Mathematical Geology, Vol. 38, No. 3, April 2006 (© 2006)
DOI: 10.1007/s11004-005-9019-3

Bayesian Mixture Modelling in Geochronology via Markov Chain Monte Carlo[1]

Ajay Jasra,[2] David A. Stephens,[2] Kerry Gallagher,[3] and Christopher C. Holmes[4,5]

In this paper we develop a generalized statistical methodology for characterizing geochronological data, represented by a distribution of single mineral ages. The main characteristics of such data are the heterogeneity and error associated with its collection. The former property means that mixture models are often appropriate for their analysis, in order to identify discrete age components in the overall distribution. We demonstrate that current methods (e.g., Sambridge and Compston, 1994) for analyzing such problems are not always suitable due to the restriction of the class of component densities that may be fitted to the data. This is of importance, when modelling geochronological data, as it is often the case that skewed and heavy tailed distributions will fit the data well. We concentrate on developing (Bayesian) mixture models with flexibility in the class of component densities, using Markov chain Monte Carlo (MCMC) methods to fit the models. Our method allows us to use any component density to fit the data, as well as returning a probability distribution for the number of components. Furthermore, rather than dealing with the observed ages, as in previous approaches, we make the inferences of components from the "true" ages, i.e., the ages had we been able to observe them without measurement error. We demonstrate our approach on two data sets: uranium-lead (U-Pb) zircon ages from the Khorat basin of northern Thailand and the Carrickalinga Head formation of southern Australia.

KEY WORDS: Bayesian statistics; mixture modelling; reversible jump Markov chain Monte Carlo; geochronology.

INTRODUCTION

Radiometric dating of individual crystals or grains in rocks is an important procedure in geology, and the subdiscipline is known as geochronology. It relies on

[1]Received 1 December 2004; accepted 14 October 2005; Published online: 27 May 2006.
[2]Department of Mathematics, Imperial College London; e-mail: ajay.jasra@imperial.ac.uk, d.stephens@imperial.ac.uk.
[3]Department of Earth Science and Engineering, Imperial College London, South Kensington, London SW7 2AS, United Kingdom; e-mail: kerry@imperial.ac.uk.
[4]Oxford Centre for Gene Function, Department of Statistics, University of Oxford, United Kingdom; e-mail: cholmes@stats.ox.ac.uk.
[5]Mammalian Genetics Unit, MRC Harwell, United Kingdom; e-mail: cholmes@stats.ox.ac.uk

the measurement of the abundance of a parent isotope (I), and its decay product or daughter isotope (D) within a mineral or rock. As the decay rate (η) is effectively constant under geological conditions, the ratio of the daughter to the parent is indicative of the age (x). The ages are given by the following (generalized) equation

$$x = \frac{1}{\eta} \log\left(1 + \frac{D}{I}\right) \tag{1}$$

The raw data are typically counts of the number of ions of a given mass (i.e., an isotope) over a given time. The mean counts are estimated for each isotope and the ratio (e.g., D/I as in Equation (1)) is then calculated. This ratio may then be compared to a natural sample of a known age, i.e., a calibration standard. There is measurable error involved in this procedure and it is estimated by using the combination of the unknown sample and standard sample errors.

One of the most common dating methods is U-Pb dating of the mineral zircon, where ^{238}U decays to ^{206}Pb, with $\eta = 1.55 \times 10^{-10}$ year^{-1}, and ^{235}U decays to ^{207}Pb, with $\eta = 9.85 \times 10^{-10}$ year^{-1}. This method is widely used because the two decay schemes have half-lives similar to the age of the Earth and also is considered relatively robust to geological perturbations, where the parent or daughter may be preferentially lost or gained from the zircon during the geological history of the host rock.

In practice, problems can arise due to sample and standard heterogeneity, instrumental drift, and some gain or loss of U and/or Pb. However, the uncertainty in the measurement of the Pb isotopes is dominated by the counting statistics, and typical one standard deviation measurement errors are 1–3% (Stern and Amelin, 2003). Normally, many individual zircon grains will be analyzed from one rock sample and the aim is to identify either the oldest or youngest grains, or to characterize the distribution of different age components. The former situation is important when ages vary due to some physical process such as thermally activated diffusion which may cause preferential loss of either the daughter or parent isotope, leading to anomalously young or old ages, respectively. The latter situation, which we concentrate upon in this paper, is particularly useful when dealing with sedimentary rocks, which are derived from the erosion of preexisting rocks (known as source rocks), and so may inherit the age signature of the different source regions, known as detrital ages. The extraction of these detrital ages is a problem which is suitably dealt with using mixture modelling (see McLachlan and Peel (2000) for an introduction).

Current statistical methodology for mixture modelling of geochronological data, e.g., Galbraith and Green (1990), Brandon (1992), and Sambridge and Compston (1994), focuses on the estimated age x_i for sample $i = 1, \ldots, n$ and the associated error (ϵ_i, assumed to be constant and known for each datum) de-

scribed above. A mixture model is then formulated for the ages, with location-scale component densities. The unknown age for a component is interpreted as the location parameter (which is the mode for the densities chosen by Sambridge and Compston (1994)) and the standard error for the data point nearest the mode as the scale. The main drawback of using the standard error to construct the mixture density is that to obtain estimates of the modes (for each component), it has to be independent of the scale parameter. In some cases, densities which obey this constraint may not fit the data well, even under transformation. A second problem is that this approach does not provide a formal way to construct a density estimate for the age distribution. This will mean that it is difficult to assess how well the mixture has fitted the data. Other approaches for analyzing geochronological data include graphical methods such as the radial plot (Galbraith (1988) and, loosely linked to kernel density estimation, probability density plots (e.g., Brandon, 1996; Ireland and others, 1998; Sircombe, 2004). We consider these methods as tools for initial or exploratory data analysis, i.e., before any formal inferences are made. However in this context, we also note the limitations with probability density plots pointed out by Galbraith (1998), which include the fact that such plots combine measurement error with true age variation and may mask the underlying signal. It is therefore an important issue to construct new statistical models for the analysis of geochronological problems.

This article is structured as follows. In the next section we demonstrate that current methodology for geochronological data is not always appropriate; we use two uranium-lead (U-Pb) zircon age data sets. We then present our approach for mixture modelling as well as the simulation (MCMC) methods needed to perform inference from these models. In the next two sections we present two detailed examples of how to use our methodology. Finally, we conclude the paper with a discussion.

EXISTING METHODOLOGY

Model

The approach of Sambridge and Compston (1994), following partly from Galbraith and Green (1990), is as follows. The observed data, or calculated ages, $x_1, \ldots, x_n$ are assumed independently distributed as

$$p(x_i; \epsilon_i, \boldsymbol{\theta}, k) = \sum_{j=1}^{k} w_j f(x_i; \boldsymbol{\phi}_j, \epsilon_i) \quad i = 1, \ldots, n$$

where a generic probability mass/density function is denoted by $p(\cdot)$, $f(\cdot)$ is denoted as the component density, the weights $\boldsymbol{w} = (w_1, \ldots, w_k)$ are such that

$w_j \geq 0 \; \forall j = 1, \ldots, k$, $\sum_{j=1}^{k} w_j = 1$, ϕ_j are the component specific parameters (here location parameters), $\phi = (\phi_1, \ldots, \phi_k)$, and $\theta = (\mathbf{w}, \phi)$.

More specifically, Sambridge and Compston (1994) take $f(\cdot)$ to be (under a slightly different parameterization) a member of the exponential power family:

$$f(x_i; \phi_j, \epsilon_i) \propto \exp\left\{-\frac{1}{2}\left|\frac{x_i - \phi_j}{\epsilon_i}\right|^q\right\} \quad q \in [1, 2]$$

that is, $q = 2$ gives the normal distribution and $q = 1$ the double exponential distribution (location ϕ_j, scale ϵ_i). The latter is often preferred as it is more robust to outliers, and has higher probability in the tails than the normal distribution. Sambridge and Compston (1994) seek to find the parameters, $\boldsymbol{\theta}$, which maximize the log likelihood $L(\boldsymbol{\theta}, k) = \log\{\prod_{i=1}^{n} p(x_i; \epsilon_i, \boldsymbol{\theta}, k)\}$ for various k. To select the number of components k they use, for example, a relative misfit criterion based upon the optimized model for progressively increasing number of components. To see that this approach does not always perform well, we consider two data sets. Firstly, we describe the data, then secondly, we analyze them using the approach of Sambridge and Compston (1994).

Carrickalinga Head Formation

The first data set considers 100 U-Pb ages from the Carrickalinga Head formation, which is part of the Kanamantoo group in southern Australia. The ages are quoted in Mega-Annum (Ma), or millions of years. The data are presented in Ireland and others (1998). Figure 1 (a) demonstrates that there are possibly three groups in the data: the samples aged up to 1000 Ma, the middle six data points (approximately 1500 Ma), and the rest of the data. This is supported by the histogram in Figure 2 (a).

Ireland and others (1998) constructed a kernel density estimate (with normal densities) with the standard error as the scale parameter and bandwidth as 1 Ma. They inferred that there were two major components, the Delamerian (500–600 Ma) and Grenvillean (1000–1200 Ma).

Khorat Basin

The second data set relates to 251 U-Pb zircon ages from the Khorat Plateau basin in northern Thailand. The data are reported in Carter and Moss (1999), and Carter and Bristow (2003) who discuss its geological relevance. The main issue was the likely geological age and original source region of these sediments.

