Input Modelling for Multimodal Data: Using Bayesian Model Selection to Fit Finite Mixture Models

Russell C.H. Cheng, University of Southampton
Christine S.M. Currie, University of Southampton

We describe a method that can be used to fit finite mixture models to multimodal simulation input data. Mixture models provide a proper representation of an input stream that is an amalgam of data from different sources but are also convenient for describing multimodal data from a single source. A key problem in fitting finite mixture models is identifying the different components in the mixture and determining how many components there are. This is known to be a non-regular/non-standard problem in the statistical sense and is difficult to handle properly using classical inferential methods. The problem is most acute when there are components with a small variance. We describe a Bayesian approach particularly suited to handling this latter situation, which we have encoded in a publicly available program FineMix. Numerical examples are given showing its application and comparing it with other approaches showing the advantages of the method.

Additional Key Words and Phrases: Mixture models; input modelling; simulation; Bayesian statistics

1. INTRODUCTION

Input modelling for simulation aims to identify appropriate probability distributions for characterising the behaviour of the streams of random variables that represent the inputs to simulation models. There is a large literature on this topic and a good general reference is [Law 2007]. Input modelling literature tends to discuss relatively simple situations where input random variables are independently and identically distributed and drawn from well-known distributions such as the normal, lognormal, gamma or Weibull; although [Kuhl et al. 2010] discuss input modelling for a wider range of distributional shapes. Two generalisations have been studied in some detail, namely: (i) where the random variables are multivariate, and (ii) where they are correlated. See for example [Nelson and Yamnitsky 1998; Deler and Nelson 2001; Ghosh and Henderson 2001]. A third generalisation has not been so well discussed, where input random variables have a multimodal distribution, and most likely have a so-called finite mixture distribution. The purpose of this article is to discuss such distributions and their modelling.

A finite mixture distribution is the weighted sum of a finite number, denoted by k in this paper, of component distributions; the latter usually all belonging to one family like the normal. A finite mixture distribution is well suited to modelling data samples that are multimodal. This occurs quite naturally if the data is a mixture of different input sources each with a distinct distribution. Finite mixture models have a wide range of application; in this paper we give just three examples, two from real

Author's addresses: R.C.H. Cheng and C.S.M. Currie, Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

applications in manufacturing and finance, but in the Online Supplement we provide real data samples from over a dozen other application areas.

Computer sampling from finite mixture models is easily implemented to provide input models for discrete event simulation.

There is a very extensive literature on fitting finite mixture models to data and this is well reviewed in [McLachlan and Peel 2000] which lists many packages for fitting such models. Most of these are likelihood based including both maximum likelihood (ML) and Bayesian methods.

When fitting such a finite mixture model to data, a problem of particular interest is the estimation of k, the number of components that the model should have. Though several of the existing packages have provision for handling this problem, there still seems room for improvement. The EMMIX program described in [McLachlan and Peel 2000] is an example of a good implementation using ML estimation, carried out using the EM algorithm, and which uses the Akaike Information Criterion (AIC) for selecting k. However a problem can occur with the likelihood becoming infinite, discussed for example in [McLachlan and Basford 1988]. This is most likely when there are one or more components with a small variance, so that the sample contains one or more subsamples with tightly clustered observations. Any method based purely on the likelihood can become unstable when fitting a mixture model to such a sample.

The difficulty can be avoided using a Bayesian formulation. In this case either the Bayes Information Criterion (BIC) or the posterior probability distribution of k can be used for selecting k. A leading method using a Bayesian formulation is the reversible jump Markov chain Monte Carlo (RJMCMC) method described in [Richardson and Green 1997]. Though theoretically attractive, we have found that in practice RJMCMC seems over cautious when handling tightly clustered data; this is illustrated in our examples. In this paper we use a different Bayesian approach which overcomes this problem with RJMCMC.

In our Bayesian approach we use importance sampling (IS) to estimate k. Our method includes an initial point estimation of all the parameters of the model based on optimization of the posterior distribution. This initial optimization allows tightly clustered observations to be readily identified, thus avoiding the conservatism of RJMCMC.

The present paper concentrates on the methodology of our approach, but given that the fitting method needs to be sufficiently stable for reliable practical use, we have made a serious attempt to implement our method in a sufficiently robust form that can be used on genuinely demanding real data sets. Our implementation, which we have called FineMix, is a computer program written in C with an Excel interface, and is available for download at http://www.soton.ac.uk/ ccurrie/.

IS is a well known approach but has not been fully considered for finite mixture modelling in the literature. In contrast Markov chain Monte Carlo (MCMC) methods have received considerable attention. We believe however that use of parameter point estimators when carrying out IS has a definite advantage when it comes to fitting components with small variances. We will be discussing this and other advantages that we feel IS has over MCMC in what follows.

We consider a random sample $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ drawn from a finite mixture distribution, i.e. a distribution with probability density function (PDF) that is the weighted sum of a finite number of continuous PDFs:

$$f(y|\boldsymbol{\psi}(k), \mathbf{w}(k), k) = \sum_{i=1}^{k} w_i g(y|\boldsymbol{\psi}_i), \tag{1}$$

where $\boldsymbol{\psi}(k) = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_k)^T$ and $\mathbf{w}(k) = (w_1, w_2, ..., w_k)^T$. The w_i are the weights, satisfying $0 \leq w_1, w_2, ..., w_k \leq 1$, $\sum_{i=1}^{k} w_i = 1$. Each component has the same form of density $g(\cdot)$. We call $g(\cdot)$ the *base* density. The cases we consider are: normal, lognormal, extreme value (EV), negative extreme value (NEV), Weibull, gamma, and inverse gaussian (IG). We consider all these in their two parameter form, these two parameters being denoted individually by α and β, and in vector form by $\boldsymbol{\psi}_i$ in (1). Table 1 lists the densities of all these cases except the NEV (this being simply the negative of the EV), and the way that these distributions are conventionally defined. The quantities k, $\boldsymbol{\psi}(k)$ and $\mathbf{w}(k)$ are all assumed to be unknown. For reasons to be discussed it is useful to consider an alternative parametrization in terms of the mean μ and standard deviation σ. Also tabulated in Table 1 are the transformations expressing the more standard parameters α and β as functions of μ and σ.

Table I. Conventional parametrizations of base distributions considered in the paper, and these parameters as functions of the mean, μ, and standard deviation, σ, of the distribution; γ_E is Euler's constant, $\omega(\cdot)$ is as in eqn. 3

Base Distribution	PDF	$\alpha(\mu, \sigma)$	$\beta(\mu, \sigma)$
Normal	$\frac{1}{\sqrt{2\pi\beta^2}} \exp\left[-(y-\alpha)^2/2\beta^2\right]$	μ	σ
Lognormal	$\frac{1}{\beta\sqrt{2\pi}y} \exp\left[-\frac{1}{2}(\frac{\ln y - \alpha}{\beta})^2\right]$	$\ln\mu - \frac{1}{2}\ln(1+(\frac{\sigma}{\mu})^2)$	$\sqrt{\ln(1+(\frac{\sigma}{\mu})^2)}$
EV	$\frac{1}{\beta}\exp\{-(\frac{y-\alpha}{\beta}) - \exp[-(\frac{y-\alpha}{\beta})]\}$	$\mu - (\gamma_E\sqrt{6}/\pi)\sigma$	$(\sqrt{6}/\pi)\sigma$
Weibull	$\frac{\alpha}{\beta}(y/\beta)^{\alpha-1}\exp[-(y/\beta)^{\alpha}]$	$\omega(\sigma/\mu)$	$\mu/\Gamma\left[1+\frac{1}{\omega(\sigma/\mu)}\right]$
Gamma	$\frac{y^{\alpha-1}\beta^{-\alpha}\exp(-y/\beta)}{\Gamma(\alpha)}$	$(\mu/\sigma)^2$	σ^2/μ
IG	$\sqrt{\frac{\alpha}{2\pi y^3}}\exp\left[-\frac{\alpha(y/\beta-1)^2}{2y}\right]$	μ^3/σ^2	μ

The estimation of the number of components k in (1) is non-standard, at least when using the frequentist likelihood approach. Two problem issues have been discussed in detail in the literature.

P1. *Ambiguous parametric model specification.* This is not simply the replicated parameter space problem (discussed for example in [Titterington et al. 1985] and in [Richardson and Green 1997]), and which is easily handled, but concerns the fact that the parametrization of (1) does not provide an injective mapping of the parameter space onto the set of actual models (i.e. different parameter combinations can give rise to the same model). This problem has been discussed in [Feng and McCulloch 1996] and in [Cheng and Liu 2001] from the classical frequentist viewpoint.

P2. *Statistical consistency.* Suppose a random sample is drawn from a mixture distribution (1) where there is a 'true' but unknown value for k, $\boldsymbol{\psi}(k)$ and $\mathbf{w}(k)$. We would expect and want an estimate of these quantities to converge to these true values as the sample size increases. This is mainly a theoretical problem and has been discussed in [Barron et al. 1999], but it is one which is not universally accepted as pertinent in the Bayesian context.

The above references provide a good understanding of problems P1 and P2, and we will not need to discuss them further here. However there is one important third problem (discussed for example in [McLachlan and Basford 1988]) which deserves mention especially in view of our wish to be able to fit components with a small variance. The problem does not really arise from the above two theoretical aspects, but stems from the fact that, by their nature, mixture models are extremely flexible. This very flexibility gives rise to the following problem

P3. *Limiting Discrete Components.* A finite mixture model includes limiting mixed discrete/continuous component models at the boundaries of the parameter space.

These models correspond to situations where one or more of the components in (1) collapse into delta functions representing discrete probability atoms located at individual sample points y. The likelihood, instead of tending to zero as such points are approached, actually tends to infinity. For example, consider fitting a normal mixture with just two components to a sample $\{y_i\}$ using maximum likelihood. Consider the line in the parameter space obtained by varying $\sigma_2 > 0$, but with $w_1, w_2, \mu_1, \sigma_1$ all fixed and positive and with $\mu_2 = y_i$ for any fixed i. On this line the loglikelihood takes the form

$$L = \log\{w_1\varphi(y_1; \mu_1, \sigma_1)] + w_2(2\pi)^{-1/2}\sigma_2^{-1}\} + \sum_{i=2}^{n} \log\{w_1\varphi(y_i; \mu_1, \sigma_1) + w_2\varphi(y_i; \mu_2, \sigma_2)\},$$

where $\varphi(\cdot, \mu, \sigma)$ is the normal PDF with mean μ and SD σ. It is evident the first term tends to infinity as $\sigma_2 \to 0$, whilst the summation is bounded below with

$$\sum_{i=2}^{n} \log\{w_1\varphi(y_i; \mu_1, \sigma_1) + w_2\varphi(y_i; \mu_2, \sigma_2)\} > \sum_{i=2}^{n} \log\{w_1\varphi(y_i; \mu_1, \sigma_1)\}.$$

Thus $L \to \infty$ as $\sigma_2 \to 0$. As $\mu_2 = y_i$ for any i, this shows that globally maximizing the likelihood is not meaningful.

The practical problem which arises is that we would wish our fitting method to be able to fit components with a small variance (these giving rise to a sharp 'spike' in the full mixture density) when such components are really present, but at the same time avoiding the method becoming unstable in trying to fit delta function spikes to what may be just random clustering in the data.

A simple but somewhat arbitrary way of handling the difficulty is to impose constraints on parameter values that bar such discrete component spikes. Sieve methods (see for example [Barron et al. 1999]) act in this way, but subtly, allowing the constraints to be relaxed as sample size increases. However for finite samples a sieve method still excludes small portions of the parameter space.

This difficulty is one of the reasons why we have used a Bayesian approach. We still need to ensure that boundary points of the parameter space, corresponding to unrealistic spikey models, are avoided. However with a Bayesian approach, this can be done in a fairly natural and automatic way by choice of an appropriate prior distribution for the parameters.

In our approach we have used IS to estimate the posterior distribution. This is done by sampling from a *importance sampling distribution* and reweighting the sampled values. (See e.g. [Hammersley and Handscomb 1964] or [Robert and Casella 1999] for more details.) It is well known (see [Geweke 1989] for example) that the IS method is most reliable if the IS distribution is a good approximation of the posterior distribution. We have followed [Geweke 1989] in the way we construct the IS distribution, but modified to allow k to be treated as a parameter as well.

Before discussing our method in more detail, we review two alternative approaches which we shall be comparing with our approach.

There has been much work where the Bayesian approach is implemented using the Markov Chain Monte Carlo (MCMC) method to find the posterior probability distribution of the data [Richardson and Green 1997; Phillips and Smith 1996; Stephens 2000; Raftery 1996; Cheng 1998; Escobar and West 1995]. An MCMC algorithm is proposed in [Stephens 2000] that creates a Markov birth-death process, in which components of the mixture model are allowed to be born and to die, allowing movement between models corresponding to different values of k. Other methods of estimating the marginal likelihood such as the Laplace-Metropolis estimator, the candidate's estimator and the data augmentation estimator are discussed in [Raftery 1996].

A particularly attractive Bayesian method is the reversible jump (RJMCMC) version, as described in [Richardson and Green 1997] and [Phillips and Smith 1996]. In RJMCMC, the steps in the Markov chain can either be discrete transitions between different models with different k (jumps) or, between jumps, are simply changes in model-specific parameter values (diffusion). Reversibility of the jumps (as defined in Richardson and Green) ensures that the Markov chain is ergodic with stationary distribution equal to the joint posterior distribution of the parameters, including k.

Though RJMCMC is arguably the most highly regarded of the MC methods to date, the reversible jump calculations are quite elaborate to set up. The formulas used are very dependent on the form of the base density, and to date only the normal and exponential cases have been considered in any detail.

