PREDICTIVE INFERENCE, RARE EVENTS AND HIERARCHICAL MODELS

© Richard L. Smith [1]

October 1 1997

Summary

Many problems of statistics are naturally formulated in terms of the predictive distribution of an as yet unobserved random variable, rather than the more traditional questions of parameter estimation and hypothesis testing. Both frequentist and Bayesian approaches to predictive inference may be considered, but it is argued that the Bayesian approach is more general and flexible. This leads us to consider the frequentist properties of Bayesian procedures. A second-order asymptotic theory is developed, for comparing the risks of different procedures under a variety of loss functions. Particular attention is paid to the tails of the predictive distribution. This leads to some surprising but general conclusions. The usual Bayesian method may well be inferior to a crude maximum likelihood "plug-in" approach, but in most cases the comparison is reversed with a more general specification of the Bayesian procedure. An extension is outlined to empirical Bayes problems. The classical James-Stein theorem is reinterpreted as an asymptotic result about the properties of Bayesian procedures, and it is shown how the same ideas may be applied to a general class of Bayesian hierarchical models.

1 INTRODUCTION

We begin with two real-data examples, intended to illustrate the need for a theory of predictive inference, and to demonstrate the practical efficacy of a Bayesian approach.

Fig. 1.1(a) shows the five best performances by different athletes in the women's 3000 metre running event, for each year from 1972 to 1992. The dotted lines connect up the annual minima. The final (right-hand) point of the plot is the remarkable record achieved by the Chinese athlete Wang Junxia in 1993. Many questions were raised about Wang's record, in particular whether she could have been taking illegal drugs. However neither Wang herself nor any member of her training group ever failed a drugs test, so the only evidence for such an assertion is that the record itself was too good to have been achieved by normal training methods.

[1] Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260, USA; email rs@stat.unc.edu. This research was supported in part by NSF grant DMS-9705166. I would like to thank Alastair Young and Jim Berger for many conversations about decision theory, and Jonathan Tawn and Dougal Goodman for discussions and data related to the two examples used in Section 1.

Robinson and Tawn (1995) proposed a statistical test. They fitted models based on extreme value theory to the five best performances in each year, including an exponentially decaying trend, and obtained confidence intervals for x_{ult}, a parameter representing the best possible long-term performance. They did this for a number of models, but in every case, the 95% confidence interval for x_{ult} included Wang's performance. Thus, while their analysis strengthened the assertion that Wang's record was very unusual — which might be interpreted as indirect evidence of drug use — it failed to provide conclusive evidence that Wang's performance was inconsistent with previous data.

Smith (1997a) argued that a more appropriate formulation of the problem is in terms of the *predictive* distribution of the best performance for 1993, given the years 1972–1992. Although Smith considered both Bayesian and frequentist approaches, the Bayesian method apparently yielded more satisfactory answers. Let Z denote the (as yet unobserved) best performance for 1993, and consider the conditional distribution of Z given that a new record is set. Thus define $G(z;\theta) = \Pr\{Z \leq z \mid Z \leq z_0;\theta\}$ where $z_0 (= 502.62)$ is the existing record and θ are the unknown model parameters. Then the standard Bayesian definition of the predictive distribution function (Aitchison and Dunsmore, 1975) is

$$\hat{G}(z) = \int G(z;\theta)\pi(\theta|\mathbf{X})d\theta. \tag{1.1}$$

Here $\pi(\cdot|\cdot)$ is the posterior density of θ given observed data $\mathbf{X}$. The analysis was simplified by fitting a model without trend to the data from 1980 onwards. In this context, the model is equivalent to assuming a three-parameter Weibull distribution for the best performance in each year, but fitted to the five best performances. An uninformative prior distribution was taken. The resulting $\hat{G}(z)$ is shown in Fig. 1.1(b). In particular, from this we read off that for $z = 486.11$, which was the actual record achieved by Wang, we have $\hat{G}(z) \approx 0.0006$, a slight revision of the value .00047 quoted in Smith (1997a). Looked at from this point of view, it indeed appears that Wang's record was extremely unusual.

Bayesian statisticians sometimes argue that it is more meaningful to look at the full posterior distribution of a quantity of interest than any single summary value such as the posterior mean. However, in the present case it is hard to see how one would use the additional information. The quantity of interest is $\phi = G(486.11;\theta)$ and the posterior probability that $\phi = 0$ is .925. Given $\phi > 0$, the posterior density of $-\log\phi$ is as in Fig. 1.1(c). The mode of this density is at $-\log\phi$=6 (ϕ=.0025) but there seems no reason to quote this as the "Bayes estimator".

There are a number of other issues such as whether it was really correct to ignore the trend, a very reasonable point raised by Robinson and Tawn in their reply to Smith (1997a), but the main focus of the present paper is on broader theoretical aspects of this kind of analysis. In particular, it seems highly pertinent to ask whether the quoted predictive tail probability of .0006 has any interpretation in terms of long-run frequency properties of the procedure.

For our second example, Fig. 1.2(a) depicts, on a logarithmic scale, 425 large insurance claims made against a company over a 15-year period. Once again there are many questions over an appropriate model for such data, and Smith (1997b) considers a number of alternatives. The present discussion, for illustrative purposes, is based on the simplest form of model for exceedances over thresholds developed by Smith (1989) and Davison and Smith (1990). We fix a threshold $u = 5$, over which there are 73 exceedances. The exceedance times are taken to form a Poisson process with constant rate λ, estimated as $\hat{\lambda} = 73/15 = 4.87$. The excesses over the threshold are taken to follow a generalised Pareto distribution (GPD) with distribution function

$$F(x; \psi, \xi) = 1 - \left(1 + \frac{\xi x}{\psi}\right)^{-1/\xi}, \quad x \geq 0, \tag{1.2}$$

where ξ is a shape parameter. The maximum likelihood estimates are $\hat{\psi} = 6.3$, $\hat{\xi} = 0.89$. If $\xi \geq \frac{1}{2}$ the variance of the claim size is infinite, and indeed we are close to the value $\xi = 1$ for which the mean becomes infinite. The estimate of ξ is reduced somewhat if mixing over different claim types is taken into account (Smith 1997b), but even so, this is an extremely long-tailed distribution, though no more so than appears to be typical in the insurance industry.

In this case the relevant question for the company appears to be, "How much of a financial reserve is needed to cover all insurance losses (with specified probability) over the next N years?" This is again a question about a predictive distribution, rather than the value of any particular parameter, and in Smith (1997b) it was answered by combining Bayesian estimation of the parameters with Monte Carlo simulation of the distribution of total cumulative claim. For the present discussion, we confine ourselves to the computationally simpler question of determining the predictive distribution of Z, defined as the largest claim over a ten-year period. The distribution function of the annual maximum and hence that of Z may be written down explicitly (equation 9.1 of Davison and Smith, 1990) and Fig. 1.2(b) shows the Bayesian predictive distribution of $\Pr\{Z > z\}$, computed exactly as in (1.1) except for the change in direction of the inequality. The Bayesian approach taken here differs sharply from the usual actuarial approach to such problems, which is much closer to the "plug-in" method (equation 1.3 below).

There are a number of other recent examples of predictive inference for extreme-value problems (Coles and Powell 1996, Coles and Tawn 1996), but the foregoing examples are intended to illustrate some points of a more general nature:

1. Many statistical problems are more meaningfully formulated as being about the predictive distribution of some unobserved random variable than about tests of hypotheses or parameter estimates.

2. The Bayesian approach to prediction problems, as in (1.1), is simple, powerful and flexible. However this does raise questions — for Bayesians as well as frequentists — about what properties this procedure has.

The simplest form of non-Bayesian procedure is the "plug-in" approach where estimates are substituted for unknown parameters. If the estimates are maximum likelihood then we call this the MLE approach:

$$\hat{G}_{MLE}(z) = G(z; \hat{\theta}) \tag{1.3}$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ. Although this is widely criticised for taking no account of the uncertainty in estimating θ, it should be pointed out that since $G(z;\theta)$ is (for fixed z) simply a nonlinear function of θ, (1.3) *is* the maximum likelihood estimator of $G(z;\theta)$, so it may not be all bad. On the other hand Cox (1975) pointed out that prediction intervals based on $\hat{G}_{MLE}(z)$ will typically have coverage probabilities that are too small, compared with their nominal values derived under the assumption that θ is known, and the main point of his contribution was to derive an asymptotic formula to correct this.

There is an extensive literature on Bayesian approaches to predictive inference, well represented by Geisser (1993). Barndorff-Nielsen and Cox (1994, pp. 316–317) have described a number of non-Bayesian approaches, but most of these are available only under specialised circumstances. Butler's (1986) predictive likelihood depends on the existence of a low-dimensional sufficient statistic. There is an exact approach due to Cox (1975), also discussed on p. 242–245 of Cox and Hinkley (1974), but this requires the existence of an exact similar test for the equality of the parameters determining the distributions of $\mathbf{X}$ and Z. A third method requires an exact pivotal statistic. None of these is applicable to the kinds of parametric families we have considered here. Thus, the asymptotic approach of Cox (1975) appears to be the only one which has any hope of applicability to our problems. There is also an asymptotic variant on Butler's approach due to Davison (1986), not requiring sufficient statistics, but Davison's approach would be more accurately described as an approximate Bayesian method and so leads us anyway into questions about the properties of Bayesian methods.

There is an extensive and growing literature on comparisons between Bayesian and frequentist procedures. Much attention in particular has been given to the question of constructing Bayesian interval estimates of parameters which achieve good coverage probability in the conventional frequentist sense. The classical results of Welch and Peers (1963), Welch (1965) and Peers (1965), together with Stein (1985), have been supplemented by many new results in recent years, for example Tibshirani (1989), Ghosh and Mukerjee (1992, 1993), Nicolau (1993), Efron (1993, 1996), Mukerjee and Dey (1993), Datta and Ghosh (1995), Datta (1996). One paper of particular interest is that of Liseo (1993), who argued through a series of examples that procedures based on the reference prior approach of Berger and Bernardo (1992) have superior frequentist properties to those based on a Jeffreys prior or on various forms of conditional inference. For reviews covering both these and other aspects of the Bayes-frequentist link, see Reid (1996) and Sweeting (1996). Within this literature, however, comparatively little attention has been given to predictive inference. By analogy with Welch-Peers and their successors, one could derive conditions under which Bayesian predictive distributions have asymptotically correct frequentist coverage probabilities, thus achieving by Bayesian means what Cox (1975) sought to achieve with an asymptotic correction, but one of the points of the present paper is to argue that this is too narrow a viewpoint by which to assess predictive procedures. An alternative approach was introduced by Aitchison (1975), who used Kullback-Leibler distance as a measure of how well the predictive density of Z approximates its true unknown density. This idea has been recently extended by Komaki (1996), who used differential geometric means to derive an optimal perturbation of the plug-in estimator. He also gave an asymptotic expansion of the Bayesian predictive density. By combining both results, it is possible in principle to compare different Bayesian predictive distributions from the point of view of Kullback-Leibler distance. However, the examples he gave were very limited, and his differential geometric approach seems harder to apply than the direct expansion in terms of cumulants of the log likelihood given in Section 3 of the present paper.

The plan of the rest of the paper is as follows. Section 2 develops detailed results for a particular problem, the one-parameter exponential distribution. This shows in particular that the usual definition of Bayesian predictive distribution might not be optimal for some natural criteria, and raises the question of whether such results might be true for more general models. These questions are explored in detail in Sections 3 and 4, where an expansion for a general parametric family is derived, and applied to the tail probabilities of the predictive distribution. Section 5 discusses the Poisson–GPD model used for the insurance example. Sections 6 and 7 discuss applications to empirical Bayes problems and hierarchical models, and Section 8 contains concluding discussion. To avoid interrupting the text with detailed derivations, all the more intricate calculations of the paper are deferred to a series of technical appendices.

2 PREDICTION BASED ON THE EXPONENTIAL DISTRIBUTION

In this section, a detailed theory is developed for the one-parameter exponential distribution. This is an appealing example to consider because both Bayesian and conditional frequentist formulae may be computed explicitly, and this greatly simplifies theoretical comparison of the procedures. The exponential distribution is also of practical relevance to the problem of exceedances over a high threshold, since it arises as a limiting case of (1.2) with $\xi = 0$, and there is a long history of the exponential distribution being applied to exceedance data on either the original or a logarithmic scale (Davison and Smith 1990).

Let $X_1, ..., X_n$ be independent and identically distributed (i.i.d.) from the density $\theta e^{-\theta x}$, $x > 0$, where $\theta > 0$ is an unknown parameter. Let $S_n = X_1 + ... + X_n$; the MLE of θ is $\hat{\theta} = n/S_n$. Let Z denote some hypothetical future value derived from the same distribution. If θ were known, then the exact predictive distribution for Z would be given by $\Pr\{Z > z\} = \phi(z;\theta) = e^{-\theta z}$ for each z. We seek an estimator $\{\hat{\phi}(z),\ z > 0\}$ which approximates $\{\phi(z;\theta),\ z > 0\}$ as closely as possible, in some suitably defined sense. We shall concentrate primarily, though not exclusively, on a mean squared error criterion: an estimator is good if $\mathrm{E}\{(\hat{\phi}(z) - \phi(z;\theta))^2\}$ is small. Where there is no ambiguity about the value of z being considered, we shall simply write ϕ in place of $e^{-\theta z}$ and $\hat{\phi}$ for an estimator of ϕ.

The simplest and most obvious estimator is the "plug-in" estimator:

$$\hat{\phi}_{MLE} = e^{-\hat{\theta} z}. \tag{2.1}$$

A second estimator is Bayesian. With gamma prior, $\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$ for $\theta > 0$, the posterior distribution of θ is gamma with parameters (α, β) replaced by $(\alpha + n, \beta + S_n)$, so the posterior mean of ϕ is

$$\hat{\phi}_{BAY,0} = \left(\frac{\beta + S_n}{\beta + S_n + z} \right)^{\alpha+n}. \tag{2.2}$$

The reason for the suffix 0 will appear shortly. The limiting case $\alpha = \beta = 0$ is the Jeffreys prior and therefore of particular interest.

A third approach is based on the fact that $W = Z/(Z + S_n)$ has density $n(1 - w)^{n-1}$ on $0 < w < 1$, which is known exactly and does not depend on θ. One consequence of this is that if we fix $p \in (0, 1)$ and choose constants a and b such that $0 \leq a < b \leq 1$ and $(1-a)^n - (1-b)^n = p$, then $(aS_n/(1-a), bS_n/(1-b))$ is an exact $100p\%$ prediction interval for Z. We also note that $\Pr\{Z > tS_n\} = (1 + t)^{-n}$ for each $t > 0$; substituting $z = tS_n$, it therefore seems reasonable to write $(1 + z/S_n)^{-n}$ as an estimator of ϕ; this is exactly the Jeffreys version of (2.2). The connection is not accidental, because what we have done is equivalent (in this special case) to Fisher's (1934) derivation of an exact conditional distribution given the maximal ancillary statistic of a location-scale family, which has long been known to be operationally equivalent to Bayesian analysis under a Jeffreys prior. The method may also be derived as Cox's (1975) exact method, or more simply by noting that W is pivotal. Thus, this is a case where the various frequentist approaches mention in Section 1 do work, and apparently agree with the Bayesian approach with Jeffreys prior.

Thus from both Bayesian and frequentist viewpoints, the estimator (2.2) with $\alpha = \beta = 0$ seems superior to (2.1). The following argument, however, shows that the comparison is not so straightforward.

By writing $S_n = n/\hat{\theta}$ and Taylor expanding –log$(\hat{\phi}_{BAY,0})$ about $\hat{\theta} z$, one deduces that

$$\hat{\phi}_{BAY,0} = e^{-\hat{\theta} z} \left\{ 1 + \frac{1}{n} \left(-\alpha \hat{\theta} z + \frac{1}{2} \hat{\theta}^2 z^2 + \beta \hat{\theta}^2 z \right) + O\left(\frac{1}{n^2} \right) \right\}. \tag{2.3}$$

Both $\hat{\phi}_{MLE}$ and $\hat{\phi}_{BAY,0}$ are therefore of the form

$$\hat{\phi} = e^{-\hat{\theta} z} \left\{ 1 + \frac{1}{n} \left(A \hat{\theta} z + B \hat{\theta}^2 z^2 \right) + O\left(\frac{1}{n^2} \right) \right\} \tag{2.4}$$

for constants A and B, possibly depending on z, but not on θ or $\hat{\theta}$.

Our main result in this section is that, for (2.4),

$$\mathrm{E}\left\{(\hat{\phi}-\phi)^2\right\} = \frac{\theta^2 z^2 e^{-2\theta z}}{n} + \frac{\theta^2 z^2 e^{-2\theta z}}{n^2} q(A,B,\theta z) + O\left(\frac{1}{n^3}\right), \qquad (2.5)$$

where, writing y in place of θz,

$$q(A,B,y) = 5 - 4A + A^2 - y(7 + 6B - 2AB - 3A) + y^2\left(B^2 + 3B + \frac{7}{4}\right). \qquad (2.6)$$

The leading term in (2.5) is the asymptotic variance of the MLE and this term will be present in all expressions of this nature. Therefore, for sufficiently large n, the comparative performance of different estimators is determined by the function $q(A,B,y)$. For $\hat{\phi}_{MLE}$ we have $A = B = 0$ and therefore

$$q(A,B,y) = q^* = 5 - 7y + \frac{7}{4}y^2.$$

For $\hat{\phi}_{BAY,0}$ with $\alpha = \beta = 0$, we have $A = 0,\ B = \frac{1}{2}$ and so

$$q(A,B,y) = q_0 = 5 - 10y + \frac{7}{2}y^2.$$

Thus $q_0 < q^*$ if and only if $y < \frac{12}{7}$, or $\phi > e^{-12/7} = 0.1801$. In other words, $\hat{\phi}_{BAY,0}$ indeed outperforms $\hat{\phi}_{MLE}$ for moderately large values of ϕ, but $\hat{\phi}_{MLE}$ performs better when the true value of ϕ is small ($<.1801$). Since a part of our interest is in cases when ϕ is very small, this is a disturbing conclusion.