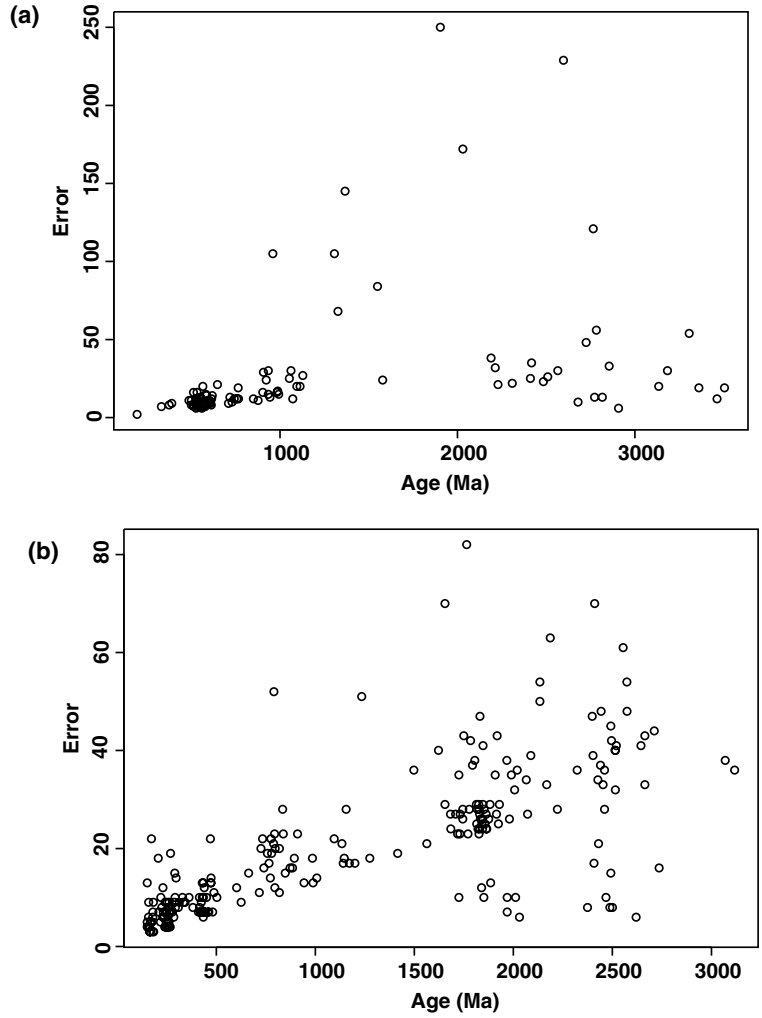

Figure 1. Plot of the U-Pb zircon age data analyzed in this paper. Plot (a) is a plot of the error against ages for the Carrickalinga Head formation. Plot (b) is a plot of the error against ages for the Khorat basin.

Figure 1 (b) illustrates the (fixed) errors that are provided with the data. From Figure 1 (b) it does not appear that the measurement error is too severe, with the error and variability of error broadly increasing with age. A histogram of the data can be seen in Figure 2 (b), which shows the large number of modes in the estimated data.

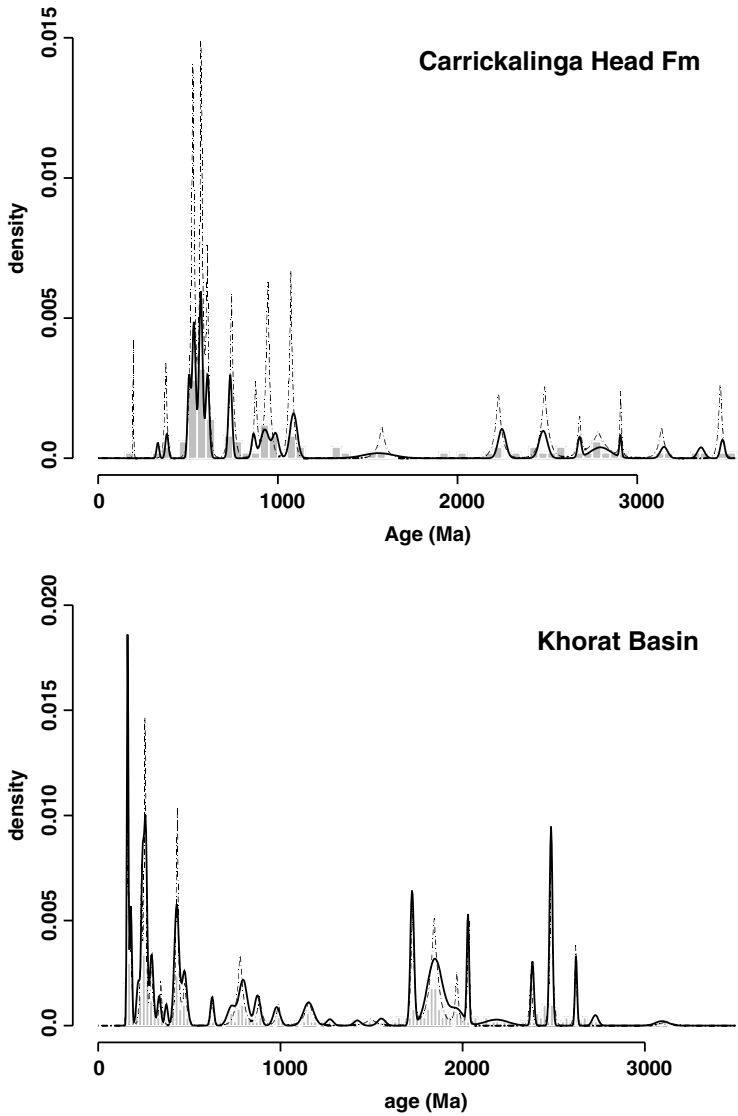

Figure 2. Histogram of the U-Pb zircon age data analyzed in this paper. The data is overlaid by a density estimate based upon the normal model (—) and the double exponential model (–·–) of Sambridge and Compston (1994).

Carter and Bristow (2003) broke the data into four subgroups based on the stratigraphic relationships of the host sediments and applied the method of Sambridge and Compston (1994) to each group individually, then, using a somewhat ad-hoc procedure, recombined the data and inferred that there were five mixture components in the data. Two other components were described as "lesser."

Results Using Sambridge and Compston's Approach

We now apply the method of Sambridge and Compston (1994) and demonstrate some of the problems that can arise with this approach.

Normal and Double Exponential Components

For both data sets, we used either normal or double exponential component densities, and considered an increasing number of components. To select the optimal number of components we use a formal statistical procedure: the Bayesian information criterion (BIC) (see McLachlan and Peel (2000)), as opposed to the relative misfit criterion or detecting the number of components by eye as suggested by Sambridge and Compston (1994).

The analysis of the Carrickalinga Head formation yielded optimal solutions of 22 and 17 components (maximum number of components taken to be 30) for the normal and double exponential models, respectively. A lower number of components may be selected by choosing a smaller maximum number of components (see Fig. 3 (a)). This is problematic, under the BIC criterion, since the optimal solution will often be this maximal value; thus, by limiting the maximum number of components to a relatively small value, in effect we are choosing the number of components *a priori*, without considering competing models.

For the Khorat basin, the optimal number of components for the normal model was found to be 29 and double exponential 22 (allowing a maximum of 30 in both cases). We note that if we use the number of components for which the BIC flattens off (i.e., when increasing the number of components does not lead to a significant reduction in BIC, known as the elbow) the solutions are 18 and 11 components, respectively (Fig. 3 (b)).

The optimal mixture densities for both data sets can be seen in Figure 2. It is clear that a large number of the components fitted to the data appear to be spurious, i.e., do not provide any additional explanation of the data. To construct the density estimates in Figure 2 we took the optimal location parameters and the standard error for the data point closest to it and then evaluated the density on a grid of 1000 equally spaced points on [0, 3500]. Since this procedure is unsatisfactory from a statistical point of view, it is inadvisable to determine the number of components from visual inspection of the density estimate.

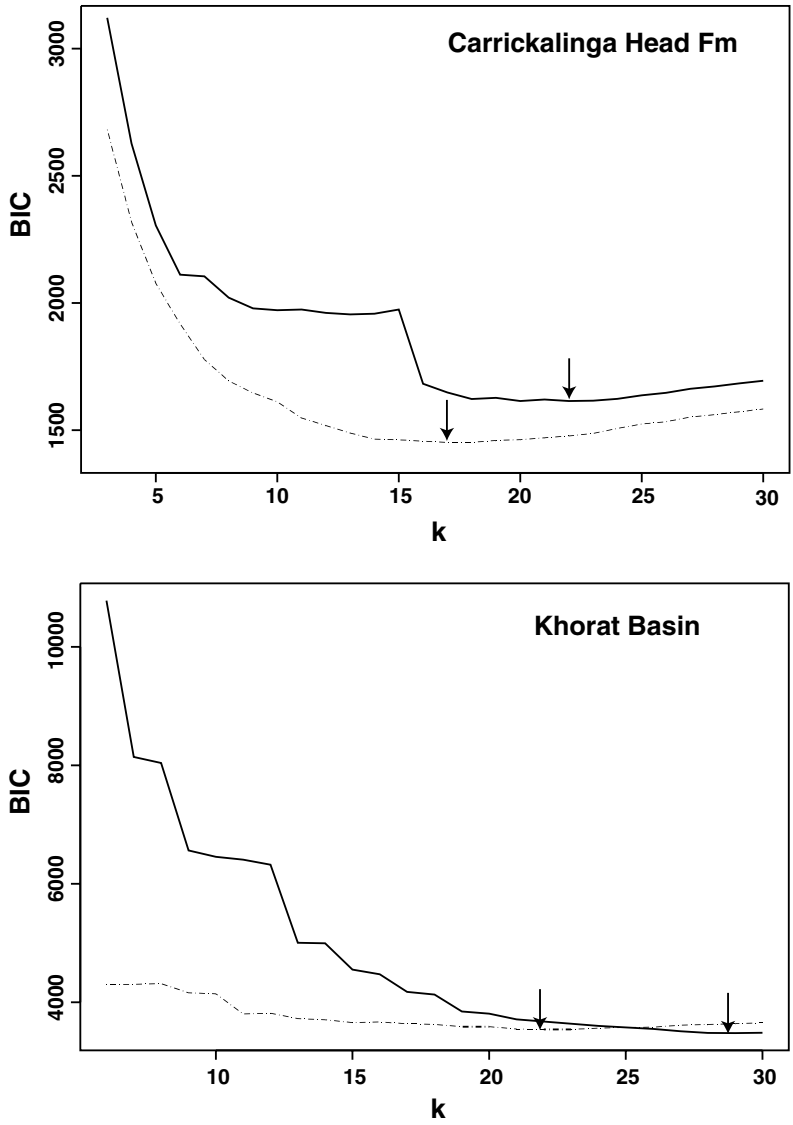

Figure 3. Plot of the BIC values for the U-Pb zircon age data analyzed in this paper. We used the approach of Sambridge and Compston (1994) to fit the models. (—) is the BIC for the normal model and (–·–) for the double exponential model. The *arrows* mark the optimal solutions according to the BIC criterion.