For all MCMC methods, the steps in a run of the MC are not independent. This means that the convergence of estimators is not easy to determine. One consequential complication is that sampling in MCMC requires a 'burn-in' or 'warm-up' period to achieve steady state in the chain before sampled observations can be recorded. In contrast when using IS the sampling replications are essentially independent. The observations are not completely independent as all observations involve a normalizing constant dependent on all the observations. Convergence statistics are nevertheless straightforward to calculate in IS. An interesting consequence, which we mention here but do not consider further, is that independence of replications allows IS to be easily extended to parallel thread computing environments without loss of efficiency.

Though the Bayesian approach does not target point estimators as being of special interest, we feel that in the case of finite mixtures such estimators are valuable in assessing how individual components contribute to the mixture. Our implementation of IS makes use of such point estimators and highlights their use in the interpretation of the results. Such estimators can be obtained using RJMCMC, but it is clear from the discussion in [Richardson and Green 1997], that the nature of the RJMCMC process means that the contribution of individual components is not so easy to assess. We will return to this point when considering the examples.

Perhaps the most important comparison of RJMCMC and IS is how they cope with P3, the difficulty of limiting discrete components discussed earlier. Our IS method involves optimization of the posterior distribution for each k, with k increasing stepwise, prior to carrying out the sampling. This ensures that potential components with small variances are quickly identified. Possible instability in the optimization process arising from the difficulty discussed in P3, can be controlled by choice of prior. One of the numerical examples discussed contains a very spikey component which our IS method readily identifies. In contrast the RJMCMC method does not involve an optimization process, and this seems to render it very conservative in identifying components with small variances. Rather than being affected by P3 it fails to correctly identify spikey components in the numerical examples, fitting an overall mixture density that is overly smooth. We have observed this not only in the numerical examples presented but with other similar examples, and it seems to be a characteristic difference between the two methods.

A rather different approach to fitting a distribution to multimodal data has been proposed by Wagner and Wilson [Wagner and Wilson 1996b] using a sum of Bézier curves to describe the cumulative distribution function (CDF). The approach is very practically oriented and Wagner and Wilson have provided a very user-friendly implementation with a graphical interface, called PRIME, publicly available at www.ise.ncsu.edu/jwilson/page3. It is able to optimise the number of control points used in forming the Bézier distribution function and find their optimal positions using one of a number of different fitting criteria. Bézier distributions provide flexible fits to multimodal distributions, although the authors do admit that it is often necessary to do

some work setting up the initial conditions for PRIME when fitting to multimodal data ([Wagner and Wilson 1996a] and personal communication). In the numerical section we give an example to compare our IS method with Wagner and Wilson's method. It should be said that the components of a mixture model arguably have a more intuitive meaning than the parameters in the Bézier distribution. One immediate consequence is that, in terms of generating random variates for simulation use, it would seem easier to implement a mixture model.

In the following sections, we first set out the general Bayesian approach to the fitting models more formally, we then discuss the case of finite mixture models, describing the priors to be used in this case, and then our proposed IS approach for such models. Three examples are discussed in detail in the Examples Section.

2. BAYESIAN ANALYSIS

It is simplest to summarize the Bayesian approach in general terms first, before discussing its application to finite mixture models.

Let $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ denote a random sample drawn from a continuous distribution with density $f(y|\boldsymbol{\theta})$ depending on a d-dimensional vector of parameters $\boldsymbol{\theta}$, with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, some region in d dimensional space. We assume that $\boldsymbol{\theta}$ is not precisely known, but is stochastic with some *prior distribution* having density $\pi(\boldsymbol{\theta})$. Denote the *likelihood* of the sample by $f(\mathbf{y}|\boldsymbol{\theta})$. The *posterior distribution* then has density

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Thus the posterior distribution incorporates the best information that we have about the distribution of $\boldsymbol{\theta}$ from the data and our initial prior information, and the main objective of Bayesian analysis is to estimate this posterior distribution.

Consider the estimation of $\pi(\boldsymbol{\theta}|\mathbf{y})$ by IS. The general methodology is set out clearly in [Geweke 1989], which we summarize here. Note that $\pi(\boldsymbol{\theta}|\mathbf{y})$ is proportional to $p(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Most quantities of interest associated with $\pi(\boldsymbol{\theta}|\mathbf{y})$ can be expressed as an expectation under the posterior:

$$E[h(\boldsymbol{\theta})] = \frac{\int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}},$$

where $h(\boldsymbol{\theta})$ is a suitably defined function of $\boldsymbol{\theta}$. In particular this includes probabilities of the form $\Pr(\boldsymbol{\theta} \in \mathbf{A})$, which we can express as

$$\Pr(\boldsymbol{\theta} \in \mathbf{A}) = E[I_A(\boldsymbol{\theta})],$$

where $I_A(\boldsymbol{\theta})$ is the indicator function with $I_A(\boldsymbol{\theta}) = 1$ if $\boldsymbol{\theta} \in \mathbf{A}$, $I_A(\boldsymbol{\theta}) = 0$ otherwise.

Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..$ be a sequence of independent and identically distributed random variables from a continuous distribution with density $c(\boldsymbol{\theta})$, which we call the *IS distribution*. Define the *IS ratios* $\rho(\boldsymbol{\theta}_i) = p(\boldsymbol{\theta}_i|\mathbf{y})/c(\boldsymbol{\theta}_i)$. Then, provided $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a proper density function defined on Θ (i.e. $\pi(\boldsymbol{\theta}|\mathbf{y}) \geq 0$ and integrable on Θ, with $\int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = 1$), $c(\boldsymbol{\theta})$ has support that includes Θ, and $E[h(\boldsymbol{\theta})]$ is finite, we have (Theorem 1 in [Geweke 1989])

$$\bar{h}_m = \sum_{i=1}^{m} h(\boldsymbol{\theta}_i)\rho(\boldsymbol{\theta}_i) / \sum_{i=1}^{m} \rho(\boldsymbol{\theta}_i) \rightarrow E[h(\boldsymbol{\theta})] \text{ almost surely as } m \rightarrow \infty \tag{2}$$

In practice convergence can be slow if, loosely speaking, $c(\boldsymbol{\theta})$ does not mimic $\pi(\boldsymbol{\theta}|\mathbf{y})$, or $p(\boldsymbol{\theta}|\mathbf{y})$, sufficiently well. The main requirement is that $c(\boldsymbol{\theta})$ does not tail off too fast compared with $p(\boldsymbol{\theta}|\mathbf{y})$.

In the next Section we describe the prior distributions that we propose using when fitting finite mixture models.

3. PRIOR DISTRIBUTIONS

Turning now to the model (1), we consider first the prior distributions to be used for k, $\boldsymbol{\psi}(k)$ and $\mathbf{w}(k)$. We focus on situations where little or no prior information exists about these parameter values. We will however *not* use improper priors that are fully non-informative, but proper priors which we can adjust in order to handle the already mentioned problem P3.

To make it explicit that k, $\boldsymbol{\psi}(k)$ and $\mathbf{w}(k)$ are being treated as stochastic variables we denote them by capitals, reserving lower case for specific values. We write K for the number of components, $\boldsymbol{\Psi}(K) = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, ..., \boldsymbol{\Psi}_K)^T$ for the component distribution parameters, and $W(K) = (W_1, W_2, ..., W_K)^T$ for the weights.

For the six base densities $g(.)$ considered in this paper: the normal, lognormal, EV, Weibull, gamma and IG, all in their two parameter form, we shall use the mean M and standard distribution S as the parameters (as with K and $W(K)$, we use upper case M and S to indicate that they are stochastic variables, reserving μ and σ for specific values of M and S). Thus $\boldsymbol{\Psi}_i = (M_i, S_i)$. We use this parametrization rather than more conventional ones for the following reasons: (i) it is usually easier to study and discuss the behaviour of different components in terms of their location and spread, and indeed this use of the mean and variance has been made by previous authors; (ii) it also enables the fits obtained using different base distributions to be more easily compared; and (iii) finally, and perhaps most importantly, we found that use of M and S gave rise to significantly more stable and consistent behaviour in the numerical optimization methods used in calculating the posterior distribution.

For the six base distributions we consider, it is easy to express the mean μ and standard deviation σ in terms of the standard parametrizations appearing in the literature, and, except in the case of the Weibull, these relationships are easily inverted to give the conventional parameters in terms of μ and σ. Table 1 lists these relationships. Thus it is easy to set out our numerical procedures in terms of how μ and σ are updated, but calculate actual density and probability values in terms of the conventional parametrization.

For the Weibull case, the shape parameter, α in Table 1, is an explicit function of the coefficient of variation $\gamma = \sigma/\mu$. We write this function as $\alpha = \omega(\gamma)$. A simple approximation for $\omega(\gamma)$ is

$$\alpha = \exp\left(0.5282 - 0.7565t - 0.3132\sqrt{6.179 - 0.5561t + 0.7057t^2}\right) \tag{3}$$

where $t = \ln(1 + \gamma^2)$, which has a relative error of less than 1% in the range $0.0001 \leq \gamma \leq 1000$. This is derived in Appendix A.2. Using this approximation we are thus able to express the usual parameters in terms of μ and σ over a reasonably practical range of values, so that in the Bayesian analysis the Weibull distribution can be handled in exactly the same way as the other base distributions.

The choice of priors even just for finite mixture distributions has been addressed by a number of authors. Most of the issues they raise are not of major concern in our work, so we cite here only authors whose work is directly relevant to the priors that we chosen. For the interested reader we provide a fuller review of the choice of priors in the Online Supplement.

For the priors of K and $W(K)$ we follow [Richardson and Green 1997] and [Roeder and Wasserman 1997].

We use a discrete uniform distribution as our prior distribution for K, namely

$$p_K(k) = \Pr\{K = k\} = 1/k_{max},\ 1 \le k \le k_{max}$$

and zero for all other values of k. Here $k_{\max}$ is a prescribed maximum number of components, which it is assumed will definitely not be exceeded.

The prior for the component weights $W(K)$, is defined by conditioning on K. For given $K = k$ we use the Dirichlet distribution with density

$$f_{W(k)}(w(k)) = \frac{\Gamma[(k+1)\delta]}{[\Gamma(\delta)]^{k+1}} \prod_{j=1}^{k} w_j^{\delta-1},\ \ 0 \le w_1, w_2, ..., w_k \le 1. \tag{4}$$

If the parameter δ is set greater than unity this prevents the weights being equal to zero. This guards against encountering degeneracy, when we are conditioning on $K = k$, with a component vanishing because its weight becomes zero. We used $\delta = 1.5$, at least as a starting value, in our analysis of the example data described below.

Consider now the priors for the parameters, M and S, of the component distributions themselves. The majority of previous work, as in [Richardson and Green 1997], has concentrated on mixtures of normal distributions, with use of a normal prior for the means of the components and a gamma distribution for the inverse variances (or equivalently an inverse gamma distribution for the variances). This choice of distributions gives some advantages of conjugacy. Our main concern however is to have priors with ranges appropriate to the parameters concerned.

We use two forms of prior for the parameter M, depending on whether M is unrestricted in range or whether it has to be positive.

In the case of the normal and EV distributions where M is unrestricted in range we use a uniform prior for M. This prior is the least sensitive of the priors and has density

$$f_M(\mu) = (2\kappa s)^{-1} I_{[\bar{y}-\kappa s,\ \bar{y}+\kappa s]}(\mu), \tag{5}$$

where $I(\cdot)$ is the previously defined indicator function, $\bar{y}$ and s are the sample mean and standard deviation, and κ is an arbitrary constant made sufficiently large ($\kappa = 10$ in the examples) to ensure that the range over which the density is positive is greater than the sample range. Strictly speaking priors should be set completely independently of the data; however, given that we wish our procedure to be robust over a wide compass of data samples, we have given ourselves a little latitude by allowing the support of this prior to be data dependent.

For the lognormal, gamma, Weibull and IG distributions, we require $M \ge 0$. In these cases we use, for the prior, the beta distribution of the second kind with density

$$f_M(\mu) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \frac{r(r\mu)^{\frac{\nu_1}{2}-1}}{(1+r\mu)^{\frac{\nu_1+\nu_2}{2}}},\quad \mu > 0, \tag{6}$$

where $r = \nu_1/(\gamma\nu_2)$, with γ, ν_1, ν_2 the three parameters of the distribution to be chosen. We note that the density has mean $m = \gamma\nu_2/(\nu_2 - 2)$, and that we can write $\nu_2 = 4 + 2(\nu_1 + 2)/(c^2\nu_1 - 2)$, where c is the coefficient of variation of the distribution. In our choice of priors, setting appropriate values for these three parameters was the most critical to obtaining a satisfactory final mixture model. Our suggested choice is as follows.

(i) Choose ν_1 sufficiently large so that there is sufficient degree of contact of the density with the abscissa to ensure that the posterior probability becomes increasingly small, tending to zero as the parameter value tends to zero. This requires choosing $\nu_1 > 2$.

(ii) Set ν_2 using $\nu_2 = 4 + 2(\nu_1 + 2)/(c^2\nu_1 - 2)$ with c reasonably large, corresponding to significant prior uncertainty. We used $c^2 = 10$.

(iii) Set γ so that $\gamma\nu_2/(\nu_2 - 2) = \bar{y}$, where $\bar{y}$ is the sample mean of the sample $\mathbf{y}$. Thus the prior mean is set to a value that reflects what one might expect the mean such a prior would have for the sample $\mathbf{y}$.

Finally consider $f_S(\sigma)$, the prior distribution of S. As S is a standard deviation we require $S \geq 0$. We therefore use the same form for $f_S(\sigma)$ that we use for M when this had to be positive. Thus we take $f_S(\sigma)$ to be exactly the same form as (6), with three parameters γ, ν_1, ν_2 to be chosen. We set ν_1 and ν_2 to the same values as in (6). However in this case we set γ to satisfy $\gamma\nu_2/(\nu_2 - 2) = s$, where s is the sample standard deviation of the sample $\mathbf{y}$.

