  # set coeffs. to start values
  Bstart <- t(t(Bprior)); B <- Bstart; b <- bstart
  #
  counter <- 0
  for(i in 1:sims){
    mu <- exp(Z%*%B)
    # sample \theta <- (\theta_1,...,\theta_p)
    theta <- rgamma(p,y + b*mu,b + 1)
    #
    # MH_beta(sims,Bprior,Sigma,Z,theta,b)
    B.sims <- MH_beta(subsims,Bprior,Sigma,Z,theta,b)
    B <- colMeans(B.sims[subburnin:subsims,]);rm(B.sims)
    #
    # MH_b(sims,bstart,zeta,eta,B,theta,Z)
    b.sims <- MH_b(subsims,bstart,zeta,eta,B,theta,Z)
    b <- mean(b.sims[subburnin:subsims]);rm(b.sims)
    #
    out[i,] <- c(t(B),b,theta)
    counter <- counter + 1
    if(counter%%10==0){
      cat(counter/sims*100,"%","\n")
      }
    }
    return(out)
}
####
```

## B.2 Simulation example of pure Bayes regression

Here is the R-script used for the simulation example in Figure 4.1. This script shows how the MCMC-algorithm `gibbs()` can be applied. The script also produces the MCMC convergence diagnostics plots in Figure B.1 and Figure B.2.

```
# Section 4.2 R-script
#----------------------------------------
#
# Import Gibbs-sampler and simulate data
#----------------------------------------
source("Ch4_GIBBS.R")
# simulate data
#--------
p <- 40
z <- 1:p/5
Z <- matrix(NA,ncol=2,nrow=p)
Z[,1] <- 1; Z[,2] <- z
b0 <- 0.2 ; b1 <- 0.5 ; bb <- 1/5
thetas <- rgamma(p,bb*exp(b0+b1*z),bb)
y <- rpois(p,thetas)
#----------------------------------------
#----------------------------------------
#
# Run the Gibbs-sampler
#----------------------------------------
burnin <- 2*10^3
sims <- 10^4
#
# gibbs(sims,y,Z,Bprior,Sigma,bstart,zeta,eta,subsims=5*10^2,subburnin=2*10^2)
ptm <- proc.time()
mcmc <- gibbs(sims,y,Z,Bprior=c(0,0),Sigma=diag(2),bstart=3,zeta=2,eta=3)
print(proc.time() - ptm)
```

```
# get estimates
mcmc <- mcmc[burnin:sims,]
mcmc <- data.frame(mcmc)
beta.mcmc <- colMeans(mcmc)[1:2]
b.mcmc <- colMeans(mcmc)[3]
theta.mcmc <- colMeans(mcmc)[4:ncol(mcmc)]
#
# finding HPD regions
hpd <- function(k,para){
  l <- as.numeric(quantile(para,k))
  q <- .9 + k
  u <- as.numeric(quantile(para,1-q))
  return(abs(u-l))
}
alpha <- seq(0,0.1,by=.001)
dists <- hpd(alpha,mcmc$b);a.b <-alpha[which(dists==min(dists))]
dists <- hpd(alpha,mcmc$beta1);a.beta1 <-alpha[which(dists==min(dists))]
dists <- hpd(alpha,mcmc$beta2);a.beta2 <-alpha[which(dists==min(dists))]
#
cat("beta.hat",beta.mcmc,"\n","b.hat",b.mcmc,"\n")
cat(quantile(mcmc[,1],c(a.beta1,1-a.beta1)),"\n",
    quantile(mcmc[,2],c(a.beta2,1-a.beta2)),"\n",
    quantile(mcmc[,3],c(a.b,1-a.b)),"\n")
# beta.hat 0.04549282 0.5097152
# b.hat 0.1727066
# -0.4784945 0.5477131
#  0.4305788 0.5906947
#  0.1292277 0.2229868
for(i in seq(1,39,by=2)){
  cat(sprintf("theta%s",i),theta.mcmc[i],
  sprintf("theta%s",(i+1)),theta.mcmc[i+1],"\n")}
#-----------------------------------------
#-----------------------------------------
#
# Figure 4.1
#-----------------------------------------
postscript("figure4_1.eps")
par(mfrow=c(1,1))
plot(z,y,frame.plot=FALSE,ylab="mle and shrinkage estimates")
lines(z,exp(Z%*%c(b0,b1)),lty=1)
points(z,theta.mcmc,pch=2,col="green")
lines(z,exp(Z%*%beta.mcmc),lty=2)
dev.off()
#-----------------------------------------
#-----------------------------------------
#
# MCMC diagnostics Figures B.1 and B.2
#-----------------------------------------
```