Student-t Components

Under the BIC criterion, if the choice of component density is "incorrect," we will tend to fit too many components (McLachlan and Peel, 2000, p. 209). It is probably the case that this occurred above, so to test this we use student-t components. The t density allows heavy tailed behavior, which is likely to reduce the number of components fitted to the data. This will occur when there are apparent outliers (as a result of the measurement process or subtle geological perturbations) as opposed to representing additional component age modes.

We fitted the student-t model using a trans-dimensional simulated annealing algorithm (Brooks, Friel, and King, 2003), essentially the MCMC algorithm described in the next section. The results can be seen in Figure 4. We used the same method as above to obtain the density estimate.

For the student-t density we obtain an intuitively more reasonable solution than for the normal or double exponential model. The optimal values of k were 3 and 6 for the Carrickalinga and Khorat data sets. However, from Figure 4, it is difficult to ascertain how well the model fits the data, but it appears that neither model is entirely appropriate in that some peaks in the histogram are not well represented in the densities.

We feel that the restriction of the component density to those with modes independent of the scale parameter (i.e., symmetrical), is the reason why these models do not fit well (i.e., the inability to model skewness). Therefore, we now focus on producing a more general statistical model.

NEW MIXTURE MODEL FOR GEOCHRONOLOGICAL DATA

Following the remarks made in the previous section, regarding the role of standard errors in the mixture likelihood, we will model the actual, unknown data $y_1, \ldots, y_n$ as a mixture. That is, the data we would have recorded had there been no measurement error. In effect, we are considering the problem as a simple form of image reconstruction (Besag, 1986), where the estimated data provide the degraded image and the error suggests how close it is to the true image.

Likelihood Construction

We will use methods similar to those outlined in Richardson and others (2002), that is:

$$x_i | y_i; \epsilon_i \sim p(x_i | y_i; \epsilon_i)$$

$$y_i | \boldsymbol{\theta}, k \sim \sum_{j=1}^{k} w_j f(y_i; \phi_j).$$

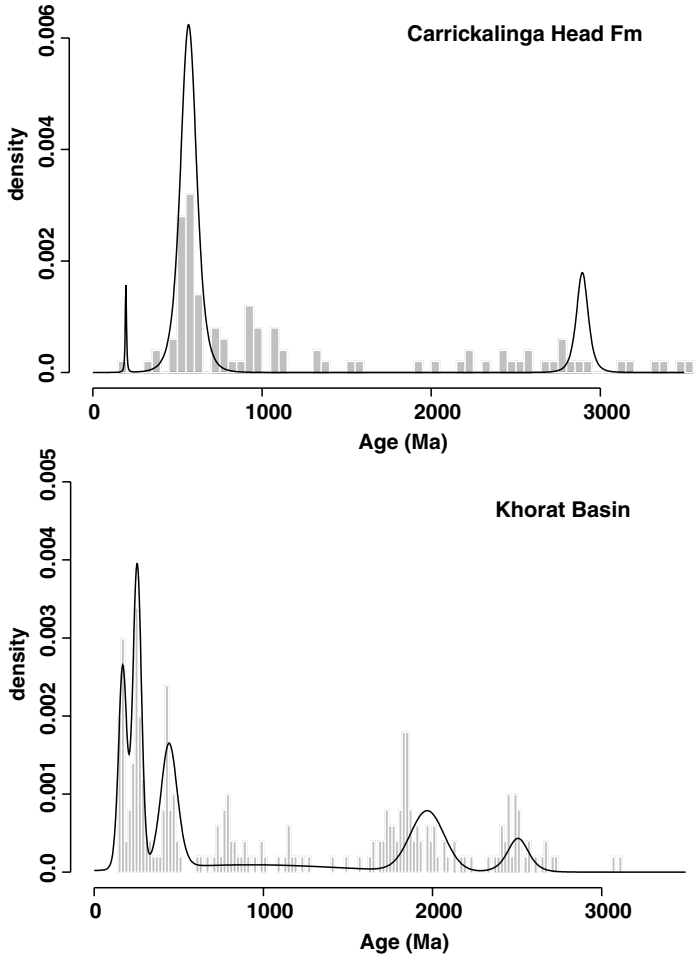

Figure 4. Density estimates of the U-Pb zircon age data analyzed in this paper. We used the approach of Sambridge and Compston (1994) under student-t component densities.

For simplicity we assume the following conditional independence structure

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}, k; \boldsymbol{\epsilon}) = \prod_{i=1}^{n} p(x_i|y_i; \epsilon_i) p(y_i|\boldsymbol{\theta}, k).$$

We choose $x_i|y_i; \epsilon_i$ to be $\mathcal{N}(y_i, \epsilon_i^2)$, where $\mathcal{N}(y_i, \epsilon_i^2)$ is the normal distribution, mean y_i, standard deviation ϵ_i. We allow the component densities, $f(\cdot)$, to be either normally distributed or, to provide further flexibility, skew-t distributions

(Jones and Faddy, 2003). In practice, allowing for skewness allows us to consider distributions which may have been geologically disturbed to varying degrees leading to differential loss or gain of the parent of daughter elements. This will lead to greater dispersion in the data than expected for a single component age.

We denote a skew-t distribution as $\mathcal{S}t(\mu, \lambda, \nu, \zeta)$, with location μ, inverse scale λ, and skew/tail parameters ν and ζ. The skew-t density is:

$$f(y; \mu, \lambda, \nu, \zeta) \propto \left\{ 1 + \frac{\lambda(y-\mu)}{(\nu + \zeta + (\lambda(y-\mu))^2)^{\frac{1}{2}}} \right\}^{\nu+\frac{1}{2}} \times \left\{ 1 - \frac{\lambda(y-\mu)}{(\nu + \zeta + (\lambda(y-\mu))^2)^{\frac{1}{2}}} \right\}^{\zeta+\frac{1}{2}}$$

where the normalizing constant is $c_{\nu,\zeta} = \lambda/(2^{\nu+\zeta-1} B(\nu, \zeta)(\nu + \zeta)^{\frac{1}{2}})$, $y, \mu \in \mathbb{R}$, $\lambda, \nu, \zeta \in \mathbb{R}^+$, and $B(\cdot, \cdot)$ is the beta function. If $\nu = \zeta$ the density is the symmetric-t density on 2ν degrees of freedom. When $\nu > \zeta$ or $\nu < \zeta$ f is negatively or positively skew, respectively. The density is unimodal with mode

$$\varrho = \mu + \frac{(\nu - \zeta)\sqrt{\nu + \zeta}}{\lambda\sqrt{2\nu + 1}\sqrt{2\zeta + 1}}.$$

Jones and Faddy (2003) report that high absolute values of skewness are associated with small values of ν and ζ, as are heavy tails of f. A model with normal or skew-t components is referred to as normal or skew-t model.

A Bayesian Model

To formally encode our prior beliefs about the geological problem at hand, we construct a Bayesian model (see Robert (2001) and Appendix A for an introduction). This requires the specification of prior distributions, which we take to be similar to those adopted in Richardson and Green (1997).

The priors on the locations and inverse scales are taken to be i.i.d (independently and identically distributed) for each component $j = 1, \ldots, k$, with $\mu_j \sim \mathcal{N}(\xi, \kappa^{-1})$, $\lambda_j | \beta \sim \mathcal{G}a(\alpha, \beta)$. We also assume $\beta \sim \mathcal{G}a(g, h)$ and $\mathbf{w}|k \sim \mathcal{D}(\delta)$. Our notation is such that $\mathcal{G}a(\alpha, \beta)$ is the gamma distribution, shape α, scale β, and $\mathcal{D}(\delta)$ is the symmetric Dirichlet distribution parameter δ. The prior on k is either a truncated Poisson prior, $p(k) \propto \tau^k/k!,\ k = 1, \ldots, k_{\max}$ or uniform on the integers $1, \ldots, k_{\max}$; See Richardson and Green (1997) for a detailed discussion of this prior structure.

For the skew-t density we consider two priors on the skew/tail parameters, both assuming they are i.i.d given k.

Prior I

Firstly, a prior that can be considered weakly informative. We take $\nu_j|b \sim \mathcal{U}_{[a,b]}$, a to be fixed, $b \sim \mathcal{E}x(\rho, a)$ and

$$p(\zeta_j|\nu_j, b) = \frac{1}{3}\{\mathbb{I}(\nu_j = \zeta_j) + \mathbb{I}(\zeta_j < \nu_j)\mathcal{U}_{[a,\nu_j]} + \mathbb{I}(\zeta_j > \nu_j)\mathcal{U}_{[\nu_j,b]}\}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{U}_{[a,b]}$ is the uniform density on $[a, b]$, and $\mathcal{E}x(\rho, a)$ is the exponential distribution parameter ρ, location a.

Prior II

Secondly, $\nu_j \sim \mathcal{G}a(\psi, \omega, a)$ (where $\mathcal{G}a(\psi, \omega, a)$ is a gamma distribution with shape ψ scale ω and location parameter a) and

$$p(\zeta_j|\nu_j) = \frac{1}{3}\left\{\mathbb{I}(\nu_j - \zeta_j) + \mathbb{I}(\zeta_j < \nu_j)\mathcal{U}_{[a,\nu_j]} + \mathbb{I}(\zeta_j > \nu_j)\mathcal{E}x(\rho, \nu_j)\right\}.$$

Discussion of Priors for Skew/Tail Parameters

We adopt the above priors from a pragmatic point of view, as they both provide equal support for symmetric, positively and negatively skew-shaped densities, i.e., they are all equally probable a priori).

