BAYESIAN STATISTICS 6, pp. 000–000
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)
Oxford University Press, 1998

Bayesian and Frequentist Approaches to Parametric Predictive Inference

=

RICHARD L. SMITH
University of North Carolina, USA

SUMMARY

Suppose we want to estimate predictive probabilities for an unobserved random variable, whose distribution depends on an unknown finite-dimensional parameter which is estimated from past data. A crude "plug-in" approach is to substitute a point estimate of the unknown parameter into the predictive distribution. This may be criticized as failing to take into account the uncertainty of the unknown parameter, whereas a Bayesian approach incorporates such uncertainty in a natural way. Therefore, we might expect that a Bayesian procedure will be superior to a plug-in approach even when assessed in purely frequentist terms. The analysis given in this paper, however, shows that this is too simple a conclusion. For many models, when assessed from the point of view of mean squared error of predictive probabilities, the plug-in approach is superior to the Bayesian approach in the extreme tail of the distribution. We may also consider alternative loss functions besides mean squared error, but in most cases a similar phenomenon occurs. However, one may also generalize the class of Bayesian procedures by considering alternatives to the posterior mean of the predictive distribution. With such generalizations, the comparison swings back in favor of Bayesian procedures. These results are initially developed for simple models, in particular the exponential distribution, but in later discussion they are used to study the properties of hierarchical modeling approaches to empirical Bayes problems. The mathematical development is only sketched in the present paper, but relies heavily on asymptotic expansions and second-order properties.

Keywords: DECISION THEORY; MEAN SQUARED ERROR; COVERAGE PROBABILITY OF PREDICTION INTERVALS; SECOND-ORDER ASYMPTOTICS; EXPONENTIAL DISTRIBUTION; EXTREME VALUES; EMPIRICAL BAYES; HIERARCHICAL MODELS.

1. INTRODUCTION

This paper is motivated by the belief that *predictive* statements about as yet unobserved random variables are often more meaningful for applications than conventional point and interval estimates of parameters. Bayesian predictive distributions are easily calculated using modern Monte Carlo techniques, but corresponding frequentist procedures are often lacking. Even from a purely Bayesian perspective, it can be far from clear how we should specify a suitable prior distribution, and as we shall see during the course of the paper, the loss function can also play a critical role. Therefore, I believe both Bayesians and frequentists should be concerned about the sampling properties of Bayesian predictive procedures.

1.1 *A motivating example*

The following example, based on an ongoing project at the National Institute of Statistical Sciences in North Carolina, will serve to illustrate the practical usefulness of Bayesian predictive procedures.

A well-known oil company asked us to assess the probability that it would suffer an insurance claim of one billion US dollars, over a time span of one year, as the result of an oil spill. The data provided included an inventory of past oil spills within the industry, detailing the location of each spill, the ship size, and amount of oil spilled, and the cost of the eventual settlement. The company also provided projections of its own future shipping movements. By constructing separate parametric models for the probability of a spill, the size of a spill, and the cost of a spill, we were able to characterize the distribution of the company's future costs by a function $p(z;\theta)$, which represents the probability that the company will suffer a claim of greater than z in a one-year period, in terms of various model parameters θ, which are estimated based on the past data X. One approach to the problem is to estimate θ by its maximum likelihood estimate (MLE) based on the past data, $\hat{\theta}$ say, and then to employ a MLE or "plug-in" approach to the predictive estimate:

$$\hat{p}_{\mathrm{MLE}}(z) = p(z;\hat{\theta}). \tag{1}$$

An alternative approach is Bayesian: assuming a prior density $\pi(\theta)$, calculate the posterior $\pi(\theta \mid X)$, and then determine the Bayes predictive distribution

$$\hat{p}_{\mathrm{BAY}}(z) = \int p(z;\theta)\pi(\theta \mid X)d\theta. \tag{2}$$

One apparent advantage of (2) over (1) is that (2) allows explicitly for the uncertainly in θ, whereas (1) makes no such allowance.

The results for one version of the analysis are shown in Figure 1, where the solid curve represents $\hat{p}_{\mathrm{BAY}}(z)$ computed from (2), and the two sets of dashed curves represent 50% and 95% Bayes credible intervals for $p(z;\theta)$, computed separately for each z, measured in millions of US dollars. The results show, for example, that for $z = 1{,}000$, the posterior mean $\hat{p}_{\mathrm{BAY}}(z)$ is about .012, but the 95% credible interval ranges from .004 to .028. In this way we are able to make an explicit assessment of the uncertainty in the whole procedure. However, it is far from clear what we can say about the frequentist properties of estimates of this general structure.

1.2 *General formulation*

In general we consider the case where $X = (X_1, ..., X_n)$ consists of n observations from a process whose distribution is indexed by θ, and Z is some future observation, not necessarily from the same distribution, but whose distribution is determined by the same θ. We assume X and Z are independent except for their common dependence on θ; this formulation excludes many problems in time series analysis or spatial statistics in which X and Z are dependent, which are intended for future investigation. The random variables $X_1, ..., X_n$ need not be independent observations from a common distribution, though all our examples will be of that form. We let $p(z;\theta)$ denote either the distribution function or the survivor function of Z. All asymptotic results are as $n \to \infty$.

Within this general framework, we consider theoretical properties of the predictive estimates (1) and (2), along with different formulations of the Bayes procedure.

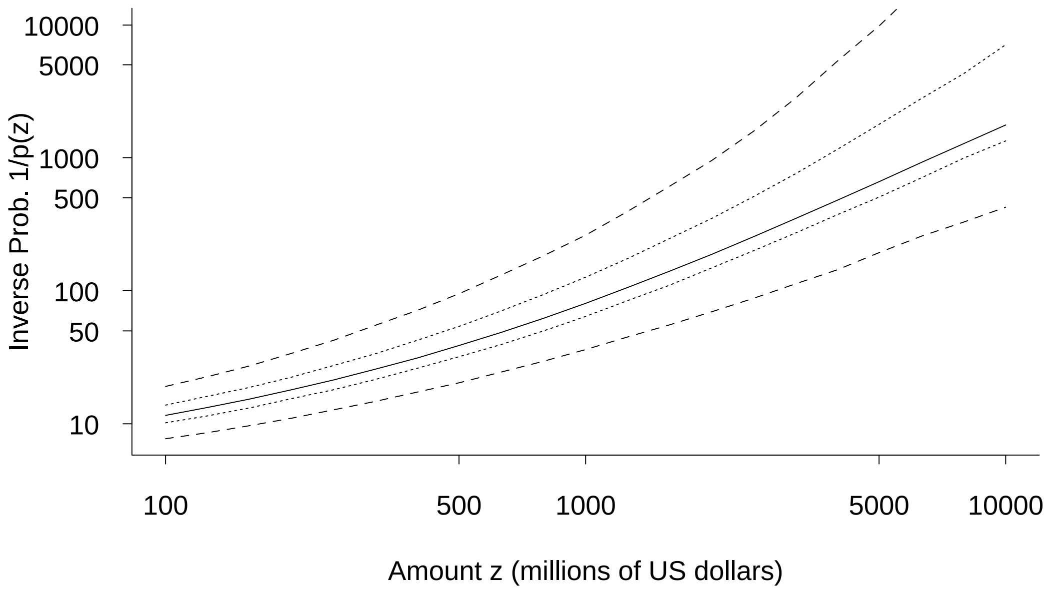

Figure 1. *Predictive distributions for the oil spill example. The solid curve gives the Bayes predictive distribution of* $p(z;\theta)$*, and the short-dashed and long-dashed curves define respectively the 50% and 95% Bayes credible intervals, computed separately for each* z*. with 50% (short-dashed curves) and 95% (long-dashed curves) credible intervals. Note that it is actually* $1/\hat{p}(z)$ *rather than* $\hat{p}(z)$ *which is plotted.*

