Applications of a Method for the Efficient Computation of Posterior Distributions

J. C. Naylor; A. F. M. Smith

Applied Statistics, Vol. 31, No. 3 (1982), 214-225.

Stable URL:
http://links.jstor.org/sici?sici=0035-9254%281982%2931%3A3%3C214%3AAOAMFT%3E2.0.CO%3B2-9

Applied Statistics is currently published by Royal Statistical Society.

Appl. Statist. (1982),
31, *No.* 3, *pp.* 214–225

Applications of a Method for the Efficient Computation of Posterior Distributions

By J. C. NAYLOR and A. F. M. SMITH

Derby Lonsdale College, Derby, UK *University of Nottingham, Nottingham, UK*

[Received September 1980. Final revision March 1982]

SUMMARY

For routine implementation with complicated likelihood functions, statistical procedures based on posterior distributions, or integrated likelihoods, require an efficient approach to numerical integration. In this paper we shall outline a numerical integration method using Gaussian quadrature which leads to efficient calculation of posterior densities for a rather wide range of problems. Several illustrative examples are provided, including a re-analysis of the Stanford heart transplant data. Among other things, these examples reveal that inferences based upon integrated likelihoods may differ substantially from those based on maximized likelihoods and the standard normal form of approximation.

Keywords: BAYESIAN INFERENCE; POSTERIOR DISTRIBUTIONS; INTEGRATED LIKELIHOOD; NUMERICAL INTEGRATION; GAUSS–HERMITE FORMULAE; HEART TRANSPLANT DATA; SURVIVAL DATA; CENSORING

1. INTRODUCTION

THE implementation of Bayesian inference procedures can be made to appear deceptively simple. Given a likelihood function $l(\mathbf{x};\ \boldsymbol{\theta})$ and prior density $p(\boldsymbol{\theta})$, we simply apply Bayes' theorem to obtain the joint posterior density for the parameter vector $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{l(\mathbf{x};\ \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int l(\mathbf{x};\ \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \tag{1}$$

If we are interested in the marginal density of $\boldsymbol{\theta}_I$, where $I \subseteq (1, \ldots, k)$ are the subscripts of the components of interest, we then simply integrate over $\boldsymbol{\theta}_{I'}$, where I' is the complement of I in $(1, \ldots, k)$, to obtain

$$p(\boldsymbol{\theta}_I \mid \mathbf{x}) = \int p(\boldsymbol{\theta} \mid \mathbf{x})\, d\boldsymbol{\theta}_{I'}. \tag{2}$$

In both (1) and (2) the integrations are to be understood as having the appropriate dimensions and ranges.

In cases where $l(\mathbf{x};\ \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ belong to the exponential family and the corresponding conjugate family, respectively, it is well known that the integrations required in (1) and (2) can be performed analytically, leading to simple, tractable forms of analysis (see, for example, DeGroot, 1970). In general, however, the forms of likelihoods and/or prior densities do not permit such a tractable analysis and the required integrations must either be performed numerically, or analytic approximations found.

Reilly (1976) presented a straightforward approach to this integration problem, evaluating the function at a large number of grid points, and replacing the integrations by appropriate summations in order to find the normalizing constants (i.e. the reciprocal of the denominator

 0035–9254/82/31214 $2.00

of (1)) and marginal densities. While this approach might be adequate for relatively straightforward likelihoods (of the type illustrated in Reilly's paper) it is easy to find important classes of problems for which large-scale function evaluation of this kind would be prohibitively expensive in computer time. Examples are given in Section 3 involving complications due to censoring and high posterior correlations among some parameters.

In Section 2 we outline an approach to numerical integration which has proved highly efficient and successful for analysing a number of problems. Various aspects of this approach will be illustrated in Section 3 by re-analysing several data sets which have already appeared in the literature—including the Stanford heart transplant data (see, for example, Turnbull *et al.*, 1974). These re-analyses serve to underline the message that approximate analyses based on maximum likelihood estimates and ranges of uncertainty obtained by appealing to "asymptotic" normality can be rather misleading—as judged in relation to exactly (i.e. efficiently numerically) calculated posterior densities based on "non-informative" priors. The posterior densities are often *not* normal in appearance and exact posterior correlations and variances are often substantially different from the values obtained using the matrix of second-derivatives of the log-likelihood.

In Section 4 we give a short summary of the kinds of problems that may be analysed within the Bayesian framework using this numerical integration approach. Some general remarks follow in Section 5.

2. Efficient Computation of Posterior Distributions

2.1. *Forms of Integral Required*

For given likelihood and prior, $l(\mathbf{x};\ \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$, let us define the operator S_I such that

$$S_I(q(\boldsymbol{\theta})) = \int q(\boldsymbol{\theta})\, l(\mathbf{x};\ \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{I'}, \tag{3}$$

writing S in place of S_I if $I' = (1, \ldots, k)$, and leaving the dimension of integration implicitly defined and the range understood to be the full parameter space. Then all the integrals required for calculating and summarizing (1) and (2) are special cases of (3) for particular choices of I and $q(\boldsymbol{\theta})$.