```r
emp.points <- function(x){
  u <- max(density(x)$y)
  for(i in 1:length(x)){
    segments(x[i],0,x[i],u*.02)}}
postscript("diagnostic1.eps")
par(mfrow=c(3,2))
ts.plot(mcmc[,1],xlab="iterations",main=expression(beta[1]))
plot(density(mcmc[,1]),main=expression(beta[1]));emp.points(mcmc[,1])
ts.plot(mcmc[,2],xlab="iterations",main=expression(beta[2]))
plot(density(mcmc[,2]),main=expression(beta[2]));emp.points(mcmc[,2])
ts.plot(mcmc[,3],xlab="iterations",main=expression(b))
plot(density(mcmc[,3]),main=expression(b));emp.points(mcmc[,3])
dev.off()
postscript("diagnostic2.eps")
par(mfrow=c(4,2))
for(j in c(1,20,30,40)){
ts.plot(mcmc[,j+3],xlab="iterations",main=bquote(expression(theta[.(j)])))
plot(density(mcmc[,j+3]),main=bquote(theta[.(j)]))
emp.points(mcmc[,j+3])
}
dev.off()
#------------------------------------------
```

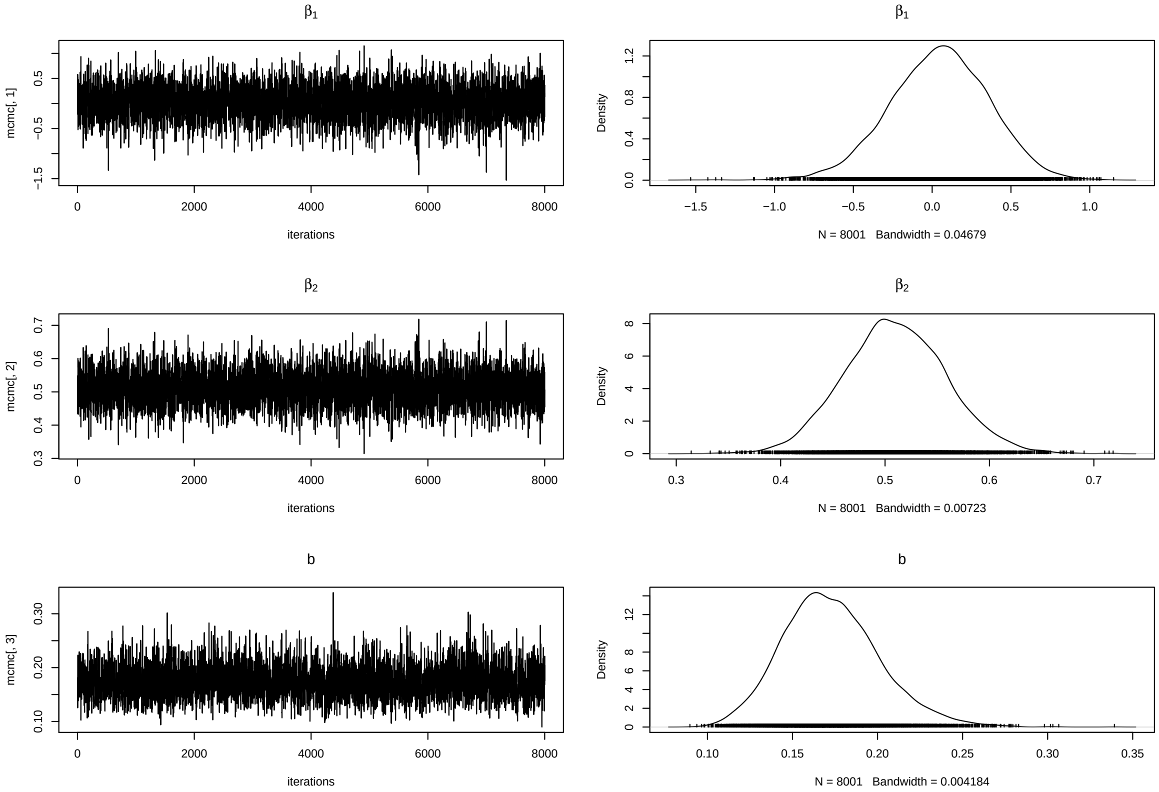

Figure B.1: Trace plots and non-parametric estimates of the densities for the parameters $\beta_1$, $\beta_2$ and $b$ in the hierarchical Bayesian regression model of Section 4.2.

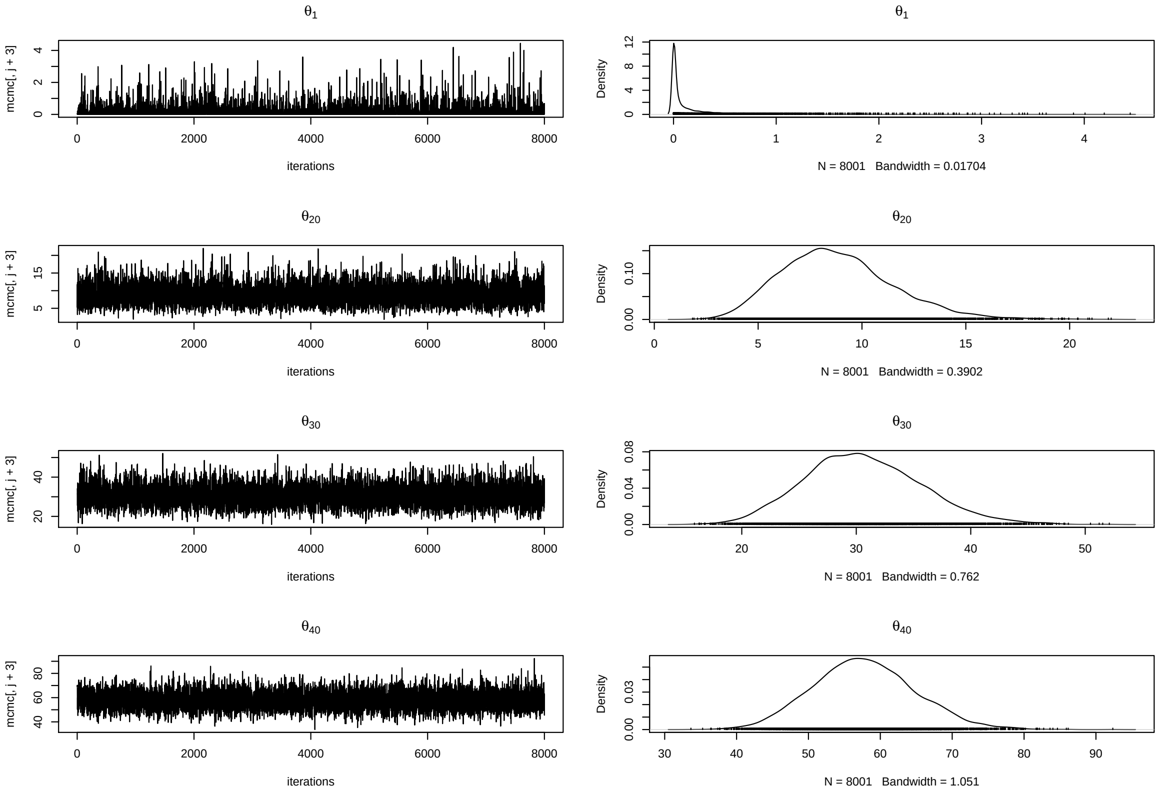

Figure B.2: Trace plots and non-parametric estimates of the densities for four of the estimated Poisson means in the hierarchical Bayesian regression model of Section 4.2.

# Appendix C

# Scripts used in simulations

In this section I have included the scripts used for the simulation studies. The statistical programming language `R` (`R` Development Core Team, 2008) is used for all the scripts.

## C.1 Script used in Section 2.4

The script below generates Table 2.1, Table 2.2 and Table 2.3 in Section 2.4. In addition, the script was used for the empirical comparison of $\delta_1^c$ and $\delta^{CZ}$ in the same section, and to make Table A.1. The data `y` and `true` are from Clevenson and Zidek (1975, 704).

```
# Section 2.4 R-script
#-----------------------------------------
#
# Figure (2.1)
#-----------------------------------------
p <- 100; c <- 40 ;z <- 0:10^3
risk <- function(gamma,p,c){
  r <- p+c+sum(((p-1)^2/(p-1+(1+c)*z) - 2*(p-1)^2/(p-1+(1+c)*z))*dpois(z,gamma))
  return(r)
}
#
risk.list1 <- c()
for(gamma in 0:100){
  risk.list1 <- append(risk.list1,risk(gamma,p,c))
}
##
postscript("figure2_1.eps")
plot(0:100,risk.list1,type="l",xlab=expression(sum(theta[i],i)==gamma),
      ylab="risk",ylim=c(min(risk.list1),p+c+15),frame.plot=FALSE)
axis(side=1); abline(p+c,0,lty=2)
```

```r
dev.off()
#-----------------------------------------
#-----------------------------------------
#
# Table (2.1) and (2.2)
#-----------------------------------------
# comparing \delta_1^c and \delta^{CZ}
cLoss <- function(est,theta,c){
  gamma <- sum(theta)
  l1 <- sum(1/theta*(est-theta)^2)
  l2 <- c/gamma*(sum(est)-sum(theta))^2
  return(l1+l2)}

# Ranges
range <- matrix(NA,nrow=6,ncol=2)
range[1,] <- c(0,4);range[2,] <- c(0,8)
range[3,] <- c(8,12);range[4,] <- c(12,16)
range[5,] <- c(0,12);range[6,] <- c(4,16)
p <- c(5,10,15)
##
L1.losses = Lc.losses = matrix(NA,nrow=18,ncol=3)
colnames(L1.losses) = colnames(Lc.losses) = c("d.cz","d.c","ml")
sims <- 10^5
L1.dcz = L1.dc = L1.ml = 0
Lc.dcz = Lc.dc = Lc.ml = 0
for(j in 1:3){
  for(i in 1:6){
    theta <- runif(p[j],range[i,1],range[i,2])
    cat(min(theta),max(theta),"\n")
    for(s in 1:sims){
      y <- rpois(p[j],theta)
      z <- sum(y)
      # estimators
      d.cz <- (1-(p[j]-1)/(p[j]-1+z))*y
      c <- 5 # c-parameter
      d.c <- (1-(p[j]-1)/(p[j]-1+(1+c)*z))*y
      # L_1-loss (c = 0)
      L1.dcz <- L1.dcz + cLoss(d.cz,theta,0)
      L1.dc <- L1.dc + cLoss(d.c,theta,0)
      L1.ml <- L1.ml + cLoss(y,theta,0)
      # L_c-loss
      Lc.dcz <- Lc.dcz + cLoss(d.cz,theta,5)
      Lc.dc <- Lc.dc + cLoss(d.c,theta,5)
      Lc.ml <- Lc.ml + cLoss(y,theta,5)