The problem of predictive inference has been studied for many years. Aitchison and Dunsmore (1975) considered predictive inference from both frequentist and Bayesian points of view, but with a heavy emphasis on the Bayesian approach. More generally, S. Geisser developed a theory of predictive inference over a long series of papers, summarized by the book Geisser (1993), which also focusses on Bayesian solutions.

Even from a purely frequentist viewpoint, there are many reasons for believing that the predictive density given by (2) might have better properties than (1). Cox (1975) evaluated predictive distributions from the point of view of how closely the resulting prediction intervals achieved their nominal coverage probabilities. He argued that prediction intervals derived from (1) typically have too low a coverage probability because they fail to take into account the sampling variability of $\hat{\theta}$, and he derived an asymptotic $O(\frac{1}{n})$ correction formula. He also proposed various exact and asymptotic improvements on (1), which under some circumstances coincide with a Bayesian solution. An example will be seen in section 2. The recent paper by Barndorff-Nielsen and Cox (1996) greatly extended the range of asymptotic formulae for prediction intervals, though without explicitly considering the connection with Bayesian formulations. From a different point of view, Aitchison (1975) showed that under certain circumstances (2) is better than (1) from the point of view of minimizing the Kullback-Leibler distance between the true and estimated predictive distributions. One case where he was able to show this was the exponential distribution, considered in section 2. Here again there have been recent asymptotic developments, as Komaki (1996) gave asymptotic formulae for Kullback-Leibler distance in predictive distributions based on differential geometric ideas.

Other frequentist approaches include Butler's (1986) development of "predictive likelihood", a Bayes-motivated approximation to this due to Davison (1986), and papers of Harris (1989) and Vidoni (1995). Some of this literature would appear to point to strong interconnections between the Bayesian and frequentist approaches, parallel to similar interconnections in more traditional parametric inference (Reid 1996, Sweeting 1996).

Thus one might well conclude that the plug-in approach is thoroughly discredited, and that the Bayesian solution (2), if not the preferred choice of all authors, is nevertheless a general-purpose alternative which performs well in cases where its properties have been investigated. In the present paper, we shall show that matters are by no means as simple as that. Under some circumstances and from certain points of view, (1) is a better estimator than (2). However in all cases where this happens, some alternative Bayesian formulation performs better than either. These alternative formulations involve computing the Bayesian decision rule under alternative loss functions. The conclusion is not that one should abandon Bayesian methods in favor of much more simple-minded approaches such as (1), but that one needs to give much more careful attention to the formulation of the Bayesian approach and in particular to the choice of loss function.

2. THE ONE-PARAMETER EXPONENTIAL DISTRIBUTION

The one-parameter exponential distribution will serve as an example which illustrates most of the ideas we want to discuss, while being simple enough to permit explicit calculations of many quantities of interest.

Let $X = (X_1, ..., X_n)$ denote an i.i.d. sample from the density $\theta e^{-\theta x}$ $(x > 0,\ \theta > 0)$ and let Z denote some future observation from the same density. We focus on the predictive probability that $Z > z$ for some given level z. When θ is known, this is given by $\phi = \phi(z;\theta) = e^{-\theta z}$; we also write this as $\phi(\theta)$ when there is no ambiguity about the value of z.

Although simplified, this problem may in itself have applications in areas such as flood prediction or insurance. For example, exceedances over a high threshold are often modeled via an exponential distribution, which is a special case of the generalized Pareto distribution (Davison and Smith 1990, Smith and Shively 1995). Insurance claims are often assumed to have a distribution with a Pareto tail, whose index may be estimated by the so-called Hill estimator (Resnick 1997, Embrechts *et al.* 1997). However, a Pareto tail is equivalent to an exponential tail for log claims, and Hill's estimator is equivalent to the maximum likelihood estimator for θ based on all exceedances over a given threshold (Smith 1987). In this context, the predictive question of interest is whether we will observe a claim over some given very large value, which may for example represent the limit of liability under a reinsurance contract.

The MLE of θ is $\hat{\theta}_n = \frac{n}{S_n}$ $(S_n = X_1 + ... + X_n)$ so the MLE of ϕ is

$$\hat{\phi}_{\text{MLE}} = \exp\left(-\frac{nz}{S_n}\right). \tag{3}$$

A Bayesian solution to this problem is to assume $\pi(\theta) \propto \theta^{a-1}e^{-b\theta}$, the Gamma prior with parameters (a, b). Then the posterior density of θ is also Gamma with parameters $(a+n, b+S_n)$, and the posterior expectation of ϕ is easily calculated to be

$$\hat{\phi}_{\text{BAY},0} = \left(\frac{b+S_n}{b+S_n+z}\right)^{a+n}. \tag{4}$$

The reason for the subscript 0 will appear shortly. When $a = b = 0$ (the Jeffreys prior), (4) reduces to

$$\hat{\phi}_{\mathrm{BAY},0} = \left(\frac{S_n}{S_n + z}\right)^n. \tag{5}$$

It turns out that (5) also has an exact frequentist interpretation. Suppose we are trying to find a level $1-\alpha$ prediction interval for Z of the form $(0, z_\alpha)$ for some data-dependent statistic z_α. Treating (5) as a predictive tail function, we may solve $\hat{\phi}_{\mathrm{BAY},0} = \alpha$ to obtain the Bayesian solution

$$z_\alpha = S_n\left(\alpha^{-1/n} - 1\right). \tag{6}$$

However a procedure suggested by Cox (1975) (see also Cox and Hinkley, 1974, pp. 242–245) suggests that we try to find some function of S_n and Z whose distribution is independent of θ. In this case it turns out that $W = S_n/(S_n + Z)$ is such a function, with distribution function $P\{W \le w\} = w^n$ for $0 \le w \le 1$. Therefore we set $w = \alpha^{1/n}$ to obtain

$$1 - \alpha = P\left\{W \ge \alpha^{1/n}\right\} = P\left\{Z \le S_n\left(\alpha^{1/n} - 1\right)\right\}, \tag{7}$$

thus confirming that (6) is an exact solution of the (frequentist) prediction interval problem, and not merely an asymptotic solution as might have been supposed. This seems to vindicate the use of (5) as a predictive distribution for Z.

