Statistics and Computing (1995) **5**, 297–305

A Bayesian predictive approach to determining the number of components in a mixture distribution

DIPAK K. DEY[1], LYNN KUO[1], and SUJIT K. SAHU[2]

[1]*Department of Statistics, University of Connecticut, Storrs, CT 06269-3120, USA and*
[2]*Statistical Laboratory, University of Cambridge, Cambridge, UK*

Received May 1994 and accepted January 1995

This paper describes a Bayesian approach to mixture modelling and a method based on predictive distribution to determine the number of components in the mixtures. The implementation is done through the use of the Gibbs sampler. The method is described through the mixtures of normal and gamma distributions. Analysis is presented in one simulated and one real data example. The Bayesian results are then compared with the likelihood approach for the two examples.

Keywords: Bootstrap procedures, conditional predictive ordinate, gamma mixtures, Gibbs sampler, likelihood ratio (LR) statistic, Metropolis algorithm, Monte Carlo methods, normal mixtures, predictive distribution, pseudo Bayes factor

1. Introduction

The use of mixture models in statistics has proved to be of considerable interest in recent years, in terms of methodological development and in their application to the most disparate fields of interest. This has led to many published papers and books by Everitt and Hand (1981), Titterington *et al.* (1985) and McLachlan and Basford (1988). The primary reason for the interest in mixture distributions is that mixture models provide an interesting alternative to non-parametric modelling, while being less restrictive than the usual distributional assumptions; located between these two extremes, they enjoy simultaneously a greater freedom and the simplicity of the parametric approach. Mixture distributions are applied in several branches of science, including actuarial sciences, agriculture, biology, economics, fisheries, genetics, medicine, and psychology.

In this paper, we consider two rich classes of finite mixture distributions, the mixture of normal distributions and the gamma distributions. Both of these classes of mixture distributions play an important role in statistical inference. We show that Bayesian methods can be easily implemented for these models. We also see that Bayes estimators associated with proper prior distributions always exist, unlike the maximum likelihood estimators. Bayes estimators under 'conjugate' prior distributions are expressible in closed form and are easily interpretable. Moreover, the performances of the Bayesian methods are usually superior to the maximum likelihood methods for small sample sizes.

The main criticism of the Bayesian approach in mixture problems is that it leads to prohibitive computation times for moderate-sized or large problems. However, we show that a sampling based approach is attractive in that implementation is easy and computing is reasonably efficient. Such simulation approaches might be non-iterative, such as standard Monte Carlo (see for example Geweke, 1989) or iterative, as for example using the Gibbs sampler or other Markov chain Monte Carlo technique (see for example Gelfand and Smith, 1990; Tierney, 1994).

The model we consider is a parametric family of finite mixture densities; i.e. a family of probability density functions of the form

$$f^{(k)}(x|\boldsymbol{\theta}) = \sum_{j=1}^{k} p_j f_j(x|\boldsymbol{\theta}_j), \tag{1}$$

where the densities f_j $(1 \leq j \leq k)$ are entirely known and parametrized by $\boldsymbol{\theta}_j$, $j = 1, \ldots, k$, the proportions

0960-3174 © 1995 Chapman & Hall

$0 < p_j < 1$ satisfy $\sum_{j=1}^{k} p_j = 1$ and k is the number of components. We denote $\boldsymbol{p} = (p_1, \ldots, p_k)$, $\boldsymbol{\theta} = (p_1, \ldots, p_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_j$ can be a vector of parameters. We observe $\boldsymbol{x} = (x_1, \ldots, x_n)$, a random sample of size n from (1).

One of the major problems in mixture models is to choose k, the number of components. In most analysis of mixture modelling, k is assumed to be known. If k is unknown, one can consider a discrete prior on possible values of k and perform a Bayesian analysis. However, it is difficult to agree on an appropriate choice of prior in this case; for example, if discrete uniform is proposed, an upper bound on k needs to be assumed.

In this paper we consider k to be unknown but not random and determine it from a predictive distribution approach. It is clear that every distinct values of k gives rise to a new statistical model. The predictive distribution suggests, for a given data, how to obtain an appropriate choice of k by selecting the best fitted model.

The use of predictive distributions in some form has long been recognized as the correct Bayesian approach to model determination. In particular, Box (1980) notes the complementary roles of the posterior and predictive distributions, arguing that the posterior is used for 'estimation of parameters conditional on the adequacy of the model' while the predictive distribution is used for 'criticism of the entertained model in light of the current data'. In comparing several models (with different number of parameters), it is clear that the predictive distributions are comparable while the posteriors are not.

Box and others have encouraged a less formal approach to Bayesian model choice, resulting in alternative predictionist criteria to the Bayes factor. Using cross-validation ideas (Stone, 1974; Geisser, 1982) gives rise to the pseudo-Bayes factor (Geisser and Eddy, 1979). This cross-validation method with its asymptotic approximations and exact calculations using Markov chain Monte Carlo approaches are explored in Gelfand and Dey (1994) and Gelfand *et al.* (1992).

In this paper we consider the cross-validation approach of model selection using the pseudo-Bayes factor and CPO (conditional predictive ordinate) to determine the number of components. In addition, we also consider exploratory data analysis approaches using graphical displays. The cross-validation approach has the advantage of being able to incorporate improper priors whereas the Bayes factor approach may not always work because marginal densities may not exist or may be arbitrary under improper prior specifications.

There are many articles on Bayesian analysis for mixture distributions (see, for example Titterington *et al.* 1985), and here we mention only a few recent papers. Diebolt and Robert (1994) study sampling-based approaches to approximating the Bayes estimates for finite mixtures of normal distributions assuming the number of components k is known. Crawford (1991) proposes a modification of the Laplace method to estimate the Bayes estimators. In addition, she considers the problem of estimating the number of components k, where a Bayesian formulation treats k as an unknown parameter with a given prior distribution. West (1992) proposes an adaptive method for estimating the posterior distribution and mixture pruning methods, useful for reducing the number of components of large mixtures. Evans *et al.* (1992) consider Bayesian inference for a mixture of two normal distributions. They derive a prior by invariance consideration. They propose importance sampling and Gibbs sampling algorithms to compute the Bayes estimates.

