ON THE BAYESIAN APPROACH FOR FREQUENTIST COMPUTATIONS

Rahul Mukerjee and N. Reid

Abstract

We review a Bayesian approach, based on a shrinkage argument, for frequentist computations. The approach can substantially simplify the computations underlying the frequentist higher order asymptotic theory. An illustrative example, on the approximate evaluation of the tail probability associated with the maximum likelihood estimator, is given.

Keywords: Edgeworth expansion; maximum likelihood estimators; posterior; prior; shrinkage; tail probability.

1 Introduction

In recent years, there has been a considerable interest in the study of the frequentist coverage of Bayesian credible sets; see Mukerjee and Reid (1999) and the references therein. In addition to yielding probability matching priors, this research has led to the development of a very powerful Bayesian route for frequentist computations. The Bayesian route is reminiscent of Dawid (1991) and was suggested originally by J.K. Ghosh to one of the present authors; cf. Ghosh and Mukerjee (1991). It is applicable in other contexts, including purely frequentist ones, as well. For example, as noted in Ghosh and Mukerjee (1994), it can be used in simplifying the computation of higher order power of standard statistical tests under contiguous alternatives.

Although Li (1998), in his unpublished PhD. thesis, gave an account of this Bayesian approach, no detailed description of it is as yet available in the published literature. The papers which utilize thie approach present only a rather brief outline. In the present article, we attempt to fill this gap by providing some of the underlying details and an illustrative example. It is hoped that this will stimulate the development of further applications of the Bayesian approach under consideration.

2 The basic steps

Let X be a possibly vector-valued random variable with a probability density function $g(\cdot\,;\theta)$ where the parameter θ belongs to $\mathcal{R}^p$ or some open subset thereof. Suppose it is intended to find an expression for $E_\theta\{h(X,\theta)\}$ where h is a measurable function and the expectation is known to exist. In particular, as in the next section, h may be an indicator functon in which case $E_\theta\{h(X,\theta)\}$ represents a probability. The Bayesian approach for the evaluation of $E_\theta\{h(X,\theta)\}$ consists of the following steps.

Step 1. Obtain the posterior density of θ under a prior density $\xi(\cdot)$ for θ. Hence obtain $E^\xi\{h(X,\theta)|X\}$, which is the expectation of $h(X,\theta)$ in the posterior setup.

Step 2. Find $E_\theta E^\xi\{h(X,\theta)|X\}$ $(=\lambda(\theta)$, say).

Step 3. Integrate $\lambda(\cdot)$ with respect to $\xi(\cdot)$ and then allow $\xi(\cdot)$ to converge weakly to the degenerate prior at θ. This yields $E_\theta\{h(X,\theta)\}$.

The above steps can be justified as follows. First note that the posterior density of θ under the prior $\xi(\cdot)$ is given by $g(X;\theta)\xi(\theta)/N(X)$, where

$$N(X) = \int g(X;\theta)\xi(\theta)d\theta \ . \tag{2.1}$$

Hence Step 1 yields

$$E^{\xi}\{h(X,\theta)|X\} = K(X)/N(X) \ , \tag{2.2}$$

where

$$K(X) = \int h(X,\theta)g(X;\theta)\xi(\theta)d\theta \ . \tag{2.3}$$

In view of (2.2), Step 2 yields

$$\lambda(\theta) = \int \{K(x)/N(x)\}g(x;\theta)dx \ .$$

Hence in Step 3, integrating $\lambda(\cdot)$ with repsect to $\xi(\cdot)$, one gets

$$\begin{aligned}
\int \lambda(\theta)\xi(\theta)d\theta &= \int\int \{K(x)/N(x)\}g(x;\theta)\xi(\theta)dxd\theta \\
&= \int \{K(x)/N(x)\}\{\int g(x;\theta)\xi(\theta)d\theta\}dx \\
&= \int K(x)dx \\
&= \int\int h(x,\theta)g(x;\theta)\xi(\theta)d\theta dx \\
&= \int \{\int h(x,\theta)g(x;\theta)dx\}\xi(\theta)d\theta \\
&= \int [E_\theta\{h(X,\theta)\}]\xi(\theta)d\theta \ ,
\end{aligned} \tag{2.4}$$

using (2.1) and (2.3). The changes in the order of integration in (2.4) are justified as $E_\theta\{h(X,\theta)\}$ exists. From the last expression in (2.4), the validity of the claim made in Step 3 is clear.

