Ann. Inst. Statist. Math.
Vol. 46, No. 2, 203–220 (1994)

BAYESIAN AND LIKELIHOOD INFERENCE FROM EQUALLY WEIGHTED MIXTURES

TOM LEONARD[1], JOHN S. J. HSU[2], KAM-WAH TSUI[1] AND JAMES F. MURRAY[3]

[1]*Department of Statistics, University of Wisconsin-Madison, 1210 West Dayton Street, Madison, WI 53706-1693, U.S.A.*
[2]*Department of Statistics and Applied Probability, University of California - Santa Barbara, Santa Barbara, CA 93106-3110, U.S.A.*
[3]*Graduate Program in Hospital and Health Administration, University of Iowa, Iowa City, IA 52242, U.S.A.*

(Received September 28, 1992; revised July 27, 1993)

Abstract. Equally weighted mixture models are recommended for situations where it is required to draw precise finite sample inferences requiring population parameters, but where the population distribution is not constrained to belong to a simple parametric family. They lead to an alternative procedure to the Laird-DerSimonian maximum likelihood algorithm for unequally weighted mixture models. Their primary purpose lies in the facilitation of exact Bayesian computations via importance sampling. Under very general sampling and prior specifications, exact Bayesian computations can be based upon an application of importance sampling, referred to as Permutable Bayesian Marginalization (PBM). An importance function based upon a truncated multivariate t-distribution is proposed, which refers to a generalization of the maximum likelihood procedure. The estimation of discrete distributions, by binomial mixtures, and inference for survivor distributions, via mixtures of exponential or Weibull distributions, are considered. Equally weighted mixture models are also shown to lead to an alternative Gibbs sampling methodology to the Lavine-West approach.

Key words and phrases: Equally weighted mixtures, survivor distribution, maximum likelihood, EM algorithm, binomial mixtures, Bayesian marginalization, importance sampling, Gibbs sampler.

1. Equally weighted mixtures for density estimation

Let $x_1, \ldots, x_n$ denote independent random variables with common density $f(t)$ and cumulative distribution function (c.d.f) $F(t)$, for $t \in (-\infty, \infty)$. Suppose that the observed data y consists of those x_i falling outside a specific "censoring region", Ω_i, for the i-th observation, and $d_1, \ldots, d_n$, where

$$d_i = \begin{cases} 1 & \text{if } x_i \notin \Omega_i \\ 0 & \text{if } x_i \in \Omega_i \end{cases} \quad (i = 1, \ldots, n). \tag{1.1}$$

An example, where $x_1, \ldots, x_n$ are censored survivor times, is discussed in Section 6.

Following Leonard (1984), assume that the unknown density f may be represented as an equally weighted mixture of the form

$$f_m(t; \boldsymbol{\xi}) = m^{-1} \sum_{k=1}^{m} \lambda(t, \xi_k), \tag{1.2}$$

where $\xi_1, \ldots, \xi_m$ are unknown scalar parameters, and, for each ξ, $\lambda(t, \xi)$ is a specified density in t, assumed twice differentiable with respect to ξ, for each t. It would also be possible to let the density λ depend upon one or two common unknown parameters, without unduly complicating our analysis.

The assumption (1.2) generalizes kernel estimators (e.g., Rosenblatt (1956), Tapia and Thompson (1978)) which, in the uncensored case, rather restrictively locate n kernels over the observed data points, leading to problems with estimation in the tails, and the identification of bandwidth parameters. The current assumption enables us to estimate the m locations $\xi_1, \ldots, \xi_m$ from the data, e.g., by maximization of $L(f_m)$ with respect to $\xi_1, \ldots, \xi_m$, where $L(f)$ denotes the log-likelihood functional

$$L(f) = \sum_{i:d_i=1} \log f(x_i) + \sum_{i:d_i=0} \log G(\Omega_i), \tag{1.3}$$

and $G(\Omega) = \int_\Omega dF(t)$.

This procedure smoothes f without assuming that f belongs to a simple parametric family; unconditional maximization of (1.3) would (e.g., Efron (1967)) provide a non-parametric step function for F. We hence consider a simple alternative to the semi-parametric procedures recommended by Leonard (1978) and Lenk (1991). There are some similarities between this approach and the Gaussian sums suggested by Sorenson and Alspach (1971), and Alspach (1975).

For moderate to large m, the model (1.2) includes, as special cases, unequally weighted mixture models of the form

$$f(t) = \sum_{j=1}^{q} \phi_j \lambda(t, b_j) \tag{1.4}$$

($-\infty < t < \infty$, $\phi_1 + \cdots + \phi_q = 1$, $\phi_j \geq 0$ and $-\infty < b_j < \infty$, for $j = 1, \ldots, q$; $q = 1, 2, \ldots$), for many convenient choices of $q \leq m$, but where each ϕ_j is constrained to be a multiple of m^{-1}. Hence, the model (1.2) provides a convenient way of saying that (1.4) may hold, but that we are unwilling to constrain q to assume a specific value. Applications of models (1.3) and (1.4) include situations where ϕ_j may be interpreted as the probability that an observation belongs to a subpopulation j, where the observation would possess density $\lambda(t, b_j)$.

Since the model (1.2) does not contain unknown unequal mixing probabilities, it is generally easier to analyze, from either a likelihood or Bayesian perspective, when compared with (1.4). Note that (1.2) includes as a special case ($\xi_1 = \xi_2 = \cdots = \xi_m$) a parametric model represented by the density λ. It is therefore possible

to use (1.2) to investigate whether the data supports a working hypothesis λ. If this hypothesis is untrue, the data can suggest a much more general estimate, based upon (1.2).

Laird (1978, 1982) and DerSimonian (1986), who published Laird's algorithm, seek to maximize (1.3) among the class of densities (1.4), thus seeking a maximum likelihood estimate $\hat{f}(t)$ for $f(t)$, which possesses Efron's self consistency property. The geometry of this solution is considered by Lindsay (1981, 1983), and a related algorithm in the context of random coefficient regression models, is discussed by Mallet (1986). The algorithm proposed by DerSimonian incorporated a search for the global maximum, proposed by Simar (1978), which does not however always return a global maximum (in practical problems many local maxima may exist). While our procedures will not completely guarantee a global maximum, they permit a quite exhaustive search for a global maximum, and also assist the choice of q.

Titterington *et al.* (1985) propose a variety of inferential procedures for unequally weighted mixtures. Their recursive formulae (Ch. 6) are proposed as potential approximations to a Bayesian solution, and parallel Sorenson and Alspach (1971). However, only Lavine and West (1992) have suggested an exact Bayesian solution for useful special cases of the model (1.4), with q fixed. They refer to the Gibbs sampler, as introduced to the Bayesian literature by Gelfand and Smith (1990), and Carlin and Gelfand (1991); this simulates from a succession of conditional distributions, rather than directly simulating from the exact posterior distribution of the parameters.

In Section 5, we will show that an exact Bayesian solution for the model (1.2) can be calculated, under a very broad range of assumptions, using a variation of importance sampling referred to as permutable Bayesian marginalization (PBM). Importance sampling is much more broadly applicable than Gibbs sampling, which requires a collection of conditional distributions to assume technically simple forms.