The problem of determining the number of components can also be studied using the likelihood ratio (LR) statistic. The difficulty of this approach is that the regularity conditions do not hold. The appropriate LR statistic for mixture models will fail to have its usual null distribution of chi-squared. However, many authors have recognized the power of bootstrap procedures to overcome this difficulty, see for example McLachlan (1987). As described there, the LR statistic for testing the null hypothesis of $k = k_1$ versus the alternative $k = k_2$ can be bootstrapped as follows. First, find the maximum likelihood estimate, $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$ from the given data under the assumption of $k = k_1$ components. This maximization can be done using the E-M algorithm. Then generate a bootstrap sample from $f^{(k_1)}(x|\hat{\boldsymbol{\theta}})$. Now, the mixture models with $k = k_1$ and k_2 can be fitted to this sample and the LR statistic can be computed. This method may be repeated to simulate the null distribution and subsequently an approximate LR test can be carried out.

The outline of the paper is thus the following. In Section 2, we develop Bayesian formulation of the mixture models and the conditional distributions needed for the Gibbs sampler. Section 3 is devoted to the development of predictive distribution using the cross-validation approach and its Monte Carlo estimates. In Section 4, we apply our method for the normal mixtures. A simulated example is provided for the determination of the number of mixtures. In Section 5, we describe the Gibbs sampling procedure for the gamma mixtures. In this section we consider Halley's mortality data (see Nelson (1982), p. 17) and model it by mixture of gamma distributions. The hazard function for this data is bathtub shaped and our method suggests the need for mixtures of three gamma distributions. For both the examples we compare our Bayesian results with those of approximate LR tests performed by using the bootstrap procedure to simulate from the null distribution. We use the E-M algorithm to carry out the maximization of the mixture likelihood needed to perform the LR tests in both the examples. We conclude with some summarizing remarks in Section 6.

2. The Bayesian formulation of mixture models

Dempster *et al.* (1977) pointed out that a mixture model can always be represented in terms of incomplete data. Suppose $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, is a k-dimensional vector of indicator variables, i.e. $z_{ij} = 1$ if x_i is the jth group and 0 otherwise with $\sum_{j=1}^{k} z_{ij} = 1$, then the density of the complete data (x_i, z_i) is given as

$$\prod_{j=1}^{k} p_j^{z_{ij}} f_j^{z_{ij}}(x_i|\boldsymbol{\theta}_j), \qquad i = 1, \ldots, n. \tag{2}$$

In Bayesian framework, z_{ij} can be treated as parameters. This makes a difference from a classical point of view. The advantage of using z_{ij} as a parameter is that it allows us to specify conditional distributions needed in the Gibbs sampler. The only disadvantage is the computational burden. However, with the use of the recently developed Markov chain Monte Carlo approach, e.g. the Gibbs sampler and the Metropolis algorithm, the computational steps are straightforward to implement.

Let us now describe our computational steps. We use the Gibbs sampling scheme to generate the samples from the posterior distribution and then use the Monte Carlo approach to estimate the predictive density. The Gibbs sampling is a Monte Carlo integration method which proceeds by Markovian updating scheme. It was developed by Geman and Geman (1984) in the context of image restoration. More recently, Gelfand and Smith (1990) showed applicability to general parametric Bayesian computations. The literature on Gibbs sampling is now vast, and here we only describe the conditional distributions needed to draw samples. We can decompose the parameter $\boldsymbol{\theta}$ into s components, $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s)$, where $\boldsymbol{u}_i$, $i = 1, \ldots, s$, can be a vector of parameters. Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$. Now at step t, we draw samples as follows:

$$\begin{aligned}
\boldsymbol{Z}^{(t+1)} &\sim f(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{u}^{(t)}) \\
\boldsymbol{u}_1^{(t+1)} &\sim f(\boldsymbol{u}_1|\boldsymbol{x}, \boldsymbol{Z}^{(t+1)}, \boldsymbol{u}_2^{(t)}, \ldots, \boldsymbol{u}_s^{(t)}) \\
\boldsymbol{u}_2^{(t+1)} &\sim f(\boldsymbol{u}_2|\boldsymbol{x}, \boldsymbol{Z}^{(t+1)}, \boldsymbol{u}_1^{(t+1)}, \boldsymbol{u}_3^{(t)}, \ldots, \boldsymbol{u}_s^{(t)}) \\
&\vdots \quad \vdots \\
\boldsymbol{u}_s^{(t+1)} &\sim f(\boldsymbol{u}_s|x, \boldsymbol{Z}^{(t+1)}, \boldsymbol{u}_1^{(t+1)}, \ldots, \boldsymbol{u}_{s-1}^{(t+1)}).
\end{aligned}$$

We repeat this process and after T such iterations arrive at $(\boldsymbol{Z}^{(T)}, \boldsymbol{u}_1^{(T)}, \ldots, \boldsymbol{u}_s^{(T)})$. We run B independent parallel replications and obtain $(\boldsymbol{Z}^{(T)j}, \boldsymbol{u}_1^{(T)j}, \ldots, \boldsymbol{u}_s^{(T)j})$, $j = 1, \ldots, B$, i.i.d. (approximately) samples of size B. The resulting sample can lead to an approximation of any well-defined posterior quantity by the ergodic theorem.