For example, $S(1)$ gives the denominator of (1) and hence the normalizing constant of the joint posterior density. Moreover, if $l(\mathbf{x};\ \boldsymbol{\theta})$ is taken to be the density of the full sampling distribution, $f(\mathbf{x} \mid \boldsymbol{\theta})$, then $S(1)$ gives $f(\mathbf{x})$, the marginal distribution of $\mathbf{x}$. By taking $q(\boldsymbol{\theta})$ equal to θ_i, θ_j, $\theta_i \theta_j$, respectively, and forming the obvious expressions involving $S(1)$, $S(\theta_i)$, $S(\theta_j)$ and $S(\theta_i \theta_j)$, posterior means, variances and covariances are easily obtained. If $q(\boldsymbol{\theta})$ is taken to be $f(\mathbf{y} \mid \boldsymbol{\theta})$, the density of the sampling distribution of future data $\mathbf{y}$, then $S(f(\mathbf{y} \mid \boldsymbol{\theta}))/S(1)$ is the predictive density for $\mathbf{y}$ given $\mathbf{x}$.

2.2. *Basic Assumptions and Method of Integration*

A very wide range of problems are such that $l(\mathbf{x};\ \boldsymbol{\theta})$, $p(\boldsymbol{\theta})$ satisfy regularity conditions which ensure the asymptotic posterior normality of $p(\boldsymbol{\theta} \mid \mathbf{x})$. From this and other considerations, particularly if the components of $\boldsymbol{\theta}$ are represented sensibly (for example, possibly using $\log(\sigma)$ in place of σ), it is reasonable to suppose that, at least for moderate samples, the forms $p(\boldsymbol{\theta} \mid \mathbf{x})$ may be adequately approximated by the product of a k-dimensional multivariate normal probability density and a polynomial in $(\theta_1, \ldots, \theta_k)$.

Now it is well known that univariate integrals of the form

$$\int_{-\infty}^{\infty} e^{-t^2} f(t)\, dt \tag{4}$$

may be approximated with a Gaussian-type formula

$$\sum_{i=1}^{n} \omega_i f(t_i), \tag{5}$$

where

$$\omega_i = \frac{2^{n-1}\, n!\sqrt{\pi}}{n^2[H_{n-1}(t_i)]^2} \tag{6}$$

and t_i is the ith zero of the Hermite polynomial $H_n(t)$; see, for example, Davis and Rabinowitz (1967). Moreover, the remainder function has the form

$$R_n = \frac{n!\sqrt{\pi}}{2^n(2n)!} f^{(2n)}(\xi), \tag{7}$$

for some ξ, so that if $f(t)$ is actually a polynomial of degree $2n-1$, the remainder will be zero and the approximation exact.

If $h(t)$ is a suitably regular function and

$$g(t) = h(t)(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\},$$

simple manipulation shows that

$$\int_{-\infty}^{\infty} g(t)\, dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} h(\mu + \sqrt{2}\sigma t)\, e^{-t^2}\, dt, \tag{8}$$

which has the form (4). Utilizing the Gauss–Hermite formula, we therefore obtain

$$\begin{aligned} \int_{-\infty}^{\infty} g(t)\, dt &\simeq \sum_{i=1}^{n} \frac{1}{\sqrt{\pi}} \omega_i h(\mu + \sqrt{2}\sigma t_i) \\ &= \sum_{i=1}^{n} \sqrt{2}\sigma \exp(t_i^2)\, \omega_i g(\mu + \sqrt{2}\sigma t_i) \\ &= \sum_{i=1}^{n} m_i g(z_i), \end{aligned} \tag{9}$$

where

$$m_i = \omega_i \exp(t_i^2)\sqrt{2}\sigma, \quad z_i = \mu + \sqrt{2}\sigma t_i. \tag{10}$$

Tables of t_i, w_i and $w_i \exp(t_i^2)$ are available for $n = 1(1)20$ (Salzer, *et al.*, 1952) and the error term will be small if $h(z)$ is (approximately) a polynomial.

Letting $t = \theta$, we may apply (9) to integrals of the form (3) with $k = 1$ and $q(\theta) = 1$ provided that we can find a normal density which, when multiplied by a polynomial in θ, gives an adequate approximation to $p(\theta \mid \mathbf{x})$, or, equivalently, to $l(\mathbf{x};\, \theta)\, p(\theta)$, as in (3). One possible choice is to use a normal density with the posterior mean and variance of θ substituted for μ and σ^2, respectively. Approximate values for these posterior moments may be available by maximum likelihood, prior knowledge of the situation, or even a crude "informed guess' based on a few evaluations of the likelihood function.

