Multivariate estimation of Poisson parameters

Emil Aas Stoltenberg [*], Nils Lid Hjort

Department of Mathematics, University of Oslo, P.B. 1053, Blindern, N-0316 Oslo, Norway

ABSTRACT

This paper is devoted to the multivariate estimation of a vector of Poisson means. A novel loss function that penalises bad estimates of each of the parameters and also the sum (or equivalently the mean) of the parameters is introduced. Under this loss function, a class of minimax estimators that uniformly dominate the maximum likelihood estimator is derived. Crucially, these methods have the property that for estimating a given component parameter, the full data vector is utilised. Estimators in this class can be fine-tuned to limit shrinkage away from the maximum likelihood estimator, thereby avoiding implausible estimates of the sum of the parameters. Further light is shed on this new class of estimators by showing that it can be derived by Bayesian and empirical Bayesian methods. In particular, we exhibit a generalisation of the Clevenson–Zidek estimator, and prove its admissibility. Moreover, a class of prior distributions for which the Bayes estimators uniformly dominate the maximum likelihood estimator under the new loss function is derived. A section is included involving weighted loss functions, notably also leading to a procedure improving uniformly on the maximum likelihood method in an infinite-dimensional setup. Importantly, some of our methods lead to constructions of new multivariate models for both rate parameters and count observations. Finally, estimators that shrink the usual estimators towards a data based point in the parameter space are derived and compared.

1. Introduction

Let $Y = (Y_1, \ldots, Y_p)$ be a vector of independent Poisson random variables with mean vector $\theta = (\theta_1, \ldots, \theta_p)$. In this article we consider the problem of estimating the vector θ. The obvious estimator is $\delta_0(Y) = Y$, that is, using $\delta_{0,i} = Y_i$ for each of the p individual parameters. It is well known that δ_0 is the maximum likelihood solution, that it has components with uniformly minimal variance among the unbiased estimators, and that it is admissible in the one-dimensional problem under squared error loss as well as under its weighted version, see, e.g., Lehmann [19, p. 277]. In the simultaneous or multivariate problem, however, Peng [22] and Clevenson and Zidek [6] were the first to show that δ_0 can be improved upon under the loss functions

$$L_0^*(\theta, \delta) = \sum_{i=1}^p (\delta_i - \theta_i)^2 \quad \text{and} \quad L_1^*(\theta, \delta) = \sum_{i=1}^p (\delta_i - \theta_i)^2/\theta_i, \tag{1}$$

---

* Corresponding author.
E-mail address: emilas@math.uio.no (E.A. Stoltenberg).

if $p \geq 3$ and $p \geq 2$, respectively. In particular, for the L_1^* loss, Clevenson and Zidek [6] derived the estimator

$$\delta_{CZ,i}(Y) = \left(1 - \frac{p-1}{p-1+Z}\right)Y_i, \quad Z = \sum_{i=1}^{p} Y_i, \tag{2}$$

demonstrating that it improves uniformly on the standard procedure δ_0. This Stein-type phenomenon has also been observed for other loss functions. Hwang [15] obtained results for $\sum_{i=1}^{p}(\delta_i-\theta)^2/\theta_i^{m_i}$, for fixed integers m_i. Ghosh and Yang [13] considered a loss function based on the entropy or Kullback–Leibler distance $L_e(\theta, \delta) = \sum_{i=1}^{p} \theta_i\{\delta_i/\theta_i - \log(\delta_i/\theta_i) - 1\}$. An excellent review article for admissibility issues, for multivariate Poisson means and for other models for discrete data, is [11], followed by discussion contributions by Berger [3], Morris [21], Hudson [14]. A more recent article on these issues is [5].

More broadly, the books Efron [7] and Fourdrinier et al. [10] contribute to seeing multivariate estimation, using shrinkage and empirical Bayes methods, as a global phenomenon. Central themes are variations on 'borrowing strength', either via direct constructions or in empirical Bayes setups. Our article is a contribution in these general directions, showing that methods developed for the multinormal and more generally spherically symmetric distributions have certain parallels in the world of multivariate Poisson estimation. Importantly, when estimating a particular Poisson parameter θ_i, the full multivariate vector of data is being used, as an integral part of our methods.

The estimators that have been found to be better than $\delta_0(Y) = Y$ in these earlier publications are essentially all of the shrinking type, pushing the maximum likelihood estimator closer to the bottom corner of the parameter space. A good example of the merits of this type of shrinkage is provided by Clevenson and Zidek [6], wherein 36 small θ_i – intensities of an oil-well discovery process – are estimated with the estimator in (2). Clevenson and Zidek [6] had access to 'known' θ_i and could check the actual loss incurred. The procedure in (2) did indeed give much smaller loss $\sum_{i=1}^{36}(\delta_{CZ,i} - \theta_i)^2/\theta_i$ than did δ_0; the losses are 14.33 and 39.26, respectively. However, another and perhaps not so pleasant feature of their estimation procedure is conspicuous here, namely that the sum $\gamma = \sum_{i=1}^{36} \theta_i$ (or equivalently the mean $\bar{\theta} = \gamma/36$) is seriously underestimated. In their oil-well discovery example the true sum is $\gamma = 26.98$, the usual δ_0 has $\sum_{i=1}^{36} y_i = 29$, while their $\sum_{i=1}^{36} \delta_{CZ,i} = 12.97$ is much too low.

In some situations the sum (or mean) is unimportant and all that matters is to estimate each θ_i. In many multiparameter cases one wants to keep track of the sum (or the mean) of the θ_i as well, however, as one surely would in the oil-well discovery process above. Other multiparameter cases where the sum (or the mean) are deemed as important as the individual θ_i abound; think for example of a decision maker having to make budgetary decisions concerning each of the boroughs of a city and the city as a whole, the resources allocated to each of p hospitals and the whole health budget, etc.

These considerations motivate studying loss functions that take into account the need for good individual estimates, while at the same time guarding against the underestimation of the sum. One example of such a loss function is

$$L(\theta, \delta) = \sum_{i=1}^{p}(\delta_i - \theta_i)^2 + c\left(\sum_{i=1}^{p} \delta_i - \gamma\right)^2 = L_0^*(\theta, \delta) + cp^2(\bar{\delta} - \bar{\theta})^2, \tag{3}$$

Since $\sum_{i=1}^{p} \delta_{0,i} = \sum_{i=1}^{p} Y_i$ is the admissible minimax solution under the loss function $\left(\sum_{i=1}^{p} \delta_i - \gamma\right)^2$ one might wonder whether the extra penalisation above would secure admissibility of δ_0 in the simultaneous problem. In Section 2 we show that this is not the case; δ_0 is again inadmissible when $p \geq 3$, under any given quadratic form loss function $(\delta-\theta)^\top A(\delta-\theta)$. Another loss function that takes the guarding against underestimation of the sum (or the mean) of the parameters into account is the weighted version of (3), that is, the loss functions that generalises the one used by Clevenson and Zidek [6], namely

$$L_c(\theta, \delta) = \sum_{i=1}^{p} \frac{(\delta_i - \theta_i)^2}{\theta_i} + c\frac{(\sum_{i=1}^{p} \delta_i - \sum_{i=1}^{p} \theta_i)^2}{\sum_{i=1}^{p} \theta_i} = L_1^*(\theta, \delta) + cp\frac{(\bar{\delta} - \bar{\theta})^2}{\bar{\theta}}, \tag{4}$$

where c is a user-defined constant. Under (4), the loss in Clevenson and Zidek's example for δ_0 is $39.26 + 0.35\,c$ while the loss for δ_{CZ} is $14.33 + 6.35\,c$. Again, one might wonder whether δ_0 is admissible for certain values of c. In Section 3 we show that this is not the case, and derive an estimator which is the natural generalisation of the one in (2),

$$\delta_i^c(Y) = \left\{1 - \frac{p-1}{p-1+(1+c)Z}\right\}Y_i, \quad i \in \{1, \ldots, p\}, \tag{5}$$

which is shown to belong to a class of estimators that dominate δ_0 uniformly over the parameter space when $p \geq 2$. In the somewhat similar multivariate normal setting, investigations on how to limit the shrinkage of the James–Stein estimator have been conducted by Efron and Morris [8,9] and by Stein [24].

Another type of loss functions constructed to compromise between conflicting desiderata, and as such somewhat similar to (3)–(4), are the so-called balanced loss functions studied, e.g., by Jafari Jozani et al. [16] and by Marchand and Strawderman [20]. See our Remark A in Section 9 for more on this kinship.

Our paper proceeds as follows. In Section 2 we study the loss function (3) and derive a class of estimators that uniformly dominate the maximum likelihood estimator. Section 3 concerns the weighted version of this loss function, the L_c loss

function of (4), and we derive the already mentioned generalisation δ^c of δ_{CZ}. In Section 4 further light is shed on this new class of estimators by showing that it can be derived by Bayesian and empirical Bayesian methods. We are also able to prove that the estimator δ^c of (5) is admissible. Classes of alternative estimators are then studied in Sections 5 and 6, involving Bayes and empirical Bayes strategies which shrink the raw data towards data-driven structures for the θ_i, such as the data mean, or a submodel. Some of these procedures succeed in having risks well below the minimax benchmark, in large regions of the parameter space, though without achieving uniform dominance. In Section 7 weighted loss functions are studied, which admit relative weights of importance; notably, an infinite-dimensional setup is also included. Then in Section 8 we demonstrate how some of our Bayesian constructions also lead to new multivariate models for rate parameters and for count observations, of interest in their own right, pointing to models for spatially dependent count data. Finally, Section 9 offers a list of concluding remarks.

2. General quadratic loss function

We may write the loss function of (3) as $L(\delta, \theta) = (\delta - \theta)^\top A_0 (\delta - \theta)$, where A_0 is the matrix with $1 + c$ down its diagonal and c elsewhere. The natural generalisation is

$$L(\delta, \theta) = (\delta - \theta)^\top A (\delta - \theta), \tag{6}$$

where A is symmetric and positive definite. Below we obtain some results for general A and apply these to two examples. Note that in situations where there is no ordering of the individual θ_i and no reason to estimate some of them more precisely than the others, the loss function in (3), that is, using A_0 with an appropriate choice of c, is the natural choice. Our method of proving inadmissibility of δ_0 resembles that of Tsui and Press [25] and Hwang [15], where A is diagonal.

Let $\delta(y) = y - \phi(y)$ be a competitor. The difference in risk between these two estimators is then

$$
\begin{aligned}
R(\delta, \theta) - R(\delta_0, \theta) &= \mathrm{E}_\theta \{ (\delta - \theta)^\top A (\delta - \theta) - (Y - \theta)^\top A (Y - \theta) \} \\
&= \mathrm{E}_\theta \{ -2(Y - \theta) A \phi(Y) + \phi(Y)^\top A \phi(Y) \} = \mathrm{E}_\theta \{ -2(Y - \theta) \psi(Y) + \psi(Y)^\top A^{-1} \psi(Y) \},
\end{aligned}
\tag{7}
$$

writing $\psi(y) = A\phi(y)$. Furthermore, since $\mathrm{E}_\theta \theta_i g(Y) = \mathrm{E}_\theta Y_i g(Y - e_i)$ for any g with $\mathrm{E}_\theta |g(Y)| < \infty$, using e_i to denote the unit vector with the ith element equal to one, we have that $\mathrm{E}_\theta (Y_i - \theta_i) \psi(Y) = \mathrm{E}_\theta Y_i \{ \psi(Y) - \psi(Y - e_i) \}$, and the risk difference (7) can be written $\mathrm{E}_\theta D(Y, \phi)$, in which

$$D(y, \phi) = - \sum_{i=1}^p 2 y_i \{ \psi(y) - \psi(y - e_i) \} + \psi(y)^\top A^{-1} \psi(y).$$

If a function $\psi(y)$ can be found such that $D(y, \phi)$ is non-positive for all y, with strict inequality for at least one datum y, then $\delta_0 = Y$ is inadmissible, being outperformed by $\delta(y) = y - A^{-1} \psi(y)$.

Theorem 1. *Let A be symmetric and positive definite. Then $\delta_0 = Y$ is inadmissible under loss function (6), if $p \geq 3$. It is dominated by $\delta(Y) = Y - A^{-1} \psi(Y)$, where $\psi(\cdot)$ is any member of the following class*

$$\psi_i(y) = \frac{d(y)}{B(y)} T(y_i), \quad i \in \{1, \ldots, p\},$$

where $T(0) = 0$ and $T(y) = \sum_{j \leq y} 1/j$ for $y \geq 1$, where $B(y) = \sum_{i=1}^p T(y_i) T(y_i + 1)$, and $d(y)$ is nondecreasing in each argument and obeying

$$0 \leq d(y) \leq (2/M) \{ N(y) - 2 \}_+,$$

writing $a_+ = \max(a, 0)$ for truncating to zero. Here $N(y) = \sum_{i=1}^p I\{ y_i \leq 1 \}$ and M is the inverse of the smallest eigenvalue of A.