Nevertheless, from other points of view, (5) may be inferior to (3). Suppose we compare the two estimators directly, using mean square error (MSE). In the case of (3) we have

$$E\{(\hat{\phi}_{\mathrm{MLE}} - \phi)^2\} = \frac{\theta^2 z^2 e^{-2\theta z}}{n} + \frac{\theta^2 z^2 e^{-2\theta z}}{n^2}\left(5 - 7\theta z + \frac{7}{4}\theta^2 z^2\right) + O\left(\frac{1}{n^3}\right). \tag{8}$$

The corresponding result for (5) is

$$E\{(\hat{\phi}_{\mathrm{BAY},0} - \phi)^2\} = \frac{\theta^2 z^2 e^{-2\theta z}}{n} + \frac{\theta^2 z^2 e^{-2\theta z}}{n^2}\left(5 - 10\theta z + \frac{7}{2}\theta^2 z^2\right) + O\left(\frac{1}{n^3}\right). \tag{9}$$

The asymptotic comparison of the two procedures depends on the second-order term, and shows that $\hat{\phi}_{\mathrm{BAY},0}$ is better than $\hat{\phi}_{MLE}$ for large n if and only if $5 - 10\theta z + \frac{7}{2}\theta^2 z^2 < 5 - 7\theta z + \frac{7}{4}\theta^2 z^2$, or in other words if $\theta z < \frac{12}{7}$, or $\phi > .1801$. For smaller values of the true ϕ, MLE is better than Bayes. Moreover, simulation results show that this is a realistic comparison even for small values of n.

One can make the same comparison using different values of a and b in (4), or more generally using non-conjugate priors, but this does not change the qualitative conclusion that for z sufficiently large or ϕ sufficiently small, the MLE performs better than the Bayes estimator when assessed from the point of view of mean squared error.

There are, however, alternative formulations of the problem which are not so unfavorable to the Bayesian solution. Instead of MSE, consider the loss function

$$L(\phi, \hat{\phi}) = \frac{(\phi - \hat{\phi})^2}{\phi^r} \tag{10}$$

for some non-zero r. In particular, $r = 2$ corresponds to relative rather than absolute mean squared error. For the frequentist decision problem, this modification to the problem makes no difference to the solution, because the denominator of (10) depends only on ϕ, not $\hat{\phi}$, so if one estimator $\hat{\phi}_1$ beats another estimator $\hat{\phi}_2$ from the point of view of MSE, the same will be true under (10) whatever the value of r. For the Bayesian solution, however, the influence of r is considerable, because the Bayes decision rule based on (10) is the estimator

$$\hat{\phi}_{\mathrm{BAY},r} = \frac{\int \phi(\theta)^{1-r}\pi(\theta \mid X)d\theta}{\int \phi(\theta)^{-r}\pi(\theta \mid X)d\theta} \tag{11}$$

which in the case of our exponential prediction problem reduces to

$$\hat{\phi}_{\mathrm{BAY},r} = \left(\frac{b+S_n-rz}{b+S_n-(r-1)z}\right)^{a+n}. \tag{12}$$

If we fix $a = b = 0$ in this case, we have

$$\begin{aligned} E\{(\hat{\phi}_{\mathrm{BAY},r}-\phi)^2\} &= \frac{\theta^2 z^2 e^{-2\theta z}}{n} \\ &+ \frac{\theta^2 z^2 e^{-2\theta z}}{n^2}\left\{5-(10-6r)\theta z+\left(\frac{7}{2}-4r+r^2\right)\theta^2 z^2\right\}+O\left(\frac{1}{n^3}\right). \end{aligned} \tag{13}$$

For large z the dominant coefficient here is $\frac{7}{2} - 4r + r^2$ and this is minimized by $r = 2$. Thus there is a sense in which $r = 2$ is optimal among all Bayes estimators, for predicting tail probabilities of the exponential distribution, and in this case the Bayes estimator also dominates the MLE.

However, the comparison does not end there, because there are other loss functions besides the family represented by (10). When ϕ is small, one can criticize MSE as a criterion for comparing estimators, because an estimator which overestimates ϕ is penalized much more heavily than one which underestimates ϕ in the same proportion. This suggests we consider alternative loss functions for assessing the performance of the estimator. One possibility is the *logarithmic loss function*

$$L(\phi,\hat{\phi}) = \phi\log\frac{\phi}{\hat{\phi}} + (1-\phi)\log\frac{1-\phi}{1-\hat{\phi}} \tag{14}$$

which is sometimes used as a scoring rule in assessing probability forecasters (Seillier-Moiseiwitsch and Dawid 1993). An alternative which penalizes underestimates of ϕ even more severely is the *squared logarithmic loss function*

$$L(\phi,\hat{\phi}) = \left(\log\phi - \log\hat{\phi}\right)^2. \tag{15}$$

For our estimators, we continue to use either $\hat{\phi}_{\mathrm{MLE}}$ or $\hat{\phi}_{\mathrm{BAY},r}$, but we now use (14) or (15) as a loss function by which to evaluate them.

For both loss functions, expansions for $E_\theta\{L(\phi,\hat{\phi})\}$ similar to (8) and (13) are available, and comparisons among estimators may be made on the basis of the $O\left(\frac{1}{n^2}\right)$ terms. For (14), the MLE is again superior to the usual ($r = 0$) Bayes estimator for sufficiently large z, but the comparison is reversed for $r = 1$, which in this case is the optimal value of r as $z \to \infty$. Similarly, under (15) the optimal r is $r = \frac{1}{2}$.

In summary, the conclusions of this section are:

1. The usual Bayes estimator ($r = 0$) achieves exact coverage probability of prediction intervals but is inferior to the MLE when assessed from the point of view of MSE for small tail probabilities.

2. Other Bayes estimators, constructed from (11) for different values of r, have smaller MSE than the MLE when ϕ is small, with optimal value $r = 2$.

3. For different loss functions such as (14) and (15), again concentrating on the extreme right tail of the predictive distribution, different optimal values of r are derived: $r = 1$ in the case of (14), $r = \frac{1}{2}$ in the case of (15). By all of these criteria, the usual Bayes estimator with $r = 0$ is inferior to the MLE when used to estimate small tail probabilities.

4. All the above results are derived for the Jeffreys prior ($a = b = 0$), but the same asymptotic expressions as $z \to \infty$ are valid for any conjugate prior or indeed for any smooth non-conjugate prior.

3. ASYMPTOTIC FORMULAE FOR GENERAL PARAMETRIC FAMILIES

Consider now a general parametric family and let $\ell_n(\theta)$ denote the log likelihood of a finite-dimensional parameter θ based on n observations. The situation is not necessarily one of i.i.d. data though in all cases we assume $n \to \infty$ in such a way that the quantities denoted by κ's below are of $O(1)$. Throughout we deal in formal asymptotic expansions and make no attempt to specify precise regularity conditions. Derivations of the following results are in Smith (1997) while the general framework for the asymptotic results follows McCullagh (1987).