Equation (9) may also be applied to (3) with $q(\theta)$ chosen to give the posterior mean or variance, since such choices merely multiply the integrand by a polynomial in θ of degree at most two. For an n-point rule, we may expect, from (7), that this application of (9) will be satisfactory if $p(\theta \mid \mathbf{x})$ is well approximated by the product of a normal density and a polynomial in θ of degree at most $2n-3$. We use an iterative method in which the approximations to the posterior mean and variance found in this way on any particular iteration are used to construct the grid (z_i) and weights (m_i) for the next. This may be applied initially to quite small grid sizes (down to $n = 3$) and then the grid sizes gradually increased until satisfactory

convergence is observed, in the sense that stable answers are obtained at each grid size, together with adequate agreement between answers derived from successive grids. A selection of such iterative results is given and discussed for a three-parameter problem in Section 3.3.

For problems with more than one parameter, we use a cartesian product rule based on (9), which may be written in the form

$$\int \dots \int g(t_1, \dots, t_k)\, dt_1 \dots dt_k \simeq \sum_{i_k} m_{i_k}^{(k)} \dots \sum_{i_2} m_{i_2}^{(2)} \sum_{i_1} m_{i_1}^{(1)}\, g(z_{i_1}^{(1)}, \dots, z_{i_k}^{(k)}), \tag{11}$$

where $m_{i_j}^{(j)}$, $z_{i_j}^{(j)}$ are found using (10), with the marginal posterior mean and variance of θ_j substituted for μ and σ^2. One method of deriving this rule is to consider the use of (9) for obtaining values of a marginal density in a two-parameter problem. Point values of such a density are defined by univariate integrals of the form

$$p(\theta_2 \mid \mathbf{x}) = \int p(\theta_1, \theta_2 \mid \mathbf{x})\, d\theta_1,$$

which may be evaluated by the method described for the one-parameter case. However, the values of μ, σ needed for (10) correspond to *conditional* posterior moments of θ_1 and so require separate iterative determinations for each point value θ_2. The more tractable form appearing in (11) is obtained if these conditional moments are replaced by *marginal* moments. In a similar way, the marginal density of θ_2 may be integrated using (9) as before (repeated applications of this leading to the complete rule defined by (11)). Clearly, the order of the weighted summations in (11) is unimportant and so all the marginal densities for a k-parameter problem are available as marginal weighted sums.

The justification of the use of marginal values for μ, σ as described above involves additional assumptions of posterior independence and homoscedasticity, but, in many problems, there are high posterior correlations between some elements of $\boldsymbol{\theta}$. We overcome this difficulty by transforming the component parameters in $\boldsymbol{\theta}$ to a new orthogonal set of parameters. For example, for the three parameter case, we may form

$$\begin{aligned} \theta'_1 &= \theta_1, \\ \theta'_2 &= \beta\theta'_1 + \theta_2, \\ \theta'_3 &= \alpha_1\, \theta'_1 + \alpha_2\, \theta'_2 + \theta_3, \end{aligned}$$

where α_1, α_2 and β are chosen to make θ'_2 orthogonal to θ'_1, and θ'_3 orthogonal to the plane $\theta'_1\, \theta'_2$. These transformations have unit Jacobian and are completely determined by the posterior covariance matrix (so they can be recalculated at each stage of the iterative determination of the mean vector and covariance matrix).

We note that in the approach given above θ_1 is not transformed and so its marginal distribution is readily available as before. In a similar way, marginals for θ_2 and θ_3 may be obtained by omitting all or part of the transformation after a sufficient number of iterations for convergent values of the mean vector and covariance matrix to have been obtained. Alternatively, the components of $\boldsymbol{\theta}$ could be reordered so as to permit different sets of parameter transformations and hence different components appearing as the untransformed θ_1. Comparative results using both these techniques are mentioned in the example of Section 3.3.

The use of transformations of the parameter space as described above may be viewed as an attempt to find a parameter set having posterior spherical symmetry. If such a set could be found, it might then be argued that more economical rules which made better use of this condition could be used (see, for example, Stroud, 1971, Section 8.9). We have considered such rules for the initial stages of iterative techniques, but there are two serious disadvantages in their general use.

The first problem is that no marginal density values would be readily available. Although we have ourselves, in Section 3, summarized many of our results simply in terms of posterior moments, this is done merely for comparison with currently available results. This practice is not a real alternative to the full presentation of marginal or joint posterior densities.

The second problem relating to the use of spherical rules is that they depend on the corresponding approximation of "multivariate normal density multiplied by a polynomial in $\boldsymbol{\theta}$" being reasonable for *all* points in the parameter space. The success of the method presented here depends only on the assumption of the adequacy of the approximation for the conditional distribution in a specific set of directions in the parameter space. This suggests that certain choices of parameter orderings (and corresponding transformations) may be better than others. Such an effect has been observed and is discussed for the example of Section 3.3.