In summary, the complete prior is

$$\pi[\boldsymbol{\theta}(k), k] = \{\prod_{i=1}^{k}[f_M(\mu_i)f_S(\sigma_i)]\}f_{W(k)}[w(k)]p_K(k),$$

where $\boldsymbol{\theta}(k) = [\boldsymbol{\psi}(k), \mathbf{w}(k)] \in \boldsymbol{\Theta}(k)$, the latter being the support of the prior distribution in $[\boldsymbol{\psi}(k), \mathbf{w}(k)]$ space.

4. IMPORTANCE SAMPLING

4.1. Posterior & Importance Sampling Distributions in the Mixture Model

As with the prior distributions, the posterior distribution is most readily specified for different $K = k$, under the assumption that there is no degeneracy when we do this, so that there are precisely k and only k components in the mixture when $K = k$. This is equivalent to assuming that $\pi(\boldsymbol{\psi}(j), \mathbf{w}(j), k|\mathbf{y}) = 0$ if $j \neq k$. The only non-zero parts of the posterior density are therefore

$$\begin{aligned}\pi(\boldsymbol{\theta}(k)|\mathbf{y}) &= \frac{f[\mathbf{y}|\boldsymbol{\theta}(k), k]\pi[\boldsymbol{\theta}(k), k]}{\sum_{\kappa=1}^{k_{\max}} \int_{\boldsymbol{\Theta}(\kappa)} f[\mathbf{y}|(\kappa), \kappa]\pi[\boldsymbol{\theta}(\kappa), \kappa]d\boldsymbol{\theta}(\kappa)}, \\ \boldsymbol{\theta}(k) &= [\boldsymbol{\psi}(k), \mathbf{w}(k)] \in \boldsymbol{\Theta}(k),\ k = 1, 2, ..., k_{\max}.\end{aligned}$$

For IS we can therefore use an IS distribution of the form

$$c_k[\boldsymbol{\theta}(k)],\ \boldsymbol{\theta}(k) \in \boldsymbol{\Theta}(k),\ k = 1, 2, ..., k_{\max} \tag{7}$$

where $c_k[\boldsymbol{\theta}(k)]$ for each k is a continuous density scaled so that

$$\int_{\boldsymbol{\Theta}(k)} c_k[\boldsymbol{\theta}(k)]d\boldsymbol{\theta}(k) = k_{\max}^{-1}.$$

The IS procedure with sample size m is then as follows.

IS1. Draw a value of K, uniformly over $1, 2, ..., k_{\max}$, giving $K = k_i$, for $i = 1, 2, ..., m$

IS2. Draw a value $\boldsymbol{\theta}(k_i)$ from the distribution with density $c_{k_i}[\boldsymbol{\theta}(k_i)]$, for $i = 1, 2, ..., m$. This produces a sequence of independent and identically distributed random variables $(\boldsymbol{\theta}(k_i), k_i)\ i = 1, 2, ..., m$.

IS3. From $(\boldsymbol{\theta}(k_i), k_i)\ i = 1, 2, ..., m$, calculate the *IS ratios*

$$\rho[\boldsymbol{\theta}_i(k_i)] = p[\mathbf{y}|\boldsymbol{\theta}_i(k_i), k_i]/c_{k_i}[\boldsymbol{\theta}_i(k_i)] \text{ for } i = 1, 2, ..., m, \tag{8}$$

with $p[\mathbf{y}|\boldsymbol{\theta}_i(k_i), k_i] = f[\mathbf{y}|\boldsymbol{\theta}_i(k_i), k_i]\pi[\boldsymbol{\theta}_i(k_i), k_i]$.

IS4. Use the $\rho[\boldsymbol{\theta}_i(k_i)]$ with appropriately defined functions $h(\boldsymbol{\theta}(k), k)$, to estimate posterior quantities of interest.

An example of IS4 is the estimation of $\pi_K(k|\mathbf{y})$, the posterior probability that $K = k$. This is simply done by appealing to (2), and noting that both the prior for K and the importance sampling of K are uniform, so that we estimate $\pi_K(k|\mathbf{y})$ by

$$\hat{\pi}_K(k|\mathbf{y}) = \sum_{k_i=k} \rho[\boldsymbol{\theta}_i(k_i)] / \sum_{i=1}^{m} \rho[\boldsymbol{\theta}_i(k_i)],\ k = 1, 2, ..., k_{\max} \tag{9}$$

It remains to select $c_k[\boldsymbol{\theta}(k)]$. Here we follow [Geweke 1989] and take as $c_k[\boldsymbol{\theta}(k)]$ a multivariate Student t distribution with mean located at the mode of $p[\mathbf{y}|\boldsymbol{\theta}(k), k]$ and with variance equal to minus the inverse of the Hessian of $\ln(p[\mathbf{y}|\boldsymbol{\theta}(k), k]$ evaluated at the mode. The t distribution is preferred to a normal as the IS distribution, because its longer tail guards against the problem of the IS distribution tailing off too quickly compared with the posterior distribution being estimated.

Calculation of $c_k[\boldsymbol{\theta}(k)]$ therefore depends on estimating the maximum point $\boldsymbol{\theta}_{\max}(k)$ of $\ln(p[\mathbf{y}|\boldsymbol{\theta}(k), k]$ for each k. We discuss this in the next subsection.

4.2. Obtaining the Maximum of the Posterior

It is well known (see [Geweke 1989], for example) that the IS method is most reliable if the IS distribution is a good approximation of the posterior distribution. We obtain such an approximation by maximising the posterior distribution for each possible value of k. This gives us a set of values for the component means, shapes and weights for each k, which for convenience we shall call the *maximum posterior probability (MPP) estimates* and denote by $\hat{\boldsymbol{\psi}}(k), \hat{\mathbf{w}}(k)$. We then construct the IS distribution, conditional on k, based on these optimized posterior values. The IS distribution is specified completely by assuming that k is uniformly distributed over some suitable range for k.

Though the optimized posterior values are obtained within a Bayesian framework, they are actually rather useful in their own right as *point estimators* viewed from a frequentist standpoint. To estimate the best k we can use (see [Schwarz 1978]) the Bayesian Information Criterion (BIC)

$$B_k = \log\left(\sum_{i=1}^{n} f(y_i|\hat{\boldsymbol{\psi}}(k), \hat{\mathbf{w}}(k), k)\right) - \frac{3k}{2}\log n, \tag{10}$$

where the first term on the right is the log-likelihood evaluated at the optimized posterior values of the parameters of the k-component model and the $3k$ factor in the second term is the number of parameters in the k-component model, including the weights. A smaller value B_k indicates a better fit. This yields a simple point estimate of k and all the component weights, and the parameters associated with each fitted component. However this does not give our objective, the posterior distribution of k, which we shall obtain using IS.

We tested three methods for finding the maximum of the posterior distribution: conjugate gradient optimization, the EM algorithm and Nelder Mead [Nelder and Mead 1965] optimization.

A good review of conjugate gradient methods is given in Chapter Two of [Burley 1974]. We tried the BFGS (Broyden - Fletcher - Goldfarb - Shanno) method introduced in [Davidon 1959]. We used the algorithm to minimise both the negative of the posterior and negative of the log of the posterior. For the negative of the posterior, we found that the algorithm did not move far from its starting point, as the gradients calculated at the initial points were very small. For the negative log of the posterior, the algorithm frequently moved to areas of parameter space associated with a very low

posterior probability. The errors causing this originated in the routine updating H, the estimate of the covariance matrix and we suspect were due to the surface being a long way from being quadratic.

We also considered the EM algorithm, introduced in [Dempster et al. 1977]. A good introduction to the EM algorithm and its application to mixture models is given in [Bilmes 1998].

We found that the EM algorithm was more sensitive to the starting point and also that it was unstable for some initial solutions. Often this occurred when a large number of components were being fitted to a dataset for which only a small number of components might be required, and took the form of one of the σ_l tending to infinity for a component with a very small weighting. The sensitivity of the limiting solution to the initial solution and the convergence to local maxima or saddle points are drawbacks that have been discussed elsewhere in the literature, e.g. in [Diebolt and Ip 1996].

There is surprisingly little theory available for the Nelder Mead routine [Nelder and Mead 1965]. Despite this we found the method the most robust of the three methods tried and in addition it was certainly the simplest to implement. This was therefore the method used in our current implementation. However we stress that we do not make this a strong recommendation. The EM algorithm is much more efficient than the Nelder Mead algorithm, performing about 100 iterations per model compared with a few thousand for Nelder Mead. There is scope for further research in this area, possibly considering an adaptation of a more sophisticated version of the EM algorithm, such as that put forward in [Arcidiacono and Bailey Jones 2003], or the use of a stochastic EM algorithm, which has previously been applied to mixture models in [Diebolt and Robert 1994]. If a sufficiently good optimum could be obtained without a significant increase in the number of runs required, this method could out-perform the Nelder Mead.

We applied the Nelder Mead [Nelder and Mead 1965] routine in the following way.

The basic version of the Nelder Mead method is for unconstrained optimization. We dealt with the positivity constraints on $\boldsymbol{\psi}_i$ and w_i simply by setting a parameter to half its current value, whenever the basic Nelder Mead algorithm proposes a negative value for the next step. We ensured the sum of the weights remains equal to unity simply by not treating the weight of the last component as being a parameter of the Nelder Mead search, but directly setting its value at each step of the search so that the weights sum to one. If, at any step and with the last weight omitted, the sum of the remaining weights is greater than unity, then all these remaining weights are rescaled so that they sum to nearly unity and the last weight is given a near zero value. A warning flag is raised if the routine exits with a supposed optimum, but with a weight near zero.

We used the Nelder Mead routine to maximize $\ln(p[\mathbf{y}|\boldsymbol{\theta}(k), k]$ for each k, doing this sequentially for increasing $k = 1, 2, ..., k_{\max}$, with the optimum point for the k component fit modified to provide the starting point for $k + 1$. With this sequential approach, a starting value, as required by the algorithm, is only really needed for the case $k = 1$. With the parametrization used an obvious starting point for this case is $\mu_0 = \bar{y}$ and $\sigma_0 = s$, the respective sample mean and sample standard deviation of the sample $\mathbf{y}$. The starting parameters for the model with $k + 1$ components are then determined from the best estimates for the model with k components. The first k components of the $k + 1$ model are set to be identical to those of the k component model, but with reduced weights to allow some weight to be given to the $(k + 1)$th component. The $(k+1)$th component is then chosen based on the discrepancies between the sample and the fitted k component model.

Specifically, let $y_i, i = 1, \ldots, n$ be the observations, and let $F_k(y)$ be the cumulative density function (CDF) of the fitted k-component model. We define $D_i, i = 1, \ldots, n$ to be

the difference between the empirical distribution function (EDF) and the fitted model with k components, such that

$$D_i = \frac{i - 0.5}{n} - F_k(y_i).$$

Let

$$p_0 = \max\{D_j - D_i | 1 \le i < j \le n\},$$

and suppose that this maximum is obtained at $i = i_0, j = j_0$. Also let

$$p_1 = \max\{D_j - D_i | 1 \le i < j \le n \text{ and } (i, j < i_0 \text{ or } i, j > j_0)\},$$

where this secondary maximum occurs at $i = i_1, j = j_1$. The $(k + 1)$th component is then given the initial mean

$$\mu_{k+1} = (y_{i_0} + y_{j_0})/2$$

and variance

$$\sigma_{k+1} = (y_{j_0} - y_{i_0})/2$$

and the weight of the $(k + 1)$th component is set to be p_0, while the weights of the remaining k components are multiplied by a factor $(1 - p_0)$. It is readliy verifiied that this procedure will reduce this maximum p_0, though it has to be admitted that there is some possibility that other, smaller, differences could be increased. However in extensive experimentation, not reported here, we found the procedure very reliable in producing acceptable optimizations over all k. The advantage of parametrizing the base distribution using its mean and standard deviation is evident in making this process of introducing additional components a straightforward one.

A second attempt is made with parameters i_1, j_1 and p_1 and the model with the greatest log likelihood is chosen as the starting point for the Nelder Mead optimisation routine. By considering an alternative starting point for the optimisation we are allowing for some of the multimodality of the posterior distribution, which will give us some protection against missing the mode.