3. Predictive distributions and Monte Carlo estimates

As mentioned before, we consider a cross-validation approach for the predictive distribution. Defining $\boldsymbol{x}_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, x_n)$, it follows from Gelfand *et al.* (1992) that the conditional predictive density of $X_i|\boldsymbol{x}_{(i)}$ is

$$\begin{aligned}
f^{(k)}(x_i|\boldsymbol{x}_{(i)}) &= \int f^{(k)}(x_i|\boldsymbol{\theta}, \boldsymbol{x}_{(i)}) \pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{x}_{(i)})\, \mathrm{d}\boldsymbol{\theta} \\
&= \int f^{(k)}(x_i|\boldsymbol{\theta}) \pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{x}_{(i)})\, \mathrm{d}\boldsymbol{\theta},
\end{aligned} \tag{3}$$

where $\pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{x}_{(i)})$ is the posterior distribution of $\boldsymbol{\theta} = (p_1, \ldots, p_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ given $\boldsymbol{x}_{(i)}$ under the prior $\pi^{(k)}(\boldsymbol{\theta})$. We drop the conditioning variable $\boldsymbol{x}_{(i)}$ in $f^{(k)}(x_i|\boldsymbol{\theta})$ because of conditional independence. The quantity $f^{(k)}(x_i|\boldsymbol{x}_{(i)})$ is known as the conditional predictive ordinate (CPO). After obtaining $f^{(k)}(x_i|\boldsymbol{x}_{(i)})$, we define the pseudo-predictive likelihood as

$$D_k = \prod_{i=1}^{n} f^{(k)}(x_i|\boldsymbol{x}_{(i)}) \tag{4}$$

and choose k which maximizes D_k. To compare between two models, say model 1 versus model 2, we define the pseudo-Bayes factor as the ratio of the pseudo predictive likelihoods, denote it as PB_{12} and use Jeffreys' (1961), scale of evidence. Since our procedure is more exploratory in nature, we consider several graphical plots in addition to the single summary measures like D_k and the pseudo-Bayes factor. The CPO plots describe for each observation, how much it supports the model. A second useful diagnostic display plots, for $i = 1, \ldots, n$, the pairs $|x_i - \hat{\boldsymbol{E}}^{(k)}(X_i|\boldsymbol{x}_{(i)})|$ versus $\sqrt{\hat{\boldsymbol{V}}^{(k)}(X_i|\boldsymbol{x}_{(i)})}$, where $\hat{\boldsymbol{V}}^{(k)}$ is the estimated conditional variance. A good model should reveal a point cloud near the origin. An underfitted model will tend to present large abscissas or ordinates. An overfitted model will tend to give small abscissas but large ordinates. A parsimonious choice should perform well on both axes.

Now we describe how arbitrary accurate estimates of (3) and (4) can be obtained using the Markov chain Monte Carlo technique. It follows from (3) that

$$f^{(k)}(x_i|\boldsymbol{x}_{(i)}) = \frac{\int f^{(k)}(\boldsymbol{x}|\boldsymbol{\theta}) \pi^{(k)}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}{\int f^{(k)}(\boldsymbol{x}_{(i)}|\boldsymbol{\theta}) \pi^{(k)}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}.$$

Suppose $g(\boldsymbol{\theta})$ is an importance sampling density for $f^{(k)}(\boldsymbol{x}_{(i)}|\boldsymbol{\theta})\pi^{(k)}(\boldsymbol{\theta})$ and $\{\boldsymbol{\theta}^j\}$, $j = 1, \ldots, B$ is a sample from $g(\boldsymbol{\theta})$, where B is the number of simulations. Defining $w_j = f^{(k)}(\boldsymbol{x}_{(i)}|\boldsymbol{\theta}^j)\pi^{(k)}(\boldsymbol{\theta}^j)/g(\boldsymbol{\theta}^j)$, it follows that a Monte Carlo integration for (3) is

$$\hat{f}^{(k)}(x_i|\boldsymbol{x}_{(i)}) = \sum_{j=1}^{B} f^{(k)}(x_i|\boldsymbol{\theta}^j) w_j \Big/ \sum_{j=1}^{B} w_j. \tag{5}$$

If a Markov chain Monte Carlo technique has been used, the output is usually taken to be a sample $\boldsymbol{\theta}^j$, $j = 1, \ldots, B$,

from the posterior $\pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{x})$. In that case we can take the posterior as the importance sampling density in (5) which gives rise to

$$w_j = \frac{f^{(k)}(\boldsymbol{x}_{(i)}|\boldsymbol{\theta}^j)\pi^{(k)}(\boldsymbol{\theta}^j)}{f^{(k)}(\boldsymbol{x}|\boldsymbol{\theta}^j)\pi^{(k)}(\boldsymbol{\theta}^j)/f^{(k)}(\boldsymbol{x})} = \left[\frac{f^{(k)}(x_i|\boldsymbol{\theta}^j)}{f^{(k)}(\boldsymbol{x})}\right]^{-1},$$

where $f^{(k)}(\boldsymbol{x}) = \int f^{(k)}(\boldsymbol{x}|\boldsymbol{\theta})\pi^{(k)}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the marginal density. Hence the Monte Carlo estimate of the CPO reduces to

$$\hat{f}^{(k)}(x_i|\boldsymbol{x}_{(i)}) = B\left[\sum_{j=1}^{B}(f^{(k)}(x_i|\boldsymbol{\theta}^j))^{-1}\right]^{-1}, \tag{6}$$

which is the harmonic mean of the conditional density of X_i given $\boldsymbol{\theta}$ evaluated at the posterior sample values.