It is of course possible to consider the effect of other prior parameters α and β, but this does not resolve the fundamental difficulty. In this case A and B are replaced by $-\alpha$ and $\frac{1}{2} + \frac{\beta}{z}$ respectively. However for large z, the behaviour of q_0 is still dominated by the term $\frac{7}{2}y^2$, which is larger than the corresponding term $\frac{7}{4}y^2$ in the expression for q^*. Therefore, for any α and β, we will have $q_0 > q^*$ for all sufficiently large values of y.

Up to this point, our Bayesian analysis has been based entirely on the posterior mean of ϕ. It is well known that this is the optimal Bayes estimator under squared error loss. However, it is not necessarily optimal with respect to other loss functions. Suppose, instead of $L(\phi,\hat{\phi}) = (\phi - \hat{\phi})^2$, we consider $L(\phi,\hat{\phi}) = (\phi - \hat{\phi})^2/w(\phi)$ for some function $w(\phi)$. In particular, we might consider $w(\phi) = \phi^2$ in order to concentrate on *relative* error in the lower tail as $\phi \to 0$. More generally, taking $w(\phi) = \phi^r$ allows for a general class of loss functions for which the Bayes estimator is still explicitly computable.

It is useful to draw a distinction between the loss function used by the statistician to derive a Bayes estimator, and the loss function by which the procedure will ultimately be evaluated. The first will be called the *statistician's loss function*, the second the *client's loss function*. For the moment, the client's loss function is still $(\phi - \hat{\phi})^2$, but the statistician's loss function is $(\phi - \hat{\phi})^2/\phi^r$ to allow a wider range of Bayesian procedures.

For a general loss function, the Bayes estimator is the value of $\hat{\phi}$ which minimises $\int L(\phi, \hat{\phi})\pi(\phi|\mathbf{X})d\phi$, where $\pi(\phi|\mathbf{X})$ is the posterior density of parameter ϕ based on data $\mathbf{X}$. This is the "conditional Bayes decision principle", cf. Berger (1985), page 16. For $L(\phi, \hat{\phi}) = (\phi - \hat{\phi})^2/w(\phi)$, we have

$$\hat{\phi} = \frac{\int \{\phi/w(\phi)\}\pi(\phi|\mathbf{X})d\phi}{\int \{1/w(\phi)\}\pi(\phi|\mathbf{X})d\phi}. \tag{2.7}$$

When applied to predictive inference with $\phi(z;\theta)$ either the distribution function or the survivor function of a random variable Z, we shall call this the tail-weighted predictive distribution, in contrast with (1.1) which represents the unweighted case. Thus for example, if $w(\phi) = \phi^r$ with $r = 1$ then our point estimator of ϕ is $1/\mathrm{E}\{\phi^{-1}|\mathbf{X}\}$, while if $r = 2$ it is $\mathrm{E}\{\phi^{-1}|\mathbf{X}\}/\mathrm{E}\{\phi^{-2}|\mathbf{X}\}$, where in both cases $\mathrm{E}\{...|\mathbf{X}\}$ means posterior expectation conditioned on the data $\mathbf{X}$. For $w(\phi) = \phi^r$ with general r, and the same prior density as before, it is readily verified that the Bayes estimator becomes

$$\hat{\phi}_{BAY,r} = \frac{\mathrm{E}\{\exp((r-1)z\theta)|\mathbf{X}\}}{\mathrm{E}\{\exp(rz\theta)|\mathbf{X}\}} = \left\{\frac{\beta + S_n - rz}{\beta + S_n - (r-1)z}\right\}^{\alpha+n}. \tag{2.8}$$

Analogously to (2.3), we now have

$$\hat{\phi}_{BAY,r} = e^{-\hat{\theta}z}\left[1 + \frac{1}{n}\left\{-\alpha\hat{\theta}z - \left(r - \frac{1}{2}\right)\hat{\theta}^2 z^2 + \beta\hat{\theta}^2 z\right\} + O\left(\frac{1}{n^2}\right)\right] \tag{2.9}$$

which is again of form (2.4) with $A = -\alpha$, $B = \frac{1}{2} - r + \frac{\beta}{z}$. Therefore, (2.5)–(2.6) still apply. The coefficient of y^2 in $q(A, B, y)$ is

$$r^2 - 4r + \frac{7}{2}$$

which is minimised when $r = 2$. Thus there is a sense in which $r = 2$ is optimal for estimating very small probabilities.

Fig. 2.1 shows simulated mean squared errors, based on 5,000 replications, for the maximum likelihood estimator and three Bayes estimators, corresponding to $r = 0$, 1 and 2 with Jeffreys prior. The estimates are evaluated at values of z corresponding to true ϕ values .5, .25, .1, .01 and .001. Also, they are repeated for three sample sizes, $n = 10$, 40 and 200. To aid comparison, the MSEs are standardised so that the maximum likelihood estimator has MSE=1 in each instance. Although the three plots have very different scales on the vertical axis, they are of similar shape, and confirm that while the usual Bayes estimator ($r = 0$) performs best in the middle of the distribution, represented by $\phi = .5$

here, it is the worst of the four in the tail as $\phi \to 0$, and that the $r = 2$ estimator performs best in the tail. For comparison, the four quadratic curves derived from (2.6) are shown in Fig. 2.2, and confirm the comparative merits of the different procedures apparent from the simulations.

Another question raised by the foregoing analysis is whether the squared error loss function is in fact the most appropriate specification of what we have called the client's loss function. An alternative is the logarithmic loss function

$$L(\phi, \hat{\phi}) = \phi \log(\phi/\hat{\phi}) + (1-\phi)\log\{(1-\phi)/(1-\hat{\phi})\}. \tag{2.10}$$

This is motivated, in part, by the theory of proper scoring rules in assessing probability forecasters (Seillier-Moiseiwitsch and Dawid 1993), where (2.10) or an equivalent expression is often preferred to the squared error loss function for assessing tail probabilities. It is readily verified that if (2.10) is also taken as the statistician's loss function, then the conditional Bayes decision principle again leads to the posterior mean of ϕ as the Bayes estimator. However, in the spirit of our earlier discussion, it is possible to vary the statistician's loss function in this case as well, and it is natural to consider the same class of tail-weighted Bayes estimators, indexed by r, as previously. For this loss function, the result analogous to (2.5) is

$$\mathrm{E}\{L(\phi, \hat{\phi})\} = \frac{\theta^2 z^2 e^{-\theta z}}{2n(1-e^{-\theta z})} + \frac{\theta^2 z^2 e^{-\theta z}}{24n^2(1-e^{-\theta z})^3} q^\dagger(A, B, \theta z) + O\left(\frac{1}{n^3}\right), \tag{2.11}$$

where

$$\begin{aligned} q^\dagger(A,B,y) = {} & 60 - 28y - 48A + 12A^2 + 3y^2 - 72By + 12B^2y^2 + 24ABy + 12By^2 \\ & + 12Ay + 60e^{-2y} - 120e^{-y} - 12Aye^{-2y} - 72Bye^{-2y} - 12By^2e^{-2y} \\ & - 24B^2y^2e^{-y} + 12B^2y^2e^{-2y} + 144Bye^{-y} + 12y^2e^{-y} + 3y^2e^{-2y} - 24A^2e^{-y} \\ & + 96Ae^{-y} - 48Ae^{-2y} + 12A^2e^{-2y} + 28ye^{-2y} - 48ABye^{-y} + 24ABye^{-2y}. \end{aligned} \tag{2.12}$$

The dominant term as $y \to \infty$ is $(3 + 12B + 12B^2)y^2$ which is minimised when $B = -\frac{1}{2}$. For the Jeffreys prior with general r we have $B = \frac{1}{2} - r$, so the optimal value of r, for prediction in the tail of the distribution, is $r = 1$ (statistician's loss function $(\phi - \hat{\phi})^2/\phi$). However, in this case we have to go both to larger n and smaller ϕ to achieve excellent agreement between simulations and asymptotic theory. Fig. 2.3 shows normalised MSE for four estimators, based on 10,000 replications with $n = 500$, and a range of values of $y = \theta z = -\log\phi$ from 2 to 20. Fig. 2.4 shows the corresponding values of $q^\dagger$ derived from (2.12). In this case, among these four estimators, the standard Bayes estimator ($r = 0$) performs best up to about $y = 4$, then the MLE performs best up to around $y = 12$, after which $r = 1$ is best.

The results of this section show that the optimal Bayes procedures depend on the client's loss function in unexpected ways. If the objective is to obtain a prediction interval whose coverage probability exactly matches its nominal value — in the next section this property will be called *unbiased in coverage probability* — then the conditional frequentist arguments show we should take the Jeffreys prior with $r = 0$. This is an exact result. If the loss function is mean squared error, then for accurate prediction of tail probabilities, best results are obtained with $r = 2$. Under the logarithmic loss function, the best value is $r = 1$.

A brief description of the derivation of (2.5)–(2.6) and (2.11)–(2.12) is given in Appendix A.

3 PREDICTION IN GENERAL PARAMETRIC MODELS

The arguments of the previous section may be extended to very general classes of parametric families. The results rely heavily on formal manipulations of asymptotic expressions and no attempt is made to provide either a precise statement of regularity conditions or rigorous proofs. Most of the technical details are removed to appendices.

Suppose we have a finite-parameter family with log likelihood function $\ell_n(\theta)$ where n denotes sample size. Suppose we are interested in a one-dimensional function of θ, denoted ϕ or $\phi(\theta)$. In later discussion, this will be the value of a predictive distribution function $G(z;\theta)$ evaluated at a given z. A typical Bayes estimator will be of the form

$$\hat{\phi}_{BAY} = \frac{\int \phi(\theta) e^{\ell_n(\theta)+Q(\theta)} d\theta}{\int e^{\ell_n(\theta)+Q(\theta)} d\theta} \tag{3.1}$$

where $Q(\theta)$ is some function determined by both the prior density and the loss function. For example, in the situation of Section 2, we have $Q(\theta) = (\alpha - 1)\log\theta + (rz - \beta)\theta$.

For the log likelihood and its derivatives, we adopt the notational conventions used in Chapter 7 of McCullagh (1987). Suppose $\kappa_{ij} = \frac{1}{n}\mathrm{E}\{\partial^2 \ell_n(\theta)/\partial\theta^i\partial\theta^j\}$, $\kappa_{ijk} = \frac{1}{n}\mathrm{E}\{\partial^3 \ell_n(\theta)/\partial\theta^i\partial\theta^j\partial\theta^k\}$, where θ^i, θ^j... denote components of the vector θ. Where suffixes are separated by commas, this denotes a covariance, for example $\kappa_{i,jk} = \frac{1}{n}\mathrm{Cov}\{\partial\ell_n(\theta)/\partial\theta^i,\ \partial^2\ell_n(\theta)/\partial\theta^j\partial\theta^k\}$. We shall not require any higher-order cumulants. The Fisher information matrix has entries $\{\kappa_{i,j}\}$ and we denote the entries of its inverse by $\{\kappa^{i,j}\}$. All these quantities are defined exactly (independent of n) whenever ℓ_n is formed from n i.i.d. random variables but the theory holds in non-i.i.d. cases provided the foregoing quantities are all of $O(1)$. Where other quantities are concerned we use suffixes to denote partial differentiation, for example $Q_i = \partial Q/\partial\theta^i$, $\phi_{ij} = \partial^2\phi/\partial\theta^i\partial\theta^j$. All these quantities are evaluated at the true θ unless denoted otherwise. The maximum likelihood estimator (MLE) is denoted $\hat{\theta}$ with components $\hat{\theta}^i$. The MLE of ϕ is $\hat{\phi} = \phi(\hat{\theta})$. We adopt the summation convention whereby repetition of an index denotes summation over that index. Then a direct extension of the argument of p. 209 of McCullagh (1987), given in more detail in Appendix B, shows that

$$\mathrm{E}(\hat{\phi} - \phi) = n^{-1}\left\{\kappa^{i,j}\kappa^{k,\ell}(\kappa_{\ell,jk} + \frac{1}{2}\kappa_{jk\ell})\phi_i + \frac{1}{2}\kappa^{i,j}\phi_{ij}\right\} + O(n^{-3/2}). \tag{3.2}$$

The main result of this section is an expression for the mean squared error of the Bayes estimator $\hat{\phi}_{BAY}$, relative to that of the MLE $\hat{\phi}$. The result is

$$E\{(\hat{\phi}_{BAY}-\phi)^2\}-E\{(\hat{\phi}-\phi)^2\}=\frac{\mathcal{A}}{n^2}+o\left(\frac{1}{n^2}\right). \tag{3.3}$$

To evaluate $\mathcal{A}$, first define a quantity $\mathcal{C}$ by

$$\mathcal{C}=\kappa_{ijk}\kappa^{i,k}\kappa^{j,\ell}\phi_\ell+\kappa^{i,j}(\phi_{ij}+2\phi_i Q_j) \tag{3.4}$$

and let $\mathcal{C}_s=\partial\mathcal{C}/\partial\theta_s$. We then have

$$\begin{aligned}\mathcal{A}=&\frac{\mathcal{C}^2}{4}+\mathcal{C}_s\kappa^{s,t}\phi_t\\ &+\mathcal{C}\left\{\kappa^{i,j}\kappa^{k,\ell}\left(\kappa_{\ell,jk}+\frac{1}{2}\kappa_{jk\ell}\right)\phi_i+\frac{1}{2}\kappa^{i,j}\phi_{ij}\right\}\\ &+\kappa^{i,k}\kappa^{j,\ell}\kappa^{s,t}\phi_i\Big\{\kappa_{k,\ell t}(\phi_{js}+2\phi_s Q_j)\\ &\quad+(\kappa_{k,\ell st}+\kappa^{u,v}\kappa_{k,su}\kappa_{\ell tv}+\kappa^{u,v}\kappa_{k,\ell s}\kappa_{utv})\phi_j\Big\}.\end{aligned} \tag{3.5}$$

A derivation of (3.3)–(3.5) is given in Appendix C.

As a by-product of the same calculation, we also have an expression for the bias of $\hat{\phi}_{BAY}$:

$$\mathrm{E}\{\hat{\phi}_{BAY}-\phi\}=n^{-1}\left\{\kappa^{i,j}\kappa^{k,\ell}(\kappa_{\ell,jk}+\kappa_{jk\ell})\phi_i+\kappa^{i,j}(\phi_{ij}+\phi_i Q_j)\right\}+O(n^{-3/2}). \tag{3.6}$$

We may also consider other loss functions $L(\phi,\hat{\phi})$. Suppose $L(\phi,\phi)=0$ and let $L_j=\partial^j L/\partial\hat{\phi}^j$ evaluated at $\hat{\phi}=\phi$. We assume L_j exists up to $j=4$ and that $L_1=0,\ L_2>0$ so that $L(\phi,\hat{\phi})$ is minimised with respect to $\hat{\phi}$ when $\hat{\phi}=\phi$. Then (Appendix D)

$$E\{L(\phi,\hat{\phi}_{BAY})\}-E\{L(\phi,\hat{\phi})\}=\frac{\mathcal{A}_L}{n^2}+o\left(\frac{1}{n^2}\right) \tag{3.7}$$

where

$$\mathcal{A}_L=\frac{1}{2}L_2\mathcal{A}+\frac{1}{4}L_3\mathcal{C}(\phi_i\phi_j\kappa^{i,j}). \tag{3.8}$$

For example, in the case of logarithmic loss (2.10) this reduces to

$$\mathcal{A}_L=\frac{\mathcal{A}}{2\phi(1-\phi)}-\frac{(1-2\phi)\mathcal{C}\phi_i\phi_j\kappa^{i,j}}{2\phi^2(1-\phi)^2}. \tag{3.9}$$

For yet another example of a loss function, consider the *squared logarithmic loss function*

$$L(\phi, \hat{\phi}) = (\log \phi - \log \hat{\phi})^2. \tag{3.10}$$

When the true ϕ is very small, the squared logarithmic loss function penalises underprediction of ϕ more severely than the logarithmic loss function, and much more severely than squared error loss. Thus it might be a useful loss function in contexts such as insurance, where it is desirable to be conservative in estimating loss probabilities. In this case $L_2 = 2/\phi^2$, $L_3 = -6/\phi^3$, so

$$\mathcal{A}_L = \frac{\mathcal{A}}{\phi^2} - \frac{3\mathcal{C}\phi_i\phi_j\kappa^{i,j}}{2\phi^3}. \tag{3.11}$$

The final criterion considered in this section is *bias in coverage probability.* Suppose the parameter ϕ is in fact one of a family of parameters indexed by a scalar z, for which we write $\phi(z;\theta)$ to emphasise the dependence on θ as well as z. We use primes to denote differentiation with respect to z and, as usual, suffixes to denote differentiation with respect to components of θ. For example, $\phi_i'(z;\theta)$ is the same thing as $\partial^2\phi(z;\theta)/\partial z\partial\theta_i$. In the context of prediction intervals, either $\phi(z;\theta)$ or $1-\phi(z;\theta)$ may denote the distribution function of an as yet unobserved random variable Z, and we may be interested in deriving a prediction interval for Z with specified coverage probability. This problem may be characterised as that of finding a statistic $\tilde{z}$ for which $\phi(\tilde{z};\theta)$ is as close as possible to some target value α. The *bias in coverage probability* is then defined to be $\mathrm{E}\{\phi(\tilde{z};\theta)\} - \alpha$.

Suppose $\phi(z;\theta)$ is estimated for each z by an estimator $\tilde{\phi}(z)$, where for this part of the discussion, $\tilde{\phi}(z)$ may be used to denote either the MLE or the Bayes estimator. Then the obvious approach is to define $\tilde{z}$ by the equation

$$\tilde{\phi}(\tilde{z}) = \alpha. \tag{3.12}$$

We assume that $\phi(z;\theta)$ is a monotonic function of z and in particular, that $\phi'(z;\theta) \neq 0$ at the true value when $\phi(z;\theta) = \alpha$.

Both the MLE and Bayes estimator have variances given, to first order, by the Fisher information matrix, so the variance of $\tilde{\phi}(z)$ is of the form $\phi_i(z;\theta)\phi_j(z;\theta)\kappa^{i,j}/n + o(1/n)$. We also assume the bias is $b(z;\theta)/n + o(1/n)$, where $b(z;\theta)$ is derived from (3.2) or (3.6). Then the leading term of the coverage probability bias of $\tilde{z}$ is

$$\frac{1}{n}\left\{\frac{\phi_i(z;\theta)\phi_j'(z;\theta)\kappa^{i,j}}{\phi'(z;\theta)} - b(z;\theta)\right\}. \tag{3.13}$$

A derivation of (3.13) is given in Appendix E. This formula is equivalent to the main formula on p. 49 of Cox (1975); the present version appears to be more convenient for numerical evaluation.