4.3. Estimation of the Covariance Matrix

With the mean of the IS distribution $c_k[\boldsymbol{\theta}(k)]$ located at the mode of $p[\mathbf{y}|\boldsymbol{\theta}(k), k]$ we continue following [Geweke 1989] and set its variance equal to minus the inverse of the Hessian of $L(\boldsymbol{\psi}(k), \mathbf{w}(k), k) = \ln(p[\mathbf{y}|\boldsymbol{\theta}(k), k]$ evaluated at the mode. This is non-trivial because the weights, w_i must sum to one. Throughout this subsection, we focus on the kth component, where k is given. To simplify the notation we therefore write, throughout this subsection, L for $L(\boldsymbol{\psi}(k), \mathbf{w}(k), k)$, $\boldsymbol{\psi}$ for $\boldsymbol{\psi}(k)$, and $\mathbf{w}$ for $\mathbf{w}(k)$. The dimension of $\boldsymbol{\psi}(k)$ is $l = 2k$. Suppose the maximum of L occurs at $(\hat{\boldsymbol{\psi}}, \hat{\mathbf{w}})$, where this optimum has been obtained subject to $\sum w_j = 1$. Let the negative *unconstrained* Hessian of second partial derivatives be

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\boldsymbol{\psi},\boldsymbol{\psi}} & \mathbf{H}_{\boldsymbol{\psi},\mathbf{w}} \\ \mathbf{H}^T_{\boldsymbol{\psi},\mathbf{w}} & \mathbf{H}_{\mathbf{w},\mathbf{w}} \end{pmatrix} \tag{11}$$

with, in particular,

$$\mathbf{H}_{\mathbf{w},\mathbf{w}}(\hat{\mathbf{w}}) = -\frac{\partial^2 L(\mathbf{w})}{\partial \mathbf{w}^2}\bigg|_{\mathbf{w}=\hat{\mathbf{w}}}. \tag{12}$$

These partial derivatives in $\mathbf{H}$ are unconstrained in that they are obtained ignoring the restriction that $\sum w_i = 1$. To include this restriction we write w_i as

$$w_i = \lambda_i + k^{-1}\left(1 - \sum_{j=1}^{k} \lambda_j\right),\ i = 1, ..., k. \tag{13}$$

This ensures that

$$\sum_{i=1}^{k} w_i = 1. \tag{14}$$

The Jacobian matrix of the transformation is

$$\mathbf{J} = \frac{\partial \mathbf{w}}{\partial \boldsymbol{\lambda}} = (\mathbf{I}_k - k^{-1}\mathbf{1}_k\mathbf{1}_k^T), \tag{15}$$

where $\mathbf{I}_k$ is the $k-$component identity matrix and $\mathbf{1}_k = (1, 1, ..., 1)^T$ is the k-component vector with unit entries. The log posterior density in terms of this parameterization, $L = L(\boldsymbol{\psi}(k), \boldsymbol{\lambda}(k), k)$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2..., \lambda_k)^T$, has Hessian

$$\mathbf{A}(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \mathbf{A}(\boldsymbol{\psi}, \mathbf{w}) = \begin{pmatrix} \mathbf{H}_{\boldsymbol{\psi},\boldsymbol{\psi}} & \mathbf{H}_{\boldsymbol{\psi},\mathbf{w}}\mathbf{J}^T \\ \mathbf{J}\mathbf{H}_{\boldsymbol{\psi},\mathbf{w}}^T & \mathbf{J}\mathbf{H}_{\mathbf{w},\mathbf{w}}\mathbf{J}^T \end{pmatrix}, \tag{16}$$

which we write as $\mathbf{A}$ from now on. This is the required Hessian because the inverse of $\mathbf{A}$ gives the covariance of $(\hat{\boldsymbol{\psi}}, \hat{\mathbf{w}})$ subject to $\sum_{i=1}^{k} \hat{w}_i = 1$.

The matrix $\mathbf{A}$ must clearly be singular, and indeed the submatrix $\mathbf{J}\mathbf{H}_{\mathbf{w},\mathbf{w}}\mathbf{J}^T$ is singular as $\det(\mathbf{J}) = 0$. Thus $\mathbf{A}$ does not have a full inverse, but it does have a generalised inverse, $\mathbf{G}$, which by definition will satisfy

$$\mathbf{AGA} = \mathbf{A}. \tag{17}$$

To find the generalised inverse, we consider $\mathbf{P}$, the orthogonal matrix formed from the eigenvectors of $\mathbf{A}$ as defined in the following Lemma.

Lemma

(i) The vector

$$\mathbf{p}_0 = \begin{pmatrix} \mathbf{0}_l \\ \mathbf{1}_k \end{pmatrix} \begin{matrix} \}l \\ \}k \end{matrix} \tag{18}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 0, where $\mathbf{0}_l$ is the l-dimensional column vector of zeros.

(ii) All other eigenvectors of $\mathbf{A}$, which we write as

$$\mathbf{p}_j = \begin{pmatrix} \mathbf{p}_j^{\boldsymbol{\psi}} \\ \mathbf{p}_j^{\mathbf{w}} \end{pmatrix} \begin{matrix} \}l \\ \}k \end{matrix},\ j = 1, 2, ..., \nu \tag{19}$$

where $\nu = l + k - 1$, satisfy

$$\mathbf{1}_k^T \mathbf{p}_j^{\mathbf{w}} = 0. \tag{20}$$

Proof The matrix $\mathbf{A}$ is singular and so, by definition, has at least one eigenvalue that is equal to zero. Therefore, in order to prove part (i) of the lemma, we need only show that

$$\mathbf{A}\mathbf{p_0} = \mathbf{0}_{(l+k)}. \tag{21}$$

Using the expansion given in (16), we can rewrite this condition as

$$\mathbf{A}\mathbf{p_0} = \begin{pmatrix} \mathbf{H}_{\psi,w}\mathbf{J}^T\mathbf{1}_k \\ \mathbf{J}\mathbf{H}_{w,w}\mathbf{J}^T\mathbf{1}_k \end{pmatrix}. \tag{22}$$

The expression for the Jacobian $\mathbf{J}$ is given in (15) and it is easy to show that $\mathbf{J}^T\mathbf{1}_k$ is equal to $\mathbf{0}_k$. Hence, (21) holds and part (i) of the lemma is proved.

To prove part (ii) we simply note that the matrix $\mathbf{A}$ is symmetric and therefore has distinct orthogonal eigenvectors $\mathbf{p}_j$, $j = 1, 2, \ldots, \nu$, for which $\mathbf{p}_j\mathbf{p}_0 = \mathbf{0}_{\nu+1}$. As the upper l components of $\mathbf{p}_0$ are zero, the orthogonality condition reduces to $\mathbf{p}_j\mathbf{1}_k = \mathbf{0}_k$, hence proving part (ii) of the lemma. □

Using the Lemma we write $\mathbf{P}$ as

$$\mathbf{P} = (\ \mathbf{P}_1\ \ \mathbf{p}_0\ ) \tag{23}$$

with $\mathbf{p}_0$ in the last column of $\mathbf{P}$ and

$$\mathbf{P}_1 = (\ \mathbf{p}_1\ \ \mathbf{p}_2\ \ \ldots\ \ \mathbf{p}_\nu\ ) \tag{24}$$

the matrix comprising the other eigenvectors.

Then

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D} = \begin{pmatrix} \mathbf{\Lambda} & 0 \\ 0 & 0 \end{pmatrix} \begin{matrix} \}\nu \\ \}1 \end{matrix}, \tag{25}$$

where $\mathbf{D}$ is the diagonal matrix of eigenvalues corresponding to the eigenvectors forming $\mathbf{P}$.

From now on we assume that $\mathbf{A}$ is positive semidefinite so that all its eigenvalues are non-negative. The main diagonal entries of $\mathbf{D}$ and $\mathbf{\Lambda}$ will therefore all be non-negative. In practice it will usually be the case, in maximizing the logposterior density subject to $\sum w_i = 1$, that *all* the main diagonal entries in $\mathbf{\Lambda}$ will be strictly positive, but our construction of the generalized inverse does not require this. However the last main diagonal entry of $\mathbf{D}$ is definitely zero by construction.

Let $\mathbf{S}$ be the $(l + k) \times (l + k)$ diagonal matrix

$$\mathbf{S} = \begin{pmatrix} \mathbf{R} & \mathbf{0}_\nu \\ \mathbf{0}_\nu^T & 0 \end{pmatrix} \begin{matrix} \}\nu \\ \}1 \end{matrix}$$

where

$$\mathbf{R} = \ \mathrm{diag}(l_{ii} |\ l_{ii} = 1/\sqrt{\lambda_{ii}} \text{ if } \lambda_{ii} > 0,\ l_{ii} = 0 \text{ if } \lambda_{ii} = 0)$$

and λ_{ii} is the ith main diagonal entry of $\mathbf{\Lambda}$. Define $\mathbf{L}$ as

$$\mathbf{L} = \mathbf{P}\mathbf{S}. \tag{26}$$

Explicitly we have

$$\mathbf{L} \ = \ (\ \mathbf{P}_1\ \ \mathbf{p}_0\ ) \begin{pmatrix} \mathbf{R} & \mathbf{0}_\nu \\ \mathbf{0}_\nu^T & 0 \end{pmatrix} \tag{27}$$

$$= \ (\ \mathbf{P}_1\mathbf{R}\ \ \mathbf{0}_{l+k}\ ). \tag{28}$$

Define $\mathbf{G}$ as

$$\mathbf{G} = \mathbf{P}\mathbf{S}\mathbf{S}\mathbf{P}^T = \mathbf{L}\mathbf{L}^T. \tag{29}$$

From the definition of $\mathbf{S}$ we have

$$\mathbf{D}\mathbf{S}\mathbf{S}\mathbf{D} = \mathbf{D} \tag{30}$$

and from (25),

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T. \tag{31}$$

Using these two expressions we have

$$\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T\mathbf{P}\mathbf{S}\mathbf{S}\mathbf{P}^T\mathbf{P}\mathbf{D}\mathbf{P}^T = \mathbf{P}\mathbf{D}\mathbf{S}\mathbf{S}\mathbf{D}\mathbf{P}^T = \mathbf{P}\mathbf{D}\mathbf{P}^T = \mathbf{A}. \tag{32}$$

So $\mathbf{G}$ satisfies (17) and is thus a generalised inverse of $\mathbf{A}$.

4.4. IS in the Finite Mixture Model

We can now describe explicitly our proposed IS method of generating the parameters $\boldsymbol{\theta}(k)$ of Step IS2 from a modified multivariate t-distribution. Specifically we generate this as

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\psi} \\ \mathbf{w} \end{pmatrix} \sim \mathrm{StudentT}\left(\begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\mathbf{w}} \end{pmatrix}, \mathbf{L}\mathbf{L}^T \right), \tag{33}$$

where $\mathrm{StudentT}(\tilde{\boldsymbol{\theta}}, \mathbf{V})$ is the multivariate t-distribution with mean $\tilde{\boldsymbol{\theta}}$ and variance $\mathbf{V}$. A variate from this distribution can be generated using

$$\boldsymbol{\theta} = \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\mathbf{w}} \end{pmatrix} + \mathbf{P}_1 \mathbf{R} \mathbf{z}_\nu, \tag{34}$$

where $\mathbf{z}_\nu$ is a vector of independent Student-t variates, each normalized to have mean zero and variance unity. These can have arbitrary degrees of freedom, d, and are derived by rescaling non-standardised t-variates. The variance-covariance of $\boldsymbol{\theta}$ generated in this way is then

$$\mathrm{Var}(\boldsymbol{\theta}) = \mathrm{E}(\mathbf{P}_1\, \mathbf{R}\mathbf{z}_\nu \mathbf{z}_\nu^T \mathbf{R}\mathbf{P}_1^T) = \mathbf{P}_1 \mathbf{R}\mathbf{R}\mathbf{P}_1^T. \tag{35}$$

But from (28) and (29) we have that $\mathbf{P}_1\mathbf{R}\mathbf{R}\mathbf{P}_1^T = \mathbf{P}\mathbf{S}\mathbf{S}\mathbf{P}^T$. Therefore

$$\mathrm{Var}(\boldsymbol{\theta}) = \mathbf{P}_1\mathbf{R}\mathbf{R}\mathbf{P}_1^T = \mathbf{P}\mathbf{S}\mathbf{S}\mathbf{P}^T = \mathbf{G}.$$

Moreover, using the result (20) that

$$(\mathbf{0}_l^T, \mathbf{1}_k^T)\mathbf{P}_1 = \mathbf{0}_\nu^T, \tag{36}$$

the sum of the component weights is given by

$$\begin{aligned} \sum_{i=1}^{k} w_i &= (\mathbf{0}_l^T, \mathbf{1}_k^T) \begin{pmatrix} \boldsymbol{\psi} \\ \mathbf{w} \end{pmatrix} && (37) \\ &= (\mathbf{0}_l^T, \mathbf{1}_k^T) \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\mathbf{w}} \end{pmatrix} + (\mathbf{0}_l^T, \mathbf{1}_k^T)\mathbf{P}_1\mathbf{R}\mathbf{z}_\nu && (38) \\ &= (\mathbf{0}_l^T, \mathbf{1}_k^T) \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\mathbf{w}} \end{pmatrix} + \mathbf{0}_\nu^T\mathbf{R}\mathbf{z}_\nu && (39) \\ &= \sum_{i=1}^{k} \hat{w}_i = 1. && (40) \end{aligned}$$

Thus under this sampling we are restricted to the simplex $\sum_{i=1}^{k} w_i = 1$.

