Approximate Bayesian Methods

D.V. LINDLEY
University College London

SUMMARY

This paper develops asymptotic expansions for the ratios of integrals that occur in Bayesian analysis: for example, the posterior mean. The first term omitted is $0(n^{-2})$ and it is shown how the term $0(n^{-1})$ can be of importance.

Keywords: ASYMPTOTIC EXPANSIONS; STEEPEST DESCENT; BAYESIAN METHODS; POSTERIOR MOMENTS; ONE-WAY ANALYSIS OF VARIANCE.

1. GENERAL DEVELOPMENT

In this paper we discuss the approximate evaluation of the ratio of integrals of the form

$$\int w(\theta)e^{L(\theta)}d\theta / \int v(\theta)e^{L(\theta)}d\theta. \qquad (1)$$

Here $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ is a parameter and

$$L(\theta) = \sum_{i=1}^{n} \log\, p(x_i|\theta)$$

is the logarithm of the likelihood for n observations $x_1, x_2, \dots, x_n$, forming a random sample from a density $p(\bullet|\theta)$. The functions $w(\bullet)$ and $v(\bullet)$ are arbitrary. A simple example is where $w(\theta) = \theta_s v(\theta)$ and $v(\bullet)$ is a prior distribution for θ, when (1) is the posterior mean of θ_s. Notice that the notation $L(\theta)$ suppresses the dependence on $x_1, x_2, \dots, x_n$. This is convenient because, in a Bayesian analysis, the x's, as observed data, are fixed and variation with respect to them is of no interest.

We shall be concerned with the asymptotic behaviour as $n \rightarrow \infty$ under regularity conditions, which will not be spelt out, in which $L(\theta)$ concentrates around the unique maximum likelihood value $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$, obtaining an asymptotic series in inverse powers of n as far as the term of order n^{-1}. Integrals of the form occurring in the numerator and denominator of (1) were considered by Lindley (1961) for univariate θ, $(m = 1)$. He obtained asymptotic expansions as far as the term of order n^{-1}. We here show that the asymptotic results for *ratios* of integrals are simpler than those for the separate integrals; and we illustrate the use of the expansions in several situations.

In the multivariate case the notation requires care. The basic idea is to expand the functions involved about $\hat{\theta}$ so obtaining terms involving $(\theta_i - \hat{\theta}_i)$, $(i = 1, 2, \ldots, m)$. We write this deviation simply as θ_i, effectively using $\hat{\theta}_i$ as the origin. Many partial derivatives occur and we write, for example, $\partial^3 L/\partial\theta_i\partial\theta_j\partial\theta_k$ as L_{ijk}. Hence each suffix denotes differentiation once with respect to the variable having that suffix. Thus L_{222} is the third derivative with respect to θ_2. All these are evaluated at $\hat{\theta}$. Notice that the order of the suffixes is irrelevant. Similar notations are used for v and w. With these conventions, the Taylor series expansion for L, say, about $\hat{\theta}$ may be written

$$L(\theta) = L + \Sigma L_i\theta_i + \tfrac{1}{2!}\Sigma L_{ij}\theta_i\theta_j + \tfrac{1}{3!}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \ldots$$

where all summations run over all suffixes from 1 to m, the dimensionality of θ. We begin by considering the numerator of (1) deriving the multivariate extension of the univariate results of Lindley (1961). It is important in collecting terms of like order together, to remember that L, and all of its derivatives, are $0(n)$, whereas θ_i, for all i, is $0(n^{-1/2})$. On expansion to $0(n^{-1})$ we have

$$\begin{aligned}
&\int w(\theta)e^{L(\theta)}d\theta \\
&\quad = \int \left[w + \Sigma w_i\theta_i + \tfrac{1}{2!}\Sigma w_{ij}\theta_i\theta_j + \ldots\right] \exp\left[L + \Sigma L_i\theta_i + \tfrac{1}{2!}\Sigma L_{ij}\theta_i\theta_j + \right. \\
&\qquad\qquad \left. \tfrac{1}{3!}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \tfrac{1}{4!}\Sigma L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \ldots\right]d\theta \\
&\quad = we^L \int \left[1 + \Sigma W_i\theta_i + \tfrac{1}{2}\Sigma W_{ij}\theta_i\theta_j + \ldots\right] \exp\left[\tfrac{1}{2}\Sigma L_{ij}\theta_i\theta_j\right] \\
&\quad \times \left[1 + \tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \tfrac{1}{24}\Sigma L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \tfrac{1}{2}\left\{\tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k\right\}^2 + \ldots\right]d\theta.
\end{aligned}$$

Here $W_i = w_i/w$, etc., $L_i = 0$, since the expansion is about the maximum likelihood value, and all functions are evaluated at $\hat{\theta}$. It is assumed that $w = w(\hat{\theta})$ does not vanish: the case where it is zero will be discussed below. Collecting terms of like order together, the integral is easily seen to be

$$we^{L}\int e^{\Sigma L_{ij}\theta_i\theta_j/2}\Big[1 \ + \ \Sigma W_i\theta_i+\tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k \ + \ \tfrac{1}{2}\Sigma W_{ij}\theta_i\theta_j$$
$$+(\Sigma W_i\theta_i)\tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k+R\Big]d\theta.$$

The orders of the terms in square brackets are respectively 1, $n^{-1/2}$, $n^{-1/2}$, n^{-1}, n^{-1} and n^{-1}, with the final term R not involving W or its derivatives. In subsequent calculations R will disappear, so we have not spelt it out.