2. Maximum likelihood methods

It is straightforward to find local maxima of (1.3), under the class of equally weighted mixtures (1.2), by considering $\xi_1, \ldots, \xi_m$ satisfying

$$\hat{\xi}_k = \sum_{i=1}^{n} d_i P_{ik} / \sum_{i=1}^{n} u_i P_{ik} \qquad (k = 1, \ldots, m), \tag{2.1}$$

with $u_i = \min(x_i, c_i)$,

$$P_{ik} = \ell_i(\hat{\xi}_k) / \sum_{g=1}^{m} \ell_i(\hat{\xi}_g), \tag{2.2}$$

and

$$\ell_i(\xi_k) = \begin{cases} \lambda(x_i, \xi_k) & \text{if } d_i = 1 \\ \int_{\Omega_i} \lambda(t, \xi_k) dt & \text{if } d_i = 0. \end{cases} \tag{2.3}$$

As a special case of the EM algorithm, discussed by Laird, trial values for $\tilde{\xi}_1, \ldots, \tilde{\xi}_m$ may be substituted, via (2.2) and (2.3) into the right hand sides of (2.1), new values obtained from the left hand sides, and the process may be repeated until convergence. The values $\hat{\xi}_1, \ldots, \hat{\xi}_m$ will always exactly collapse into $\hat{q}$ groups, with $\hat{q} \leq m$. In practice $\hat{q}$ is typically much smaller than m, so that the $\hat{\xi}$'s tend to exactly cluster into a small number of groups. This is to be intuitively anticipated as model (1.2) simply constrains the ϕ_j in model (1.4) to be integer multiples of m^{-1}, so that we would expect the maximum likelihood estimates for the two models to behave somewhat similarly. Suppose that $\hat{m}_{(j)}$ of the ξ_k are set equal to $\hat{b}_j$, for $j = 1, \ldots, \hat{q}$. Then this solution will yield a member of the class (1.4), but with q estimated by $\hat{q}$, ϕ_j estimated by $\hat{\phi}_j = \hat{m}_{(j)}/m$, and b_j estimated by $\hat{b}_j$. In practice, we estimate q, by putting any two ξ's in the same group if they are differ by no more than some value ϵ (e.g., $\epsilon = 0.0001 \times$ the smallest difference between two distinct observations, for the example in Section 6 involving a mixture of exponential densities), and then find $\hat{q}$ simply by counting the number of distinct groups of the $\hat{\xi}$'s. In most numerical examples, we have found the distinction between the groups to be remarkably clear as long as the iterative procedure for the ξ's is allowed to completely converge, e.g., to five decimal place accuracy. Note that, in some applications, $\hat{m}_{(j)} l_i(\hat{b}_j) / \sum_{k=1}^{\hat{q}} \hat{m}_{(k)} l_i(\hat{b}_k)$ estimates the probability that the i-th observation belongs to subpopulation j.

In the special case where the quantities

$$r_k = \sum_{i=1}^{n} P_{ik} \qquad (k = 1, \ldots, m), \tag{2.4}$$

are constant in k, the above procedure will also provide an exact solution to Laird's maximum likelihood equations for the model (1.3), with $q = \hat{q}$. This can easily be demonstrated by substituting our solutions into Laird's equations. By increasing m, it is possible to ensure that $r_1 = r_2 = \cdots = r_m$ to any required degree of accuracy. Numerical comparison of $r_1, r_2, \ldots, r_m$, for any finite m, permits us to judge how closely the procedure defined by (2.1)–(2.3), is likely to approximate Laird's solution.

As long as $\xi_1, \ldots, \xi_m$ are scalars, our procedure however permits a straightforward systematic search for a global maximum, for any fixed m, and as m increases the search may be made narrower, thus permitting a reasonably exhaustive search for a global maximum of the likelihood under Laird's model (1.3). If a single specified value of m is of interest we recommend

(a) Start with several different sets of initial values for $\hat{\xi}_1, \ldots, \hat{\xi}_m$, but where $\hat{\xi}_1 < \hat{\xi}_2 < \cdots < \hat{\xi}_m$. For each set, use the above iterations to find $\hat{\xi}_1 \leq \hat{\xi}_2 \leq \cdots \leq \hat{\xi}_m$, $\hat{q}$, and $\hat{\phi}_j$ and $\hat{b}_j$, for $j = 1, \ldots, \hat{q}$. Choose the solution, and the corresponding $\hat{q}$ maximizing the log likelihood (1.2). Arrange that $\hat{b}_1 < \hat{b}_2 < \cdots < \hat{b}_q$.

(b) Repeat the solution of (2.1), ..., (2.3) using a variety of initial values based upon the quantities calculated in (a). For example, if $\hat{q} = 2$,

(i) Use $\hat{b}_1$ as initial value for first $\hat{m}_1 - 1$ of the ξ's, and $\hat{b}_2$ as initial values for remaining $\hat{m}_2 + 1$ of the ξ's. If this yields a larger log-likelihood than before, use new $\hat{b}_1$ as initial value for first $\hat{m}_1 - 2$ of the ξ's and new

$\hat{b}_2$ as initial value for remaining $\hat{m}_2 + 2$ of the ξ's. Keep subtracting a parameter from the first group, and adding to the second group, until the log-likelihood starts decreasing.

(ii) Use original $\hat{b}_1$ as initial value for the first $\hat{m}_1 + 1$ of the ξ's and original $\hat{b}_2$ as initial value for remaining $\hat{m}_2 - 1$ of the ξ's. Keep subtracting a parameter from the second group, and adding to the first group, using the latest values for $\hat{b}_1$ and $\hat{b}_2$, until the log-likelihood starts decreasing. For all runs in (i) and (ii) choose the values of $\hat{b}_1$, $\hat{b}_2$ and $\hat{m}_{(1)}$ and $\hat{m}_{(2)}$ maximizing the log-likelihood. To be completely thorough, all values of $\hat{m}_{(1)}$ and $\hat{m}_{(2)}$ satisfying $\hat{m}_{(1)} + \hat{m}_{(2)} = m$ should be considered.

For general $\hat{q}$, obvious generalizations of the above scheme may be constructed, which successively perturb $\hat{m}_{(1)}, \hat{m}_{(2)}, \ldots, \hat{m}_{(\hat{q})}$. Again, to be completely thorough, all integer values of $\hat{m}_{(1)}, \ldots, \hat{m}_{(\hat{q})}$, summing to m, should be considered.

When searching for Laird's solution ($m \to \infty$) we recommend seeking the global maximum for $m = m^*, 2m^*, 4m^*, 8m^*, \ldots$ until the quantities in (2.4) become constant in k. The choices $m^* = 25$ i.e., $m = 25, 50, 100, 200, 400, \ldots$, often work well in practice. In this case we recommend a very thorough search for the global maximum, for the smallest value, $m = m^*$. Suppose that this yields the solution $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{m^*}$ which collapse into $\hat{q}$ groups. Then, as initial values for $\xi_1, \ldots, \xi_m$, when $m = 2m^*$, use $\hat{\xi}_1, \hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_2, \ldots, \hat{\xi}_{m^*}, \hat{\xi}_{m^*}$. However, also compare with other, unequal, initial values, to check that $\hat{q}$ has not changed. Once an initial maximum likelihood solution has been obtained, when $m = 2m^*$, perturb in similar fashion to the procedures described in (i) and (ii), to seek a larger log-likelihood. However, if a fairly exhaustive search has been completed, when $m = m^*$, it is only necessary to slightly perturb the solution, when $m = 2m^*$. Keep doubling m, and proceed in similar fashion, always maximizing (1.2), until the condition based upon (2.4) is satisfied.