    }
    row <- c(0,6,12)
    L1.losses[i+row[j],1] <- L1.dcz/sims
    L1.losses[i+row[j],2] <- L1.dc/sims
    L1.losses[i+row[j],3] <- L1.ml/sims
    #
    Lc.losses[i+row[j],1] <- Lc.dcz/sims
    Lc.losses[i+row[j],2] <- Lc.dc/sims
    Lc.losses[i+row[j],3] <- Lc.ml/sims
  }
}
#
# percentage saved
ch2.table <- function(loss,out.name){
  pcs <- matrix(NA,nrow=18,ncol=2)
  colnames(pcs) <- c("d.cz","d.c")
  pcs[,1] <- (loss[,3] - loss[,1])/loss[,1]
  pcs[,2] <- (loss[,3] - loss[,2])/loss[,2]
  pcs <- round(pcs*100,2)
  sink(out.name)
  r <- c("(0,4)","(0,8)","(8,12)","(12,16)","(0,12)",
      "(4,16)")
  for(k in 1:6){
  cat(r[k],"&",pcs[1+(k-1),1],"&",pcs[1+(k-1),2],"&",
      pcs[7+(k-1),1],"&",pcs[7+(k-1),2],
      "&",pcs[13+(k-1),1],"&",pcs[13+(k-1),2],"\\\\","\n")
    }
  sink()
}
# write the tables
ch2.table(L1.losses,"table2_1.txt")
ch2.table(Lc.losses,"table2_2.txt")
#-----------------------------------------
#-----------------------------------------
#
# Table A.1
#-----------------------------------------
# Clevenson and Zidek (1975,704)  data
y <- c(0,0,0,1,2,1,0,2,0,0,0,1,3,0,0,
        3,0,2,1,2,0,0,0,1,5,0,1,0,0,1,
              0,1,0,1,0,1)
true <- c(1.17,0.83,0.50,1.00,0.83,0.83,
      1.17,0.83,0.67,0.17,0.00,0.33,
      1.50,0.50,1.17,1.33,0.50,1.17,
      0.50,0.50,1.33,0.83,0.33,1.50,
      1.33,0.67,0.67,0.33,0.33,0.33,
      0.50,0.83,0.67,0.33,0.00,0.50)

# make Table A.1
## make a table
p <- length(y);
c <- 40
z <- sum(y)
cz <- (1 - (p-1)/(p-1+z))*y
delta.c <- (1 - (p-1)/(p-1+(1+c)+z))*y
#
```

```r
table <- matrix(NA,38,8)
table[1:36,1] <- 1:36; table[1:36,2] <- y; table[1:36,3] <- true
table[1:36,4] <- round(cz,2); table[1:36,5] <- round(delta.c,2)
table[1:36,6] <- round((1/true)*(y-true)^2,2)
table[1:36,7] <- round((1/true)*(cz-true)^2,2)
table[1:36,8] <- round((1/true)*(delta.c-true)^2,2)
# zero is estimated by zero (replace NaN)
table[11,6:8] <- 0.00;table[35,6:8] <- 0.00
true[true==0] <- 1/10^3
table[37,] <- c("$L_1$","","","","",round(cLoss(y,true,0),2),
          round(cLoss(cz,true,0),2),
              round(cLoss(delta.c,true,0),2))
table[38,] <- c("$L_c$",sprintf("$Z = %s$",round(sum(y),2)),
          sprintf("$\\gamma = %s$",round(sum(true),2)),
          "","",round(cLoss(y,true,c),2),round(cLoss(cz,true,c),2),   round(cLoss(delta.c,true,c),2))

# write to file
sink("tableA_1.txt")
for(row in 1:38){
  out<-sprintf("%s & %s & %s & %s & %s & %s & %s & %s \\\\",table[row,1],
  table[row,2],table[row,3],table[row,4],table[row,5],table[row,6],
  table[row,7],table[row,8])
  cat(out)
  if(row == 36){
    cat("\\hline")}
  cat("\n")
}
sink()
#-----------------------------------------
#-----------------------------------------
#
# Simulations with CZ-data in Section 2.4
#-----------------------------------------
Lc.c = Lc.cz = Lc.ml = L1.c = L1.cz = L1.ml = 0
sims <- 10^5
true[true==0] <- 1/10^3
for(i in 1:sims){
  p <- 36
  y.sim <- rpois(p,true)
  z <- sum(y.sim)
  # estimators
  c <- 40
  delta.cz <- (1 - (p-1)/(p-1+z))*y.sim
  delta.c <- (1 - (p-1)/(p-1+(1+c)*z))*y.sim
  # Lc-losses
  Lc.cz <- Lc.cz + 1/sims*cLoss(delta.cz,true,c)
  Lc.c <- Lc.c + 1/sims*cLoss(delta.c,true,c)
  Lc.ml <- Lc.ml + 1/sims*cLoss(y.sim,true,c)
  #
  # L1-losses
  L1.cz <- L1.cz + 1/sims*sum(1/true*(delta.cz-true)^2)
  L1.c <- L1.c + 1/sims*sum(1/true*(delta.c-true)^2)
  L1.ml <- L1.ml + 1/sims*sum(1/true*(y.sim-true)^2)
}
cat("L1-losses","\n")
cat("CZ",(L1.ml-L1.cz)/L1.ml*100,"\n")
cat("delta.c",(L1.ml-L1.c)/L1.ml*100,"\n")
cat("Lc-losses","\n")
cat("CZ",(Lc.ml-Lc.cz)/Lc.ml*100,"\n")
cat("delta.c",(Lc.ml-Lc.c)/Lc.ml*100,"\n")
# L1-losses
# CZ 57.16046
# delta.c 6.017174
# Lc-losses
# CZ -387.0966
# delta.c 1.583304
#-----------------------------------------
#-----------------------------------------
```

## C.2 Script used in Section 2.5

This is the script used in Section 2.5 for finding an optimal value of $c$, and for the simulation study reported in Table 2.3.