However, given that we may be dealing with a skewed distribution, the first prior does not favor moderately or severely skewed densities, allowing us to consider the range of support for the parameters (i.e., heavy or light tailed). This behavior can be seen in Figure 5 (a). The settings of the hyperparameter are as for the next example, and we can observe (in Fig. 5 (a)) the large number of heavy tailed shapes permitted by the prior.

The second prior is constructed with the intention of favoring heavy tailed, moderately or severely skewed component densities (given the skewness). That is if ν_j is assumed to be small, then $|\nu_j - \zeta_j|$ will be similarly small (for large ρ), with most of the prior probability being focused on this region. Some densities can be seen in Figure 5 (b). Here we see the extreme skewness allowed.

We constrain ν_j and ζ_j to be bigger than a to avoid fitting components that have extremely heavy tails (as noted by Jones and Faddy (2003)). It also prevents numerical errors in the code.

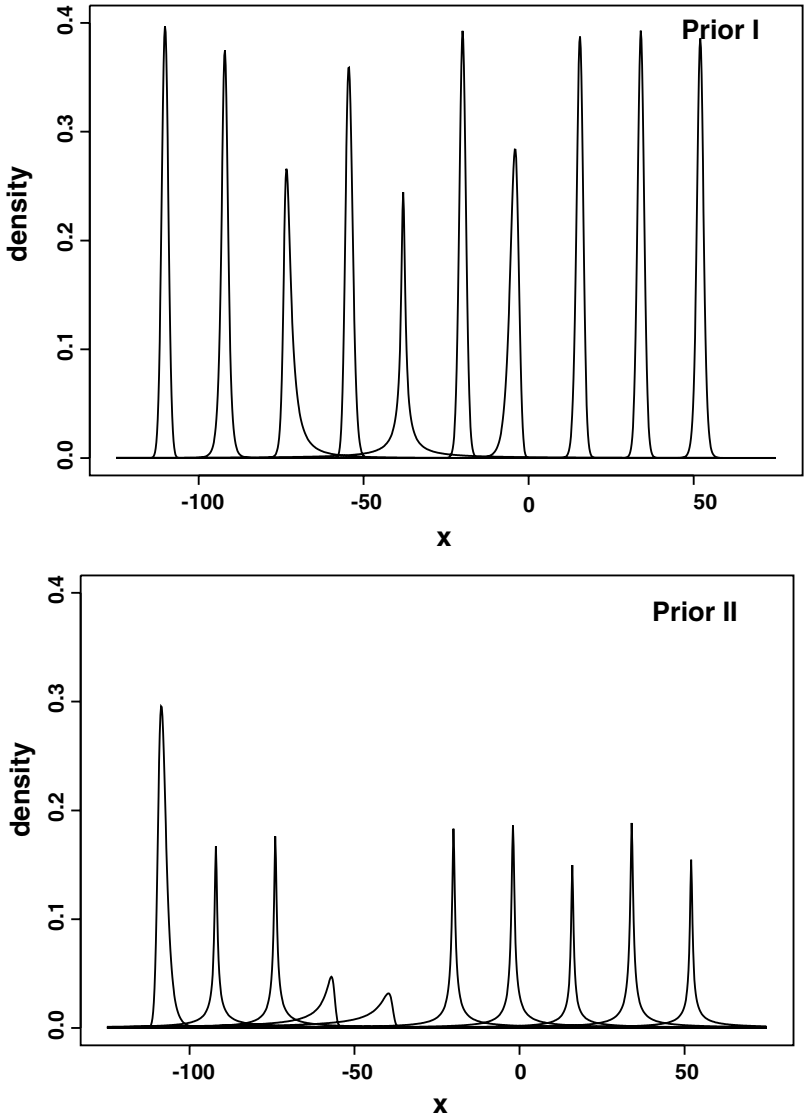

Figure 5. Skew-t densities with skew/tail parameters drawn from the prior to demonstrate the potential range of distributions. We show 10 densities plotted on $[-125, 75]$ with location parameter increasing by 18 from *left* to *right* and scale 1, (ν, ζ) drawn from both priors with specification of the relevant hyperparameters (a, ρ, as described for the two examples.

Posterior

The posterior $\pi(\boldsymbol{\theta}, \mathbf{y}, \beta, k|\mathbf{x}; \boldsymbol{\epsilon})$ for the skew-t model under priors I and II are:

$$\pi(\boldsymbol{\theta}, \mathbf{y}, \beta, k|\mathbf{x}; \boldsymbol{\epsilon}) \propto \prod_{i=1}^{n}\{p(x_i|y_i; \epsilon_i)p(y_i|\boldsymbol{\theta}, k)\}\prod_{j=1}^{k}\{p(\mu_j)p(\lambda_j|\beta)p(\nu_j, \zeta_j|b)\}$$
$$\times p(\mathbf{w}|k)p(\beta)p(b)p(k)$$
$$\pi(\boldsymbol{\theta}, \mathbf{y}, \beta, k|\mathbf{x}; \boldsymbol{\epsilon}) \propto \prod_{i=1}^{n}\{p(x_i|y_i; \epsilon_i)p(y_i|\boldsymbol{\theta}, k)\}\prod_{j=1}^{k}\{p(\mu_j)p(\lambda_j|\beta)p(\nu_j, \zeta_j)\}$$
$$\times p(\mathbf{w}|k)p(\beta)p(k).$$

Simulation Details and Markov Chain Monte Carlo Inference

We are interested in calculating posterior expectations to estimate, for example, the posterior probability that $k = j,\ j = 1, \ldots, k_{\max}$. To approximate the appropriate integrals, we rely upon Monte Carlo integration via reversible jump MCMC (Green, 1995). We provide a simple introduction to this approach in Appendix B, for comprehensive details see, for example, Robert and Casella (2004).

The reversible jump sampler for the skew-t model is now described, with obvious modifications for normal component densities in terms of the prior distributions and model parameters. We denote a proposed value of x (here a generic random variable) by x' and w.p. means "with probability." All updates are standard Metropolis–Hastings moves, involving either a single parameter or blocks of parameters, unless otherwise stated.

1. *Update* $\mathbf{y}$: We update y_i individually, in the order of the indices. We use an additive Cauchy/normal random walk via delayed rejection (Green and Mira, 2001). Here we make an initial random walk proposal based upon a large variance. If the move is rejected, make a second random walk proposal with a smaller variance. We center both moves at the current state of the chain.
2. *Update the locations*, μ: This is done in one block (i.e., $(\mu_1, \ldots, \mu_k)$) via an additive Cauchy or normal random walk.
3. *Update the inverse scale parameters*, λ: Again, in one block via a multiplicative log-normal random walk.
4. *Update the weights*, $\mathbf{w}$: In one block via an additive normal random walk on the logit scale.

5. *Update* β: We use a Gibbs kernel, full conditional $\mathcal{G}a(g + k\alpha, h + \sum_{j=1}^{k} \lambda_j)$.
6. *Update the skew/tail parameters*, (ν, ζ): We update $(\nu_1, \zeta_1, \ldots, \nu_k, \zeta_k)$ in one block via a Metropolis–Hastings move. For prior I on (ν_j, ζ_j), we perform random walks on $\log\{(\nu_j - a)/(b - \nu_j)\}$ (similarly for ζ_j) and for prior II on $\log\{\nu_j - a\}$. If ν_j and ζ_j are currently equal w.p. υ_s we propose them to be equal updating via a Cauchy or normal transformed random walk, or w.p. $1 - \upsilon_s$ that they are not equal, using a normal/Cauchy transformed random walk. Similarly if ν_j and ζ_j are not equal, w.p. υ_a we propose them not to be equal updating via a Cauchy or normal transformed random walk, w.p. $1 - \upsilon_a$ equal and a normal transformed random walk.
7. *Birth/Death of a component*: This move largely follows Richardson and Green (1997). For the birth in state k, selected with probability b_k, propose from current state $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{w}, \boldsymbol{\nu}, \boldsymbol{\zeta}, k)$ to $(\boldsymbol{\mu}', \boldsymbol{\lambda}', \mathbf{w}', \boldsymbol{\nu}', \boldsymbol{\zeta}', k + 1)$ $(k = 1, \ldots, k_{\max} - 1)$ by setting the diffeomorphism (the function and its inverse are differentiable) linking $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{w}, \mu, \lambda, w, \nu, \zeta, k)$ and $(\boldsymbol{\mu}', \boldsymbol{\lambda}', \mathbf{w}', \boldsymbol{\nu}', \boldsymbol{\zeta}', k + 1)$ as the identity for all random variables other than the weights, for which we set $\mathbf{w}' = (w_1(1 - w), \ldots, w_k(1 - w), w)$. The extra random variables needed to match dimension $(\mu, \lambda, \nu, \zeta)$ are drawn from the prior and $w \sim \mathcal{B}e(1, k)$ where $\mathcal{B}e(\cdot, \cdot)$ is the beta distribution. This move has acceptance probability $\min\{1, A\}$, with

$$A = \frac{p(\mathbf{y}|\boldsymbol{\theta}', k + 1)p(k + 1)}{p(\mathbf{y}|\boldsymbol{\theta}, k)p(k)} B(k\delta, \delta)^{-1} w^{\delta - 1}(1 - w)^{k(\delta - 1)} \frac{(k + 1)!}{k!} \times \frac{d_{k+1}}{(k + 1)b_k} \frac{(1 - w)^{k-1}}{\mathcal{B}e(w; 1, k)}$$

   where $p(\mathbf{y}|\boldsymbol{\theta}, k) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}, k)$, d_k is the probability of proposing to perform a death move in state k and $\mathcal{B}e(w; \cdot, \cdot)$ is the beta density evaluated at w. The reverse death move is achieved by selecting a component with uniform probability to die and inverting the jump function. The death move is available for $k = 2, \ldots, k_{\max}$.
8. *Update* b: A normal random walk on the $\log(m - b)$ scale, $m = \max_j\{\zeta_j, \nu_j\}$.

We perform the algorithm in a deterministic sweep in the order 1–6 followed by the random choice of a birth or death. Step 8 is added if prior I on (ν_j, ζ_j) is used.