The integrations all involve the moments of the multivariate normal distribution with density proportional to $\exp((1/2)\Sigma L_{ij}\theta_i\theta_j)$. The precision matrix has elements $-L_{ij}$. The elements of the matrix inverse to this are written σ_{ij}, forming a matrix Σ. It is well-known that for this distribution, $E(\theta_i) = 0$, $E(\theta_i\theta_j) = \sigma_{ij}$ and $E(\theta_i\theta_j\theta_k) = 0$. It is not perhaps so well-known that $E(\theta_i\theta_j\theta_k\theta_l) = \sigma_{ij}\sigma_{kl}+\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}$: see, for example, Anderson (1958: equation (26) of §2.6). The result of the integration is that

$$\int w(\theta)e^{L(\theta)}d\theta \sim we^{L}(2\pi)^{m/2}|\Sigma|^{1/2} \text{ x}$$
$$[1+\tfrac{1}{2}\Sigma W_{ij}\sigma_{ij}+\tfrac{1}{6}\Sigma L_{ijk}W_l(\sigma_{ij}\sigma_{kl}+\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}) \ + \ R^*], \qquad (2)$$

where R^* arises from R: the terms in square brackets being of order n^{-1} apart from the first. The second summation can be simplified since all three terms in it are equal. To see this, remember L_{ijk} is unaffected by permutation of its suffixes, so that permuting j and k in the first term gives $\Sigma L_{ikj}W_l\sigma_{ij}\sigma_{kl}$, and then interchanging the roles of j and k makes this equal to $\Sigma L_{ijk}W_l\sigma_{ik}\sigma_{jl}$, the second term: the third follows similarly.

This result is of interest in its own right but is complicated if the term R^* is spelt out. However, if we pass to a ratio (1), of such integrals with the same likelihood, the terms outside the square brackets in (2) cancel except for $w = w(\hat{\theta})$ and $v = v(\hat{\theta})$: and on expanding the ratio of the two terms in square brackets to order n^{-1}, R^*, which does not involve w, cancels with the *same* term R^* in the denominator, so that finally we have

$$\int w(\theta)e^{L(\theta)}d\theta/\int v(\theta)e^{L(\theta)}d\theta \ \sim$$
$$\frac{w}{v}\,[1 \ + \ \tfrac{1}{2}\Sigma(W_{ij}-V_{ij})\sigma_{ij} \ + \ \tfrac{1}{2}\Sigma L_{ijk}(W_l-V_l)\sigma_{ij}\sigma_{kl} \ + \ \ldots] \ .$$

(It has been assumed that $v \neq 0$.)

In the applications we have in mind, $v(\theta) = \pi(\theta)$, the prior distribution for θ , so that the denominator is the normalizing constant in Bayes theorem; and $w(\theta) = u(\theta)\pi(\theta)$, so that the ratio is $E[u(\theta)|x_1, x_2,\ldots,x_n]$. Simple calculation then shows that $W_{ij}-V_{ij} = u_{ij}/u+(u_i\pi_j+u_j\pi_i)/u\pi$ and $W_i-V_i = u_i/u$. If we write $\varrho(\theta) = \log\pi(\theta)$, a little more calculation finally gives

$$\int u(\theta)e^{L(\theta)+\rho(\theta)}d\theta / \int e^{L(\theta)+\rho(\theta)}d\theta \sim$$

$$u + \tfrac{1}{2}\,\Sigma(u_{ij} + 2u_i\varrho_j)\sigma_{ij} + \tfrac{1}{2}\Sigma L_{ijk}u_l\sigma_{ij}\sigma_{kl} \tag{3}$$

to order n^{-1}. This is our basic result. The first term is $0(1)$: the next are all $0(n^{-1})$ and will be referred to as *correction* terms. Notice that, because of the vanishing of all moments of odd orders for a multivariate normal distribution, the first term neglected is $0(n^{-2})$, not $0(n^{-3/2})$. Remember that on the right-hand side of (3) all functions are evaluated at the maximum likelihood value of θ, and that summation is over all suffixes and from 1 to m. One feature of immediate interest in (3) is that it does not involve the second derivatives of the prior, but that those of u do occur. Secondly, the prior is absent from the last correction term incorporating the third derivatives of the log-likelihood.

An alternative form is available for the final term in (3). Since the matrix of elements σ_{ij} is inverse to that of elements $-L_{ij}$, we have $\sum_k L_{ik}\sigma_{kl} = -\delta_{il}$. On differentiating with respect to θ_j, we obtain $\sum_k L_{ijk}\sigma_{kl} + \sum_k L_{ik}(\sigma_{kl})_j = 0$. Hence

$$\begin{aligned} \Sigma L_{ijk}u_l\sigma_{ij}\sigma_{kl} &= -\,\Sigma u_l\sigma_{ij}L_{ik}(\sigma_{kl})_j \\ &= \Sigma u_l\delta_{jk}(\sigma_{kl})_j, \text{ on summing over } i, \\ &= \sum_{k,l} u_l(\sigma_{kl})_k, \text{ on summing over } j. \end{aligned} \tag{4}$$

Although it appears simpler, we have found this form less convenient than that in (3) because it uses the algebraic inversion of $-L_{ij}$, in order to find $(\sigma_{kl})_k$, whereas the other only requires the numerical inversion in any application.