To illustrate how these formulae may be evaluated, let us return to the exponential example of Section 2. In this example all quantities are one-dimensional so we revert to the usual notation whereby z^2 or θ^3 are powers of a scalar rather than components of a vector.

We have $\phi = e^{-\theta z}$, $Q = (\alpha - 1)\log\theta + (rz - \beta)\theta$ and $\ell_n(\theta) = n\log\theta - \theta S_n$, so with Z_1 denoting an asymptotically $N(0, 1/\theta^2))$ random variable, we have

$$\frac{\partial \ell_n}{\partial \theta} = \frac{n}{\theta} - S_n = n^{1/2} Z_1, \qquad \frac{\partial^2 \ell_n}{\partial \theta^2} = -\frac{n}{\theta^2}, \qquad \frac{\partial^3 \ell_n}{\partial \theta^3} = \frac{2n}{\theta^3},$$

and hence $\kappa_{11} = -1/\theta^2$, $\kappa^{1,1} = \theta^2$, $\kappa_{111} = 2/\theta^2$, etc. Also, all terms of the form $\kappa_{1,11}$, $\kappa_{1,111}$ and so on are 0. Referring to (3.4), we have $\mathcal{C} = 2\theta\phi_1 + \theta^2(\phi_{11} + 2\phi_1 Q_1)$ which turns out after a little manipulation to be the same as $2A\theta z e^{-\theta z} + 2B\theta^2 z^2 e^{-\theta z}$, where $A = -\alpha$ and $B = \frac{1}{2} - r + \frac{\beta}{z}$ as in Section 2.

It then follows that (3.5) is the same as

$$\mathcal{A} = \frac{\mathcal{C}^2}{4} + \theta^2 \mathcal{C}_1 \phi_1 + \theta \mathcal{C} \phi_1 + \frac{\theta^2 \mathcal{C} \phi_{11}}{2}$$

which evaluates to

$$\mathcal{A} = \theta^2 z^2 e^{-2\theta z}\{A^2 - 4A + (3A - 6B + 2AB)\theta z + (B^2 + 3B)\theta^2 z^2\}$$

and this in turn is the same as

$$\theta^2 z^2 e^{-2\theta z}\{q(A, B, z) - q(0, 0, z)\}$$

in the notation of (2.5)–(2.6).

A similar correspondence can be shown between (3.9) and (2.11)–(2.12). The corresponding calculation of (3.13) leads to a coverage probability bias which is asymptotically

$$\frac{1}{n}\left\{\alpha\theta z e^{-\theta z} + \left(r - \frac{\beta}{z}\right)\theta^2 z^2 e^{-\theta z}\right\}.$$

If $\alpha = \beta = r = 0$ then this expression is 0. Of course, in this case we know that the coverage probability bias is exactly 0 for all n, not just in the asymptotic sense considered here.

4 PROBABILITIES OF RARE EVENTS

Despite the complexity of the formulae in Section 3, it turns out that for a wide class of models, it is possible to pick out certain dominant terms as z tends to a limit, and so to obtain asymptotic results for the optimal tail weight function in the same fashion as was done in Section 2 for the exponential distribution. In this section the main assumptions and results will be outlined; technical details are deferred to Appendix F.

Suppose $\phi(z;\theta)$ is the tail distribution function of a random variable Z, and we are interested in its limit as $z \uparrow z_\omega$ where z_ω is the (finite or infinite) right-hand endpoint of the distribution of Z.

We consider functions $\phi(z;\theta)$ for which there exists a function $f(z;\theta)$, a function $\psi(\theta)$ (not depending on z) and constants $G,\ G_1$ such that $|f(z;\theta)| \to \infty$ as $z \uparrow z_\omega$ and

$$\begin{aligned} \phi_i(z;\theta) &\sim f(z;\theta)\psi_i(\theta)\phi(z;\theta), \\ \phi_{ij}(z;\theta) &\sim G\ f^2(z;\theta)\psi_i(\theta)\psi_j(\theta)\phi(z;\theta), \\ \phi_{ijk}(z;\theta) &\sim G_1\ f^3(z;\theta)\psi_i(\theta)\psi_j(\theta)\psi_k(\theta)\phi(z;\theta). \end{aligned} \tag{4.1}$$

In many cases we will find $G = G_1 = 1$, but we do not assume that at the outset.

The simplest case of a distribution satisfying (4.1) is when $\phi(z;\theta) \sim \exp\{-z\psi(\theta)\}$ as $z \to \infty$, for some scalar function $\psi(\theta)$. This is a direct generalisation of the exponential distribution. However (4.1) also applies to many other classes of distributions, including normal distributions with unknown mean and variance either known or unknown, and extreme value distributions. However, there are exceptions; for instance, the mirror image of an exponential distribution with known endpoint does not satisfy (4.1), and this is relevant for the example considered later in Section 7. Further details of the examples are in Appendix F.

For distributions satisfying (4.1), and a tail-weighted predictive distribution with weight $w(\phi) = \phi^r$, one can derive the optimal r as $z \to z_\omega$, for a variety of criteria. The results are summarised in the following table:

Criterion	Optimal r
Mean squared error	$3G - 1$
Logarithmic Loss	$3G - 2$
Squared Logarithmic Loss	$3G - \frac{5}{2}$
Bias (3.6)	G
CPB (3.13)	0

When $G = 1$, the most common case, these results are exactly the same as for the exponential distribution. The conclusion is that the results of Section 2 are very generally applicable; for a variety of criteria, the usual Bayes estimator with $r = 0$ performs worse than the plug-in approach, but the result is reversed with suitable choice of the parameter r.

5 POISSON-GPD MODEL

As a more complicated example of how the formulae in Sections 3 and 4 may be evaluated, let us return to the model used for the insurance data in Section 1. In this model, the number N of exceedances of the threshold u within a period of n years has a Poisson distribution with mean λn, and conditionally on N, the excesses $X_1, ..., X_N$ are i.i.d. with generalised Pareto distribution (1.2). The log likelihood is

$$\ell_n = N \log \lambda - \lambda n - N \log \psi - \left(1 + \frac{1}{\xi}\right) \sum \log \left(1 + \frac{\xi X_i}{\psi}\right). \tag{5.1}$$

Identifying (λ, ψ, ξ) with $(\theta^1, \theta^2, \theta^3)$, we calculate all derivatives of (5.1) up to third-order in $(\theta^1, \theta^2, \theta^3)$, and then evaluate their means and covariances using the formulae

$$\begin{aligned} \mathrm{E}\left\{\left(1 + \frac{\xi X_i}{\psi}\right)^r\right\} &= \frac{1}{1 - r\xi}, \\ \mathrm{E}\left\{\left(1 + \frac{\xi X_i}{\psi}\right)^r \log\left(1 + \frac{\xi X_i}{\psi}\right)\right\} &= \frac{\xi}{(1 - r\xi)^2}, \\ \mathrm{E}\left\{\left(1 + \frac{\xi X_i}{\psi}\right)^r \log^2\left(1 + \frac{\xi X_i}{\psi}\right)\right\} &= \frac{2\xi^2}{(1 - r\xi)^3}. \end{aligned} \tag{5.2}$$

The resulting cumulants are then collected together to evaluate $\mathcal{A}$, $\mathcal{A}_L$ for the logarithmic loss function, and the CPB as in (3.5), (3.9), (3.6) and (3.13).

These formulae have been computed using the computer algebra language MAPLE, but the resulting expressions are extremely unwieldy. In one calculation, $\mathcal{A}$ was translated into a Fortran programme, but the resulting code occupied over 1,500 lines. A more computationally efficient approach is to substitute numerical values for unknown parameters, within the computer algebra programme, before simplifying the expressions. This approach has been taken for the numerical results which follow.

As an example, suppose we set $\lambda = 5$, $\psi = 1$, $\xi = 0.5$, $\phi(z; \theta) = 1 - \exp[-\lambda\{1 + \xi(z-u)/\psi\}^{-1/\xi}]$ which corresponds to the tail distribution function of the annual maximum (equation (9.1) of Davison and Smith, 1990). A series of values of z is taken, corresponding to different predetermined values of ϕ between 10^{-1} and 10^{-9}. For Q, we consider

$$Q = -\log \lambda - r \log \phi \tag{5.3}$$

corresponding to the improper prior density λ^{-1}, but including the tail weight ϕ^r. To aid numerical stability, both $\mathcal{A}$ and $\mathcal{A}_L$ are normalised by dividing by $\kappa^{i,j}\phi_i\phi_j$ (the asymptotic variance of the MLE). In the case of CPB, the result is expressed as a ratio to that of the MLE — this does not create any ambiguity about signs because in our calculations the MLE always has a positive CPB.

In Fig. 5.1(a), it can be seen that from the point of view of MSE, $r = 0$ is the worst of the four estimators when $\phi < 10^{-4}$, with $r = 2$ eventually best for very small ϕ, but

$r = 1$ best across most of the range depicted. For logarithmic loss, in Fig. 5.1(b), $r = 0$ is best across most of the range but $r = 1$ is better than $r = 0$ at $\phi = 10^{-9}$. We would need an even smaller ϕ to show that that $r = 1$ is better than the MLE. Finally looking at Fig. 5.1(c) for CPB, we see that as $n \to \infty$, the Bayes estimator with $r = 0$ has a smaller CPB than the MLE for sufficiently small ϕ, though this does not appear to be the case in the middle range of ϕ.

We now consider to what extent these results can be confirmed by simulations of the actual models used for the examples of Section 1. It would be unrealistic to expect the theoretical results to be numerically accurate for sample sizes of the order of these two examples, and we indeed find that they are not accurate. Simulations for the insurance example fixed $\xi = -.89$. Simulations of the records example followed the same model as that of Robinson and Tawn (1995) and Smith (1997a), which though not the same as the Poisson-GPD model has a lot of features in common with it. In particular, the same parameter ξ arises in this model, and for the simulations the value $\xi = -0.2$ was fixed, rather arbitrarily. For regular maximum likelihood estimation, the Fisher information matrix was calculated by Tawn (1988) and requires $\xi > -0.5$. For the current problem, the existence of all the quantities defined in Section 3 requires $\xi > -0.25$, but even this "regularity condition" ignores certain features of the problem such as $\phi = 0$ with positive posterior probability. For this reason, none of the results are strictly correct for the records example, but it is nevertheless of interest to see to what extent qualitative features of the theoretical results are reflected in simulations.

It should also be pointed out that the simulations themselves are far from easy to conduct, since the original data analyses involved extensive MCMC sampling and it is not possible to do this to the same extent in a simulation. The results are based on 1,000 replications and, as a protection against occasional wild results, the largest and smallest 2.5% of the replications are deleted.

Fig. 5.2 plots the mean square error and the logarithmic risk of the Bayes estimator for the insurance model, and the mean squared error for the records model. In each case the estimate is calculated for a sequence of values of ϕ and the quantity

$$\frac{R_B(r, \phi)}{R_M(\phi)} - 1$$

is plotted, where R_B denotes the risk of the Bayes estimator, indexed by r as well as ϕ, and R_M is the risk of the MLE. No calculation is made for the logarithmic risk in the records model because in this case the MLE is 0 for many simulations, in which case the logarithmic risk is infinite, and no meaningful comparison is possible.

In spite of all these difficulties with both the asymptotic and simulated results, the two have a number of qualitative features in common, and from this point of view the asymptotics may be said to provide some practical guidance to the choice of procedure. For both the insurance and records simulations, the usual estimator with $r = 0$ performs poorly from the point of view of mean squared error, but either $r = 1$ or $r = 2$ is much better. For logarithmic loss in the insurance simulation, $r = 0$ is best across most of the range of ϕ. This is also the case by default for the records model, since when $\phi = 0$ with positive posterior probability, it is automatically the case that $\hat{\phi}_{BAY} = 0$ for any $r > 0$, so in this case these estimators, like the MLE, have infinite logarithmic risk.

A final simulation was made to assess CPB. This was done for three values for the target α (.1, .01 and .001), and for six estimators (the MLE, and Bayes with five values of r). In each case, $\hat{z}$ was determined so that $\hat{\phi}(\hat{z}) = \alpha$ and $\mathrm{E}\{\phi(\hat{z})\}$ was estimated by simulation. Fig. 5.3(a) shows that for the insurance model, the CPB is positive for the MLE, for each of the three values of α. The best agreement between the true and nominal coverage probability is obtained using the Bayes estimator with $r = 0$. In the case of the records simulation (Fig. 5.3(b)), the same conclusions may be drawn for $\alpha = .1$, but the results for $\alpha = .01$ and .001 are poor for all the estimators, with perhaps still a rather weak case for saying that $r = 0$ is best. From these simulations we may be led to conclude that the Bayes estimator with $r = 0$ is superior to the MLE in general, but the asymptotic results shown in Fig. 5.2(c) do not point towards such a clear-cut conclusion. In this case, the asymptotic theory serves to act as a warning against too hastily drawing general conclusions from limited simulations.

6 HIERARCHICAL MODELS: THE NORMAL MEANS PROBLEM

The results so far in this paper have implicity assumed a single homogeneous sample. However, they are also applicable in multi-sample problems, in which the parameters of the model are possibly different from one sample to another. Such problems lead to what are usually called empirical Bayes methods of analysis. In recent years it has become more common to solve such problems from a fully Bayesian point of view, using a hierarchical model structure to link together the parameters of the different subsamples. This is the point of view taken, for example, in the excellent recent monograph by Carlin and Louis (1996).

Despite the very rapid growth of this field, there has been comparatively little study of the frequentist properties of Bayesian procedures in this setting. Berger and Strawderman (1996) established some admissibility results, which have the advantage of not relying on any kind of asymptotics, and which provide guidance on the choice of prior particularly where improper priors are concerned. On the other hand, the class of models to which their results apply is restrictive, and admissibility results do not necessarily help to pick out a prior distribution which has good properties under particular conditions. In contrast, the results of the present paper are asymptotic (letting sample size $n \to \infty$ while the number of samples remains fixed) but they do allow explicit computations to be made under a veriety of circumstances.

In the present section, these ideas are worked out in some detail for the simplest problem in this class: the case of p normal distributions with unknown means and known common variance. In the next section, a more complicated example is considered.

Suppose there are p subgroups and the data in the j'th subgroup follow a $N(\theta_j, 1)$ distribution. Here the vector of subgroup means $\theta = (\theta_1, ..., \theta_p)$ are the parameters of interest. The remarkable result of James and Stein (1961) showed that from the point of view of squared error loss, $L(\theta, \hat{\theta}) = \sum_{j=1}^{p}(\theta_j - \hat{\theta}_j)^2$, for $p \geq 3$ the vector of sample means (or MLE) is inadmissible as an estimator of θ, and they exhibited an estimator which dominates the MLE. This stimulated a rich stream of developments in decision theory. In recent years, especially since the papers of Morris (1983a, 1983b), practical attention has shifted away from the decision-theoretic viewpoint towards such issues as confidence intervals for the individual subgroup means. The present focus on predictive inference may be regarded as a logical continuation of this shift of emphasis, but it is also our objective to bring out the connections between this and the decision-theoretic approach.

Suppose we have n observations in each subgroup and suppose, initially, that we are interested in a particular subgroup mean, $\phi(\theta) = \theta_k$ for given $k \in \{1, ..., p\}$. In the notation of Section 3, we have $\kappa_{i,j} = 1$ if $i = j$, 0 if $i \neq j$, and hence also $\kappa^{i,j} = \kappa_{i,j}$ for all i and j. Moreover the higher order cumulants, $\kappa_{i,jk}$, κ_{ijk} and $\kappa_{i,jk\ell}$, are all 0. Then it is quickly established that for a Bayesian procedure with prior e^Q,

$$\mathcal{A} = Q_k^2 + 2Q_{kk}. \tag{6.1}$$

This immediately extends to the loss function $L(\theta, \hat{\theta}) = \sum_{j=1}^{p}(\theta_j - \hat{\theta}_j)^2$, for which

$$\mathcal{A} = \sum_{j=1}^{p}(Q_j^2 + 2Q_{jj}). \tag{6.2}$$

Note that we are no longer employing the summation convention.

As an example, consider the prior density

$$\pi(\theta) = e^{Q(\theta)} \propto \left(\sum \theta_j^2\right)^{-\alpha} \tag{6.3}$$

with $\alpha > 0$. Then (6.2) becomes

$$\mathcal{A} = \frac{4\{\alpha^2 - (p-2)\alpha\}}{\sum \theta_j^2}. \tag{6.4}$$

Now (6.4) is negative for all θ provided $p > 2$ and $0 < \alpha < p-2$, with an optimal $\alpha = \frac{p-2}{2}$. Thus in this case we obtain a very simple asymptotic version of the James-Stein theorem, whereby the Bayes estimator dominates the maximum likelihood estimator uniformly over all θ. However another conclusion from (6.4) is that the improvement is largest when $\sum \theta_j^2$ is very small.

In practice, empirical Bayes methods are often applied in situations where we believe that the θ_j's are roughly equal, but not necessarily roughly equal to 0. As an alternative to (6.3), such information may be represented in the prior density

$$\pi(\theta) \propto \left\{\sum(\theta_j - \bar{\theta})^2\right\}^{-\alpha} \tag{6.5}$$

where $\bar{\theta} = \frac{1}{p}\sum \theta_j$. In this case (6.2) reduces to

$$\mathcal{A} = \frac{4\{\alpha^2 - (p-3)\alpha\}}{\sum(\theta_j - \bar{\theta})^2}. \tag{6.6}$$

In contrast with (6.4), (6.6) imposes stricter conditions on p and α ($\mathcal{A} < 0$ requires $p > 3$ and $0 < \alpha < p - 3$, with best value $\alpha = \frac{p-3}{2}$), but we get a large improvement from using the Bayes estimator whenever $\sum(\theta_j - \bar{\theta})^2$ is small, which is a less stringent on θ than requiring $\sum \theta_j^2$ to be small.

