ORIGINAL ARTICLE

Input Modelling for Multimodal Data

Russell C.H. Cheng[a] and Christine S.M. Currie[a]

[a]Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

ARTICLE HISTORY
Compiled April 1, 2019

ABSTRACT
Multimodal data occurs frequently in discrete-event simulation input analysis, typically arising when an input sample stream comes from different sources. A finite mixture distribution is a simple input model for representing such data, but fitting a mixture distribution is not straightforward as the problem is well-known to be statistically non-standard. Even though much studied, the most common fitting approach, Bayesian reversible jump Markov Chain Monte Carlo (RJMCMC), is not very satisfactory for use in setting up input models. We describe an alternative Bayesian approach, MAPIS, which uses maximum a posteriori estimation with importance sampling, showing it overcomes the main problems encountered with RJMCMC. We demonstrate use of a publicly-available implementation of MAPIS, which we have called FineMix, applying it to practical examples coming from finance and manufacturing.

Keywords Simulation; statistics; input analysis; mixture models

1. Introduction

Input modelling for discrete-event simulation (DES) aims to identify appropriate probability distributions for characterising the behaviour of the streams of random variables that represent the inputs to DES models. A recent review of input modelling has been given by Cheng (2017a) showing how the topic has grown so that there is now an extensive literature. The most basic situation is the simple one where input random variables are independently and identically distributed and drawn from well-known distributions such as the normal, lognormal, gamma or Weibull. This situation is discussed in detail in Law (2007). A wider range of distributional shapes has been discussed in Kuhl et al. (2010). Two generalizations, reviewed by Cheng (2017a), have been studied in some detail, namely: (i) where the random variables are multivariate, and (ii) where they are correlated. See for example Deler and Nelson (2001); Ghosh and Henderson (2001); Nelson and Yamnitsky (1998). A third generalization, though mentioned in Cheng (2017a) and Kuhl et al. (2010), has not been so well discussed, where input random variables have a multimodal distribution, and most likely follow a finite mixture distribution. The purpose of this article is to discuss such distributions and their modelling in DES.

The article extends preliminary work in which we considered only mixtures of normal distributions in Cheng and Currie (2003) and relates the ideas presented in Cheng

CONTACT Christine S.M. Currie Email: christine.currie@soton.ac.uk

(2017b) to the specific application of input modelling for DES. We provide additional theoretical insights into the problem that are not included in Cheng (2017b) and compare our proposed method MAPIS, which uses maximum a posteriori estimation with importance sampling, with the most common fitting approach, Bayesian reversible jump Markov Chain Monte Carlo (RJMCMC), on a selection of examples.

The most common form of finite mixture distribution studied in the literature is where the probability density function (PDF) is a weighted sum of a finite number of continuous distributions, which we call the *component* distributions, each with the *same* form for their PDF or base density, $g(\cdot)$. This is a special case of the more general finite mixture distribution in which each of the base densities may take a different form.

We can write the PDF of the overall finite mixture as

$$f(y|\psi(k), \mathbf{w}(k), k) = \sum_{j=1}^{k} w_j g(y|\psi_j), \tag{1}$$

where $\psi_j(k) = (\mu_j, \sigma_j)$ are the parameters of the base density for component j, $j = 1, \ldots, k$ and $\psi(k) = (\psi_1, \psi_2, ..., \psi_k)^T$. Note that the number of components k is included as a parameter. The w_j are component weights, satisfying $0 \leq w_1$, $w_2, ...,$ $w_k \leq 1$, $\sum_{j=1}^{k} w_j = 1$ and written as $\mathbf{w}(k) = (w_1, w_2, ..., w_k)^T$. We omit the k dependence of the individual component parameters $\psi_j, w_j, j = 1, ..., k$, to avoid clumsy notation. When we do not need to consider $\psi(k)$ and $\mathbf{w}(k)$ separately we shall write $\theta(k) = (\psi(k), \mathbf{w}(k))$, referring to this simply as the vector of component parameters with the weights tacitly included.

There is a focus in the literature on the case where $g(\cdot)$ is the normal density with just two parameters so that $\psi_j = (\mu_j, \sigma_j)$, where μ_j and σ_j are the mean and standard deviation (SD) of component j. However our method also covers other two-component base densities: lognormal, extreme value (EV), negative extreme value (NEV), Weibull, gamma, and inverse Gaussian (IG).

We consider the DES situation where we wish to use an input model obtained under the following assumption.

Assumption A0: The finite mixture exists and has PDF of the form

$$f_0 = f(\cdot|\theta_0(k_0), k_0) \tag{2}$$

where f is as in Equation (1), with k_0, $\theta_0(k_0)$ regarded as a fixed set of true values, *but which are all unknown.* We have a random sample $\mathbf{y} = (y_1, y_2, ..., y_N)^T$, drawn from the distribution (2), that we can use to estimate the parameters, including k_0.

To obtain an input model based on Assumption A0, we have three requirements for the output of the fitting algorithm:

Requirement R1: Point estimates $\hat{k}$, $\hat{\theta}(\hat{k})$. We can then take the input model as $f(\cdot|\hat{\theta}(\hat{k}), \hat{k})$, generating any random variates required in the simulation from this fitted distribution.

Requirement R2: An estimate of the distribution of the point estimates. This is needed if we wish to estimate the uncertainty in the estimated input model.

Requirement R3: The base density is assumed to be unimodal.

The third requirement might be considered to be part of Assumption A0, and is true for the two-component base densities mentioned previously. Making this explicit simplifies discussion to unimodal fitted components with parameters whose posterior distributions are also unimodal.

Finite mixture models have a wide range of application (see for example McLachlan and Peel (2000)) and are of particular relevance where the combined overall input is made up of component distributions representing different types of input. Provided k and $\theta(k)$ are completely known, it is straightforward to generate variates from the distribution, so that a finite mixture model provides an easily implementable input model.

To date the most widely studied methods of fitting the finite mixture (1) are maximum likelihood (ML) and Bayesian Markov chain Monte-Carlo (MCMC), most notably the reversible jump version (RJMCMC) introduced in Richardson and Green (1997). Detailed discussion of both approaches are given by McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). Estimation of the parameters when k is unknown is recognised to be a non-standard problem. In brief, the issue is that if one attempts to estimate parameters conditionally for a given k, where k is greater than the true value k_0, (what is called an overfitted model by Rousseau and Mengersen (2011)) we have an indeterminate problem in which $(k-k_0)$ components of the model are non-estimable as they do not exist. This causes issues for both Bayesian and classical (ML) analysis of the problem. In the Bayesian case, using MCMC is unsatisfactory when the true value k_0 is unknown, as estimates of the posterior distributions of the component parameters and weights are very difficult to interpret, often becoming multimodal. Richardson and Green (1997) discuss multimodality in posteriors when using RJMCMC without coming to any definitive conclusions, and although Rousseau and Mengersen (2011) provide a much more definitive explanation of the observed multimodality of posteriors of component parameters, this latter work does not mention multimodality explicitly. In Section 2.2, we discuss the issue of multimodality in more detail, extending the discussion in Cheng (2017b) to give a new characterization of multimodality in the posterior distributions, suggesting why and how it arises using MCMC methods.

In this paper we use a Bayesian approach, which we call the MAPIS method, to fit mixture models to multimodal data. This does not rely on MCMC. Instead point estimators of $\theta(k)$ conditional on k, which we shall write as $\tilde{\theta}(k)$, are obtained for each k using the *maximum a posteriori* (MAP) method with k being increased systematically stepwise. Sufficiently informative priors are used to ensure that the posterior distributions of $\theta(k)$ conditional on k, are all unimodal with modes estimable by MAP. Thus precisely k components are fitted at each k. Posterior distributions of $\theta(k)$ conditional on k are then obtained using importance sampling (IS), with the MAP estimators $\tilde{\theta}(k)$ playing a key role in constructing the candidate distributions used in IS.

We believe MAPIS is sufficiently reliable for practical use, and an implementation of MAPIS known as FineMix (written in C with an Excel interface) is available for download at `http://www.curries.org.uk/christine/`. In this article we apply MAPIS to four examples coming from finance and manufacturing. A library of real data samples including these examples and over a dozen others from different application areas are provided with FineMix.