Suppose we are interested in a scalar function $\phi(\theta)$ — later on this will denote the tail of a predictive distribution function, but for now ϕ may be any one-dimensional function of θ. The MLE of θ is denoted $\hat{\theta}$ and the MLE of ϕ is of course

$$\hat{\phi}_{\mathrm{MLE}} = \phi(\hat{\theta}). \tag{16}$$

The Bayes estimator of ϕ is typically given by

$$\hat{\phi}_{\mathrm{BAY}} = \frac{\int \phi(\theta) e^{\ell_n(\theta)+Q(\theta)} d\theta}{\int e^{\ell_n(\theta)+Q(\theta)} d\theta} \tag{17}$$

where $Q(\theta)$ depends on both the prior density and the loss function. In section 2, $Q(\theta) = (a-1)\log\theta + (rz - b)\theta$.

Let $U_i = \frac{\partial \ell_n(\theta)}{\partial \theta_i}$, $U_{ij} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta_i \partial \theta_j}$ etc., and let $\{U^{ij}\}$ denote the matrix inverse of $\{U_{ij}\}$. These quantities are assumed to be evaluated at the true θ unless they have hats on, in which case they are evaluated at $\hat{\theta}$. This in particular, $\{\hat{U}_{ij}\}$ is the observed information matrix and $\{\hat{U}^{ij}\}$ is its inverse. Let ϕ_i, ϕ_{ij}, Q_k, etc., denote partial derivatives of ϕ or Q with respect to the indicated components of θ. We adopt the summation convention whereby a repeated index implies summation over that index. With this convention define

$$\mathcal{D} = \frac{1}{2} U_{ijk} U^{ik} U^{j\ell} \phi_\ell - \frac{1}{2}(\phi_{ij} + 2\phi_i Q_j) U^{ij} \tag{18}$$

and let $\hat{\mathcal{D}}$ denote the corresponding expression evaluated at $\theta = \hat{\theta}$, as just described. Our first main result is the asymptotic formula

$$\hat{\phi}_{\mathrm{BAY}} \approx \hat{\phi}_{\mathrm{MLE}} + \hat{\mathcal{D}} \tag{19}$$

with an error of $O_p\left(\frac{1}{n^2}\right)$. Formula (19) is equivalent to the same order of magnitude as the formula of Tierney and Kadane (1986), though it is not the same as Tierney-Kadane. Loosely speaking, Tierney-Kadane is in most cases the easier formula to evaluate numerically, not requiring any explicit derivation of an asymptotic expansion, but (19) is more convenient for the kinds of theoretical manipulations we have in mind here.

Let $\kappa_{i_1 i_2 \ldots i_k} = \frac{1}{n} E\left\{U_{i_1 i_2 \ldots i_k}\right\}$, $\kappa_{i_1 i_2 \ldots i_k, j_1 j_2 \ldots j_\ell} = \frac{1}{n} \mathrm{Cov}\left\{U_{i_1 i_2 \ldots i_k}, U_{j_1 j_2 \ldots j_\ell}\right\}$; as noted previously, these are assumed to be of $O(1)$ as $n \to \infty$. Also let $\{\kappa^{i,j}\}$ denote the matrix inverse of $\{\kappa_{i,j}\}$ (the Fisher information matrix). Define quantities $\mathcal{C}$, $\mathcal{C}_s$ and $\mathcal{A}$ by

$$
\begin{aligned}
\mathcal{C} &= \kappa_{ijk}\kappa^{i,k}\kappa^{j,\ell}\phi_\ell + \kappa^{i,j}(\phi_{ij} + 2\phi_i Q_j), \\
\mathcal{C}_s &= \frac{\partial \mathcal{C}}{\partial \theta_s}, \\
\mathcal{A} &= \frac{\mathcal{C}^2}{4} + \mathcal{C}_s \kappa^{s,t}\phi_t \\
&\quad + \mathcal{C}\left\{\kappa^{i,j}\kappa^{k,\ell}\left(\kappa_{\ell,jk} + \frac{1}{2}\kappa_{jk\ell}\right)\phi_i + \frac{1}{2}\kappa^{i,j}\phi_{ij}\right\} \\
&\quad + \kappa^{i,k}\kappa^{j,\ell}\kappa^{s,t}\phi_i\left\{\kappa_{k,\ell t}(\phi_{js} + 2\phi_s Q_j)\right. \\
&\qquad \left. + (\kappa_{k,\ell st} + \kappa^{u,v}\kappa_{k,su}\kappa_{\ell tv} + \kappa^{u,v}\kappa_{k,\ell s}\kappa_{utv})\phi_j\right\}.
\end{aligned}
\tag{20}
$$

Our second main result is the formula

$$
E\{(\hat{\phi}_{\mathrm{BAY}} - \phi)^2\} - E\{(\hat{\phi}_{\mathrm{MLE}} - \phi)^2\} = \frac{\mathcal{A}}{n^2} + o\left(\frac{1}{n^2}\right). \tag{21}
$$

Recall that in section 2, we showed that in the case of the exponential distribution, the difference in MSEs for $\hat{\phi}_{\mathrm{MLE}}$ and $\hat{\phi}_{\mathrm{BAY}}$ was of $O\left(\frac{1}{n^2}\right)$; (21) is therefore the corresponding general result for comparison of the two estimators.

Suppose we have some other loss function $L(\phi, \hat{\phi})$ such as (14) or (15). Assume that, for each ϕ, $L(\phi, \hat{\phi})$ is minimized with respect to $\hat{\phi}$ when $\hat{\phi} = \phi$, and assume $L_j = \partial^j L(\phi, \hat{\phi})/\partial \hat{\phi}^j \mid_{\hat{\phi}=\phi}$ exists for $j = 2, 3, 4$. Then

$$
E\{L(\phi, \hat{\phi}_{\mathrm{BAY}})\} - E\{L(\phi, \hat{\phi}_{\mathrm{MLE}})\} = \frac{\mathcal{A}_L}{n^2} + o\left(\frac{1}{n^2}\right) \tag{22}
$$

where

$$
\mathcal{A}_L = \frac{1}{2}L_2\mathcal{A} + \frac{1}{4}L_3\mathcal{C}(\phi_i\phi_j\kappa^{i,j}). \tag{23}
$$

The final criterion we consider is *coverage probability bias* (CPB) and is equivalent to the main criterion adopted by Cox (1975) and by Barndorff-Nielsen and Cox (1996). We consider now the case when $\phi = \phi(z;\theta) = P\{Z > z;\theta\}$ is the tail probability of a predictand Z at given level z. The one-sided prediction interval problem is to determine a statistic $\tilde{z}_\alpha$ such that $E\{\phi(\tilde{z}_\alpha;\theta)\}$ is as close as possible to some prescribed value α. In the situation of section 2, this problem has an exact solution, but in general there is no exact solution and the simplest way to proceed is through some estimator $\tilde{\phi}(z)$ of $\phi(z;\theta)$. In that case we define $\tilde{z}_\alpha$ by solving the equation $\tilde{\phi}(\tilde{z}_\alpha) = \alpha$. The CPB is then $E\{\phi(\tilde{z}_\alpha;\theta)\} - \alpha$. An adaptation of the main formula in Cox (1975) shows that the leading term of the CPB is

$$
\frac{1}{n}\left\{\frac{\phi_i(z;\theta)\phi_j'(z;\theta)\kappa^{i,j}}{\phi'(z;\theta)} - b(z;\theta)\right\}. \tag{24}
$$