```
# Finding optimal c-values (Table A.2)
#-----------------------------------------
#
phi <- function(k,c){
  return((p-1)/(p-1+(1+c)*k))
}
Ei <- function(z,c,gamma){
  D <- ((phi(z+1,c)^2 - 2*phi(z+1,c))*(z+1) + 2*phi(z,c)*z)*
    dpois(z,gamma)
  return(D)
}
c.table40_8 <- list()
sample.size <- c(40,8)
c <- 0:(2*10^2)
G <- 1:50
for(u in 1:2){
  p <- sample.size[u]
  c.table <- matrix(NA,nrow=length(G),ncol=5+1)
  c.table[,1] <- G ; counter <- 2
  for(tol in c(2,5,10,15,20)){
    c.hat <- numeric(length(G))
    for(gamma in G){
      loss.det <- numeric(length(c))
      for(j in c){
        loss.det[j] <- sum(Ei(1:10^3,j,gamma))*100
      }
    c.hat[gamma] <- which(loss.det ==  max(loss.det[loss.det<=tol]))-1
    }
    c.table[,counter] <- c.hat
    counter <- counter + 1}
  c.table40_8[[u]] <- c.table
}
sink("optimal_c.txt")
t1 <-c.table40_8[[1]];t2 <-c.table40_8[[2]]
for(r in G){
  cat(t1[r,1],"&",t1[r,2],"&",t1[r,3],"&",t1[r,4],"&",t1[r,5],"&",
    t1[r,6],"&",t2[r,2],"&",t2[r,3],"&",t2[r,4],"&",t2[r,5],
    "&",t2[r,6],"\\\\","\n")}
sink()
#-----------------------------------------
#-----------------------------------------
#
# A comparison of \delta^{CZ} and
# \delta_1^c under L_1 (Table 2.3)
#-----------------------------------------
p <- 10   # sample size
gamma <- 28 # prior guess
K <- 10   # tolerance
loss.det <- numeric(length(c))
for(j in c){
  loss.det[j] <- sum(Ei(1:10^3,j,gamma))*100
}
c.hat <- which(loss.det ==  max(loss.det[loss.det<=K]))-1
#
p <- 10
theta.s <- runif(p-2,0,2)
theta.b <- runif(2,5,8)
theta <- c(theta.s,theta.b)
sims <- 10^5
L1.cz = L1.c = numeric(p+1) # (p+1)th is total loss
c <- c.hat
for(j in 1:sims){
  y <- rpois(p,theta)
  z <- sum(y)
  d.cz <- y - (p-1)/(p-1+z)*y
  d.c <- y - (p-1)/(p-1+(1+c)*z)*y
  #
  # Loss, \delta^{CZ}
  L1.cz[1:p] <- L1.cz[1:p] + 1/sims*(1/theta*(d.cz - theta)^2)
  L1.cz[p+1] <- sum(L1.cz[1:p])
  # \delta^{c}
  L1.c[1:p] <- L1.c[1:p] + 1/sims*(1/theta*(d.c - theta)^2)
  L1.c[p+1] <- sum(L1.c[1:p])
}
## table
sink("table2_3.txt")
for(i in 1:p){
cat(round(theta[i],2),"&",round((1-L1.cz[i])*100,2),
      "&",round((1-L1.c[i])*100,2),"\\\\","\n")
}; cat("\\hline","\n")
cat("$L_1$","&",round((p-L1.cz[p+1])/p*100,2),"&",
      round((p-L1.c[p+1])/p*100,2),"\\\\")
#sink()
##
#-----------------------------------------
#-----------------------------------------
```

## C.3 Scripts used in Chapter 3

In the R-script below I implement the estimators described in Section 3.1 and Section 3.2. In addition, the script includes a function makePlot() that is used for the plots in Figure 3.1 and Figure 3.3. The script is called PoissonEstimators and is sourced into the two simulation scripts that follow.

```
#-------------------------------------------
# Squared error loss estimators L_0
#-------------------------------------------
#
# \delta^{G1}
# Ghosh et al (1983, 354) Example 2.1
#-------------
N <- function(y,nu.i){
  return(sum(y>nu.i))
}

h <- function(y.i){
  # h = \sum_{k=1}^{Y_i}1/k
  out <- 0
  if(y.i>0){
    out <- sum(1/(1:y.i))}
    return(out)
}
d.i <- function(y.i,nu.i){
  d <- 0
  if(y.i < nu.i){
    d <- (h(y.i)-h(nu.i))^2+.5*max(3*h(nu.i)-2,0)}
  else{
    d <- (h(y.i)-h(nu.i))*(h(y.i+1)-h(nu.i))}
  return(d)
}
```

```
# the estimator
delta.G1 <- function(y,nu){
  p <- length(y)
  ests <- numeric(length(y))
  D <- 0
  for(j in 1:length(y)){
    D <- D + d.i(y[j],nu[j])
  }
  if(D==0){
    # Y_i = \nu_i \forall i
    ests[1:p] <- nu}
  else{
    for(i in 1:length(y)){
      ests[i] <- y[i] - max(N(y,nu[i])-2,0)*(h(y[i])-h(nu[i])
    )/D}
  }
  return(ests)
}
##
#---------------------------------------------
# \delta^{G2}
# Ghosh et al (1983, 355) Example 2.2
#-------------
H.i <- function(y,k){
  y1 <- min(y); out <- 0; h.y1 <- 0
  if(y1 != 0){
    h.y1 <- sum(1/(1:y1))}
  if(y[k] != y1){
    out <- sum(1/(1:y[k])) - h.y1}
  return(out)
}
e.i <- function(n,i){
  e <- numeric(n);e[i] <- 1
  return(e)
}
# the estimator
deltaG2 <- function(y){
  ests <- numeric(length(y))
  p <- length(y); D <- 0
  for(j in 1:p){
    D <- D + sum(H.i(y,j)*H.i(y+e.i(p,j),j))
  }
  y1 <- min(y) # min of observations
  if(length(unique(y))==1){
    ests[1:p] <- y1
  }
  else{
    for(i in 1:p){
      ests[i] <- y[i] - max(N(y,y1)-2,0)*H.i(y,i)/D}
  }
  return(ests)
```

```
}
##
#---------------------------------------------
# \delta^{G3}
# Ghosh et al (1983, 355) Example 2.3
#-------------
deltaG3 <- function(y){
  p <- length(y)
  m <- round(median(y))
  if(sum(y<=m)<p/2){
    m <- m - 1}
  ##
  return(delta.G1(y,rep(m,p)))
}
##
#---------------------------------------------
# \delta^{P} Peng (1975)
#-------------
delta.Peng <- function(y){
  p <- length(y)
  ests <- numeric(p)
  N0 <- sum(y==0)
  const <- max(p-N0-2,0)
  H2 <- 0
  for(j in 1:p){
    H2 <- H2 + h(y[j])^2}
  if(H2==0){
    ests <- rep(0,p)}
  if(H2>0){
    for(i in 1:p){
      ests[i] <- y[i] - const*h(y[i])/H2
  }}
  return(ests)
}
##
#---------------------------------------------
# delta^{Lp}, Lindley type estimator
# Ghosh et al (1983,355) Equation (2.13)
#-------------
delta.Lp <- function(y){
  p <- length(y)
  if(sum(y)==0){
    ests <- rep(0,p)}
  else{
    h.bar <- 0
    for(j in 1:p){
      h.bar <- h.bar + 1/p*h(y[j])}
    h.s <- numeric(p)
    for(i in 1:p){
      h.s[i] <- h(y[i])}
    N.bar <- sum(h.s > h.bar)
    const <- max(N.bar-2,0)
    D <- sum((h.s-h.bar)^2)
    if(D==0){
      ests<-h.bar}
    else{
    ests <- y - const*(h.s-h.bar)/D}
  }
  return(ests)
}
##
#--------------------------------------------
# delta^{m} Equation (3.8)
# Shrink to mean estimator
#-------------
delta.m <- function(y){
  p <- length(y)
  z <- sum(y)
  if(z==0){
    ests <- rep(0,p)}
  else{
    Syy <- (p-1)/p*var(y)
    y.bar <- mean(y)
    Bhat <- y.bar/(max(Syy-y.bar,0)+y.bar)
    ests <- Bhat*mean(y) + (1 - Bhat)*y
  }
  return(ests)
}
##
#--------------------------------------------
#--------------------------------------------
#
#--------------------------------------------
# Weighted squared error loss estimators L_1
#--------------------------------------------
#
# delta^{m1} Equation (3.13)
#-------------
delta.m1 <- function(y){
  p <- length(y); z <- sum(y)
  if(z==0){
    ests <- rep(0,p)}
  else{
    Syy <- (p-1)/p*var(y)
    y.bar <- mean(y)
    Bhat <- y.bar/(max(Syy-y.bar,0)+y.bar)
    ests <- Bhat*mean(y) + (1-Bhat)*(y-1)
    ests[which(y==0)] <- 0
  }
  return(ests)
}
##
```