We now demonstrate our methodology on the two data sets we described earlier. Note that the code is available from the first author on request.

EXAMPLE I: CARRICKALINGA HEAD FORMATION

Firstly we consider the analysis of the Carrickalinga Head formation data. For illustration we adopt skew-t densities with prior I on (ν_j, ζ_j) and a uniform prior on k.

In order to fix ξ, κ and h we define the following. Let R be the range of the data defined to be $R = \max_i\{x_i + 2\epsilon_i\} - \min_i\{x_i - 2\epsilon_i\}$ and M be the midpoint of the data defined to be $M = [\max_i\{x_i + 2\epsilon_i\} + \min_i\{x_i - 2\epsilon_i\}]/2$. We take $\xi = M, \kappa = 1/R^2, h = 10/R^2$. We also let $\alpha = 2, \delta = 1, g = 0.2, k_{\max} = 30$ all these settings being consistent with Richardson and Green (1997). For ρ, we set $\rho = 1/10$, which states that we prefer components with heavy tails—i.e. we wish to avoid spurious components introduced by the light tails inherent in normal distributions. Finally we set $a = 0.01$, which was not so small as to introduce numerical errors into our code.

Performance of the Sampler

We ran our MCMC sampler for 200,000 sweeps taking the burn in to be 100,000 sweeps (see Robert and Casella (2004) for a discussion of convergence issues in MCMC). The acceptance rates of all fixed dimensional moves were in the range $(0.3, 0.6)$. The birth/death move was accepted 10% of the time indicating good mixing over k (the birth/death probabilities were uniform among the moves allowed given the current state of the chain).

Simple convergence measures can be seen in Figure 6. In Figure 6 (a) the occupancy probabilities can be seen ($p(k \leq j|\mathbf{x}; \epsilon)$) every 100 sweeps of the algorithm. This quantity quickly stabilizes, indicating the fast convergence of the algorithm. Figure 6 (b) reiterates the rapid mixing of the algorithm: our sampler is able to visit a large number of potential models with little difficulty.

Inference

Now that we are satisfied that our MCMC sampler has converged we now seek to draw inference from our statistical model. This is complicated by the *label switching* problem, see Stephens (2000) or Jasra, Holmes, and Stephens (2005) for a review and our approach is described in detail below. In practice, this is manifested as two parameters (e.g., 1 and 2) which might represent age modes, swapping over so component 1 may become younger than component 2. The parameters associated with these age components also swap. Thus the identifiability of parameters needs to be considered.

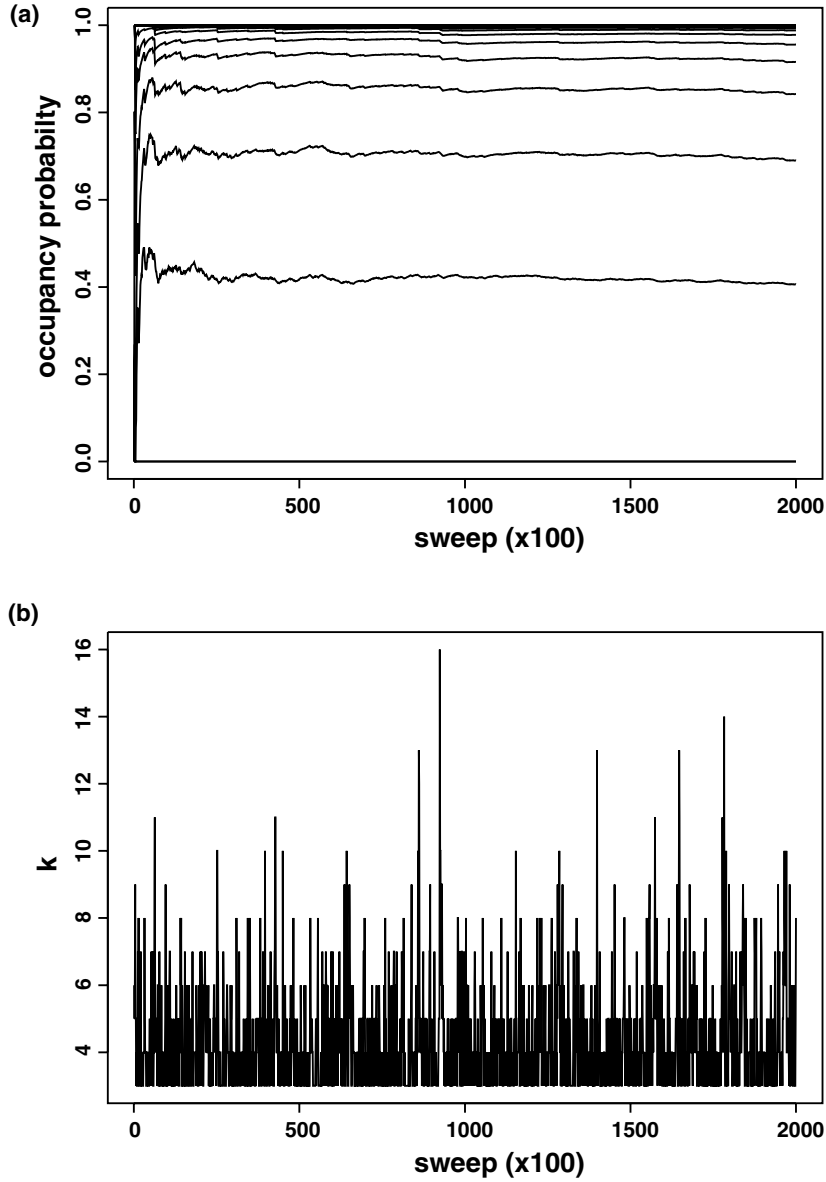

Figure 6. Convergence of MCMC algorithm; Carrickalinga Head Formation. (a) displays the cumulative occupancy probability ($p(k \leq j|\mathbf{x};\epsilon)$): the line at 0.4 is $p(k \leq 3|\mathbf{x};\epsilon)$. (b) displays the sampled k. Both (a) and (b) are recorded every 100 sweeps.

Label Invariant Quantities

We firstly concentrate on inference which is not subject to label switching. To find the posterior distribution for k, we use the Monte Carlo estimate:

$$p(k = j|\mathbf{x}; \epsilon) \approx \frac{1}{N} \sum_{t=1}^{N} \mathbb{I}\big(k^{(t)} = j\big)$$

where N is the number of MCMC samples taken and $k^{(t)}$ are the (post burn-in) sampled value of k at sweep t.

The posterior distribution for k can be seen in Table 1, column 1. We observe that the maximal posterior probability is for $k = 3$; one more component than described by Ireland and others (1998). We discuss this point later in this section

To compute a density estimate (denote this $p(y|\boldsymbol{\theta}, k)$) we use:

$$p(y|\boldsymbol{\theta}, k) \approx \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{k^{(t)}} w_j^{(t)} f\big(y; \phi_j^{(t)}\big)$$

evaluated over a grid of 1000 equally spaced points on $[0, 3500]$. The estimate can be seen in Figure 7 (a). It displays the three components represented in the posterior distribution for k, and is significantly more smooth than the estimate in Figure 2 (a). Comparing with the solution under the (symmetrical) student-t model (Fig. 4 (a)), we can see that we have picked out modes in higher density regions (e.g. the mode at 190 Ma in Fig. 4 (a)).

Table 1. Posterior Distribution for the Number of Components Using Various Skew-t Models; Carrickalinga Formation

	Distributions for the following values of ρ:				
k	$\rho = 1/10$	$\rho = 1/20$	$\rho = 1/30$	$\rho = 1/40$	$\rho = 1/50$
≤ 2	0.000	0.000	0.000	0.000	0.000
3	0.404	0.381	0.381	0.396	0.387
4	0.287	0.272	0.276	0.276	0.275
5	0.156	0.160	0.161	0.157	0.156
6	0.076	0.086	0.089	0.082	0.085
7	0.037	0.047	0.047	0.043	0.046
≥ 8	0.040	0.054	0.046	0.046	0.051

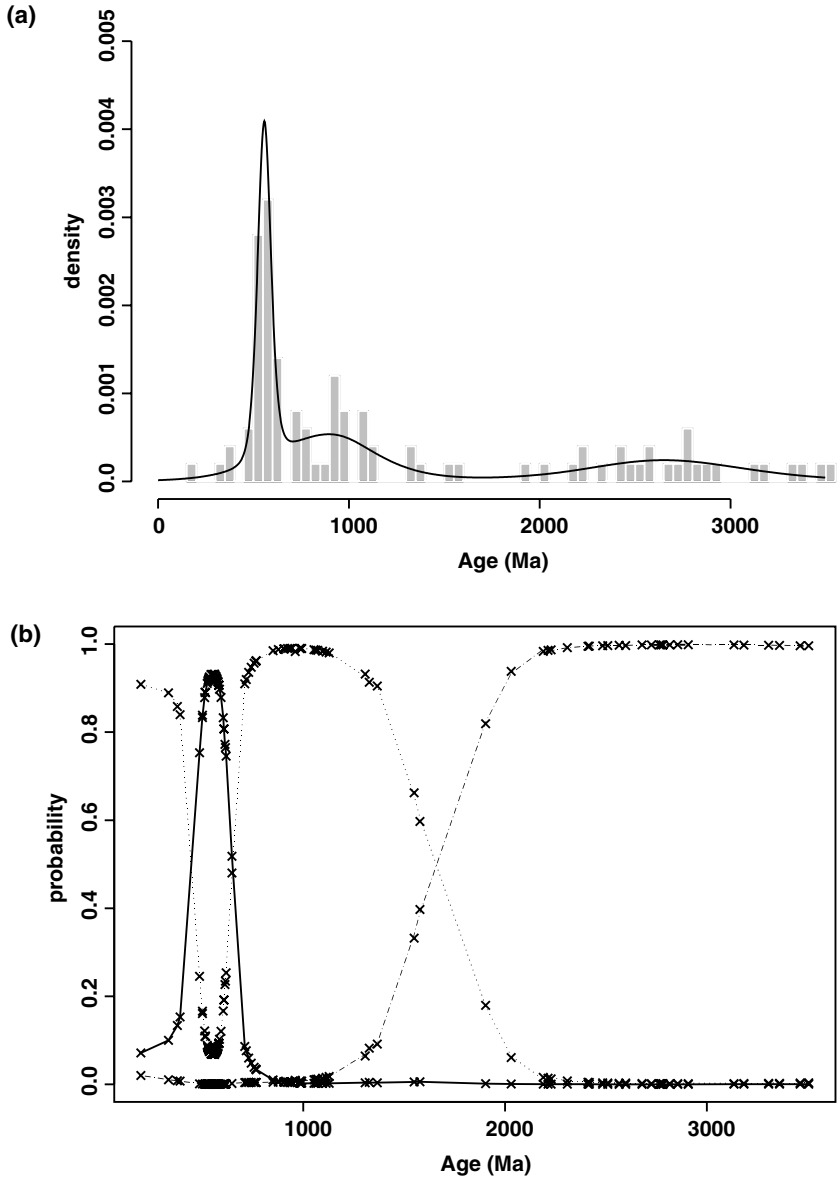

Figure 7. Density estimate and classification probabilities; Carrickalinga Head Formation. For plot (b), the crosses are data points and the lines are the probability that a data point is in the component, for component 1 (—), 2 (· · ·) and 3 (–·–).