The vector of weights, $\mathbf{W} = (W_1, W_2, ..., W_k)$ clearly has a singular distribution. Let $\boldsymbol{\Omega}$ be the $(k-1)$ dimensional vector formed from the first $(k-1)$ components of $\mathbf{W}$ and write

$$\boldsymbol{\Phi} = (\boldsymbol{\Psi}, \boldsymbol{\Omega}) \tag{41}$$

for the vector of component distribution parameters and this *reduced set of weights*, with $\boldsymbol{\phi} = (\boldsymbol{\psi}, \boldsymbol{\omega})$ for a particular instance. In the importance sampling process IS1-IS4 we can think of the PDF $c_k[\boldsymbol{\theta}(k), k]$ of (7) as being completely determined just by $\boldsymbol{\phi}$. Therefore, we can replace $c_k[\boldsymbol{\theta}(k), k]$ appearing in the IS ratio (8) by $f_{\boldsymbol{\Phi}}(\boldsymbol{\psi}, \boldsymbol{\omega})$, the IS

PDF of ϕ, which is nondegenerate and given by

$$f_{\boldsymbol{\Phi}}(\boldsymbol{\psi},\boldsymbol{\omega}) = f_{\mathbf{z}_\nu}(\mathbf{z}_{\boldsymbol{\nu}}) \left| \frac{\partial \mathbf{z}_{\boldsymbol{\nu}}}{\partial(\boldsymbol{\psi},\boldsymbol{\omega})} \right| \tag{42}$$

$$= f_{\mathbf{z}_\nu}(\mathbf{z}_{\boldsymbol{\nu}}) \left| \frac{\partial(\boldsymbol{\psi},\boldsymbol{\omega})}{\partial(\mathbf{z}_{\boldsymbol{\nu}})} \right|^{-1}, \tag{43}$$

where $f_{\mathbf{z}_\nu}(\mathbf{z}_{\boldsymbol{\nu}})$ is the joint PDF of sampling the ν standard t-variates. From the form of (34) it is clear that

$$\partial(\boldsymbol{\psi},\boldsymbol{\omega})/\partial(\mathbf{z}_{\boldsymbol{\nu}}) = \mathbf{M}$$

where $\mathbf{M}$ is the matrix $\mathbf{P}_1\mathbf{R}$ but with the last row omitted. Thus $|\partial(\boldsymbol{\psi},\boldsymbol{\omega})/\partial(\mathbf{z}_{\boldsymbol{\nu}})| = \det(\mathbf{M})$, so that

$$f_{\boldsymbol{\Phi}}(\boldsymbol{\psi},\boldsymbol{\omega}) = [\det(\mathbf{M})]^{-1} f_{\mathbf{z}_\nu}(\mathbf{z}_{\boldsymbol{\nu}}). \tag{44}$$

Use of (34) to generate IS variates does not guarantee that parameters which should be positive necessarily are positive, nor that all weights are necessarily less than unity. This is easily handled by rejecting any $\boldsymbol{\theta}$ sample where *any* such constraint which should be satisfied is not. This restricts the support of the IS distribution to precisely the region where *all* parameter constraints are satisfied. The IS sampling is therefore an acceptance/rejection procedure. Thus given $K = k$, the IS distribution actually sampled is modified from (44) to

$$c_k[\boldsymbol{\psi}(k),\boldsymbol{\omega}(k)] = [\det(\mathbf{M}(k))]^{-1} f_{\mathbf{z}_\nu}(\mathbf{z}_{\boldsymbol{\nu}})/R(k) \tag{45}$$

where we have included dependency on k explicitly, and $R(k)$ is the probability that a value sampled from (44) is accepted (because it falls in the support of the k component form of the mixture model being fitted). The value of $R(k)$ is easily estimated in from the IS sampling by

$$\hat{R}(k) = (\#\text{ of replications sampled from (44) for the given } k \text{ and accepted})\ /m_k$$

where

$$m_k = (\#\text{ of replications sampled from (44) for the given } k\ ).$$

Acceptance/rejection adds to the computational effort, but would only be problematic if $R(k)$ were ever to be small, which we have not encountered. Our proposed IS procedure appears acceptably fast in practice. We have tried more elaborate IS distributions which directly satisfy the required parameter constraints, but such distributions made the calculations significantly more complicated and less transparent and actually slowed the IS procedure.

We summarize the IS as it applies to estimation of the posterior distribution of K.

In the previous section we assumed, for ease of exposition, that the number of components k in each IS replication was sampled independently. However we can remove the inherent variability in this sampling of k by using *stratified* sampling. We thus replace IS1 by:

IS1'. Sample k *cyclically* with $k = 1, 2, .., k_{\max}, 1, 2, ..., k_{\max}, 1, 2, ...$ and so on, so that if m replications are drawn, where for simplicity we assume that m is divisible by $k_{\max}$, we sample the *same* number of replications for each possible k, i.e. if m_k is the number of replications where $K = k$ then

$$m_k = m/k_{\max},\ k = 1, 2, ..., k_{\max}$$

so that all the m_k are equal. All the IS formulas derived in the previous section are unchanged by this.

IS2'. For each k_i obtained in IS1', sample the mixture model parameters $\boldsymbol{\psi}(k_i), \boldsymbol{\omega}(k_i)$ using (34) but applying acceptance/rejection so that each accepted $\boldsymbol{\psi}(k_i), \boldsymbol{\omega}(k_i)$ satisfies all parameter constraints for the given k_i component mixture model.

IS3'. For each replication calculate the IS ratio, ρ_i, as given in 8 with the divisor given by (45).

IS4'. Estimate $\hat{\pi}_K(k|\mathbf{y})$, $k = 1, 2, .., k_{\max}$, the posterior distribution of the number of components from (9)

Other quantities of interest such as the PDF of the parameters $\boldsymbol{\psi}(k), \mathbf{w}(k)$ conditional on k, can then be estimated by appropriate weighted frequency histograms using the IS ratios as the weights.

4.5. Convergence Statistics

An advantage of the use of IS over MCMC is that convergence statistics are more readily obtained because the observations of the IS sample are essentially independent. The analysis is not completely straightforward because the estimator of (2) involves a ratio of two means; however [Geweke 1989] shows that this can be characterized by a central limit theorem. Geweke's Theorem 2 shows that if

$$\hat{\sigma}_m^2 = \sum_{i=1}^{m} [h(\theta_i) - \bar{h}_m]\rho^2(\theta_i) / \sum_{i=1}^{m} \rho^2(\theta_i) \tag{46}$$

then

$$m^{1/2}\{\bar{h}_m - E[h(\theta)]\} \Longrightarrow N(0, \sigma^2)$$

and

$$m\hat{\sigma}_m^2 \to \sigma^2$$

for a suitably defined variance σ^2. (The formula for σ^2 appearing in the statement of Geweke's Theorem 2 is different from the version appearing in the proof. The one in the proof is correct. Further details are in A.3). Put simply, for large m, $\hat{\sigma}_m^2$ estimates the variance of $\bar{h}_m$.

Applying this to the estimate, $\hat{\pi}_K(k|\mathbf{y})$, of the posterior probability that the number of components in the model is equal to k as given in (9), we only need take $h(\theta_i)$ to be

$$h_k(\theta_i) = \begin{cases} 1 \text{ if the number of components is } k, \text{ i.e. if } k_i = k \\ 0 \text{ otherwise} \end{cases} \tag{47}$$

and

$$\bar{h}_m(k) = \pi_K(k|\mathbf{y}),$$

in (46) so that $\hat{\sigma}_m^2$ specialises to

$$\hat{\sigma}_m^2(k) = \frac{\left\{ \sum_{k_i=k} [(1 - \bar{h}_m(k))\omega(\boldsymbol{\theta}_i(k_i))]^2 + \sum_{k_i \neq k} [\bar{h}_m(k)\omega(\boldsymbol{\theta}_i(k_i))]^2 \right\}}{\left\{ \sum_{i=1}^{m} \omega[\boldsymbol{\theta}_i(k_i)] \right\}^2}, \tag{48}$$

$$k = 1, \ldots, k_{max}.$$

This provides an indication of the accuracy of the estimates $\pi_K(k|\mathbf{y})$. Note that the $\hat{\sigma}_m^2(k)$ are not mutually independent.

5. EXAMPLES

We consider three examples. Our main comparison is between our proposed optimization/IS approach. which in this section we refer to simply as the 'IS method', and the RJMCMC method, which we refer to simply as the 'MC method'.

Note that the RJMCMC method is availabe in a Fortran code implementation called Nmix, downloadable from http://www.stats.bris.ac.uk/~peter/Nmix/. We have included Nmix as an option in the FineMix implementation, so both the 'IS method' and the 'MC method' can be run using just FineMix, and this is what we did in the examples.

The first example is an artificial data set comprising a set of 100 observations generated from a normal mixture distribution with three quite distinct components. The three true (μ_i, σ_i, w_i) vectors used to generate the components and form the sample were $(12.0, 0.125, 0.25)$, $(12.5, 0.02, 0.2)$, and $(13.0, 0.3, 0.55)$. We shall refer to this sample as the '3ClearNorms' example. The second component has a small variance and is specifically included to test if the methods can accurately identify such a component.

The other two samples are fairly large compared with most examples considered in previously published work. One comprises the lot-sizes of surface mounted capacitors, with sample size 2083, to which Wagner and Wilson [Wagner and Wilson 1996b] fitted a multimodal distribution using PRIME, their proposed Bézier fitting method. We refer to this sample as the 'LotSize' data set. This example allows our IS method to be directly compared with this very different approach.

The third data set is a complex financial one comprising the loss given defaults (LGD) of 7051 clients. This sample, like the first also has a small but distinctive cluster, due possibly to a component with a small variance.

We consider all three examples using the FineMix program in order to illustrate some of the features incorporated in it. FineMix should be able to detect and model components with small variances and/or small weights. Two of the parameters in FineMix specifying the priors are adjustable to ensure that any particular sample can be handled with a balance of sensitivity and stability.

One is the shape parameter, δ, of the Dirichlet distribution in the prior (4) for the component weights. We need $\delta \geq 1$ to avoid any of the weights tending to zero in the Nelder-Mead optimization, but if δ is set too large this can over-restrict the search in its choice of small weight values. This can have the computational side effect of components 'merging' so that some of the components, though calculated separately, end up having the same parameter values. Recalling that the Nelder-Mead routine is carried out on individual k, with the k being sequentially increased, this would mean that a final fit for a given k would be indistinguishable from a fit with a lower k, because two or more of the components are effectively the same. Though undesirable, this is actually not all that serious from a practical point of view because it tends only to occur when k is larger than needed. The posterior probabilities for such k values will then be small, so that inaccuracies in their estimated value have negligible practical consequence. Nevertheless it would be preferable not to have this problem occur, and in the FineMix implementation a check is made to ensure that for any fitted k the estimated component values are all distinct.

The other parameter we have allowed to be adjustable is the shape parameter ν_1 used in the priors for component SDs and for component means that have to be positive. This parameter behaves like δ in that we would prefer $\nu_1 \simeq 2$ to express prior uncertainty, but need $\nu_1 > 2$ to avoid component parameters that should be positive tending to zero in the Nelder-Mead routine, rendering degenerate the component in question. The FineMix implementation checks component variances in particular, issuing a warning if such a value is near zero.

Table II. Estimated posterior distribution of k for the '3ClearNorms' sample using the IS and MC methods. For the IS case the SD of the estimate of $\hat{p}_k$ is included

Method	k	1	2	3	4	5	6	7	8	9	10
IS	$\hat{p}_k$			0.996	0.004						
	σ_{IS}			0.0033	0.0033						
MC	$\hat{p}_k$	0.001	0.023	0.329	0.281	0.177	0.098	0.050	0.022	0.010	0.005

We have found that δ and ν_1 can almost invariably be adjusted together, with (δ, ν_1) in the range $(1.1, 2.2)$ to $(2, 4)$ providing enough flexibility.

The FineMix implementation includes one other diagnostic check. At the end of the Nelder-Mead routine, the eigenvalues of the negative of the Hessian of the posterior distribution evaluated at the optimal point, are examined and any found to be negative are reported. If all of the eigenvalues are positive, this is an indication that at least a local optimum has been obtained. It is possible, especially when k is much larger than needed, for the posterior to become rather flat and the Nelder-Mead routine can terminate before all the eigenvalues become positive. The IS sampling can still return a useful estimate of the posterior distribution of k in this case as a negative eigenvalue is usually associated with a k that is an extreme value for which p_k is very small and the IS sampling will reflect this, so that the overall distribution of k is still correctly estimated. However if there is any concern then it is usually easiest to refit using a smaller $k_{\max}$, so that the problem is not encountered for the range of k considered.

FineMix can be downloaded from http://www.soton.ac.uk/ ccurrie/, including fuller details of how to run the user interface and of the output that it produces. In addition to the examples described in the Examples Section of this paper, we have considered a large number of other mainly rather smaller data sets coming from diverse application areas. The FineMix interface includes a spreadsheet containing over a dozen of these, including the three data sets considered in [Richardson and Green 1997].

5.1. Example 1: Mixture of Three Normals

In this example we set $k_{\max} = 8$ and $\delta = 1.5$, $\nu_1 = 3$ for the main smoothing parameters. We applied the IS method with a sampling size of $m = 50,000$ to estimate the posterior distribution of k. This gave the estimates $\hat{p}_3 = 0.996$ and $\hat{p}_4 = 0.004$ as shown in Table 2, with other posterior probabilities negligible in comparison. The probabilities must sum to unity and as in effect only $\hat{p}_3$ and $\hat{p}_4$ are positive this means that their SDs (calculated from (48) must be effectively the same.

We can estimate what is called the *predictive density conditional on* k in [Richardson and Green 1997], i.e. the posterior distribution of the full mixture model conditional on k, in two ways: (i) The *MPP parameters method* where for each k the posterior distribution for that particular k is calculated with the parameters set equal to their MPP estimates for that k, calculated as described in subsection 4.2; (ii) The *averaged IS parameters method* which calculates the full mixture density for the given k with parameter values set equal to the average of the values obtained just at IS observations with $K = k$. The predictive density, obtained both ways, for the normal mixture model with $k = 3$, when fitted to the 3ClearNorms dataset using IS, is shown Figure 1.

It will be seen that both ways of calculating the predictive density clearly identify the component with the small variance, with the MPP method perhaps appearing to be the more accurate visually in this case.

We also fitted the normal mixtures model using the MC method. We used a burn-in of $100,000$ steps and $50,000$ recorded observations. As shown in Table 2 the MC method does not give a very clear result for the estimated posterior distribution of k, assigning higher posterior probabilities to a larger number of components than the data would suggest.