Observe that in Step 3, $\xi(\cdot)$ is allowed to converge weakly to a degenerate prior. This is why the present Bayesian approach is said to be based on a shrinkage argument. As illustrated in the next section, explicit specification of the prior $\xi(\cdot)$ is not needed in applying Steps 1-3 above. When executed up to the desired order of approximation, these steps can significantly simplify the algebra underlying frequentist higher order asymptotic computations. This simplification arises from two counts. First, the Bayesian approach avoids the explicit use of an Edgeworth expansion and the associated, often cumbersome, derivation of approximate cumulants. Second, it helps in expressing the final result in terms of derivatives (cf. (3.13) below) and hence in a compact form which is often easy to interpret especially when θ is vector-valued. The example presented in Section 3 below illustrates the above points.

3 An example

Consider the problem of evaluating the frequentist tail probability $P_\theta\{\sqrt{n}(\hat{\theta}-\theta) \leq u\}$ with margin of error $o(n^{-1/2})$, where $\hat{\theta}$ is the maximum likelihood estimator of a scalar parameter θ on the basis of n independent and identically distributed observations $X_1, \ldots, X_n$ with common density $f(\cdot;\theta)$, and the constant u is free from n and θ. We work under the regularity conditions of Bickel and Ghosh (1990) which include Edgeworth assumptions. Let $X = (X_1, \ldots, X_n)'$, $\ell(\theta) = n^{-1}\sum_{i=1}^{n} \log f(X_i;\theta)$,

$$c = -\ell''(\hat{\theta}), \qquad a = \ell'''(\hat{\theta}), \tag{3.1}$$

and $\xi(\cdot)$ be a prior density for θ. The prior density is assumed to be twice continuously differentiable. All formal expansions for the posterior, as used here, are valid for sample points in a set S with P_θ-probability $1 + o(n^{-1/2})$

uniformly over compact sets of θ. The set S may be defined following Bickel and Ghosh (1990). The quantity c is positive over S.

Observe that

$$P_\theta\{\sqrt{n}(\hat{\theta}-\theta) \leq u\} = E_\theta\{h(X,\theta)\} \ , \tag{3.2}$$

where the indicator $h(X,\theta)$ equals unity if $\sqrt{n}(\hat{\theta}-\theta) \leq u$ and zero otherwise. We now execute Steps 1-3 of the last section up to the desired order of approximation.

Step 1. The posterior density of θ under the prior $\xi(\cdot)$ is given by

$$\frac{\xi(\theta)\exp\{n\ell(\theta)\}}{\int \xi(\theta)\exp\{n\ell(\theta)\}d\theta} = \frac{\xi(\theta)\exp[n\{\ell(\theta)-\ell(\hat{\theta})\}]}{\int \xi(\theta)\exp[n\{\ell(\theta)-\ell(\hat{\theta})\}]d\theta} \ .$$

Hence the posterior density of $y=(nc)^{1/2}(\theta-\hat{\theta})$ under the prior $\xi(\cdot)$ is given by, say,

$$\xi^*(y|X) = b(y,X)/\{\int b(y,X)dy\} \ , \tag{3.3}$$

where

$$b(y,X) = \xi\{\hat{\theta}+(nc)^{-1/2}y\}\exp(n[\ell\{\hat{\theta}+(nc)^{-1/2}y\}-\ell(\hat{\theta})]) \ . \tag{3.4}$$