3. Numerical comparison with the Laird-DerSimonian algorithm

Consider the situation where, conditional on $\theta_1, \ldots, \theta_n$, our observations $y_1, \ldots, y_n$ possess independent binomial distributions, with respective probabilities $\theta_1, \ldots, \theta_n$ and sample sizes $t_1, \ldots, t_n$, and where $\theta_1, \ldots, \theta_n$ are a random sample from a discrete distribution. Under Laird's model this common mixing distribution assigns probabilities $\phi_1, \ldots, \phi_q$ to the values $b_1, \ldots, b_q$. Under our equally weighted model, the mixing distribution instead assigns equal probabilities to each of the points $\xi_1, \ldots, \xi_m$. Similar techniques to those indicated in Sections 1 and 2 may be applied to the present situation. Binomial mixture sampling models provide alternatives to discrete exponential family models (e.g., Hsu *et al.* (1991)).

Consider the gender data reported by Leonard (1972), with $n = 10$, and y_i denoting the number of females attending course i out of t_i males and females ($i = 1, \ldots, 10$). For these data, the DerSimonian algorithm converges very quickly, using just two iterations on the location parameters, with a search for the mixing probabilities at each stage. This algorithm however reported a local maximum of $\hat{q} = 3$, $\hat{b}_1 = 0.152$, $\hat{b}_2 = 0.158$, and $\hat{b}_3 = 0.439$, with mixing probabilities $\hat{\phi}_1 = 0.540$, $\hat{\phi}_2 = 0.015$, and $\hat{\phi}_3 = 0.445$, and log-likelihood $L = -27.9972$, and

returned an error code stating that it was unable to find a global maximum. The DerSimonian algorithm, for the unconstrained model (1.4), does not fail due to lack of model identification, but rather due to computational problems during the search for the global maximum. This procedure finds the maximum with the ϕ_j fixed and attempts a grid search on the ϕ_j, for the global maximum.

We firstly applied our procedure, with $m = m^* = 25$. Our starting values included a set of 25 equi-distant values for $\xi_1, \ldots, \xi_m$. The corresponding iterations converged to a local maximum, correct to four decimal places, in 24 iterations, with log-likelihood $L = -28.2044$. This solution set 11 of the ξ's equal to 0.150 and 14 of the ξ's equal to 0.435, giving $\hat{q} = 2$ and estimated mixing probabilities $\hat{\phi}_1 = 11/25$ and $\hat{\phi}_2 = 14/25$ for the locations $\hat{b}_1 = 0.150$ and $\hat{b}_2 = 0.435$.

We then perturbed this solution, once according to step (i) of Section 2, and four times according to step (ii). This yielded an improved solution of $\hat{q} = 2$, $\hat{\phi}_1 = 14/25$, $\hat{\phi}_2 = 11/25$, $\hat{b}_1 = 0.153$, and $\hat{b}_2 = 0.439$, with $L = -27.9974$. Just to make sure, we explored all possibilities of $\hat{\phi}_1$, as an integer multiple of m^{-1}, with $\hat{q} = 2$. Since we did not achieve a larger log-likelihood we tried a few different sets of unequal starting values for $\xi_1, \ldots, \xi_n$, in order to check $\hat{q}$. We conclude that our solution, with $L = -27.9974$ provides a global maximum, when $m = 25$.

With $m = 50$, the solution for $m = 25$ seemed to remain a global maximum, and in this case we attempted perturbations to the right and left of the $m = 25$ solution. With $m = 100$, we increased, by a single perturbation to $L = -27.997479$, with $\hat{q} = 2$, $\hat{\phi}_1 = 55/100$, $\hat{\phi}_2 = 45/100$. With $m = 200$, we again increased L, via a single perturbation to $L = -27.9966$. With $m = 400$, the global maximum remained the same, with $\hat{q} = 2$, $\hat{\phi}_1 = 222/400 = 0.555$, $\hat{\phi}_2 = 178/400 = 0.445$, $\hat{b}_1 = 0.152$ and $\hat{b}_2 = 0.439$.

We conclude, via our analyses for different values of m, that we have provided a reasonable exhaustive search for the global maximum when $m = 400$. For each value of m, we also checked $\hat{q}$, by considering other, unequal, initial values for $\xi_1, \ldots, \xi_m$. Since the r_k in (2.4) all become equal, when $m = 400$, to four decimal places, we conclude that, to a reasonable level of accuracy, (within about $1/400 = 0.0025$ on the mixing probabilities) we have achieved an adequate approximation for the global maximum, as $m \to \infty$ i.e., for the maximum likelihood estimates of q, the ϕ's and b's under the model (1.3).

Since the DerSimonian computer program gives a smaller log-likelihood L, we conclude that her program, published in Applied Statistics, for Laird's solution, does not always achieve a global maximum. It similarly does not achieve the global maximum for the baseball batting data analyzed by Laird (1982), even though Laird achieves the correct solution in her paper. It is clear, however, that the different solutions yield only small changes in the likelihood. Therefore DerSimonian's solution can still provide sensible estimates which are not precisely maximum likelihood estimates. A broader application of the algorithm by DerSimonian might well produce the correct answer.

Our more detailed analysis does not require a large amount of computer time, and does not increase dramatically, as m increases. For example, all the above computations were completed within a total of 15.2 seconds of CPU time on a Sunsparc station, while the DerSimonian algorithm continued indefinitely, or until

reaching the maximum number of cycles allowed to reach convergence to a global maximum. While it might be possible to achieve the global maximum, by Laird's methodology, and other search procedures, our own search procedure seems more direct. However, the maximum likelihood procedure for the class of models (1.2) can readily be extended to the calculation of Bayesian posterior modes, which are simple enough to provide the basis for a Bayesian importance sampling procedure.

4. Simulation results

Each result in this section is based upon 100 simulations from some true distribution. Only our own procedure is considered, as the Laird-DerSimonian computer program used too much computer time to facilitate repeated simulation. We firstly simulated from mixtures of normal distributions with unit variance, and applied the methodology of Section 2, but with (1.2) replaced by mixtures of $m = 20$ normal densities, each with unit variance.

The first column of Table 1 describes the true value of q, for a particular set of simulations from the model (1.3), with the locations of the normal densities in model (1.3) described in the second column of Table 1, and each mixing probability equal to $1/q$. The third column of Table 1 gives the sample size, and the fourth through ninth columns describe the frequencies $N(1), \ldots, N(6)$, which relate to the numbers of occasions, during the 100 simulations, that our procedure estimated q to be respectively 1, 2, 3, 4, 5, and 6. The tenth column of Table 1 describes the simulated mean integrated squared error (MISE) of our density estimator, when compared with the true density. The last column describes the simulated MISE, when the estimator is instead a single normal density, with correct variance, and location replaced by the sample mean.