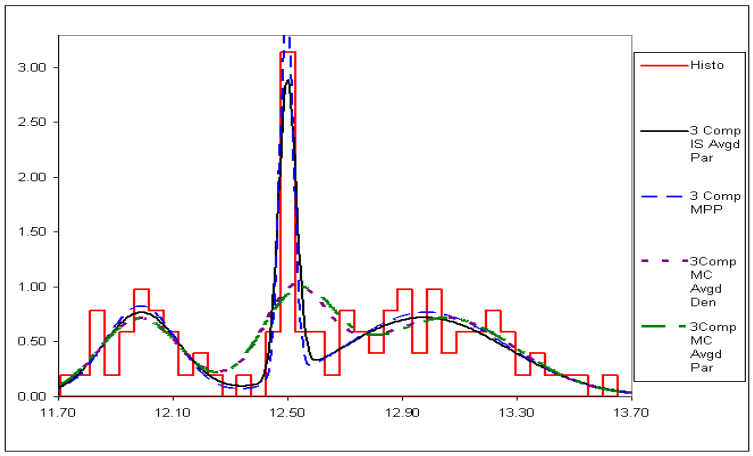

Fig. 1. 3ClearNorms example: frequency histogram; IS 3 component normal fits using averaged parameter values and MPP estimators; 3 component MC normal fits using averaged density values and averaged parameter values

Richardson and Green give two methods when using the MC method for estimating the predictive densities of the full mixture distribution for each given k: (i) The *averaged density method* where the full mixture density is calculated at all those steps of the Markov chain in which the given k value is obtained, and the predictive density conditional on k is the average of these values; (ii) The *averaged MC parameters method* which is the analogue of the averaged IS parameters method, calculating the full mixture density for the given k with parameters set equal to the average of the values obtained at just those observations of the Markov chain where $K = k$. The predictive densities for $k = 3$ obtained using both these MC based methods is also depicted in Figure 1. Neither predictive density is very satisfactory, with an over-smooth density at the histogram peak evident at the location of the second component.

As pointed out in [Richardson and Green 1997] the averaged MC parameters method can result in a predictive density that is too smooth as has occurred in this example. A possible reason is that an averaged parameter value can be an unrepresentative estimate of the values sampled. Figure 4 in Appendix A.4 shows the histogram plots of the posterior parameter and weight values obtained in the MC method for the case $k = 4$. Though this model has only one component more than the best value $k = 3$, the shape of the distribution of a parameter or weight can be rather different - bimodal in many cases, so that the average value is atypical.

For comparison Figure 5 in Appendix A.4 shows the histograms of the posterior parameter and weight values of the 4-component model as calculated using IS sampling showing these to be unimodal, so that the average parameter value is more meaningful in this case. Also included in the Appendix is Figure 6, the analogue of Figure 1 but showing the 4-component rather than the 3-component predictive densities. It will be seen that inclusion of the extra component leaves the difference in quality of the fits essentially unchanged, with the densities obtained using the IS method identifying the second component much more clearly than the densities obtained using the MC method.

As mentioned in the Introduction, the Bayesian Influence Criterion (BIC) of (10) can be used to compare different point estimates. Table 3 shows the values of B_k, $k = 1, 2, ..., 10$ for the present example, when fitting the mixture model using the normal

Table III. BIC Values of k-component mixture models for the '3ClearNorms' sample

k	1	2	3	4	5	6	7	8	9	10
normal	662.5	549.2	**525.2**	536.8	543.7	551.7	554.7	565.3	573.4	579.8

as base distribution. It will be seen that the minimum B_k was obtained at $k = 3$, corroborating the IS analysis.

5.2. Example 2: Capacitor Lot Sizes

Wagner and Wilson [Wagner and Wilson 1996b] consider a sample comprising the lot sizes in thousands for 2083 lots of surface mounted capacitors being stored in a facility while waiting for their insulation resistance to be tested. In their paper Wagner and Wilson [Wagner and Wilson 1996b] described the sample as being bimodal. The frequency histogram of the data set depicted in Figures 6 and 8 of their paper appears bimodal, but its full nature is somewhat masked as the cell size appears large. We include in Appendix A.5 in the Online Supplement a figure giving the fit we obtained using PRIME, which is bimodal and similar to that shown in Figure 6 of Wagner and Wilson. However their Figure 10, which also depicts the frequency histogram but using a smaller cell size, seems to show a more multimodal behaviour in the frequency histogram. Given the fairly large sample size it is therefore of interest to examine if some of the more detailed multimodality is part of a systematic pattern rather than being the result of random variation.

We fitted the normal mixture model using the IS method with $\delta = 1.05$, $\nu_1 = 2.1$. We were able to set these close to their lower limits, because the large sample size leads to a more stable Nelder-Mead search. For smaller sample sizes setting δ and ν_1 can lead to unstable fits with some component σ values tending to zero. The number of IS replications was $50,000$. The estimated posterior distribution of k is shown in Table 4 with probabilities that are negligibly small not shown. There was some spread of positive $\hat{p}_k$ values with the peak at $k = 6$. The PDF of this fitted 6-component model is shown in Figure 2.

We also fitted the normal mixture model using the MC method, with $50,000$ recorded MC steps. The resulting estimate of the posterior distribution of k is also shown in Table 4.

The predictive density for the 5-component mixtures model obtained by the averaged density method from the MC run is also shown in Figure 2.

We also fitted the Weibull mixture model using the IS approach, again with $\delta = 1.05$, $\nu_1 = 2.1$ and $50,000$ IS replications and this gave a definitive result with $\hat{p}_4 = 0.9995$, and other probabilities negligible. The reason for this is probably because the Weibull has quite a flexible shape range including both negative and positive skewness. The 4-component Weibull fit is also shown in Figure 2. It will be seen that at least 4 components are needed, with the IS method suggesting that some of the smaller fluctuations might be due to extra components.

Table IV. Estimated posterior distribution of k for the 'LotSize' sample. The values of $\hat{p}_k$ are given for the normal and Weibull cases fitted by the IS method. Values of $\hat{p}_k$ are also shown for the normal case fitted by the MC method.

Base Model	Method	k	3	4	5	6	7	8	9	10
normal	IS	$\hat{p}_k$			0.075	0.466	0.367	0.091		
Weibull	IS	$\hat{p}_k$		1.000						
normal	MC	$\hat{p}_k$		0.396	0.424	0.142	0.031	0.006	0.001	

The BIC values for the k-component normal and Weibull fits using IS are shown in Table 5 with the lowest values highlighted. In the normal case the BIC values for k

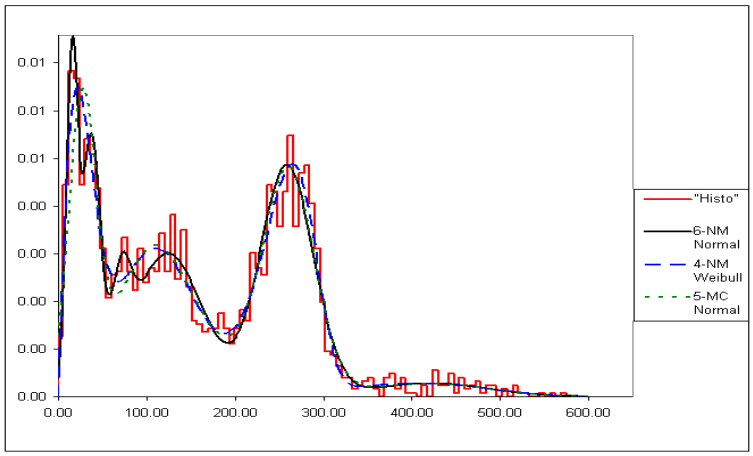

Fig. 2. LotSize example: frequency histogram; IS 6 component normal fit; 5 component MC normal fit using averaged density values; IS 4 component Weibull fit

Table V. BIC Values $\times 10^{-1}$ of k-component mixture models for the 'LotSize' data set

k	1	2	3	4	5	6	7	8	9	10	11	12
normal	2561	2500	2481	2451	**2448**	**2448**	2449	2451	2453	2454	2455	2457
Weibull	2512	2449	2445	**2442**	2444	2445	2447	2449	2450	2452	2454	2456

between 4 to 7 are not very different. Though only $k = 6$ is depicted, the fits for $k = 4$, 5 and 7 are quite similar.

The data set illustrates one of the difficulties of using a package like PRIME. Little knowledge is needed to use PRIME especially given the very user-friendly interface that allows easy movement of the control points around the screen showing how this impacts on the fit. However as the actual parameter values have little meaning outside of how they influence the shape of the PDF, it is less easy to decide at the outset how much of the detail of a data set should be taken into account in the fitting.

In using PRIME we found that choice of initial conditions strongly influenced the fitting process. This is backed up by Wagner and Wilson's own comment [Wagner and Wilson 1996a] "Some manual intervention beyond routine application of the likelihood ratio is often required to obtain adequate fits to multimodal data sets". The last example which we consider in the next subsection, is another instance that is hard to model due to a spike in the data.

5.3. Example 3: Credit Risk

In this example the data are the loss given default (LGD) for clients at a bank. Approximately 30% of debtors paid in full and so had an LGD of zero. We have removed these values from the data set and just consider the non-zero losses, leaving us with 7051 data points. An LGD of 1 corresponds to the debtor having paid off their loan in full, but if fees and legal costs have been incurred the LGD can be greater than 1, which is the case for approximately 15% of the non-zero losses.

The data histogram includes a small, but statistically significant, spike at just less than 0.15, representing the behaviour of a certain kind of client. As the sample size is quite large, we are able to use a relatively high value for the smoothing parameters with $\delta = 2$ and $\nu_1 = 4$, to avoid fitting to spurious clusters. We again set the number of IS relpications at $50,000$. As can be seen from Table 6 the minimum value of the BIC, taken over the seven base distributions considered, corresponds to a mixture of EV

Table VI. Credit risk data in Example 3, showing for each base distribution, the number of components giving the lowest BIC value, and BIC value obtained.

Component Distribution	Number of Components	BIC
Normal	7	1296.23
Lognormal	7	1297.24
EV	5	1249.44
NEV	5	1514.97
Weibull	5	1331.88
Gamma	7	1291.81
IG	7	1318.83

distributions, with BIC value minimized at $k = 5$. The IS method yielded the estimate of the posterior distribution of k as given in Table 7 with the maximum of $\hat{p}_5 = 0.682$. We also used the IS method to estimate the posterior distribution of k for the mixture model with Weibull base distribution; this being when the minimized BIC is one of the largest, i.e. worst cases. The estimates are also given in Table 7. In this case the result is very clear cut with the largest probability of $\hat{p}_5 = 0.998$, still at $k = 5$.

The plots of the 5-component fitted distributions for both the EV and Weibull cases are shown in Figure 3. Both provide a similar fit, with the EV fit perhaps being slightly better. Both fit a component to the observations clustered just below 0.15 with estimated weight $w = 0.0244$; in the EV case this component has $\mu = 0.147$ and $\sigma = 0.0047$ and in the Weibull case $\mu = 0.146$ and $\sigma = 0.0051$.

We also used the MC method to analyse this data set. The estimated the posterior distribution, which has a wider spread, is again shown in Table 7. Despite estimating a higher number of components than that obtained using the IS method, the MC method did not accurately identify the cluster at 0.15. Figure 3 includes the 6-component fit using the averaged density method, and it will be seen that the cluster at 0.15 is not well represented. The MC method does not seem to identify such clusters all that well. Though not discussed here, the FineMix program also includes a real data sample from car manufacturing giving the times of a certain activity cycle, which includes a remarkably tight cluster. The MC method also fails to properly identify the subsample involved in this data set. In contrast the IS method fitted a clear component to this subsample.

Summarizing, using the IS method to fit a mixture model in our present sample, even using base distributions with rather different properties, has still allowed a small but important component to be identified quite clearly.

Table VII. Estimated posterior distribution of k for the 'CreditRisk' sample. The values of $\hat{p}_k$ are given for the EV and Weibull cases fitted by IS and for the normal case fitted by the MC method

Base Model	Method	k	3	4	5	6	7	8	9	10
EV	IS	$\hat{p}_k$		0.015	0.682	0.303				
Weibull	IS	$\hat{p}_k$			0.998	0.002				
normal	MC	$\hat{p}_k$			0.351	0.363	0.194	0.075	0.013	0.004

6. FINAL SUMMARY

Finite mixture models are ideally suited to describing multimodal data and are particularly appropriate in simulation input modelling because of the ease of incorporating them into any simulation package. This paper describes a Bayesian method that has an established theoretical basis, which can be used to fit finite mixture models to multimodal data including the fitting of spikes and which can handle quite large sample sizes. A program implementing this method, FineMix, can be downloaded from the

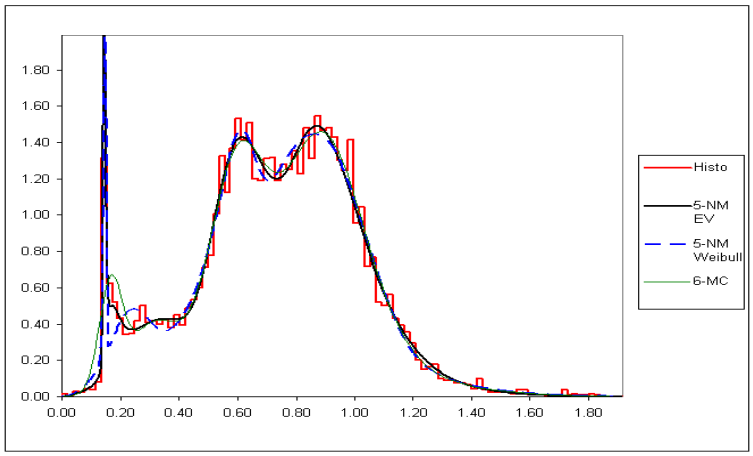

Fig. 3. Credit Risk example: frequency histogram; IS 5 component EV and Weibull fits and MC 6 component fit

authors' website, with a number of different options for the base component distribution. The fitting process offers a clear alternative to RJMCMC with the particular advantage of allowing easy implementation of other base distributions. An alternative, PRIME, proposed by Wagner and Wilson [Wagner and Wilson 1996b] though easy to use produces results that are less easy to interpret and to implement in simulation studies.

In this paper we have focussed discussion on the simple situation where we have already chosen our base distribution and are concentrating on identifying an appropriate number of components to use in the mixture. We have not discussed how to choose the base distribution. A simple approach would be a more formal extension of what we considered in the third example, where we fitted mixtures for a number of different base distributions and chose from these fits. Such a comprehensive approach would seem rather over elaborate in most applications and would have taken the discussion further than we wish in the present paper. In many applications the context may suggest an appropriate base. For example if the data is obviously positively skewed, then the lognormal, gamma or IG may be suitable, with the EV appropriate if there is no need to bound the support of the distribution from below. The Weibull is useful as it can be negatively as well as positively skewed.