Proof. The choice of M entails $\psi^\top A^{-1} \psi \leq M \sum_{i=1}^p \psi_i^2$ so that $D(\phi, y) \leq -2 \sum_{i=1}^p y_i \{ \psi_i(y) - \psi_i(y - e_i) \} + \sum_{i=1}^p M \psi_i(y)^2$. The general Theorem 2.1 in Hwang [15], with accompanying corollaries, can then be used to find $\psi_i(\cdot)$ functions that make $D(\phi, y)$ non-positive for all y. We skip details but record that Hwang's method gives

$$D(\phi, y) \leq -2 d(y) \frac{\{ N(y) - 2 - M d(y)/2 \}_+}{B(y)}, \tag{8}$$

which is non-positive for each $\psi(\cdot)$ described in the theorem. $\square$

A natural choice for $d(y)$ is $d_0(y) = (1/M) \{ N(y) - 2 \}_+$, minimising the upper bound in (8). This means using $\phi_0(y) = A^{-1} \psi_0(y)$ with $D(\phi_0, y) \leq -M^{-1} \{ N(y) - 2 \}^2 / B(y)$, which shows that

$$\delta(y) = y - A^{-1} \psi_0(y), \quad \psi_{0,i}(y) = \frac{1}{M} \frac{T(y_i)}{B(x)} \{ N(y) - 2 \}_+,$$

achieves $R(\delta, \theta) - R(Y, \theta) \leq -M^{-1} \mathrm{E}_\theta \{ (N(Y) - 2)_+ \}^2 / B(Y) < 0$, for all θ. Note in particular that the same $\psi_0(\cdot)$ function works for a large class of loss functions (6).

Example 2.1. Let A be the square matrix with $1 + c$ down its diagonal and c elsewhere. This matrix might be written $A = I + cee^\top$, with I the identity matrix and $e = (1, \dots, 1)^\top$ the $p \times 1$ vector of ones. This choice of A gives the loss function (3). Here $M = 1$ and $A^{-1} = I - \{c/(1 + cp)\}ee^\top$, and $\psi^\top A^{-1}\psi \le \sum_{i=1}^p \psi_i^2$ follows. The natural estimator is then $\delta(y) = y - \phi_0(y)$, where

$$\phi_0(y) = \psi_{i,0}(y) - \frac{c}{1+cp}\sum_{j=1}^p \psi_{0,j}(y) = \{N(y) - 2\}_+ \frac{T(y_i) - \frac{cp}{1+cp}\bar{T}(y)}{B(y)},$$

in which $\bar{T}(y) = p^{-1}\sum_{i=1}^p T(y_i)$. Note that δ does not shrink the y in any particular direction, but rather pushes the components y_i in different directions according to the sign of $T(y_i) - \{cp/(1 + cp)\}\bar{T}(y)$. Note further that if the y_i are moderate or large, then

$$\delta_i(y) \doteq y_i - \frac{p-2}{\sum_{j=1}^p(\log y_j)^2}\left\{\log y_i - \frac{cp}{1+cp}\frac{1}{p}\sum_{j=1}^p \log y_j\right\}.$$

Example 2.2. Samples of independent Poisson variables arise naturally when one or more Poisson processes are observed over time. Dividing the time interval into p parts gives counts $y_1, \dots, y_p$ with certain means $\theta_1, \dots, \theta_p$. In the nonparametric setting, where the intensity of the process is unknown, these parameters are also completely unknown. If one wishes to estimate the cumulative intensity of the process, then $(\lambda_1, \dots, \lambda_p)$ are more important than $(\theta_1, \dots, \theta_p)$, where $\lambda_i = \sum_{j=1}^i \theta_j$. This suggests using the loss function $L(\theta, \delta) = \sum_{i=1}^p(\widehat{\lambda}_i - \lambda_i)^2$, where $\widehat{\lambda}_i = \sum_{j=1}^i \delta_j$. But this is seen to be a special case of (6), with elements $a_{i,j} = p + 1 - \max(i, j)$ filling the A matrix. Its inverse A^{-1} has first row $(1, -1, 0, \dots, 0)$, last row $(0, \dots, 0, -1, 2)$, and in between we find $(0, \dots, 0, -1, 2, -1, 0, \dots, 0)$. One has $\psi^\top A^{-1}\psi \le 4\psi^\top \psi$, and can use $M = 4$ when applying the theorem. Hence the following estimator improves on $\delta_0 = Y$:

$$\delta_1^*(y) = y_1 - \psi_{0,1}(y) + \psi_{0,2}(y),$$
$$\delta_i^*(y) = y_i + \psi_{0,i-1}(y) - 2\psi_{0,i}(y) + \psi_{0,i+1}(y), \quad i \in \{2, \dots, p-1\},$$
$$\delta_p^*(y) = y_p + \psi_{0,p-1}(y) - 2\psi_{0,p}(y),$$

where this time

$$\psi_{0,i}(y) = \frac{1}{4}\frac{T(y_i)}{B(y)}\{N(y) - 2\}_+.$$

Notice finally that the corresponding improved estimators for the cumulatives λ_i become $\widehat{\lambda}_i = \sum_{j=1}^i Y_j - \psi_{0,i}(Y) + \psi_{0,i+1}(Y)$ for $i \le p - 1$, while $\widehat{\lambda}_p = \sum_{j=1}^p Y_j - \psi_{0,p}(Y)$.

3. The L_c loss function

The main consideration leading to the loss function $(\delta_i - \theta_i)^2/\theta_i$ is that the statistician seeks precise estimates of small values of θ_i. A loss function that penalises heavily for bad estimates of small parameters is then a natural choice. Related to this is the obvious fact that when the parameters are small, they can only be badly overestimated, zero being the boundary of the parameter space. The corresponding multiparameter version of this is $\sum_{i=1}^p(\delta_i - \theta_i)^2/\theta_i$, and is the one treated by Ghosh, Hwang, and Tsui [11], Hwang [15], Tsui and Press [25], Clevenson and Zidek [6], and others. Note that $\delta_0 = Y$ has constant risk p with this loss, and it is not difficult to establish that it is minimax.

The above mentioned authors obtain classes of estimators that perform uniformly better than δ_0 if only $p \ge 2$. As discussed in Section 1 and illustrated by the oil-well example, these shrinkage estimators do not take into account the additional desideratum, namely a precise estimate of the sum $\gamma = \sum_{i=1}^p \theta_i$. We will now study the L_c loss function of (4), $L_c(\theta, \delta) = L_1^*(\theta, \delta) + c(\sum_{i=1}^p \delta_i - \gamma)^2/\gamma$. If we consider the second term in (4) by itself, we recognise the one-dimensional loss function $(\delta - \gamma)^2/\gamma$. It is well known that $\widehat{\gamma} = Z$ is admissible and the unique minimax solution under this loss function, and can therefore not be uniformly improved upon [19, p. 277]. Consequently, since z is the sum of the y_i, higher values of c will result in estimators that lie closer the δ_0. On the other hand, we know that for $c = 0$ the estimator in (2) uniformly dominates δ_0. Hence, the user defined constant c determines how to compromise between δ_0 and δ_{CZ}.

Before we derive a class of estimators that dominate δ_0 under L_c in Section 3.2, we derive formulae for the Bayes solution and show that δ_0 is minimax.

3.1. The $\delta_0 = Y$ estimator is minimax

The maximum likelihood estimator $\delta_0 = Y$ has constant risk $p + c$ under L_c and is a natural candidate for being minimax. We demonstrate minimaxity by exhibiting a sequence of priors with minimum Bayes risks $BR(\delta, \pi) = E R_c(\delta, \theta)$ which converge towards $p + c$; that this is sufficient follows from well-known arguments, as exposited, e.g., in Robert [23, Ch. 2.4].

Some analysis is required to characterise the Bayes solution. We first find the values $\delta_1, \ldots, \delta_p$ that minimise the posterior expected loss. Given a dataset $y = (y_1, \ldots, y_p)$, and with $\gamma = \sum_{i=1}^{p} \theta_i$, introduce

$$a_i = \{E(\theta_i^{-1} \mid y)\}^{-1}, \quad a = \sum_{i=1}^{p} a_i, \quad b = \{E(\gamma^{-1} \mid y)\}^{-1}.$$

Then

$$E\{L_c(\theta, \delta) \mid y\} = \sum_{i=1}^{p} \{\delta_i^2/a_i - 2\delta_i + E(\theta_i \mid y)\} + c\left\{\left(\sum_{i=1}^{p} \delta_i\right)^2/b - 2\sum_{i=1}^{p} \delta_i + E(\gamma \mid y)\right\}, \tag{9}$$

assuming the moments to exist. Some analysis shows that the minimum takes place for

$$\delta_i^B(y) = \frac{1+c}{1+ca/b} a_i = \frac{1+c}{1+ca/b} \frac{1}{E\{(1/\theta_i) \mid y\}}, \quad i \in \{1, \ldots, p\}. \tag{10}$$

This generalises the familiar result that the Bayes solution is $\{E(\theta_i^{-1} \mid Y)\}^{-1} = a_i$ under $L_1^*(\theta, \delta)$, that is, when $c = 0$. Note also that if $E\theta_i^{-1} = \infty$ for some i, and $E\theta_i$ is finite, then only $\delta_i = 0$ gives a finite risk, which means that (10) is correct even in such cases.

To illustrate this, suppose θ_i has a Gamma prior with parameters (α_i, β), which we write as $\mathcal{G}(\alpha_i, \beta)$, i.e., θ_i has prior mean α_i/β, and that these are independent. Then $\theta_i \mid y \sim \mathcal{G}(\alpha_i + y_i, \beta + 1)$, and some calculations lead to the Bayes estimators

$$\delta_i^B = \frac{1+c}{1+c(p\bar\alpha + z - p)/(p\bar\alpha + z - 1)} \frac{\alpha_i + y_i - 1}{\beta + 1} = \frac{(1+c)(p\bar\alpha + z - 1)}{(1+c)(p\bar\alpha + z - 1) - c(p-1)} \frac{\alpha_i + y_i - 1}{\beta + 1},$$

for $i \in \{1, \ldots, p\}$, writing $\bar\alpha = (1/p)\sum_{i=1}^{p} \alpha_i$ and again $z = \sum_{i=1}^{p} y_i$. If $\alpha_i \leq 1$, then the Bayes estimate is zero if $y_i = 0$, by the comment above about (10). We note that for large c, corresponding to the loss being essentially related to estimating the sum $\gamma = \sum_{i=1}^{p} \theta_i$ well under $(\widehat\gamma - \gamma)^2/\gamma$ loss,

$$\delta_i^B \doteq \frac{p\bar\alpha + z - 1}{\beta + 1} \frac{\alpha_i + y_i - 1}{p\bar\alpha + z - p}, \quad \sum_{i=1}^{p} \delta_i^B \doteq \frac{p\bar\alpha + z - 1}{\beta + 1},$$

tying in with $\gamma \mid y$ being a Gamma with parameters $(p\bar\alpha + z, \beta + 1)$.

Now consider the special case where the θ_i are independent $\mathcal{G}(1, \beta)$. The Bayes solution then takes the form

$$\delta_i^B = \frac{1+c}{1+cz/(p-1+z)} \frac{y_i}{1+\beta} = \frac{(1+c)(p-1+z)}{p-1+(1+c)z} \frac{y_i}{1+\beta}. \tag{11}$$

It now remains to show that the minimum Bayes risk for this prior, say MBR$(1, \beta)$, tends to $p + c$ as $\beta \to 0$. Using that Y_i given Z is binomial with mean $Z\theta_i/\gamma$, provided $Z \geq 1$, and that the Y vector and hence the δ^B estimator are equal to zero when $Z = 0$, the risk of δ_i^B can be expressed as

$$R_c(\delta_i^B, \theta) = E_\theta E_\theta \{L(\theta, \delta_i^B) \mid Z\} = E_\gamma \left\{\frac{(1+c)^2(p-1+Z)^2}{p-1+(1+c)Z} \frac{Z}{\gamma(1+\beta)^2} - 2\frac{(1+c)^2(p-1+Z)}{p-1+(1+c)Z} \frac{Z}{1+\beta} + (1+c)\gamma\right\}.$$

Since the risk depends on the θ_i only through the sum γ, the minimum Bayes risk may be written

$$\text{MBR}(1, \beta) = E_\gamma E_\gamma \{L(\theta, \delta^B) \mid Z\} = \frac{1+c}{1+\beta} E_\gamma \frac{p(p-1) + (p+c)Z}{p-1+(1+c)Z},$$

in which the expectation on the right is with respect to the marginal distribution of Z. Since Z tends in probability to infinity as $\beta \to 0$, the function $\{p(p-1) + (p+c)Z\}/\{p-1+(1+c)Z\}$ converges to $(p+c)/(1+c)$ in probability when $\beta \to 0$. Furthermore, this function is bounded by p, so by the bounded convergence theorem MBR$(1, \beta)$ tends to $p + c$ as β tends to zero, as was to be shown.