Part of the efficiency of our method stems from the fact that the same grid and weights may be used for approximating integrals of the form (3) for a wide choice of functions $q(\boldsymbol{\theta})$. The polynomial choices required for the mean vector and covariance matrix have already been discussed and are used at each stage of the iteration. In the case of predictive distributions, $q(\boldsymbol{\theta}) = f(\mathbf{y} \mid \boldsymbol{\theta})$, the typical form of $f(\mathbf{y} \mid \boldsymbol{\theta})$ is such that its product with the likelihood $l(\mathbf{x};\ \boldsymbol{\theta})$ is the likelihood for the sample $(\mathbf{x}, \mathbf{y})$ and hence behaves like $l(\mathbf{x};\ \boldsymbol{\theta})$ in situations where the dimension of $\mathbf{y}$ is much less than that of $\mathbf{x}$ (often, of course, $\mathbf{y}$ is simply one-dimensional).

This integration method has now been tested on a wide range of both real and artificial problems, a selection of which are described in Section 3. In particular, we have compared our approach with other methods currently available. Experience so far supports the view that the method is generally "fail-safe", in the sense that if the assumptions underlying the approach are not reasonably satisfied, then convergent results cannot be found. This is generally not the case, for example, with approximations based on maximum likelihood methods for which some sort of numerical answer is always available, irrespective of the adequacy of the "asymptotic normality" assumption.

In some problems, it is very advantageous to use functional transformations (e.g. logarithmic) of some components of the parameter vector in order to enhance the spherical normal approximation: for example, it is usually most sensible to work with $\log(\sigma)$ instead of σ. In the examples presented here, however, this has not always been done, in order to facilitate comparison with previous analyses of the data sets.

Techniques are also incorporated into the method to prevent real arithmetic overflow (since the integral (3) with $q(\boldsymbol{\theta}) = 1$ is generally very small) and for the accurate summation of expressions such as (11) (since with some choices of $q(\boldsymbol{\theta})$ the terms in this sum may be of different magnitudes, but none can be assumed to make an insignificant contribution). These techniques will not be discussed here. The graphs of marginal posterior densities presented in this report have been drawn using natural cubic splines (see, for example, Ahlberg *et al.*, 1967).

3. Some Illustrative Examples

3.1. *Timing Comparison with Reilly's Method*

Reilly (1976) illustrated his method with a small data set of six $(x_i,\ y_i)$ pairs from the model $\log y_i = \log(\alpha + \beta x_i) + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$. Taking σ^2 to be known, so that $\boldsymbol{\theta} = (\alpha, \beta)$, Reilly produced joint posterior contours for (α, β), marginal densities for α and β, and for α/β, based on function evaluations using a 101×101 grid. The site for this grid was found by manipulating an initial grid of 11×11 points. Although this problem does not seem ideal as a possible test case for our approach (since the sample is rather small), results corresponding to and *agreeing exactly* (to within the limits of graphical presentation) with those given by Reilly have been obtained for this two-parameter problem using our method. A deliberately poor starting point was chosen and two initial iterations on 5×5 grids, followed by four 7×7 grids, produced the required convergence. A second run was needed with reparametrization to obtain the marginal distribution of α/β.

Using the 101×101 grid, a direct timing comparison with Reilly's method was not possible on the computer used in our study. However, a succession of runs using a range of smaller grid sizes were used to obtain an estimate of the time required to compute the function evaluations for a grid of 101×101 points. Using this deliberately conservative estimate of the time for Reilly's method, we found a time ratio of about 4 : 1 in favour of our approach. In the three-parameter problem, the time ratio in favour of our method was found to be in excess of 25 : 1 when all bivariate and univariate marginal posterior densities were produced.

In addition to these favourable time comparisons, which, of course, become even more extreme as the dimensions of the problem increase, we note also the very minimal storage requirement of our method as compared with one which relies on large grids of function evaluations. (A grid of 101^3 single precision real values requires over four mega-bytes of core storage!)

3.2. *Re-analysis of Some Leukaemia Data*

The data in Table 1 (Gehan, 1965) are the remission times (in weeks) of two groups of leukaemia patients, censored observations being indicated by "†".

TABLE 1
Remission times for leukaemia patients

Group 1	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11
	12	12	15	17	22	23									
Group 2	6	6	6	7	10	13	16	22	23						
	6†	9†	10†	11†	17†	19†	20†	25†	32†	32†	34†	35†			

These data have been re-examined recently by Aitkin and Clayton (1980) using GLIM and the EM algorithm to analyse a two-sample Weibull model with a covariate (group membership) coded to be $z_1 = \frac{1}{2}$ in the first group and $z_2 = -\frac{1}{2}$ in the second group. The survival times can then be regarded as a single sample of $n+m$ values, the last m being censored ($n = 30$, $m = 12$). Given a survival time density function $f(t)$ and probability of survival $R(t) = \int_t^\infty f(u)\,du$, the likelihood function is defined by

$$L = \prod_{i=1}^{n} [f(t_i)] \prod_{i=n+1}^{n+m} [R(t_i)].$$

The explanatory variable z is incorporated through a proportional hazards model in which the hazard function $h(t) = f(t)/R(t)$ is assumed to be of the form

$$h(t_i) = \lambda(t_i) \exp(\beta_0 + \beta_1 z_i),$$

where, with $\lambda(t) = \alpha t^{\alpha-1}$, it can be shown that

$$f(t_i) = \alpha t_i^{\alpha-1} \exp\{\beta_0 + \beta_1 z_i - t_i^\alpha e^{\beta_0 + \beta_1 z_i}\}$$

(which is the standard Weibull density if $\beta_0 + \beta_1 z_i = 0$).