Another form of (3) may be obtained by writing $\Lambda(\theta) = L(\theta) + \varrho(\theta)$ which, apart from an additive constant, is the logarithm of the posterior distribution of θ, given $x_1, x_2, \ldots, x_n$. Then, instead of expanding about the maximum likelihood value, $\Lambda(\theta)$ may be expanded about its maximum, the posterior mode. Consideration of each of the individual steps in the argument that led to (3) shows that they apply when Λ replaces L. Effectively in (1), v becomes 1 and w, u. Hence

$$\begin{aligned} &\int u(\theta)e^{\Lambda(\theta)}d\theta / \int e^{\Lambda(\theta)}d\theta \sim \\ &u + \tfrac{1}{2}\Sigma\; u_{ij}\tau_{ij} + \tfrac{1}{2}\Sigma\Lambda_{ijk}u_l\tau_{ij}\tau_{kl}. \end{aligned} \tag{5}$$

Here $\tau_{ij} = -\Lambda^{ij}$ and all quantities are evaluated at the posterior mode, $\tilde{\theta}$, instead of the maximum likelihood value, $\hat{\theta}$. An alternative form is available using a result parallel to (4). (5) is simpler than (3), but the latter has the advantage of explicitly displaying the separate roles of u and π.

An important special case is where $u(\theta) = \theta_s$, $1 \leq s \leq m$, so that the ratio of integrals is the posterior mean of θ_s, $\bar{\theta}_s$ say. Since $u_s = 1$, $u_t = 0$ for $t \neq s$ and $u_{ij} = 0$, (5) immediately shows that the difference between the posterior mean and mode for θ_s is

$$\bar{\theta}_s - \widetilde{\theta_s} \sim \tfrac{1}{2} \sum_{i,j,k} \Lambda_{ijk}\tau_{ij}\tau_{ks}. \tag{6}$$

A similar result for the maximum likelihood values is, from (3),

$$\bar{\theta}_s - \hat{\theta}_s \sim \sum_i \varrho_i \sigma_{is} + \tfrac{1}{2} \sum_{i,j,k} L_{ijk}\sigma_{ij}\sigma_{ks}. \tag{7}$$

Similar calculations using $u(\theta) = \theta_s\theta_t$ give results which, when combined with (6), show that the posterior dispersion matrix for θ has elements τ_{ij} to $0(n^{-1})$, so requiring no correction from the corresponding modal values. Equivalent use of (3) shows that τ_{ij} may be replaced by σ_{ij} to the same order. Thus there is an order n^{-1} correction to the mean but not to the dispersion. An alternative way of obtaining this result is to use $u(\theta) = (\theta\text{-}\hat{\theta}_s)\,(\theta\text{-}\hat{\theta}_t)$, but this, and its first derivatives, vanish at $\hat{\theta}$, so that our expressions are no longer valid. The modifications necessary in this case are a little tedious, though straightforward in principle, and we therefore do not provide a general treatment but discuss special cases below: from these, the reader will be able to see how a general discussion would proceed.

The results simplify if the parameters are locally orthogonal: that is, if $L_{ij} = 0$, and hence $\sigma_{ij} = 0$, for all $i \neq j$. For example, the right-hand side of (3) reduces to

$$u + \tfrac{1}{2}\Sigma(u_{ii} + 2u_i\,\varrho_i)\sigma_{ii} + \tfrac{1}{2}\Sigma L_{iik}u_k\sigma_{ii}\sigma_{kk}\,,$$

and (7), for the mean, is simply

$$\bar{\theta}_s - \hat{\theta}_s \sim \varrho_s\sigma_{ss} + \tfrac{1}{2}\sum_i L_{iis}\sigma_{ii}\sigma_{ss}.$$

Local orthogonality can always be obtained by a locally orthogonal transformation of the parameter space at $\hat{\theta}$, or $\tilde{\theta}$.

Parameters are usually said to be orthogonal if $EL_{ij}(\theta) = 0$ for all $i \neq j$ and all θ; the expectation being over $x_1, x_2, \ldots, x_n$ (Jeffreys (1961)). Since L and its derivatives are sums of n terms, and hence of order n, they will, by the central limit theorem, differ from their expectations by a term of order $n^{-1/2}$. Hence replacement of L_{ij}, or L_{ijk}, by expectations will not, as many writers have noticed, affect the order of the correction terms, but it will affect the order of the terms discarded. As pointed out above, at the moment these are $0(n^{-2})$: if expectations are used they will rise to $0(n^{-3/2})$. Consequently the replacements should be used with care. Actually they violate the likelihood principle and are hence incoherent. In any case, as we try to show by example below, they are not needed in the numerical analysis of data. If they are used and the parameters are orthogonal, then further reductions occur: (3) reducing to

$$E(u) \sim u + \tfrac{1}{2}\Sigma(u_{ii}+2u_i\varrho_i)\sigma_{ii} + \tfrac{1}{2}\Sigma L_{iii}u_i\sigma_{ii}^2$$

and (7) to

$$\bar{\theta}_s - \hat{\theta}_s = \varrho_s\sigma_{ss} + \tfrac{1}{2}L_{sss}\sigma_{ss}^2 \, .$$

These reductions arise because the vanishing of the mixed second derivatives for all θ implies zero values for the mixed third derivatives.