In addition to handling the problem of multimodal posterior distributions of component parameters, MAPIS is also able to detect so-called *spikes* in the frequency histogram, tight clusters of observations with very similar values. These correspond to components with small variance and weight. Spikes are hard to detect using MCMC, even when they are due to a real component, because of multimodality in the posterior distributions, as we illustrate in Section 4.

We formulate the Bayesian model in Section 2 and discuss multimodality of posterior distributions in Section 2.2, before describing MAPIS in more detail in Section 3. Its practical performance is discussed in Section 4 and compared with RJMCMC using

examples from a range of applications. We then conclude in Section 5.

2. Bayesian Estimation for Finite Mixture Models

In considering Bayesian methods we shall write K for the number of components to indicate that it is being treated as a Bayesian variable with a prior distribution, and likewise we write $\boldsymbol{\Theta}(K) = (\boldsymbol{\Psi}(K), \mathbf{W}(K))$ for the component parameters and weights. We use the lower case versions k, $\theta(k)$ to indicate given values.

In Bayesian modelling of finite mixtures, the main objective is to estimate the posterior distributions of K and $\boldsymbol{\Theta}(K)$. This is most conveniently done by using some appropriate method to estimate $\pi(k|\mathbf{y})$, the posterior distribution of K, and also to estimate $\pi_\theta[\theta(k)|k, \mathbf{y}]$, the posterior distributions of the component parameter $\boldsymbol{\Theta}(k)$ conditional on k. To satisfy requirement R1, we need to obtain point estimates $\tilde{k}$, $\tilde{\theta}(k)$ from these posterior distributions.

We assume that the posterior distribution of $\boldsymbol{\Theta}(K)$ and K has the form

$$\pi(\theta(k), k|\mathbf{y}) = \frac{f[\mathbf{y}|\theta(k), k]\pi[\theta(k), k]}{\sum_{l=1}^{k_{\max}} J(l|\mathbf{y})}, \ \theta(k) \in \mathbf{E}(k), \ k = 1, 2, ..., k_{\max} \tag{3}$$

where

$$J(k|\mathbf{y}) = \int_{\mathbf{E}(k)} f[\mathbf{y}|\theta(k), k]\pi[\theta(k), k]d\theta(k), \tag{4}$$

and $\pi[\theta(k), k]$ is the prior distribution of k and $\theta(k)$; with $\mathbf{E}(k)$ the Euclidean region containing all possible $\theta(k)$ values, and $k_{\max}$ some finite upper limit that will not be exceeded by K.

We use the mean M and standard deviation S as the parameters for the seven base densities $g(\cdot)$ considered in this article: the normal, lognormal, extreme value (EV), negative extreme value (NEV), Weibull, gamma and inverse gamma (IG). The mean and standard deviation are not the standard parameters for some of the distributions we consider and Table 5 in the Appendix includes the transformations from the more standard parameters. 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; (ii) it 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 the mean and standard deviation gave rise to significantly more stable and consistent behaviour in the numerical optimization methods used in calculating the posterior distribution.

2.1. Prior Distributions

The choice of priors for finite mixture distributions has been addressed by a number of authors and we, as far as possible, use the same functional forms for the prior distributions in MAPIS as given in Richardson and Green (1997) for RJMCMC. This set up for the prior distributions works well in the situation we consider and enables a fair comparison of the two methods, as we discuss further in Section 4.

Prior for K: we use a discrete uniform distribution for K, namely

$$p_K(k) = \Pr\{K = k\} = 1/k_{max},\ k = 1, 2, \ldots, k_{max}$$

where $k_{\max}$ is a prescribed maximum number of components, which it is assumed will definitely not be exceeded so that $p_K(k) = 0$ for all other values of k.
Prior for Component Weights $W(K)$: for given $K = k$ we use the Dirichlet distribution with density

$$f_{W(k)}(w(k)) = \frac{\Gamma[(k)\delta]}{[\Gamma(\delta)]^k} \prod_{j=1}^{k} w_j^{\delta-1},\ 0 \leq w_1, w_2, ..., w_k \leq 1. \tag{5}$$

We discuss appropriate values for δ in the following section.
Prior for Component Means: In order for MAPIS to be robust to different data sets, we set the prior distributions for the component means so that they have both mean and variance that are of the same order as the data. 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 is where, in handling the multimodality problem, the corresponding posteriors are least sensitive to the choice of prior. We therefore use the prior

$$f_M(\mu) = \begin{cases} (2\kappa s)^{-1}, & \bar{y} - \kappa s < \mu < \bar{y} + \kappa s \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where $\bar{y}$ and s are the sample mean and standard deviation of the data set, 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.

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

$$f_M(\mu) = \frac{\Gamma(g+\alpha)}{\Gamma(g)\Gamma(\alpha)} \frac{r(r\mu)^{g-1}}{(1+r\mu)^{g+\alpha}},\ \mu > 0, \tag{7}$$

where r is a scaling parameter and α, g are two shape parameters (g not to be confused with the base density $g(\cdot)$, this latter always being written with the argument). All three need to be chosen and we discuss practical choices in Section 4.
Prior for Component SD: For SDs, S, we use the prior

$$f_S(\sigma) = 2\frac{\Gamma(\alpha+g)}{\Gamma(\alpha)\Gamma(g)} \frac{h^g \sigma^{2g-1}}{(1+h\sigma^2)^{\alpha+g}}, \tag{8}$$

as given by Cheng (2017b), where α, g, and h are the parameters used by Richardson and Green (1997), only they do not give this PDF explicitly, couching their discussion in terms of a hyperparametrized prior for S. The values used for the three parameters are discussed in Section 4.

In summary, the complete prior is

$$\pi[\theta(k), k] = \left\{ \prod_{j=1}^{k} [f_M(\mu_j) f_S(\sigma_j)] \right\} f_{W(k)}[w(k)] p_K(k), \;\; k = 1, 2, ..., k_{\max},$$

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

2.2. Multimodality and Overfitted Models

Rousseau and Mengersen (2011) show that f_0, the true PDF of the finite mixture of Equation (1), is consistently estimable under Assumption A0 but their Theorem 1 shows that when using MCMC in an *overfitted* model ($k > k_0$), the component parameters and weights are *nonidentifiable*. What happens depends critically on δ, the shape parameter in the Dirichlet weight prior, (5).

We summarize Part (ii) of their Theorem 1 as follows. When $\delta > 1$, so that the weight prior can be regarded as informative, the sum of the weights of the $(k-k_0)$ fitted components with the smallest weights ($\Omega = \min_\lambda \sum_{j=1}^{k-k_0} w_{\lambda(k)}$, where $\lambda(k)$ indicates the position of weight w_k in a non-decreasing list of the weight values), will not tend to zero but will remain positive asymptotically. Thus k' component weights, where $k' > k_0$ will remain positive, so that identification of the k_0 true components will be very difficult, no matter how large the data sample size.

Rousseau and Mengersen (2011) do not discuss the multimodality of posterior of component parameters. However, as they show that f_0 is consistently estimable, their Theorem 1(ii) must mean that, in overfitted models, f_0 is a mixture of $k' > k_0$ components, even asymptotically. The k_0 true components must therefore be split up over the k' fitted components, and this is how multimodality arises in the posterior distributions for the component parameters. We show in the Appendix how the true mixture $f_0(y) = \sum_{j=1}^{k_0} w_{0j} g(y|\psi_{0j})$ has many *alternative* representations involving $k' > k_0$ components:

$$f_0(y) = f[y|k', \theta(k')] = \sum_{i=1}^{k'} w_i h_i(y|\mathbf{w}_0, \rho_i, \psi_0), \tag{9}$$

where each h_i (see Equation 13 in the Appendix) is made up of different fragments of the true components $g(y|\psi_{0j}), j = 1, 2, ..., k_0$, explaining why fitted posterior parameter distributions may be multimodal.

Procedures that result in multimodal posterior distributions for parameters do not satisfy requirements R1 and R2 listed previously.

As pointed out by Richardson and Green (1997), an immediate effect of parameter posterior distributions being multimodal is that parameter values are more spread out. This makes an estimate of $f_0(y)$ over-smooth if it is calculated using parameter and weight values that are average values based on such posteriors. Often such an estimate does not even correspond sensibly to the shape of the sample histogram.

There is a further ramification of multimodality, that the posterior distribution of K will have probabilities $\pi(k|\mathbf{y})$ that are biased high for $k > k_0$. This means that the overall posterior distribution $\pi(k|\mathbf{y})$ $k = 1, 2, ..., k_{\max}$ is an unreliable indication of the comparative merits of the different k component fits. In particular it is quite likely

that the largest $\pi(k|\mathbf{y})$ will occur at a value of $k > k_0$ so that the point estimate $\tilde{k}$ is biased high.