The logarithm of the likelihood function may be written as

$$\ln L = n \ln \alpha + \sum_{i=1}^{n+m} (\omega_i \ln \mu_i - \mu_i) - \sum_{i=1}^{n} \ln t_i,$$

where $\ln \mu_i = \alpha \ln t_i + \beta_0 + \beta_1 z_i$ and ω_i is an indicator variable taking the value 1 for uncensored times and 0 for censored times.

An improper locally uniform prior, $p(\beta_0, \beta_1, \alpha) =$ constant, was used, and Table 2 gives the posterior means, standard errors and correlations for $\boldsymbol{\theta} = (\beta_0, \beta_1, \alpha)$, obtained with our

method, based mainly on a $7 \times 7 \times 7$ grid which worked with β_0, β_1 and, as a third parameter, a linear combination of all three parameters chosen to avoid problems arising from the high correlation between β_0 and α. The table also summarizes the results obtained by Aitkin and Clayton using GLIM and maximum likelihood (and we are very grateful to them for supplying details not given in their paper).

There is considerable agreement in this case between the numerically "exact" posterior means and standard errors (i.e. those based on the efficient numerical procedure) and those based (equivalently) on the standard posterior normal approximation. In fact, direct inspection of the log-likelihood contours reveals that they are quite smoothly elliptical over a large range, despite the high correlation between β_0 and α.

3.3. *Re-analysis of the Stanford Heart Transplant Data*

Data relating to the Stanford heart transplant programme have been discussed and analysed using a variety of methods by Turnbull *et al.* (1974). One of the models proposed (referred to in Section 4.3 of their paper as the Pareto model) assumes that individual patients in the non-transplant group have exponential lifetime distributions with parameters ϕ drawn from a gamma population density $\lambda^p \phi^{p-1} e^{-\lambda\phi}/\Gamma(p)$, and that patients in the transplant group have similar distributions, but with $\tau\phi$ in place of ϕ. With $\boldsymbol{\theta} = (\tau, \lambda, p)$ the likelihood is then given by

$$\prod_{i=1}^{n} \frac{p\lambda^p}{(\lambda + x_i)^{p+1}} \cdot \prod_{i=n+1}^{N} \left(\frac{\lambda}{\lambda + x_i}\right)^p \cdot \prod_{j=1}^{m} \frac{\tau p \lambda^p}{(\lambda + y_j + \tau z_j)^{p+1}} \cdot \prod_{j=m+1}^{M} \left(\frac{\lambda}{\lambda + y_j + \tau z_j}\right)^p,$$

where the x_i are the survival times in days of the $N = 30$ non-transplant patients, $n = 26$ of whom died, and y_j, z_j are the times to transplant and survival times, respectively, for the $M = 52$ transplant patients, $m = 34$ of whom died.

Although it is actually possible in this case to integrate analytically over the p parameter, we shall illustrate our numerical procedure for the full three-parameter likelihood. Using an improper uniform prior for $\boldsymbol{\theta}$, we have calculated the posterior means, standard errors and correlations for τ, λ and p. These are shown in Table 3, together with the corresponding maximum likelihood estimates and "asymptotic" correlation matrix derived from covariance estimates given by Turnbull *et al.* It is clear that there are significant differences between the two sets of results and we shall consider the reason for this in more detail. This example also provides a good illustration of the comments made in Section 2 regarding the performance and operation of the iterative scheme.

In Fig. 1, marginal posterior densities of τ, λ and p are shown (constructed using natural cubic spline curves fitted through the final 9 and 10 points for the parameter in each case) together with normal densities based on the ML estimates. Again, there is considerable discrepancy between the corresponding pairs of densities, particularly for λ. We see that all the exact marginal densities are markedly non-normal and in particular show positive skewness.