To obtain other diagnostic plots we need to calculate the Monte Carlo estimates of $\boldsymbol{E}^{(k)}(X_i|\boldsymbol{x}_{(i)})$ and $\mathrm{Var}^{(k)}(X_i|\boldsymbol{x}_{(i)})$. It follows that

$$\begin{aligned}\boldsymbol{E}^{(k)}(X_i|\boldsymbol{x}_{(i)}) &= \int \boldsymbol{E}^{(k)}(X_i|\boldsymbol{\theta})\pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{x}_{(i)})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{f^{(k)}(\boldsymbol{x})}{f^{(k)}(\boldsymbol{x}_{(i)})}\int \boldsymbol{E}^{(k)}(X_i|\boldsymbol{\theta})\frac{\pi^{(k)}(\boldsymbol{\theta}|x)}{f^{(k)}(x_i|\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta},\end{aligned}$$

hence a Monte Carlo estimate is given as

$$\hat{\boldsymbol{E}}^{(k)}(X_i|\boldsymbol{x}_{(i)}) = \hat{f}^{(k)}(x_i|\boldsymbol{x}_{(i)})B^{-1}\sum_{j=1}^{B}\frac{\boldsymbol{E}^{(k)}(X_i|\boldsymbol{\theta}^j)}{f^{(k)}(x_i|\boldsymbol{\theta}^j)}. \tag{7}$$

Similar argument produces a Monte Carlo estimate of the conditional variance as

$$\hat{\boldsymbol{V}}^{(k)}(X_i|\boldsymbol{x}_{(i)}) = \hat{f}^{(k)}(x_i|\boldsymbol{x}_{(i)})B^{-1}\sum_{j=1}^{B}\frac{\mathrm{Var}^{(k)}(X_i|\boldsymbol{\theta}^j)}{f^{(k)}(x_i|\boldsymbol{\theta}^j)}. \tag{8}$$

The quantities $\boldsymbol{E}^{(k)}(X_i|\boldsymbol{\theta})$ and $\mathrm{Var}^{(k)}(X_i|\boldsymbol{\theta})$ are expressible in closed form for the mixture problem. All the Monte Carlo estimates mentioned above are simulation consistent.

4. The normal mixtures

Here we consider the mixture of normal distributions. We assume

$$x_i \sim f^{(k)}(x|\boldsymbol{\theta}) = \sum_{j=1}^{k} p_j f_j(x_i|\boldsymbol{\theta}_j), \qquad i = 1,\ldots,n,$$

$$0 < p_j < 1, \quad \sum_{j=1}^{k} p_j = 1$$

and f_j is a normal density with parameter $\boldsymbol{\theta}_j = (\xi_j, \sigma_j^2)$, $j = 1,\ldots,k$.

Further we assume that the prior information can be modelled through a conjugate prior on the unknown parameters. In practice the specification of hyperparameters may be difficult, so we take the values of hyperparameters in such a way that we get non-informative priors in the limiting case. We assume $\boldsymbol{p}$ follows a Dirichlet distribution $\mathscr{D}(\alpha_1,\ldots,\alpha_k)$, ξ_j given σ_j^2 follows a $\mathrm{N}(\mu_j, \sigma_j^2/n_j)$ distribution and σ_j has a modified inverse gamma distribution with density $f(\sigma_j) \propto \sigma_j^{-\gamma_j - 1}\exp(-s_j^2/2\sigma_j^2)$, i.e. $\sigma_j \sim \mathrm{IG}(\gamma_j, s_j^2)$. All the hyperparameters are assumed to be known. Although we choose all the priors to be proper, our analysis would work for improper prior on any of the parameters, e.g. $f(\sigma_j) = \sigma_j^{-1}$. The above prior specification may be slightly unrealistic, see Berger (1985), but we work with these values for ease of computations.

4.1. *Steps for Gibbs sampler*

The above conjugate prior structure leads to the following Gibbs steps which are easy to implement.

Step I. Generate $\boldsymbol{Z}_i = (z_{i1},\ldots,z_{ik}) \sim \mathrm{MN}(1; q_{i1},\ldots,q_{ik})$ (multinomial) independently for each i, $i = 1,\ldots,n$, where

$$q_{ij} = \frac{\frac{p_j}{\sigma_j}\exp\{-(x_i-\xi_j)^2/2\sigma_j^2\}}{\sum_{j=1}^{k}\frac{p_j}{\sigma_j}\exp\{-(x_i-\xi_j)^2/2\sigma_j^2\}}$$

for $j = 1,\ldots,k$.

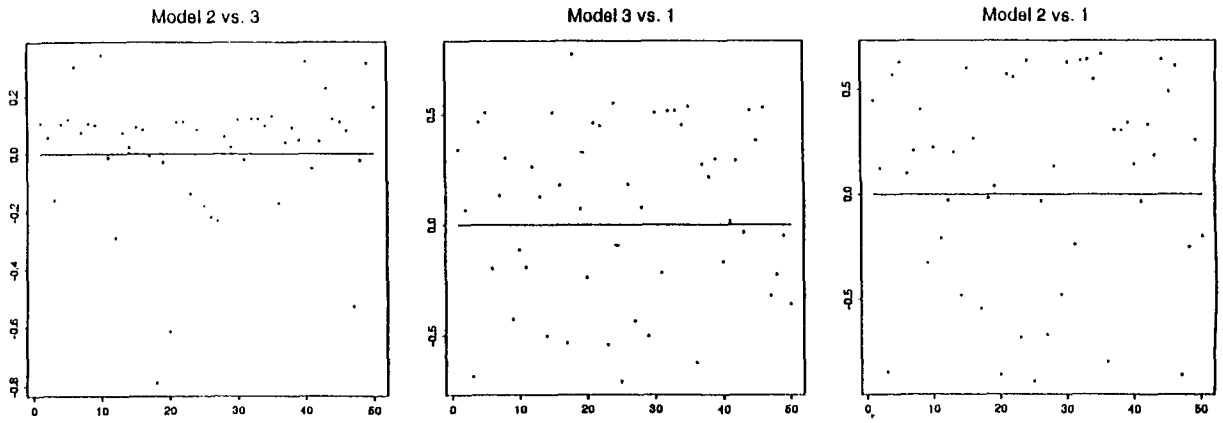

Fig. 1. *Log CPO ratios for normal mixture models*

Step II. Generate

$$\boldsymbol{p} \sim \mathcal{D}(\alpha_1 + z_{.1}, \ldots, \alpha_k + z_{.k}),$$

$$\text{where } z_{.j} = \sum_{i=1}^{n} z_{ij}, \quad j = 1, \ldots k.$$

Step III. Generate $\xi_j \sim \mathrm{N}(\delta_j, \tau_j^2)$ independently for each $j, j = 1, \ldots, k$, where

$$\delta_j = \frac{n_j \mu_j + \sum_{i=1}^{n} x_i z_{ij}}{n_j + z_{.j}} \quad \text{and} \quad \tau_j^2 = \frac{\sigma_j^2}{n_j + z_{.j}}, \qquad j = 1, \ldots, k.$$