3.2. A dominating class of estimators

We will now develop a class of estimators with uniformly smaller risk than $p + c$ under the L_c loss function, that is, estimators that uniformly dominate the maximum likelihood estimator. Consider estimators of the form $\delta_i(Y) = \{1 - \phi(Z)\}Y_i$. Write $D(\phi, y) = L_c(\theta, \delta) - L_c(\theta, \delta_0)$ for the difference in loss. Then

$$D(\phi, y) = \sum_{i=1}^{p} \left\{\frac{\phi(z)^2 y_i^2 - 2\phi(z)y_i(y_i - \theta_i)}{\theta_i} + c\frac{\phi(z)^2 z^2 - 2\phi(z)z(z - \gamma)}{\gamma}\right\},$$

Now, use the fact that for any real valued function h with finite mean $E_\theta h(Y)$, and with the property that $h(y) = 0$ whenever $y_i = 0$, the following identity holds:

$$E_\theta h(Y)/\theta_i = E_\theta h(Y + e_i)/(Y_i + 1). \tag{12}$$

Using this identity we obtain an expression for the difference in risk $E_\theta D(\phi, Y) = R(\delta^*, \theta) - R(Y, \theta)$, namely

$$E_\theta D(\phi, Y) = E_\theta E_\theta \{D(\phi, Y) \mid Z\} = E_\theta \left[\{\phi^2(Z) - 2\phi(Z)\} \frac{Z\{(p-1) + (1+c)Z\}}{\gamma} + 2(1+c)\phi(Z)Z \right]$$

$$= E_\theta \{(\phi^2(Z+1) - 2\phi(Z+1))\{(p-1) + (1+c)(Z+1)\} + 2(1+c)\phi(Z)Z\}.$$

This can hence be expressed as $E_\gamma D^*(\phi, Z)$, with the $D^*(\phi, z)$ function not depending on the parameters; in particular, the risk function $R_c(\delta, \theta) = p + c + E_\gamma D^*(\phi, Z)$ depends on the parameter vector only via $\gamma = \sum_{i=1}^p \theta_i$. Also, any function $\phi(\cdot)$ that ensures that $D^*(\phi, z) \leq 0$ for all z, with strict inequality for at least one datum z, yields an estimator that uniformly dominates the δ_0. This leads to the following result.

Theorem 2. *For each function $\phi(\cdot)$ such that*

$$0 < \phi(z) < \frac{2(p-1)}{p-1+(1+c)z}, \quad \phi(z)z < \phi(z+1)(z+1),$$

for all z, the estimator $\delta_c = \{1 - \phi(Z)\}Y$ uniformly dominates $\delta_0 = Y$. These conditions are met for functions of the type $\phi(z) = \psi(z)/\{p - 1 + (1 + c)z\}$ where $\psi(\cdot)$ is nondecreasing with $0 < \psi(z) < 2(p - 1)$. In particular, δ_0 is inadmissible if $p \geq 2$.

Proof. Using the expression for $D^*(\phi, Z)$ derived above, the following holds:

$$E_\gamma D^*(\phi, Z) = E_\gamma \left[\{\phi(Z)^2 - 2\phi(Z)\} \frac{Z\{(p-1) + (1+c)Z\}}{\gamma} + 2(1+c)\phi(Z)Z \right]$$

$$= E_\gamma \phi(Z)Z \left[\{\phi(Z) - 2\} \frac{p-1+(1+c)Z}{\gamma} + 2(1+c) \right]$$

$$= \frac{1}{\gamma} E_\gamma \phi(Z)Z \big[\phi(Z)\{p - 1 - (1+c)Z\} - 2\{p - 1 + (1+c)(Z - \gamma)\} \big]$$

$$\leq \frac{1}{\gamma} E_\gamma \phi(Z)Z \big[2(p-1) - 2\{p - 1 + (1+c)(Z - \gamma)\} \big]$$

$$= -\frac{2(1+c)}{\gamma} E_\gamma \phi(Z)Z(Z - \gamma)$$

$$= -2(1+c) E_\gamma \{\phi(Z)Z^2/\gamma - \phi(Z)Z\}$$

$$= -2(1+c) E_\gamma \{\phi(Z+1)(Z+1) - \phi(Z)Z\} < 0.$$

This is valid for all γ since $\phi(z)z$ is a strictly increasing function of z. □

We denote by $\mathcal{D}_c$ the class of estimators

$$\delta(Y) = \{1 - \phi(Z)\}Y, \quad \phi(z) = \psi(z)/\{p - 1 + (1 + c)z\},$$

with $\psi(\cdot)$ satisfying the conditions of Theorem 2. The optimal choice of $\psi(\cdot)$ in terms of minimising risk, based on the simple upper bound of $E_\theta D^*(\phi, Z)$, is $\psi(z) = p - 1$, leading to the estimator

$$\delta_i^c(Y) = \left\{ 1 - \frac{p-1}{p-1+(1+c)Z} \right\} Y_i, \quad i \in \{1, \dots, p\}. \tag{13}$$

Note that δ^c appropriately generalises the Clevenson–Zidek estimator of (2). Importantly, it is clearly seen how fine-tuning of c determines the amount of shrinkage away from the δ_0. We can use the expression for $D^*(\phi, z)$ derived above to find the risk function for the estimator in (13),

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

Note that the risk depends on θ_i only through the sum γ, and that numerical evaluation is easy because Z is Poisson with mean γ. The risk function starts at

$$R_c(\delta^c, \theta) = p + c - \frac{(p-1)^2}{p+c} - 2(1+c)\frac{p-1}{p+c} = \frac{(1+c)^2}{p+c},$$

for $\theta = 0$, and then increases continuously towards the minimax risk $p + c$. As illustrated in Fig. 1, the improvement over $\delta_0 = Y$ is substantial for small to moderate values of γ, and always lies below the risk of the usual estimator δ_0.

3.3. Loss function robustness

Robustness of performance statements with respect to the specific loss function used is often overlooked in the literature, as if the loss function worked with had been handed down from above with absolute precision. The matter
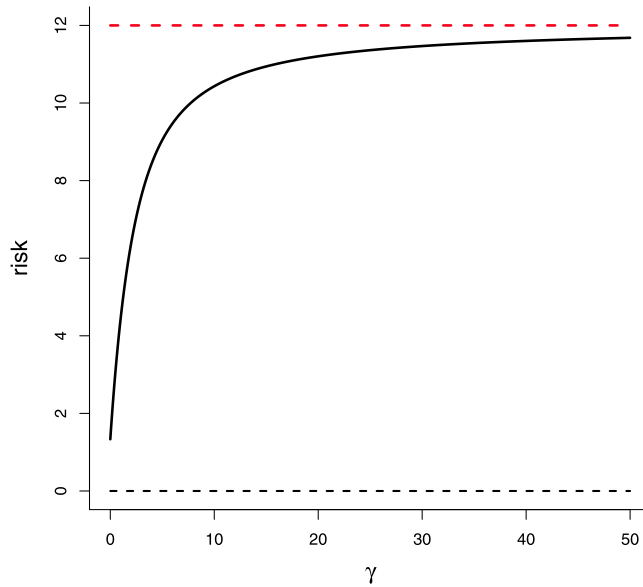
Fig. 1. The risk functions of δ^c of (13) (full curve) and of $\delta_0 = Y$ (constant, slanted), with $p = 9$ and $c = 3$.

is clearly of importance, however, as pointed to in comments by both Berger [3] and Morris [21]. If an estimator performs well with respect to one loss function Lo_a, but not for another loss function Lo_b, say, even when these two are close, it is a cause for concern. We argue here, however, that our estimator (13), derived under loss function L_c of (3), is somewhat robust regarding the precise value of c.

First consider the estimator δ_{CZ} of (2), which uses $\phi_0(z) = (p-1)/\{p-1+z\}$. It satisfies the second requirement of Theorem 2, i.e., $\phi_0(z)z < \phi_0(z+1)(z+1)$ for all z, but it does not necessarily satisfy the first requirement, namely $0 < \phi(z) \leq 2(p-1)/\{p-1+(1+c)z\}$ for all z. It is easy to verify that if $0 \leq c \leq 1$, then it does satisfy this requirement, however, showing that δ_{CZ} has certain robustness properties with respect to the L_c loss function: It is minimax and uniformly dominates δ_0 under L_c, provided $0 \leq c \leq 1$.

One can similarly study how the $\delta^{c_0}(Y) = [(1+c_0)Z/\{p-1+(1+c_0)Z\}] Y$ of (13) fares when the loss function is not quite the L_{c_0} under which it was derived, but rather L_c, with another penalty value of c, that is, with a somewhat different penalty paid to incorrect estimation of the sum γ. From the first condition of Theorem 2 we see that δ^{c_0} is still minimax and uniformly better than $\delta_0 = y$ under L_c, provided that $0 \leq c \leq 2c_0 + 1$. An immediate implication of this is that all estimators in the class $\mathcal{D}_c$ are minimax and uniformly better than δ_0 under the L_1^* loss function, showing that the more prudent estimation strategy δ^c, in the sense that it shrinks less that δ_{CZ}, is robust to $c = 0$.

4. Bayes, empirical Bayes, and admissibility

In this section a certain class of priors is studied along with Bayes and empirical Bayes consequences. The estimator δ^c of (13) will be shown to be both a generalised Bayes estimator with respect to a certain noninformative prior, and an empirical Bayes estimator with respect to independent Gamma priors. In addition, a class of proper Bayes estimators belonging to the class $\mathcal{D}_c$ is derived from another prior construction. Finally, we are also able to show that for each given c, the δ^c of (13) is an admissible estimator, under the L_c loss function. In particular, the Clevenson–Zidek estimator (2) is admissible under the L_0 function for which it was derived. Our result hence generalises that of Johnstone [17, Theorem 4.1].

4.1. Priors with sum independent of proportions

In the following we model the means $(\theta_1, \ldots, \theta_p)$ in terms of the sum $\gamma = \sum_{i=1}^p \theta_i$ and proportions $\pi_i = \theta_i/\gamma$ for $i \in \{1, \ldots, p\}$.

Lemma 1. *Suppose that γ and $\pi = (\pi_1, \ldots, \pi_p)$ are independent with simultaneous prior density $q(\gamma)r(\pi_1, \ldots, \pi_{p-1})$. This corresponds to a density*

$$q\left(\sum_{i=1}^p \theta_i\right) \frac{1}{(\sum_{i=1}^p \theta_i)^{p-1}} \, r\left(\frac{\theta_1}{\sum_{i=1}^p \theta_i}, \ldots, \frac{\theta_{p-1}}{\sum_{i=1}^p \theta_i}\right),$$

for $(\theta_1, \ldots, \theta_p)$. *The posterior distribution for θ given the data is on the same form: γ and $(\pi_1, \ldots, \pi_p)$ are still independent, and*

$$\gamma \mid y \sim \text{const.} \, \gamma^z e^{-\gamma} q(\gamma), \quad \pi \mid y \sim \text{const.} \, \pi_1^{y_1} \cdots \pi_{p-1}^{y_{p-1}} (1 - \pi_1 - \cdots - \pi_{p-1})^{y_p} r(\pi_1, \ldots, \pi_{p-1}).$$

Proof. The first part in an exercise in the transformation of random variables, involving calculating the determinant of Jacobi matrix $|\partial(\gamma, \pi_1, \ldots, \pi_{p-1})/\partial(\theta_1, \ldots, \theta_p)| = \left(\sum_{i=1}^p \theta_i\right)^{-(p-1)}$. The second part follows because the combined Poisson likelihood $\prod_{i=1}^p \theta_i^{y_i} e^{-\theta_i}/y_i!$ is proportional to $\gamma^z e^{-\gamma} \pi_1^{y_1} \cdots \pi_{p-1}^{y_{p-1}} (1 - \pi_1 - \cdots - \pi_{p-1})^{y_p}$. $\square$

If in particular $(\pi_1, \ldots, \pi_p)$ has a Dirichlet prior distribution with parameters $(\alpha_1, \ldots, \alpha_p)$, then the posterior is another Dirichlet with updated parameters $(\alpha_1 + y_1, \ldots, \alpha_p + y_p)$, and this holds regardless of the prior used for γ. The important case of independent θ_i from a Gamma (α, β) corresponds to a Gamma prior $(p\alpha, \beta)$ for γ and an independent Dirichlet $(\alpha, \ldots, \alpha)$ for the proportions.

Suppose $(\pi_1, \ldots, \pi_p)$ comes from a symmetric Dirichlet $(\alpha, \ldots, \alpha)$ independent of γ, the latter coming from a suitable prior $q(\cdot)$. The Bayes estimator under the L_c loss function takes the form (10), i.e., $\delta_i = \{(1 + c)/(1 + ca/b)\}a_i$, with

$$a_i = \left\{ \mathrm{E}(\gamma^{-1}\pi_i^{-1} \mid y) \right\}^{-1} = \left\{ \mathrm{E}(\gamma^{-1} \mid y) \right\}^{-1} \left\{ \mathrm{E}(\pi_i^{-1} \mid y) \right\}^{-1} = \frac{K(z)}{K(z-1)} \frac{\alpha + y_i - 1}{p\alpha + z - 1},$$

and $b = \left\{ \mathrm{E}(\gamma^{-1} \mid Y) \right\}^{-1} = K(z)/K(z-1)$, writing $K(z) = \int_0^\infty \gamma^z e^{-\gamma} q(\gamma)\, d\gamma$. Letting in particular π be uniform over the simplex we obtain the estimator

$$\delta_{B,i}(y) = \frac{1 + c}{1 + ca/b} a_i = \frac{K(z)}{K(z-1)} \frac{1 + c}{p - 1 + (1 + c)z} y_i. \tag{14}$$

We are now in a position to give three pleasing interpretations of the estimator δ^c of (13).

First, it is a generalised Bayes estimator. For if γ is given a flat prior $q(\gamma) = 1$ on the halfline, then $K(z) = \Gamma(z + 1)$, which inserted in (14) gives $\delta_B = \delta^c$.