REFERENCES

ARCIDIACONO, P. AND BAILEY JONES, J. 2003. Finite mixture distributions, sequential likelihood and the EM algorithm. *Econometrica 71*, 933–946.

BARRON, A., SCHERVISH, M. J., AND WASSERMAN, L. 1999. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics 27*, 536–561.

BERKHOF, J., VAN MECHELEN, I., AND GELMAN, A. 2003. A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica 13*, 423–442.

BILMES, J. 1998. A gentle tutorial of the EM algorithm and its applications to parameter estimation for gaussian mixture and hidden Markov models. TR-97-021, U.C. Berkeley.

BURLEY, D. 1974. *Studies in Optimization*. International Textbook Co. Ltd.

CHENG, R. 1998. Bayesian model selection when the number of components is unknown. In *Proceedings of the Winter Simulation Conference*, D. Medeiros, E. Watson, J. Carson, and M. Manivannan, Eds. 653–659.

CHENG, R. C. H. AND LIU, W. B. 2001. The consistency of estimators in finite mixture models. *Scandinavian Journal of Statistics 28*, 603–616.

DAVIDON, W. 1959. Variable metric method for minimization. Argonne Nat. Lab. report ANL-5990 Rev.

DELER, B. AND NELSON, B. L. 2001. Modeling and generating multivariate time series with arbitrary marginals and autocorrelation structures. In *Proceedings of the Winter Simulation Conference*, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, Eds. 275–282.

DEMPSTER, A., LAIRD, N., AND RUBIN, D. 1977. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B 39*, 1–38.

DIEBOLT, J. AND IP, E. 1996. Stochastic EM: method and application. In *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. Spiegelhalter, Eds. Chapman and Hall, Chapter 15.

DIEBOLT, J. AND ROBERT, C. 1994. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B 56*, 363–375.

ESCOBAR, M. AND WEST, M. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

FENG, Z. AND MCCULLOCH, C. 1996. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B 58*, 609–617.

GEWEKE, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

GHOSH, S. AND HENDERSON, S. G. 2001. Chessboard distributions. In *Proceedings of the Winter Simulation Conference*, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, Eds. 385–393.

HAMMERSLEY, J. AND HANDSCOMB, D. 1964. *Monte Carlo Methods*. Methuen.

KUHL, M., IVY, J., LADA, E., STEIGER, N., WAGNER, M., AND WILSON, J. 2010. Univariate input models for stochastic simulation. *Journal of Simulation 4*, 81–97.

LAW, A. M. 2007. *Simulation Modeling and Analysis (Fourth Edition)*. McGraw-Hill.

MCLACHLAN, G. AND PEEL, D. 2000. *Finite Mixture Models*. John Wiley and Sons.

MCLACHLAN, G. J. AND BASFORD, K. E. 1988. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.

NELDER, J. AND MEAD, R. 1965. A simplex method for function minimization. *Computer Journal 7*, 308–313.

NELSON, B. L. AND YAMNITSKY, M. 1998. Input modeling tools for complex problems. In *Proceedings of the Winter Simulation Conference*, D. Medeiros, E. Watson, J. Carson, and M. Manivannan, Eds. 105–112.

PHILLIPS, D. AND SMITH, A. 1996. Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. Spiegelhalter, Eds. Chapman and Hall, Chapter 13.

RAFTERY, A. 1996. Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. Spiegelhalter, Eds. Chapman and Hall, Chapter 10.

RICHARDSON, S. AND GREEN, P. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B 59*, 731–792.

ROBERT, C. AND CASELLA, G. 1999. *Monte Carlo Statistical Models*. Springer-Verlag.

ROEDER, K. AND WASSERMAN, L. 1997. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92*, 894–902.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

STEPHENS, M. 2000. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics 28*, 40–74.

TITTERINGTON, D. M., SMITH, A. F. M., AND E., M. U. 1985. *Statistical analysis of finite mixture distributions*. Wiley.

WAGNER, M. A. F. AND WILSON, J. R. 1996a. Recent developments in input modeling with Bézier distributions. In *Proceedings of the Winter Simulation Conference*, J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, Eds. 1448–1456.

WAGNER, M. A. F. AND WILSON, J. R. 1996b. Using univariate Bézier distributions to model simulation input processes. *IIE Transactions 28*, 699–711.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

Online Appendix to: Input Modelling for Multimodal Data: Using Bayesian Model Selection to Fit Finite Mixture Models

Russell C.H. Cheng, University of Southampton
Christine S.M. Currie, University of Southampton

A. APPENDIX

A.1. Choice of priors

Consider first the prior distribution for K, the number of components. A modified Poisson distribution is proposed in [Phillips and Smith 1996], which places greater probability mass on values close to the input parameter of the distribution, a value set by the user. This introduces some influence on the values of K obtained. In [Escobar and West 1995] a Dirichlet process is used to provide the prior distribution for K. With this choice of prior, the expected number of components in the mixture model for a sample of size n is proportional to $\ln(1 + n/\alpha)$; therefore, as the sample size increases, the expected number of components also increases. Although to a certain extent this is logical, it does lead to some influence in the prior distribution.

Consider now the priors for the parameters, M and S, of the component distributions themselves. The majority of previous work has concentrated on mixtures of normal distributions. Most authors use a normal prior for the means of the components and a gamma distribution for the inverse variances (or equivalently and inverse gamma distribution for the variances). This choice of distributions gives some advantages of conjugacy. This structure is extended in [Richardson and Green 1997] to include a hyperprior structure for the shape parameter in the gamma distribution for the inverse variance. Adding such a hyperprior makes the prior distribution for the component variances a little vaguer and can result in more sensible posterior distributions [Berkhof et al. 2003]. However, hyperpriors add to the complexity of the set up for the model and make the prior structure less transparent to a non-expert user. For example adding a hyperprior gamma distribution for the shape parameter in the gamma prior is equivalent to assuming a three parameter prior which includes a factor that is a modified Bessel function of the second kind.

The authors in [Roeder and Wasserman 1997] use what they describe as partially proper priors for the means and standard deviations of the component parameters. These are partially proper in the sense that the overall scale and location of the parameters require no subjective input but the parameters for different components are linked. The means are loosely linked through a Markov Chain, which means that the prior distribution for the position of an individual component mean in parameter space is flat but the distribution describing the distance between two component means is not. The joint prior distribution for the component variances is a product of scaled inverse-chi distributions with a common scale parameter and common degrees of freedom. This has the effect of pushing all of the component standard deviations towards some common, unspecified value. The prior requires two hyperparameters, one influencing the distance between the component means and the other affecting the difference in the scale of the component variances.

Although this choice of prior distribution could be used in many different applications without adaptation, it does impose some structure on the problem through having

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

non-flat distributions describing the distance between the component means and the difference in scale of the component variances. A prior distribution that imposes some scale on the component means and variances but treats them independently may actually impart less information. Further problems occur with Roeder and Wasserman's approach if the data being modelled comes from a mixture of components when two or more of those components have the same mean. The prior that they use has zero probability of this occurring and so prevents the correct posterior probability distribution being obtained.

The choice of prior for the base distribution parameters moreover does have some effect on the prior probability for the number of components in the model and so choosing a suitably vague prior for this problem is more difficult than it might first appear. The way the prior probability for the number of components can vary, dependent on the variance of the prior distributions for the component means, is described in [Stephens 2000] . For a very small variance, representing a strong belief that the prior information about the means is correct, the prior distribution favours models with a low number of components. As the variance is increased, to represent vaguer prior knowledge of the position of the component means, initially more components are fitted with means spread across the range of the data, but continuing to increase the variance will eventually favour fitting fewer components. In the limit of the variance tending to ∞, the distribution of k becomes independent of the data [Stephens 2000] and this heavily favours a one component model.

A.2. Approximation for $\omega(\cdot)$ function

Consider the Weibull distribution with PDF

$$f(y) = \frac{\alpha}{\beta}(\frac{y}{\beta})^{\alpha-1}\exp(-(\frac{y}{\beta})^{\alpha})$$

This has mean $\mu = \beta\Gamma(\frac{1}{\alpha}+1)$ and variance $\sigma^2 = \beta^2(\Gamma(\frac{2}{\alpha}+1)-(\Gamma(\frac{1}{\alpha}+1))^2)$. We therefore have

$$\begin{aligned}\ln\ln(1+(\sigma/\mu)^2) &= \ln\{\ln[\Gamma(2z+1)] - 2\ln[\Gamma(z+1)]\}\\ &= R(z), \text{ say, where } z = 1/\alpha.\end{aligned}$$

Consider first the behaviour of $R(z)$ as $z \to 0$. Expanding $R(z)$ as a power series, we have

$$\ln(\ln(\Gamma(2z+1)) - 2\ln(\Gamma(z+1))) = \ln(\pi^2/6) + 2\ln z + O(z). \tag{49}$$

Now consider $R(z)$ as $z \to \infty$. Using a standard asymptotic formula, as given by Abramowitz and Stegun (1965, 6.1.41), we have

$$\begin{aligned}r_1 &= \ln(\Gamma(2z+1)) \sim (2z+1-\frac{1}{2})\ln(2z+1) - (2z+1) + \frac{1}{2}\ln(2\pi)\\ &\quad + \frac{1}{12(2z+1)} - \frac{1}{360(2z+1)^3} + O(\frac{1}{z^4})\end{aligned}$$

and

$$\begin{aligned}r_2 &= 2\ln(\Gamma(z+1)) \sim 2(z+1-\frac{1}{2})\ln(z+1) - 2(z+1) + \ln(2\pi)\\ &\quad + \frac{1}{6(z+1)} - \frac{1}{180(z+1)^3} + O(\frac{1}{z^4}).\end{aligned}$$

The log factor in the first term in the expression for r_1 is

$$\begin{aligned}\ln(2z+1) &= \ln 2z + \ln(1+\frac{1}{2z}) = \ln 2z + \frac{1}{2z} - \frac{1}{2(2z)^2} \\ &\quad + \frac{1}{3(2z)^3} - \frac{1}{4(2z)^4} + O(\frac{1}{z^5}).\end{aligned}$$

Therefore the first term in r_1 is

$$\begin{aligned}(2z+1-\frac{1}{2})\ln(2z+1) &= (2z+\frac{1}{2})(\ln 2z + \frac{1}{2z} - \frac{1}{2(2z)^2} \\ &\quad + \frac{1}{3(2z)^3} - \frac{1}{4(2z)^4} + O(\frac{1}{z^5})) \\ &= 2(\ln 2)z + 2z\ln z + 1 + \frac{1}{48z^2} - \frac{1}{96z^3} \\ &\quad + \frac{1}{2}\ln 2 + \frac{1}{2}\ln z + + O(\frac{1}{z^4}).\end{aligned}$$

We also have that the first term in r_2 is

$$\begin{aligned}2(z+1-\frac{1}{2})\ln(z+1) &= 2(z+1-\frac{1}{2})(\ln z + \frac{1}{z} - \frac{1}{2(z)^2} + \frac{1}{3(z)^3} \\ &\quad - \frac{1}{4(z)^4} + O(\frac{1}{z^5})) \\ &= 2z\ln z + \ln z + 2 + \frac{1}{6z^2} - \frac{1}{6z^3} + O(\frac{1}{z^4}).\end{aligned}$$

The difference in these two first terms is therefore

$$\begin{aligned}&2(\ln 2)z + 2z\ln z + 1 + \frac{1}{48z^2} - \frac{1}{96z^3} + \frac{1}{2}\ln 2 + \\ &\frac{1}{2}\ln z - (2z\ln z + \ln z + 2 + \frac{1}{6z^2} - \frac{1}{6z^3}) + O(\frac{1}{z^4}) \\ &= 2(\ln 2)z - \frac{1}{2}\ln z + \frac{1}{2}\ln 2 - 1 - \frac{7}{48z^2} + \frac{5}{32z^3} + O(\frac{1}{z^4}).\end{aligned}$$

Thus

$$\begin{aligned}r_1 - r_2 &= \ln(\Gamma(2z+1)) - 2\ln(\Gamma(z+1)) \\ &= 2(\ln 2)z - \frac{1}{2}\ln z + \frac{1}{2}\ln 2 - 1 + O(\frac{1}{z^2}).\end{aligned}$$

Hence

$$R(z) = \ln(r_1 - r_2) = \ln(z(2(\ln 2) - \frac{1}{2z}\ln z + (\frac{1}{2}\ln 2 - 1)\frac{1}{z}) + O(\frac{1}{z^3}))) \tag{50}$$

$$= \ln z + \ln(2(\ln 2) - \frac{1}{2z}\ln z + (\frac{1}{2}\ln 2 - 1)\frac{1}{z} + O(\frac{1}{z^3})) \tag{51}$$

$$= \ln z + \ln(2(\ln 2)) + O(\frac{\ln z}{z}) \quad \text{as } z \to \infty. \tag{52}$$

If we write $x = -\ln z$, so that $\alpha = \exp(x)$, we have from (49) and (52) that

$$\begin{aligned} R(\exp(-x)) &= \begin{cases} g(x) + O(\frac{x}{\exp(-x)}) \text{ as } x \to -\infty \\ h(x) + O(\exp(-x)) \text{ as } x \to \infty \end{cases} \\ &= y(x), \text{ say,} \end{aligned}$$

where

$$\begin{aligned} g(x) &= 0.3266 - x \text{ (using } \ln(2(\ln 2)) \simeq 0.326\,63\text{)}, \\ h(x) &= 0.4977 - 2x \text{ (using } \ln(\pi^2/6) \simeq 0.49770\text{)}. \end{aligned}$$

The function $y(x) = R(\exp(-x))$ can be represented by one arm of the hyperbola

$$(y + x - a)(y + 2x - b) = A,$$

for suitable chosen coefficients A, a, and b. Use of such a hyperbolic approximation allows inversion to express x in terms of y. The required solution is

$$x = \frac{1}{2}a + \frac{1}{4}b - \frac{3}{4}y - \frac{1}{4}\sqrt{4a^2 - 4ab + b^2 + 8A + (2b - 4a)y + y^2},$$

with a and b having values similar to $a = \ln(2(\ln 2))$, $b = \ln(\pi^2/6)$. The coefficients in (3) used in the spreadsheet version (the function subroutine WeibAlfa) correspond to the approximation

$$x = 0.5282 - 0.7565y - 0.3132\sqrt{6.180 - 0.5561y + 0.7057y^2}$$

which gives an α relative accuracy within 1% over the range coefficient of variation range $0.0001 \le \sigma/\mu \le 1000$.