Table 1. Simulated results (mixtures of normal densities with unit variance).

q	Locations	n	$N(1)$	$N(2)$	$N(3)$	$N(4)$	$N(5)$	$N(6)$	MISE ($m = 20$)	MISE ($m = 1$)
2	3, 6	200	0	41	46	12	1	0	0.00232	0.01566
2	3, 6	400	0	37	49	14	0	0	0.00141	0.01553
2	4, 5	200	1	67	30	2	0	0	0.00163	0.00081
2	4, 5	400	1	58	39	2	0	0	0.00100	0.00042
3	3, 6, 9	200	0	0	46	43	11	0	0.00240	0.01284
3	3, 6, 9	400	0	0	56	33	10	1	0.00122	0.01277
4	3, 6, 9, 12	400	0	0	0	32	44	24	0.00143	0.01063

It can be concluded that, while this methodology does not perfectly recover the number of terms in the mixture, it does possess quite appealing MISE properties. Similar results are repeated in Table 2, when the true density is a mixture of exponential densities. Note that, whenever the true density is close in numerical terms to a single exponential density, whose mean can be estimated by the sample

Table 2. Simulated results (mixtures of exponential densities).

q	Locations	n	$N(1)$	$N(2)$	$N(3)$	$N(4)$	MISE ($m = 20$)	MISE ($m = 1$)
2	3, 6	200	8	76	16	0	0.00133	0.00086
2	3, 6	400	5	76	19	0	0.00064	0.00067
2	4, 5	200	38	54	7	1	0.00099	0.00030
2	4, 5	400	39	51	10	0	0.00055	0.00014
2	3, 9	200	1	69	30	0	0.00115	0.00293
2	3, 9	400	0	70	30	0	0.00055	0.00274
3	3, 6, 9	200	4	67	29	0	0.00152	0.00141
3	3, 6, 9	400	0	56	39	5	0.00058	0.00127
4	3, 6, 9, 12	400	0	47	47	6	0.00053	0.00163

mean, the latter will also provide an excellent estimator (see the simulated MISE's in the last column of Table 2). However, when the true density is radically different from a single exponential density, our mixture method can provide substantial savings in MISE (e.g., the $q = 4$ results in Table 2, or when the $q = 2$ locations are as far apart as 3 and 9). The performance of our procedure is therefore highly sensitive to the locations chosen in the mixture for the true density. It appears to work best when the true density is quite complex, but still works well (e.g., Table 1, when the locations are 4 and 5) when the true density is well approximated by a simple curve. It is anticipated that our procedure will detect q more accurately, when the sample size is very large.

5. Exact Bayesian analysis

For fixed m, the posterior density of $\xi_1, \ldots, \xi_m$ is denoted by

$$\pi(\boldsymbol{\xi} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\xi}) \prod_{i=1}^{n} \sum_{k=1}^{m} \ell_i(\xi_k), \tag{5.1}$$

where $\pi(\boldsymbol{\xi})$ denotes the prior density and the $\ell_i(\xi_k)$ satisfy (2.3). Since the likelihood contribution to (1.1) is a permutable function of $\xi_1, \ldots, \xi_m$, we assume that $\pi(\boldsymbol{\xi})$, and hence $\pi(\boldsymbol{\xi} \mid \boldsymbol{y})$, is also a permutable function of $\xi_1, \ldots, \xi_m$, and address the problem of computing the posterior density, or expectation, of any parameter of interest

$$\eta = \tau(\boldsymbol{\xi}) = \tau(\xi_1, \ldots, \xi_m), \tag{5.2}$$

which can be expressed as a permutable function of $\xi_1, \ldots, \xi_m$. Special cases of permutable η's include the population moments, and, for fixed t, the population density (1.2), and the corresponding population c.d.f. Our procedures are not applicable if $\eta = \tau(\boldsymbol{\xi})$ is not a permutable function of $\xi_1, \ldots, \xi_m$.

Importance sampling (Rubinstein (1981), Geweke (1988, 1989), Leonard *et al.* (1989) and Leonard and Hsu (1992)) works well whenever the exact posterior

density can be well approximated by another joint density which yields straightforward simulations for $\xi_1, \ldots, \xi_m$. While no useful approximating density is obviously available under the general mixture model (1.3), quite general classes of distributions are available under the model (1.4). A moderate value of m e.g., $m = 20$ is recommended, to ensure that the particular importance sampling procedure introduced below leads to feasible computations.

We suggest firstly seeking one to one transformations $\gamma_1 = h(\xi_1), \ldots, \gamma_m = h(\xi_m)$ of $\xi_1, \ldots, \xi_m$, such that the permutable piecewise multivariate t-approximation (PPMT) describe below is most reasonable for the transformed parameters. In particular, the vector of $\gamma = (\gamma_1, \ldots, \gamma_m)^T$, should be unconstrained in m-dimensional real space. Obvious transformations like log or logit will often suffice. Other useful transformations are considered by Bates and Watts ((1988), Ch. 6), since these are introduced in the context of providing suitable normalizing transformations for likelihood functions.

Suppose that the posterior density

$$\pi(\gamma \mid \boldsymbol{y}) \propto B(\gamma) \qquad (\gamma \in R^m), \tag{5.3}$$

of γ is proportional to a function $B(\gamma)$ which can be fully specified on m-dimensional real space R^m, and consider a parameter of interest η which can be expressed as a permutable function $\tau(\boldsymbol{\xi})$ of $\boldsymbol{\xi}$, or equivalently, as a permutable function $\tau^*(\gamma)$ of γ.

Let $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_m)^T$ provide a global maximum of $B(\gamma)$, satisfying $\tilde{\gamma}_1 \leq \tilde{\gamma}_2 \leq \cdots \leq \tilde{\gamma}_m$. As $B(\gamma)$ is a permutable function of $\gamma_1, \ldots, \gamma_n$, any permutation of the elements of $\tilde{\gamma}$ will also provide a global maximum i.e., $B(\gamma)$ can possess up to $m!$ global maxima. Unless $\tilde{\gamma}_1 = \tilde{\gamma}_2 = \cdots = \tilde{\gamma}_m$, a multivariate normal or multivariate t-approximation to (5.3) will therefore be inappropriate. This property leads to considerable practical complications which need to be circumvented by a series of theoretical devices. Note that Leonard *et al.* (1989) indicate that the approach by Tierney and Kadane (1986) cannot be reasonably extended to the computation of approximate posterior distribution of many non-linear functions of the parameters. Moreover, the procedure introduced by Leonard *et al.*, is virtually impossible to apply to the current situation, as very complicated conditional maximization procedure generalizing the methodology of Section 2, are required, together with complicated simulations for the f-contribution to their approximations.

Let Δ_m denote the set of all possible permutations $(i) = (i_1, \ldots, i_m)$ of the integers $(1, \ldots, m)$. For $i = (i_1, \ldots, i_m) \in \Delta_m$, consider

$$\Lambda^{(i)} = \{\gamma = (\gamma_1, \ldots, \gamma_m)^T : \gamma_{i_1} \leq \gamma_{i_2} \leq \cdots \leq \gamma_{i_m}\}. \tag{5.4}$$

Then, for each $(i) \in \Delta_m$, there exists $\tilde{\gamma}^{(i)}$ maximizing $B(\gamma)$, for $\gamma \in \Lambda^{(i)}$, such that $\tilde{\gamma}^{(i)}$ permutes the elements of $\tilde{\gamma}$. For each $(i) \in \Delta_m$ consider a Taylor Series expansion of $[B(\gamma)]^{-a}$, with $a = 2/(\nu + m)$, about $\gamma = \gamma^{(i)}$, for all $\gamma \in \Lambda^{(i)}$. Neglecting cubic and higher terms, in these $m!$ expansions, and raising the remaining terms to the power $(\nu + m)/2$, yields an approximation to (5.3) which we refer to as PPMT.