```
#--------------------------------------------
# \delta^{CZ}
#-------------
delta.cz <- function(y){
  p <- length(y); z <- sum(y)
  ests <- (1-(p-1)/(p-1+z))*y
  return(ests)
}
##
#--------------------------------------------
# \delta^{Gm}
# Ghosh et al (1983,357) example 2.5
#-------------
delta.Gm <- function(y){
  p <- length(y);y.m <- min(y)
  g <- y - y.m; D <- sum(g)
  const <- max(sum(y>y.m)-1,0)
  ests <- rep(p,0)
  if(D != 0){
    ests <- y - const*g/D}
  return(ests)
}
##
#--------------------------------------------
# \delta^{\nu} Equation (3.14)
#-------------
delta.nu <- function(y,nu){
  ind <- function(vec,cut){
    # returns I(vec \geq cut)
    return(as.numeric(vec >= cut))
  }
  p.l <- sum(y >= nu)
  Z.l <- sum(ind(y,nu)*y)
  ests <- y
  if(p.l >= 2){
    ests <- y - ind(y,nu)*(p.l-1)*(y-nu)/(p.l-1+Z.l-p.l*nu)
  }
  return(ests)
}
#--------------------------------------------
#--------------------------------------------
#
# make a plot for one estimator
#--------------------------------------------
makePlot <- function(estimator,name,yrange){
plot(NA,NA,xlim=c(1,6),ylim=yrange,xlab="",ylab="% savings in risk",frame.plot=FALSE,xaxt='n',main=name)
p <- c("5","10","15")
Axis(side=1,at=c(1:6),labels=c("(0,4)","(0,8)","(8,12)","(12,16)","(0,12)",
      "(4,16)"))
```

```r
segments(0.2,0,5.8,0,lty=2)
for(j in 1:length(savings)){
  s <- savings[[j]]
  for(i in 1:3){
    text(j,s[i,estimator],p[i])}
  }
}
#-------------------------------------------
```

## C.4 Script used in Section 3.1

The squared error loss simulation study reported in Section 3.1. The script generates Figure 3.1 and Figure 3.2.

```r
# Section 3.1 R-script
# Squared error loss simulations
#----------------------------------------
source("PoissonEstimators.R")
#
#----------------------------------------
sims <- 10^4
loss <- c("LG1","LG2","LG3","LPeng","LLindley","Lm")
# ranges
range <- matrix(NA,nrow=6,ncol=2)
range[1,] <- c(0,4);range[2,] <- c(0,8)
range[3,] <- c(8,12);range[4,] <- c(12,16)
range[5,] <- c(0,12);range[6,] <- c(4,16)
size <- c(5,10,15)
savings <- list()
Estm = EstLindley = matrix(NA,nrow=sims,ncol=5)
for(j in 1:nrow(range)){
  pcs <- matrix(NA,nrow=3,ncol=length(loss))
  for(pp in 1:length(size)){
    LMLe = LG1 = LG2 = LG3 = LPeng = LLindley = Lm = 0
    theta <- runif(size[pp],range[j,1],range[j,2])
      if((j==3)&(pp==1)){
        save.theta <- theta}
    for(k in 1:sims){
      y <- rpois(size[pp],theta)
      LG1 <- LG1 + 1/sims*sum((delta.G1(y,round(theta,0)) - theta)^2)
      LG2 <- LG2 + 1/sims*sum((deltaG2(y) - theta)^2)
      LG3 <- LG3 + 1/sims*sum((deltaG3(y) - theta)^2)
      LPeng <- LPeng + 1/sims*sum((delta.Peng(y) - theta)^2)
      LLindley <- LLindley + 1/sims*sum((delta.Lp(y) - theta)^2)
      Lm <- Lm + 1/sims*sum((delta.m(y) - theta)^2)
      LMLe <- LMLe + 1/sims*sum((y-theta)^2)
      if((j==3)&(pp==1)){
        # save estimates for comparison
        # of the variance of estimators
        # save.theta is true theta values
        EstLindley[k,] <- delta.Lp(y)
        Estm[k,] <- delta.m(y)
      }
    }
    for(q in 1:length(loss)){
      pcs[pp,q] <- round((LMLe-get(loss[q]))/LMLe*100,2)
  }
}
savings[[j]] <- pcs
}
#-----------------------------------------
#-----------------------------------------
#
# Figure 3.1
#-----------------------------------------
# write (Figure 3.1)
postscript("figure3_1.eps")
par(mfrow=c(3,2))
makePlot(1,"delta^{G1}",c(-5,82))
makePlot(2,"delta^{G2}",c(-5,82))
makePlot(3,"delta^{G3}",c(-5,82))
makePlot(4,"delta^{Peng}",c(-5,82))
makePlot(5,"delta^{Lp}",c(-5,82))
makePlot(6,"delta^{m}",c(-5,83))
dev.off()
#-----------------------------------------
#-----------------------------------------
#
# Figure 3.2 (box plot)
#-----------------------------------------
# compare variances of \delta^{Lp} and \delta^{m}
myBox <- function(x,yup,title){
  boxplot(x,frame.plot=FALSE,xaxt="n",main=title,
    outline=FALSE,whisklty=0,staplelty=0,ylim=yup)
  for(i in 1:5){
    points(i,save.theta[i],pch=1,cex=1.2)
    q<-quantile(x[,i],c(.025,.975))
    segments(i,q[1],i,q[2],lty=2)
    segments(i-.2,q[1],i+.2,q[1],lwd=1.2)
    segments(i-.2,q[2],i+.2,q[2],lwd=1.2)
  }
}
postscript("figure3_2.eps")
par(mfrow=c(1,2))
myBox(Estm,c(4,20),"delta^{m}")
myBox(EstLindley,c(4,20),"delta^{Lp}")
dev.off()
##
# the savings in (8,12)
```

```
savings[[3]]
#       [,1]  [,2] [,3] [,4]  [,5]  [,6]
# [1,] 4.34 13.13 0.00 0.50 14.74 62.68
# [2,] 3.44 14.90 2.70 0.80 39.65 69.80
# [3,] 3.91 14.50 3.64 0.89 49.58 76.59
##
# the savings in (12,16)
savings[[4]]
#       [,1]  [,2] [,3] [,4]  [,5]  [,6]
# [1,] 3.51 13.75 0.00 0.24 15.02 65.37
# [2,] 2.53 16.77 1.84 0.41 44.04 78.54
# [3,] 2.61 15.79 2.26 0.52 52.70 82.01
#-----------------------------------------
#-----------------------------------------
```

## C.5 Script used in Section 3.2

The weighted squared error loss simulation study reported in Section 3.2. The script generates Figure 3.3 and Table 3.2.