Here primes denote differentiation with respect to z, while $b(z;\theta)$ is the limit as $n \to \infty$ of $nE\{\tilde{\phi}(z) - \phi(z;\theta)\}$. In the case of the MLE this is given by

$$b(z;\theta) = \kappa^{i,j}\kappa^{k,\ell}(\kappa_{\ell,jk} + \frac{1}{2}\kappa_{jk\ell})\phi_i + \frac{1}{2}\kappa^{i,j}\phi_{ij} \tag{25}$$

while the corresponding expression for $\hat{\phi}_{\mathrm{BAY}}$ is

$$b(z;\theta) = \kappa^{i,j}\kappa^{k,\ell}(\kappa_{\ell,jk} + \kappa_{jk\ell})\phi_i + \kappa^{i,j}(\phi_{ij} + \phi_i Q_j) \tag{26}$$

Equation (25) is an alternative version of a formula given by McCullagh (1987), p. 209; (26) is similar, based on (19).

In view of the results of section 2, it is of particular interest to examine how these formulae behave as $z \to \infty$. This of course depends on the form of the function $\phi(z;\theta)$ as $z \to \infty$, but in many cases there are asymptotic expressions of the form

$$\begin{aligned} \phi_i(z;\theta) &\sim f(z;\theta)\psi_i(\theta)\phi(z;\theta), \\ \phi_{ij}(z;\theta) &\sim G\ f^2(z;\theta)\psi_i(\theta)\psi_j(\theta)\phi(z;\theta), \\ \phi_{ijk}(z;\theta) &\sim G_1\ f^3(z;\theta)\psi_i(\theta)\psi_j(\theta)\psi_k(\theta)\phi(z;\theta), \end{aligned} \tag{27}$$

in terms of some function $f(z;\theta)$ which tends to $\pm\infty$ as $z \to \infty$, a function $\psi(\theta)$ which is independent of z, and constants G and G_1. In many cases, including the exponential, we find $G = G_1 = 1$, but this is not universal. We assume $Q(\theta) = Q(z;\theta) = \log \pi(\theta) - r \log \phi(z;\theta)$ where π is the prior density (independent of z) and r has exactly the same interpretation as in section 2. Under these conditions, one can obtain asymptotic expressions as $z \to \infty$ for $\mathcal{A}$, $\mathcal{A}_L$ and the leading term in the CPB, and one can thus determine the value of r which minimizes these expressions. The answers in the case of mean squared error, logarithmic loss (14), squared logarithmic loss (15), and CPB, are respectively $3G-1$, $3G-2$, $3G-\frac{5}{2}$ and 0. All of these agree with the exponential case when $G=1$.

As a final comment, we discuss briefly the application of these results to empirical Bayes problems. In such problems there are typically p subpopulations, with the same structure but possibly different values $\theta_1, ..., \theta_p$ of the model parameter(s) θ. It is convenient to redefine θ as an overall parameter obtained by concatenating $\theta_1, ..., \theta_p$. Modern Bayesian methods typically impose some hierarchical structure of the form

$$\begin{aligned} \theta &\sim \pi(\theta \mid \eta) = \prod_{j=1}^{p} \pi_j(\theta_j \mid \eta), \\ \eta &\sim \pi_0(\eta), \end{aligned} \tag{28}$$

where η is a hyperparameter. After integrating out η, (28) may be combined into a single prior

$$\pi(\theta) = \int \pi(\theta \mid \eta)\pi_0(\eta)d\eta \tag{29}$$

and the formulae of this section applied to π. The asymptotics here assume that the number of populations p remains fixed but the sample sizes are proportional to some index n which tends to ∞. This provides a theoretical basis for comparing different prior specifications in a hierarchical model.

4. EXAMPLES

4.1 *The Poisson-GPD model*

In section 2 we briefly mentioned the use of the exponential distribution for exceedances over a high threshold. A more complicated model is to assume exceedances over a threshold u occur according to a Poisson process of rate λ and, conditionally on the number of exceedances, that the individual *excess values* over the threshold are independent with a generalized Pareto distribution (GPD)

$$P\{X \leq x+u \,|\, X>u\} = F(x;\psi,\xi) = 1-\left(1+\frac{\xi x}{\psi}\right)_+^{-1/\xi}. \tag{30}$$

This model was adopted by Davison and Smith (1990) as a general model for extremes over a high threshold; the model used for oil spill costs in section 1.1 was also based on this, though in a more complicated form because that example also included covariates. Adopting one year as the unit of time, the tail distribution function for the annual maximum is given by

$$\phi(z;\theta) = 1-\exp\left[-\lambda\left\{1+\frac{\xi(z-u)}{\psi}\right\}_+^{-1/\xi}\right], \quad z \geq u$$

(Davison and Smith 1990, equation 9.1). This is a three-parameter model and the evaluation of the formulae in section 3 is highly complicated. For the purpose of asymptotic calculations, we assume that the process has been observed for n years. We set $Q = -\log\lambda - r\log\phi$ corresponding to the improper prior density λ^{-1}, but including the tail weight ϕ^r, with $r = 0,\ 1$ and 2 as well as the MLE case. Thus the family of Bayes estimators is equivalent to (10). Numerical evaluation is based on the typical values $\lambda = 5,\ \psi = 1,\ \xi = 0.5$.

Figure 2 shows normalized values of $\mathcal{A}$ (for squared error loss), $\mathcal{A}_L$ (for logarithmic loss) and the CPB for various values of z corresponding to different true tail probabilities ϕ; the normalization (not described in detail here) has been chosen to achieve approximate stability across different values of ϕ, but for any given ϕ, the comparison among the four estimators is accurately shown by the plot. Note that the asymptotic CPB of the MLE is always positive, confirming that in this case the plug-in approach leads to prediction intervals with too low a coverage probability.

Qualitatively, Figure 2 exhibits many of the results found for the exponential distribution in section 2, though the asymptotic results as $z \to \infty$ tend to require much larger z or smaller ϕ to take effect. For example, when assessed from the point of view of mean squared error, $r = 1$ is the best estimator across much of the range of ϕ, with $r = 2$ taking over only at $\phi = 10^{-9}$. The same comparison for logarithmic loss function shows $r = 0$ as best for most of the range of ϕ — thus the usual Bayes estimator appears after all to be the best in practice for this loss function. The CPB results imply that the usual Bayes estimator with $r = 0$ performs very well when assessed from this point of view.