Second, it is a limit of Bayes estimators. Let the proportions $(\pi_1, \ldots, \pi_p)$ have a flat Dirichlet $(1, \ldots, 1)$ prior, with an independent γ from a Gamma $(1, \beta)$. The Bayes solution is then

$$\delta_i^B = \left(1 - \frac{p - 1}{p - 1 + (1 + c)z}\right) \frac{y_i}{1 + \beta},$$

and the limit as $\beta \to 0$ is again δ_i^c of (13).

Thirdly, it is a natural empirical Bayes estimator. To see one of several such constructions let the θ_i be independent $\mathcal{G}(1, \beta)$. The corresponding exact Bayes solution is given in (11). Now β is to be estimated from the data. The marginal distribution of Z is found from the facts that Z given γ is Poisson with mean γ, and γ is Gamma distributed with parameters (p, β), so

$$\Pr\{Z = z\} = \frac{\beta^p}{\Gamma(p)} \frac{1}{z!} \frac{\Gamma(p + z)}{(1 + \beta)^{p+z}}, \quad z \in \{0, 1, 2, \ldots\}.$$

The sum Z is sufficient for β and $Z/(p - 1 + Z)$ is found to be an unbiased estimator for $1/(1 + \beta)$, for any $p \geq 2$. Inserting this data-based value in (11) produces once again δ^c.

Finally, the Gamma $(1, \beta)$ prior construction may be extended to an hierarchical setup where a prior is put on the hyperparameter β. This extends the approach of Ghosh and Parsian [12] from the L_1^* of (1) to the L_c setting. Let the parameter β be distributed according to

$$s(\beta) \propto \beta^{\eta - 1}(\beta + 1)^{-(\eta + \zeta)}. \tag{15}$$

Utilising that $\mathrm{E}(\theta_i^{-1} \mid y) = \mathrm{E}\,\mathrm{E}(\theta_i^{-1} \mid y, \beta) = \mathrm{E}\{(1 + \beta)/y_i \mid y\}$, the Bayes solution is then

$$\delta_i^B(y) = \frac{(1 + c)(p - 1 + z)}{p - 1 + (1 + c)z} \frac{y_i}{\mathrm{E}(1 + \beta \mid z)}, \tag{16}$$

since Z is sufficient for β. Also, β given $Z = z$ is distributed as

$$\beta \mid z \sim \{B(\eta + p, \zeta + z)\}^{-1} \beta^{\eta + p - 1}(1 + \beta)^{-(\eta + \zeta + z + p)},$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the Beta function. Provided β comes from the $s(\beta)$ of (15), and $0 < \eta \leq (p - 2 - c)/(1 + c)$, then the Bayes solution in (16), by virtue of belonging to the class $\mathcal{D}_c$, is minimax and uniformly dominates δ_0. To see this, insert $\mathrm{E}(1 + \beta \mid z) = (p + \eta + \zeta + z - 1)/(z + \zeta - 1)$ in (16), yielding

$$\delta_i^B(y) = \frac{(1 + c)(p - 1 + z)}{p - 1 + (1 + c)z} \frac{z + \zeta - 1}{p + \eta + \zeta + z - 1} y_i. \tag{17}$$

By some algebra we obtain that for the Bayes solution we here consider

$$\psi(z) = (1 + c)\frac{(p - 1 + z)(p + \eta)}{p - 1 + \eta + \zeta + z} - c(p - 1).$$

This function is non-decreasing for all $z \geq 0$. Moreover, we see that $\sup_{z \geq 0} \psi(z) = (1+c)(p+\eta) - c(p-1) \leq 2(p-1)$, since $\eta \leq (p-2-c)/(1+c)$. This means that the class of Bayes solutions in (17), with η obeying the constraint mentioned, satisfy both conditions of Theorem 2.

Before coming to our demonstration of admissibility of the (13) estimator, we note that constructions of probability distributions for multivariate count data somewhat similar to those treated above have recently been considered in Jones and Marchand [18], involving concepts of sums-and-shares.

4.2. Admissibility

So far we have studied estimators that dominate the maximum likelihood estimator under the L_c loss function. In this section we prove that the estimator δ^c of (13) cannot be uniformly improved upon, that is, δ^c is admissible. The basic ingredient in this proof is the characterisation of admissibility given by Brown and Farrell [4, Theorem 2.6]. According to this theorem an estimator δ is admissible for θ under $(\delta - \theta)^2/\theta$ if and only if there exists a sequence of finite measures ν_n such that the Bayes solution with respect to ν_n, say δ_n, converges to δ as $n \to \infty$, and

$$\lim_{n \to \infty} \{\mathrm{BR}(\delta, \nu_n) - \mathrm{MBR}(\nu_n)\} = 0.$$

This prior sequence has to satisfy certain conditions, the details of which are stated in Brown and Farrell [4] and Johnstone [17, pages 237-238]. For our purpose, it is sufficient to know that such a sequence exists if δ is admissible.

Consider the class of estimators given by

$$\delta_i = \frac{(1+c)\kappa(Z)}{p-1+(1+c)Z} Y_i = \frac{(1+c)Y_i}{p-1+(1+c)Z} \kappa(Z), \quad i \in \{1, \ldots, p\}. \tag{18}$$

The Bayes estimators of (14) are on this form with $\kappa(z) = K(z)/K(z-1)$; in particular the estimator δ^c of (13) is on this form, with $\kappa(z) = z$. As in Johnstone [17], it turns out that estimators of the class (18) are admissible provided that $\kappa(Z)$ is admissible for γ under the loss function $(\kappa - \gamma)^2/\gamma$. The theorem below is in part a restatement of his Theorem 4.1.

Theorem 3. *Estimators of the form* (18) *are admissible for* $(\theta_1, \ldots, \theta_p)$ *under* L_c *if and only if* $\kappa(Z)$ *is admissible for* γ *under* $L(\gamma, \delta) = (\delta - \gamma)^2/\gamma$.

Proof. The risk function of (18) under L_c can be written

$$R_c(\delta, \theta) = \mathrm{E}_\gamma \left\{ \frac{(1+c)^2 Z}{p-1+(1+c)Z} L(\gamma, \kappa(Z)) + \frac{(p-1)(1+c)\gamma}{p-1+(1+c)Z} \right\},$$

with $L(\gamma, \delta) = (\kappa - \gamma)^2/\gamma$. Introduce $u(z) = (1+c)^2 z/\{p-1+(1+c)z\}$, and write $\bar{L}(\delta, \gamma, z) = u(z)L(\gamma, \delta)$. Let δ and δ' be of the form (18) with $\kappa(Z)$ and $\kappa'(Z)$ respectively. If $\mathrm{E}_\gamma\{\bar{L}(\kappa'(Z), \gamma, Z)\} - \mathrm{E}_\gamma\{\bar{L}(\kappa(Z), \gamma, Z)\} \leq 0$, with strict inequality for some γ, then $R_c(\delta, \theta) - R_c(\delta', \theta) \leq 0$ with strict inequality for some θ. This shows that if $\kappa(Z)$ is inadmissible under L^*, then δ is inadmissible under L_c. The contrapositive statement is more enlightening: If δ is admissible under L_c then $\kappa(z)$ is admissible under $\bar{L}$. Conversely, assume that $\kappa(Z)$ is admissible for γ under L, let $\{\nu_n(\mathrm{d}\gamma)\}_{n \geq 1}$ be a sequence of prior measures satisfying the conditions of Brown and Farrell [4, Theorem 2.6], and let $\rho_n(A) = \Gamma(p) \int_A \nu_n(\mathrm{d}\gamma) \prod_{i=1}^p \mathrm{d}\pi_i$ be the prior measure over $(0, \infty) \times S$, where S is the $(p-1)$-dimensional simplex. Then, since $0 < u(z) < 1+c$ for all $z \geq 1$,

$$\mathrm{BR}_c(\delta, \rho_n) - \mathrm{BR}_c(\delta_n, \rho_n) \leq (1+c) \int \mathrm{E}_\gamma \{L(\gamma, \kappa(Z)) - L(\gamma, \kappa_n(Z))\} \, \mathrm{d}\nu_n(\gamma).$$

The right hand side is non-negative for all n since δ_n is the Bayes solution, and it also tends to zero by Theorem 2.6 in Brown and Farrell [4] because $\kappa(z)$ is admissible. This implies admissibility of δ under L_c. $\square$

The immediate corollary to Theorem 3 is that the estimator δ^c of (13) is admissible under L_c, because $\kappa(Z) = Z$ is admissible under the weighted squared error loss function $L(\gamma, \delta) = (\delta - \gamma)^2/\gamma$. As a special case, the Clevenson–Zidek estimator (2) is admissible under the $\sum_{i=1}^p (\delta_i - \theta_i)^2/\theta_i$ loss function.

5. Smoothing towards the mean

In the following we consider different strategies for smoothing the maximum likelihood estimator towards the mean of the observations. Pushing the maximum likelihood estimator towards a data-based point should in many cases yield more reduction in risk than pushing δ_0 towards the origin, particularly when the θ_i are not small and not too spread out. This is clearly visible in Fig. 1 where the improvement in risk of δ^c compared to δ_0 becomes smaller as $\gamma = \sum_{i=1}^p \theta_i$ grows; the improvement in risk is, in other words, biggest near the 'point of attraction'.

Ghosh et al. [11] considered a modification of δ_{CZ} that shrinks the maximum likelihood estimator towards the minimum of the observations, and were able to prove uniform dominance under the weighted loss L_1^* of (1), for their modified estimator. Albert [2] took the Bayes estimator under $L_0(\theta, \delta) = \sum_{i=1}^p (\delta_i - \theta_i)^2$ as his point of departure, and developed an

estimator that pushes the observations towards the mean $\bar{y}$ and performs better than the usual estimator in large parts of the parameter space.

A complication when working with the L_c loss function of (4) is that we compete with the maximum likelihood estimator on two turfs, so to speak, namely under $\sum_{i=1}^{p}(\delta_i-\theta_i)^2/\theta_i$ and under $\left(\sum_{i=1}^{p}\delta_i-\gamma\right)^2/\gamma$. One reason for choosing the mean as our point of attraction is that several of the estimators we construct preserve the mean, that is, $\sum_{i=1}^{p}\delta_i(Y)=Z$. In this way such new estimators will always match the risk performance of δ_0 when it comes estimating γ, and the penalty parameter c becomes immaterial.

5.1. A restricted minimax estimator

Consider estimators of the form

$$\delta_i(Y) = Y_i - g(Z)(Y_i - \bar{Y}), \tag{19}$$

where $\bar{Y} = Z/p$. Notice that $\sum_{i=1}^{p}\delta_i = Z = \sum_{i=1}^{p}\delta_{0,i}$, which means that in calculations of the difference in risk between estimators of the form (19) and $\delta_0 = Y$, the c-term in the L_c loss function disappears. The value of $g(0)$ is immaterial, since δ is then equal to zero, and we may take $g(0) = 0$ for convenience.

Using that Y given $Z = z$ is multinomial with cell probabilities π_i, for each $z \geq 1$, the risk difference $R_c(\delta, \theta) - R_c(\delta_0, \theta)$ can be expressed as

$$d(\theta) = E_\theta\Big(g(Z+1)^2\left[p-1+(Z+1)\{B(\pi)/p-1\}\right] - 2(p-1)g(Z+1)\Big),$$

writing

$$B(\pi) = p^{-1}\sum_{i=1}^{p}\pi_i^{-1}, \tag{20}$$

for the mean of the inverse proportions. It is not possible to find a function $g(\cdot)$ such that the estimator in (19) dominates $\delta_0 = Y$ over the entire parameter space $\Theta = (0,\infty)^p$. This is the 'tyranny of the boundary' phenomenon; as a single $\theta_i \to 0$ the sum $B(\pi) \to \infty$ and the risk blows up for non-null choices of g.

Estimators can be found, however, that dominate δ_0 over large proportions of the parameter space. Let $\Theta(b_0) \subset \Theta$ be the subset of the parameter space where $B(\pi) \leq pb_0$. The minimum value of $B(\pi)$ is p, so $b_0 > 1$. In some situations one might have prior grounds for believing that the θ_i are somewhat similar, that is, not too far from the mean $\bar{\theta}$. One might for example have that each $\theta_i \geq \varepsilon\bar{\theta} = \varepsilon\gamma/p$, or equivalently $\pi_i \geq \varepsilon/p$, for some small $\varepsilon > 0$. This implies that $B(\pi) \leq p/\varepsilon$, so $b_0 = 1/\varepsilon$ may be used, so the risk difference can be bounded:

$$d(\theta) \leq E_\theta\, g(Z+1)\left\{g(Z+1)[p-1+(Z+1)(b_0-1)] - 2(p-1)\right\}. \tag{21}$$

Based on this upper bound we derive the estimator with

$$\delta_i^*(Y) = Y_i - \frac{p-1}{p-1+(b_0-1)Z}(Y_i - \bar{Y}), \quad i \in \{1,\dots,p\}. \tag{22}$$

This estimator succeeds in having

$$R_c(\delta^*,\theta) \leq R_c(\delta_0,\theta) - E_\theta\frac{(p-1)^2}{p-1+(b_0-1)(Z+1)} < p+c-(p-1)^2/\{p-1+(b_0-1)(\gamma+1)\},$$

for all $\theta \in \Theta(b_0)$, with the last inequality following from Jensen's.