Label Switching

The label switching problem occurs because the likelihood is invariant to permutations of the labels of the parameters. If the prior exhibits this property, so will the posterior distribution. As a consequence, the marginal posterior distribution for the component specific parameters, given k, are identical for each component. This phenomenon is demonstrated in Figure 8 (a), where we conditioned our MCMC samples such that $k = 3$ (our samples were taken post burn-in, with a minimum of five sample thinning). In this case it is simple to remove label switching by applying the constraint $\varrho_1 < \varrho_2 < \varrho_3$, as shown in Figure 8 (b) in which the constraint is applied offline. We note that this not always appropriate, as these constraints are not always effective in removing label switching; see Stephens (2000) for example. The estimated modes with 95% credible intervals are: $\varrho_1 = 556.660$, $C_{0.95}(\varrho_1) = (545, 744, 567.590)$, $\varrho_2 = 881.543$, $C_{0.95}(\varrho_2) = (779.013, 982.353)$ and $\varrho_3 = 2651.60$, $C_{0.95}(\varrho_3) = (2473.680, 2818.540)$, where $C_{0.95}(\cdot)$ denotes the 95% credible interval for a parameter.

If we are interested in classifying the data, i.e.. allocating which rock belongs to a particular component, we can use the classification probabilities:

$$p(z_i = j|\mathbf{x};\epsilon) \approx \frac{1}{N_{k^*}} \sum_{t=1}^{N} \mathbb{I}\big(k^{(t)} = k^*\big) \frac{w_j^{(t)} f\big(y_i^{(t)}; \phi_j^{(t)}\big)}{\sum_{l=1}^{k} w_l^{(t)} f\big(y_i^{(t)}; \phi_l^{(t)}\big)} \tag{2}$$

where $N_{k^*} = \sum_{t=1}^{N} \mathbb{I}(k^{(t)} = k^*)$, z_i is an allocation variable, i.e., $z_i = j$ if data point i is in component j and k^* is some selected value of k. The probabilities can be seen in Figure 7 (b). We see that the model classifies the youngest data as being in component 2 (on the basis of maximal classification probability), because it is inconsistent with the first component (mode 556 Ma). That this occurs is of little interest in terms of identifying age components since Ireland and others (1998) state that this data is due to Pb loss. However, this does demonstrate that our approach behaves gracefully in such situations, and potentially may be used to identify such geological perturbations.

Comparison of Results With Inference of Ireland and Others (1998)

We now compare our inferences with those in Ireland and others (1998). Ireland and others (1998) state that there is evidence for two major components, which we find (component 1, 556 Ma, component 2, 881 Ma), we also find evidence for a third component, mode 2651 Ma. Using our methodology we have been able to precisely find an estimate of second mode at 881 Ma (compared to the approximate 1000–1200 Ma quoted by Ireland and others (1998)). This is

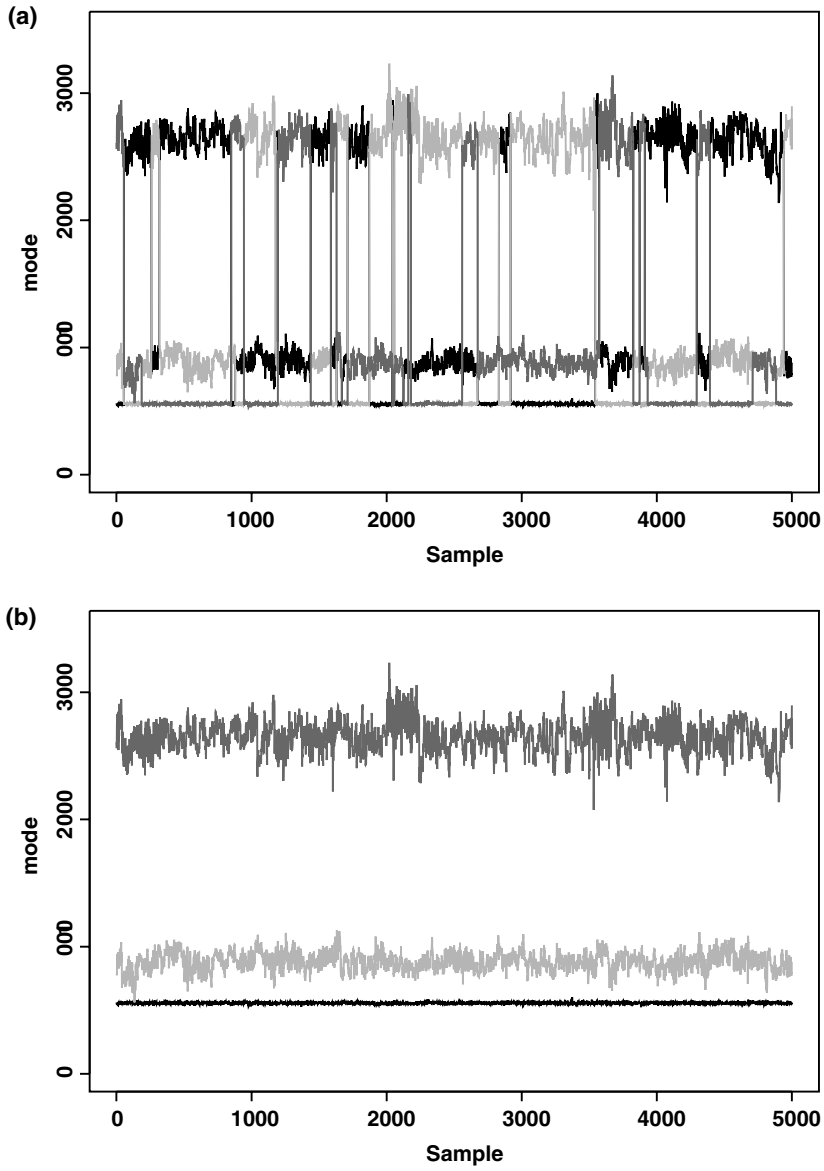

Figure 8. Sampled modes; Carrickalinga Head Formation. (a) displays the sampled modes as returned by our MCMC algorithm. (b) are the sampled modes after permuting the samples by ordering. component 1—*black*, 2—*light grey*, 3—*dark grey*.

potentially of more use to a geologist when attempting to find sediment provenance for example. Ireland and others (1998) do not infer the third component indicated by our results. It may be that, while this component is of statistical importance, it is of little interest to geologists. However, this latter inference is best made having first recognized the statistically significant components. Similarly, the inference of various age components (e.g., 881 Ma) needs to be rationalized with other geological information, although this aspect is beyond the aims of this paper.

Stability With Respect to the Prior

One aspect of particular importance in Bayesian mixture modelling is the sensitivity of the posterior for the number of components, to changes in the value of prior parameters (especially for the prior on the component specific parameters; see Jennison (1997) for example).

Our model is based upon that of Richardson and Green (1997) and the sensitivity of inferences related to the posterior distribution for k are investigated there. They are as follows. In the limit as $\beta \to \infty, \kappa \to 0$, Jennison (1997) notes that the posterior distribution for k starts to favor one component, which is a form of Lindley's paradox (Lindley, 1957) (that is, as we seemingly place less prior information on the parameters, the prior becomes highly informative for the dimensionality of the model, in that it puts more posterior weight on models with fewer components). We found similar behavior for our model. Therefore, we only investigate the sensitivity to changes in ρ; this can be seen in Table 1.

In Table 1, we can observe that, under a reasonable departure from the default prior, the posterior inference for the number of components does not change significantly (i.e., the relative probabilities of models with different numbers of components is the same for different values of ρ). This is reassuring, as the only part of our prior intended to be informative is that on (ν_j, ζ_j) (in that we prefer heavy tailed densities to take into account the measurement error in the data) and the posterior for k is fairly robust to changes in this prior.

EXAMPLE II: KHORAT BASIN

For our second example we return to the Khorat basin data. We consider analysis of the data using both normal and skew-t (with prior II) component densities and a truncated Poisson prior on k. We adopt an informative prior specification, in that we favor shapes apparent in the data histogram. We do this both for illustration and because the prior settings (for ν and ζ) in the previous example did not appear to fit the data well in this example.

We set the prior parameters $\xi, \kappa, \alpha, \delta, g, h, k_{\max}$ as for Example I. For the truncated Poisson prior, we set $\tau = 5$, which gives prior expectation of k to be approximately 5 (i.e., we center on the result of Carter and Bristow (2003)). The prior parameters for the skew-t model are as follows. For ν_j the value of a is 0.05 (set so as to be low as possible while avoiding numerical errors in the code as a tends to zero) and the prior mean ($= \psi/\omega + a$) and variance ($= \psi/\omega^2$) are 0.5 and 5, respectively. To model $\zeta_j|\nu_j$ we set $\rho = 1/\sqrt{5}$. This prior favors component shapes that are moderately, positively skewed, and extremely negatively skewed—with heavy tails. This specification is chosen since we found that the weakly informative settings of the previous example, did not favor components with the extreme skewness apparent in the histogram of the data (Fig. 2 (b)). See Figure 5 (b), to get an idea of the density shapes favored by this choice of prior.