TABLE 2
Summary of inferences for $\boldsymbol{\theta} = (\beta_0, \beta_1, \alpha)$

	Estimates and standard errors			*Correlations*		
	β_0	β_1	α		β_0	β_1
Bayes	−4·05	1·77	1·39	β_1	−0·38	
	(0·61)	(0·42)	(0·20)	α	−0·94	0·26
ML	−3·93	1·73	1·37	β_1	−0·38	
	(0·66)	(0·41)	(0·29)	α	−0·94	0·26

TABLE 3
Comparison of Bayes and ML inferences for (τ, λ, p)

	Estimates and standard errors				*Correlations*	
	τ	λ	p		τ	λ
Bayes	1·04	32·5	0·50	λ	−0·08	
	(0·47)	(16·2)	(0·14)	p	−0·44	0·78
ML	0·81	22·0	0·44	λ	−0·46	
	(0·34)	(6·0)	(1·10)	p	−0·46	0·70

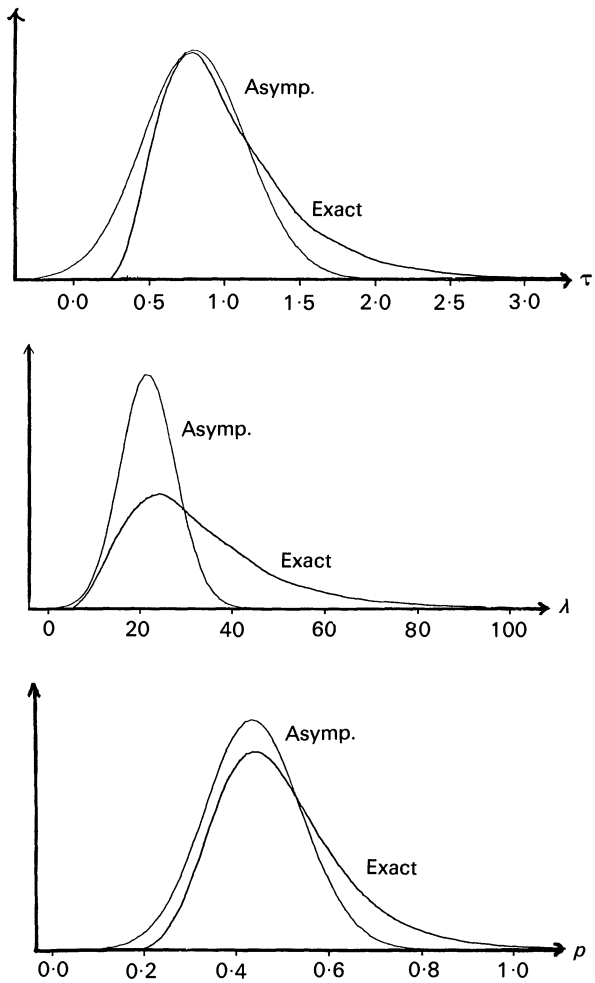

FIG. 1. Exact and "asymptotic" posterior densities for the Stanford data.

Such a degree of non-normality was anticipated after a run on a series of $5 \times 5 \times 5$ grids failed to show satisfactory convergence.

For larger grid sizes, satisfactory convergence is observed and a summary of convergent results for a series of grid sizes is shown in Table 4. In each case iterations were continued until there was seven-digit agreement in the successive values for the logarithm of the reciprocal of the normalizing constant, i.e. $f(\mathbf{x})$. Because of the relatively high posterior correlations in this problem, a full orthogonal transformation of the parameter space was used. Points for the graphs were mainly obtained by changing the order in which the parameters were presented to the program (as shown in Table 4). However, in each case one set of points was obtained (with the chosen parameter in the θ_2 or θ_3 position) by omitting part of the transformation. There was good agreement between the density points obtained using the two approaches.

TABLE 4

Convergent results at a series of grid sizes for different parameter orderings

	τ		λ		p					
Grid Sizes	*Mean* (*Taken as* θ_1)	*s.d.*	*Mean* (*Taken as* θ_2)	*s.d.*	*Mean* (*Taken as* θ_3)	*s.d.*	*Correlations* $\tau.\lambda$	$\tau.p$	$\lambda.p$	*Norm. const.* $10^{164} \times f(\mathbf{x})$
8^3	1·0675	0·4502	33·15	15·28	0·4960	0·1342	−0·070	−0·429	0·751	3·7447
9^3	1·0127	0·4875	31·70	16·55	0·4963	0·1480	−0·078	−0·444	0·779	4·2499
10^3	1·0577	0·4707	32·93	16·00	0·4971	0·1420	−0·072	−0·433	0·761	3·8472
	(Taken as θ_3)		(Taken as θ_1)		(Taken as θ_2)					
8^3	1·0297	0·4545	33·28	15·40	0·5036	0·1382	−0·074	−0·447	0·752	3·9197
9^3	1·0414	0·4743	31·59	16·48	0·4892	0·1437	−0·082	−0·424	0·786	4·1748
10^3	1·0379	0·4706	32·97	16·04	0·4999	0·1411	−0·081	−0·437	0·769	3·9724
	(Taken as θ_3)		(Taken as θ_2)		(Taken as θ_1)					
8^3	1·0273	0·4564	32·67	15·69	0·4998	0·1370	−0·073	−0·437	0·766	4·0035
9^3	1·0378	0·4698	32·60	15·85	0·4971	0·1394	−0·076	−0·437	0·766	4·0407
10^3	1·0375	0·4732	32·47	16·17	0·4967	0·1411	−0·084	−0·437	0·775	4·0590