```
# Section 3.2 R-script
# Weighted squared error loss simulations
#-----------------------------------------
source("PoissonEstimators.R")
#
#-----------------------------------------
sims <- 2*10^3
t.sims <- 10^1
loss <- c("Lcz","LGm","Lnu","Lm1")
# ranges of \theta_1,...,\theta_p
range <- matrix(NA,nrow=6,ncol=2)
range[1,] <- c(0,4);range[2,] <- c(0,8)
range[3,] <- c(8,12);range[4,] <- c(12,16)
range[5,] <- c(0,12);range[6,] <- c(4,16)
# sample sizes
size <- c(5,10,15)
savings <- list()
#
counter <- 1
for(j in 1:nrow(range)){
  pcs <- matrix(NA,nrow=3,ncol=length(loss))
  for(pp in 1:length(size)){
    Lcz = LGm = Lnu = Lm1 = LMLe = 0
    for(tt in 1:t.sims){
      theta <- runif(size[pp],range[j,1],range[j,2])
      if(sum(theta==0)>=1){
        print("trouble"); break}
      for(k in 1:sims){
        y <- rpois(size[pp],theta)
        Lcz <- Lcz + 1/sims*sum(1/theta*(delta.cz(y)-theta)^2)
        LGm <- LGm + 1/sims*sum(1/theta*(delta.Gm(y)-theta)^2)
        Lnu <- Lnu + 1/sims*sum(1/theta*(delta.nu(y,median(y)-1)-theta)^2)
        Lm1 <- Lm1 + 1/sims*sum(1/theta*(delta.m1(y)-theta)^2)
        LMLe <- LMLe + 1/sims*sum(1/theta*(y-theta)^2)
      }
      Lcz <- 1/t.sims*Lcz; LGm <- 1/t.sims*LGm;
      Lnu <- 1/t.sims*Lnu; Lm1 <- 1/t.sims*Lm1;
      LMLe <- 1/t.sims*LMLe;
    }
    for(q in 1:length(loss)){
      pcs[pp,q] <- round((LMLe-get(loss[q]))/LMLe*100,2)}
    colnames(pcs) <- c("cz","min","nu","m1")
    rownames(pcs) <- c("5","10","15")
  }
  savings[[j]] <- pcs
  cat(counter/nrow(range),"%","\n")
  counter <- counter + 1
}
#----------------------------------------
#----------------------------------------
#
# Figure 3.3
#----------------------------------------
postscript("figure3_3.eps")
par(mfrow=c(2,2))
makePlot(1,"delta^{CZ}",c(-5,90))
makePlot(2,"delta^{Gm}",c(-5,90))
makePlot(3,"delta^{nu}",c(-5,90))
makePlot(4,"delta^{m1}",c(-5,90))
dev.off()
#----------------------------------------
#----------------------------------------
#
# Table 3.1
#-------------------------------------
save1<-cbind(t(savings[[1]]),t(savings[[2]]),t(savings[[3]]))
save2<-cbind(t(savings[[4]]),t(savings[[5]]),t(savings[[6]]))
est <- c("$\\delta^{CZ}$","$\\delta^{Gm}$","$\\delta^{\\nu}$","$\\delta^{m1}$")
sink("table3_1i.txt")
  for(j in 1:4){
    row <- save1[j,]
    cat(est[j],"&",row[1],"&",row[2],"&",row[3],"&",
      row[4],"&",row[5],"&",row[6],"&",
      row[7],"&",row[8],"&",row[9],"\\\\","\n")
    }
```

```
sink()
sink("table3_1ii.txt")
  for(j in 1:4){
    row <- save2[j,]
    cat(est[j],"&",row[1],"&",row[2],"&",row[3],"&",
      row[4],"&",row[5],"&",row[6],"&",
      row[7],"&",row[8],"&",row[9],"\\\\","\n")
    }
sink()
#-----------------------------------------
#-----------------------------------------
```

## C.6 Script used in Section 3.3

Here is the script that compares normal- and Poisson theory estimators under $L_0$ and $L_1$. The script generates Table 3.2.

```
# Section 3.3 R-script
#-----------------------------------------
source("PoissonEstimators.R")
#
# Compare Normal and Poisson estimators
#-----------------------------------------
#
p <- 31
sims <- 10^4
ranges <- rbind(c(0,8),c(4,8))
cz.L1 = m1.L1 = js.L1 = Lindley.L1 = numeric(2)
Peng.L0 = m.L0 = js.L0 = Lindley.L0 = numeric(2)
for(j in 1:2){
  theta <- runif(p-2,ranges[j,1],ranges[j,2])
  theta <- c(ranges[j,1]+.01,theta,8) # set boundaries
  L1cz = L1m1 = L1js = L1Lindley = L1ML = 0
  L0Peng = L0m = L0js = L0Lindley = L0ML = 0
  for(i in 1:sims){
    y <- rpois(p,theta)
    # CZ-estimator
    z <- sum(y)
    cz <- y - (p-1)/(p-1+z)*y
    # James-Steins and Lindley
    x <- 2*sqrt(y);
    js.norm <- x - (p-2)/sum((x)^2)*x
    js <- 1/4*js.norm^2
    #
    x.bar <- mean(x)
    Lindley.norm <- x.bar + (p-3)/sum((x-x.bar)^2)*(x-x.bar)
    Lindley <- 1/4*Lindley.norm^2
    #
    L1cz <- L1cz + 1/sims*sum(1/theta*(cz-theta)^2)
    L0Peng <- L0Peng + 1/sims*sum((delta.Peng(y)-theta)^2)
    #
    # delta^{m} estimators
    L1m1 <- L1m1 + 1/sims*sum(1/theta*(delta.m1(y)-theta)^2)
    L0m <- L0m + 1/sims*sum((delta.m(y)-theta)^2)
    #
    L1js <- L1js + 1/sims*sum(1/theta*(js-theta)^2)
    L0js <- L0js + 1/sims*sum((js-theta)^2)
    #
    L1Lindley <- L1Lindley + 1/sims*sum(1/theta*(Lindley-
   theta)^2)
    L0Lindley <- L0Lindley + 1/sims*sum((Lindley-theta)^2)
    #
    L1ML <- L1ML + 1/sims*sum(1/theta*(y-theta)^2)
    L0ML <- L0ML + 1/sims*sum((y-theta)^2)
  }
  cz.L1[j] <- round((L1ML-L1cz)/L1ML*100,2)
  m1.L1[j] <- round((L1ML-L1m1)/L1ML*100,2)
  js.L1[j] <- round((L1ML-L1js)/L1ML*100,2)
  Lindley.L1[j] <- round((L1ML-L1Lindley)/L1ML*100,2)
  #
  Peng.L0[j] <- round((L0ML-L0Peng)/L0ML*100,2)
  m.L0[j] <- round((L0ML-L0m)/L0ML*100,2)
  js.L0[j] <- round((L0ML-L0js)/L0ML*100,2)
  Lindley.L0[j] <- round((L0ML-L0Lindley)/L0ML*100,2)


}
#-----------------------------------------
#-----------------------------------------
#
# Table 3.2
#-----------------------------------------
sink("table3_2.txt")
cat("& $\\delta^{CZ}$ & $\\delta^{m1}$ & $\\delta^{JS}$ & $\\
   delta^{L}$ & $\\delta^{CZ}$ & $\\delta^{m1}$ & $\\delta^{
   JS}$ & $\\delta^{L}$\\\\","\n")
cat("$L_1$","&",cz.L1[1],"&",m1.L1[1],"&",js.L1[1],"&",
   Lindley.L1[1],"&",
  cz.L1[2],"&",m1.L1[2],"&",js.L1[2],"&",Lindley.L1[2],"\\\\"
   ,"\n")
cat("\\hline","\n")
cat("& $\\delta^{P}$ & $\\delta^{m}$ & $\\delta^{JS}$ & $\\
   delta^{L}$ & $\\delta^{P}$ & $\\delta^{m}$ & $\\delta^{JS}
   $ & $\\delta^{L}$\\\\","\n")
cat("$L_0$","&",Peng.L0[1],"&",m.L0[1],"&",js.L0[1],"&",
   Lindley.L0[1],"&",
  Peng.L0[2],"&",m.L0[2],"&",js.L0[2],"&",Lindley.L0[2],"\\\\
   ")
sink()
#-----------------------------------------
```

```
#-----------------------------------------
```

## C.7 Script used in Section 4.3

This is the script for the simulated regressions in Section 4.3. This script produces Table 4.2 and Figure 4.2.