Since $\ell'(\hat{\theta})=0$, by (3.1) and (3.4), a Taylor's expansion about $\hat{\theta}$ yields

$$b(y,X) = \xi(\hat{\theta})\{1+n^{-1/2}(A_1y+A_3y^3)\}\exp\left(-\frac{1}{2}y^2\right)+o(n^{-1/2}) \ , \tag{3.5}$$

where

$$A_1 = c^{-1/2}\xi'(\hat{\theta})/\xi(\hat{\theta}) \ , \quad A_3 = \frac{1}{6}c^{-3/2}a \ . \tag{3.6}$$

Hence

$$\int b(y,X)dy = \xi(\hat{\theta})\sqrt{2\pi}+o(n^{-1/2}) \ . \tag{3.7}$$

By (3.3), (3.5) and (3.7),

$$\begin{aligned}\xi^*(y|X) &= \phi(y)\{1+n^{-1/2}(A_1 y + A_3 y^3)\} + o(n^{-1/2}) \\ &= \phi(y)[1+n^{-1/2}\{G_1 J_1(y) + G_3 J_3(y)\}] + o(n^{-1/2}) \ , \qquad (3.8)\end{aligned}$$

where

$$G_1 = A_1 + 3A_3 \ , \quad G_3 = A_3 \ , \qquad (3.9)$$

$\phi(\cdot)$ is the standard normal density and $J_i(\cdot)$ is the Hermite polynomial of degree i. The right hand side of (3.8) resembles the frequentist Edgeworth expansion but has a much simpler derivation, as it only involves derivatives of the loglikelihood function with respect to θ.

Let $P^\xi(\cdot|X)$ be the posterior probability measure under the prior $\xi(\cdot)$. Then, with $h(X,\theta)$ defined as in the context of (3.2), from (3.8) and (3.9) we get

$$\begin{aligned}&E^\xi\{h(X,\theta)|X\} = P^\xi\{\sqrt{n}(\hat\theta-\theta) \le u|X\} = P^\xi\{y \ge -u\sqrt{c}|X\} \\ =\ & \int_{-u\sqrt{c}}^{\infty} \phi(y)[1+n^{-1/2}\{G_1 J_1(y) + G_3 J_3(y)\}]dy + o(n^{-1/2}) \\ =\ & \Phi(u\sqrt{c}) + n^{-1/2}\{G_1 + G_3(u^2 c - 1)\}\phi(u\sqrt{c}) + o(n^{-1/2}) \\ =\ & \Phi(u\sqrt{c}) + n^{-1/2}\{A_1 + A_3(u^2 c + 2)\}\phi(u\sqrt{c}) + o(n^{-1/2}), \qquad (3.10)\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Step 2. Let

$$\begin{aligned} I &\equiv I(\theta) = -E_\theta\{d^2 \log f(X_1;\theta)/d\theta^2\}, \\ L &\equiv L(\theta) = E_\theta\{d^3 \log f(X_1;\theta)/d\theta^3\}.\end{aligned}$$

Then by (3.1), (3.6) and (3.10),

$$\lambda(\theta) = E_\theta E^\xi\{h(X,\theta)|X\}$$

$$= \Phi(u\sqrt{I}) + n^{-1/2}\{I^{-1/2}\frac{\xi'(\theta)}{\xi(\theta)} + \frac{1}{6}I^{-3/2}L(u^2I+2)\}\phi(u\sqrt{I}) + o(n^{-1/2}) \ . \quad (3.11)$$

Step 3. By (3.11),

$$\begin{aligned} & \int \lambda(\theta)\xi(\theta)d\theta \\ = & \int \{\Phi(u\sqrt{I}) + \frac{1}{6}n^{-1/2}I^{-3/2}L(u^2I+2)\phi(u\sqrt{I})\}\xi(\theta)d\theta \\ + & \ n^{-1/2}\int I^{-1/2}\xi'(\theta)\phi(u\sqrt{I})d\theta + o(n^{-1/2}) \end{aligned} \quad (3.12)$$