5.2. A restricted Bayes estimator

The estimator in (22) was derived with the aim of risk function dominance in a given large region $\Theta(b_0)$ of the parameter space. We may also derive the restricted Bayes solution, that is, the best estimator among those of the form (19), under a prior of the type discussed in Section 4. Since γ and π are independent, the Bayes risk of such an estimator is

$$BR_c(\delta) = p+c-E^q\,E_\gamma\left\{2(p-1)g(Z+1)-g(Z+1)^2[p-1+\{E^r B(\pi)/p-1\}(Z+1)]\right\}, \tag{23}$$

where $E^q(\cdot)$ and $E^r(\cdot)$ are the expectations of γ and π with respect to their prior distributions, see Lemma 1. If we now let $E^r B(\pi) = pb_0$, this reproduces the estimator in (22), but with a differently interpreted b_0. The risk function is

$$R_c(\delta^B,\theta) = p+c-(p-1)E_\gamma\,g(Z+1)\left[2-\frac{p-1+\{B(\pi)/p-1\}(Z+1)}{p-1+(b_0-1)(Z+1)}\right],$$

with $g(z) = (p-1)/\{p-1+(b_0-1)z\}$. Consider again the prior construction of Section 4. If $(\pi_1,\dots,\pi_p)$ comes from a Dirichlet distribution with parameters $(\alpha,\dots,\alpha)$, then $b_0 = (\alpha-1/p)/(\alpha-1)$, provided that $\alpha > 1$. Recall that the inequality in (21) only holds if $b_0 > 1$, which means that the subspace $\Theta(b_0)$ is empty if our prior knowledge dictates

$0 < \alpha \leq 1$. On the other side of the spectrum, as $\alpha \to \infty$ this estimator would assign the estimate $1/p$ to each of the proportions, and since b_0 goes to one as α goes to infinity, the estimator (22) tends to $\bar{y} = z/p$. In other words, having smaller values of $\alpha > 1$ expands the space $\Theta(b_0)$ but results in less gain in risk.

The (22) estimator may also use a data-based value for b_0. With π and γ independent, the marginal distribution of the data is

$$\int_S \int_0^\infty f(y \mid \pi, \gamma) r(\pi) q(\gamma) \, d\gamma \, d\pi_1 \cdots d\pi_{p-1} \propto \frac{\Gamma(p\alpha)}{\Gamma(\alpha)^p} \frac{\prod_{i=1}^p \Gamma(\alpha + y_i)}{\Gamma(p\alpha + z)}.$$

This likelihood can be maximised to obtain an estimate $\widehat{\alpha}$ which is plugged into $b_0 = (\alpha - 1/p)/(\alpha - 1)$, again provided that $\widehat{\alpha} > 1$.

5.3. More careful smoothing towards the mean

We now consider the construction

$$\widehat{\theta}_i = Y_i - g(Z)(Y_i - \bar{Y})h(Y),$$

where $h(y)$ is a function such that if one or more $y_i = 0$, then $h(y) = h(y_1, \ldots, y_p) = 0$. The intention is that of more careful smoothing towards the mean than with (19), to achieve risk improvement in potentially larger parts of the parameter space. Note that $\sum_{i=1}^p \widehat{\theta}_i = \sum_{i=1}^p y_i$, so any risk difference $R_c(\widehat{\theta}, \theta) - R_c(Y, \theta)$ does not depend on the c term of the loss function. Also, the value of $g(z)$ at $z = 0$ is immaterial, so we may take $g(0) = 0$, for convenience. Using now familiar techniques, the risk difference $R_c(\widehat{\theta}, \theta) - R_c(Y, \theta)$ may hence be expressed as

$$E_\theta \sum_{i=1}^p \{g(Z)^2(Y_i - \bar{Y})^2 h(Y)^2 - 2g(Z)(Y_i - \bar{Y})h(Y)(Y_i - \theta_i)\}/\theta_i = E_\theta D(Y),$$

with

$$D(y) = g(z+1)^2 \left\{ \sum_{i=1}^p (y_i + 1)h(y + e_i)^2 - 2\frac{z+1}{p} \sum_{i=1}^p h(y + e_i)^2 + \frac{(z+1)^2}{p^2} \sum_{i=1}^p \frac{h(y + e_i)^2}{y_i + 1} \right\}$$

$$- 2g(z+1) \left\{ \sum_{i=1}^p (y_i + 1)h(y + e_i) - \frac{z+1}{p} \sum_{i=1}^p h(y + e_i) \right\}.$$

The property that $y_i = 0$ implies $h(y) = 0$ is actively used here; without such a constraint, more complicated terms need to be added to the $E_\theta D(Y)$ here.

Several choices may be considered and worked with for further fine-tuning, regarding the $h(y)$ function. For the present article we limit attention to the special case of $h_0(y) = I\{$each $y_i \geq 1\}$. We need to study and bound the $D(y)$ function with this choice of $h(y)$. Note that $\sum_{i=1}^p h_0(y + e_i)$ is p, if all $y_i \geq 1$; is 1, if one and only one of $y_i = 0$; and is 0, otherwise. Similarly, $\sum_{i=1}^p y_i h_0(y + e_i) = z$, if all $y_i \geq 1$; and 0, otherwise. Furthermore,

$$\sum_{i=1}^p \frac{h_0(y + e_i)}{y_i + 1} \begin{cases} \leq p/2, & \text{if all } y_i \geq 1, \\ = 1, & \text{if only one } y_i = 0, \\ = 0, & \text{otherwise.} \end{cases}$$

(i) Assume first that all $y_i \geq 1$. Then

$$D(y) = g(z+1)^2 \left\{ z + p - 2(z+1) + \frac{(z+1)^2}{p^2} \sum_{i=1}^p \frac{1}{y_i + 1} \right\} - 2g(z+1)(p-1)$$

$$\leq g(z+1)^2 \left\{ p - 1 - (z+1) + \frac{1}{2}\frac{(z+1)^2}{p} \right\} - 2g(z+1)(p-1).$$

The function $p - 1 - x + \frac{1}{2}x^2/p$ for $x \geq 1$ has its minimum value at position $x = p$, with minimum value $\frac{1}{2}p - 1$, which is positive as long as $p \geq 3$. (ii) When there is only one $y_i = 0$, the rest $y_j \geq 1$, we find

$$D(y) = g(z+1)^2 \left\{ 0 + 1 - 2\frac{z+1}{p}1 + \frac{(z+1)^2}{p^2}1 \right\} - 2g(z+1)\left\{ 0 + 1 - \frac{z+1}{p}0 \right\}$$

$$= g(z+1)^2 \left\{ -2\frac{z+1}{p} + \frac{(z+1)^2}{p^2} \right\} - 2g(z+1).$$

(iii) Otherwise, which means that the number of $y_i = 0$ is between 2 and $p - 1$, we find that $D(y) = 0$. Our best choice for $g(z)$, based on the upper risk bound for the case of all $y_i \geq 1$, is $g_0(z) = (p - 1)/\{p - 1 - z + z^2/(2p)\}$. The estimator

$$\widehat{\theta}_i = Y_i - g_0(Z)(Y_i - \bar{Y})h_0(Y) = Y_i - \frac{p-1}{p - 1 - Z + \frac{1}{2}Z^2/p}(Y_i - \bar{Y})I\{\text{all } Y_i \geq 1\}, \tag{24}$$

therefore, has risk function $R_c(\widehat{\theta}) = p + c + \mathrm{E}_\theta D_0(Y)$, where an exact expression for $D_0(Y)$ is found via the above. We also know that

$$D_0(y) \leq -\frac{(p-1)^2}{p - 1 - (z+1) + \frac{1}{2}(z+1)^2/p}, \quad y_i \geq 1 ;$$

and that

$$D_0(y) = \left\{ \frac{p-1}{p - 1 - (z+1) + \frac{1}{2}(z+1)^2/p} \right\}^2 \left\{ -2\frac{z+1}{p} + \frac{(z+1)^2}{p^2} \right\} - 2\frac{p-1}{p - 1 - (z+1) + \frac{1}{2}(z+1)^2/p},$$

for the cases of y where precisely one $y_i = 0$, the other $y_j \geq 1$; and finally that $D_0(y) = 0$ for those y for which the number of $y_i = 0$ is between 2 and $p - 1$. As long as all θ_i are at least moderate, so that $h_0(Y) = 1$ with high probability, there is clear risk improvement on the minimax risk $p + c$. The corner cases, however, where one θ_i is small and the others not, are the 'bad cases' for the (24) estimator, where the risk might become larger than $p + c$. Since $D_0(y)$ flattens to zero when $z = \sum_{i=1}^p y_i$ increases, the risk converges to $p + c$ for all $\theta = \gamma(\pi_1, \ldots, \pi_p)$ where γ tends to infinity.

6. Bayes and empirical Bayes with more structure

In the previous sections we have constructed and analysed estimators essentially symmetric in the observations. Sometimes some structure is anticipated in the parameters, however, as with setting up regressions or log-linear models for Poisson tables, see, e.g., Agresti [1], or with classes of priors. The present section briefly complements our earlier efforts by examining risk function consequences for estimators that favour asymmetric structures.

6.1. Risk functions with Gamma priors

A natural class of priors takes independent $\mathcal{G}(\alpha_i, \beta)$ priors for the θ_i. As was seen in Section 3, the Bayes estimator then takes the form

$$\delta_i^B = h_c(z)\frac{\alpha_i + y_i - 1}{\beta + 1}, \quad h_c(z) = \frac{(1+c)(p\bar{\alpha} + z - 1)}{(1+c)(p\bar{\alpha} + z - 1) - c(p-1)}, \tag{25}$$

writing $\bar{\alpha}$ for $(1/p)\sum_{i=1}^p \alpha_i$; also, for cases where $\alpha_i < 1$ and $y_i = 0$, the estimator is zero.

The present task is to study the consequent risk functions, for such estimators, outside the special and simplest case where each $\alpha_i = 1$. The point will be that estimators then lose the minimax property, with the risk exceeding the $p + c$ benchmark level when one or more of the θ_i come close to zero, but that the risk otherwise can be lower than $p + c$ in big and reasonable parts of the parameter space. For simplicity of presentation we restrict attention here to the case of $c = 0$. Similar results and insights may be reached for the general loss function L_c, using more laborious methods we develop and exploit for somewhat different purposes in Section 6.2. For the ensuing estimator $(\alpha_i + Y_i - 1)/(\beta + 1)$, some calculations yield

$$R(\delta^B, \theta) = \sum_{i=1}^p \frac{1}{\theta_i}\left[\left\{\frac{\alpha_i + \theta_i - 1 - (\beta+1)\theta_i}{\beta + 1}\right\}^2 + \frac{\theta_i}{(\beta+1)^2}\right] = \frac{1}{(\beta+1)^2}\left\{\sum_{i=1}^p \frac{(\alpha_i - 1 - \beta\theta_i)^2}{\theta_i} + p\right\}.$$

This is smaller than or equal to $R(Y, \theta) = p$ when

$$A(\theta) = \frac{1}{p}\sum_{i=1}^p \frac{(\alpha_i - 1 - \beta\theta_i)^2}{\beta\theta_i} \leq \frac{(\beta+1)^2 - 1}{\beta} = 2 + \beta.$$

This defines a fairly large parameter region, $\{\theta: A(\theta) \leq 2 + \beta\}$, with the $1/\theta_i$ not being too far away from the prior mean values $\beta/(\alpha_i - 1)$, and where using the Bayes estimator hence is better than with the $\delta_0 = Y$ method. Under the prior itself, the random $A(\theta)$ has mean

$$\frac{1}{p}\sum_{i=1}^p \left\{\frac{(\alpha_i - 1)^2}{\beta}\frac{\beta}{\alpha_i - 1} - 2(\alpha_i - 1) + \beta\frac{\alpha_i}{\beta}\right\} = 1,$$

and variance of order $O(1/p)$, showing that only rather unlikely values of θ will have risk above the benchmark p.