An obvious advantage of some form of orthogonality is the diagonal form of the matrix of elements $-L_{ij}$ and the consequent ease of its inversion to give σ_{ij} : $\sigma_{ii} = -L_{ii}^{-1}$ and $\sigma_{ij} = 0$ for $i \neq j$.

But an additional advantage is the reduction in the numbers of third derivatives that have to be considered. These are $m(m+1)(m+2)/6$ if all distinct ones are needed; m^2 with local orthogonality; and m with full orthogonality. Full orthogonality cannot usually be achieved for $m > 3$.

2. UNIVARIATE APPLICATIONS

In this section the case is considered of a single parameter, written θ, hence $m = 1$. The notation L_{ijk} etc., for the derivatives is cumbersome, all suffixes necessarily being 1, and we revert to the more usual form in which L_3, for example, denotes the third derivative; previously L_{111}. The basic result (3) is that

$$E(u \,|\, x_1, x_2, \ldots, x_n) \sim u + \tfrac{1}{2}(u_2 + 2u_1\varrho_1)\sigma^2 + \tfrac{1}{2} L_3 u_1 \sigma^4 \qquad (8)$$

whereas in posterior mode form (5)

$$E(u \,|\, x_1, x_2, \ldots, x_n) \sim u + \tfrac{1}{2}u_2\tau^2 + \tfrac{1}{2}\Lambda_3 u_1 \tau^4 \, . \qquad (9)$$

The results for $u(\theta) = \theta$, giving the posterior mean $\bar{\theta}$, are

$$\bar{\theta} - \hat{\theta} = \varrho_1\sigma^2 + \tfrac{1}{2}L_3\sigma^4 \qquad (10)$$

and

$$\bar{\theta} - \tilde{\theta} = \tfrac{1}{2}\Lambda_3\tau^4 \, . \qquad (11)$$

It is clear from these formulas that there would be some advantage in arranging for L_3, or Λ_3, to be zero. This can be done in the case of the exponential family with a single sufficient statistic. In the canonical form, the density exp $[-x\theta - g(\theta) - h(x)]$ gives a log-likelihood $L(\theta) = -X\theta - ng(\theta)$ with $X = \Sigma x_i$ the sufficient statistic, and $L_i = -ng_i$ for $i > 1$, irrespective of the sample values. Suppose the parameterization is altered from θ to ϕ where $d\phi/d\theta = L_2^{1/3}$. Then $d\theta/d\phi = L_2^{-1/3}$ and $d^2\theta/d\phi^2 = -\frac{1}{3} L_3/L_2^{5/3}$. Consequently

$$\frac{d^3L}{d\phi^3} = L_3\left(\frac{d\theta}{d\phi}\right)^3 + 3L_2 \frac{d\theta}{d\phi} \frac{d^2\theta}{d\phi^2},$$

since $L_1 = 0$, vanishes. Hence a change from the canonical parameter θ to ϕ, where $d\phi/d\theta = L_2^{1/3}$, or $\phi = \int L_2^{1/3} (\theta) d\theta$ will make the final correction terms in (8) and (10) vanish. If the conjugate family is used for the prior to the exponential family, the same arguments will apply to Λ and, from (11), the posterior mean and mode will be the same to order n^{-1}.

As an example consider the gamma distribution with $p(x|\theta) \sim \theta^r e^{-\theta x}$, $g(\theta) = -r \log \theta$, so that $L_2 = ng_2 = nr\theta^{-2}$. Then $d\phi/d\theta = \theta^{-2/3}$, the constant being irrelevant, and hence $\phi = \theta^{1/3}$. With this parametric form, $L(\phi) = -X\phi^3 + 3nr \log \phi$ and $d^3L/d\phi^3 = 0$. This is the Wilson-Hilferty transformation, though applied to the parameter rather than the data.

It is a curious feature of the exponential family that in canonical form the derivatives of the log-likelihood above the first do not involve the data. An important effect of this is that the sampling theorist's violation of the likelihood principle in taking expectations over the sample space does no damage to the principle when applied to these higher derivatives: in particular, the large-sample variance, $\sigma^2 = -L_2^{-1}$, is unaffected. In general the derivatives will be data dependent and a transformation that makes L_3 zero is not available. An argument similar to that used above shows that a change to $\phi = \int \{EL(\theta)\}^{1/3} d\theta$ will make $EL_3 = 0$. As explained above, a change from L_i to EL_i will change the order of the neglected terms.

Transformations associated with L_3 are sometimes used to control skewness. It is therefore of interest to examine the third moment of θ. To do this we need the case $u(\theta) = (\theta - \hat{\theta})^3$ in the univariate form of (3). But $u = u(\hat{\theta})$ vanishes and our results do not apply. We therefore develop an expansion analogous to (2) valid when $w = 0$, confining ourselves to the univariate case. Multivariate extensions follow straightforwardly. Suppose that the first non-vanishing derivative of w at $\hat{\theta}$ is the s^{th}, $s > 0$. The derivatives will be written w_s etc. Then as in the derivation of (2)

$$\int w(\theta)e^{L(\theta)}d\theta = \int \left[\frac{1}{s!} w_s\theta^s + \frac{1}{(s+1)!} w_{s+1}\theta^{s+1} + \dots \; e^{L(\theta)}\right] d\theta$$

$$= \frac{w_s e^L}{s!}\left[\int \quad \theta^s + \frac{W_{s+1}}{s+1} \quad \theta^{s+1} + \ldots \right] e^{(1/2)L_2\theta^2}\left[1 + \frac{1}{6} L_3\theta^3 + 0(n^{-1})\right] d\theta \quad .$$