Both (6.3) and (6.5) may be developed as limiting cases of the hierarchical model

$$\begin{aligned} \theta_1, ..., \theta_p | \beta, \tau &\sim N(\beta, \tau), \\ \beta | \tau &\sim N(m, c\tau), \\ \tau &\sim IG(a, b) \end{aligned} \tag{6.7}$$

where IG denotes the inverse gamma distribution with density $b^a \tau^{-a-1} e^{-b/\tau}/\Gamma(a)$ and m, c, a, b are fixed hyperparameters. In this case, direct integration of β and τ from the joint density of (θ, β, τ) shows that

$$\pi(\theta) \propto \left\{ \frac{1}{2} \sum (\theta_j - \bar{\theta})^2 + \frac{1}{2} \frac{p}{cp+1} (\bar{\theta} - m)^2 + b \right\}^{-a-p/2}. \tag{6.8}$$

Then (6.3) may be obtained from (6.8) by letting $m = 0,\ b \to 0,\ c \to 0$, and similarly (6.5) with $m = 0,\ b \to 0,\ c \to \infty$.

One advantage of rewriting (6.3) or (6.5) in terms of the hierarchical model (6.7) is that it allows the computations to be made using Gibbs sampling (Gelfand and Smith 1990). However (6.7) also suggests the way towards much broader classes of prior distributions based on different specifications of the marginal densities of β and τ (Berger and Strawderman 1996).

Now let us return to the problem of predictive inference. Suppose, for definiteness, we are interested in predictive inference for a specific subgroup, indexed by k. In accordance with our earlier point of view, this may also be regarded as a problem about estimating $\phi(z; \theta) = 1 - H(z - \theta_k)$ where H is the standard normal distribution function. Let $h = H'$. Suppose we do this for fixed $v = z - \theta_k$. Then after taking into account such properties as $h'(v) = -vh(v)$, $h''(v) = (v^2 - 1)h(v)$, we find that

$$\mathcal{A} = h^2(v) \left(\frac{7v^2}{4} - 1 + 4vQ_k + Q_k^2 + 2Q_{kk} \right). \tag{6.9}$$

As an example, consider $Q = -\alpha \log\{\sum(\theta_j - \bar{\theta})^2\} - r \log \phi(z; \theta_k)$ which corresponds to the prior (6.5) with tail weight $w(\phi) = \phi^r$. Then

$$\begin{aligned} \mathcal{A} = h^2(v) \Bigg[& \frac{7v^2}{4} - 1 - \frac{8\alpha v(\theta_k - \bar{\theta})}{\sum(\theta_j - \bar{\theta})^2} - \frac{6rvh(v)}{1 - H(v)} + \left\{ \frac{2\alpha(\theta_k - \bar{\theta})}{\sum(\theta_j - \bar{\theta})^2} + \frac{rh(v)}{1 - H(v)} \right\}^2 \\ & - \frac{4\alpha(p-1)}{p\sum(\theta_j - \bar{\theta})^2} + \frac{8\alpha(\theta_k - \bar{\theta})^2}{\{\sum(\theta_j - \bar{\theta})^2\}^2} + 2r \left(\frac{h(v)}{1 - H(v)} \right)^2 \Bigg]. \end{aligned} \tag{6.10}$$

Although our focus here is on predicting observations within a single subgroup, it is still natural to evaluate the results by averaging the risk over all p subgroups. Therefore we replace $\mathcal{A}$ by its average over all k, to give

$$\bar{\mathcal{A}} = h^2(v)\left[\frac{7v^2}{4} - 1 - \frac{6rvh(v)}{1-H(v)} + (r^2+2r)\left\{\frac{h(v)}{1-H(v)}\right\}^2 + \frac{4\alpha(\alpha-p+3)}{p\sum(\theta_j-\bar{\theta})^2}\right]. \qquad (6.11)$$

The last term in (6.11) is the same except for a multiplicative constant as (6.6), and is minimised when $\alpha = \frac{p-3}{2}$. This confirms that the influence of the prior density is essentially the same as in our earlier discussion when the sole objective was to estimate θ with squared error loss. The remaining terms in (6.11) represent the effect of prediction at a specific value of v. For large v, $h(v)/\{1-H(v)\} \sim v$ so

$$\bar{\mathcal{A}} \sim \left(\frac{7}{4} - 4r + r^2\right) v^2 h^2(v) \quad (v \to \infty) \qquad (6.12)$$

which is minimised when $r = 2$, as in several earlier examples.

By similarly averaging over all subgroups, using (3.9), it is possible to compute an averaged version of $\mathcal{A}_L$ for logarithmic loss in the form

$$\bar{\mathcal{A}}_L = \frac{\bar{\mathcal{A}}}{2H(v)\{1-H(v)\}} - \frac{\{2H(v)-1\}h^3(v)}{2H(v)^2\{1-H(v)\}^2}\left\{v - 2r\frac{h(v)}{1-H(v)}\right\}. \qquad (6.13)$$

The second term in (6.13) does not depend on α, so once again we are led to the optimal value $\alpha = \frac{p-3}{2}$, as is the case for squared error loss.

Fig. 6.1 shows normalised versions of $\mathcal{A}$ and $\mathcal{A}_L$, for $r = 0$, 1 and 2, ignoring the contribution of the prior density (or equivalently, setting $\alpha = 0$ in (6.11)). The plots show that many of the characteristics of earlier plots, such as Figs. 2.2 and 2.4, also hold for the normal distribution.

In the case of CPB, if we are interested in the k'th subgroup then it follows from (3.6) and (3.13) that the resulting asymptotic bias is given by

$$b_k^*(z_k) = h(v)\left\{\frac{2\alpha(\theta_k-\bar{\theta})}{\sum(\theta_j-\bar{\theta})^2} + \frac{rh(v)}{1-H(v)}\right\} \qquad (6.14)$$

where again $v = z - \theta_k$.

In this case it is less logical to average over all k, since we would clearly be concerned about the situation if even one subgroup had a large CPB, and in any case (6.14) is easy enough to interpret on its own. Clearly if our only objective is to minimise CPB then we should not do any kind of Bayesian analysis, since the standard single-population frequentist analysis gives an exactly unbiased prediction interval. On the other hand, if α and r are chosen to improve the performance of the prediction interval as measured by $\bar{\mathcal{A}}$ or $\bar{\mathcal{A}}_L$, then (6.14) provides a measure of the penalty incurred in terms of the CPB.

In the case of confidence rather than prediction intervals, the possibility of reducing CPB while maintaining the desirable shrinkage properties of empirical Bayes procedures has in the past been considered from a bootstrap perspective (Laird and Louis 1987, Carlin and Gelfand 1990, 1991). The present discussion shows how these contrasting criteria may be assessed when the procedures considered are fully Bayesian.

7 POISSON-GAMMA MODEL

The final example in this paper is a model for exchangeable Poisson processes developed by Gaver and O'Muircheartaigh (1987) and extended to a hierarchical framework by Gelfand and Smith (1990), who described a Gibbs sampling framework for carrying out the Bayesian calculations. 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

$$\mathrm{E}\{X_j\} = nt_j\theta_j. \tag{7.1}$$

Compared with Gelfand and Smith, the only change is the inclusion of a parameter n, to enable asymptotic calculations as $n \to \infty$.

Gelfand and Smith proposed a hierarchical prior structure based on gamma distributions, with

$$\begin{aligned} \pi(\theta|\beta) &= \prod_{j=1}^{p} \frac{\theta_j^{\alpha-1} e^{-\theta_j/\beta}}{\beta^\alpha \Gamma(\alpha)}, \\ \pi(\beta) &= \frac{\delta^\gamma e^{-\delta/\beta}}{\beta^{\gamma+1}\Gamma(\gamma)}, \end{aligned} \tag{7.2}$$

in terms of hyperparameters α, γ and δ. By combining (7.2) into a single joint prior density and integrating out β, we have

$$\pi(\theta) = \frac{\delta^\gamma \Gamma(p\alpha+\gamma)}{\Gamma^p(\alpha)\Gamma(\gamma)} \cdot \frac{(\prod \theta_j)^{\alpha-1}}{(\sum \theta_j + \delta)^{p\alpha+\gamma}}. \tag{7.3}$$

and hence $Q = \log \pi$. This example does not fall within the framework of Section 4 and there seems no advantage in considering a tail-weighted estimator; therefore, we do not consider any r parameter in this model.

As a predictand for this problem, there are clearly many possibilities but one which seems natural in the context of the problem is the time Z to first failure by any of the machines, with particular concern if this first failure time is very small. Thus we consider the distribution function of Z,

$$\phi(z;\theta) = 1 - e^{-z\sum_1^p \theta_j} \tag{7.4}$$

with particular focus on the limit $z \to 0$.

It is then possible to derive expressions for $\mathcal{A}$, $\mathcal{A}_L$ and the asymptotic coverage probability bias using the formulae from Section 3; details of the calculation are in Appendix G.

For this model and choice of ϕ, a simulation exercise has been conducted with the parameters $n = p = 5,\ t_j = 1,\ \theta_j = j$. All simulations are based on 10,000 replications and used 250 iterations of a Gibbs sampler with the first 50 iterations discarded as warm-up iterations. The observed $\mathcal{A}$ or $\mathcal{A}_L$ is defined as n^2 times the difference in either squared error or logarithmic risk for the Bayes and ML estimators. This may then be compared with the theoretical value derived from the asymptotic theory. To assess the simulation-induced error, the calculations of observed risk are combined with 95% confidence bounds obtained by bootstrap resampling from the 10,000 replications. This does not take into account the possibility that 250 Gibbs iterations might be inadequate, but this particular model has been well studied from the point of view of convergence of the Gibbs sampler, and this number of iterations appears to be quite adequate.

For the specification of α, γ and δ, Gelfand and Smith (1990) took the "vague prior" values $\gamma = 0.1$, $\delta = 1$, and fixed α by an empirical Bayes argument designed to match the estimated standard deviation of $\theta_1, ..., \theta_p$. They were at pains to emphasise that this choice of hyperparameters was for illustrative purposes only and they did not claim that these represent "optimal" choices in any sense whatsoever. Nevertheless it will be argued that their specification of γ was misguided and that better results may be obtained by suitable intelligent specifications of both α and γ.

If $\theta_j = j,\ j = 1, .., 5$ are considered as five observations from the gamma distribution (7.2), then the method of moments estimators are $\alpha = 3.6$, $\beta = 0.8333$; the maximum likelihood estimators are only slightly different from those. Moreover the posterior mean of $1/\beta$ in (7.2) is γ/δ. Motivated by these considerations we consider α in the range 1–9 and a selection of choices for (γ, δ) satisfying $\gamma = \delta$.

Fig. 7.1 shows the theoretical $\mathcal{A}$ and $\mathcal{A}_L$ functions, as a function of α for fixed $\gamma = 5$, and as a function of γ for fixed $\alpha = 5$. The plots are shown for two values of z corresponding to $\phi = 0.1$ and 0.001; there is very little difference between the two values of z. Since the objective is to choose hyperpriors for which $\mathcal{A}$ or $\mathcal{A}_L$ is small, the plots suggest strongly that one should not make the "vague prior" choices, in which α or γ are set close to 0, but that it is worth seeking out optimal values, of which the choice $\alpha = \gamma = 5$ is a reasonable rough guess.

In Fig. 7.2 this is explored further, with simulated and theoretical values of $\mathcal{A}$, $\mathcal{A}_L$ and the CPB for a variety of priors corresponding to different choice of α and γ. In the case of CPB, the agreement between theoretical and simulated values ((a) and (b)) is remarkably good. For the remaining plots, the overall shape of the simulated curves (solid lines) is the same as that of the theoretical curves (dotted lines). Nevertheless, the agreement is still not ideal. In contrast with the example of Section 5, which is messy from every point of view, this is one for which the asymptotics ought to be reasonably well behaved, so it is worth exploring the reasons for the discrepancy.

Some investigation has been made into the source of the approximation error, from which it is clear that no single part of the calculation is primarily responsible for the error. It is possible that the calculation could be improved by formal calculation of the next term in the asymptotic expansion, but this would involve considerable additional analytical work and has not been attempted. An intermediate step has been attempted in which the sequence of approximations given in the Appendices by equations (B.3), (C.4) and (C.5) are taken as the starting point, and (C.6) evaluated exactly from that point on. The modification from $\mathcal{A}$ to $\mathcal{A}_L$ again uses (3.9), without any additional correction terms. This produces the dashed curves in Fig. 7.2(c–j). The modification substantially improves the agreement between the theoretical and simulated results. The moral of this story may be that for specific models of interest, it is possible to improve the approximations by detailed consideration of individual error terms, but it looks to be a formidable task to derive general-purpose formulae which are better than the ones given in this paper.

As far as the optimal choices of α and γ are concerned, the simulation results suggest values of α and γ that are somewhat smaller than those suggested by the asymptotic theory, but they confirm the general convex shape of the plots and in particular the message that middle-range values of α and γ, in the range 3–5 say, are superior from those at either extreme.

Of course, this whole comparison is artificial in the sense that the values of θ_j are known, whereas in a real data problem the specification of hyperparameters would have to be dealt with simultaneously as the estimation of the θ_j's. Nevertheless our results suggest that some strategy based on intelligent selection of the hyperparameters may be superior to the vague prior approach. This recommendation appears to be at variance with what is currently preferred in the literature on hierarchical models, see e.g. Christiansen and Morris (1997).

8 CONCLUSIONS

The main purpose of this paper has been to argue that theoretical comparisons of Bayesian predictive procedures are both possible and worthwhile. Comparisons may be decision-theoretic using one of a variety of loss functions, or based on the coverage probability bias of prediction intervals. The asymptotic calculations do not always yield accurate numerical results, but nevertheless provide practical guidance in the choice of statistical procedures. Whenever possible, the results of such an analysis should be supplemented by simulations, but in complicated problems requiring extensive Monte Carlo sampling to compute the Bayes estimates, an accurate simulation study is in itself very difficult to conduct.

For the estimation of tail probabilities, the results suggest that substantial gains may be made by introducing a tail weighting function in the calculation of the Bayesian predictive distribution. However the tail weight itself depends on the criterion used to assess the estimator, so practical implementation of these ideas would require more attention to the loss function than is currently fashionable in either the frequentist or Bayesian literatures. The superiority of the plug-in approach under some circumstances is not seen as an argument in favour of that approach, but rather as a warning against the automatic assumption that Bayes estimates must be superior.

For hierarchical models, the results provide a tool for comparing different specifications of the hyperprior. In the normal means problem these may be considered an asymptotic version of well-known formulae dating back to James and Stein, but the present approach allows corresponding formulae to be developed in the context of predictive inference and for more complicated models such as the Poisson-Gamma model discussed in Section 7.

The limited examples that have been discussed suggest that intelligent specifications of hyperprior parameters may be superior to the vague prior approach. One possible direction for future work is to consider more formally the consequences of data-based priors.

The main drawback of the methodology is the reliance on asymptotic approximations. In certain cases it may be possible to improve on these by *ad hoc* arguments, but as far as general results are concerned, the theory needs to be viewed with the same qualifications as are present with any asymptotic theory.

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. (1994), *Inference and Asymptotics.* Chapman and Hall, London.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis* (second edition). Springer, New York.

Berger, J.O. and Bernardo, J.M. (1992), On the development of reference priors (with discussion). In *Bayesian Statistics 4*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, pp. 35–60.

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

Bleistein, N. and Handelsman, R.A. (1986), *Asymptotic Expansion of Integrals.* Second edition, Dover, New York.

Bollobás, B. (1986) (editor), *Littlewood's Miscellany.* Cambridge University Press, Cambridge.

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

Carlin, B.P. and Gelfand, A.E. (1990), Approaches for empirical Bayes confidence intervals. *J. Amer. Statist. Assoc.* **85**, 105–114.

Carlin, B.P. and Gelfand, A.E. (1991), A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J.R. Statist. Soc. B* **53**, 189–200.

Carlin, B.P. and Louis, T.A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman and Hall.

Christiansen, C.L. and Morris, C.N. (1997), Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.* **92**, 618–632.

Coles, S.G. and Powell, E.A. (1996), Bayesian methods in extreme value modelling: a review and new developments. *Internat. Statist. Rev.* **64**, 119–136.

Coles, S.G. and Tawn, J.A. (1996), A Bayesian analysis of extreme rainfall data. *Applied Statistics* **45**, 463–478.

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.

Datta, G.S. (1996), On priors providing frequentist validity of Bayesian inference for multipe parametric functions. *Biometrika* **83**, 287–298.

Datta, G.S. and Ghosh, J.K. (1995), On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 37–46.

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.

Efron, B. (1993), Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Efron, B. (1996), Empirical Bayes methods for combining likelihood (with discussion). *J. Amer. Statist. Assoc* **91**, 538–565.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, Vol. I. (3rd ed.). Wiley, New York.

Fisher, R.A. (1934), Two new properties of mathematical likelihood. *Proc. R. Soc. London A* **144**, 285–307.

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.

Ghosh, J.K. and Mukerjee, R. (1991), Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multiparameter case. *J. Mult. Anal.* **38**, 385–393.

Ghosh, J.K. and Mukerjee, R. (1992), Non-informative priors. In *Bayesian Statistics 4*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, pp. 195–210.

Ghosh, J.K. and Mukerjee, R. (1993), On priors that match posterior and frequentist distribution functions. *Canadian J. Statistics* **21**, 89–96.

James, W. and Stein, C. (1961), Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium in Mathmatical 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.

Laird, N. and Louis, T. (1987), Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Amer. Statist. Assoc.* **82**, 741–757.

Liseo, B. (1993), Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

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

Morris, C.N. (1983a), Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78**, 47–65.

Morris, C.N. (1983b), Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis and Robustness.* New York, Academic Press, 25–50.

Mukerjee, R. and Dey, D.K. (1993), Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher-order asymptotics. *Biometrika* **80**, 499–506.

Nicolau, A. (1993), Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J.R. Statist. Soc. B* **55**, 377–390.

Peers, H.W. (1965), On confidence points and Bayesian probability points in the case of several parameters. *J.R. Statist. Soc. B* **27**, 9–16.

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.

Robinson, M.E. and Tawn, J.A. (1995), Statistics for exceptional athletics records. *Applied Statistics* **44**, 499–511.

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. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science* **4**, 367-393.

Smith, R.L. (1997a), Statistics for exceptional athletics records: Letter to the editor. *Applied Statistics* **46**, 123–127.