Step IV. Generate $\sigma_j \sim \mathrm{IG}(z_{.j} + \gamma_j + 1, \beta_j)$ independently for each $j, j = 1, \ldots, k$, where

$$\beta_j = \sum_{i=1}^{n} z_{ij}(x_i - \mu_j)^2 + n_j(\xi_j - \mu_j)^2 + s_j^2.$$

4.2. *Simulation for the normal mixtures*

First we simulated 50 observations from $0.36 \times \mathrm{N}(111, 529) + 0.64 \times \mathrm{N}(190, 324)$ as given in Diebolt and Robert (1991). As mentioned earlier we set $\alpha_j = 0.5$, $s_j^2 = 0.000002$, $\gamma_j = 3.0$ for all j to have an approximate non-informative prior selection. We choose $\alpha_j = 0.5$ by considering Jeffreys' prior selection method. Also we set $n_1 = 3$ for $k = 1$; $n_1 = 3$, $n_2 = 2$ for $k = 2$ and $n_1 = 3$, $n_2 = 2$, $n_3 = 5$ for $k = 3$.

We took the starting values for different parameters to be very dispersed around the true values. We have used 500 parallel chains and considered the criterion of Gelman and Rubin (1992) to detect convergence. For $k = 1$ and 2, the Gelman–Rubin scale reduction factor came down to 1 within 50 iterations; however, for $k = 3$, the scale reduction factor was slightly larger than 1 even after 2000 iterations. Then we considered the log of the density for a few parallel chains and following Gelman and Rubin we detected convergence within 500 iterations.

In Fig. 1 we plot log of the CPO ratios for different models with respect to others against the observation number. Positive values of log CPO ratios indicate the preference of the first model with respect to the other. For example, in Fig. 1, the model 2 vs. 3 plot indicates 34 out of 50 observations support model 2 over model 3. Similarly, 30 observations support model 3 over model 1, and 31 observations support model 2 over model 1. Thus, in conclusion, this criterion says that model 2 is the best.

The same conclusion follows from Figs 2(a) and 2(b), i.e. model 2 improves upon model 3 and both improve upon model 1 substantially. In terms of single summary measure our calculation shows that $\log_{10} D_1 = -114.62$, $\log_{10} D_2 = -112.85$, and $\log_{10} D_3 = -113.13$. In terms of the log of the pseudo-Bayes factor we obtain $\log_{10} PB_{23} = 0.28$, $\log_{10} PB_{31} = 1.49$ and $\log_{10} PB_{21} = 1.77$.

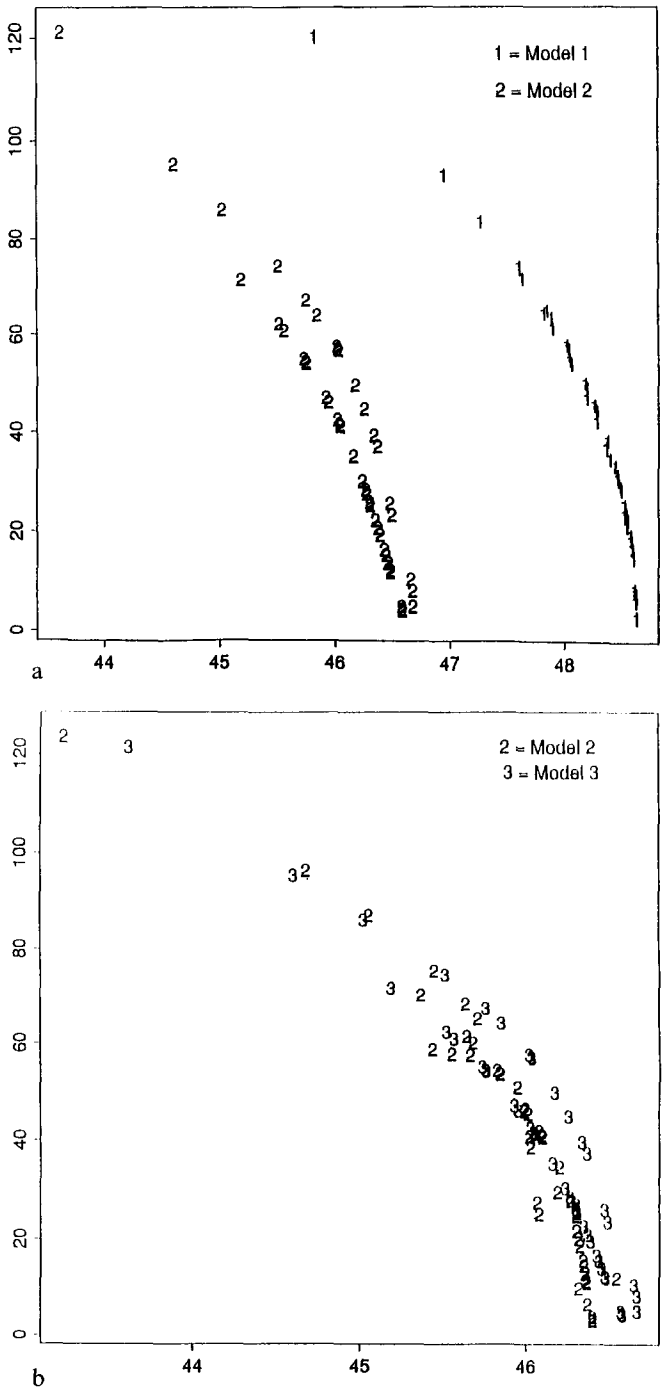

Fig. 2. *(a) Plot of* $|X_{i,\mathrm{obs}} - E(X_i|x_{(i),\mathrm{obs}})|$ *versus* $\sqrt{\mathrm{var}(X_i|x_{(i),\mathrm{obs}})}$ *for normal mixture models. (b) Plot of* $|X_{i,\mathrm{obs}} - E(x_{(i),\mathrm{obs}})|$ *versus* $\sqrt{\mathrm{Var}(X_i|x_{(i)})}$ *for normal mixture models*