Now suppose $\xi(\cdot)$ is so chosen that it vanishes on the boundaries of an interval containing θ. Then, upon integration by parts,

$$\int I^{-1/2}\xi'(\theta)\phi(u\sqrt{I})d\theta = -\int [\frac{d}{d\theta}\{I^{-1/2}\phi(u\sqrt{I})\}]\xi(\theta)d\theta \ .$$

Using the above in (3.12) and then allowing $\xi(\cdot)$ to converge weakly to the degenerate prior at θ, by (3.2) one obtains

$$\begin{aligned} & P_\theta\{\sqrt{n}(\hat{\theta}-\theta) \le u\} = E_\theta\{h(X,\theta)\} \\ = & \ \Phi(u\sqrt{I}) + n^{-1/2}[\frac{1}{6}I^{-3/2}L(u^2I+2)\phi(u\sqrt{I}) - \frac{d}{d\theta}\{I^{-1/2}\phi(u\sqrt{I})\}] \\ & + o(n^{-1/2}) \ . \end{aligned} \quad (3.13)$$

Since $dI/d\theta = -(L+M)$, where

$$M = E_\theta[\{d\log f(X_1;\theta)/d\theta\}\{d^2\log f(X_1;\theta)/d\theta^2\}] \ ,$$

it can be seen that (3.13) is in agreement with what one obtains via a direct Edgeworth expansion, given for example in McCullagh (1987, Ch. 6). However, the use of a direct Edgeworth expansion calls for obtaining a stochastic expansion for $\sqrt{n}(\hat{\theta}-\theta)$ and then evaluation of approximate cumulants. The present Bayesian approach avoids all this algebra and thus simplifies the computation considerably. The gains, via the use of the Bayesian approach, can be even more significant in frequentist computations with vector valued θ.

Some examples of the use of the shrinkage argument with vector valued parameters can be found in the literature. Bickel and Ghosh (1990) use the argument to derive the validity of Bayesian and frequentist Bartlett correction factors, and Ghosh and Mukerjee (1991) to obtain an explicit expression for the frequentist Bartlett correction factor. Mukerjee and Reid (1999) use the argument in a problem of inference for a component parameter of interest in the presence of nuisance parameters. Datta et al. (1999) apply the method in investigating higher order expansions of posterior prediction intervals. Ghosh and Mukerjee (1999) highlight the argument as a means of obtaining probability matching priors in their review paper on that topic.

Acknowledgement

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Centre for Management and Development Studies, Indian Institute of Management Calcutta. It was partly done when RM as visiting the University of San Paulo, Brazil.

References

[1] Bickel, P.J. and Ghosh, J.K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction – a Bayesian argument. *Ann. Statist.* **18**, 1070-1090.

[2] Datta, G.S., Mukerjee, R., Ghosh, M. and Sweeting, T.J. (1999). Bayesian prediction with approximate frequentist validity. preprint.

[3] Dawid, A.P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *J. Roy. Statist. Soc. B* **53**, 79-109.

[4] 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.

[5] Ghosh, J.K. and Mukerjee, R. (1994). Higher order comparison of tests: a Bayesian route. In: Essays on Probability and Statistics (S.P. Mukherjee et al. eds.), pp. 92-101, Presidency College, Calcutta.

[6] Ghosh, M. and Mukerjee, R. (1999). Recent developments on probability matching priors. preprint.

[7] Li, J. (1998). Some Aspect of Bayesian and Frequentist Asymptotics. Ph.D. Thesis, University of Toronto.

[8] McCullagh, P. (1987). *Tensor methods in Statistics.* Chapman & Hall, London.

[9] Mukerjee, R. and Reid, N. (1999). On a property of probability matching priors: matching the alternative coverage probabilities. *Biometrika*, **86**, 333–340.