For this reason Rousseau and Mengersen (2011) prefer the noninformative choice $\delta < 1$ although recent work by van Havre et al. (2015) demonstrates that this choice has its own problems. In contrast, Frühwirth-Schnatter (2006) considers the benefits of informative priors, and we focus on such priors in this paper. In the following section we describe why MAPIS does not suffer from the same issues of multimodality in the posterior distributions using the informative choice $\delta > 1$.

Label switching is another awkward problem that arises with MCMC when using symmetric priors, as is the case in finite mixtures. Our proposed alternative approach uses MAP estimation to directly estimate $\psi(k)$ and $\mathbf{w}(k)$ conditional on k. As pointed out by Jasra et al. (2005), MAP does not suffer from the label switching problem. Although Jasra et al. (2005) do not recommend MAP, our method of applying MAP circumvents the concerns they raise.

3. The MAPIS Method

As already mentioned, in MAPIS two steps are used to determine the posterior distribution for the parameters of the mixture model. In the first step, we use the MAP method of to obtain the point estimate $\tilde{\theta}(k)$, of $\theta(k) = (\psi(k), \mathbf{w}(k))$, corresponding to the mode of the posterior distribution conditional on k for each of the allowed values of K. Then in the second step, the posterior distributions of component parameters are estimated using importance sampling (IS), making use of the estimates $\tilde{\theta}(k)$ to set up importance sampling distributions. The structure of MAPIS has previously been described in Cheng (2017b) but we provide a short description here for convenience.

3.1. MAP estimation

In the MAP step, posterior point estimates of parameter components are obtained directly by fitting k-component mixtures sequentially for increasing $k = 1, 2, ..., k_{max}$, with MAP reoptimization carried out at each k. We use the Nelder Mead optimization routine (Nelder and Mead (1965)) to maximize the log posterior, $\ln(p[\mathbf{y}|\theta(k), k])$ for each k.

For $k = 1$, the Nelder Mead is started at $\mu_0 = \bar{y}$ and $\sigma_0 = s$, the respective sample mean and sample standard deviation of the sample $\mathbf{y}$. At each subsequent step the difference between the k-component fit obtained so far and the data sample is examined to see how a *meaningful* additional component might be added that would best reduce the discrepancy between the data and fitted model. This is then used as the starting point for the Nelder Mead optimization. Details of the optimization method are given in the Appendix, but the advantage of parametrizing the base distribution using its mean and standard deviation is now evident, as it makes this process of introducing additional components a straightforward one.

Given k, the MAP estimator $\tilde{\theta}(k)$ comprises the parameter values which maximize the posterior distribution conditional on k. Calculation of $\tilde{\theta}(k)$ is simplified by noting that in the maximization of $\pi(\theta(k), k|\mathbf{y})$ we can omit the denominator $\sum_{l=1}^{k_{\max}} J(l|\mathbf{y})$ as it is a summation over all k. The MAP estimator, conditional on each k, $k =$

$1, 2, ..., k_{\max}$, is therefore

$$\tilde{\theta}(k) = \arg\max_{\theta(k)} \{ f[\mathbf{y}|\theta(k), k] \pi[\theta(k), k] \} \tag{10}$$

for $k = 1, 2, ..., k_{\max}$, where we can write

$$p[\mathbf{y}|\theta(k), k] = f[\mathbf{y}|\theta(k), k] \pi[\theta(k), k]$$

as the posterior of $\theta(k)$ conditional on k.

At each k, using MAP, the problem is thus effectively a standard estimation of parameters of the mixture (1) with precisely k components, so that posteriors become unimodal as data sample size increases. The only possible alternative representation (9) is where two fitted components are identical with exactly the same original functional form $g(\cdot)$ and identical ψ. This can happen when $k' > k_0$ but even then only rarely as the MAP procedure favours fitting to different features. This is easily detected and allowed for, as is done in *FineMix*, the MAPIS implementation described in Section 4.

Note that although MAPIS is still Bayesian, it is different from MCMC in that the objective of the MAP stage is to produce point estimates of the set of component parameter values $\theta(k)$ for each k, satisfying requirement R1. This is achieved by estimating precisely k possible components of base component form $g(\cdot|\psi_j)$, $j = 1, 2, ..., k$ and a corresponding weight w_j at each given k, each component being non-degenerate and unimodal as in R3. (The ψ_j and w_j being different at different k.)

In order to ensure components are non-degenerate and unimodal, prior parameters are chosen such that the prior distributions are sufficiently informative to ensure the maximum of the posterior distribution of $\theta(k)$ tends to zero at the boundary of the parameter space $\mathbf{\Theta}(k)$. The MAP estimator $\tilde{\theta}(k)$, with k given, will then be an internal point of $\mathbf{\Theta}(k)$ corresponding to a k component mixture with all components nondegenerate.

The boundary is approached if any component mean $\mu_j \to \mu_0$ (where μ_0 is a given lower bound of μ), or if any SD $\sigma_j \to 0$, or if any weight $w_j \to 0$. Setting $\delta > 1$ and $g > 0.5$ in the Dirichlet and SD priors of Equations (5) and (8) ensures that all priors remain bounded and at least one tends to zero as the boundary of $\mathbf{\Theta}(k)$ is approached. This guarantees that the maximum of the posterior distribution of $\theta(k)$ tends to zero as the boundary of $\mathbf{\Theta}(k)$ is approached.

3.2. Importance Sampling

To evaluate the full posterior distribution (3) we need to estimate the $J(k|\mathbf{y})$ integral of (4) for all k. IS is a numerical method for evaluating a general integral $\int_\Theta h(\theta) d\theta$, an early reference for which is Hammersley and Handscomb (1964). In IS, samples are drawn from a candidate distribution and weighted by the ratio of the integrand $h(\theta)$ evaluated at the sample point, to the value of the candidate distribution at that point. If the candidate distribution is chosen correctly, this results in the sampling being concentrated in parts of parameter space at which the integrand is large, i.e. more *important* parts of parameter space.

We use the method introduced by Geweke (1989) to estimate a posterior distribution using importance sampling. As pointed out in Cheng (2017b), because each sampled point in IS is obtained independently of all other points, we can find the normalising constant for each of the posterior distributions, conditional on k, for each value of k,

independently of one another. Therefore, when running the importance sampling, we first sample the number of components, k, with equal probability $1/k_{max}$ of choosing k in the range $1, \ldots, k_{max}$. Then we sample from the candidate distribution for the model with k components $c_k[\theta(k)]$, a multivariate Student t-distribution, centred on the mode of the posterior, as calculated by the optimization routine, with covariance matrix set equal to minus the inverse of the Hessian of the log posterior density evaluated at the mode of the posterior distribution conditional on k. Details of this procedure are given in the Appendix.

One of the benefits of our IS routine is that both the prior for k and the importance sampling of k are uniform, therefore we have no need to calculate the normalising integrals of the posterior distribution over the $(\psi(k), \mathbf{w}(k))$ space explicitly. This allows us to take the most likely k,

$$\tilde{k} = \arg\max_k \tilde{\pi}(k|y), \tag{11}$$

as the best estimate of k.

To cope with the possibility that the resulting IS distribution is a poor representation of the posterior distribution, Geweke (1989) suggests adjustments of the IS distribution in each direction of each parameter axis. We have not implemented this more elaborate version but report results using just the Student t-distribution. We would expect results to be satisfactory when $k = k_0$, but for k different from k_0 it is likely to introduce a bias in estimating $\pi(k|\mathbf{y})$ making it smaller than the true value. Thus our method will produce an estimate of the posterior distribution of k that is likely to be more concentrated about k_0 than with an MCMC method.

4. Examples

This section describes how MAPIS is used in practice via FineMix, our implementation of the theory described previously. It also serves as a comparison between MAPIS and RJMCMC, demonstrating examples of multimodality in the posterior distributions as discussed previously in Section 2.2. We consider four examples: (1) An artificial sample of size 500 drawn from a four-component normal mixture with all components and weights known; (2) a Lot-Size example: this is given in Wagner and Wilson (1996a) and is a sample of 2083 lot-sizes, in thousands, of surface mounted capacitors being stored in a facility while waiting for their insulation resistance to be tested; (3) an Activity-Cycle example: 1500 activity-cycle times observed in the production process of a car manufacturer; (4) a Credit Risk example: a subsample of size 2000 drawn from a larger complex financial data set comprising the loss given defaults (LGD) of 7051 clients.