```
# Section 4.3 R-script
#-----------------------------------------
source("Ch4_GIBBS.R")
#
# Program regression models
#-----------------------------------------
#
# \delta^H Hudson (1985) estimator
#---------
delta.H <- function(y,Z){
  p <- length(y); H <- numeric(p)
  k <- dim(Z)[2]; N <- sum(y==0) #observed zeros
  est <- numeric(p) # the estimates
  for(i in 1:p){
    if(y[i] > 0){
      H[i] <- sum(1/1:y[i])}
    else{
      H[i] <- 0}}
    #
    beta.hat <- solve(t(Z)%*%Z)%*%t(Z)%*%H
    H.hat <- Z%*%beta.hat
    y.hat <- .56*(exp(H.hat) - 1); y.hat[y.hat < 0] <- 0
    S <- sum((H - H.hat)^2)
    shrink <- max(p-N-k-2,0)/S
    for(i in 1:p){
      if(y[i]+0.56 > shrink){
        est[i] <- y[i] - shrink*(H[i]-H.hat[i])}
      else{
        est[i] <- y.hat[i]}
      }
  # transform beta.hat to Poisson world
  beta.hat <- .56*(exp(beta.hat) - 1)
  return(list(est,beta.hat))
}
##
# \delta^{EB} (James-Stein type)
#---------
delta.EB <- function(y,Z){
  p <- length(y); k <- dim(Z)[2]
  x <- 2*sqrt(y)
  beta.hat <- solve(t(Z)%*%Z)%*%t(Z)%*%x
  xi.hat <- Z%*%beta.hat
  shrink <- (p-k-2)/sum((x-xi.hat)^2)
  est <- xi.hat + (1 - shrink)*(x-xi.hat)
  return(list(est,beta.hat))
}
##
# Empirical Bayes regression model
#---------
eb.fun <- function(y,start.params){
  logLikNegbin <- function(params){
    # params = c(beta,b)
    b0 <- params[1];b1 <- params[2]
    b <- params[3]; w <- b/(b+1)
    mu <- exp(Z%*%c(b0,b1))
    ll <- sum(lgamma(y + b*mu)-lfactorial(y) - lgamma(b*mu) +
    b*mu*log(w) + y*log(1-w))
    # returns negative
    return(-ll)}
  fit.eb <- nlm(logLikNegbin,start.params,hessian=TRUE)
  beta.eb <- fit.eb$estimate[1:2]
  b.eb <- fit.eb$estimate[3]
  delta.eb <- exp(Z%*%beta.eb)*b.eb/(b.eb+1) + (1-b.eb/(b.eb
   +1))*y
  #
  return(list(beta.eb,b.eb))
}
#-----------------------------------------
#-----------------------------------------
#
# Simulation study
#--------------------------------------
# just give me some truth
p <- 40; z <- 1:p/5
Z <- matrix(NA,ncol=2,nrow=p)
Z[,1] <- 1; Z[,2] <- z
b0 <- 0.2 ; b1 <- 0.5 ; bb <- 1/5
theta <- rgamma(p,bb*exp(b0+b1*z),bb)
#
sims <- 500
counter <- 1
b.eb = b.Bayes = numeric(sims) # collect the b.hat-params
beta.Bayes = beta.Preg = beta.H = beta.eb = beta.EB = matrix(
   NA,nrow=sims,ncol=2)
LBayes = LPreg = LHudson = LEmpBayes = Lefronmorris = Lml = 0
for(i in 1:sims){
  # draw data
  y <- rpois(p,theta)
  # Run the Gibbs-sampler
  #---------------------
  burnin <- 2*10^2; mcmc.sims <- 5*10^2
  ptm <- proc.time()
  mcmc <- gibbs(mcmc.sims,y,Z,Bprior=c(0,.3),Sigma=diag(2),
   bstart=3,zeta=2,eta=3)
  print(proc.time() - ptm)
  # get estimates
  mcmc <- mcmc[burnin:mcmc.sims,]
  beta.Bayes[i,] <- colMeans(mcmc)[1:2]
  b.Bayes[i] <- mean(mcmc[,3])
  theta.Bayes <- colMeans(mcmc)[4:ncol(mcmc)]
  ##
  # standard Poisson
  #----------------------
  beta.Preg[i,] <- glm(y~z,family="poisson")$coeff
  theta.Preg <- exp(Z%*%beta.Preg[i,])
  ##
  # \delta^H Hudson (1985)
  #----------------------
  Hudson <- delta.H(y,Z)
  theta.Hudson <- Hudson[[1]]
  beta.H[i,] <- Hudson[[2]]
  ##
  # Empirical Bayes
  #----------------------
  eb.hat <- eb.fun(y,c(.3,.8,1/2))
  beta.eb[i,] <- eb.hat[[1]]
  b.eb[i] <- eb.hat[[2]]
  w.hat <- b.eb[i]/(b.eb[i]+1)
  theta.eb <- exp(Z%*%beta.eb[i,])*w.hat + (1-w.hat)*y
  ##
  # delta.EB var.stab. transform
  em <- delta.EB(y,Z)
  # transform to Poisson world
  theta.EB <- 1/4*em[[1]]^2
  beta.EB[i,] <-  1/4*em[[2]]^2
  ##
  # Loss
  #-----------------
  LBayes <- LBayes + 1/sims*sum((theta.Bayes-theta)^2)
  LEmpBayes <- LEmpBayes + 1/sims*sum((theta.eb-theta)^2)
  LHudson <- LHudson + 1/sims*sum((theta.Hudson-theta)^2)
  LPreg <- LPreg + 1/sims*sum((theta.Preg-theta)^2)
  Lefronmorris <- Lefronmorris + 1/sims*sum((theta.EB-theta)^2)
  Lml <- Lml + 1/sims*sum((y-theta)^2)
  ##
  cat(counter/sims*100,"% -------------","\n")
  counter <- counter + 1
}
#
L0 <- round(c(LBayes,LEmpBayes,LHudson,LPreg,Lefronmorris,Lml),2)
est <- c("Pure Bayes","Emp. Bayes","$\\delta^{H}$",
    "Poisson reg.","$1/4(\\delta^{EB})^2$","MLE")
sink("table4_2.txt")
```

```
for(i in 1:length(est)){
  cat(est[i],"&",L0[i],"&",round((Lml-L0[i])/Lml*100,2),"\\\\",
    "\n")
}
sink()
#-----------------------------------------
#-----------------------------------------
#
# Figure 4.2 \beta box plot
#-----------------------------------------
beta.hats <- cbind(b.Bayes,b.eb,beta.Bayes,beta.Preg,beta.H,beta.eb,beta.EB)
# estimated b's in Bayes and empirical Bayes
for(i in 1:2){
cat(mean(beta.hats[,i]),quantile(beta.hats[,i],c(.025,.975)),"\n")}
# 0.2292195 0.1648957 0.3219973 # Bayes
# 0.2002152 0.1403232 0.2915222 # empirical Bayes
#
quickBox <- function(w,para){
  boxplot(beta.hats[,w[1]],beta.hats[,w[2]],beta.hats[,w[3]],
  beta.hats[,w[4]],beta.hats[,w[5]],frame.plot=FALSE,whisklty=0,
  staplelty=0,names=c("Bayes","Poisreg","H","EB","1/4dEB^2"),
  main=bquote(expression(beta[.(para)]))); i <- 1
  for(j in w){
    q<-quantile(beta.hats[,j],c(.025,.975))
    segments(i,q[1],i,q[2],lty=2)
    segments(i-.2,q[1],i+.2,q[1],lwd=1.2)
    segments(i-.2,q[2],i+.2,q[2],lwd=1.2); i <- i + 1}
}
postscript("figure4_2.eps")
par(mfrow=c(1,2))
quickBox(c(3,5,7,9,11),0)
quickBox(c(4,6,8,10,12),1)
dev.off()
#-----------------------------------------
#-----------------------------------------
```

## C.8 Script used in Section 5.1

This is the script used to make Figure 5.1 in Section 5.