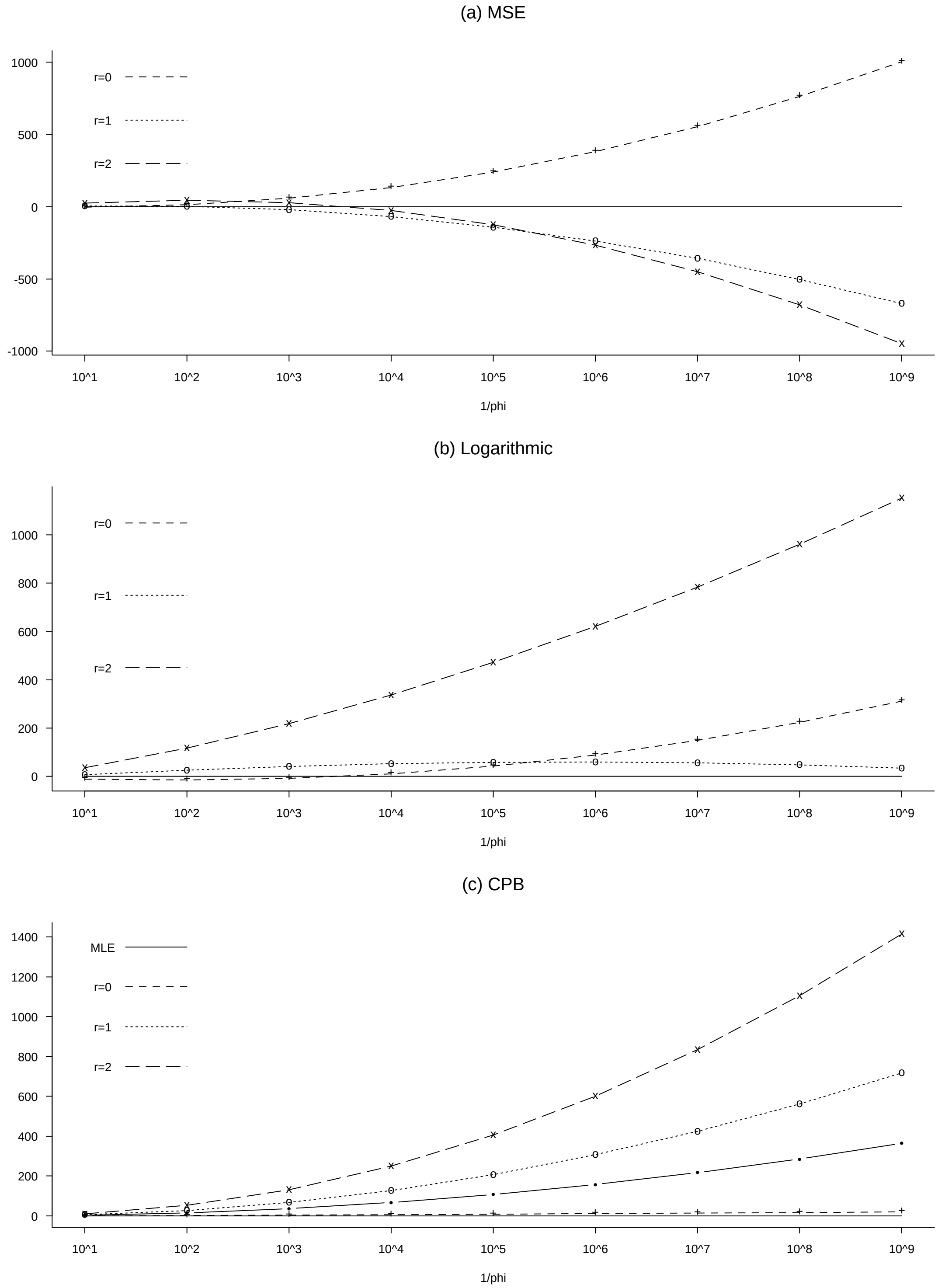

Figure 2. *Normalized values of (a)* $\mathcal{A}$*, (b)* $\mathcal{A}_L$ *and (c) CPB for three Bayes estimators and the MLE; Poisson-GPD model.*

Given that all of these results are asymptotic as $n \to \infty$, it is natural to ask how well they perform for small n. In Smith (1997) some simulations results are shown for $n = 15$. The numerical agreement between asymptotic and theoretical results at this n is not at all good, but the foregoing results nevertheless do a reasonable job of describing comparisons among the four estimators. Simulated mean squared errors show both $r = 1$ and $r = 2$ beating the usual Bayes estimator $r = 0$ across a wide range of ϕ. On the other hand, from the point of view of logarithmic loss, $r = 0$ is good except in the very extreme lower tail of ϕ. Simulated results for CPB show the Bayes estimator with $r = 0$ consistently showing a smaller bias than any of the other three estimators.

4.2 *Estimation of several normal means*

Suppose $X = \{X_{im},\ 1 \le i \le p,\ 1 \le m \le n\}$, where $X_{im} \sim N(\theta_i, 1)$ (mutually independent). This problem for fixed n was studied by James and Stein (1961), who proved the famous result that the maximum likelihood estimator $\hat{\theta}_i = \bar{X}_i = \frac{1}{n}\sum_m X_{im}$ is inadmissible with respect to squared error loss when $p \ge 3$, and established the dominating estimator

$$\hat{\theta}_i^{(JS)} = \hat{\theta}_i \left(1 - \frac{p-2}{n\sum \bar{X}_j^2}\right). \tag{31}$$

(The usual James-Stein estimator is for $n = 1$, but trivial rescaling produces (31) as the estimator for general n.) It is also well known that there is a strong connection between these ideas and Bayesian methods — if the prior on $\theta = (\theta_1, ..., \theta_p)$ is proper, then the Bayes posterior mean is necessarily admissible, and some similar results are known for hierarchical models with improper priors (Berger and Strawderman 1996). The purpose of the present discussion is to relate these results to the asymptotic formulae of earlier in the paper, with a view to providing some insight into more complicated problems for which exact admissibility results are not available.

Consider the improper prior

$$\pi(\theta) = e^{Q(\theta)} \propto \left(\sum_j \theta_j^2\right)^{-\alpha}, \qquad \alpha > 0. \tag{32}$$

Suppose we want to estimate $\phi = \theta_i$ for given index i. In this case it may be shown that the approximate Bayes estimator given by (19) exactly coincides with the James-Stein estimator (31), but with $p-2$ replaced by 2α. Combining the results for different values of i, we find

$$E\left\{\sum_i \left(\hat{\theta}_{i,\mathrm{BAY}} - \theta_i\right)^2 - \sum_i \left(\hat{\theta}_{i,\mathrm{MLE}} - \theta_i\right)^2\right\} \sim \frac{4\alpha(\alpha+2-p)}{n^2\sum\theta_j^2}. \tag{33}$$

Thus from the point of view of the classical mean squared error criterion, the Bayes estimator $\tilde{\theta}$ is asymptotically superior to the MLE $\hat{\theta}$ whenever $0 < \alpha < p-2$, the optimal value being $\alpha = \frac{p-2}{2}$. This choice also results in the usual James-Stein estimator being the approximate Bayesian estimator given by (19).