There are two cases according as s is odd or even. In the even case the leading term is $w_s e^L \sqrt{2\pi}\sigma E(\theta^s)/s!$. In the odd case, two terms need consideration and we have

$$\left\{ w_s e^L \sqrt{2}\pi \ \sigma/s! \right\} \left\{ \frac{W_{s+1}}{s+1} E(\theta^{s+1}) + \frac{1}{6} L_3 E(\theta^{s+3}) \right\} .$$

We next need to combine the results for the numerator, for w, with those for the denominator, for v. In applications $v = e^\rho$ is the prior. We shall suppose that this nowhere vanishes, in line with the principle that a Bayesian should never assign zero probability to any value, because to do so would commit him to zero irrespective of any data. This being so, the dominant term in the denominator is $ve^L \sqrt{2}\pi \ \sigma$ giving

$$\frac{\int w(\theta) e^{L(\theta)} d\theta}{\int v(\theta) e^{L(\theta)} d\theta} \quad \sim \quad \begin{array}{ll} w_s E(\theta^s)/s!v & \text{s even} \\ \{w_{s+1}E(\theta^{s+1})/(s+1) + w_s L_3 E(\theta^{s+3})/6\}/s!v, & \text{s odd} \end{array} \tag{12}$$

of order $n^{-s/2}$ for s even, and $n^{-(s+1)/2}$ for s odd.

To obtain the posterior moments we write $w(\theta) = (\theta\text{-}\hat{\theta})^s e^{\rho(\theta)}$ and $v(\theta) = e^{\rho(\theta)}$. For $s = 2$, we immediately obtain σ^2, a result discussed in the general development. The third moment is a little more complicated. The first non-vanishing derivative is $w_3 = 3!e^\rho$ and $w_4 = 4!e^\rho \varrho_1$. Hence

$$E(\theta\text{-}\hat{\theta})^3 \sim E(\theta^4)\varrho_1 + \tfrac{1}{6} L_3 E(\theta^6) = 3\sigma^4\varrho_1 + \tfrac{5}{2}\sigma^6 L_3 ,$$

of order n^{-2}. The fourth moment is easily seen to $3\sigma^4$. To obtain the moments about the mean write

$$\begin{aligned} E(\theta\text{-}\theta)^3 &= E(\theta\text{-}\hat{\theta} + \hat{\theta}\text{-}\bar{\theta})^3 \\ &= E(\theta\text{-}\hat{\theta})^3 + 3E(\theta\text{-}\hat{\theta})^2(\hat{\theta}\text{-}\bar{\theta}) + 2(\bar{\theta}\text{-}\hat{\theta})^3 \\ &= 3\sigma^4\varrho_1 + \tfrac{5}{2}\sigma^6 L_3 + 3\sigma^2\{\text{-}\varrho_1\sigma^2\text{-}\tfrac{1}{2}L_3\sigma^4\} + 0(n^{-3}) \\ &= L_3\sigma^6 + 0(n^{-3}) , \end{aligned} \tag{13}$$

also of order n^{-2}. Similarly $E(\theta\text{-}\bar{\theta})^4 = 3\sigma^4 + 0(n^{-3})$.

It is interesting to see that neither of these involve the prior distribution and that the fourth moment is that predicted by assuming a normal distribution for θ. Skewness would seem to be a more important feature of posterior distributions than kurtosis.

We now consider some examples, excluding the exponential family which, as we have seen, is somewhat unusual. The first is a sample from a t-distribution of unknown location, but known spread and degrees of freedom: the sample size is $n = 7$, and the degrees of freedom are 5. The log-density for x is therefore $C - 3\log\{1 + (x-\theta)^2/5\}$. With true value $\theta = 0$ the sample is:

$$-1.0\ ,\ -0.3\ ,\ -0.1\ ,\ +0.4\ ,\ +0.9\ ,\ +1.6\ ,\ +3.0\ . \qquad (14)$$

The upper 1% point of t_5 is 3.36, so that the last value is unusual and almost deserves the title of an outlier: it would certainly be an outlier for the corresponding normal distribution with $\nu = \infty$. Table 1 gives the value of the log-likelihood and its first three differences around the maximum value. Interpolation gives $\hat{\theta} = 0.4954$, and L_2 at this value is -4.923 from the second differences. Hence $\sigma^2 = 0.2031$ and $\sigma = 0.451$. Simple calculation for the t-distribution shows that $E(L_2) = -n(\nu+1)/(\nu+3)$, giving here an average value of σ^2 of 0.190, slightly less than the sample value obtained here, so that the sample is a little less informative than an average one. Assuming $\varrho_1 = 0$ corresponding to a *flat* prior at $\hat{\theta}$, the correction for $\hat{\theta}$, equation (10), is $\frac{1}{2}L_3\sigma^4$. With $L_3 = 0.724$, by interpolation in the third differences, the correction to $\hat{\theta}$ is 0.0149, so that $\bar{\theta} = 0.5103$. The correction is negligible in comparison with the standard deviation. Notice, however, that $\bar{\theta}$ is very different from the arithmetic mean of the sample at 0.643, which is unduly swayed by the outlier. The correction for the prior need not be negligible. Suppose that $\pi(\theta)$ is such that θ/K is t_ν: that is, centred at the true value of $\theta = 0$ but with variance $K^2\nu/(\nu-2)$ for $\nu>2$. It is easy to establish that $\varrho_1 = -\hat{\theta}(\nu+1)/(K\nu+\hat{\theta}^2)$. For example with $K = 1$ and $\nu = 5$, roughly making the prior equivalent to an extra value at 0, the correction term $\varrho_1\sigma^2 = -0.115$, giving $\bar{\theta} = 0.395$. Increasing K to 2, making one initially less sure about θ, gives a correction of -0.0589 and $\bar{\theta} = 0.451$.