The PPMT approximation refers to the posterior dispersion matrices $\boldsymbol{D}^{(i)}$ satisfying

$$(\boldsymbol{D}^{(i)})^{-1} = -\frac{\partial^2 \log B(\gamma)}{\partial(\gamma\gamma^T)}\bigg|_{\gamma=\tilde{\gamma}^{(i)}} \qquad ((i) \in \Delta_m). \tag{5.5}$$

Then each $\boldsymbol{D}^{(i)}$ involves the obvious permutations of the rows and columns of the matrix $\boldsymbol{D}$ satisfying

$$\boldsymbol{D}^{-1} = -\frac{\partial^2 \log B(\gamma)}{\partial(\gamma\gamma^T)}\bigg|_{\gamma=\tilde{\gamma}}. \tag{5.6}$$

Let $t\gamma(\nu, \tilde{\gamma}, \boldsymbol{T})$ denote a multivariate t-density for γ, with ν degrees of freedom, mean vector $\tilde{\gamma}$, and precision matrix $\boldsymbol{T}$. Then our possibly multimodal PPMT approximation may be expressed in the form

$$\pi^*(\gamma \mid \boldsymbol{y}) = \sum_{(i)\in\Delta_m} I[\gamma \in \Lambda^{(i)}] t\gamma(\nu, \tilde{\gamma}^{(i)}, \boldsymbol{T}^{(i)}) \qquad (\gamma \in R^m), \tag{5.7}$$

where $\tilde{\gamma}^{(i)}$ maximizes $B(\gamma)$, for $\gamma \in \Lambda^{(i)}$, $I[A]$ denotes the indicator function for the event A, and

$$\boldsymbol{T}^{(i)} = \nu(\boldsymbol{D}^{(i)})^{-1}/(\nu + m). \tag{5.8}$$

The approximation (5.7) preserves permutability and global modality properties, of the exact density (5.3), and is continuous at all boundary points of each $\Lambda^{(i)}$. The degrees of freedom ν should be chosen pragmatically, to ensure that the PBM simulation procedure, described below, works well. The choice $\nu = \infty$ can be inefficient if the tails of (5.3) are much thicker than the tails of a multivariate normal density. With $m = 20$, the choice $\nu = 40$ often works well, and leads to somewhat smoother convergence when compared with Hsu (1990).

The approximation (5.7) may be used to compute the marginal posterior distribution, or moments, of the permutable parameter of interest η in (5.2), providing an exact purely Bayesian approach, as follows:

Simulation Procedure (PBM)

(a) Simulate a large number M of independent vectors $\gamma_1, \ldots, \gamma_M$, from a distribution for γ which truncates a multivariate t-distribution with ν degrees of freedom, mean vector $\tilde{\gamma}$ and precision matrix $\boldsymbol{T} = \nu\boldsymbol{D}^{-1}/(\nu+m)$, satisfying (5.6), to the region $\Gamma = \{\gamma :$ the elements of γ match the ordering of the elements of $\tilde{\gamma}\}$. The set Γ is fully explained during rejection step (a2) below.

(b) Importance Sampling: Simulate the exact posterior c.d.f. of any permutable function $\eta = \tau^*(\gamma)$, of γ from

$$F(\eta) = \sum_{j=1}^{M} I[\tau^*(\gamma_j) \leq \eta] W(\gamma_j) / \sum_{j=1}^{M} W(\gamma_j), \tag{5.9}$$

with

$$W(\gamma) = B(\gamma)/t\gamma(\nu, \tilde{\gamma}, \boldsymbol{T}), \tag{5.10}$$

where $B(\gamma)$ satisfies (5.3).

Step (a) of the above procedure can be performed as follows:

(a1) Simulate a large number $\gamma_1, \gamma_2, \ldots$ of independent multivariate t-vectors with ν degrees of freedom, mean vector $\tilde{\gamma}$ and precision matrix $\boldsymbol{T} = \nu \boldsymbol{D}^{-1}/(\nu+m)$, satisfying (5.6),

(a2) Rejection Step. In (a1) reject any simulated γ's which are inconsistent with the ordering $\tilde{\gamma}_1 \leq \tilde{\gamma}_2 \leq \cdots \leq \tilde{\gamma}_m$ of the elements of the vector $\tilde{\gamma}$. If the $\tilde{\gamma}_k$ have collapsed into $\tilde{q}$ subsets (see Section 2) then consistency with their ordering is only required between subsets. This defines matching rule for the set Γ introduced at step (a). Compute M unrejected vectors $\gamma_1, \gamma_2, \ldots, \gamma_M$.

In order to understand the above simulation procedure, it is important to note that

(i) The simulation in (a) do not provide simulations from the PPMT distribution (5.7), owing to the truncation to the region Γ. However, if $\eta = \tau^*(\gamma)$ is a permutable function of γ, then simulating values $\eta_1, \ldots, \eta_M$ for η in this way, is equivalent to simulating $\gamma_1, \gamma_2, \ldots, \gamma_M$. from (5.7) and then calculated $\eta_1 = \tau^*(\gamma_1), \ldots, \eta_M = \tau^*(\gamma_M)$. Furthermore, the denominator of (5.10) should remain the same in both cases.

(ii) Rejection Step (a2) is essential for the current procedure. Without this step, we would just be simulating from a straightforward multivariate t-distribution. In practice, we have found that this does not lead to convergence of the simulation procedure, within any reasonable time limits.

(iii) Steps (a1) and (a2) could be replaced by any efficient procedure for simulating from the truncated multivariate t-distribution, described at Step (a), for example, based upon conditional distributions of the elements of γ, or upon the Gibbs sampler.

The above approach permits a variety of prior formulations, though a proper prior distribution is often needed to ensure a proper posterior distribution. One flexible possibility is to assume that γ belongs to a piecewise permutable multivariate normal (PPMN) family, with permutable density of the form

$$\pi(\gamma \mid \mu, \boldsymbol{C}) = \sum_{(i) \in \Delta_m} I[\gamma \in \Delta^{(i)}] \Psi_\gamma(\mu^{(i)}, \boldsymbol{C}^{(i)}) \tag{5.11}$$

where $\mu^{(i)}$ and $\boldsymbol{C}^{(i)}$ provide appropriate permutations of a specified prior modal vector μ and prior dispersion matrix $\boldsymbol{C}$, and the Ψ contribution to (5.11) represents a multivariate normal density for γ, with mean vector $\mu^{(i)}$, and covariance matrix $\boldsymbol{C}^{(i)}$. Note that the assumption of a simple multivariate normal prior density could cause considerable problems concerning posterior multimodality.