An alternative treatment uses the prior

$$\pi(\theta) \propto \left\{\sum_j \left(\theta_j - \bar{\theta}\right)^2\right\}^{-\alpha}, \qquad \alpha > 0 \quad (\bar{\theta} = \frac{1}{p}\sum \theta_j) \tag{34}$$

which leads to the approximate formula for the Bayes estimator

$$\hat{\theta}_{i,\mathrm{BAY}} \approx \bar{X}_i - \frac{2\alpha}{n\sum(\bar{X}_j - \bar{X})^2}(\bar{X}_i - \bar{X}) \qquad (\bar{X} = \frac{1}{p}\sum \bar{X}_i) \tag{35}$$

(shrinkage towards the mean rather than towards 0) and for which

$$E\left\{\sum_i \left(\hat{\theta}_{i,\mathrm{BAY}} - \theta_i\right)^2 - \sum_i \left(\hat{\theta}_{i,\mathrm{MLE}} - \theta_i\right)^2\right\} \sim \frac{4\alpha(\alpha+3-p)}{n^2\sum(\theta_j - \bar{\theta})^2}. \tag{36}$$

In this case the optimal α is $\frac{p-3}{2}$, which shows $\hat{\theta}_{\mathrm{BAY}}$ to be superior to $\hat{\theta}_{\mathrm{MLE}}$ whenever $p \geq 4$. Comparing (36) to (33), if $\bar{\theta}$ is small then the improvement on the MLE is greater using (32) than with (34). However for $\bar{\theta}$ far from 0 we would prefer to use (34) — roughly speaking, the Bayesian analysis with prior (32) produces good results only if all the θ_j's are close to 0, whereas the analysis with (34) is good whenever the θ_j's are roughly equal, a much weaker requirement.

So far in this subsection we have said nothing about predictive inference. Suppose, however, that in the spirit of earlier sections we define ϕ_i to be the predictive probability that an observation in the i'th subpopulation exceeds a given threshold z_i. Thus $\phi_i = 1 - H(z_i - \theta_i)$ where H is the standard normal CDF; we also write $h = H'$. The results take on the simplest mathematical form if we assume the formulae are evaluated at $z_i = \theta_i + v$ where v is a constant. Assume also that the Bayes estimate of ϕ_i is based on the prior (34), but weighted by ϕ_i^{-r}, as in (11). In that case, the result analogous to (36) is

$$\begin{aligned} E\left\{\sum_i \left(\hat{\phi}_{i,\mathrm{BAY}} - \phi_i\right)^2 - \sum_i \left(\hat{\phi}_{i,\mathrm{MLE}} - \phi_i\right)^2\right\} &\sim \frac{ph^2(v)}{n^2}\left[\frac{7v^2}{4} - 1 - \frac{6rvh(v)}{1-H(v)}\right. \\ &\left. + (r^2+2r)\left\{\frac{h(v)}{1-H(v)}\right\}^2 + \frac{4\alpha(\alpha+3-p)}{p\sum(\theta_j - \bar{\theta})^2}\right]. \end{aligned} \tag{37}$$

Looking at (37), it can be seen that the last term is the only one that depends on α and is minimized at $\alpha = \frac{p-3}{2}$, exactly as in (36). However, the remaining terms depend on r and v. Using the asymptotic relation $\frac{h(v)}{1-H(v)} \sim v$ as $v \to \infty$, it can be shown that (37) behaves in a manner similar to what we saw in section 2 — in particular, the "optimal" value of r, in the sense of minimizing (37) for very large v, is $r = 2$. Similar results are available for logarithmic loss, producing $r = 1$ as the limiting optimal value. Thus it can be seen that optimal predictive inference for this problem requires a combination of some of the traditional ideas related to James-Stein inference for quadratic loss, with the tail modifications of Bayes estimators which we have already seen in earlier sections.

4.3 *Poisson-Gamma model*

To conclude this paper we give an example of a non-normal empirical Bayes model for which exact calculations are seemingly impossible but which nevertheless falls within the framework of our asymptotic approach. The model was originally discussed by Gaver and O'Muircheartaigh (1987) but developed further by Gelfand and Smith (1990) and subsequently by numerous other authors. Most of the discussion, however, has been concerned with computational aspects of optimizing the Gibbs sampler, rather than the more fundamental questions surrounding the properties of the resulting Bayes inferences.

Suppose p machines are observed for known sampling times $nt_1, ..., nt_p$ and the number of observed failures on the j'th machine is X_j. If θ_j denotes the unknown failure rate for machine j, then $X_1, ..., X_p$ are independent Poisson random variables where

$$\mathrm{E}\{X_j\} = nt_j\theta_j. \tag{38}$$

Compared with earlier authors, this formulation includes a parameter n which is there primarily to enable asymptotic statements. Our numerical results are all for $n = 1$. Gelfand and Smith proposed a hierarchical prior structure based on gamma distributions, with

$$\begin{aligned} \pi(\theta \mid \beta) &= \prod_{j=1}^{p} \frac{\theta_j^{\alpha-1} e^{-\theta_j/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \\ \pi_0(\beta) &= \frac{\delta^{\gamma} e^{-\delta/\beta}}{\beta^{\gamma+1}\Gamma(\gamma)}, \end{aligned} \tag{39}$$

in terms of hyperparameters α, γ and δ. This may also be written in the form

$$\pi(\theta) = \frac{\delta^{\gamma}\Gamma(p\alpha + \gamma)}{\Gamma^p(\alpha)\Gamma(\gamma)} \cdot \frac{(\prod \theta_j)^{\alpha-1}}{(\sum \theta_j + \delta)^{p\alpha+\gamma}}. \tag{40}$$

Smith (1997) considered predictive inference for the time Z to first failure by any of the machines, with particular concern if this first failure time is very small. The distribution function of Z is given by

$$\phi(z; \theta) = 1 - e^{-z\sum_1^p \theta_j}. \tag{41}$$

For this model it is possible to evaluate the expressions for $\mathcal{A}$, $\mathcal{A}_L$ and the CPB given in section 3. To save space we shall not give the algebraic expressions but concentrate on their interpretation. The tail behavior is different from that assumed in (27) and in this case there seems to be no advantage in considering a ϕ^{-r} tail weighting; therefore we ignore that possibility here.

We consider a particular data set consisting of numbers of failures in ten pumps at a nuclear power plant (Table 3 of either Gaver and O'Muircheartaigh (1987) or Gelfand and Smith (1990)). A critical question is how best to specify the hyperparameters α, γ and δ. Gelfand and Smith took $\gamma = 0.1$, $\delta = 1.0$ and specified α by an empirical Bayes argument which, for the data set in question, leads to $\alpha = 1.52$. We shall refer to these values as prior 1. They were careful to avoid any claims of optimality for these values and, likewise, no such claim is made for what follows here. Nevertheless, as an illustration of what the theory can achieve, it will be shown alternative specification $\alpha = \gamma = \delta = 5$ (prior 2) produces superior results. All simulations are carried out assuming that the sample estimates of θ_j are also the correct parameter values.