Smith, R.L. (1997b), Extreme value analysis of insurance risk. Submitted for publication.

Stein, C. (1985), On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics.* Banach Center Publications Vol. 16.

PWN-Polish Scientific Publishers, Warsaw, pp. 485–514.

Sweeting, T.J. (1996), Approximate Bayesian computation based on signed roots of log-density ratios (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. 427–444.

Tawn, J.A. (1988), An extreme-value theory model for dependent observations. *J. Hydrology* **101**, 227–250.

Tibshirani, T. (1989), Non-informative priors for one parameter of many. *Biometrika* **76**, 604–608.

Welch, B.L. (1965), On comparisons between confidence point procedures in the case of a single parameter. *J.R. Statist. Soc. B* **27**, 1–8.

Welch, B.L. and Peers, H.W. (1963), On formulae for confidence points based on integrals of weighted likelihoods. *J.R. Statist. Soc. B* **25**, 318–329.

Appendix A: Derivation of (2.5) and (2.11)

First we need asymptotic expressions for $\mathrm{E}\{(\hat\theta-\theta)^t\}$ for $t=1,2,3,4$; we must retain terms as far as $O(n^{-2})$. Noting that $\hat\theta = n/S_n$, from elementary properties of the gamma distribution we have

$$\mathrm{E}\left\{\left(\frac{\hat\theta}{\theta}\right)^t\right\} = \frac{n^t}{(n-1)(n-2)...(n-t)}$$

for integer $t<n$. We quickly deduce

$$\mathrm{E}\left\{\left(\frac{\hat\theta-\theta}{\theta}\right)^t\right\} = \begin{cases} n^{-1}+n^{-2}+O(n^{-3}), & t=1,\\ n^{-1}+5n^{-2}+O(n^{-3}), & t=2,\\ 7n^{-2}+O(n^{-3}), & t=3,\\ 3n^{-2}+O(n^{-3}), & t=4,\\ O(n^{-3}), & t>4. \end{cases}$$

To derive (2.5), we expand $(\phi-\hat\phi)^2$ as far as the fourth power in $\hat\theta-\theta$, taking expected values, and collecting terms. The same method is used for (2.11) based on Taylor expansion of $L(\phi,\hat\phi)$ given by (2.10). Computer algebra has been used to assist in the manipulations involved.

Appendix B: Derivation of (3.2)

Equation (7.10), page 209, of McCullagh (1987), may be written in the form

$$\hat\theta^i-\theta^i = n^{-1/2}\kappa^{i,j}Z_j + n^{-1}(\kappa^{i,j}\kappa^{k,\ell}Z_{jk}Z_\ell + \frac{1}{2}\kappa^{i,s}\kappa^{j,t}\kappa^{k,u}\kappa_{stu}Z_jZ_k) + O_p(n^{-3/2}).$$

From this we deduce that

$$\mathrm{E}\{\hat\theta^i-\theta^i\} = n^{-1}(\kappa^{i,j}\kappa^{k,\ell}\kappa_{jk,\ell} + \frac{1}{2}\kappa^{i,j}\kappa^{k,\ell}\kappa_{jk\ell}) + O(n^{-3/2}) \qquad \text{(B.1)}$$

and

$$\mathrm{E}\{(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j)\} = n^{-1}\kappa^{i,j} + O(n^{-3/2}). \tag{B.2}$$

We also have

$$\begin{aligned}
\hat{\phi} - \phi &= \phi_i(\hat{\theta}^i - \theta^i) + \frac{1}{2}\phi_{ij}(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j) + O_p(n^{-3/2}) \\
&= n^{-1/2}\kappa^{i,j}Z_j\phi_i + n^{-1}(\kappa^{i,j}\kappa^{k,\ell}Z_{jk}Z_\ell + \frac{1}{2}\kappa^{i,s}\kappa^{j,t}\kappa^{k,u}\kappa_{stu}Z_jZ_k)\phi_i \\
&\quad + \frac{1}{2}n^{-1}\kappa^{i,k}\kappa^{j,\ell}Z_kZ_\ell\phi_{ij} + O_p(n^{-3/2}).
\end{aligned} \tag{B.3}$$

Combining (B.1), (B.2) and the first equality in (B.3), we deduce (3.2).

Appendix C: Derivation of (3.3)–(3.5)

We need some additional notation. As in McCullagh (1987), let $U_i = \partial\ell_n(\theta)/\partial\theta_i$, $U_{ij} = \partial\ell_n(\theta)/\partial\theta_i\partial\theta_j$, etc. Also $U_i = n^{1/2}Z_i$, $\quad U_{ij} = n\kappa_{ij} + n^{1/2}Z_{ij}$, and so on, where $Z_i, Z_{ij}, ...$ are $O_p(1)$ random variables with mean 0. We write $\{U^{ij}\}$ for the matrix inverse of $\{U_{ij}\}$. Any expression with a hat on it, such as $\hat{U}_{ijk}$, means that it is to be evaluated at the MLE $\hat{\theta}$ rather than the true value θ. Define

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

and let $\hat{\mathcal{D}}$ denote the same expression where all terms have hats. With these conventions, two applications of Laplace's integral expansion yield the formula

$$\hat{\phi}_{BAY} = \hat{\phi} + \hat{\mathcal{D}}. \tag{C.2}$$

To derive (C.2), consider an integral of the form

$$I_n(g, \psi) = \int_D g(\theta)e^{n\psi(\theta)}d\theta,$$

where $D \subseteq \mathcal{R}^p$ is a bounded simply-connected domain, the function ψ has a unique maximum attained at an interior point θ_0 of D, and both ψ and g are sufficiently differentiable in a neighborhood of θ_0. Then by equations (8.3.50–8.3.55) of Bleistein and Handelsman (1986),

$$\begin{aligned}
I_n(g, \psi) = \left(\frac{2\pi}{n}\right)^{p/2} e^{n\psi(\theta_0)}|\det\ B|^{1/2}\Bigg[g - \frac{1}{2n}\Bigg\{&-g_i\psi_{jk\ell}b^{ij}b^{k\ell} + g_{ij}b^{ij} \\
&+ g\psi_{ijk}\psi_{stu}\left(\frac{1}{4}b^{is}b^{jk}b^{tu} + \frac{1}{6}b^{is}b^{jt}b^{ku}\right) - \frac{1}{4}g\psi_{ijk\ell}b^{ij}b^{k\ell}\Bigg\} + O\left(\frac{1}{n^2}\right)\Bigg].
\end{aligned} \tag{C.3}$$

Here suffixes denote differentiation with respect to the components of θ, all functions are evaluated at θ_0, and the matrix B with entries $\{b^{ij}\}$ is the inverse of the matrix whose entries are $\{\psi_{ij}\}$.

Let $\psi(\theta) = \frac{1}{n}\ell_n(\theta)$, which by assumption is of $O_p(1)$. The formula (C.2) follows by applying (C.3) twice, once with $g(\theta) = \phi(\theta)e^{Q(\theta)}$ and a second time with $g(\theta) = e^{Q(\theta)}$, and rearranging terms.

The next step of the argument takes (C.2) as its starting point. If matrices A, B and C, with entries $\{a_{ij}\}$, $\{b_{ij}\}$, $\{c_{ij}\}$, are related by

$$a_{ij} = b_{ij} + \epsilon c_{ij} + O(\epsilon^2)$$

then the inverses of A and B, with entries $\{a^{ij}\}$, $\{b^{ij}\}$, are related by

$$a^{ij} = b^{ij} - \epsilon b^{ik} b^{j\ell} c_{k\ell} + O(\epsilon^2).$$

Applying this to

$$U_{ij} = -n\kappa_{i,j} + n^{1/2} Z_{ij},$$

we deduce

$$U^{ij} = -n^{-1}\kappa^{i,j} - n^{-3/2}\kappa^{i,s}\kappa^{j,t} Z_{st} + O_p(n^{-2}).$$

Hence

$$\begin{aligned}
\mathcal{D} &= \frac{1}{2}(n\kappa_{ijk} + n^{1/2} Z_{ijk})(n^{-1}\kappa^{i,k} + n^{-3/2}\kappa^{i,s}\kappa^{k,t} Z_{st})(n^{-1}\kappa^{j,\ell} + n^{-3/2}\kappa^{j,u}\kappa^{\ell,v} Z_{uv})\phi_\ell \\
&\quad + \frac{1}{2}(\phi_{ij} + 2\phi_i Q_j)(n^{-1}\kappa^{i,j} + n^{-3/2}\kappa^{i,s}\kappa^{j,t} Z_{st}) \\
&= \frac{1}{2n}\left\{\kappa_{ijk}\kappa^{i,k}\kappa^{j,\ell}\phi_\ell + \kappa^{i,j}(\phi_{ij} + 2\phi_i Q_j)\right\} \\
&\quad + \frac{1}{2n^{3/2}}\Big\{\kappa^{i,k}\kappa^{j,\ell} Z_{ijk}\phi_\ell + \kappa_{ijk}\kappa^{i,s}\kappa^{k,t}\kappa^{j,\ell} Z_{st}\phi_\ell + \kappa_{ijk}\kappa^{i,k}\kappa^{j,s}\kappa^{\ell,t} Z_{st}\phi_\ell \\
&\quad + \kappa^{i,s}\kappa^{j,t} Z_{st}(\phi_{ij} + 2\phi_i Q_j)\Big\} + O_p\left(\frac{1}{n^2}\right).
\end{aligned} \tag{C.4}$$

The next step is to pass from $\mathcal{D}$ to $\hat{\mathcal{D}}$. If we write $\mathcal{D} = \frac{1}{2}\mathcal{C}n^{-1} + \frac{1}{2}\mathcal{E}n^{-3/2} + O_p(n^{-2})$, where $\mathcal{C}$ is the same as in (3.4), it follows that

$$\hat{\mathcal{D}} - \mathcal{D} = (\frac{1}{2}\mathcal{C}_s n^{-1} + \frac{1}{2}\mathcal{E}_s n^{-3/2})(\hat{\theta}^s - \theta^s) + O_p(n^{-2}).$$

But $\hat{\theta}^s - \theta^s = n^{-1/2}\kappa^{s,t} Z_t + O_p(n^{-1})$ as in Appendix B, so

$$\hat{\mathcal{D}} = \mathcal{D} + \frac{1}{2}n^{-3/2}\mathcal{C}_s\kappa^{s,t} Z_t + O_p(n^{-2}). \tag{C.5}$$

The second equality in (B.3) gives an expansion for $\hat{\phi}-\phi$. The quantity we are trying to calculate is

$$\mathrm{E}\{(\hat{\phi}_{BAY}-\phi)^2\}-\mathrm{E}\{(\hat{\phi}-\phi)^2\}=2\mathrm{E}\{(\hat{\phi}-\phi)\hat{\mathcal{D}}\}+\mathrm{E}\{\hat{\mathcal{D}}^2\}. \tag{C.6}$$

The rest of the argument is just algebraic manipulation: by combining (C.4), (C.5) and (B.3), rearranging terms and taking expectations, we deduce that (C.6) is indeed of the form (3.3), with $\mathcal{A}$ given by (3.5).

Appendix D: Derivation of (3.7)–(3.8)

Suppose $\hat{\phi}=\phi+Rn^{-1/2}+Sn^{-1}+Tn^{-3/2}+Un^{-2}+o_p(n^{-2})$ with $R,\ S,\ T$ and U all random quantities of $O_p(1)$. Thus $R=\phi_i\kappa^{i,j}Z_j$, etc. After taking a Taylor expansion with respect to $\hat{\phi}$ in $L(\phi,\hat{\phi})$, we find that

$$\begin{aligned}L(\phi,\hat{\phi})=&\ \frac{1}{2}L_2R^2n^{-1}+(L_2RS+\frac{1}{6}L_3R^3)n^{-3/2}\\&+(L_2RT+\frac{1}{2}L_2S^2+\frac{1}{2}L_3R^2S+\frac{1}{24}L_4R^4)n^{-2}+o_p(n^{-2}).\end{aligned}$$

Now let us compare two estimators $\hat{\phi}_0$ and $\hat{\phi}_1$ with corresponding $R_0,...,U_0,R_1,...,U_1$; however, we assume $R_0=R_1$ since Bayes estimators and MLE are equivalent to first order. Then

$$\begin{aligned}L(\phi,\hat{\phi}_1)-L(\phi,\hat{\phi}_0)=&\ \{L_2R_0(S_1-S_0)\}n^{-3/2}\\&+\{L_2R_0(T_1-T_0)+\frac{1}{2}L_2(S_1^2-S_0^2)+\frac{1}{2}L_3R_0^2(S_1-S_0)\}n^{-2}+o_p(n^{-2}).\end{aligned} \tag{D.1}$$

But if $\hat{\phi}_1$ is the Bayes estimator and $\hat{\phi}_0$ the MLE, then S_1-S_0 is the constant $\frac{\mathcal{C}}{2}$ by (C.2)–(C.5), and since $\mathrm{E}\{R_0\}=0$, the lead term in (D.1) has expectation 0. Consequently

$$\mathcal{A}_L=L_2\mathrm{E}\{\frac{1}{2}(S_1^2-S_0^2)+R_0(T_1-T_0)\}+\frac{1}{2}L_3\mathrm{E}\{R_0^2(S_1-S_0)\}. \tag{D.2}$$

However if we evaluate this expression for $L(\phi,\hat{\phi})=(\hat{\phi}-\phi)^2$, for which $L_2=2,\ L_3=0$, we deduce that

$$\mathcal{A}=\mathrm{E}\{S_1^2-S_0^2+2R_0(T_1-T_0)\}. \tag{D.3}$$

We also have that $S_1-S_0=\frac{1}{2}\mathcal{C}$ and $\mathrm{E}R_0^2=\phi_i\phi_j\kappa^{i,j}$. Combining these expressions with (D.2) and (D.3), we deduce the result.

Appendix E: Derivation of (3.13)

Write $\tilde{\phi}(z) = \phi(z;\theta) + n^{-1/2}R(z) + n^{-1}S(z) + o_p(n^{-1})$ where $R(z) = \phi_i(z;\theta)\kappa^{i,j}Z_j$ has mean 0 and $\mathrm{E}\{S(z)\} = b(z;\theta)$. Then

$$\begin{aligned}
0 &= \tilde{\phi}(\tilde{z}) - \phi(z;\theta) \\
&= \tilde{\phi}(\tilde{z}) - \tilde{\phi}(z) + \tilde{\phi}(z) - \phi(z;\theta) \\
&= (\tilde{z}-z)\tilde{\phi}'(z) + \frac{1}{2}(\tilde{z}-z)^2\tilde{\phi}''(z) + n^{-1/2}R(z) + n^{-1}S(z) + ... \\
&= (\tilde{z}-z)\phi'(z;\theta) + n^{-1/2}(\tilde{z}-z)R'(z) + \frac{1}{2}(\tilde{z}-z)^2\phi''(z;\theta) + n^{-1/2}R(z) + n^{-1}S(z) + ...
\end{aligned}$$

Solving for $\tilde{z}$, we find

$$\tilde{z} - z = -n^{-1/2}\frac{R(z)}{\phi'(z;\theta)} + n^{-1}\frac{R(z)R'(z)}{\{\phi'(z;\theta)\}^2} - n^{-1}\frac{R^2(z)\phi''(z;\theta)}{2\{\phi'(z;\theta)\}^3} - n^{-1}S(z) + ...$$

and hence

$$\begin{aligned}
\phi(\tilde{z};\theta) - \phi(z;\theta) &= (\tilde{z}-z)\phi'(z;\theta) + \frac{1}{2}(\tilde{z}-z)^2\phi''(z;\theta) + ... \\
&= -n^{-1/2}R(z) + n^{-1}\frac{R(z)R'(z)}{\phi'(z;\theta)} - n^{-1}S(z) + ...
\end{aligned}$$

Taking expectations, noting that $\mathrm{E}\{R(z)\} = 0$ and $\mathrm{E}\{R(z)R'(z)\} = \mathrm{E}\{\phi_i(z;\theta)\kappa^{i,k}Z_k\cdot$ $\cdot\phi_j'(z;\theta)\kappa^{j,\ell}Z_\ell\} = \phi_i(z;\theta)\phi_j'(z;\theta)\kappa^{i,j}$, we deduce (3.13).

Appendix F: Detailed calculations for Section 4

First we consider a number of examples of (4.1), and one counterexample.

Example 1. If $\phi(z;\theta) \sim e^{-\psi(\theta)z}$, then set $f(z;\theta) = -z,\ G = G_1 = 1$.

Example 2. Let H (here and subsequently) denote the distribution function of a standard normal random variable and suppose $\phi(z;\theta) = 1 - H(z-\mu)$ where $\mu = \mu(\theta)$ is an unknown location parameter. Also let $h = H'$ and note the relations $1 - H(x) \sim h(x)/x$ as $x \to \infty$ (see, e.g. Feller 1968, p. 193), $h'(x) = -xh(x)$. So $\phi_i(z;\theta) = \mu_i h(z-\mu) \sim \mu_i\cdot(z-\mu)\cdot\phi(z;\theta)$, $\phi_{ij}(z;\theta) = \mu_{ij}h(z-\mu) - \mu_i\mu_j h'(z-\mu) = \{\mu_{ij} + \mu_i\mu_j\cdot(z-\mu)\}h(z-\mu) \sim \mu_i\mu_j\cdot(z-\mu)^2\cdot\phi(z;\theta)$, and by extension of the same argument $\phi_{ijk}(z;\theta) \sim \mu_i\mu_j\mu_k\cdot(z-\mu)^3\cdot\phi(z;\theta)$. So in this case, $f(z;\theta) = z-\mu,\ G = G_1 = 1$.

Example 3. Now suppose $\phi(z;\theta) = 1 - H((z-\mu)/\sigma)$ with both μ and σ functions of unknown θ. In this case it is readily verified that $\phi_i \sim \{\sigma_i/\sigma\}\cdot\{(z-\mu)/\sigma\}^2\cdot\phi(z;\theta),\ \ \phi_{ij} \sim$

$\{\sigma_i/\sigma\}\cdot\{\sigma_j/\sigma\}\cdot\{(z-\mu)/\sigma\}^4\cdot\phi(z;\theta)$, $\phi_{ijk}\sim\{\sigma_i/\sigma\}\cdot\{\sigma_j/\sigma\}\cdot\{\sigma_k/\sigma\}\cdot\{(z-\mu)/\sigma\}^6\cdot\phi(z;\theta)$, which is again of form (4.1) with $G=G_1=1$, but a different $f(z;\theta)$ compared with Example 2.