A.3. Geweke's Theorem 2

Theorem 2 in [Geweke 1989], which is cited in the main paper, concerns the use of importance sampling (IS) to estimate

$$\bar{g} = \int g(\theta)\rho(\theta)d\theta$$

In this subsection alone, the notation is as in Geweke, except for $\rho(\theta)$ which is not used in [Geweke 1989] and which we use, again in this subsection alone, to represent the posterior distribution of θ.

Let the IS distribution have density $I(\theta)$. Normalizing constants are avoided by working with $p(\theta)$ and $I^*(\theta)$ instead, where $p(\theta) = c\rho(\theta)$, and $I^*(\theta) = d^{-1}I(\theta)$, the constants c and d being unknown. An estimator for $\bar{g} = E[g(\theta)]$ is

$$\bar{g}_n = \sum_{i=1}^{n} g(\theta_i)w(\theta_i) / \sum_{i=1}^{n} w(\theta_i)$$

where the θ_i have been drawn from $I(\theta)$, and

$$w(\theta) = p(\theta)/I^*(\theta).$$

Theorem 2 in [Geweke 1989] states that

$$\begin{aligned} n^{1/2}(\bar{g}_n - \bar{g}) &\Rightarrow N(0, \sigma^2) \\ n\hat{\sigma}_n^2 &\to \sigma^2 \end{aligned}$$

where

$$\hat{\sigma}_n^2 = \sum_{i=1}^{n}(g(\theta_i) - \bar{g}_n)^2 w^2(\theta_i)/(\sum_{i=1}^{n} w(\theta_i))^2$$

is the sample variance of the individual terms used to form $\bar{g}_n$. This is correct but the statement of the Theorem actually gives a formula for σ^2 that is different from that derived in the proof. The version in the proof, namely

$$\sigma^2 = c^{-2}d^{-1}\int_{\Theta}[g(\theta) - \bar{g}]^2 w(\theta)p(\theta)d\theta, \tag{53}$$

is the correct one. We note that σ^2 should not depend on the constants c and d. From (53) we have

$$\begin{aligned}
\sigma^2 &= c^{-2}d^{-1}\int_{\Theta}[g(\theta) - \bar{g}]^2 \frac{p(\theta}{I^*(\theta)}p(\theta)d\theta, \\
&= c^{-2}d^{-1}\int_{\Theta}[g(\theta) - \bar{g}]^2 \frac{(c^2 d)\rho^2(\theta)}{I(\theta)}d\theta \\
&= \int_{\Theta}[g(\theta) - \bar{g}]^2 \frac{\rho^2(\theta)}{I^2(\theta)}I(\theta)d\theta \\
&= \mathrm{Var}_I\left\{[g(\theta) - \bar{g}]\frac{\rho(\theta)}{I(\theta)}\right\}
\end{aligned}$$

which is independent of c and d, as required.

Geweke does not actually demonstrate the second part of the Theorem, that $n\hat{\sigma}_n^2 \rightarrow \sigma^2$. We can show the weak version of this by writing $n\hat{\sigma}_n^2$ as

$$n\hat{\sigma}_n^2 = \left(n^{-1}\sum_{i=1}^{n}[(g(\theta_i) - \bar{g}) + (\bar{g} - \bar{g}_n)]^2 w^2(\theta_i)\right) \Big/ \left(n^{-1}\sum_{i=1}^{n} w(\theta_i)\right)^2$$

The numerator is

$$[n^{-1}\sum_{i=1}^{n}(g(\theta_i) - \bar{g})^2 w^2(\theta_i)] + 2(\bar{g} - \bar{g}_n)[n^{-1}\sum_{i=1}^{n} g(\theta_i) - \bar{g})w^2(\theta_i)] + (\bar{g} - \bar{g}_n)^2[n^{-1}\sum_{i=1}^{n} w^2(\theta_i)].$$

The three terms in the square brackets all tend to constants in probability. In the first we have that

$$\begin{aligned}
E_I[(g(\theta_i) - \bar{g})^2 w^2(\theta_i)] &= \int [g(\theta) - \bar{g}]^2 \frac{p^2(\theta)}{I^{*2}(\theta)}I(\theta)d\theta \\
&= c^2 d^2 \int [g(\theta) - \bar{g}]^2 \frac{p^2(\theta)}{I^2(\theta)}I(\theta)d\theta \\
&= c^2 d^2 \mathrm{Var}_I\left\{[g(\theta) - \bar{g}]\frac{\rho(\theta)}{I(\theta)}\right\}.
\end{aligned} \tag{54}$$

Both the second and third terms contain the factor $\bar{g}_n - \bar{g}$ which tends to zero in probability. Thus, applying Slutsky's Theorem, the second and third terms also tend to zero in probability, so that the numerator, applying Slutsky's theorem again, tends (54).

The denominator is the product of two identical random variables each tending to the constant cd. By Slutsky's theorem their product tends to the square of this constant. Thus combining numerator and denominator and applying Slutsky's Theorem one more time, yields

$$n\hat{\sigma}_n^2 \to c^2 d^2 \mathrm{Var}_I \left\{ [g(\theta) - \bar{g}] \frac{\rho(\theta)}{I(\theta)} \right\} / (cd)^2 = \sigma^2$$

in probability as required.

A.4. 3ClearNorms Example, Additional Plots

Figures 4 and 5 contrast the posterior distributions of the component mean and SD parameters and weights for the 4-component normal mixture distribution fitted by the IS and MC methods to the 3ClearNorms data set. The distributions are unimodal in the IS case. The weights in the MC case are bimodal.

Figure 5 may explain why it may be unsatisfactory to estimate the value of a given parameter in the k-component mixture, for given k, by taking the average of the values obtained for this parameter over all steps of an MC run where the component value sampled is k. When the distribution of the parameter is not unimodal, this average value is not a very representative value.

Figure 6 is the analogue of Figure 1 but showing the predictive densities conditional on $k = 4$ instead of $k = 3$. The densities obtained using the IS averaged parameters method and the MPP parameters method identify the sharp peak due to the second component. The densities obtained using the MC method are too smooth, with the density obtained using the MC averaged parameters method particularly so.

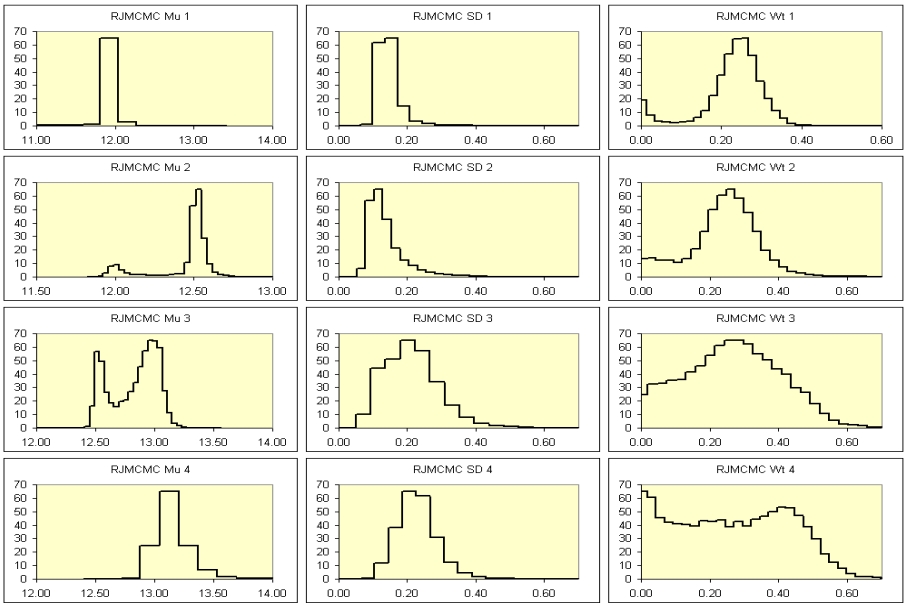

Fig. 4. 3ClearNorms Example: Posterior distributions of component mean and SD parameters and weights for the 4-component MC normal mixture fit

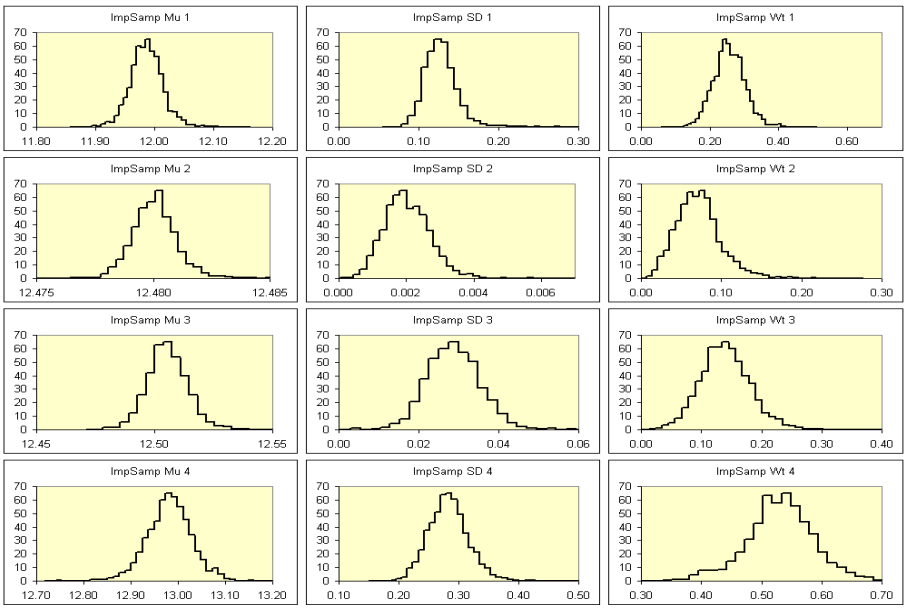

Fig. 5. 3ClearNorms Example: Posterior distributions of component mean and SD parameters and weights for the 4-component IS normal mixture fit

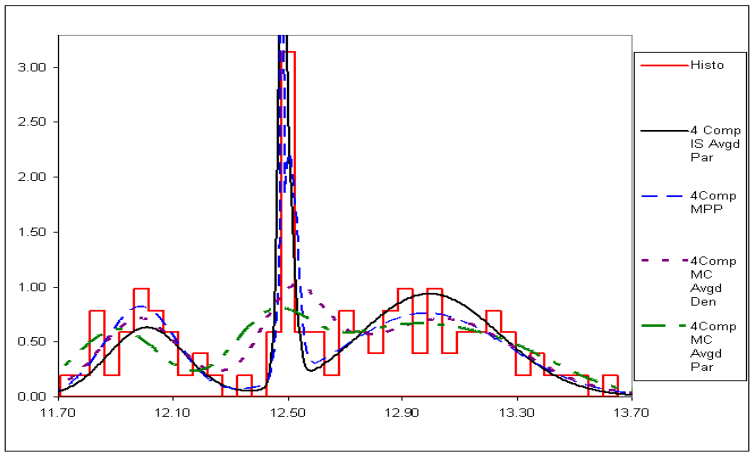

Fig. 6. 3ClearNorms example: frequency histogram; IS 4 component normal fits using averaged parameter values and MPP estimators; 4 component MC normal fits using averaged density values and averaged parameter values

A.5. LotSize Example, Bezier distribution fitted using PRIME

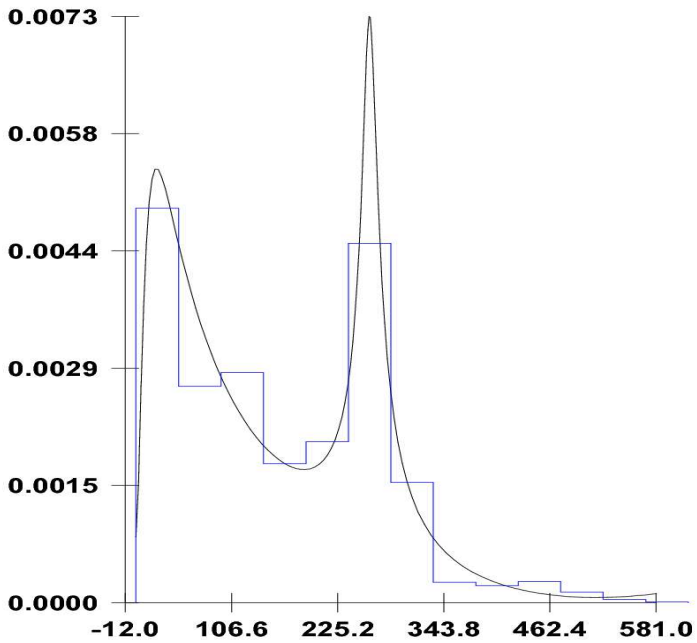

Fig. 7. LotSize Example: Bezier distribution fit using PRIME package