```
# Figure 5.1 (modelling dependent thetas)
#-----------------------------------------
library(MASS);library(mvtnorm)
A.matrix <- function(p,rho){
  A <- matrix(NA,p,p)
  for(col in 1:p){
    for(row in 1:p){
    A[row,col] <- rho^(abs(col-row))
    }
  }
  return(A)
}
#
p <- 12
postscript("figure5_1.eps") ; par(mfrow=c(3,1))
for(rho in c(.8,.6,.4)){
  A <- A.matrix(p,rho)
  sims <- 10^4
  V <- rmvnorm(sims,rep(0,p),A)
  a <- 8; b <- 2.5
  theta <- qgamma(pnorm(V),a,b)
  colnames(theta) <- paste("theta",1:p,sep="")
  sd.theta1 <- var(theta[,1])
  acf.gamma <- numeric(p)
  for(j in 1:p){
    acf.gamma[j] <- (b^2/a)*var(theta[,1],theta[,j])
  }
  rm(theta)
  char.rho <- paste("=",rho,sep=" ")
  plot(NA,NA,ylim=c(min(acf.gamma),max(acf.gamma)),xlim=c(0,p+1),
      frame.plot=FALSE,xlab="|i-j|",ylab="cf",
      main=bquote(rho ~ .(char.rho)))
  for(l in 1:p){
    segments(l-1,0,l-1,acf.gamma[l],lwd=2)
    segments(l-.95,0,l-.95,rho^(l-1),lty=1,lwd=1.4,col="green")
    }
}
dev.off()
#-----------------------------------------
#-----------------------------------------
```

# Bibliography

Albert, J. H. (1981). Simultaneous Estimation of Poisson Means. *Journal of Multivariate Analysis*, 11:400–417.

Berger, J. O. (1980). Improving on Inadmissible Estimators in Continuous Exponential Families with Applications to Simultaneous Estimation of Gamma Scale Parameters. *The Annals of Statistics*, 8:545–571.

Berger, J. O. (1983). Discussion: Construction of Improved Estimators in Multiparameter Estimation for Discrete Exponential Families. *The Annals of Statistics*, 11:368–369.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis. Second Edition.* Springer-Verlag, New York.

Brown, L. D. (1979). A Heuristic Method for Determining Admissibility of Estimators – With Applications. *The Annals of Statistics*, 7:960–994.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis. Third Edition.* Chapman and Hall/CRC, Boca Raton.

Casella, G. and Berger, R. L. (2002). *Statistical Inference. Second Edition.* Duxbury Press, Brooks/Cole.

Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson Regression Modelling. *Journal of the American Statistical Association*, 92:618–632.

Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous Estimation of the Means of Independent Poisson Laws. *Journal of the American Statistical Association*, 70:698–705.

Efron, B. (1982). Maximum Likelihood and Decision Theory. *The Annals of Statistics*, 10:340–356.

Efron, B. (1986). Why Isn't Everyone a Bayesian? *The American Statistician*, 40:1–5.

Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators–Part I: The Bayes Case. *Journal of the American Statistical Association*, 66:807–815.

Efron, B. and Morris, C. (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators–Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, 67:130–139.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, 70:311–319.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 5:119–127.

Fienberg, S. A. and Holland, P. W. (1973). Simultaneous Estimation of Multinomial Cell Probabilities. *Journal of the American Statistical Association*, 68:683–691.

Gelman, A. and Robert, C. P. (2013). "Not Only Defended But Also Applied": The Perceived Absurdity of Bayesian Inference. *The American Statistician*, 67:1–5.

Ghosh, J. K., Delampaday, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer, New York.

Ghosh, M., Hwang, J. T., and Tsui, K.-W. (1983). Construction of Improved Estimators in Multiparameter Estimation for Discrete Exponential Families. *The Annals of Statistics*, 11:351–367.

Ghosh, M. and Parsian, A. (1981). Bayes Minimax Estimation of Multiple Poisson Parameters. *Journal of Multivariate Analysis*, 11:280–288.

Hudson, H. M. (1985). Adaptive Estimators for Simultaneous Estimation of Poisson Means. *The Annals of Statistics*, 13:246–261.

Hudson, H. M. and Tsui, K.-W. (1981). Simultaneous Poisson Estimators for A Priori Hypotheses About Means. *Journal of the American Statistical Association*, 76:182–187.

Hwang, J. T. (1982). Improving Upon Standard Estimators in Discrete Exponential Families with Applications to Poisson and Negative Binomial Cases. *The Annals of Statistics*, 10:857–867.

James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379.

Johnstone, I. (1984). Admissibility, Difference Equations and Recurrence in Estimating a Poisson Mean. *The Annals of Statistics*, 12:1173–1198.

Johnstone, I. (1986). Admissible Estimation, Dirichlet Principles and Recurrence of Birth-Death Chains on $\mathbb{Z}_+^p$. *Probability Theory and Related Fields*, 71:231–269.

Johnstone, I. and MacGibbon, K. B. (1992). Minimax Estimation of a Constrained Poisson Vector. *The Annals of Statistics*, 20:807–831.

Lay, D. C. (2012). *Linear Algebra and Its Applications. Fourth Edition.* Addison-Wesley, Pearson, Boston.

Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley & Sons, New York.

Lindley, D. V. (1962). Discussion on Professor Stein's Paper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24:285–287.

Morris, C. N. (1983a). Discussion: Construction of Improved Estimators in Multiparameter Estimation for Discrete Exponential Families. *The Annals of Statistics*, 11:372–374.

Morris, C. N. (1983b). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78:47–55.

Peng, J. C.-M. (1975). Simultaneous Estimation of the Parameters of Independent Poisson Disitributions. Technical report no. 48, Stanford University, Departement of Statistics.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robert, C. P. (2001). *The Bayesian Choice: from decision theoretic foundations to computational implementation. 2nd ed.* Springer, Berlin.

Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R.* Springer, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:197–206.

Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9:1135–1151.

Stigler, S. M. (1990). The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators. *Statistical Science*, 5:147–155.

Tsui, K.-W. (1981). Simultaneous Estimation of Several Poisson Parameters Under Squared Error Loss. *Annals of the Institute of Statistical Mathematics*, 33:215–223.

Tsui, K.-W. (1984). Robustness of Clevenson-Zidek-Type Estimators. *Journal of the American Statistical Association*, 79:152–157.

Tsui, K.-W. and Press, S. J. (1982). Simultaneous Estimation of Several Poisson Means Under K-Normalized Squared Error Loss. *The Annals of Statistics*, 10:93–100.