The general form of the correction to $\bar{\theta}$ due to the prior, $\varrho_1\sigma^2$, is best appreciated by the following heuristic argument. In a quadratic approximation to the logarithm, ϱ, of the prior, it can be written $-(\theta-\theta_0)^2/2\sigma_0^2$ where θ_0 is the prior mean (or mode) and σ_0^2 the prior variance. Its derivative at $\hat{\theta}$ is $-(\hat{\theta}-\theta_0)/\sigma_0^2$. Hence, ignoring the other correction term $\frac{1}{2}L_3\sigma^4$,

$$\bar{\theta} \sim \hat{\theta} + \sigma^2(\theta_0-\hat{\theta})/\sigma_0^2 \sim \{\hat{\theta}/\sigma^2 + \theta_0/\sigma_0^2\}\{\sigma^{-2} + \sigma_0^{-2}\}^{-1}$$

to order n^{-1}. This is the usual weighted average of $\hat{\theta}$ and θ_0 with weights equal to their precisions.

Table 1. $L(\theta) = -3\Sigma \log \{1 + (x_i - \theta)^2/5\}$ and its differences for the sample (14).

θ	L	L_1	L_2	L_3
0.475	-4.869 031 221			
		+ 759 378		
485	8 271 843		- 492 996	
		+ 266 382		+ 705
495	8 005 461		- 492 291	
		- 225 909		+ 741
505	8 231 370		- 491 550	
		- 717 459		
515	8 948 829			

Returning to the sample (14), consider what happens when the outlier *increases* from 3.0 to 4.0. The maximum likelihood value *decreases* from 0.4954 to 0.4714, showing that less attention is paid to the extreme value. The variance σ^2 increases slightly from 0.2031 to 0.2058 and L_3 grows from 0.724 to 0.825, with the result that the correction $\frac{1}{2} L_3\sigma^4$ changes from 0.0149 to 0.0175. Hence $\bar{\theta} = 0.4889$. There is still little skewness in the posterior distribution. This is a result of the symmetry in the original density. To exhibit a substantial correction term it is necessary to take a skew density for θ, but before doing this there is one more remark that is worth making about the t-distribution. It can happen that the log-likelihood has two local maxima, in which case each will give a contribution in the asymptotic expansions.

To exhibit a skew distribution giving a larger correction term, consider a sample, again of size 7, from an F-distribution of unknown scale. We have taken a case with degrees of freedom, $\nu_1 = 4$ and $\nu_2 = 8$, giving a density proportional to $\theta^2 x/(8 + 4\theta x)^6$. With true value $\theta = 1$, the sample is

$$.3 \quad .5 \quad .8 \quad 1.2 \quad 1.4 \quad 2.5 \quad 4.0 \qquad (15)$$

Table 2 gives the value of the log-likelihood and its first three differences around the maximum value. Interpolation gives $\hat{\theta} = 0.8110$, and L_3 at this value is -12.399. Hence $\sigma^2 = 0.08065$ and $\sigma = 0.2840$. The value of L_3 is 42.0, so that with a *flat* prior the correction term, $\frac{1}{2} L_3\sigma^4$, is 0.1366. The result of applying this is that $\hat{\theta}$ at 0.8110 is increased to $\bar{\theta}$ at 0.9476, and the correction is almost one half the standard deviation. The posterior distribution is skew to the right, the mean exceeding the mode. The third moment, equation (13), is 0.022 and the fourth 0.0195.

Notice that in doing numerical work with the results we have not used the differential calculus to evaluate $L_i(\theta)$ and then inserted the numerical values for θ (and $x_1, x_2, \ldots, x_n$): instead $L\ (\theta)$ has been evaluated for a range of values of θ and the differences used to obtain $L_i(\hat{\theta})$. This reduces substantially the amount of work, both analytic and numeric, and has the advantage of displaying the form of the log-likelihood where it is large.

One other application of the basic results, (8) and (9), that merits attention is to obtain the predictive distribution. Let y be an, as yet unobserved, value whose density, given θ, is $q(y|\theta)$. Often q will be p, the density leading to L, and y, equivalently, x_{n+1}, but the results are general. Then, given $x_1, x_2, \ldots x_n$, the density of y is given by (8) with $u(\theta) = q(y|\theta)$. The leading term is $q(y|\hat{\theta})$ and the correction allows for the uncertainty about θ. Moments for the predictive distribution are available if the moments of q are expressible as functions of θ. A related use is in empirical Bayes problems which have been treated by Deely and Lindley (1979).

Dunsmore (1976) writes the predictive distribution as $\int q\ (y\ |\theta)\ p\ (\theta|x_1, \ldots x_n)\ d\theta$ and uses asymptotic results for the posterior distribution to obtain approximations for the univariate case that are similar to those in the present paper. The main differences are that Dunsmore's asymptotic results use $\hat{\theta}$, not $\tilde{\theta}$; σ, not τ.