MCMC Sampler Performance

For the normal model, we ran our MCMC sampler for 1 million sweeps and used the last 250,000 samples for inference. That convergence takes longer than the previous example is to be expected: We have a larger data set with more modes and therefore a more complex target space.

The skew-t MCMC algorithm was run for 4 million sweeps after a burn-in period of 500,000 sweeps. We found that such substantial sample numbers were required to ensure that the sampler not only converged but that mixing was acceptable.

Inference for k

We now consider the posterior distribution for the number of components under the normal and skew-t models (Table 2).

Table 2. Posterior Distribution for the Number of Components Using Normal and Skew-t Models; Khorat Basin

k	Normal	Skew-t
≤ 5	0.000	0.000
6	0.000	0.522
7	0.008	0.351
8	0.092	0.104
9	0.209	0.020
10	0.279	0.003
≥ 11	0.412	0.000

In Table 2 we can observe that the normal model favors many more components than for the skew-t model. We attribute this to the component shapes that can be represented by the normal density. That is, the normal distribution produces components that are symmetric and light tailed, which does not appear to be the case in the estimated data (Fig. 2 (b)). As a result, the normal mixture requires more components to fit the data, compared with the skew-t model.

The heavy tails and skewed component densities which we preferred in our prior, for the skew-t model, has led to a more parsimonious representation of the data than for the normal model. This illustrates the importance in being able to subjectively specify heavy tailed, skew behavior for geochronological data, especially when parsimony is of interest.

Inference Under the Skew-t Model

We conclude this example with some inferences using the skew-t model. We used a Bayes factor to decide that the skew-t model was preferable to the normal one. To deal with label switching, we use the method discussed in Appendix C.

We first consider the estimates of the modes and weights (Table 3). In keeping with the trend of increasing error with age (Fig. 1 (b)) the width of the credible intervals for the modes generally increases with the size of estimate (an exception is component 5). For example, $\varrho_4 = 165.285$, $|C_{0.95}(\varrho_4)| = 10$ Ma (where $|C_\cdot(\cdot)|$ is the width of a credible interval) compared with $\varrho_3 = 2475.847$, $|C_{0.95}(\varrho_3)| = 78$ Ma. Component 5 has the largest credible interval, since for some sweeps it is a symmetric density, with different mode to when it is skewed thus increasing the uncertainty of the estimate. In terms of the weights, component 2 (data around 1800 Ma) have largest contribution which is supported by the fact that most data points are allocated to this class (under the maximal classification probability, Fig. 9 (b). We note, in comparison to the solution using student-t densities that (see Fig. 4 b), our estimates are not too dissimilar. However, under our approach,

Table 3. Posterior Estimates of the Modes and Weights With 95% Credible Intervals Under the Skew-t Model; Khorat Basin

Component	Mode	Weight
1	253.054 (242.975, 263.295)	0.201 (0.147, 0.265)
2	1835.779 (1792.220,1873.010)	0.284 (0.219, 0.354)
3	2476.847 (2438.160, 2516.770)	0.128 (0.081, 0.182)
4	165.285 (160.103, 170.255)	0.110 (0.072, 0.157)
5	804.570 (742.722, 898.306)	0.159 (0.107, 0.218)
6	438.366 (426.888, 451.494)	0.117 (0.074, 0.169)

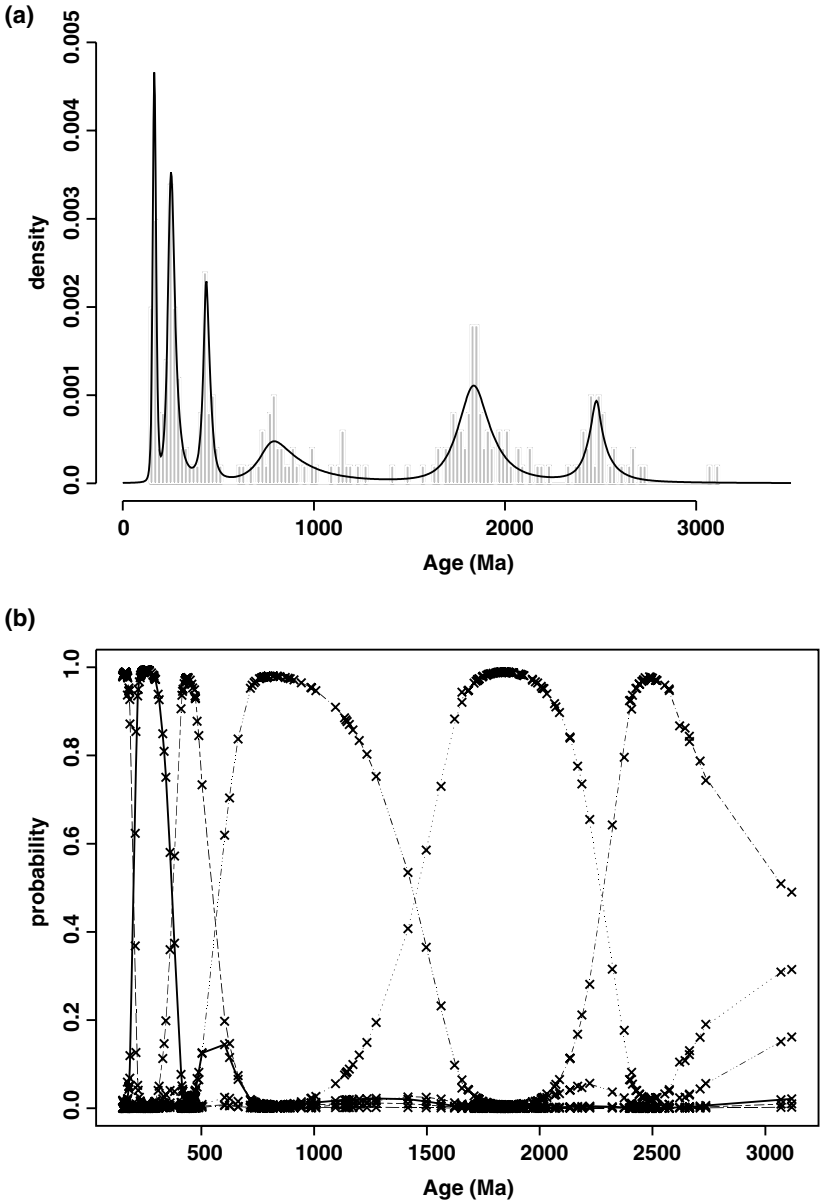

Figure 9. Density estimate and classification probabilities under the Skew-*t* model; Khorat Basin. The density estimate is for the six component skew-*t* model.

we are reasonably satisfied with the fit of our model (i.e., we have some idea of its validity).

The classification probabilities (Fig. 9 (b)) show a typical trend, that is, data points that are similar in age are likely to be clustered together. It is interesting that data points that are adjacent in terms of age are not allocated to different groups (except at the break points of course), something which occurred in the previous example. We observe that, while there is generally a clear allocation of data point to component, at the break points between classes there is some ambiguity with respect to class membership. One point of note are the outlying observations at 3100 Ma. These data points do not seem to be entirely consistent with the component they are allocated to. That is, these data are the only points which have classification probability larger than 0.1 for more than one class. Indeed if we observe Figure 2 (b) we see that the error associated with these points is approximately 40 Ma; it may be that they warrant a separate component. While it is not statistically unreasonable to fit another component, there are relatively few data to justify this.

DISCUSSION

In this paper we have demonstrated that current statistical methodology for mixture modelling of geochronological data does not always work well, often fitting an unacceptably large number of components under the BIC criterion. We constructed a new generalized approach and showed that it provided more reasonable solutions for the examples considered. The method deals with nonsymmetric (unimodal) component densities, which are likely to be an important factor in mixture modelling for geochronology. Skewness is not be an entirely random phenomenon of geochronological data, but can arise due to geological perturbations, with an effect such as loss of the daughter isotope, the significance typically depending on grain size and structure, for example.

While our model may be adopted using likelihood-based inference, the unboundedness of this quantity makes optimization methods difficult to implement (in our experience). Our approach is best applied in the Bayesian context, where we can explicitly introduce prior information into the problem, especially with respect to heavy tailed and skewed behavior. The priors adopted for the normal model and the skew-t model (prior I) can be considered as being geared toward exploring heterogeneity (Richardson and Green, 1997) and, other than the influence of Lindley's paradox, is reasonably robust (with respect to inference for the number of components) to the prior specification. Prior II is constructed so that a subjective data analysis may be performed, as for example II, which implied skewed distributions. In this case we recommend simulating from the prior to determine the component shapes which are most likely to occur. In terms of issues specific

to Pb isotope data, a referee (KS) suggested that a measure of discordance such as

$$\left(1 - \frac{^{206}\mathrm{Pb}/^{238}\mathrm{U}}{^{207}\mathrm{Pb}/^{206}\mathrm{Pb}}\right) \tag{3}$$

may be used to assess the likely skew in the data, which may arise due to some disturbance to the isotope ratios, as mentioned above.

Throughout this paper we stated we generally seek to use heavy-tailed component densities. We believe that this is appropriate in cases for which the measurement error is quite large (as for the examples in this article). When the measurement error is not too large, the approach of Sambridge and Compston (1994) is likely to perform quite well. The determination of measurement error in geochronology is somewhat subjective and beyond the scope of this paper (see Ireland and Williams 2003 for discussion). In general, the raw data can require corrections for variations in isotope ratios over time, inherited daughter element (e.g., common Pb), uncertainties in standards, and decay constants. Typically, errors are combined in quadrature (adding the variances) to produce an estimate of the error on the final age. Overall, experience with our method with respect to the measurement error shows that inferences can (as expected) become less exact when the measurement error increases, that is, that credible intervals can be wider and classification probabilities less discriminated between components. Clearly, when making inferences as described here, it is important to understand the measurement error aspect of a data set, which will differ for different analytical methods (e.g., using $^{206}\mathrm{Pb}/^{238}\mathrm{U}$ and $^{207}\mathrm{Pb}/^{206}\mathrm{Pb}$ ages). In this paper, we have dealt with making inference from the observed data as reported, under the assumption that the data are collected in an unbiased manner. An issue which we have not explicitly considered is the number of data (i.e., analyses) required to resolve an age component present in a given proportion. This is potentially important for studies of sediment provenance and stratigraphic dating. This aspect has been addressed in recent papers by Vermeesch (2004) and Andersen (2005), who provide some guidelines for the sampling and reporting strategies in this context.