6.2. Risks for a class of empirical Bayes estimators

With the independent Gamma priors used above we next note that the marginal distribution of y_i becomes

$$g(y_i \mid \alpha_i, \beta) = \int_0^\infty \frac{\beta^{\alpha_i}}{\Gamma(\alpha_i)}\theta_i^{\alpha_i - 1}\exp(-\beta\theta_i)\frac{1}{y_i!}\exp(-\theta_i)\theta_i^{y_i}\,d\theta_i = \frac{\beta^{\alpha_i}}{\Gamma(\alpha_i)}\frac{1}{y_i!}\frac{\Gamma(\alpha_i + y_i)}{(\beta+1)^{\alpha_i + y_i}},$$

for $y_i \in \{0, 1, 2, \ldots\}$. In the setup where the α_i are taken known but β an unknown parameter, we see that $Z = \sum_{i=1}^{p} Y_i$ is sufficient. Since Z given the parameters is Poisson with mean $\gamma = \sum_{i=1}^{p} \theta_i$, and γ is Gamma $(p\bar{\alpha}, \beta)$, writing $\bar{\alpha} = (1/p) \sum_{i=1}^{p} \alpha_i$, its distribution may be written

$$q(z \mid \beta) = \frac{\beta^{p\bar{\alpha}}}{\Gamma(p\bar{\alpha})} \frac{1}{z!} \frac{\Gamma(p\bar{\alpha} + z)}{(\beta + 1)^{p\bar{\alpha}+z}}, \quad z \in \{0, 1, 2, \ldots\}.$$

From this we can derive

$$\mathrm{E}\, \frac{Z}{p\bar{\alpha} - 1 + Z} = \sum_{z=1}^{\infty} \frac{1}{(z-1)!} \frac{\beta^{p\bar{\alpha}}}{\Gamma(p\bar{\alpha})} \frac{\Gamma(p\bar{\alpha} + z - 1)}{(\beta + 1)^{p\bar{\alpha}+z-1}} \frac{1}{\beta + 1} = \frac{1}{\beta + 1}, \tag{26}$$

provided only that $p \geq 2$. Hence $Z/(p\bar{\alpha} - 1 + Z)$ can be used as an estimator for the quantity $1/(\beta + 1)$. With pre-specified α_i, then, a natural empirical Bayes version of estimator (25) emerges:

$$\widehat{\theta}_i = \frac{(1+c)(p\bar{\alpha} - Z - 1)}{(1+c)(p\bar{\alpha} + Z - 1) - c(p-1)} \frac{Z}{p\bar{\alpha} - 1 + z}(\alpha_i + Y_i - 1) = h_c(Z)(\alpha_i + Y_i - 1),$$

with $h_c(Z) = \{(1+c)Z\}/\{(1+c)(p\bar{\alpha} + Z - 1) - c(p-1)\}$. In particular, under L_0 loss, with $c = 0$, the natural empirical Bayes estimator is

$$\widehat{\theta}_i = h_0(Z)(\alpha_i + Y_i - 1) = \frac{Z}{p\bar{\alpha} + Z - 1}(\alpha_i + Y_i - 1), \tag{27}$$

generalising the earlier symmetric case with all $\alpha_i = 1$, which yields the already studied minimax and admissible estimator (5).

Expressions for the risk function $R_c(\widehat{\theta}, \theta)$ can now be worked out, using the fact that $y_i \mid z$ is binomial (z, π_i), with $\pi_i = \theta_i/\gamma$. Consider the general class of estimators

$$\widehat{\theta}_i = h(Z)(\alpha_i + Y_i - 1), \quad i \in \{1, \ldots, p\}. \tag{28}$$

The $h(z)$ functions we are encountering all have $h(0) = 0$, and will in fact have the form $q(z)z$, for suitable $q(z)$. Also, they will be nondecreasing with $h(z) \to 1$ as z increases. The task now is to develop formulae for their risk functions, through suitable representations of the form

$$R(\widehat{\theta}, \theta) = \mathrm{E}_\theta\, H(Z, \theta) = \sum_{z=1}^{\infty} H(z, \theta) \exp(-\gamma)\gamma^z/z!,$$

and then showing, for relevant choices of $h(z)$, that this is less than $p + c$ for large and relevant parameter regions. We start from

$$\mathrm{E}_\theta\, \frac{1}{\theta_i}\{h(z)(\alpha_i + Y_i - 1) - \theta_i\}^2 \mid z = \frac{1}{\gamma\pi_i}[\{h(z)(\alpha_i + z\pi_i - 1) - \gamma\pi_i\}^2 + h(z)^2 z\pi_i(1 - \pi_i)]$$

$$= \frac{1}{\gamma\pi_i}\{[h(z)(\alpha_i - 1) + \{h(z)z - \gamma\}\pi_i]^2 + h(z)^2 z\pi_i(1 - \pi_i)\}.$$

For the case of $c = 0$ this leads to

$$H(z, \theta) = h(z)^2 \sum_{i=1}^{p} \frac{(\alpha_i - 1)^2}{\gamma\pi_i} + \{h(z)z - \gamma\}^2 \frac{1}{\gamma} + 2\{h(z)z - \gamma\}h(z)(p\bar{\alpha} - p)\frac{1}{\gamma} + h(z)^2 z(p-1)\frac{1}{\gamma}.$$

For the c-term part of the risk, note that $\widehat{\theta}_i = h(z)(p\bar{\alpha} + z - p)$, leading to

$$\mathrm{E}_\gamma\, \frac{\{h(Z)(p\bar{\alpha} + Z - p) - \gamma\}^2}{\gamma} = \mathrm{E}_\gamma\, \frac{\{h(Z)Z - \gamma + h(Z)(p\bar{\alpha} - p)\}^2}{\gamma}.$$

This may be exploited further using the identity $\mathrm{E}_\gamma\, r(Z)/\gamma = \mathrm{E}_\gamma\, r(Z + 1)/(Z + 1)$ for functions $r(z)$ with $r(0) = 0$.

For brevity we limit attention here to the case of $c = 0$; extensions can be worked out using the same methods. We use the identity pointed to for $\{h(z)z - \gamma\}^2/\gamma = h(z)^2 z^2/\gamma - 2h(z)z + \gamma$, and find $R(\widehat{\theta}, \theta) = \mathrm{E}_\gamma\, Q(Z, \theta)$, with

$$Q(z, \theta) = h(z)^2 \sum_{i=1}^{p} \frac{(\alpha_i - 1)^2}{\gamma\pi_i} + h(z + 1)^2(z + 1) - 2h(z)z + \gamma$$

$$+ 2\frac{\{h(z + 1)(z + 1) - \gamma\}h(z + 1)}{z + 1}(p\bar{\alpha} - p) + h(z + 1)^2(p - 1).$$

The risk function may hence be expressed as

$$R(\widehat{\theta}, \theta) = pm_1(\gamma)C(\pi)/\gamma + R_1(\gamma) + 2(p\bar{\alpha} - p)R_2(\gamma) + (p - 1)m_2(\gamma), \tag{29}$$
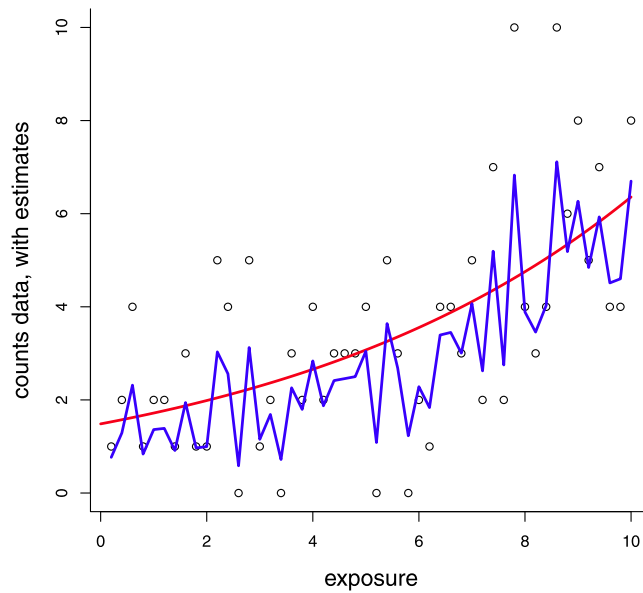
Fig. 2. Estimating $p = 50$ Poisson parameters, in a simulated regression setting with (x_i, y_i) data, with the empirical Bayes method (27), shrinking the raw estimates y_i towards prior means α_i/β, with β estimated from the data.

in which $C(\pi) = (1/p) \sum_{i=1}^{p} (\alpha_i - 1)^2 / \pi_i$, and

$$m_1(\gamma) = \mathrm{E}_\gamma\, h(Z)^2, \quad m_2(\gamma) = \mathrm{E}_\gamma\, h(Z+1)^2,$$

$$R_1(\gamma) = \mathrm{E}_\gamma \{h(Z+1)^2(Z+1) - 2h(Z)Z + \gamma\}, \quad R_2(\gamma) = \mathrm{E}_\gamma\, \frac{\{h(Z+1)(Z+1) - \gamma\}h(Z+1)}{Z+1}.$$

Here $m_1(\gamma)$ and $m_2(\gamma)$ are inside $(0, 1)$, and increase to 1; the $R_1(\gamma)$ and $R_2(\gamma)$ are bounded and converge to respectively one and zero as γ increases. For larger γ, therefore, the risk goes to the minimax risk p. The risk function (29) may exceed the minimax threshold level p if one or more of the $\pi_i = \theta_i/\gamma$ are small, but even for small π_i the risk decreases with increasing γ. Otherwise the situation is that the risk may become significantly smaller than p in parts of the parameter space not disagreeing much from what is judged likely under the prior, and that it can be smaller than p also in other larger parameter regions. An upper bound is

$$R(\widehat{\theta}, \theta) \le pm_1(\gamma)\{\max_{i \le p} |\alpha_i - 1|\}^2 B(\pi)/\gamma + R_1(\gamma) + 2(p\bar{\alpha} - p)R_2(\gamma) + (p-1)m_2(\gamma),$$

with $B(\pi)$ as in (20). This may in particular be investigated further, with the choice $h_0(z)$, corresponding to the estimator (27). The risk function is bounded; converges to the minimax value p when γ increases, regardless of proportions $(\pi_1, \ldots, \pi_p)$; may offer substantial improvement for sizeable portions of the parameter space; and its maximum value is often not much bigger than p.

An illustration of the empirical Bayes strategy (27) is provided in Fig. 2, in a situation with simulated regression data (x_i, y_i). The prior takes the θ_i to stem from $\mathcal{G}(\alpha_i, \beta)$, with $\alpha_i = \exp(\gamma_0 + \gamma_1 x_i)$, for suitable prior guess values (γ_0, γ_1), and then estimates β from data, as per (26). Similar Bayes and empirical Bayes methods can be developed for priors of the type $\theta_i \sim \mathcal{G}(d\theta_{0,i}, d)$, with either hyperpriors on the prior parameters $\theta_{0,i}$ and d, or involving estimators for these from the data.

7. Estimation with a weighted loss function

Above we have worked with our loss function L_c, a natural extension of the Clevenson–Zidek loss function $L_1^*(\theta, \delta) = \sum_{i=1}^{p} (\delta_i - \theta_i)^2/\theta_i$, to account for not shrinking the mean too much. Another useful extension is to allow for weighting, with $L_w(\theta, \delta) = \sum_{i=1}^{p} w_i(\delta_i - \theta_i)^2/\theta_i$, where $w_1, \ldots, w_p$ are fixed, positive, and context driven, reflecting relative importance. This is, e.g., important when the y_i result from different levels of exposure, as with y_i stemming from a Poisson with parameter $w_i\theta_i$. The $\delta_0 = Y$ estimator is again the natural benchmark; it has constant risk $w_0 = \sum_{i=1}^{p} w_i$, and it is minimax. To prove this second claim, one may work with the prior where the θ_i are independent and Gamma distributed, with parameters (α, β), and with $\alpha \ge 1$ to avoid a certain technical issue. Some work shows that the Bayes estimator becomes $\widehat{\theta}_i = 1/\mathrm{E}(\theta_i \mid y) = (\alpha + y_i - 1)/(\beta + 1)$, with associated minimum Bayes risk as simple as $w_0/(\beta + 1)$. Letting $\beta \to 0$ we have convergence to the constant risk of δ_0.

This estimator may be uniformly improved upon, however, as we now demonstrate, yielding another generalisation of the Clevenson–Zidek estimator. Consider estimators of the form

$$\widehat{\theta}_i = \{1 - \phi(V)\}Y_i \quad \text{for } i \in \{1, \ldots, p\}, \quad V = v(Y) = \sum_{i=1}^{p} w_i Y_i. \tag{30}$$

Using the identity (12) we may express the risk difference between $\widehat{\theta}$ and δ_0 as

$$\mathrm{E}_\theta \{L_w(\theta, \widehat{\theta}) - L_w(\theta, \delta_0)\} = \mathrm{E}_\theta \sum_{i=1}^{p} w_i\{\phi(v + w_i)^2(Y_i + 1) - 2\phi(v + w_i)(Y_i + 1) + 2\phi(V)Y_i\} = \mathrm{E}_\theta\, D_w(Y),$$

say, where $D_w(y) = \sum_{i=1}^{p} w_i\{\phi(v + w_i)^2 - 2\phi(v + w_i)\}(y_i + 1) + 2\phi(v)v$. As we have argued on previous occasions in our paper, if we succeed in finding a function $\phi(v)$ such that $D_w(y) \le 0$ for all y, with strict inequality for at least one y, we have demonstrated inadmissibility of $\delta_0 = Y$.

To work with this we set up some mild requirements regarding the w_i weights. We take all w_i to be inside some $[a, b]$ interval, situated inside $(0, 1)$, and also stipulate that $w_0 = \sum_{i=1}^{p} w_i > 1$. For estimators of the form (30), consider $\phi(v(y)) = \psi(v(y))/\{w_0 - 1 + v(y)\}$, with (i) $\psi(v)$ nondecreasing, (ii) with $0 < \psi(v) < 2(w_0 - 1)$ for all $v = v(y)$, and (iii) such that $v\phi(v)$ is increasing. We then find

$$\begin{aligned}
D_w(y) &= \sum_{i=1}^{p} w_i \frac{\psi(v + w_i)}{w_0 - 1 + v + w_i}\left\{\frac{\psi(v + w_i)}{w_0 - 1 + v + w_i} - 2\right\}(y_i + 1) + 2\phi(v)v \\
&\le \sum_{i=1}^{p} w_i \frac{\psi(v + a)}{w_0 - 1 + v + b}\left\{\frac{\psi(v + b)}{w_0 - 1 + v + a} - 2\right\}(y_i + 1) + 2\phi(v)v \\
&= \frac{\psi(v + a)}{w_0 - 1 + v + b}\left\{\frac{\psi(v + b)}{w_0 - 1 + v + a} - 2\right\}(w_0 + v) + 2\phi(v)v \\
&\le \psi(v + a)\left\{\frac{\psi(v + b)}{w_0 - 1 + v + a} - 2\right\} + 2\phi(v)v \\
&= \psi(v + a)\{\psi(v + b) - 2(w_0 - 1)\} - 2\{\phi(v + a)(v + a) - \phi(v)v\},
\end{aligned}$$

which is demonstrably negative. Our preferred generalisation of the Clevenson–Zidek estimator, to the present case of weighted loss, becomes

$$\widehat{\theta}_i = \left\{1 - \frac{w_0 - 1}{w_0 - 1 + v(Y)}\right\}Y_i, \quad i \in \{1, \ldots, p\}. \tag{31}$$

The special case of equal weights $w_i = 1$ leads back to the Clevenson–Zidek estimator (2).