6. Bayesian inference for survivor distributions

Consider the special case of the assumptions described in Section 1, where, for $i = 1, \ldots, n$, $\Omega_i = (c_i, \infty)$, and $\lambda(t, b) = b\exp\{-bt\}$ for $0 < t < \infty$, and $0 < b < \infty$, with $c_1, \ldots, c_n$ denoting fixed censoring times. Then (1.2) yields a mixture of exponential densities which constrains the density to be decreasing and concave. One possibility is to seek a power transformation δ such that $x_1^\delta, \ldots, x_n^\delta$ possess the density (1.2). In this case $x_i, x_2, \ldots, x_n$ possess a mixture of Weibull distributions with common power parameter δ. The parameter δ may be chosen pragmatically (see below).

For our equally weighted exponential mixture model, with parameters $\xi_1, \ldots, \xi_m$, we assume the following permutable prior distribution, also recommended by Gelfand and Smith (1990), for several Poisson means:

Stage I. Given α and β, $\xi_1, \ldots, \xi_m$ are independent and Gamma distributed, with common mean α/β and variance α/β^2.

Stage II. Given κ and ζ, β is Gamma distributed with mean κ/ζ and variance κ/ζ^2.

Hence, $\xi_1, \ldots, \xi_m$ are taken to possess a permutable scale transformed F-distribution, with density

$$\pi(\boldsymbol{\xi}) \propto \prod_{k=1}^{m} \xi_k^{\alpha-1} \Big/ \left(\sum_{k=1}^{m} \xi_k + \zeta\right)^{m\alpha+\kappa}. \tag{6.1}$$

The parameters $\xi_0 = \zeta\alpha/\kappa$ provides a common prior estimate for each ξ_k, $\alpha+k$ measures the closeness of each ξ_k to ξ_0, and κ measures the common variability of each ξ_k. Consider the transformations $\gamma_k = h(\xi_k) = \log \xi_k$, for $k = 1, \ldots, m$. Then a posterior mode vector $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_m)^T$ satisfying $\tilde{\gamma}_1 \leq \tilde{\gamma}_2 \leq \cdots \leq \tilde{\gamma}_m$, and maximizing the posterior density of $\gamma_1, \ldots, \gamma_m$ is straightforward to compute, together with a dispersion matrix $\boldsymbol{D}$, satisfying (5.6), by obvious generalizations of the maximum likelihood techniques described in Section 2. Problems in achieving a global maximum are not so acute, owing to the influence of the prior. Further details are described by Hsu (1990). Hence the PBM procedure of Section 5 may be readily employed. Parameters of interest include, for fixed t, the survivor function

$$\eta = \tau^*(\boldsymbol{\gamma}) = G(t) = m^{-1} \sum_{k=1}^{m} \exp\{-e^{\gamma_k} t^\delta\} \tag{6.2}$$

together with the corresponding survivor density. These procedures provide alternatives to the existing methodology referenced by Cox and Oakes (1985), and Kalbfleisch and Prentice (1980), and the Bayesian approaches due to Susarla and Van Ryzin (1978), Burridge (1981), and Sweeting (1987).

The data in Table 3 provides $n = 52$ observations and censoring times in weeks for the survival times of patients subject to an oral treatment for colon cancer (see Ansfield *et al.* (1977)). A total of 45 observations were uncensored and 7 were censored. The Kaplan-Meier estimate (see Kaplan and Meier (1958)) of the survivor function is the step function, described by curve (c) of Fig. 1. With $m = 20$,

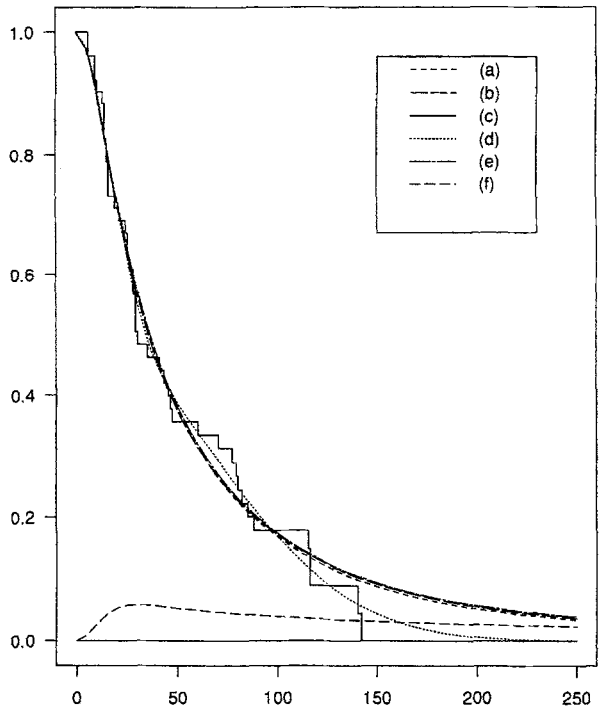

Fig. 1. Estimated survivor function. (a) Posterior mean value function (PBM, M = 200,000). (b) Posterior mean value function (PBM, M = 50,000). (c) Kaplan-Meier estimate. (d) Maximum likelihood estimate. (e) Posterior mean value function (Gibbs sampler, M = 50,000). (f) Posterior standard deviation function (PBM, M = 200,000).

Table 3. Possibly censored survival times.

Uncensored observations: 6 6 9 9 10 13 14 14 14 15 15 16 16 16 19 21 24 25 25 26 28 28 29
29 29 30 35 41 43 45 46 47 60 70 77 79 80 82 85 88 115 116 116 140 142
Censoring times for censored observations: 19 23 25 56 89 111 134

we firstly estimated $\xi_1, \ldots, \xi_m$ by the maximum likelihood procedure of Section 1. We attempted to jointly maximize the likelihood of $\xi_1, \ldots, \xi_m$ and the power parameter δ. However, the maximum likelihood estimate of δ was unconvincingly large, and did not lead to a good fit to the Kaplan-Meier estimate. A large range of choices of δ did lead to a good fit, and we chose $\delta = 2$ for simplicity, and hence fit an equally weighted mixture of exponential distributions to the squares of the observations. When $m = 20$, our procedure suggested a mixture of two exponential distributions assigning weights 0.50 and 0.50 to the (squared) locations 544.29 and 9183.59. The corresponding estimate of the survivor function of the original observations is described in curve (d) of Fig. 1, and closely fits and smoothes Kaplan-Meier estimator, in particular extrapolating beyond the last uncensored

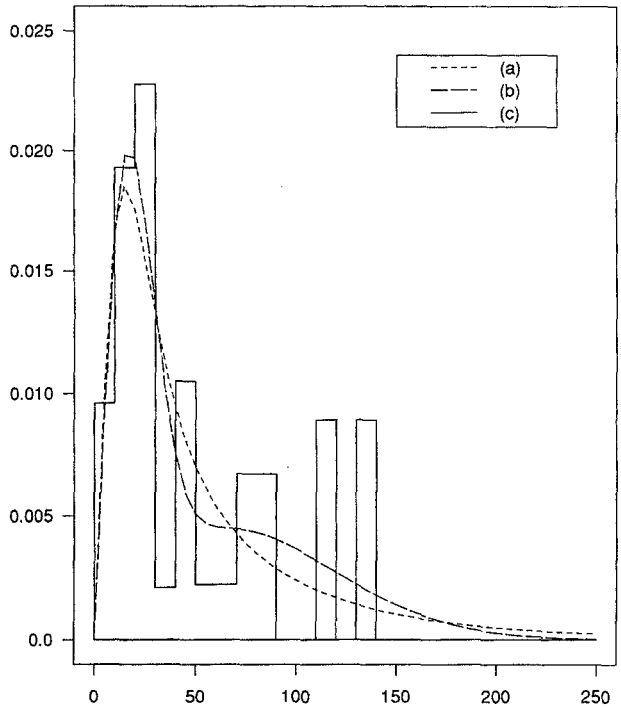

Fig. 2. Estimated survivor density. (a) Posterior mean value function (PBM, M = 200,000). (b) Maximum likelihood estimate. (c) Grouped histogram based on Kaplan-Meier.