Now using Jeffrey's scale of evidence, it follows that model 2 is best.

We compare the above results with the likelihood approach. We simulate the null distribution of the log likelihood ratio statistic for comparing two models using 500 bootstrap replicates. Let L_{ij} denote the LR statistic for testing a model with i components versus one with j components, $i, j = 1, 2, 3$, $i < j$. The three null distributions are given in Fig. 3. The observed values of the LR test statistics

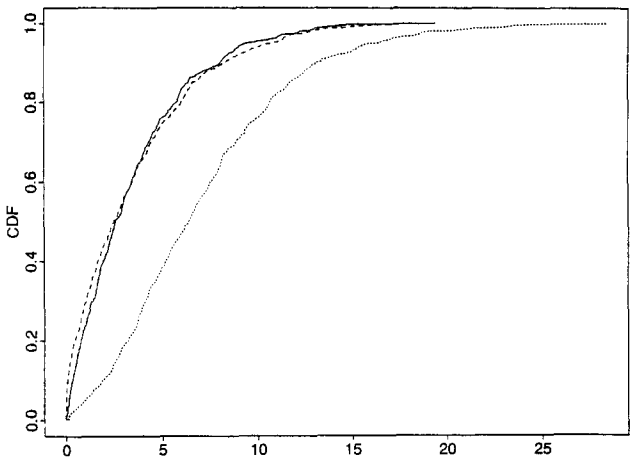

Fig. 3. *Bootstrapped null distributions for the simulated normal data example. Solid line for L_{12}; dotted line for L_{13} and dashed line for L_{23}*

are 17.05, 18.8 and 1.75 for L_{12}, L_{13} and L_{23} respectively with corresponding approximate P values 0.0001, 0.022 and 0.59. Hence, this method suggests a two-component mixture is a best fit for the data, which is in agreement with the Bayesian approach.

5. The gamma mixtures

Here we assume that

$$x_i \sim f^{(k)}(x|\boldsymbol{\theta}) = \sum_{j=1}^{k} p_j f_j(x_i|\boldsymbol{\theta}_j), \qquad i = 1, \ldots, n,$$

$$0 < p_j < 1, \quad \sum_{j=1}^{k} p_j = 1$$

where $f_j \sim \mathrm{G}(a_j, b_j)$, (gamma density) with mean a_j/b_j, $j = 1, \ldots, k$.

We assume 'conjugate' prior in the sense of George *et al.* (1993) that is closed under log-concavity for the different parameters in the model. Again we assume $\boldsymbol{p} \sim \mathscr{D}(\gamma_1, \ldots, \gamma_k)$, $a_j \sim \mathrm{G}(1, \lambda_j)$, $b_j \sim \mathrm{G}(\alpha_j, \beta_j)$ and all prior distributions are independent.

5.1. *Steps for Gibbs sampler*

For the above model we have the following Gibbs steps.

Step I. Generate $\boldsymbol{Z} = (z_{i1}, \ldots, z_{ik}) \sim \mathrm{MN}(1; q_{i1}, \ldots, q_{ik})$ independently for each i, $i = 1, \ldots, n$, where

$$q_{ij} = \frac{p_j \dfrac{b_j^{a_j}}{\Gamma(a_j)} x_i^{a_j - 1} \exp\{-b_j x_i\}}{\displaystyle\sum_{j=1}^{k} p_j \frac{b_j^{a_j}}{\Gamma(a_j)} x_i^{a_j - 1} \exp\{-b_j x_i\}}$$

for each $j = 1, \ldots, k$.

Step II. Generate

$$\boldsymbol{p} \sim \mathscr{D}(\gamma_1 + z_{\cdot 1}, \ldots, \gamma_k + z_{\cdot k}).$$

Step III. Generate

$$b_j \sim \mathrm{G}\left(\alpha_j + a_j z_{\cdot j}, \beta_j + \sum_{i=1}^{n} x_i z_{ij}\right)$$

independently for each j, $j = 1, \ldots, k$.

Step IV. Generate a_j independently for each j, $j = 1, \ldots, k$, where

$$f(a_j) \propto \frac{b_j^{a_j z_{\cdot j}} \left(\displaystyle\prod_{i=1}^{n} x_i^{z_{ij}}\right)^{a_j} \exp\{-\lambda_j a_j\}}{\{\Gamma(a_j)\}^{z_{\cdot j}}}.$$

Note that the density in step IV is not a standard one. Therefore to sample from this distribution we ran one trajectory of the random increment Metropolis algorithm. We started the algorithm at the mode and ran 50 iterations to get our sample. Alternatively, we note that the above distribution is log-concave (see George *et al.*, 1993). So the