Table 2. $L(\theta) = 14 \log \theta - 6\Sigma \log (8 + 4\theta x_i)$ and its differences for the sample (15)

θ	L	L_1	L_2	L_3
0.79	-108.814 597 6			
		+ 20 320		
.80	2 565 6		- 12 868	
		+ 7 452		+ 428
.81	1 820 4		- 12 440	
		- 4 988		+ 414
.82	2 319 2		- 12 026	
		- 17 014		
.83	4 020 6			

3. BIVARIATE APPLICATIONS

With two parameters, θ_1 and θ_2, there are only 4 third derivatives and the notation L_{30} etc., in lieu of L_{111} etc., seems preferable. The correction term $\frac{1}{2} L_{ijk}u_l\sigma_{ij}\sigma_{kl}$ (equation (3)) becomes one half

$$L_{30}\{u_1\sigma_{11}^2 + u_2\sigma_{11}\sigma_{12}\} + L_{21}\{3u_1\sigma_{11}\sigma_{12} + u_2(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2)\}$$
$$+ L_{12}\{u_1(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2) + 3u_2\sigma_{12}\sigma_{22}\} + L_{03}\{u_1\sigma_{12}\sigma_{22} + u_2\sigma_{22}^2\}.$$

An alternative form uses $U_k = \sum_l u_l\sigma_{kl}$. The whole expression (3) is then

$$u + \tfrac{1}{2}\,\Sigma u_{ij}\sigma_{ij} + \Sigma U_j\varrho_j + \tfrac{1}{2}L_{30}\sigma_{11}U_1 + \tfrac{1}{2}L_{21}(2\sigma_{12}U_1 + \sigma_{11}U_2)$$
$$+ \tfrac{1}{2}L_{12}(\sigma_{22}U_1 + 2\sigma_{12}U_2) + \tfrac{1}{2}L_{03}\sigma_{22}U_2\,. \qquad (16)$$

These expressions are complicated but well-adapted for numerical work. With L, u and ϱ evaluated, as in §2, on a grid of values of θ_1, θ_2 about $\hat{\theta}$, differences may again be used to form the derivatives, the matrix of minus the second derivatives inverted to give σ_{ij}, and then easy arithmetic gives the value of (16).

For the posterior mean of θ_1, say, we have $u(\theta) = \theta_1$ and hence $u_1 = 1$, $u_2 = 0$ and $u_{ij} = 0$ for all i, j. Hence (also from (7))

$$\bar{\theta}_1 - \hat{\theta}_1' = \varrho_1\sigma_{11} + \varrho_2\sigma_{21} + \tfrac{1}{2}L_{30}\sigma_{11}^2 + \tfrac{3}{2}L_{21}\sigma_{11}\sigma_{12} + \tfrac{1}{2}L_{12}(\sigma_{11}\sigma_{22}+2\sigma_{12}^2)$$
$$+ \tfrac{1}{2}L_{03}\sigma_{12}\sigma_{22}\,. \qquad (17)$$

We illustrate these results for the analysis of a one-way table. This example differs from those studied in §2 in two respects. First, we operate directly with the posterior distribution rather than the likelihood. Second, the case is more interesting because the modal values (and the maximum likelihood ones) are known to be misleading, so that evaluation of means by methods that avoids tedious bivariate integrations may be of real value in the appreciation of data from such a table. There are possibilities of extensions to more elaborate analyses of variance.

The data x_{ij} $(i = 1, 2,\ldots m;\ j = 1, 2,\ldots n)$ are, given $\{\mu_i\}$ and σ^2, independent with $x_{ij} \sim N(\mu_i, \sigma^2)$ that is, m groups with n observations in each group. For the prior density of the μ's, we suppose them i.i.d. $N(\mu, \tau^2)$, and independent of σ^2. This distribution can be thought of as part of the likelihood, in which case we have a Model II, rather than Model I, situation. Finally the distributions for σ^2, τ^2 and μ are supposed independent with $\nu_1\lambda_1/\sigma^2 \sim \chi^2_{\nu 1}$, $\nu_2\lambda_2/\tau^2 \sim \chi^2_{\nu 2}$ and μ uniform. The prior for σ^2 and τ^2 has not been expressed in the mathematically more convenient, conjugate form in terms of $\sigma^2 + n\tau^2$ since we believe that a prior depending on the sample size is unrealistic. Tedious calculations show that the joint posterior distribution of σ^2 and τ^2 has logarithm equal to a constant plus

$$-\tfrac{1}{2}\,(N\text{-}m+\nu_1+2)\log\sigma^2-\tfrac{1}{2}\,(\nu_2+2)\log\tau^2-\tfrac{1}{2}\,(m\text{-}1)\log\,(n\tau^2+\sigma^2)$$
$$-\,nT^2/2(n\tau^2+\sigma^2)-\nu_2\,\lambda_2/2\tau^2-(S^2+\nu_1\,\lambda_1)\,/2\sigma^2\,. \qquad (18)$$

In the notation used above this is $\Lambda(\theta_1,\ \theta_2) = \Lambda(\sigma^2,\ \tau^2)$. The unexplained notation is $N = nm$, $nT^2 = n\Sigma(x_{i\cdot}\ \text{-}x_{\cdot\cdot})^2$ and $S^2 = \Sigma(x_{ij} - x_{i\cdot})^2$, the between and within sums of squares. The modal values for σ^2 and τ^2 are easily found from (18), and these can be used to find approximate posterior means for the μ_i, which are weighted averages of $x_{i\cdot}$ and $x_{\cdot\cdot}$ with weights dependent on these modes. However the distribution (18) is skew and the preferred means may differ from their modes.