observation. Curve (b) of Fig. 2 estimates the common density of the original observations, by a mixture of Weibull densities, and this closely fits an observed histogram, obtained by grouping the steps of the Kaplan-Meier estimates. When $m = 200$, very similar estimates of the survivor function and survivor density were obtained. The maximum likelihood estimate of the distribution of the squared observations, when $m = 200$, assigns weights 0.515 and 0.485 to the locations 545.48 and 9328.99. We have provided a simple way of smoothing the Kaplan-Meier estimate which would possess the advantage of reducing to a simple parametric model if $\hat{q} = 1$.

To illustrate the Bayesian procedure, we assumed the above two stage prior distribution, with $k = \alpha = 0.5$, and $\xi_0 = 1/52^2$, corresponding to a prior estimate of 52 weeks for the mean of the survivor distribution, and $\delta = 2$. With $m = 20$, 12 of the posterior modes of $\gamma_1, \ldots, \gamma_{20}$ collapsed to $\log 760.284$, and eight collapsed to $\log 8437.747$, again suggesting a mixture of two exponential distributions.

Using PBM we simulated M = 200,000 unrejected vectors for $\gamma = (\gamma_1, \ldots, \gamma_{20})^T$ with degrees of freedom $\nu = 40$ for our PPMT approximation. However, in many applications M = 50,000 provides reasonable results. About 10% of the vectors were unrejected, and our Bayesian computations are correct to at least 3 decimal places. Curves (a) and (f) of Fig. 1 respectively describe the exact posterior mean value function and posterior standard deviation function for the

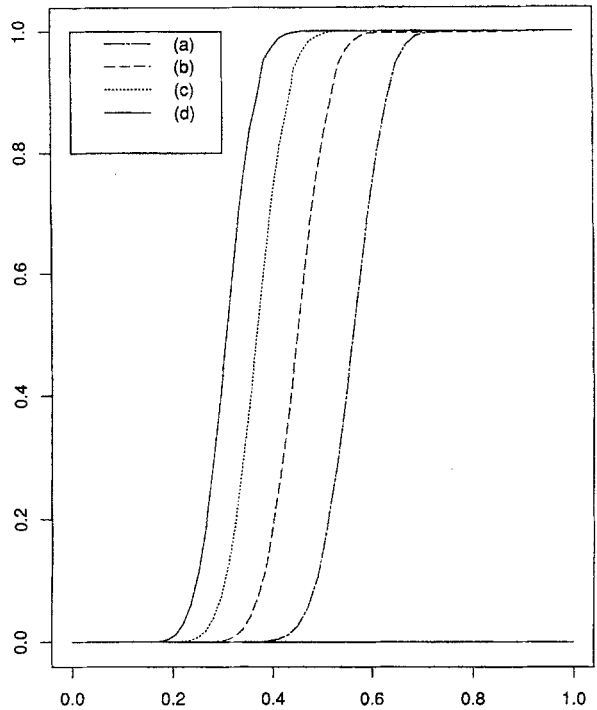

Fig. 3. Posterior c.d.f.'s of survival probabilities. (a) Posterior c.d.f. of $G(30)$. (b) Posterior c.d.f. of $G(40)$. (c) Posterior c.d.f. of $G(50)$. (d) Posterior c.d.f. of $G(60)$.

survivor function $G(t)$ in (6.2). Curve (a) of Fig. 2 describes the posterior mean value function of the survivor density. In Fig. 3, curves (a), (b), (c), and (d) respectively describe the posterior c.d.f.'s of $G(30), G(40), G(50)$, and $G(60)$, the probabilities of survival beyond 30, 40, 50, and 60 weeks.

7. Comparison with the Gibbs sampler

Lavine and West (1992) suggest an elegant computational scheme for unequally weighted mixtures of multivariate normal distributions, based on the Gibbs sampler. This, for example, permits the computation of posterior probabilities that a given observation belongs to a particular subpopulation, a side benefit not obviously available under our model (1.2). Extensions of their method require the ability to simulate from a variety of conditional distributions which therefore need to assume analytically tractable forms. In such special cases their procedure is easier to compute when compared with PBM. However, the Lavine-West approach is not generally applicable, for example (a) when (5.11) represents the prior density and the sampling distribution is not a mixture of multivariate normal densities and (b) the sampling distribution is not a mixture of distributions for which simple conjugate families of prior distributions exist. It seems difficult to apply any form of the Gibbs sampler to many situations, though it is possible to apply the Lavine-West approach to mixtures of multivariate densities in a number of special cases.

The Lavine-West approach can be extended, for example, to the special case discussed in Section 6. Moreover, by considering the equally weighted mixture model (1.2), rather than assumption (1.4), the computations can be somewhat simplified. Suppose that $z_1, z_2, \ldots, z_n$ are independent and unobservable polychotomous variables, each equal to j, with probability m^{-1} for $j = 1, \ldots, m$. Then, under the assumptions of Section 6, $x_1, x_2, \ldots, x_n$ denote independent random variables, such that, conditional on $z_i = j$, x_i has an exponential distribution with mean ξ_j^{-1} $(i = 1, \ldots, n)$.

Consequently, under the prior density (6.1)

(1) conditional on $z_1, \ldots, z_n$, and β, the parameters $\xi_1, \ldots, \xi_m$ are a posteriori independent. For $j = 1, \ldots, m$, the parameter ξ_j conditionally possesses a Gamma distribution, with parameters $\alpha + \sum_i d_i I[z_i = j]$ and $\beta + \sum_i u_i I[z_i = j]$.

(2) conditional on $\xi_1, \ldots, \xi_n$ and β, the polychotomous variables $z_1, \ldots, z_n$ are a posteriori independent, with

$$p(z_i = j) = \xi_j^{d_i} \exp\{-\xi_j u_i\} / \sum_{k=1}^{m} \xi_k^{d_i} \exp\{-\xi_k u_i\} \qquad (j = 1, \ldots, m). \tag{7.1}$$

(3) Conditional upon $\xi_1, \ldots, \xi_m$, and $z_1, \ldots, z_n$, the prior parameter β has a Gamma distribution, with parameters $m\alpha + \kappa$ and $m\xi_\cdot + \zeta$.