We shall compare MAPIS and RJMCMC in each example, but before considering the examples we make some general comments about fitting using MAPIS.

As discussed in Section 3, MAPIS requires setting $\delta > 1$. Low values of δ enable identification of components with small variance but if set too low, can result in the fitting of spurious spikes to random clusters in the sample. This latter issue is not usually a problem for $k > k_0$, so a simple fitting strategy is to select δ near to unity, checking the fit of the CDF to the EDF and ensuring that the posterior parameter and weight distributions are unimodal for all k, then stopping and reducing $k_{\max}$, if a spurious fit arises. FineMix also includes a diagnostic check on the size of the coefficient of variation of each component and suggests either using a smaller $k_{\max}$

or refitting with δ (and possibly g) increased so that the problem is not encountered for the range of k considered. We recommend a maximum value for δ of 5 based on practical experience.

In our examples we find that the fits are not sensitive to the precise values of the three remaining parameters g, α and h. Although in Section 3 we state that $g > 0.5$, in practice this does not seem as important as the condition $\delta > 1$. We have therefore used the value $g = 0.5$ as our default, but have used $g = 0.2$, the default value of Richardson and Green (1997), in one example, and g = 2 (with $g = \delta$) in another example. We set $\alpha = 2$, the default value used by Richardson and Green (1997). It will be seen from the prior of S, as given in Equation (8), that h is a scaling parameter of S^2. Richardson and Green (1997) used $h = 10/R^2$, where R is the data range and we use the same value in our examples.

FineMix includes three further diagnostic checks. The first determines how distinct the fitted components are with a flag being raised if two adjacent mean estimates m_j,m_{j+1} and corresponding SD estimates s_{j+1}, s_j satisfy

$$(m_{j+1}^2 + m_j^2)^{-1/2}(m_{j+1} - m_j) + (s_{j+1}^2 + s_j^2)^{-1/2}(s_{j+1} - s_j) < 0.001.$$

The second check is used to determine whether the optimization routine has found an optimal solution (even if only a local optimum). 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 importance 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.

The third check is designed to alert the user to the presence of very low variance and low weight fitted components. Warnings are given for any fitted component where $\sigma < R/1000$ or where $w < 1/n$ where R and n are the sample range and sample size.

For RJMCMC we used the NMix simulation implementation (downloadable from `https://people.maths.bris.ac.uk/~mapjg/Nmix/`), to which we added an Excel front-end interface, which can be accessed via FineMix. Default parameters were used to match those used in FineMix, except where specified in the examples.

4.1. Example 1: Four Normals Mixture

This is a sample, of size 500, drawn from the four-component normal mixture with known parameters (μ, w) = (5.0, 0.4), (9.0, 0.3), (13.0, 0.2) and (17.0, 0.2), all with the same SD $\sigma = 1.5$. We fitted the normal mixture model using both RJMCMC and MAPIS. In order to illustrate how the output varies for different values of δ, we give results for $\delta = 1.1$, 2, and 4, with $g = 0.5$ held fixed.

For all combinations of method and δ, run lengths of $50,000$ replications were used. In the RJMCMC cases there was an additional burn-in of $5,000$ replications. In order to improve the accuracy of run times, we measured the durations with 100,000 replications, and an additional burn-in of 10,000 replications for RJMCMC. One run took 99 seconds for RJMCMC and 123 secs for MAPIS using an Intel Pentium G2030 running at 3GHz.

Table 1. Artificial four normals example: estimated posterior distribution of k, obtained by fitting a mixture of normal distributions using RJMCMC and MAPIS using $\delta = 1.1, 2, 4; g = \delta$

δ	k	3	4	5	6	7	8	9
1.1	$\tilde{\pi}_{MC}(k\|\mathbf{y})$		0.277	**0.281**	0.198	0.120	0.061	0.031
2.0	$\tilde{\pi}_{MC}(k\|\mathbf{y})$	0.004	0.298	**0.306**	0.191	0.105	0.053	0.024
4.0	$\tilde{\pi}_{MC}(k\|\mathbf{y})$	0.002	0.288	**0.305**	0.202	0.109	0.050	0.023
1.1	$\tilde{\pi}_{IS}(k\|\mathbf{y})$	0.173	**0.815**	0.02				
2.0	$\tilde{\pi}_{IS}(k\|\mathbf{y})$	0.103	**0.883**	0.008				
4.0	$\tilde{\pi}_{IS}(k\|\mathbf{y})$	0.181	**0.775**	0.041	0.001			

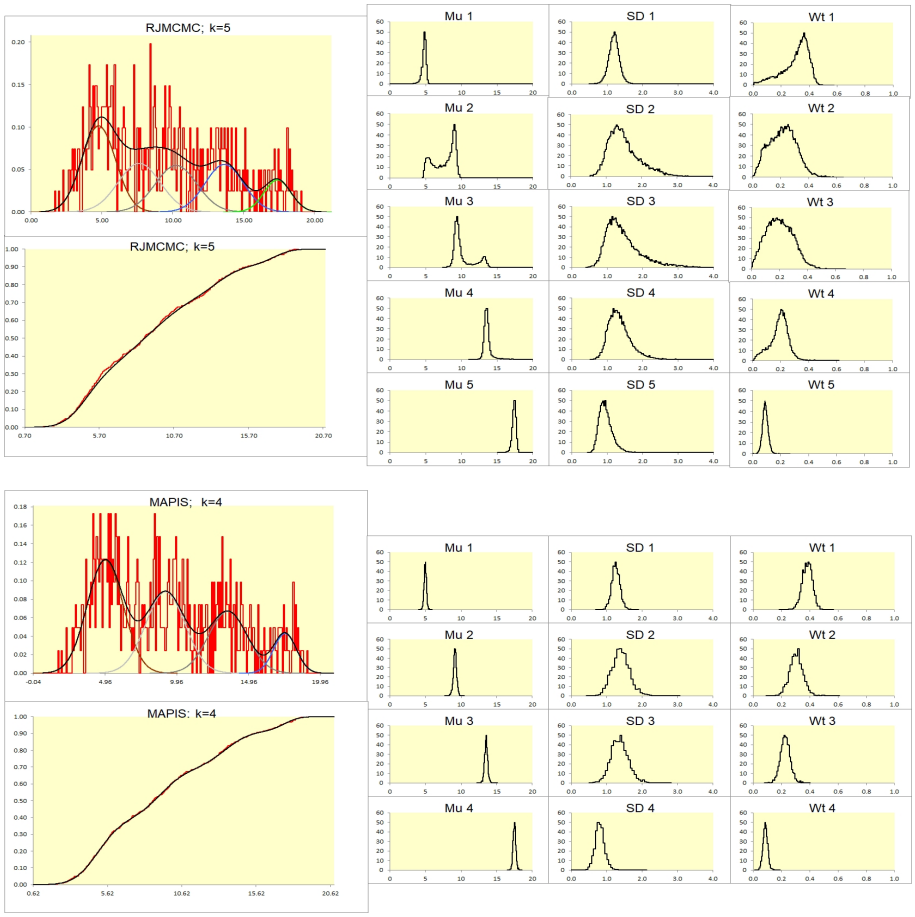

Figure 1. Four Normals Artificial Example: CDFs and PDFs and Posterior Parameter and Weight Probability Distributions for the 5-component RJMCMC best fit (upper charts) and the 4-component MAPIS best fit (lower charts). It is clear in the RJMCMC case, that the *second true* component has split in two, appearing together with a small part of the *first true* component as the *second fitted* component and together with a small part of the *third true* component as the *third fitted* component.

Table 1 shows the posterior distributions of $\tilde{\pi}_{MC}(k|\mathbf{y})$, $\tilde{\pi}_{IS}(k|\mathbf{y})$, $k = 3, 4, ..., 9$, estimated by RJMCMC and MAPIS respectively using each δ value. They show that the choice of δ makes negligible difference in terms of drawing statistical inferences. RJMCMC suggests the best choice for the number of components, the k where $\tilde{\pi}_{MC}(k|\mathbf{y}$ is maximized, to be $\tilde{k}_{MC} = 5$, whilst MAPIS gives $\tilde{k}_{IS} = 4$.