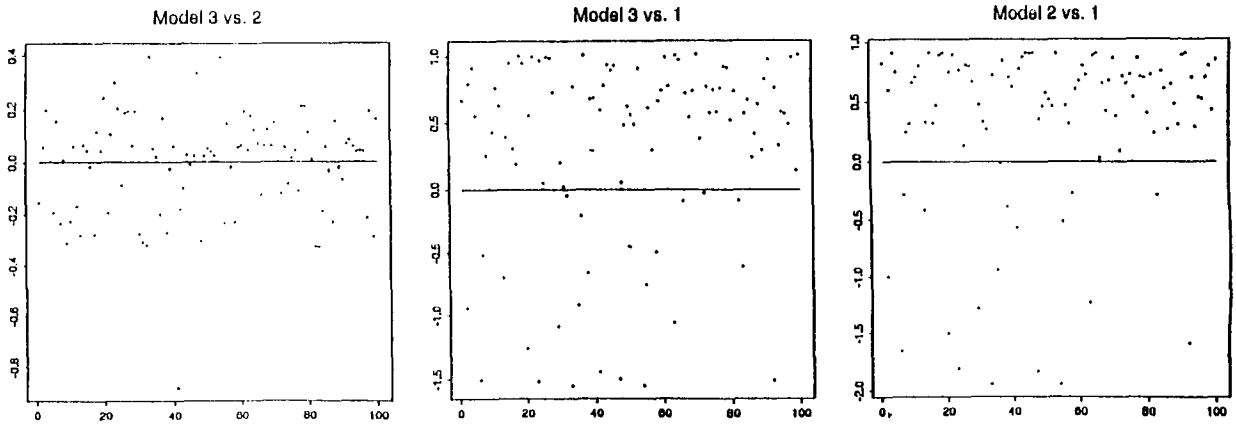

Fig. 4. *Log CPO ratios for gamma mixture models*

adaptive rejection sampling of Gilks and Wild (1992) can be considered.

5.2. *Analysis of Halley's data using gamma mixtures*

For Halley's mortality table as described in Nelson (1982, p. 17), Jaisingh *et al.* (1987) proposed a bathtub hazard model and fitted the model using the maximum likelihood

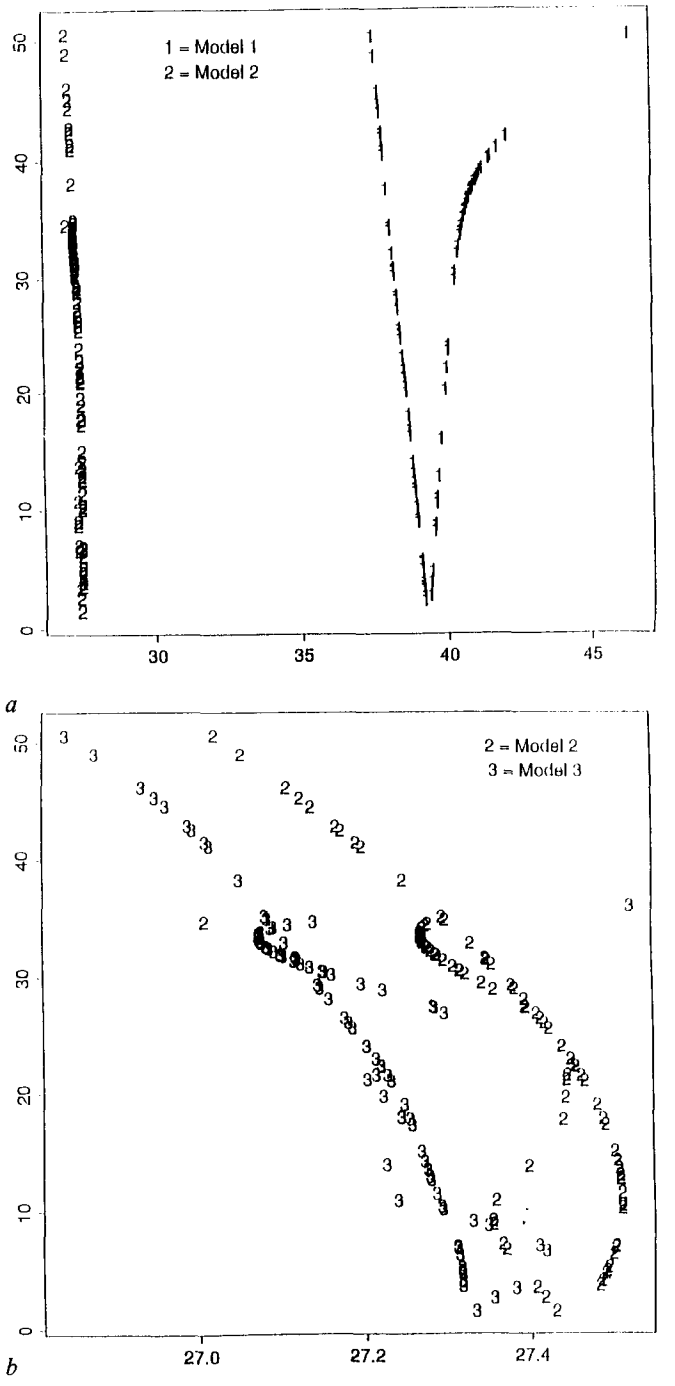

Fig. 5. *(a) Plot of* $|X_{i,\text{obs}} - E(X_i|x_{(i),\text{obs}})|$ *versus* $\sqrt{\text{Var}(X_i|x_{(i),\text{obs}})}$ *for gamma mixture models. (b) Plot of* $|X_{i,\text{obs}} - E(X_i|x_{(i),\text{obs}})|$ *versus* $\sqrt{\text{Var}(X_i|x_{(i),\text{obs}})}$ *for gamma mixture models*