In Table 4 we note that the choice (p, λ, τ) for the parameter order performs rather better from the convergence point of view (in that stable values from successive grids show better agreement) than the choice (λ, p, τ), and that both of these perform better than the basic order (τ, λ, p). The results from the choice (p, λ, τ) have been reported in Table 3. Since different choices of parameter order correspond to different transformations of the parameter space this observation is readily explained by supposing that the transformed parameters obtained with the order (p, λ, τ) have sets of conditional distribution which are more nearly normal than those for the other choices. Marginal density plots for these transformed parameters were found to give a visual confirmation of this which suggests that, since some "directions" are "better" than others, the use of methods which depend on full spherical symmetry may not be useful for problems such as this (*cf.* our earlier remarks about spherical rules in Section 2.2).

As can be seen from Table 4, the convergence for the posterior quantities summarized in Table 3 is generally rather poor, hence more results than are generally required are presented as evidence for these values. This situation could have been improved by considering additional transformations of the parameter space (for example replacing λ with $\log(\lambda)$). However, this would have been at the expense of straightforward comparability with the results of Turnbull *et al.* Detailed studies have shown that the marginal density values show much better agreement as grid size is varied and, in any case, are more useful as a basis for inference than posterior means and standard errors in the case of non-normal densities such as these.

On the basis of the ML estimate of τ, $\hat{\tau} = 0{\cdot}81$ with standard error 0·34, Turnbull *et al.* conclude that, for the transplant group, "The results indicate some reduction in hazard (19 per cent), but... nothing approaching statistical significance". Examination of the skewed marginal posterior density for τ reveals that even this tentative suggestion of a reduction is not justified. The posterior probabilities for $\tau < 1$ and $\tau > 1$ are virtually identical ($= \frac{1}{2}$) and there is no evidence in favour of the former hypothesis.

3.4. *Re-analysis of a Regression Problem with Censored Data*

Schmee and Hahn (1979) discuss the results of temperature accelerated life tests on electrical insulation in 40 motorettes. Ten motorettes were tested at each of four temperatures, the test termination (censoring) time being different at each temperature. The results (in hours) are given in Table 5.

TABLE 5
Insulation life at various test temperatures

150 °C	8064†,	8064†,	8064†,	8064†,	8064†,	8064†,	8064†,	8064†,	8064†,	8064†,
170 °C	1764,	2772,	3444,	3542,	3780,	4860,	5196,	5448†,	5448†,	5448†,
190 °C	408,	408,	1344,	1344,	1440,	1680†,	1680†,	1680†,	1680†,	1680†,
220 °C	408	408,	504,	504,	504,	528†,	528†,	528†,	528†,	528†,

The model adopted by Schmee and Hahn assumes that log-failure times are normally distributed with constant variances and means which are linear in $x = 1000/(T+273{\cdot}2)$, where T is the Centigrade temperature. If the means are denoted by $\beta_0 + \beta_1 x_i$, the log-likelihood has the form

$$-\log(\sqrt{(2\pi)}\,\sigma)\sum_{i=1}^{4} n_{0i} - (2\sigma^2)^{-1}\sum_{i=1}^{4}\sum_{j=1}^{n_{0i}}(\log y_{ij} - (\beta_0 + \beta_1 x_i))^2 + \sum_{i=1}^{4}\{c_i \log[1 - \Phi(\sigma^{-1}(\log d_i - (\beta_0 + \beta_1 x_i)))]\},$$

where at temperature T_i there are n_{0i} observed lifetimes y_{ij} and c_i items surviving longer than d_i hours.

In this case, we worked with parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$, with the "non-informative" prior form, $p(\boldsymbol{\theta}) \propto \sigma^{-1}$, and crude starting values for the mean vector and covariance matrix obtained by applying least squares under the assumption that the censored values actually failed at the censoring time (see Schmee and Hahn for details). This simple preliminary analysis revealed the possibility of very high posterior correlation between β_0 and β_1 and so the program worked with an "orthogonalized" set of parameters (as discussed in Section 2.2). We worked directly with σ, rather than $\log(\sigma)$, in order to facilitate comparison with Schmee and Hahn's results.

In the case of β_0 and β_1, the marginal posterior densities are symmetric and the posterior means (PM) ($E(\beta_0 \mid \text{data}) = -6{\cdot}2$, $E(\beta_1 \mid \text{data}) = 4{\cdot}4$) are similar to the iterative least squares (ILS) estimates (Schmee and Hahn) and the maximum likelihood (ML) estimates (Aitkin, 1980).

The main interest in this reanalysis of these data lies in the results for σ. The posterior density is rather skewed, as shown in Figure 2, and the use of the iterated least squares procedure is seen to be extremely suspect. The Bayesian analysis reinforces the points made by Aitkin (1980) concerning the bias problems that arise with ILS and ML estimates of σ from censored data. In addition, of course, such estimates would produce highly misleading statements of uncertainty for β_0 and β_1 if classical confidence interval procedures were used.