Figure 1 shows the best-fit CDFs and PDFs for RJMCMC and MAPIS, correspond-

Table 2. Lot-Size example: Posterior probability distribution of the number of components k calculated using RJMCMC and MAPIS

k	5	6	7	8	9	10	11	12
$\tilde{\pi}_{MC}(k\|\mathbf{y})$		0.004	0.177	**0.318**	0.246	0.139	0.069	0.030
$\tilde{\pi}_{IS}(k\|\mathbf{y})$	0.348	**0.375**	0.252	0.024				

ing to $\tilde{k}_{MC} = 5$ (upper charts) and $\tilde{k}_{IS} = 4$ (lower charts) respectively. The posterior distributions of the fitted component parameters are shown to the right of the figure. These are unimodal in the MAPIS case and correspond quite well to the true parameter values. Conversely for RJCMC, there is evidence of multimodality in the fitted posterior distributions where the posteriors of the means of the first three true components have split, with the first contributing to both the posteriors of the first and second fitted means, the second to the posteriors of the second and third fitted means and the third to the posteriors of the third and fourth fitted means. Although RJMCMC suggests $k = 5$ to be the optimal value, the fit of the model is actually noticeably poorer than that for $k = 4$, which we do not reproduce here, and moreover in this $k = 4$ case, all the RJMCMC component parameter posteriors are unimodal.

4.2. Example 2: the Lot-Size Example

The second example is a real data set given in Wagner and Wilson (1996b) who use it to illustrate use of PRIME, a method that they propose for fitting multimodal data using a sum of Bézier curves to represent the fitted CDF. Wagner and Wilson (1996a) used to supply a very user-friendly implementation of PRIME with a graphical interface but this no longer appears to be available. We emphasize that PRIME, though easy to use, can only be regarded as an exploratory tool, and is not suitable for our fitting purposes as it does not satisfy requirements R1 and R2, making it difficult for generating random variates to use in a simulation.

Wagner and Wilson describe the lot-size sample as being bimodal and the frequency histogram of the data set depicted in Figures 6 and 8 of their article seems to show this; however their Figure 10, which also depicts the frequency histogram but using a smaller bin width, seems to show a more multimodal behaviour and this is what we explore.

We again compare fitting a normal mixture model using the RJMCMC and MAPIS methods. In this example run lengths of $100,000$ were used, with RJMCMC using an additional $10,000$ replications for the burn-in. As with the other examples, the choice of δ and g is not too critical because of the relatively large sample size of $n = 2083$. We did try different δ as we have recommended above, but as the results were similar and to save space, we only report the comparison between MAPIS and RJMCMC in which $\delta = 2$, $g = 0.5$ were used for both methods.

Table 2 gives the probability distributions of k obtained using RJMCMC and MAPIS, which can be seen to be quite different, with $\tilde{k}_{MC} = 8$ and $\tilde{k}_{IS} = 6$. The CDFs and PDFs of the fitted models for MAPIS and RJMCMC are given in Figure 2, together with EDFs and frequency histograms. Though the fits are similar, there are small but clearly visible discrepancies between the EDF and the 8-component RJMCMC fitted CDF. In fact, though not shown here, the discrepancies are not apparent in the 6-component RJMCMC fit, which is similar to the MAPIS 6-component fit in Figure 2. So the two additional components have resulted in a counter-intuitive degradation of the RJMCMC fit. We do not show the posterior distributions for the

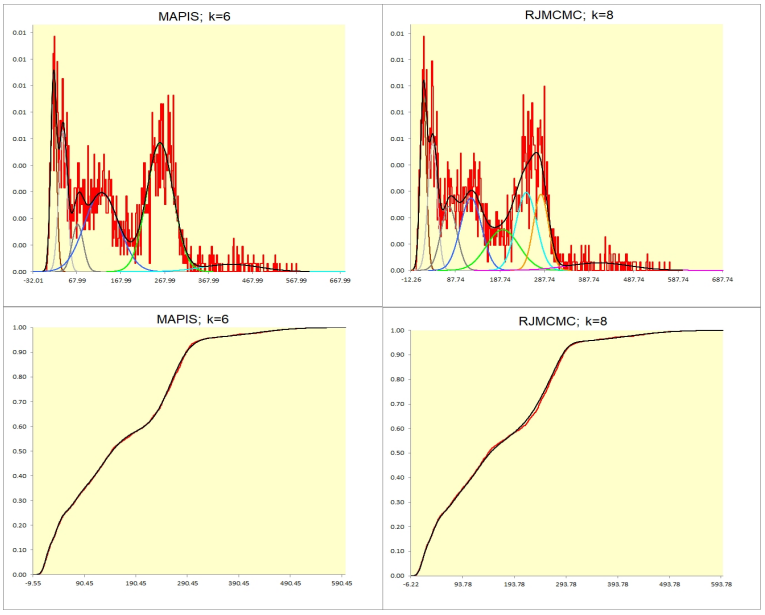

Figure 2. Lot-Size example: CDFs and PDFs of the MAPIS fitted 6-component and RJMCMC fitted 8-component normal mixture distributions with frequency histograms and EDFs.

component parameters for this example but there is again a presence of multimodality in the posterior distributions obtained by RJMCMC for $k = 8$, explaining why $k = 8$ is chosen over $k = 6$.

A more satisfactory fit can be obtained for this example using a mixture of Weibull components with $\tilde{\pi}_{IS}(5|\mathbf{y}) = 0.994$, making $\tilde{k}_{IS} = 5$ the obvious best choice for k. Though the two-parameter Weibull distribution has a fixed lower threshold of zero, its flexible shape, which includes both negative and positive skewness, makes it a more suitable base distribution for this example than the normal, but does not allow it to be compared with RJMCMC which only fits normal mixtures.

4.3. Example 3: the Activity-Cycle Example

This is a sample of an activity-cycle time observed in the production line of a UK car manufacturer. Management expected a particular informal work pattern to affect activity-cycle time, and a preliminary visual examination suggested the presence of a low-variance component of small weight corresponding to this behaviour. We use a relatively small sample to provide a stringent test for any finite mixture model fitting routine in that the small-variance component is then not very obviously present at first sight.

We fitted the normal mixture model using both RJMCMC and MAPIS. In both we used $\delta = 2$ and $g = 0.2$, where g is set to the default value used by Richardson and Green (1997).

We applied RJMCMC and MAPIS, using $100,000$ replications with each method, again with an additional $10,000$ replications for the burn-in in the RJMCMC case. In this example, a run with 100,000 replications, with an additional 10,000 replications burn-in for RJMCMC, took 415 secs for RJMCMC, and 530 secs for MAPIS using an Intel Pentium G2030 running at 3GHz. Table 3 gives $\tilde{\pi}_{MC}(k|\mathbf{y})$ and $\tilde{\pi}_{IS}(k|\mathbf{y})$, the posterior distribution of k obtained by each method. As can be seen, MAPIS gives a

Table 3. Activity-Cycle example: estimated posterior distribution of k, obtained by fitting a mixture of normal distributions using RJMCMC and MAPIS.

k	3	4	5	6	7	8	9	10
$\tilde{\pi}_{MC}(k\|\mathbf{y})$		0.002	**0.473**	0.316	0.146	0.045	0.014	0.004
$\tilde{\pi}_{IS}(k\|\mathbf{y})$	0.027	0.037	**0.934**	0.002	0.000			

clear maximum posterior value of $\tilde{\pi}_{IS}(5|\mathbf{y}) = 0.934$, making $\tilde{k}_{IS} = 5$ the best MAPIS estimate of k.

Figure 3 depicts the CDFs and PDFs of the 5-component and 6-component models fitted by each method, and the fits for MAPIS appear to be very good. In particular the fourth component fitted by MAP is a spike with estimated parameters: mean =1.25, SD=0.006 and weight=0.03, so that SD and weight are both small. In comparison the RJMCMC fit is not satisfactory, failing to find this spike, and none of its fitted components having SD this small. Table 3 shows that the maximum posterior probability of k is also obtained at $k = 5$, though less definitively as $\tilde{\pi}_{MC}(5|\mathbf{y}) = 0.447$, with $\tilde{\pi}_{MC}(6|\mathbf{y}) = 0.316$ and $\tilde{\pi}_{MC}(7|\mathbf{y}) = 0.146$ not negligible. However there is one serious concern. In contrast to MAPIS, neither the 5 nor the 6-component RJMCMC fits has identified the component found by MAPIS with the small SD.