Remarkably, the apparatus above allows extension to the case of infinitely many Poisson parameters. Suppose $Y_1, Y_2, \ldots$ are independent Poisson counts with means $\theta_1, \theta_2, \ldots$, and that loss incurred by estimators $\delta_1, \delta_2, \ldots$ is taken to be $L(\theta, \delta) = \sum_{i=1}^{\infty} w_i(\delta_i - \theta_i)^2/\theta_i$. Here the sequence of weights is such that $w_0 = \sum_{i=1}^{\infty} w_i$ is finite, and the parameter space to be considered is Ω, the set of sequences of θ_i for which $\sum_{i=1}^{\infty} w_i\theta_i$ is finite (including in particular each bounded sequence). The benchmark procedure is again δ_0, with components $\delta_{0,i} = y_i$. It has constant risk w_0, and our previous arguments may be extended to demonstrate that this procedure is minimax. Also, crucially, the estimator

$$\widehat{\theta}_i = \left\{1 - \frac{w_0 - 1}{w_0 - 1 + v(Y)}\right\}Y_i, \quad i \in \{1, 2, \ldots\}, \tag{32}$$

offers uniform risk improvement over δ_0, where now $v(y) = \sum_{i=1}^{\infty} w_i y_i$. This follows from arguments used to reach the corresponding statement for the finitely-many procedure (31), in the light of two necessary remarks. The first is that the identity (12) continues to hold, for all $h(Y_1, Y_2, \ldots)$ with finite mean, with the property that $h(y) = 0$ as long as $y_i = 0$. The second is that the upper bound reached above for $D_w(y)$, under the condition that $a \le \theta_i \le b$ for all i, applies here too, but now we need $a = 0$, since the $\sum_{i=1}^{\infty} w_i$ is finite. In other words,

$$D_w(y) \le \psi(v(y))\{\psi(v(y) + b) - 2(w_0 - 1)\}$$

remains correct.

Above we stipulated a scaling of the importance weights w_i so that their sum $w_0 = \sum_{i=1}^{\infty} w_i$ is above 1. This is partly in order for the estimator (32) to be a natural generalisation of the Clevenson–Zidek estimator. Similar reasoning goes through for estimators $[\psi(v(Y))/\{w_0 - \varepsilon + v(Y)\}]Y$, if we instead stipulate $w_0 > \varepsilon$.

8. Multivariate models for count variables

The methods developed in Section 4 utilised certain constructions which also involve multivariate models for rates and for count variables, of interest in their own right. Models can be built with both positive and negative correlations between rates θ_i and between count observations Y_i. These modelling ideas also point to Bayesian nonparametrics, see our Remark C in Section 9.

8.1. Sum and proportions models

Suppose first in general terms that given $\theta = (\theta_1, \ldots, \theta_p)$, the observations $Y_1, \ldots, Y_p$ have independent Poisson distributions with these parameters, and that the θ has a background distribution, which we for simplicity of presentation here take to be symmetric with finite variances. Let us write $\mathrm{E}\,\theta_i = \theta_0$, $\mathrm{Var}\,\theta_i = \sigma_0^2$, $\mathrm{Cov}(\theta_i, \theta_j) = \rho\sigma_0^2$ for $i \neq j$. We then deduce

$$\mathrm{E}\,Y_i = \theta_0, \quad \mathrm{Var}\,Y_i = \theta_0 + \sigma_0^2, \quad \mathrm{Cov}(Y_i, Y_j) = \rho\sigma_0^2, \quad \mathrm{Corr}(Y_i, Y_j) = \frac{\rho\sigma_0^2}{\theta_0 + \sigma_0^2}. \tag{33}$$

A class of multivariate models for the θ_i, and by implication also for the Y_i, emerges from the construction of Lemma 1, with a prior $q(\gamma)$ for $\gamma = \sum_{i=1}^p \theta_i$ and a symmetric prior $\mathrm{Dir}(\alpha, \ldots, \alpha)$ for the proportions $\pi_i = \theta_i/\gamma$. Write $\mathrm{E}\,\gamma = \gamma_0 = p\theta_0$ and $\mathrm{Var}\,\gamma = p\tau_0^2$. The π_i have means $1/p$, variances $(1/p)(1 - 1/p)/(p\alpha + 1)$, and covariances $-(1/p^2)/(p\alpha + 1)$. From these facts we first find $\mathrm{E}\,\theta_i = \mathrm{E}(\gamma\pi_i) = \theta_0$ and then

$$\mathrm{Var}\,\theta_i = \mathrm{E}(\gamma\pi_i)^2 - \theta_0^2 = (\gamma_0^2 + p\tau_0^2)\left\{\frac{1}{p^2} + \frac{1}{p}\left(1 - \frac{1}{p}\right)\frac{1}{p\alpha + 1}\right\} - \theta_0^2 = \theta_0^2\frac{p - 1}{p\alpha + 1} + \tau_0^2\frac{\alpha + 1}{p\alpha + 1}.$$

Similarly, some calculations lead to $\mathrm{Cov}(\theta_i, \theta_j) = (\tau_0^2\alpha - \theta_0^2)/(p\alpha + 1)$, so that the correlation parameter ρ of (33) may be expressed as

$$\rho = \frac{\alpha\tau_0^2 - \theta_0^2}{(\alpha + 1)\tau_0^2 + (p - 1)\theta_0^2}.$$

For the special case of $\gamma \sim \mathcal{G}(p\alpha, \beta)$, we have $\theta_0 = \alpha/\beta$ and $\tau_0^2 = \alpha/\beta^2$, the covariance is zero, and the formula for $\mathrm{Var}\,\theta_i$ gives α/β^2. This is indeed the familiar case of independent $\theta_i \sim \mathcal{G}(\alpha, \beta)$.

For other models for γ, however, the construction above leads to useful multivariate models for $(\theta_1, \ldots, \theta_p)$ and for $(Y_1, \ldots, Y_p)$. In general terms, if γ has density $q(\gamma)$, the distribution of the data vector can be written as

$$f(y_1, \ldots, y_p) = \mathrm{E}\prod_{i=1}^p \exp(-\theta_i)\theta_i^{y_i}/y_i! = \mathrm{E}\,\exp(-\gamma)\gamma^z\pi_1^{y_1}\cdots\pi_p^{y_p}/(y_1!\cdots y_p!)$$

$$= K(z)\frac{\Gamma(p\alpha)}{\Gamma(\alpha)\cdots\Gamma(\alpha)}\frac{\Gamma(\alpha + y_1)\cdots\Gamma(\alpha + y_p)}{\Gamma(p\alpha + z)}\frac{1}{y_1!\cdots y_p!}, \tag{34}$$

with $K(z) = \int_0^\infty \exp(-\gamma)\gamma^z q(\gamma)\,d\gamma$ as in Section 4. If in particular $\gamma \sim \mathcal{G}(\alpha_0, \beta_0)$, then

$$f(y_1, \ldots, y_p) = \frac{\Gamma(\alpha_0 + z)}{\Gamma(p\alpha + z)}\frac{\Gamma(p\alpha)}{\Gamma(\alpha_0)}\frac{\Gamma(\alpha + y_1)}{\Gamma(\alpha)}\cdots\frac{\Gamma(\alpha + y_p)}{\Gamma(\alpha)}\frac{\beta_0^{\alpha_0}}{(\beta_0 + 1)^{\alpha_0 + z}}\frac{1}{y_1!\cdots y_p!}.$$

The point is that this has a multiplicative independence structure only if $\alpha_0 = p\alpha$, then with negative binomial marginal distributions. In this light, the (34) construction amounts to an extended class of models for count data, allowing both positive and negative correlations. The generalisation to the nonsymmetric case takes $\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_p)$, and leads to

$$f(y_1, \ldots, y_p) = K(z)\frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_p)}\frac{\Gamma(\alpha_1 + y_1)\cdots\Gamma(\alpha_p + y_p)}{\Gamma(\alpha_1 + \cdots + \alpha_p + z)}\frac{1}{y_1!\cdots y_p!},$$

with the particular choice $\gamma \sim \mathcal{G}(\sum_{i=1}^p \alpha_i, \beta)$ yielding independent negative binomials.

8.2. Poisson processes with dependence in time and space

Suppose independent Poisson processes $Y_1(t), \ldots, Y_k(t)$ are observed over a time period $[0, \tau]$, and divide this period into p cells or windows, say $(t_{j-1}, t_j]$. We take $Y_i(t)$ to have cumulative intensity function $G_i(t)$, with consequent Poisson parameters $\theta_{i,j} = G_i(t_j) - G_i(t_{j-1})$ for the counts $Y_{i,j} = Y_i(t_j) - Y_i(t_{j-1})$, for $j \in \{1, \ldots, p\}$. There are now different ways of modelling the $k \times p$ matrix of rate parameters, using aspects of the apparatus above. For simplicity of presentation we limit these brief pointers to the neutral cases, where the vectors of fractions involved come from symmetric Dirichlet distributions.

Idea (a) is to allow for dependence over time, for each process:

$$\theta_{i,j} = \gamma_i\pi_{i,j}, \quad j \in \{1, \ldots, p\}, \quad \gamma_i = \sum_{j=1}^p \theta_{i,j},$$

and with $(\pi_{i,1}, \ldots, \pi_{i,p})$ from a $\mathrm{Dir}(\alpha_i, \ldots, \alpha_i)$. The particular case of $\gamma_i \sim \mathcal{G}(p\alpha_i, \beta_i)$ corresponds to independent $\theta_{i,j} \sim \mathcal{G}(\alpha_i, \beta_i)$, $j \in \{1, \ldots, p\}$.

Idea (b) is to build dependence structure into the sequence of processes:

$$\theta_{i,j} = \kappa_j \pi_{i,j}, \quad i \in \{1, \ldots, k\}, \quad \kappa_j = \sum_{i=1}^{k} \theta_{i,j},$$

and with $(\pi_{1,j}, \ldots, \pi_{k,j})$ from a $\mathrm{Dir}(\alpha_j, \ldots, \alpha_j)$. If in particular $\kappa_j \sim \mathcal{G}(k\alpha_j, \beta_j)$, then we have independent $\theta_{i,j} \sim \mathcal{G}(\alpha_j, \beta_j)$, $i \in \{1, \ldots, k\}$.

Both ideas (a) and (b) have Bayesian counterparts, and motivate extensions of the L_c loss function. Suppose we are interested in precise estimates of the full $k \times p$ parameter matrix and in the cumulatives $\gamma_i = G_i(\tau)$. A natural loss function is then

$$L_c^*(\theta, \delta) = \sum_{i=1}^{k} L_c(\theta_i, \delta_i) = \sum_{i=1}^{k} \left\{ \sum_{j=1}^{p} (\delta_{i,j} - \theta_{i,j})^2 / \theta_{i,j} + c \Big(\sum_{j=1}^{p} \delta_{i,j} - \gamma_i \Big)^2 / \gamma_i \right\},$$

where L_c is the loss function introduced in Section 3, and c is a positive constant set by the statistician. The Bayes solution is

$$\delta_{i,j}^B = \frac{1+c}{1 + ca_i/b_i} \, a_{i,j}, \quad j \in \{1, \ldots, p\}, \ i \in \{1, \ldots, k\},$$

where $a_{i,j} = \{\mathrm{E}(\theta_{i,j}^{-1} \mid y)\}^{-1}$, $a_i = \sum_{j=1}^{p} a_{i,j}$, and $b_i = \{\mathrm{E}(\gamma_i^{-1} \mid y)\}^{-1}$. Suppose prior knowledge dictates that intensities of the k processes are functionally somewhat alike, but at different levels; then a natural prior construction takes $(\pi_{i,1}, \ldots, \pi_{i,p}) \equiv (\pi_1, \ldots, \pi_p) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_p)$ (i.e., one draw for all i) and $\gamma_i \sim \mathcal{G}(\alpha_{0,i}, \beta_{0,i})$, $i \in \{1, \ldots, k\}$, with these being independent and also independent of $(\pi_1, \ldots, \pi_p)$. The posterior is then

$$(\pi_1, \ldots, \pi_p) \mid \text{data} \sim \mathrm{Dir}(\alpha_1 + Z_1, \ldots, \alpha_p + Z_p),$$
$$\gamma_i \mid \text{data} \sim \mathcal{G}(\alpha_{0,i} + Y_i(\tau), \beta_{0,i} + 1), \quad i \in \{1, \ldots, k\},$$