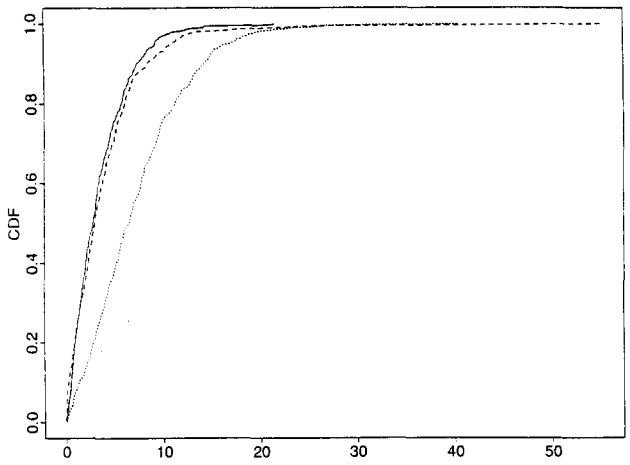

Fig. 6. *Bootstrapped null distributions for Halley's data example. Solid line for* L_{12}*; dotted line for* L_{13} *and dashed line for* L_{23}

estimates. The hazard plot for this data indicates its bathtub nature. From the density and the hazard plot it is anticipated that the density function can be approximated by appropriate gamma mixtures. The data was given in frequency counts where the number of deaths was given in categories of increment of 5 years. We simulated 100 observations from the table using a simple random sampling scheme and subsequently drawing one observation uniformly in the selected category.

We use the following values of the hyperparameters $\gamma_j = 0.5$, $\lambda_j = 0.00001$, $\alpha_j = 1.0$, $\beta_j = 0.00002$ for all j in the gamma mixture model. Note that these correspond to a non-informative prior in the limiting case.

We took the starting values for different parameters very dispersed and used 500 parallel chains. As in Section 4.2 we used the Gelman–Rubin criterion to detect convergence and in all three cases convergence was achieved within 500 iterations.

As in Section 4.2, in Fig. 4 we plot the log of the CPO ratios for different models with respect to others against the observation number. In Fig. 4, the model 3 versus model 2 plot indicates that 62 out of 100 observations support model 3 over model 2. Similarly, 77 observations support model 3 over model 1, and 81 observations support model 2 over 1. So here this criterion says that model 3 is the best fit for the data.

The above conclusion is also supported by Figs 5a and 5b. In terms of single summary measure our calculation shows the $\log_{10} D_1 = -200.5$, $\log_{10} D_2 = -186.98$, and $\log_{10} D_3 = -187.47$, and in terms of the pseudo-Bayes factors $\log_{10} PB_{32} = -0.49$, $\log_{10} PB_{31} = 13.03$ and $\log_{10} PB_{21} = 13.52$. Again, using Jeffreys' scale of evidence it suggests that both models 2 and 3 are much better than model 1 and there is slight evidence against model 3 with respect to model 2.

Here also we compare our results with the likelihood

approach. We simulate 500 replications of the three LR statistics, L_{12}, L_{13} and L_{23}. The null distributions so constructed are given in Fig. 6. The observed values of the test statistics are 42.39, 47.57 and 5.18 with corresponding approximate P values 0, 0 and 0.26. This method chooses the two-component mixture model as the best fit for the data. So the two approaches give different results for this example.

6. Concluding remarks

Our efforts here have focused on modelling data through mixtures of several distributions and the determination of the number of mixture components. We have demonstrated our procedures with normal and gamma mixtures. However, our methods extend to other mixtures, such as mixtures of beta distributions (see Gelfand *et al.*, 1995), mixtures of exponential distributions and mixtures of lognormal distributions. The results can also be extended to multivariate problems. For example, in the normal mixture problem the normal-gamma prior is replaced by the multivariate normal-Wishart prior. When observations are d-variate, display of the entire predictive distribution is not feasible; however, predictive distributions of one variable, or joint predictive distributions of two variables, conditional on values of the remaining variables can be produced.

Another generalization of the mixture models is to deal with dependent data. One way to introduce dependence in the model consists in allowing the probability p_j that an observation comes from the jth group to be a function of the preceding observations, which could be modelled through a Markov chain. This will be pursued elsewhere.

Acknowledgements

The authors thank A. E. Gelfand and B. K. Mallick for valuable discussions and an anonymous referee for comments and suggestions.

References

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer-Verlag, New York.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society*, B, **143**, 383–430.

Crawford, S. L. (1994) An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, **89**, 259–67.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

Diebolt, J. and Robert, C. (1991) Bayesian estimation of finite mixture distributions part II: sampling implementation. Unpublished Report, L. S. T. A., Université Paris VI.

Diebolt, J. and Robert, C. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*, Series B, **56**, 363–75.

Evans, M., Guttman, I. and Olkin, I. (1992) Numerical aspects in estimating the parameters of a mixture of normal distributions. *Journal of Computational and Graphical Statistics*, **1**, 351–65.

Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions*. Chapman and Hall, New York.

Geisser, S. (1982) Aspects of the predictive and estimative approaches in the determination of probabilities. *Biometrics Supplement*, **38**, 75–85.

Geisser, S. and Eddy, W. (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–60.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, Series B, **56**, 501–14.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions, with implementation via sampling-based methods. In *Bayesian Statistics* 4, J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), pp. 147–69. Oxford University Press, Oxford.

Gelfand, A. E., Mallick, B. K. and Dey, D. K. (1995) Modeling expert opinion arising as a partial probabilistic specification. *Journal of the American Statistical Association*, **90**, 398–409.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.

George, E. I., Makov, U. E. and Smith, A. F. M. (1993) Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, **20**, 147–56.

Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317–39.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–48.

Jaisingh, L. R., Kolarik, W. J. and Dey, D. K. (1987) A flexible bathtub hazard model for non-repairable systems with uncensored data. *Microelectronics and Reliability*, **27**, 87–103.

Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. Oxford University Press.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, **36**, 318–24.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

Nelson, W. (1982) *Applied Life Data Analysis*. John Wiley, New York.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, Series **B**, **36**, 111–47.

Tierney, L. (1995) Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701–1721.

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

West, M. (1992) Modeling with mixtures. In *Bayesian Statistics* 4, J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), pp. 503–24. Oxford University Press.