Figure 4 depicts, for the case $k = 6$, the estimated posterior distributions of all the component parameters and weights obtained by each method. The reason we choose to display the posterior distributions for $k = 6$ rather than $k = 5$ (the best choice of k for both methods) is to give RJMCMC as much opportunity as possible for detecting the spike with mean approximately 1.25. These show that the posterior distributions estimated by MAPIS are essentially all unimodal, mostly with a small spread, indicating unambiguous parameter values assigned to the fit, including the component with a mean of approximately 1.25 (component 4 in Figure 4), which corroborates the behaviour anticipated by the management team. RJMCMC produces component parameter and weight posterior distributions that are unsatisfactory. Although the estimated sixth component, highlighted with the label '$w_c = 0$' in the figure, is concentrated at $w = 0$, the component is not a spike as the SD is too large. Instead, this seems to imply that the true $k_0 < 6$, so that a meaningless component with weight near zero has been fitted, instead of the component with small SD and weight, and mean 1.25 identified by MAPIS and corresponding to real system behaviour. There is also evidence of multimodality in the RJMCMC posterior distributions for the case $k = 6$, making them difficult to use for input modelling.

Multimodality leads to the estimated posterior probability values $\tilde{\pi}(k|\mathbf{y})$, when estimated by RJMCMC, being overlarge when $k > k_0$. In our example, though we have not included plots, multimodality occurs in many fitted parameter posterior distributions for all $k \geq 6$ that we have examined. This would explain the high $\tilde{\pi}_{MC}(k|\mathbf{y})$ values for $k = 6, 7, ...$ given in Table 3 compared with the negligibly small corresponding $\tilde{\pi}_{IS}(k|\mathbf{y})$ values. Due at least in part to multimodality, the problem of oversmooth predictive densities also arises with the RJMCMC results. The left-hand plots in Figure 3 show this oversmoothness clearly in the case of the 6-component fit and also provide further evidence that RJMCMC has not picked up the spike component.

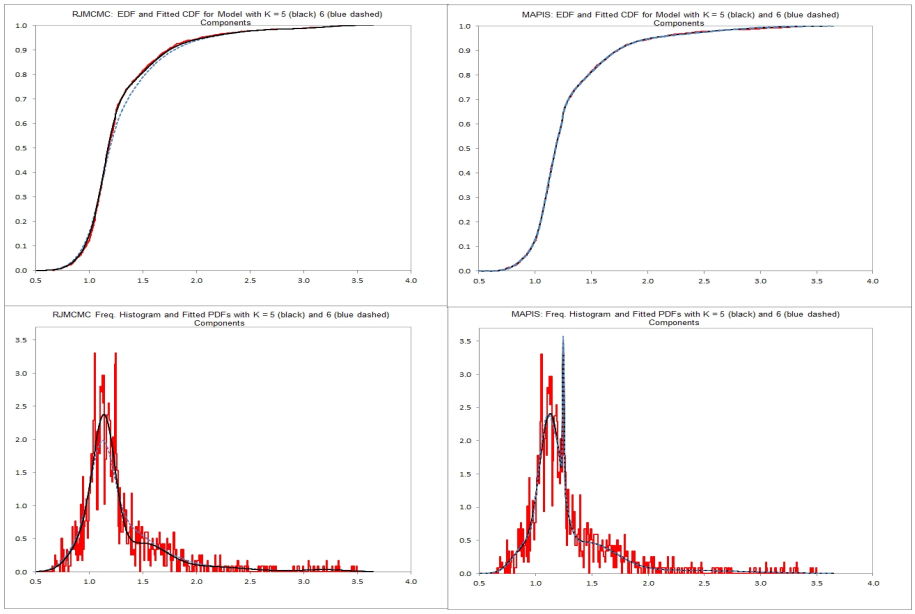

Figure 3. Activity-cycle example: frequency histogram and fitted 5 and 6-component normal mixture distributions using RJMCMC and MAPIS.

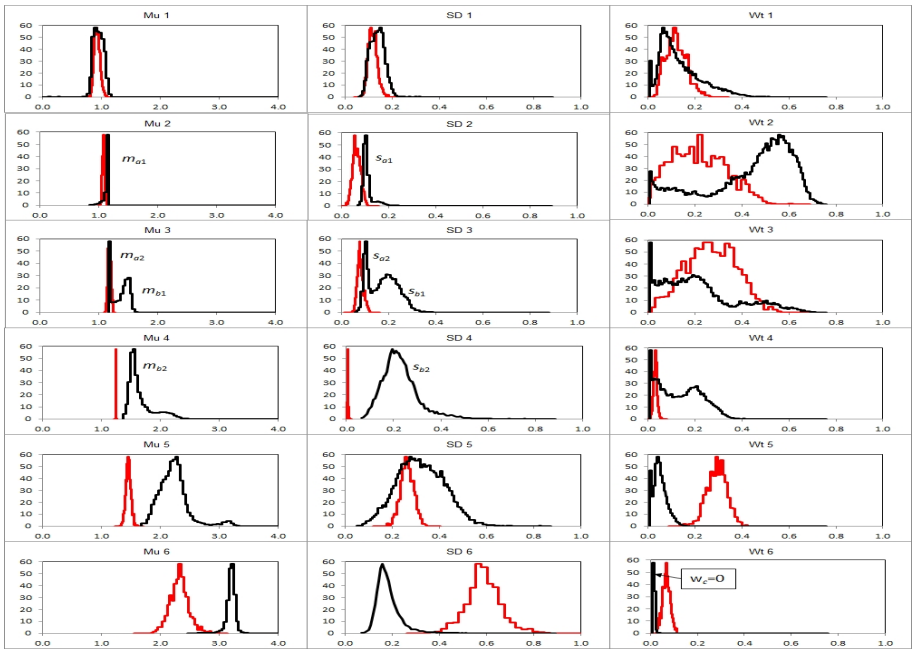

Figure 4. Activity-Cycle Example: frequency histogram and fitted parameter posterior probabilities for the 6 component model. RJMCMC posteriors - black, MAPIS posteriors - red.

4.4. Example 4: the Credit Risk Example

This example uses a data sample that, like the activity-cycle sample, contains a tightly clustered, distinctive subsample arising in a real context. We consider all seven base distributions that our method implements to illustrate how the Bayesian Information Criterion (BIC) might be used in practice when comparing fits using different base distributions. (See Schwarz (1978) and McLachlan and Peel (2000) for more discussion of the BIC). As a result we do not compare MAPIS with RJMCMC, omitting RJMCMC altogether as it is only able to fit normal mixtures.

The data are the loss given default (LGD) arising from 2000 clients at a bank. The sample includes only those cases where a non-zero loss was incurred. 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 spike with mean approximately 0.15, representing the behaviour of a certain kind of client. We used a relatively high value for the smoothing parameters with $\delta = 2$ and $g = 2$, to avoid fitting to spurious clusters.

We use MAPIS to estimate the BIC values for each of the k-component fits using MAP estimators, defining $\tilde{k}_{BIC}$ as the k at which the BIC value is maximized. The number of importance sampling replications is set at $50,000$. Table 4 lists $\tilde{k}_{BIC}$, $\tilde{k}_{IS}$ and the BIC for each base distribution. There is fairly close agreement between $\tilde{k}_{BIC}$ and $\tilde{k}_{IS}$, with only the gamma and inverse Gaussian distributions having a higher $\tilde{k}_{IS}$ value. Our advice is to use BIC principally as a method for determining which base distribution to use in the mixture. The overall maximum value of the BIC, taken over the seven base distributions, was obtained with a mixture of EV distributions, with $k = 4$.

Table 4. Credit Risk example: $\tilde{k}_{BIC}$ is the number of components corresponding to the maximum BIC value; 'BIC' is the BIC value obtained for each base model; $\tilde{k}_{IS}$ is the best k as estimated by MAPIS.