Example 4. Suppose $\phi(z;\theta)\sim\{(z-\mu)/\sigma\}^{-\alpha}$ with $\alpha>0$ unknown (as well as, possibly, μ and σ). Then $\phi_i(z;\theta)\sim-\alpha_i\{(z-\mu)/\sigma\}^{-\alpha}\log\{(z-\mu)/\sigma\}$, $\phi_{ij}(z;\theta)\sim\alpha_i\alpha_j\{(z-\mu)/\sigma\}^{-\alpha}\log^2\{(z-\mu)/\sigma\}$, $\phi_{ijk}(z;\theta)\sim-\alpha_i\alpha_j\alpha_k\{(z-\mu)/\sigma\}^{-\alpha}\log^3\{(z-\mu)/\sigma\}$. So in this case $f(z;\theta)=-\log\{(z-\mu)/\sigma\}$ and $G=G_1=1$.

Example 5. Here is an example for which $G\neq 1\neq G_1$. Suppose $z_\omega=\mu$ and $\phi(z;\theta)\sim\{(\mu-z)/\sigma\}^{\alpha}$ as $z\uparrow\mu$, $\alpha>0$, with μ a function of unknown θ. Then $\phi_i(z;\theta)/\phi(z;\theta)\sim\alpha(\mu-z)^{-1}\mu_i$, $\phi_{ij}(z;\theta)/\phi(z;\theta)\sim\alpha(\alpha-1)(\mu-z)^{-2}\mu_i\mu_j$, $\phi_{ijk}(z;\theta)/\phi(z;\theta)\sim\alpha(\alpha-1)(\alpha-2)(\mu-z)^{-3}\mu_i\mu_j\mu_k$. This satisfies (4.1) $f(z)=\alpha(\mu-z)^{-1}$, $G=1-1/\alpha$, $G_1=(1-1/\alpha)(1-2/\alpha)$.

Example 6. Finally, an example wich does not fit (4.1). Consider Example 5 but with μ and α known and the only unknown parameter $\sigma=\theta_1$. Then $\phi_1/\phi=-\alpha\sigma^{-1}$, $\phi_{11}/\phi=\alpha(\alpha+1)\sigma^{-2}$, $\phi_{111}/\phi=-\alpha(\alpha+1)(\alpha+2)\sigma^{-3}$, which do not depend on z at all; in other words, there is no $f(z;\theta)$ tending to $\pm\infty$ for which (4.1) holds. In cases such as this, the risk function does not take any special asymptotic form as $z\uparrow z_\omega$ and asymptotic relationships such as (F.1) below do not hold.

Having specified ϕ, we must next specify Q. We have $Q(\theta)=\log\pi(\theta)-\log w(\phi)$ where π is the prior density and w the tail weight, but as may easily be verified, π does not have any influence on the $z\to z_\omega$ asymptotic results, so we ignore it. Changing notation, assume $Q=Q(z;\theta)\sim\eta(-\log\phi(z;\theta))$ where η is, in principle, an arbitrary function. Then $Q_j\sim-\eta' f\psi_j$, $Q_{jk}\sim(\eta''-\eta' G+\eta')f^2\psi_j\psi_k$, where the derivatives η' and η'' are each evaluated at $-\log\phi(z;\theta)$. In the special case $\eta(x)=rx$ we have $\eta'=r$, $\eta''=0$ and if in addition $G=1$, then Q_{jk} may be ignored in the asymptotics to follow.

(a) Squared error loss function. In the notation of (3.3)–(3.5), we find that

$$\begin{aligned}
\mathcal{C}&\sim(G-2\eta')f^2(z;\theta)\psi_j\psi_k\kappa^{j,k}\phi(z;\theta),\\
\mathcal{C}_s&\sim(G_1-4G\eta'+2\eta''+2\eta')f^3(z;\theta)\psi_j\psi_k\psi_s\kappa^{j,k}\phi(z;\theta),\\
\mathcal{A}&\sim\left[2\eta''+\eta'^2+(2-6G)\eta'+\frac{3G^2}{4}+G_1\right]f^4(z;\theta)\left(\psi_j\psi_k\kappa^{j,k}\right)^2\phi^2(z;\theta).
\end{aligned}\tag{F.1}$$

The function η is arbitrary, but to keep the expression in square brackets bounded, it is desirable that both $\eta'(x)$ and $\eta''(x)$ remain bounded as $x\to\infty$. Suppose we assume $\eta'(x)\to r$, with r finite. If we assume η''' to be bounded, which is a kind of smoothness condition, then it follows from a result given by Littlewood (Bollobás 1986, page 54) that $\eta''(x)\to 0$. The expression in square brackets then reduces to $r^2+(2-6G)r+\frac{3}{4}G^2+G_1$, which is minimised when $r=3G-1$. In the most common case when $G=1$, this leads to $r=2$, thus generalising the result of Section 2 to a wide class of distributions.

(b) Logarithmic loss function. Our starting point is (3.9), where since $\phi \to 0$ we may replace $1-\phi$ and $1-2\phi$ by 1. Thus

$$\begin{aligned}\mathcal{A}_L &\sim \frac{1}{2\phi}\left(\mathcal{A} - \mathcal{C}\cdot\frac{\phi_i\phi_j\kappa^{i,j}}{\phi}\right) \\ &\sim \frac{1}{2}\left[2\eta'' + {\eta'}^2 + (4-6G)\eta' + \frac{3G^2}{4} + G_1 - G\right].f^4\left(\psi_j\psi_k\kappa^{j,k}\right)^2\phi\end{aligned}$$

Again we write r for the limit of $\eta'(x)$ as $x \to \infty$; then we must minimise $r^2 + (4-6G)r$ which leads to $r = 3G-2$.

(c) Squared logarithmic loss function. This is very similar, starting from (3.11). In this case

$$\begin{aligned}\mathcal{A}_L &\sim \frac{1}{\phi^2}\left(\mathcal{A} - \frac{3}{2}\mathcal{C}\cdot\frac{\phi_i\phi_j\kappa^{i,j}}{\phi}\right) \\ &\sim \left[2\eta'' + {\eta'}^2 + (5-6G)\eta' + \frac{3G^2}{4} + G_1 - \frac{3G}{2}\right].f^4\left(\psi_j\psi_k\kappa^{j,k}\right)^2\end{aligned}$$

Again let $\eta' \to r$, $\eta'' \to 0$. The minimum of $r^2 + (5-6G)r$ is attained when $r = 3G - \frac{5}{2}$. In the case $G = G_1 = 1$, the expression in square brackets reduces to $r^2 - r + \frac{1}{4}$, which is minimised when $r = \frac{1}{2}$, but which only then takes the value 0. In other words, for this loss function, even with our "best" Bayes estimator, we have still not demonstrated a clear superiority over the plug-in approach, in contrast with the situation in cases (a) and (b) just considered.

(d) Bias in $\hat{\phi}_{BAY}$. According to (3.6), the leading term in the bias is asymptotic to

$$(G-r)f^2(z;\theta)\psi_i\psi_j\kappa^{i,j}\phi(z;\theta).$$

Thus if our objective is to minimise the bias of the Bayes estimator of ϕ, we should take $r = G$.

(e) Bias in coverage probability. This is a little more complicated because we need assumptions on $\phi'(z)$ and its θ-derivatives, as well as $\phi(z)$ itself. However, for each of our examples 1–5 we find that

$$\frac{\phi_i'(z;\theta)}{\phi'(z;\theta)} \sim G\psi_i f(z;\theta) \quad \text{as} \quad z \uparrow z_\omega \tag{F.2}$$

so we adopt (F.2) as a general assumption. In that case, we find from (3.13) and (3.6) that the bias in coverage probability based on the Bayes estimator is asymptotic to

$$rf^2(z;\theta)\psi_i\psi_j\kappa^{i,j}\phi(z;\theta) \tag{F.3}$$

and this is minimised by taking $r = 0$. On the other hand, for the MLE, using (3.13) together with (3.2), the asymptotic form of the bias in coverage probability is of the same form as (F.3) but with r replaced by $G/2$, which is not at our liberty to set equal to 0. Thus at least in this asymptotic sense, we are able to make precise the notion that the usual Bayes estimator has smaller coverage probability bias than the MLE.

Appendix G: Derivations for Poisson-Gamma model

In this Appendix we do not use the summation convention, components of the vector θ are denoted by subscripts rather than superscripts, while superscripts denote powers as in the usual algebraic notation.

Since X_i has a Poisson distribution with mean $nt_i\theta_i$, we have, modulo constants, $\ell_n(\theta) = \sum(X_i \log\theta_i - nt_i\theta_i)$. Differentiating and taking expectations, $\kappa_{ii} = -t_i/\theta_i$, $\kappa_{iii} = 2t_i/\theta_i^2$, $\kappa_{i,ii} = -t_i/\theta_i^2$, $\kappa_{i,iii} = 2t_i/\theta_i^3$. Also $\kappa_{i,i} = t_i/\theta_i$, $\kappa^{i,i} = \theta_i/t_i$, while all terms involving combinations of different coefficients (e.g. $\kappa_{i,j}$ with $i \neq j$) are 0. By (3.4), $\mathcal{C} = \sum B_i/t_i$, where $B_i = 2\phi_i + \theta_i\phi_{ii} + 2\theta_i\phi_i Q_i$. Also $\mathcal{C}_s = \sum_i B_{is}/t_i$ where $B_{is} = \partial B_i/\partial\theta_s = 2\phi_{is} + \delta_{is}\phi_{ii} + \theta_i\phi_{iis} + 2\delta_{is}\phi_i Q_i + 2\theta_i\phi_{is}Q_i + 2\theta_i\phi_i Q_{is}$ (here δ_{is} is the Kronecker delta). To evaluate the second row of (3.5), first note that $\kappa_{i,ii} + \frac{1}{2}\kappa_{iii} = 0$, so the term multiplying $\mathcal{C}$ is $\frac{1}{2}\sum\kappa^{i,i}\phi_{ii} = \frac{1}{2}\sum\theta_i\phi_{ii}/t_i$. For the last two rows in (3.5), first note that the only non-zero terms are those for which all indices are the same, and they then reduce to

$$\begin{aligned}
&\sum_i \frac{\theta_i^3}{t_i^3}\phi_i\left\{-\frac{t_i}{\theta_i^2}(\phi_{ii} + 2\phi_i Q_i) + \left(\frac{2t_i}{\theta_i^3} - 4\frac{\theta_i}{t_i}\frac{t_i}{\theta_i^2}\frac{t_i}{\theta_i^2}\right)\phi_i\right\} \\
&= \sum_i \frac{\phi_i}{t_i^2}\left(-\theta_i\phi_{ii} - 2\theta_i\phi_i Q_i - 2\phi_i\right) \\
&= -\sum_i \frac{\phi_i B_i}{t_i^2}.
\end{aligned}$$

Hence with $\mathcal{C}$, $\mathcal{C}_s$ and B_i already defined, we have

$$\mathcal{A} = \frac{\mathcal{C}^2}{4} + \sum_s \frac{\mathcal{C}_s\theta_s\phi_s}{t_s} + \frac{\mathcal{C}}{2}\sum_i \frac{\theta_i\phi_{ii}}{t_i} - \sum_i \frac{\phi_i B_i}{t_i^2}.$$

The corresponding formula for logarithmic loss, based on (3.9), is

$$\mathcal{A}_L = \frac{\mathcal{A}}{2\phi(1-\phi)} - \frac{(1-2\phi)\mathcal{C}}{2\phi^2(1-\phi)^2}\sum_i \frac{\theta_i\phi_i^2}{t_i}.$$

To extend the computation to coverage probability bias, we note from (3.2) that for the MLE,

$$b(z;\theta) = \frac{1}{2}\sum_i \frac{\theta_i\phi_{ii}}{t_i}$$

while for the Bayes estimator, from (3.6),

$$b(z;\theta) = \sum_i \frac{1}{t_i}(\phi_i + \theta_i\phi_{ii} + \theta_i\phi_i Q_i).$$

The asymptotic expression for the CPB then follows from (3.13).

To complete the calculations, we need expressions for the derivatives of ϕ and Q. With $\phi = 1 - \exp(-z\sum\theta_j)$, it follows at once that

$$\phi_i = ze^{-z\sum\theta_j}, \qquad \phi_{is} = -z^2 e^{-z\sum\theta_j}, \qquad \phi_{ist} = z^3 e^{-z\sum\theta_j}.$$

For Q, from (7.3) we deduce

$$\begin{aligned} Q_i &= \frac{\alpha-1}{\theta_i} - \frac{p\alpha+\gamma}{\sum\theta_j+\delta}, \\ Q_{is} &= -\frac{\alpha-1}{\theta_i^2}\delta_{is} + \frac{p\alpha+\gamma}{(\sum\theta_j+\delta)^2}, \end{aligned}$$

For (3.13), we also need to calculate

$$\begin{aligned} \phi' &= (\sum\theta_j)e^{-z\sum\theta_j}, \\ \phi_i' &= (1 - z\sum\theta_j)e^{-z\sum\theta_j}, \end{aligned}$$

and hence

$$\frac{\phi_i'}{\phi'} = \frac{1}{\sum\theta_j} - z.$$

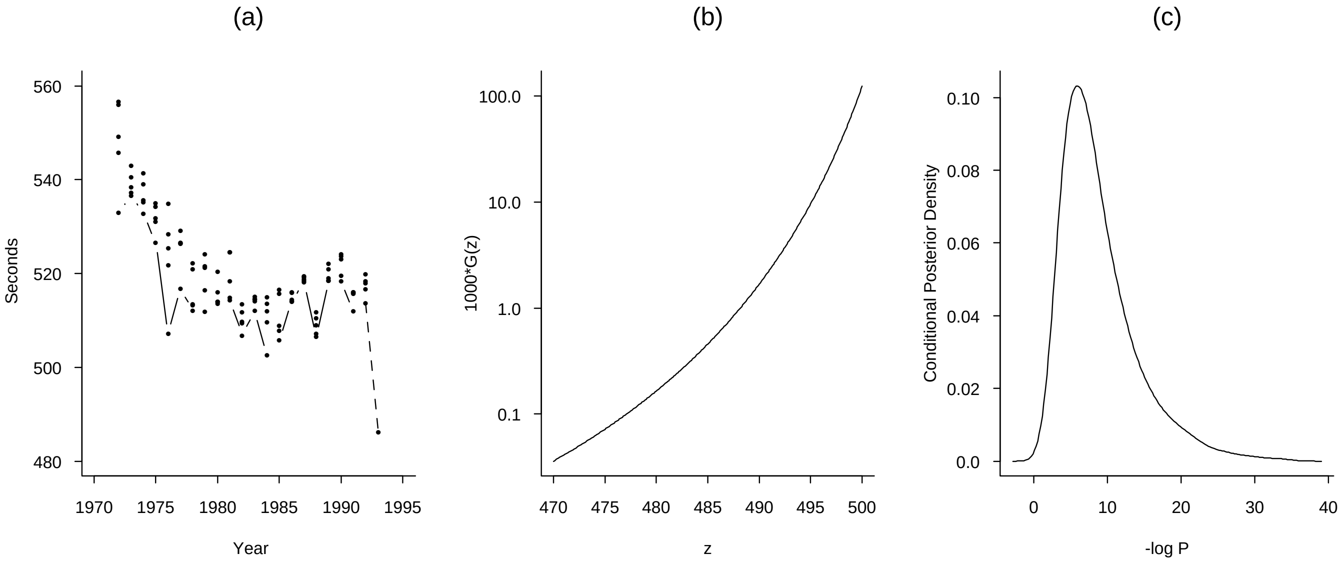


Fig. 1.1. (a) Plot of five best performances in women's 3000 metre event for each year from 1972–1992, together with Wang's 1993 record. (b) Predictive distribution function. (c) Posterior density of $\phi = G(486.11;\theta)$ conditioned on $\phi > 0$.

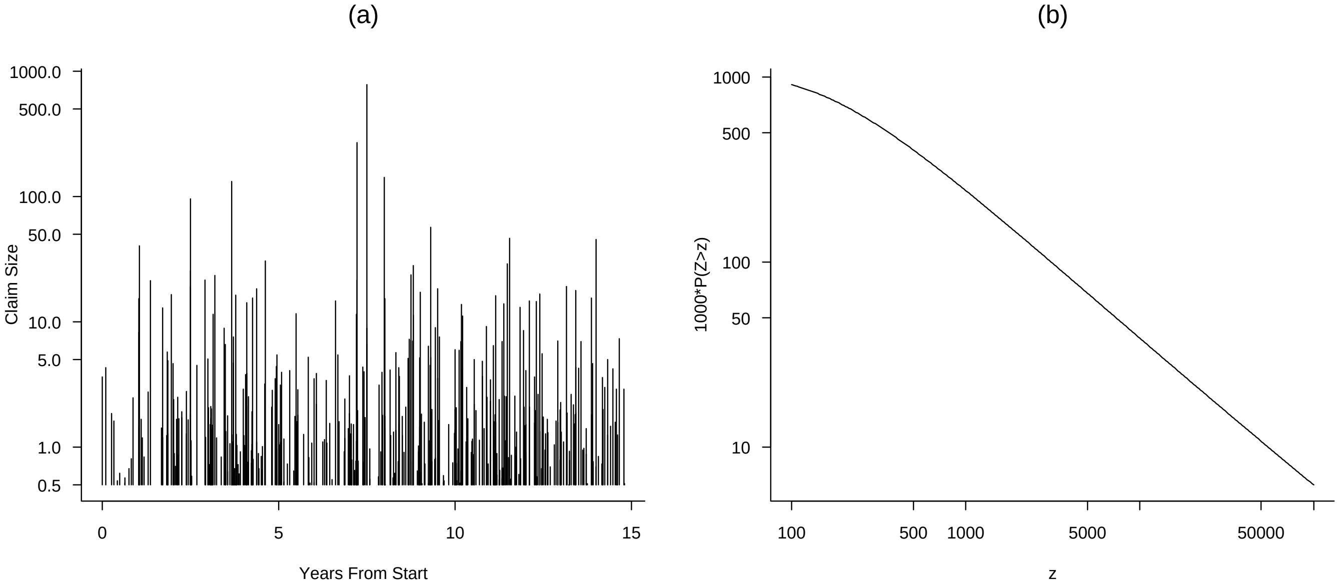


Fig. 1.2. (a) Magnitude and times of insurance claims for 15-year period. (b) Upper tail of predictive distribution function.