Based upon the conditional distributions in (1), (2), and (3), Gelfand and Smith (1990) tell us that we may obtain a simulation for $\xi_1, \xi_2, \ldots, \xi_m$ from their unconditional distribution. Starting with some initial $z_1, \ldots, z_n$ and β, simulate values for $\xi_1, \ldots, \xi_m$ from the independent Gamma distributions in (1). Then simulate new values for $z_1, \ldots, z_n$ from the polychotomous distributions in (2) based upon the latest values for $\xi_1, \ldots, \xi_m$. Next simulate a value for β from the Gamma distribution in (3), using the new $z_1, \ldots, z_n$. Return to (1), and keep cycling, always using the latest simulated values for the conditional variables. Ultimately, the values of $\xi_1, \ldots, \xi_m$ will converge to a single simulation from their unconditional distribution. There are similarities with the data augmentation approach due to Tanner and Wong (1987).

Moreover, the Ergodic theorem described by Gelfand and Smith tells us that, if we take all realizations of $\xi_1, \ldots, \xi_m$ in our iterative sequence, and the average value for any parameter $\eta = g(\xi_1, \xi_2, \ldots, \xi_m)$ of interest, then, in the long run, this average will converge to the posterior expectation of η. There are some difficulties in judging when the average value of η has converged. Note that the successive simulations for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)^T$ are serially correlated, and that this can give rise to the phenomenon of "apparent convergence" (Edward George, Nick Polson, personal communication). Sometimes, the sequence can appear to converge to three decimal places, but then slightly diverge again. However, comparison with our results based upon importance sampling (e.g., see below) suggests that 50,000 replications of η can lead to an apparent convergence which is reasonably close to actual convergence (within about two decimal places if η is a posterior probability, with possibly greater accuracy if η is a posterior moment). Note that the simulations for $z_1, \ldots, z_n$ can become tedious if n is large, in which case our importance sampling approach can become quite appealing, as approximation (5.7) with $\nu \to \infty$ can become very accurate.

The Gibbs sampling procedure for the posterior mean value function of the survivor function provided curve (e) of Fig. 1, after 50,000 simulations and this is correct to about two decimal places compared with the essentially exact curve (a) based upon PBM and extensive importance sampling, though slightly less close than curve (b), based upon PBM but with only 50,000 unrejected vectors.

We conclude that apparent convergence of the Gibbs sampler is good enough for practical purposes in this special case, and for this moderate sample size. However, PBM seems more useful as a general paradigm, or when high accuracy (based upon independent simulations) is required, since without PBM it is difficult to confirm that the Gibbs sampler has indeed converged. Note that accurate results may also be obtained by using techniques suggested by Ogata (1989, 1990), but tremendous computer time will also be needed to completed the computations.

Acknowledgements

Thanks are due to Barry Storer and Jerry Klotz, for providing the colon cancer data, Michael West for several useful technical reports and for his permission to report the Lavine-West approach, Rebecca DerSimonian for providing details of her maximum likelihood algorithm, Bradley Carlin for introducing the authors to the Gibbs sampler, Edward George and Nick Polson for discussing "apparent convergence" of the Gibbs sampler. We are also indebted to two refrees and Graham Wood for particularly sound and constructive suggestions.

References

Alspach, D. L. (1975). A Gaussian sum approach to the multi-target identification tracking problem, *Automatica*, **11**, 285–296.

Ansfield, F., Klotz, J., Nealton, T., Ramirez, G., Minton, J., Hill, G., Wilson, W., Davis, H. and Cornell, G. (1977). A phase III study comparing the clinical utility of four regimens of 5-Fluorouracil, *Cancer*, **39**, 34–38.

Bates, D. M. and Watts, D. G. (1988). *Non-Linear Regression Analysis and Its Applications*, Wiley, New York.

Burridge, J. (1981). Empirical Bayes analysis of survival time data, *J. Roy. Statist. Soc. Ser. B*, **43**, 65–75.

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

Cox, D. R. and Oakes, D. (1985). *The Analysis of Survival Data*, Chapman and Hall, New York.

DerSimonian, R. (1986). Maximum likelihood estimation of a mixing distribution algorithm, *Appl. Statist.*, **35**, 302–309.

Efron, B. (1967). The two sample problem with censored data, *Proc. Fifth Berkeley Symp. on Math. Statist. Prob.*, Vol. IV, 831–853, Univ. of California Press, Berkeley.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.*, **85**, 393–397.

Geweke, J. (1988). Antithetic acceleration of Monte-Carlo integration in Bayesian inference, *J. Econometrics*, **38**, 73–89.

Geweke, J. (1989). Exact predictive densities for linear models with arch distribution, *J. Econometrics*, **40**, 63–86.

Hsu, J. S. J. (1990). Bayesian inference and marginalization, Ph.D. Thesis, University of Wisconsin-Madison.

Hsu, J. S. J., Leonard, T. and Tsui, K. (1991). Statistical inference for multiple choice tests, *Psychometrika*, **56**, 327–348.

Kalbfleisch, T. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457–487.

Laird, N. M. (1978). Non-parametric maximum likelihood estimation of a mixing distribution, *J. Amer. Statist. Assoc.*, **73**, 361–379.

Laird, N. M. (1982). Empirical Bayes estimation using the non-parametric maximum likelihood estimate of the prior, *J. Statist. Comput. Simulation*, **15**, 211–220.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination, *Canad. J. Statist.*, **20**, 451–461.

Lenk, P. (1991). Toward a practicable Bayesian non-parametric density estimator, *Biometrika*, **78**, 531–544.

Leonard, T. (1972). Bayesian methods for binomial data, *Biometrika*, **59**, 581–589.

Leonard, T. (1978). Density estimation, stochastic processes and prior information (with Discussion), *J. Roy. Statist. Soc. Ser. B*, **40**, 113–146.

Leonard, T. (1984). Some data-analytic modifications to Bayes-Stein estimation, *Ann. Inst. Statist. Math.*, **36**, 11–21.

Leonard, T. and Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix, *Ann. Statist.*, **20**, 1669–1696.

Leonard, T., Hsu, J. S. J. and Tsui, K. (1989). Bayesian marginal inference, *J. Amer. Statist. Assoc.*, **84**, 1051–1057.

Lindsay, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution, *Statistical Distribution in Scientific Work* (eds. C. Taillie, G. Patial and B. Baldessari), **5**, 95–109, Reidel, Holland.

Lindsay, B. G. (1983). A geometry of mixture likelihoods Part II: The exponential family, *J. Amer. Statist. Assoc.*, **4**, 1200–1209.

Mallet, A. (1986). A maximum likelihood estimation method for random coefficient models, *Biometrika*, **73**, 645–656.

Ogata, Y. (1989). A Monte Carlo method for high-dimensional integration, *Numer. Math.*, **55**, 137–157.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure, *Ann. Inst. Statist. Math.*, **42**, 403–433.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.*, **27**, 832–835.

Rubinstein, R. Y. (1981). *Simulation and the Monte-Carlo Method*, Wiley, New York.

Simar, L. (1978). Maximum likelihood estimation of a compound Poisson process, *Ann. Statist.*, **4**, 1206–1209.

Sorenson, H. W. and Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums, *Automatica*, **7**, 465–479.

Sweeting, T. J. (1987). Approximate Bayesian analysis for censored survival data, *Biometrika*, **74**, 809–816.

Tanner, T. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation, *J. Amer. Statist. Assoc.*, **81**, 82–86.

Tapia, R. A. and Thompson, J. R. (1978). *Nonparametric Probability Density Estimation*, John Hopkins University Press, Baltimore.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities, *J. Amer. Statist. Assoc.*, **82**, 528–549.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.