Table 1 shows simulated values for the MSE, logarithmic loss and CPB of the MLE (plug-in) and both Bayesian procedures, based on 10,000 Monte Carlo replications and using 250 iterations of the Gibbs sampler to evaluate the Bayes estimates. Also shown in the table are theoretical values of the CPB as derived from the asymptotic procedure. These results have been obtained for the case where the true ϕ in (41) is 0.1, but similar results are obtained for other values of ϕ. It can be seen that the Bayes procedure with prior 1 has smaller MSE and logarithmic loss than the MLE, but the Bayes procedure

Table 1. *Comparison of Bayes and MLE procedures for the Poisson-gamma model.*

Estimator	MSE	Log. Loss	CPB (actual)	CPB (theoretical)
MLE	.00050	.0028	.0054	.0050
Bayes Prior 1	.00035	.0023	.0147	.0223
Bayes Prior 2	.00012	.0006	–.0026	.0001

with prior 2 does much better. The theoretical results for CPB are surprisingly accurate considering that this is an asymptotic formula being evaluated for $n = 1$. Similar good agreement between theoretical and simulated values of the CPB has been observed for other values of ϕ and for other simulations of this problem (Smith 1997).

Table 2 shows the differences between the MLE and Bayesian procedures as assessed by MSE and logarithmic loss, computed both from the simulated values in Table 1 and from the asymptotic theory. In this case the numerical agreement is not good, but the main point of the example is that the theoretical results correctly suggest that prior 2 is superior to prior 1. In fact, the original motivation behind the choice of prior 2 was based on using the asymptotic theory to compare a large number of different possible priors. After picking out prior 2 using this mechanism, the simulation results confirmed its superiority to prior 1, and thereby vindicated the use of the asymptotic theory.

Table 2. *Comparison of simulated and asymptotic results for the differences between MSE and logarithmic loss for the Bayes and MLE procedures.*

	MSE	Log. Loss
Prior 1 (simulated)	–.00015	–.00057
Prior 1 (theoretical)	–.00005	+.00058
Prior 2 (simulated)	–.00037	–.00219
Prior 2 (theoretical)	–.00233	–.01316

Finally, we consider the actual predictive distributions of $\phi(z;\theta)$ for the original data. Figure 3 shows the predictive distribution functions, evaluated using 1,000 iterations of the Gibbs sampler, for each of the three estimators. In this case, it turns out that the MLE lies between the two Bayes estimators. In this example it is debatable whether the difference among the three predictive curves is of practical importance, but if it is important, then it is clear that we must think about which of the two Bayes estimates is the better — it is not good enough simply to assume that any Bayes estimate must be superior to the plug-in approach.

5. CONCLUSIONS

The asymptotic methods developed in this paper may be used to assess the frequentist properties of Bayesian predictive inferences. One characteristic feature of the results is that they show to what extent the optimal Bayes procedure depends on the criterion being optimized — much contemporary Bayesian debate focusses on the choice of a prior distribution but these results show the importance of paying attention to the loss function as well. Obviously there are many open questions, including the extent to which asymptotic results are valid in small samples, though such simulation results as have been performed support the assertion that the asymptotic results often given the correct comparisons between different procedures even when the numerical agreement is poor. Future work may well focus on the extension to prediction in time series and spatial processes, and more detailed results for empirical Bayes analysis and hierarchical models.

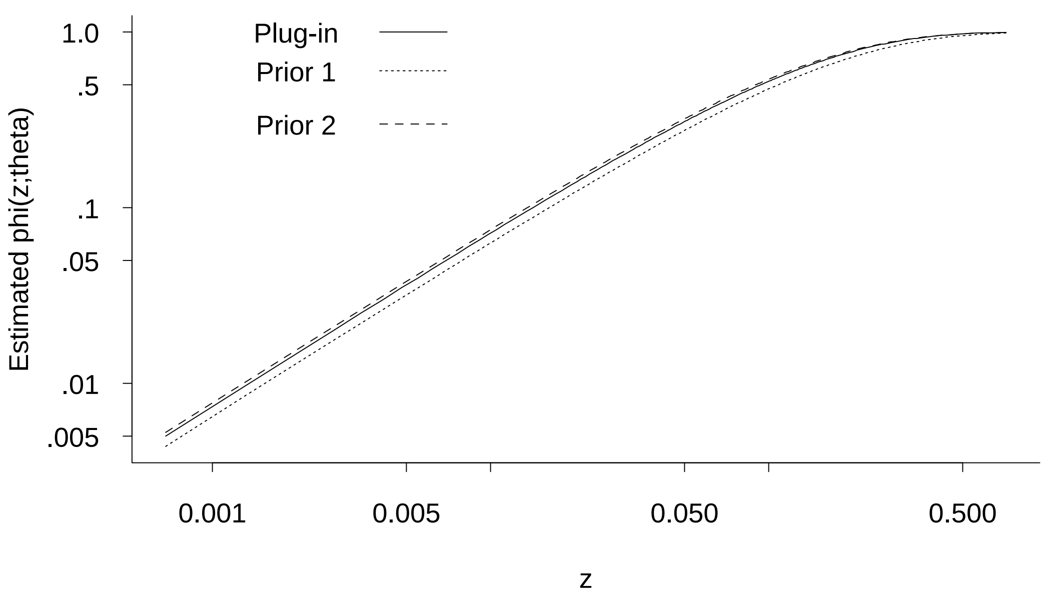

Figure 3. *MLE (plug-in) estimate of $\phi(z;\theta)$ and Bayes posterior means under priors 1 and 2.*

Acknowledgement. This research was supported in part by the National Science Foundation, grant DMS-9705166. The oil spill example is based on continuing joint work with James Hilden-Minton of the National Institute of Statistical Sciences.

REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–554.

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis.* Cambridge University Press.

Barndorff-Nielsen, O.E. and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli* **2**, 319–340.

Berger, J.O. and Strawderman, W.E. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.* **24**, 931–951.

Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion). *J.R. Statist. Soc. B* **48**, 1–38.

Cox, D.R. (1975). Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in Probability and Statistics* (ed. J. Gani). Academic Press, London, pp. 47–55.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Davison, A.C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–332.

Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc.*, **52**, 393-442.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer, New York.

Gaver, D. and O'Muircheartaigh, I. (1987). Robust empirical Bayes analysis of event rates. *Technometrics* **29**, 1–15.

Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Harris, I.R. (1989). Predictive fit for natural exponential families. *Biometrika* **76**, 675–684.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 361–379.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometirka* **96**, 299–313.

McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman and Hall, London.

Reid, N. (1996). Likelihood and approximate Bayesian methods (with discussion). In *Bayesian Statistics 5*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, pp. 351–368.

Resnick, S.I. (1997). Heavy tail modeling and teletraffic data (with discussion). *Ann. Statist.* **25**, 1805–1869.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88**, 355–359.

Smith, R.L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15**, 1174–1207.

Smith, R.L. (1997). Predictive inference, rare events and hierarchical models. Unpublished manuscript, University of North Carolina, Chapel Hill.

Smith, R.L. and Shively, T.S. (1995). A point process approach to modeling trends in tropospheric ozone. *Atmospheric Environment* **29**, 3489–3499.

Sweeting, T.J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–24.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

Vidoni, P. (1995). A simple predictive density based on the p^* formula. *Biometrika* **82**, 855–863.