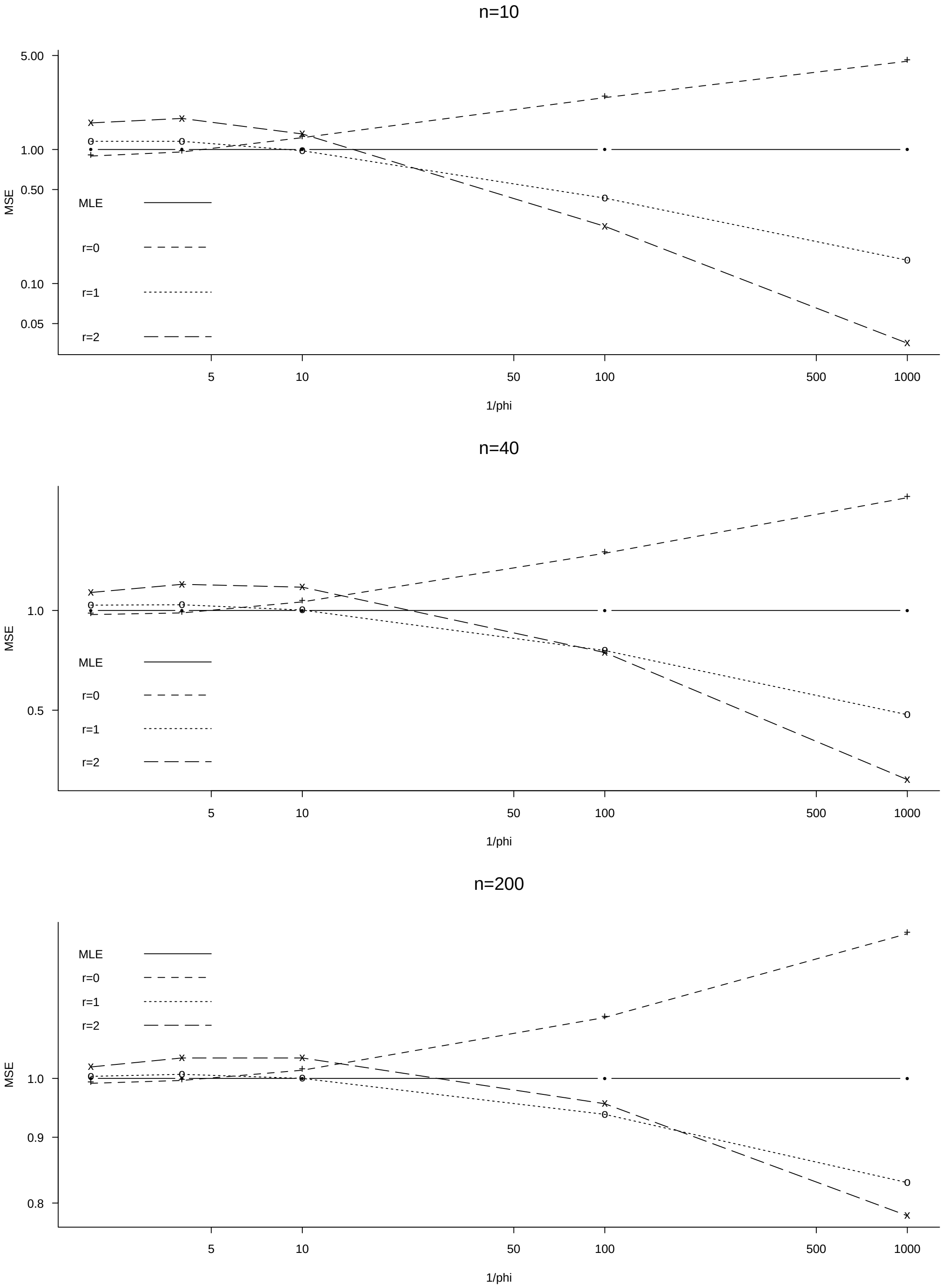


Fig. 2.1. Simulated mean squared errors of four estimators, scaled relative to the MLE for each value of ϕ, and for each of three sample sizes n.

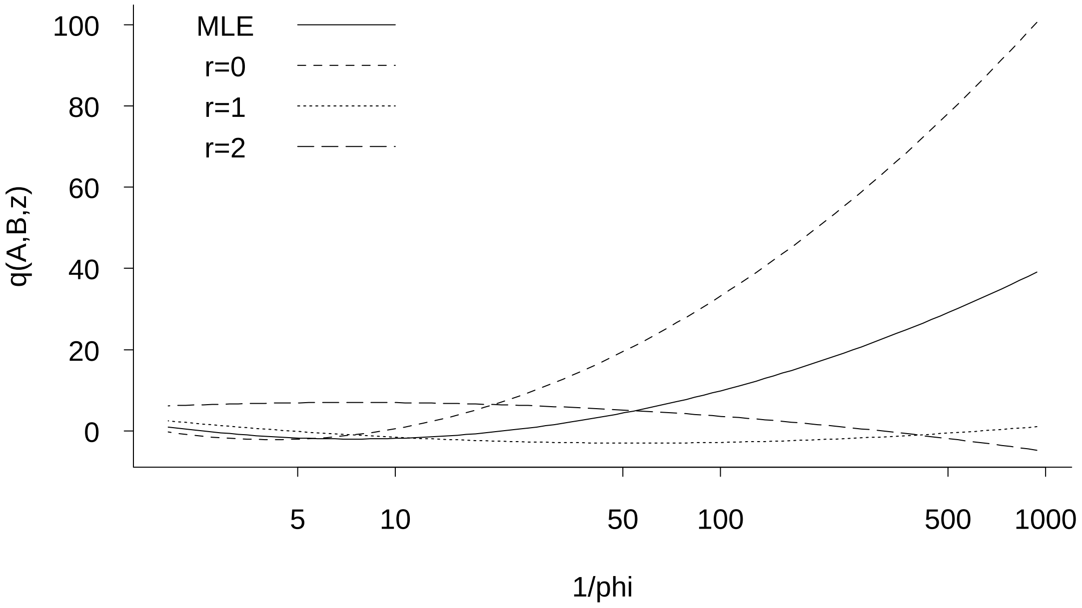


Fig. 2.2. Theoretical $q(A, B, y)$ function for the same four estimators as in Fig. 2.1.

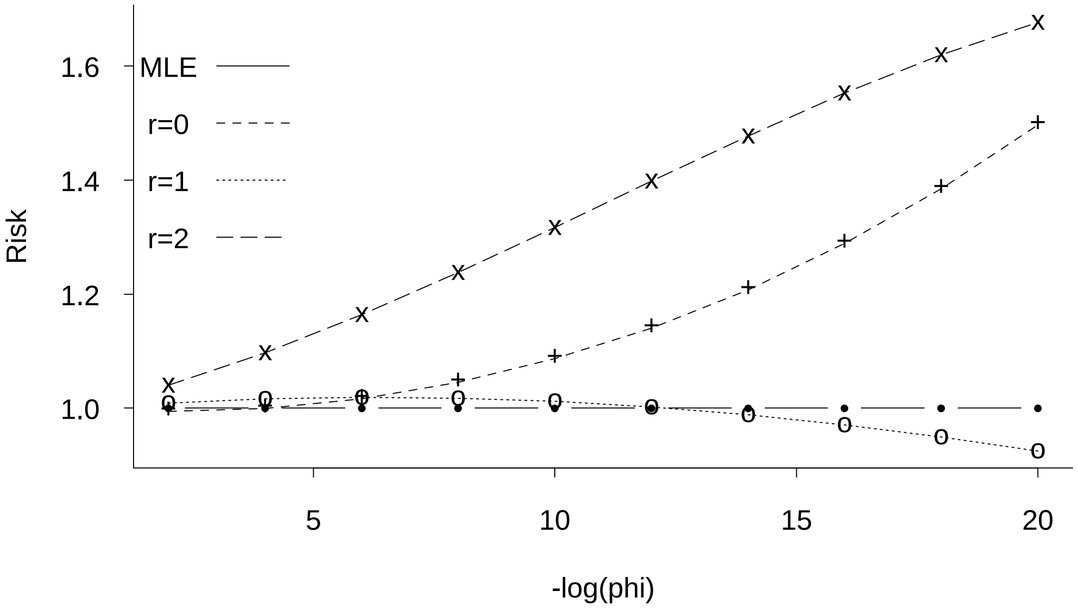


Fig. 2.3. Simulated logarithmic risk function for samples of size $n = 500$, scaled relative to the MLE for each value of ϕ.

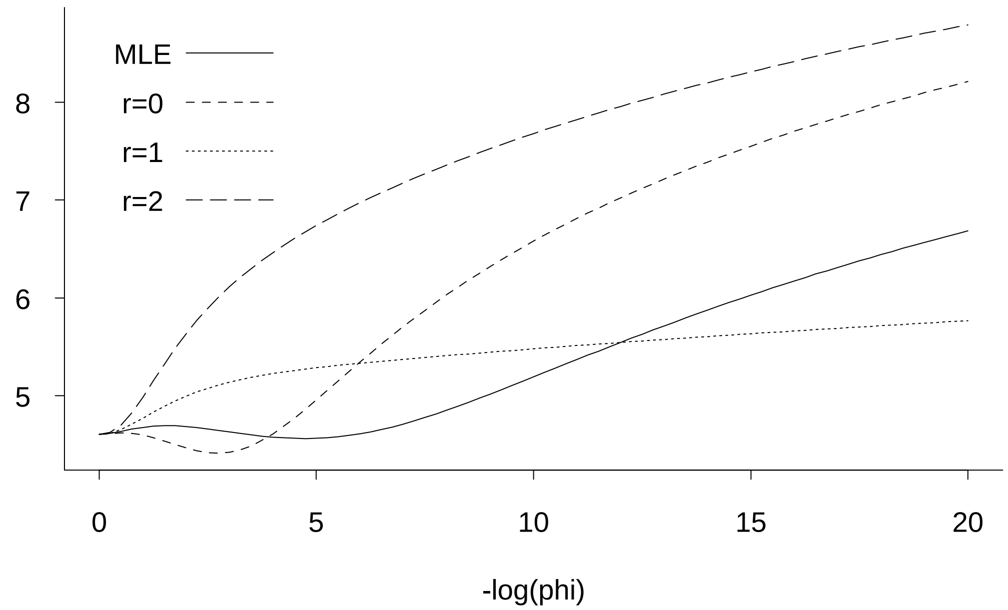


Fig. 2.4. Theoretical $q^\dagger(A, B, y)$ function for the same four estimators as in Fig. 2.3. To achieve a reasonable scaling, the actual value plotted is $\log(100 + q^\dagger(A, B, y))$.

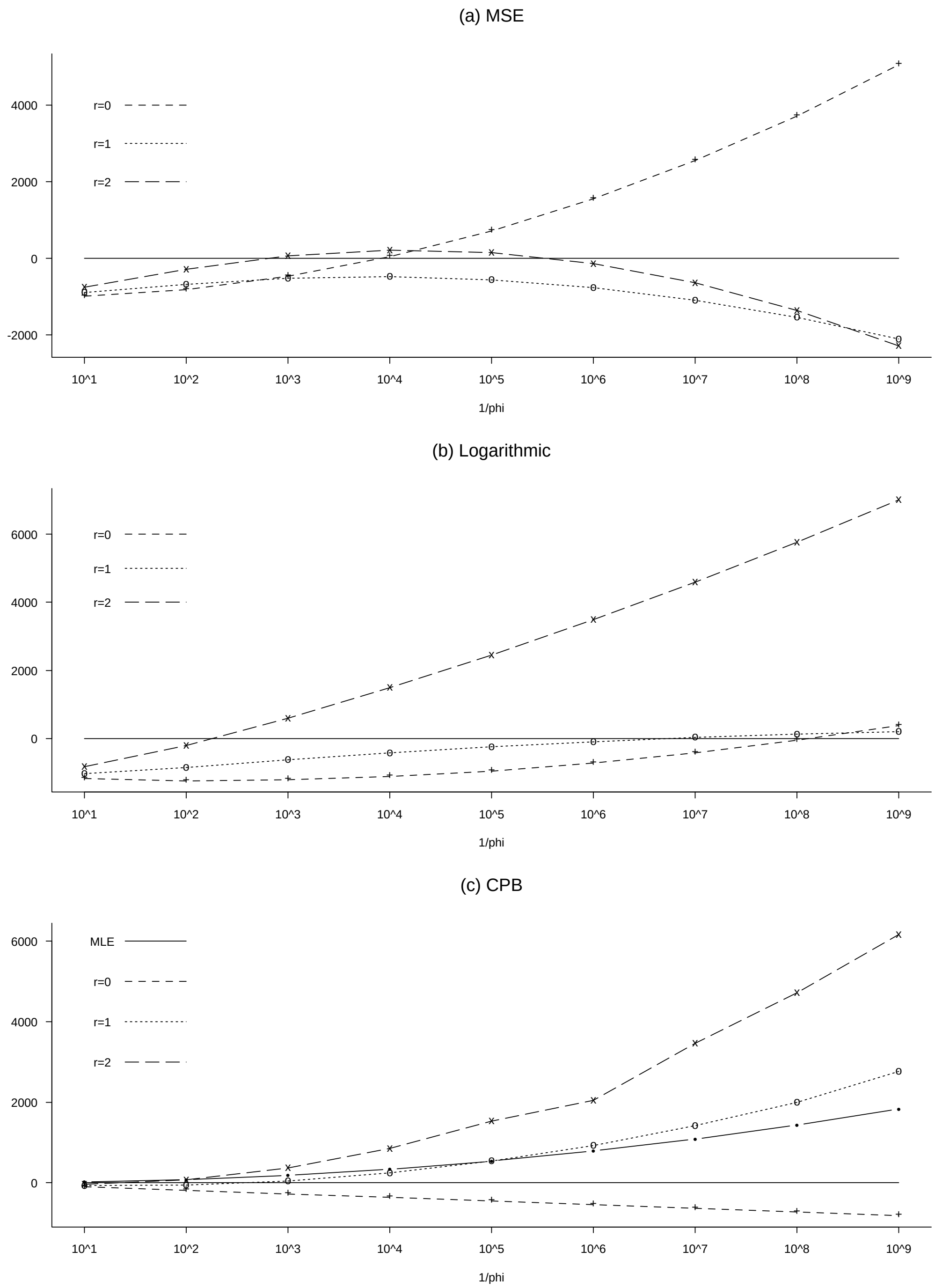


Fig. 5.1. (a) Normalised values of $\mathcal{A}$ associated with squared error loss function for three Bayes estimators. (b) Normalised values of $\mathcal{A}_L$ associated with logarithmic loss function for three Bayes estimators. (c) Theoretical calculations of the asymptotic coverage probability bias for four estimators. For each value of ϕ, the values are scaled so that the MLE always has CPB equal to 1.

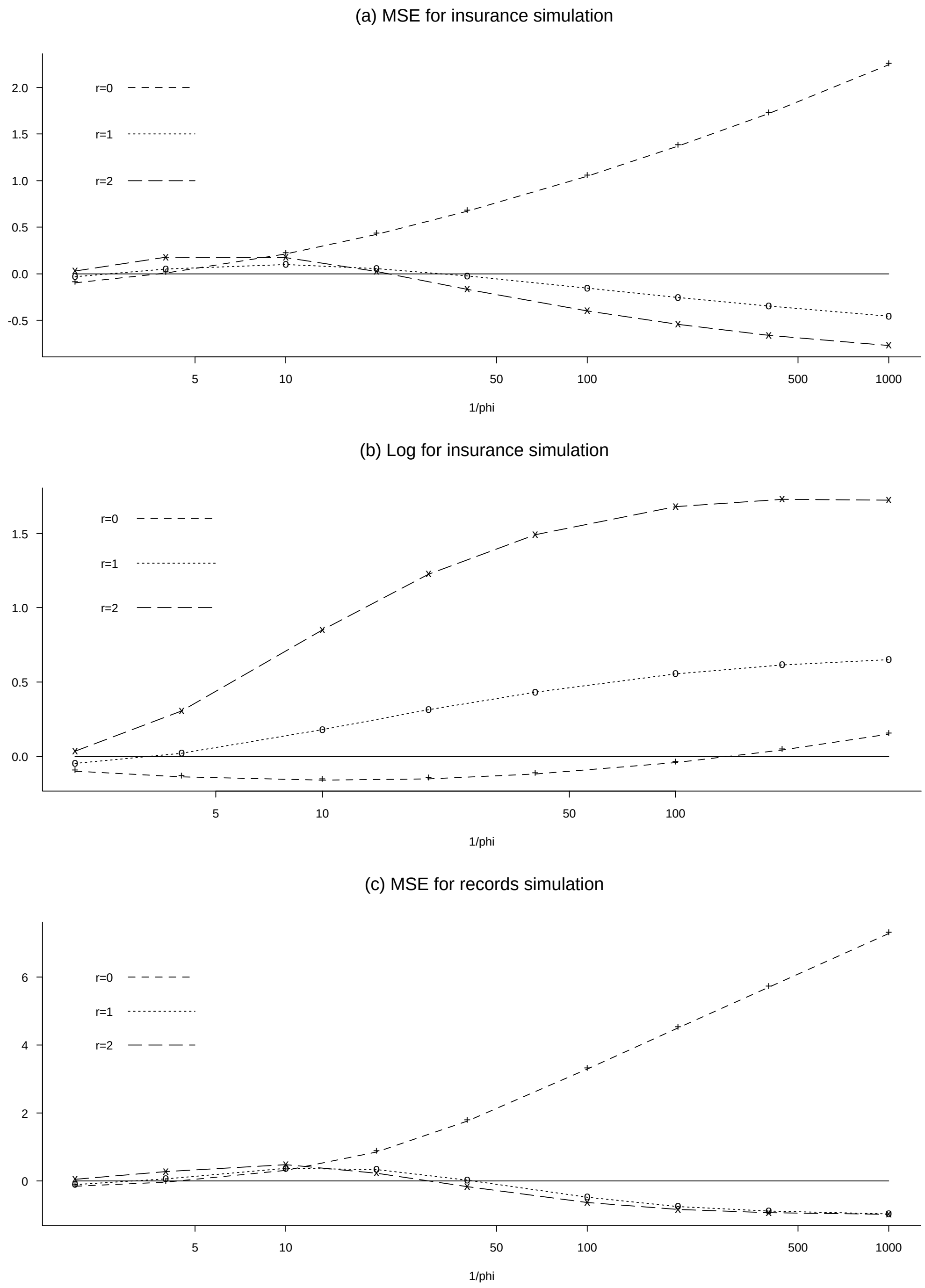
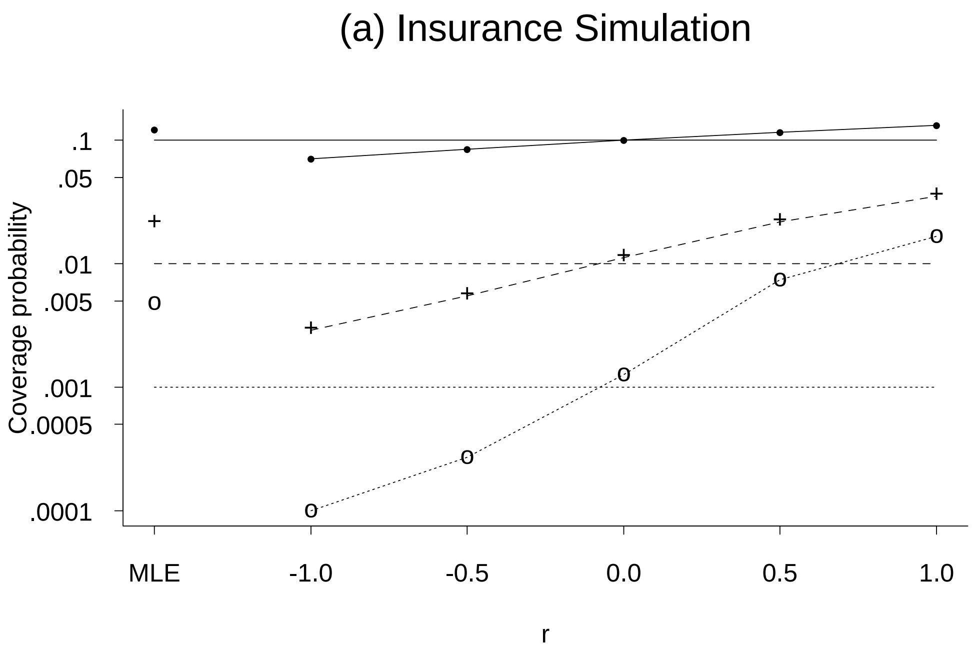