Base Model	$\tilde{k}_{BIC}$	BIC	$\tilde{k}_{IS}$	Base Model	$\tilde{k}_{BIC}$	BIC	$\tilde{k}_{IS}$
Normal	5	-422.8	5	Lognormal	4	-400.5	4
EV	**4**	**-391.8**	4	NEV	6	-466.1	6
Weibull	5	-427.6	5	Gamma	4	-403.4	5
IG	4	-402.4	5				

In a practical situation, running FineMix for each of the seven possible base distributions to obtain BIC values is a time-consuming process. Consequently, it is advisable to make a sensible initial choice for the base distribution, such as a distribution that has a shape that is similar to that of the individual peaks in the data. A visual check of the plots of the CDF of the fitted model and the EDF of the original data usually provides some indication of whether several components are being used to obtain the right shape for just one individual component.

The plots of the 4-component MAP fitted predictive densities for the EV distribution are given in Figure 5 and show that the component matching the spike of observations clustered at 0.15 with estimated parameter values $(\tilde{\mu}, \tilde{\sigma}, \tilde{w}) = (0.152, 0.0084, 0.028)$ is visible in the plot as a tall, narrow spike. The estimates for the parameters of this spike component are very stable, remaining essentially unchanged for $k = 5, 6, ..., 10$

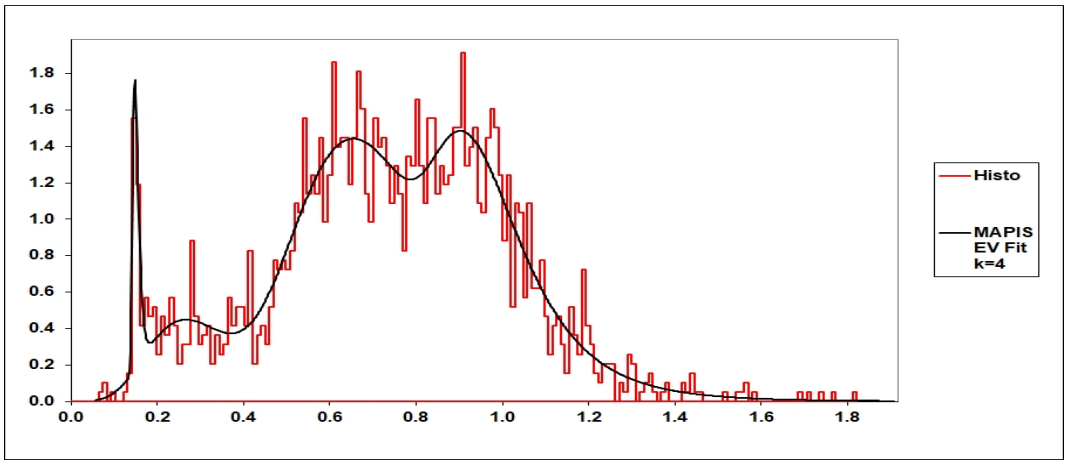

Figure 5. Credit Risk example: frequency histogram and fitted models for the Extreme Value (EV) fits with $k = 4$, using MAPIS.

and for all of the base distributions.

5. Conclusion

Finite mixture models are ideally suited for describing multimodal data and are particularly appropriate in simulation input modelling because of the ease of incorporating them into any simulation package. An entity in the simulation model will have a duration or other characteristic that follows a mixture model if the simulation first allocates it to a component with probability equal to the component weight and then samples from that component's distribution.

We describe a fitting method, MAPIS, that uses maximum a posteriori estimation of parameters and weights, conditional on k, for a range of values for k. In these examples we assume that k is unknown and MAPIS uses importance sampling to calculate its posterior distribution and from this, a point estimate of k.

The behaviour of MAPIS is compared in detail with RJMCMC in three examples: the first an artificial example with known parameter values and the remaining two arising from real data. The results demonstrate some of the issues with using RJMCMC when fitting multimodal input models; in particular oversmooth fits, a tendency to over-estimate the number of components k by including spurious components with very low weights, and the presence of multimodality in the posterior distributions for the component parameters. These issues make RJMCMC unsatisfactory when parameterizing input models for DES, in not providing reliable point estimators of k, the number of components, nor estimators of component parameters and weights. Similar issues have not been encountered with MAPIS.

The choice of base distribution is discussed in the final example describing a data set of loss given defaults. Typically a good choice of base distribution will result in a mixture with a smaller number of components but the results show that the BIC also provides a good guide when choosing a base density. Where there are discrepancies between the k value suggested by the BIC and that recommended after the IS procedure, this is usually due to the posterior probabilities for both k values being relatively similar. In these cases a little more care needs to be taken when running the simulation model to ensure that full account is taken of the posterior uncertainty in

k.

The investigations we present in Section 2.2 indicate the difficulty in dealing with overfitted models where $k > k_0$. Problems occur in overfitted models because more components k are being fitted than there are true components k_0. Rousseau and Mengersen (2011) identified two issues in this situation, one where the fitted posterior distributions indicate some component weights are small, and the other issue, which is discussed in this paper, where the posterior distributions associated with a true component become split between the posterior distributions associated with two or more fitted components. In both cases some of the supposed posterior distributions generated by MCMC methods do not correspond to meaningful possible components, so that point estimates of parameter values of true components are not properly identified. MAPIS works better because it identifies features in the data that could arise from actual components, assigning posterior distributions and weights to such possible components in a way that reflects how well the fitted posterior distributions then explain the data overall. Parameter estimates obtained using MAPIS are readily interpretable in terms of characteristics appearing in the data sample.

The FineMix implementation of MAPIS can be downloaded from the authors' website (`http://www.curries.org.uk/christine/`), with seven different options for the base component distribution. FineMix is robust and is able to cope with a wide range of data features; in particular the presence of small, but real, spikes in the distribution function. The use of seven base distributions also increases its flexibility and allows it to better model skewed data than using just a finite mixture of normal distributions.

References

Cheng, R. C. H. (2017a). History of input modeling. In Chan, W. K. V., D'Ambrogio, A., Zacharewicz, G., Mustafee, N., Wainer, G., and Page, E., editors, *Proceedings of the 2017 Winter Simulation Conference*, page To be published. Piscataway NJ: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. C. H. (2017b). *Non-standard parametric statistical inference*. Oxford University Press.

Cheng, R. C. H. and Currie, C. S. M. (2003). Prior and candidate models in the Bayesian analysis of finite mixtures. In Chick, S., Sanchez, P., Ferrin, D., and Maurice, D., editors, *Proceedings of the Winter Simulation Conference 2003*, pages 392 – 398.

Deler, B. and Nelson, B. L. (2001). Modeling and generating multivariate time series with arbitrary marginals and autocorrelation structures. In Peters, B. A., Smith, J. S., Medeiros, D. J., and Rohrer, M. W., editors, *Proceedings of the Winter Simulation Conference*, pages 275–282.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models 1st ed.* Springer.

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 Peters, B. A., Smith, J. S., Medeiros, D. J., and Rohrer, M. W., editors, *Proceedings of the Winter Simulation Conference*, pages 385–393.

Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. Methuen.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modelling. *Statistical Science*, 20:50–67.

Kuhl, M. E., Ivy, J. S., Lada, E. K., Steiger, N. M., Wagner, M. A., and Wilson, J. R. (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.

Nelder, J. A. 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 Medeiros, D., Watson, E., Carson, J., and Manivannan, M., editors, *Proceedings of the Winter Simulation Conference*, pages 105–112.

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.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society B*, 73:689–710.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015). Overfitting bayesian mixture models with an unknown number of components. *PLoS ONE*, 10:1–7.

Wagner, M. A. F. and Wilson, J. R. (1996a). Recent developments in input modeling with Bézier distributions. In Charnes, J. M., Morrice, D. J., Brunner, D. T., and Swain, J. J., editors, *Proceedings of the Winter Simulation Conference*, pages 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.

Appendix

1.1. Transformation of Base Densities from Standard Parameterisation

Table 5. 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. 12

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	$\frac{\alpha}{\beta}(y/\beta)^{\alpha-1}\exp\left[-(y/\beta)^{\alpha}\right]$	$\omega(\sigma/\mu)$	$\mu/\Gamma\left[1+\frac{1}{\omega(\sigma/\mu)}\right]$
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

For the 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 5 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 5, 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 given in Cheng (2017b)

$$\omega(\gamma)\simeq\exp\left(0.5282-0.7565t-0.3132\sqrt{6.179-0.5561t+0.7057t^2}\right) \qquad (12)$$

where $t=\ln(1+\gamma^2)$, which has a relative error of less than 1% in the range $0.0001\leq\gamma\leq 1000$. 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.