where $Z_j = \sum_{i=1}^{k} \{Y_i(t_j) - Y_i(t_{j-1})\}$, $j \in \{1, \ldots, p\}$, and the $\gamma_i \mid \text{data}$, $i \in \{1, \ldots, k\}$ are independent. The Bayes estimator that emerges is

$$\delta_{i,j}^B = \frac{(1+c)(p\bar{\alpha} + Z_j - 1)}{(1+c)(p\bar{\alpha} + Z - 1) - c(p-1)} \frac{a_{0,i} + Y_i(\tau) - 1}{\beta_{0,i} + 1}, \tag{35}$$

for $j \in \{1, \ldots, p\}$ and $i \in \{1, \ldots, k\}$, where $\bar{\alpha} = (1/p) \sum_{j=1}^{p} \alpha_j$ and Z is the total sum $\sum_{j=1}^{p} Z_j$. This and the accompanying natural frequentist estimator

$$\widehat{\theta}_{i,j} = \frac{(1+c)(p-1+Z_j)}{(1+c)(p+Z-1) - c(p-1)} Y_{i,j} = \frac{(1+c)(p-1+Z_j)}{p - 1 + (1+c)Z} Y_{i,j}, \tag{36}$$

are interesting because they borrow information in all directions, so to speak; cross-sectionally through Z_j; in time through $Y_i(\tau)$; and both horizontally and vertically via the total sum Z. From Section 3 we know that if the two parts of L_c^* are viewed separately, the estimators

$$\delta_{i,j}' = \left(1 - \frac{kp-1}{kp-1+Z}\right) Y_{i,j}, \quad \delta_i' = \left(1 - \frac{k-1}{k-1+Z}\right) Y_i(\tau),$$

uniformly dominate $Y_{i,j}$ and $Y_i(\tau)$, under the loss functions $\sum_{i=1}^{k} \sum_{j=1}^{p} (\delta_{i,j} - \theta_{i,j})^2 / \theta_{i,j}$ and $\sum_{i=1}^{k} (\delta_i - \gamma_i)^2 / \gamma_i$, respectively. The estimators (35)–(36) provide guidance on how to exploit the multivariate nature of the problem in order to compromise between $\delta_{i,j}'$ and δ_i', and thereby achieve risk dominance in large parts of the parameter space, or even uniformly.

9. Concluding remarks

We round off our paper by offering a list of concluding remarks, some pointing to further research.

A. Balanced loss functions. As pointed out by a referee the loss functions (3) and (4) are somewhat similar to the balanced loss functions studied by Jafari Jozani et al. [16] and by Marchand and Strawderman [20]. In the univariate case these are of the type $L_{\omega, \delta_{\text{target}}}(\theta, \delta) = \omega q_1(\theta)(\theta - \delta_{\text{target}})^2 + (1 - \omega) q_2(\theta)(\delta - \theta)^2$, where $0 \le \omega \le 1$, the q_1, q_2 are positive weight functions, and $\delta_{\text{target}} = \delta_{\text{target}}(Y)$ is a target estimator, i.e., a data-dependent point the statistician seeks to draw the estimates towards. Extensions to multivariate estimation are found in [20]. Note that the L_c loss function in (4) has a natural counterpart in the balanced loss function $L_{\omega, \bar{y}}(\theta, \delta) = (1 - \omega) L_1^*(\theta, \delta) + \omega p(\bar{\delta} - \bar{y})^2 / \bar{\theta}$. Furthermore, observe that the corresponding risk function is $\mathrm{E} L_{\omega, z}(\theta, \delta) = (1 - \omega) \mathrm{E} L_1^*(\theta, \delta) + \omega p \mathrm{E} (\bar{\delta} - \bar{\theta})^2 / \bar{\theta} + \omega$, which entails that the results we derive for L_c in this paper can be recast as results for $L_{\omega, z}$. A balanced loss function type idea, merging the ideas underlying (3)–(4) with those of the balanced loss functions of Marchand and Strawderman [20], are loss functions compromising between two different transformations of the parameter(s). As an example, consider the generalisation of the L_c loss function given by $L_{\omega, h}(\theta, \delta) = (1 - \omega) L_1^*(\theta, \delta) + \omega q(\theta) \{h(\delta) - h(\theta)\}^2$, where $h(\cdot)$ is a given transformation of the parameters and $q(\theta)$ is a weight function, i.e., $q(\theta) = (\sum_{i=1}^{p} \theta_i)^{-1}$ and $h(\theta) = \sum_{i=1}^{p} \theta_i$ lead back to the L_c loss of (4).

B. Normal approximations and the square-root transformation. When the θ_i are likely to not being small, normal approximations might work well, and multiparameter estimation may proceed via, e.g., the approximate model $2y_i^{1/2} \sim \mathcal{N}(2\theta_i^{1/2}, 1)$. The point is that there is a voluminous literature on shrinkage methods for normal setups, with Stein–James estimators etc. This also invites loss functions of the type $\sum_{i=1}^{p}(\delta_i^{1/2} - \theta_i^{1/2})^2$, which may also be motivated via the Hellinger distance between the real and estimated Poisson distributions. Methods of our paper, using exact Poisson calculations as opposed to normal approximations, may indeed be used to demonstrate that estimators of the type $\widehat{\theta}_i^{1/2} = \{1 - g(z)\}y_i^{1/2}$ may be found, which dominate the default method's $y_i^{1/2}$ in large parts of the parameter space. Specifically, if $p \geq 2$, and given a lower positive threshold π_0, methods exist which dominate the default method in the region characterised by having all $\pi_i = \theta_i/\gamma \geq \pi_0$.

C. Bayesian nonparametrics. Suppose a time-inhomogeneous Poisson process $Y = \{Y(t): t \geq 0\}$ is observed over some time window $[0, \tau]$, with cumulative intensity $G = \{G(t): t \geq 0\}$. We wish to estimate this function, with the loss function

$$L(G, \widehat{G}) = \int_0^{\tau} \{\widehat{G}(t) - G(t)\}^2 / G(t)\, w(\mathrm{d}t),$$

with some fixed weight measure w. Note however that this is not an obvious extension of our previous Clevenson–Zidek loss function, since it works via the cumulatives. A natural estimator is Y itself, with constant risk function $r_0 = \int_0^{\tau} w(\mathrm{d}t)$. A natural class of priors takes $G \sim \mathcal{G}(aG_0, a)$, say, a Gamma process with independent increments and $G(t) \sim \mathcal{G}(aG_0(t), a)$. Then $G(t) \mid \text{data} \sim \mathcal{G}(aG_0(t) + Y(t), a + 1)$, and the Bayes estimator becomes

$$\widehat{G}(t) = \frac{1}{\mathrm{E}\{1/G(t) \mid \text{data}\}} = \frac{aG_0(t) + Y(t) - 1}{a + 1}.$$

Working with expressions for the minimum Bayes risk one may show that this converges as $a \to 0$ to the value r_0, proving that the estimator Y is minimax.

A larger class of priors can however also be investigated, inspired by methods and results of our Section 4. Write $G = \gamma F$, with $\gamma = G(\tau)$ the full mass and $F = G/\gamma$ normalised to be a cumulative distribution function. Now construct a prior by having γ from some density $q(\gamma)$ independent of a Dirichlet process for F, say $F \sim \mathrm{Dir}(bF_0)$, i.e., prior mean F_0 and b the mass or precision parameter. An appropriate extension of our Lemma 1 then shows that (i) γ and F are independent, given data; (ii) $\gamma \mid \text{data}$ has a density proportional to $q(\gamma)\gamma^z \exp(-\gamma)$, with $z = Y(\tau)$; and (iii) $F \mid \text{data}$ is a Dirichlet $bF_0 + Y$. The Bayes estimator becomes

$$G^*(t) = \{\mathrm{E}(\gamma^{-1} \mid \text{data})\}^{-1}[\mathrm{E}\{F(t)^{-1} \mid \text{data}\}]^{-1} = \frac{K(z)}{K(z-1)} \frac{bF_0(t) + Y(t) - 1}{b + Y(\tau) - 1},$$

with $K(z) = \int_0^{\infty} \gamma^z e^{-\gamma} q(\gamma)\, \mathrm{d}\gamma$ as in Section 4. Note that this construction provides a genuine extension of the Gamma process, and still with easy control of the posterior distribution for $G = \gamma F$. The usual Gamma process, with $G(t) \sim \mathcal{G}(aG_0(t), a)$, corresponds to the special case of $G = \gamma F$, where γ has a $\mathcal{G}(aG_0(\tau), a)$ and is independent of $F \sim \mathrm{Dir}(aG_0)$, i.e., with prior mean $F_0 = G_0/G_0(\tau)$ and mass parameter $aG_0(\tau)$.

D. Separating sum and proportions. When estimating Poisson parameters $\theta_1, \ldots, \theta_p$, another type of loss function than those worked with above is to separate the sum γ and proportions $\pi_i = \theta_i/\gamma$, and then work with $(\widehat{\gamma} - \gamma)^2/\gamma + z\sum_{i=1}^{p}(\widehat{\pi}_i - \pi_i)^2/\pi_i$. The maximum likelihood procedure corresponds to $\widehat{\pi}_i = Y_i/Z$ and $\widehat{\gamma} = Z$, having constant risk function $1 + (p - 1) = p$. Procedures achieving lower risk in large parameter regions may be constructed via empirical Bayes arguments.

E. Shrinking towards submodels. There is scope for extension of our methods and constructions in several directions for multiparameter Poisson- and Poisson-related inference. It is inherently useful to shrink raw estimates towards meaningful submodels, such as with log-linear setups for analysis of tables of count-data, see e.g. Agresti [1]. There are notable Poisson-related models for small-area estimation, involving also mixed Poisson models, zero-inflated versions, etc. Generally speaking Bayes and empirical Bayes constructions can be brought to such tables, and will tend to work well for sizeable parameter regions, whereas the exact risk calculations worked with in the present article are harder to generalise.

Acknowledgements

References

[1] A. Agresti, An Introduction to Categorical Data Analysis, third ed., Wiley, New York, 2019.
[2] J.H. Albert, Simultaneous estimation of Poisson means, J. Multivariate Anal. 11 (1981) 400–417.

[3] J.O. Berger, Discussion: Construction of improved estimators in multiparameter estimation for discrete exponential families, Ann. Statist. 11 (1983) 368–369.

[4] L.D. Brown, R.H. Farrell, Proof of a necessary and sufficient condition for admissibility in discrete multivariate problems, J. Multivariate Anal. 24 (1988) 46–52.

[5] L.D. Brown, E. Greenshtein, Y. Ritov, The Poisson compound decision problem revisited, J. Amer. Statist. Assoc. 108 (2013) 741–749.

[6] M.L. Clevenson, J.V. Zidek, Simultaneous estimation of the means of independent Poisson laws, J. Amer. Statist. Assoc. 70 (1975) 698–705.

[7] B. Efron, Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Cambridge University Press, 2013.

[8] B. Efron, C. Morris, Limiting the risk of Bayes and empirical Bayes estimators – Part I: The Bayes case, J. Amer. Statist. Assoc. 66 (1971) 807–815.

[9] B. Efron, C. Morris, Limiting the risk of Bayes and empirical Bayes estimators – Part II: The empirical Bayes case, J. Amer. Statist. Assoc. 67 (1972) 130–139.

[10] D. Fourdrinier, W.E. Strawderman, M.T. Wells, Shrinkage Estimation, Springer, 2018.

[11] M. Ghosh, J.T. Hwang, K.-W. Tsui, Construction of improved estimators in multiparameter estimation for discrete exponential families, Ann. Statist. 11 (1983) 351–367.

[12] M. Ghosh, A. Parsian, Bayes minimax estimation of multiple Poisson parameters, J. Multivariate Anal. 11 (1981) 280–288.

[13] M. Ghosh, M. Yang, Simultaneous estimation of Poisson means under entropy loss, Ann. Statist. 16 (1988) 278–291.

[14] H.M. Hudson, Discussion: Construction of improved estimators in multiparameter estimation for discrete exponential families, Ann. Statist. 11 (1983) 370–371.

[15] J.T. Hwang, Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases, Ann. Statist. 10 (1982) 857–867.

[16] M. Jafari Jozani, É. Marchand, A. Parsian, On estimation with weighted balanced-type loss function, Statist. Probab. Lett. 76 (2006) 773–780.

[17] I. Johnstone, Admissible estimation, Dirichlet principles and recurrence of birth-death chains on $\mathbb{Z}_+^p$, Probab. Theory Related Fields 71 (1986) 231–269.

[18] M. Jones, É. Marchand, Multivariate discrete distributions via sums and shares, J. Multivariate Anal. 171 (2019) 83–93.

[19] E.L. Lehmann, Theory of Point Estimation, John Wiley & Sons, New York, 1983.

[20] É. Marchand, W.E. Strawderman, On shrinkage estimation for balanced loss functions, arXiv preprint arXiv:1904.03171, (2019).

[21] C.N. Morris, Discussion: Construction of improved estimators in multiparameter estimation for discrete exponential families, Ann. Statist. 11 (1983) 372–374.

[22] J.C.-M. Peng, Simultaneous estimation of the parameters of independent Poisson distributions, Technical Report No. 48, Stanford University, Departement of Statistics, 1975.

[23] C.P. Robert, The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation, second ed., Springer, Berlin, 2007.

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

[25] K.-W. Tsui, S.J. Press, Simultaneous estimation of several Poisson means under k-normalized squared error loss, Ann. Statist. 10 (1982) 93–100.