Fig. 5.2. Simulated risk functions for four estimators, rescaled relative to the MLE (solid line).



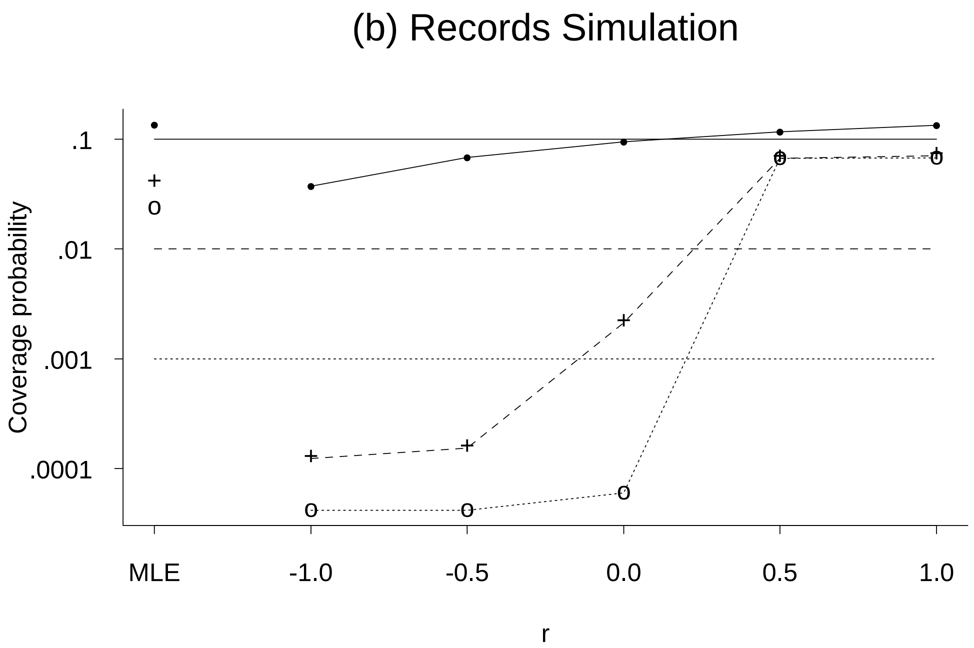


Fig. 5.3. Simulated values of CPB for three values of α and six estimators. In each case the plotted value is $\mathrm{E}\{\phi(\tilde{z};\theta)\}$ and this is compared with the nominal value of α, respectively .1 (top plot), .01 (middle), .001 (bottom).

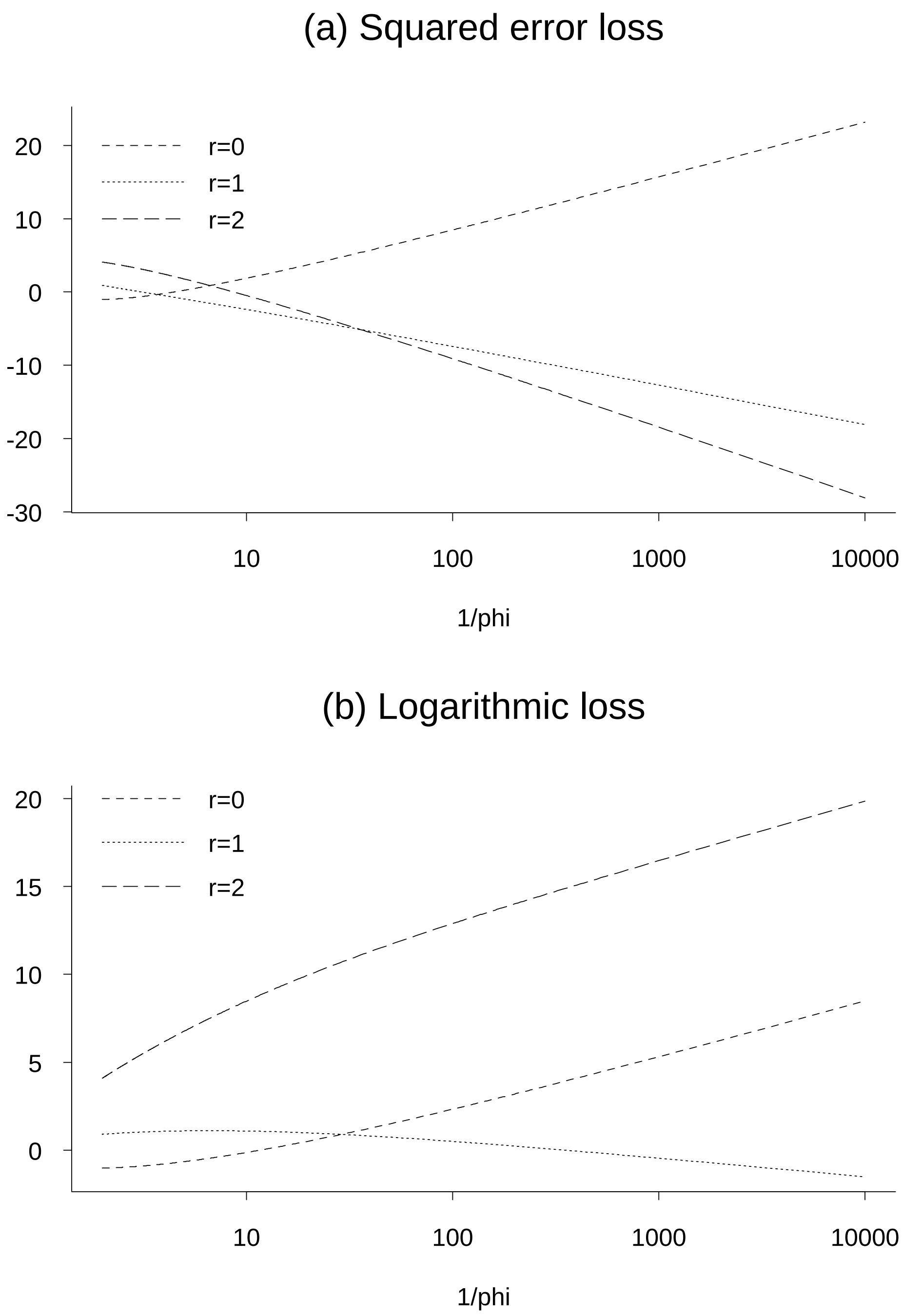


Fig. 6.1. Influence of the parameter r on predictive inference. (a) The curve $\bar{\mathcal{A}}/h^2(v)$, taken from (6.11) with $\alpha = 0$. (b) The curve $2H(v)\{1-H(v)\}\bar{\mathcal{A}}_L/h(v)^2$, taken from (6.13) with $\alpha = 0$. In each case the x axis is proportional to $1/\phi$, where $\phi = 1 - H(v)$.

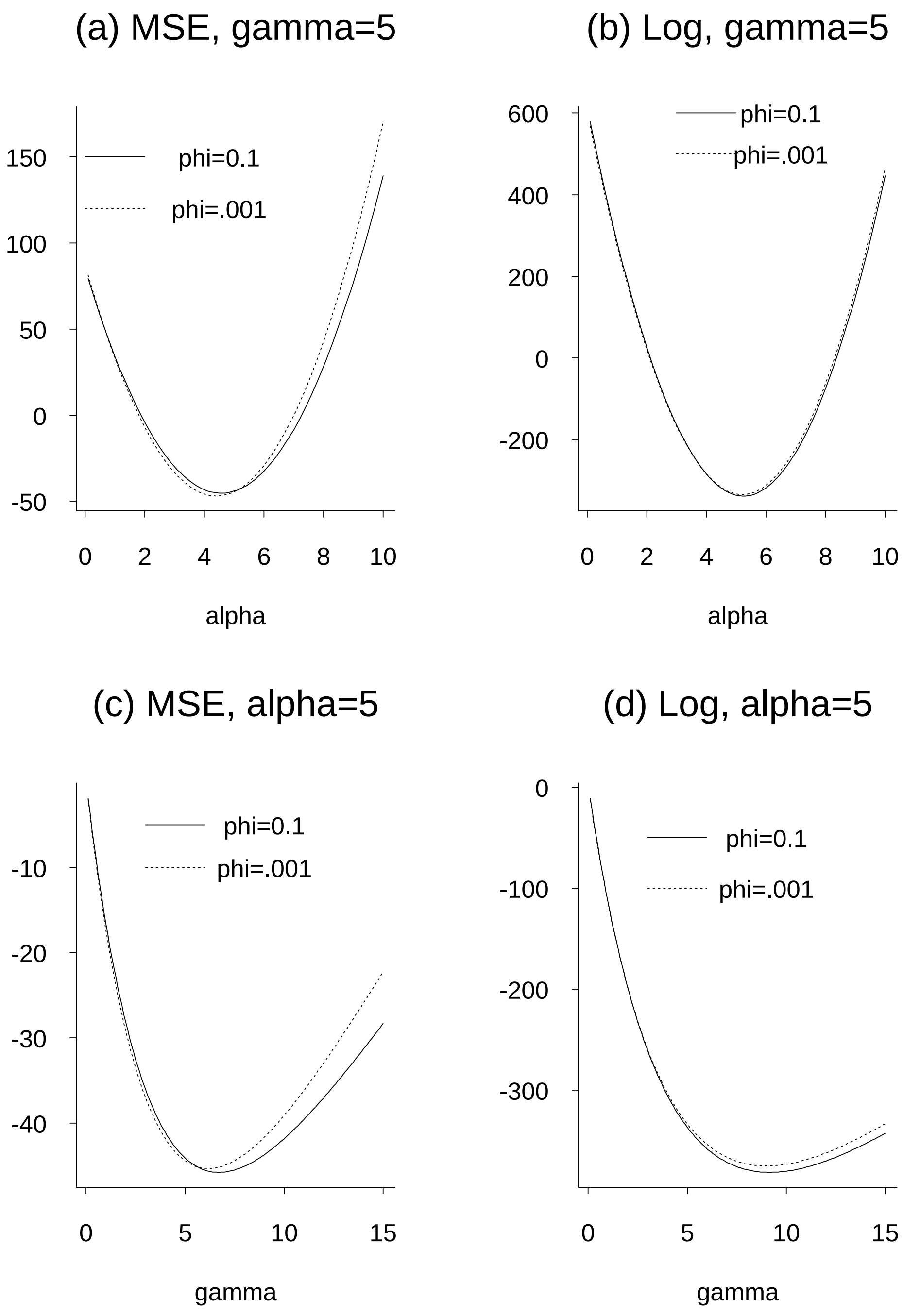


Fig. 7.1. Poisson-Gamma model: The curves $10^3 \times \mathcal{A}/\phi^2$ (a,c) and $10^4 \times \mathcal{A}_L/\phi$ (b,d), plotted as a function of α for fixed $\gamma = 5$ (a,b), or against γ for fixed $\alpha = 5$ (c,d). Curves are shown for two values of z corresponding to $\phi = 0.1$ and 0.001.

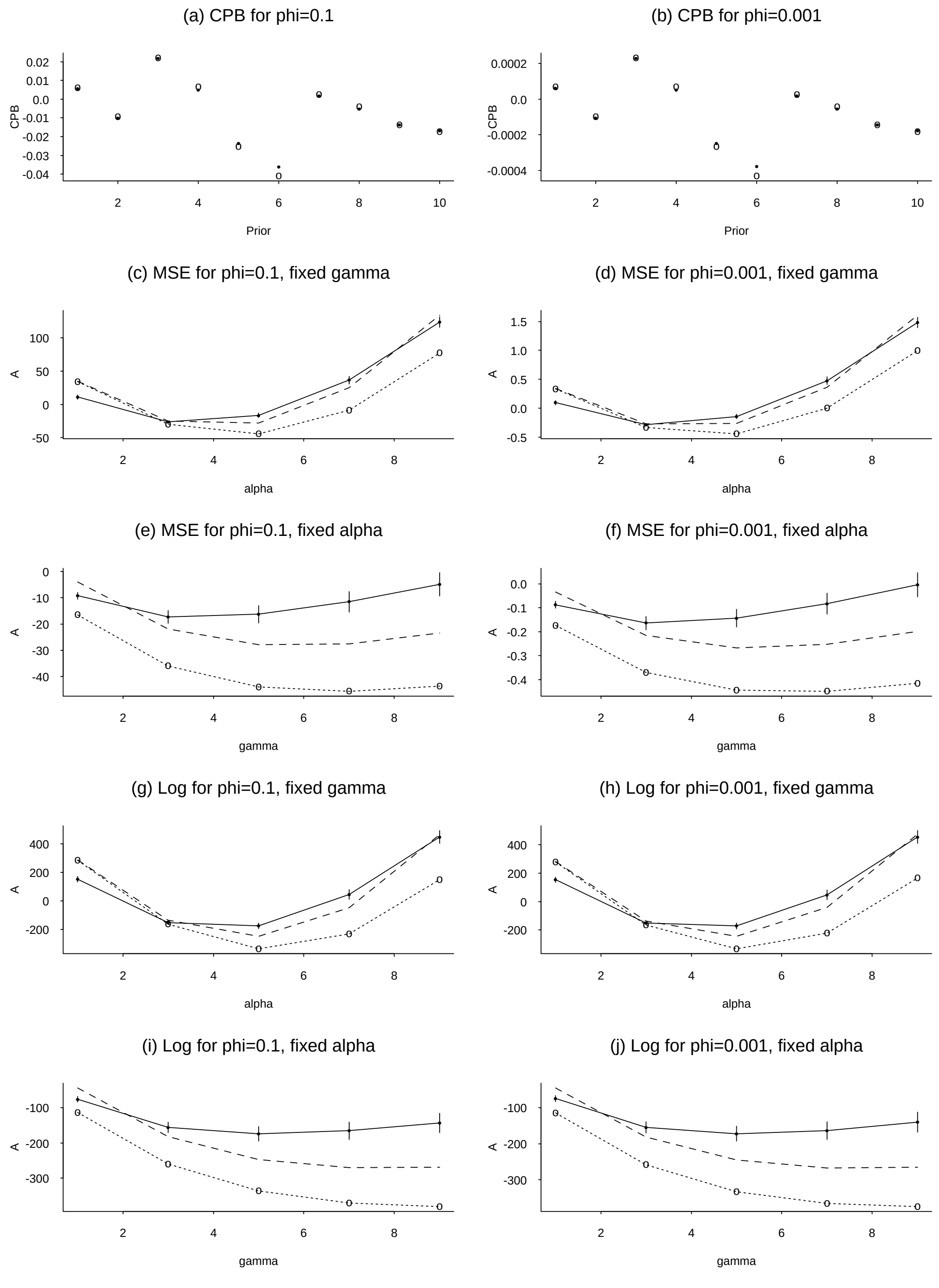


Fig. 7.2. Poisson-Gamma model. (a) and (b) show simulated (solid dots) and theoretical (circles) CPBs for MLE ("prior 1"), for Bayes with $\alpha = \gamma = 5$ (prior 2), for Bayes with $\alpha = 5$ and $\gamma = 1, 3, 7, 9$ (priors 3–6) and for Bayes with $\gamma = 5$ and $\alpha = 1, 3, 7, 9$ (priors 7–10). (c)–(j) show $10^4 \times \mathcal{A}/\phi$ (MSE) or $10^4 \times \mathcal{A}_L/\phi$ (logarithmic loss) for different scenarios. Solid lines: simulation results, with 95% confidence bands for simulation-induced error. Circles and dotted lines: theoretical calculation. Dashed lines: improved theoretical calculation.