1.2. Alternative Representations of f_0 in Overfitted Models

Each of the PDFs, $g(y|\psi_{0j})$, $j=1,2,...,k_0$ of the original mixture of k_0 true components can be viewed as a mixture of $k'>k_0$ versions of *itself*.

$$g(y|\psi_{0j})=\sum_{i=1}^{k'}\rho_{ij}g(y|\psi_{0j}) \text{ where } \sum_{i=1}^{k'}\rho_{ij}=1,$$

only with the ith version or fragment counted as part of the *fitted* ith component. Thus the true mixture $f_0(y) = \sum_{j=1}^{k_0} w_{0j} g(y|\psi_{0j})$ has *alternative* representations

$$f_0(y) = f[y|k', \theta(k')] = \sum_{i=1}^{k'} w_i h_i(y|\mathbf{w}_0, \rho_i, \psi_0), \tag{13}$$

where $w_i = \sum_{j=1}^{k_0} w_{0j}\rho_{ij}$ and $h_i(y|\mathbf{w}_0, \rho_i, \psi_0) = w_i^{-1} \sum_{j=1}^{k_0} w_{0j}\rho_{ij} g(y|\psi_{0j})$ for $i = 1, 2, ..., k'$. This makes $\int h_i = 1$ for $i = 1, 2, ..., k'$, so in (13) the h_i are all PDFs. Thus, as $\sum_{i=1}^{k'} w_i = 1$, these alternative representations are still mixtures, only with $k' > k_0$ rather than k_0 components, but each of which is a different mixture of the original components $g(y|\psi_{0j})$, $j = 1, 2, ..., k_0$.

This characterizes non-identifiability as where the mixture representation of f_0 is not unique with different mixtures possible all precisely matching f_0 overall.

1.3. Importance Sampling: Further Details

To be unambiguous we shall write the candidate distribution specifically as $\tilde{c}_k(\cdot)$ to indicate when it has been obtained using MAP estimators in this way. In what follows some care is needed to distinguish the parameters $\theta(k)$ as they appear in the mixture PDF, the MAP estimator $\tilde{\theta}(k)$ and the parameters treated as variates generated by the importance sampling, which we shall denote by $\theta^*(k)$. A typical parameter point obtained in this way has the form

$$\theta^* = \begin{pmatrix} \psi^* \\ \mathbf{w}^* \end{pmatrix} = \begin{pmatrix} \tilde{\psi} \\ \tilde{\mathbf{w}} \end{pmatrix} + \theta_0^*, \tag{14}$$

where

$$\theta_0^* = \tilde{\mathbf{P}}_1 \tilde{\mathbf{R}} \mathbf{z}_\nu^*,$$

with $\mathbf{z}_\nu^*$ a vector of $\nu = 3k - 1$ independent Student-t variates, each normalized to have mean zero and variance unity. We do not really need the asterisk in the case of $\mathbf{z}_\nu$ but we have added it just to emphasize that it is the source of the randomness in the IS samples. (All quantities should carry a k subscript but for simplicity this is omitted.) The matrices $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{R}}$ can be calculated explicitly from the eigenvectors and eigenvalues respectively of the Hessian matrix $\tilde{H} = H(\tilde{\theta}(k), k)$ of second derivatives of $L[\theta(k), k] = \ln(p[\mathbf{y}|\theta(k), k])$ evaluated at $\tilde{\theta}(k)$.

$\mathrm{Var}(\theta_0)$ is singular as the component weights satisfy $\sum_{i=1}^{k} w_i = 1$. A simple way to remove this singularity is to reduce the dimensionality of θ by one in such a way that $\sum_{i=1}^{k} w_i = 1$ is automatically satisfied. Thus we let ω be the $(k-1)$ dimensional vector of the *reduced set of weights* formed from the first $(k-1)$ components of $\mathbf{w}$, and work with $\phi = (\psi, \omega)$ instead of θ in the importance sampling. We have

$$\phi^* = \begin{pmatrix} \psi^* \\ \omega^* \end{pmatrix} = \begin{pmatrix} \tilde{\psi} \\ \tilde{\omega} \end{pmatrix} + \tilde{\mathbf{Q}} \mathbf{z}_\nu^*$$

where $\tilde{\mathbf{Q}}$ is the matrix $\tilde{\mathbf{P}}_1 \tilde{\mathbf{R}}$ but with the last row omitted. This equation is a nonsingular linear transform of $\mathbf{z}_\nu^*$ to ϕ^*. The Jacobian of the transformation is

$|\partial[\psi,\omega]/\partial(\mathbf{z}_\nu)|_{\phi=\tilde{\phi}} = \det(\tilde{\mathbf{Q}})$ so that the PDF of ϕ^*, in the importance sampling, is

$$\tilde{f}(\phi^*) = \left|\frac{\partial(\psi,\omega)}{\partial(\mathbf{z}_\nu)}\right|^{-1}_{\phi=\tilde{\phi}} g_{\mathbf{z}_\nu}(\mathbf{z}^*_\nu) = [\det(\tilde{\mathbf{M}})]^{-1} g_\nu(\mathbf{z}^*_\nu), \tag{15}$$

where g_ν is the PDF of $\mathbf{z}_\nu$.

Use of (14) to generate IS variates does not guarantee that parameters which should be positive necessarily are positive, nor that all weights necessarily satisfy $0 < w_j < 1$. This is easily handled by rejecting any θ^* sample where *any* such constraint which should be satisfied is not, restricting 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. Given k, the IS distribution actually sampled is modified from (15) to

$$\tilde{c}^R_k[\psi(k),\omega(k)] = [\det(\tilde{\mathbf{Q}}(k))]^{-1} g_\nu(\mathbf{z}_\nu)/\tilde{R}(k) \tag{16}$$

where we have now included dependency on k explicitly, and $\tilde{R}(k)$ is an estimate of $R(k)$, the probability that a parameter point sampled from (15) is accepted (because it falls in the support of the k component form of the mixture model being fitted). A simple estimate of $R(k)$ is easily obtained from the IS sampling as

$$\tilde{R}(k) = \frac{(\#\text{ of replications sampled from (15) for the given } k \text{ and accepted})}{I_k}, \tag{17}$$

where $I_k = (\#$ of replications sampled from (15) for the given $k)$.

This gives ϕ^*, with θ^*, when needed in the IS calculations simply taken as (ϕ^*, w^*_k) with the last weight $w^*_k = 1 - \sum_{j=1}^{k-1} w^*_j$.

Let $f(\cdot)$ be the mixture PDF of eqn (1). The importance sampling procedure with sample size I is as follows.

IS1. Draw I values of $k : k_i$, $i = 1, 2, ..., I$ independently and uniformly distributed over $1, 2, ..., k_{\max}$.

IS2. Draw values $\theta^*(k_i)$ from the distribution with density $\tilde{c}_{k_i}[\phi^R(k_i)]$, as in eqn (16), for $i = 1, 2, ..., I$.

This produces a sequence of independent and identically distributed random variables $(\theta^*_i(k_i), k_i)$ $i = 1, 2, ..., I$.

For each k record the acceptance probabilities $\tilde{R}(k)$ of eqn (17).

IS3. From $(\theta^*_i(k_i), k_i)$, $i = 1, 2, ..., I$, calculate the importance sampling ratios

$$\rho[\theta^*_i(k_i), k_i] = \frac{p[\mathbf{y}|\theta^*_i(k_i), k_i]}{\tilde{c}^R_{k_i}[\phi^*_i(k_i)]} \text{ for } i = 1, 2, ..., I, \tag{18}$$

with $p[\mathbf{y}|\theta_i(k_i), k_i]$ the posterior distribution and $\tilde{c}^R_k[\psi(k),\omega(k)]$ as in eqn (16).

IS4. Estimate $\pi(k|\mathbf{y})$ by

$$\tilde{\pi}(k|\mathbf{y}) = \frac{\sum\limits_{k_i=k} \rho[\theta^*_i(k_i), k_i]}{\sum\limits_{i=1}^{I} \rho[\theta^*_i(k_i), k_i]}, \; k = 1, 2, ..., k_{\max}, \tag{19}$$

where, as both the prior for k and the importance sampling of k are uniform, we have no need to calculate the normalising integrals of the posterior distribution over the $(\psi(k), \mathbf{w}(k))$ space explicitly. As with RJMCMC, we can take the most likely k,

$$\tilde{k} = \arg\max_{k} \tilde{\pi}(k|y),$$

as the best estimate of k.