We illustrate using a numerical example with $\mu = 0$, $\sigma^2 = \tau^2 = 1$, having the hyperparameters, $\nu_1 = \nu_2 = 4$, $\lambda_1 = \lambda_2 = 1$, and with data $S^2 = 37.34372$, $T^2 = 4.556774$, for $m = 8$ and $n = 5$. Notice that the prior information about τ^2, with 4 degrees of freedom, is comparable with the information from the data, through T^2, having 7. The prior expectation of τ^2 is $\lambda_2\nu_2\,/(\nu_2 - 2) = 2$, and the standard deviation is infinite, but the mode is at $\lambda_2\nu_2\ /\ (\nu_2 + 2) = 2/3$. This does not seem unrealistic in some applications, though each case must be decided in the light of practical experience. Table 3 gives the value of $\Lambda(\sigma^2,\ \tau^2)$, equation (18), for a grid of values of σ^2 and τ^2 around $\sigma^2 = 1.1$ and $\tau^2 = 0.6$. Interpolation shows that the modal values are $\hat{\sigma}^2 = 1.08$ and $\hat{\tau}^2 = 0.59$. (The estimates obtained by equating the values of S^2 and T^2 to their expectations are for σ^2, 1.17 and for τ^2, 0.42.) To evaluate the correction terms, the differences are used to obtain the derivatives. Thus $\Lambda_{20} = \{(-0.2267 + 0.1213) - (-0.1213 + 0.1727)\}\ /0.01 = -15.60$. Similarly $\Lambda_{02} = -13.75$ and $\Lambda_{11} = +0.63$. The small value of this mixed, second derivative, in

Table 3. Values of $30 + \Lambda\ (\sigma^2,\ \tau^2)$, equation (18), for the values given in the text. All entries preceded by -0.

σ^2 \ τ^2	0.4	0.5	0.6	0.7	0.8
0.9	8728	5212	4522	5287	6847
1.0	5654	2319	1727	2548	4139
1.1	4894	1718	1213	2083	3701
1.2	5728	2693	2267	3181	4823
1.3	7683	4773	4417	5370	7033

comparison with the larger values of the unmixed ones, means that σ^2 and τ^2 are almost locally orthogonal and we will treat them as such in what follows. Extending to the third derivatives $\Lambda_{30} = 59.$, $\Lambda_{03} = 97.$ and Λ_{21} and Λ_{12} are virtually zero. Hence we may use the two univariate formulae for $E(\sigma^2)$ and $E(\tau^2)$ separately. For σ^2 the mode is 1.08 and the variance is $(-\Lambda_{20})^{-1} = 0.0641$, with standard deviation 0.253. The correction term is $\frac{1}{2}59 \times (.0641)^2 = 0.12$, raising σ^2 to 1.20 as the posterior mean. For τ^2 the mode is 0.59 and the variance is $(-\Lambda_{02})^{-1} = 0.0727$, with standard deviation 0.270. The correction term is $\frac{1}{2}97 \times (.0727)^2 = 0.26$ raising τ^2 to 0.85 as the posterior mean. Notice that the two correction terms are both positive, the means exceeding the modes, and that they are comparable with the standard deviations: for σ^2 the correction is about half the standard deviation, whilst for τ^2 they are about equal. Hence the term of order n^{-1} (for the correction) is comparable with that of order $n^{-1/2}$ (for the standard deviation). The claim sometimes made that terms of smaller order may be neglected in maximum likelihood (or maximum posterior) theory may not be true for some skew distributions. Notice, that because of the large, unmixed, third derivatives, the skewness in both parameters is quite large. It is interesting that the standard deviations of σ^2 and τ^2 are about equal (0.25 and 0.27 respectively) whereas one might have expected σ^2 to be better determined than τ^2.

4. DISCUSSION

The analytic results of this paper enable one to calculate the difference between the mean and mode of certain distributions as far as the dominant term of order n^{-1} in the sample size n. The difference involves the second and third derivatives of the log-likelihood at the mode and is in a form suitable for numerical calculation. Such calculations tentatively suggest that the differences are appreciable even in comparison with the standard deviations, but much more needs to be done before these claims can be substantiated.

The method used here is essentially that of steepest descents. This tool has been used by Barndorff-Nielsen and Cox (1979) to obtain sampling distributions that enable inferences to be made about one parameter, θ_m say, in the presence of nuisance parameters $\theta_1, \theta_2, \ldots \theta_{m-1}$. It will be of interest to see how these sampling-theory approximations compare with the Bayesian results of this paper.

REFERENCES

ANDERSON, T.W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley

BARNDORFF-NIELSEN, O. and COX, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. Roy. Statist. Soc B*, **41**, 279-312

DEELY, J.J. and LINDLEY, D.V. (1979). Bayes empirical Bayes. *Tech. Report.* University of Canterbury.

DUNSMORE, I.R. (1976) Asymptotic prediction analysis. *Biometrika,* **63**, 627-630.

JEFFREYS, H. (1961) *Theory of probability.* Oxford: Clarendon Press.

LINDLEY, D.V. (1961) The use of prior probability distributions in statistical inference and decisions. *Proc 4th Berkeley Symp.* **1**, 453-468.