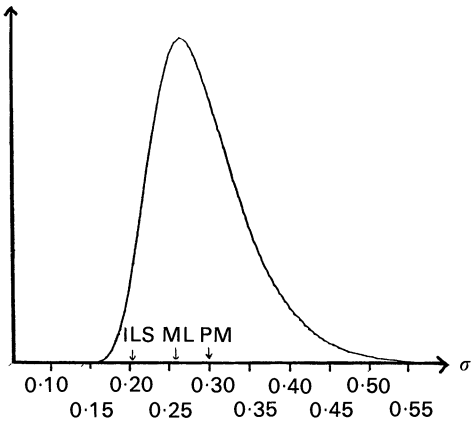

FIG. 2. Posterior density for σ, together with ILS, ML and PM estimates.

4. OTHER APPLICATIONS

A library of Fortran subroutines is currently available (on request from the first author) for the case $k = 3$. This enables one, for example, to carry out detailed numerical analysis of a robust Bayesian approach to location-scale problems using any family of sampling distributions defined by a further "shape" parameter, λ, say, so that $\boldsymbol{\theta} = (\mu, \log(\sigma), \lambda)$. We can, for example, consider the exponential-power family (Box and Tiao, 1973), or the family of t-densities (Relles and Rogers, 1977).

Using the routines for efficient numerical integration in three-dimensions, a number of problems involving $k = 4$, or $k = 5$, parameters can be satisfactorily approached by adopting suitable discretized ranges for the additional parameters. For example, robust analysis of a straight line model ($E(y) = \alpha + \beta x$, $V(y) = \sigma^2$) can be carried out by assuming the error to follow one of the families of densities described above, with "shape" parameter λ: we then work with $\boldsymbol{\theta} = (\alpha, \beta, \log(\sigma), \lambda)$ (so that $k = 4$) and a suitable discrete range of values for λ. As a further example, suppose we consider the straight-line model with errors following a normal ARMA (1, 1) process (with parameters ϕ and ψ). A satisfactory approximate analysis can then be made by taking $\boldsymbol{\theta} = (\alpha, \beta, \log(\sigma), \phi, \psi)$ (so that $k = 5$) with a suitable discrete grid of values for (ϕ, ψ).

5. GENERAL REMARKS

The numerical approach outlined in this paper is based on the assumption that for moderate (finite) samples, and under suitable parametrization, a posterior density can be well approximated by the product of a normal density and a polynomial.

In comparison with standard maximum likelihood methods, perhaps based on the EM algorithm, the following points should be noted:

(i) complex likelihoods or prior densities are handled routinely by our method;

(ii) maximum likelihood variance–covariance estimates (based on the matrix of second derivatives) are only really sensible if the log-likelihood contours are close to ellipsoidal—a much more restrictive assumption than ours;

(iii) even when the maximum likelihood approach converges, a rather subjective assessment (perhaps by inspection of log-likelihood plots) has to be made of whether the assumption of ellipsoidal contours is reasonable or not; with our approach, if the assumptions made are not reasonable for a particular problem (or parametrization) we learn this directly, since stable answers within and between successive grids are not obtainable.

Acknowledgements

This work is part of the first author's Ph.D. thesis being written at Nottingham University under the supervision of the second author. We are grateful to referees for detailed and helpful comments on earlier drafts of the paper.

References

Ahlberg, J. H., Nilson, E. N. and Walsh, J. L. (1967). *The Theory of Splines and their Applications.* New York: Academic Press.

Aitkin, M. (1981). A note on the regression analysis of censored data. *Technometrics*, **23**, 161–163.

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex survival data using GLIM. *Appl. Statist.*, **29**, 156–163.

Box, G. E. P. and Tiao,, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass: Addison-Wesley.

Davis, P. J. and Rabinowitz, P. (1967). *Numerical Integration.* Waltham, Mass.: Blaisdell.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrary singly-censored samples. *Biometrika*, **52**, 203–232.

Reilly, P. M. (1976). The numerical computation of posterior distributions. *Appl. Statist.*, **25**, 201–209.

Relles, D. A. and Rogers, W. H. (1977). Statisticians are fairly robust estimators of location. *J. Amer. Statist. Ass*, **72**, 107–111.

Salzer, H. E., Zucker, R. and Capuano, R. (1952). Table of the zeroes and weight factors of the first twenty Hermite polynomials. *J. Res. Nat. Bur. Standards*, **48**, 111–116.

Schmee, J. and Hahn, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–432.

Stroud, A. H. (1972). *Approximate Calculation of Multiple Integrals.* Englewood Cliffs: Prentice-Hall.

Turnbull, B. W., Brown, B. W. Jr and Hu, M. (1974). Survivorship analysis of heart transplant data. *J. Amer. Statist. Ass.*, **69**, 74–80.