We end with a short summary of what our methodology adds to the widely used approach outlined by Sambridge and Compston (1994). The main advantages are:

(A) Our method is more flexible, since we are able to use any component density to model the data.
(B) Our method provides a formal way to construct a density estimate of the data.
(C) Our method allows us to explicitly incorporate prior information (e.g., skewness, or the likely number of components. into the data analysis.

(D) Our method provides a more satisfactory way to interpret model parameters (e.g., age modes, number of components).

We have demonstrated (A) and (B) in examples I and II. Point (C), may be of more interest to geologists who have a reasonable idea of what they expect from their data, but would perhaps find it difficult to specify in practice. Point (D) relates to the fact that all inferences under Sambridge and Compston's approach are related to the *estimated data* and not the actual data, had we been able to obtain these. This, perhaps, appears only of interest from a conceptual point of view, but helps to clarify that our approach is a formal statistical procedure. We feel that the only drawback of using our methodology are the computational aspects (label switching, prior sensitivity) associated with Bayesian mixture modelling.

ACKNOWLEDGMENTS

The first author was supported by an Engineering and Physical Sciences Research Council Studentship. We would like to thank Andy Carter for providing the data and Trevor Ireland for his insight into U-Pb dating.

REFERENCES

Andersen, T., 2005, Detrital zircons as tracers of sedimentary provenance: Limiting conditions from statistics and numerical simulation: Chem. Geol., v. 216, p. 249–270.

Besag, J., 1986, On the statistical analysis of dirty pictures: J. R. Stat. Soc. B, v. 48, no. 3, p. 259–279.

Brandon, M. T., 1992, Decomposition of fission-track age distributions: Am. J. Sci. v. 292, p. 535–564.

Brandon, M. T., 1996, Probability density plot for fission-track grain-age samples: Rad. Meas., v. 26, no. 5, p. 663–676.

Brooks, S. P., Friel, N., and King, R., 2003, Classical model selection via simulated annealing: J. R. Stat. Soc. B, v. 65, no. 2, p. 503–520.

Carter, A., and Bristow, C. S., 2003, Linking hinterland evolution and continental basin sedimentation by using detrital zircon thermochronology: A study of the Khorat Plateau basin, Eastern Thailand: Basin Res., v. 15, no. 2, p. 271–285.

Carter, A., and Moss, S. J., 1999, Combined detrital-zircon fission track and U-Pb dating: A new approach to understanding hinterland evolution: Geology, v. 27, no. 3, p. 235–238.

Galbraith, R. F., 1998, Graphical display of estimates having differing standard errors: Technometrics, v. 30, p. 271–281.

Galbraith, R. F., 1998, The trouble with "probability density" plots of fission track ages: Radiat. Meas., v. 29, no. 2., p. 125–131.

Galbraith, R. F., and Green, P. F., 1990, Estimating the component ages in a finite mixture: Nucl. Track Radiat. Meas., v. 1, no. 3, p. 197–206.

Green, P. J., 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination: Biometrika, v. 82, no. 4, p. 711–732.

Green, P. J., and Mira, A., 2001, Delayed rejection in reversible jump Metropolis–Hastings: Biometrika, v. 88, no. 4, p. 1035–1053.
Ireland, T. R., Flöttmann, T., Fanning, C. M., Gibson, G. M., and Preiss, W. V., 1998, Development of the early Paleozoic Pacific margin of Gondwana from detrital-zircon ages across the delamerian orogen: Geology, v. 26, no. 3, p. 243–246.
Ireland, T. R., and Williams, I. S. 2003, Considerations in zircon geochronology by SIMS, *in* Hanchar, J. M., and Hoskin, P. W. O., eds., Zircon: Reviews in mineralogy and geochemistry: Mineralogical Society of America, Washington, DC, v. 53, p. 215–241.
Jasra, A., Holmes, C. C., and Stephens, D. A., 2005, Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling: Stat. Sci., v. 20, no. 1, p. 50–67.
Jennison, C., 1997, Discussion of on Bayesian analysis of mixtures with an unknown number of components: J. R. Stat. Soc. B., v. 59, no. 4, p. 778–779.
Jones, M. C., and Faddy, M. J., 2003, A skew extension of the *t*-distribution, with applications: J. R. Stat. Soc. B., v. 65, no. 1, p. 159–174.
Lindley, D. V., 1957, A statistical paradox: Biometrika, v. 44, no. 1, p. 187–192.
McLachlan, G. J., and Peel, D., 2000, Finite mixture models: Wiley, Chichester, UK, 419 p.
Richardson, S., and Green, P. J., 1997, On Bayesian analysis of mixture models with an unknown number of components: J. R. Stat. Soc. B, v. 59, no. 4, p. 731–792.
Richardson, S., Leblond, L., Jaussent, I., and Green, P. J., 2002, Mixture models in measurement error problems, with reference to epidemiological studies: J. R. Stat. Soc. A, v. 165, no. 3, p. 549–566.
Robert, C. P., 2001, The Bayesian choice: From decision-theoretic foundations to computational implementation, 2nd ed.: Springer, New York, 604 p.
Robert, C. P., and Casella, G., 2004, Monte Carlo statistical methods, 2nd ed.: Springer, New York, 645 p.
Sambridge, M. S., and Compston, W., 1994, Mixture modelling of multi-component data sets with application to ion-probe zircon ages: Earth Planet. Sci. Lett., v. 128, no. 3, p. 373–390.
Sircombe, K. N., 2004, AGEDISPLAY: An EXCEL workbook to evaluate and display univariate geochronological data using binned frequency histograms and probability density distributions: Comp. Geosci., v. 30, p. 21–31.
Stephens, M., 2000, Dealing with label switching in mixture models: J. R. Stat. Soc. B, v. 62, no. 4, p. 795–809.
Stern, R. A., and Amelin, Y., 2003, Assessment of errors in SIMS zircon U-Pb geochronology using a natural zircon standard and NIST SRM 610 glass: Chem. Geol., v. 197, no. 1, p. 111–142.
Vermeesch, P., 2004. How many grains are needed for a provenance study?: Earth. Planet. Sci. Lett., v. 224. p. 441–451.

APPENDIX A: BAYESIAN STATISTICS

This appendix features a short introduction to Bayesian statistics; see Robert (2001) for a thorough account.

Consider a typical statistical problem where we have observed data $x = (x_1, \ldots, x_n)$ assumed to be drawn from some probability density $p(x|\theta)$, where θ is an unknown real-valued (vector) parameter (for example the mean rock age). Bayesian statistics treats θ as a random variable and places a *prior* density $\pi(\theta)$

upon θ. All inferences on θ are then via the *posterior* density $\pi(\theta|x)$:

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta)$$

for example, to estimate θ we may use the posterior mean.

The prior $\pi(\theta)$ encapsulates the beliefs that we have before seeing the data, and the posterior can be thought of how those beliefs are changed given the data. However, there is not often substantial information to construct a prior. As a result, we often use a prior which minimally influences our posterior inferences.

APPENDIX B: MARKOV CHAIN MONTE CARLO

Introduction

In this appendix we briefly describe MCMC. Suppose that we wish to simulate from a highly complex multivariate probability density $\pi(x),\ x \in E$ and that conventional methods are not possible. Then the idea of MCMC is to construct an *ergodic Markov chain with stationary distribution* $\pi(\cdot)$ (Robert and Casella, 2004).

Metropolis–Hastings and the Gibbs Sampler

The Metropolis-Hastings chain is as follows. Suppose that the current state of the chain is x, and we wish to generate a new state so that in the long run our samples are drawn from π; this is achieved in the following way. Generate $X' \sim q$ where q is a *proposal* density, such that $E \subseteq \cup_{x \in E} \text{supp}[q(x, \cdot)]$, then accept x' as the new value of the chain with probability $\min\{1, A\}$ where

$$A = \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}.$$

An easy way to construct a proposal that will be accepted reasonably often is the random walk move. That is, the additive and multiplicative random walks are (for some specified σ)

$$x' = x + \sigma u$$
$$x' = x\sigma u$$

where $u \sim f$ is a density chosen to ensure the proposal gives full support to the target.

Typically $x = (x_1, \ldots, x_k)$, and it will often be difficult to find proposals over large dimensional spaces to yield a reasonable acceptance rate. One way to avoid such a difficulty is blocking the variables.

Suppose we choose any combination of the $x_1, \ldots, x_k$ e.g. $x_1, \ldots, x_l$ and $x_{l+1}, \ldots, x_k$, then a valid way to simulate from $\pi(\cdot)$ is to generate from the *full conditionals* $\pi(x_1, \ldots, x_l \,|x_{l+1}, \ldots, x_k)$ and $\pi(x_{l+1}, \ldots, x_k|x_1, \ldots, x_l)$. If we can generate exactly from the full conditional distributions then this is called the *Gibbs sampler*, otherwise it is valid to use Metropolis–Hastings.

Reversible Jump MCMC

Suppose that the support π is of variable dimension, (k is unknown) then none of the above methods can be used. One solution to this problem is that of reversible jump MCMC Green (1995).

To move from x_k in k dimensional space, propose to move to $k+1$ dimensional space in the following way. Generate $U \sim q$ from some density such that $k + \dim(u) = k + 1$ where $\dim(u)$ is the dimension of u. Then set $x_{k+1} = g(x_k, u)$, where $g(\cdot)$ is the *jump function* (it has to be invertible and differentiable). We suppose that the probability of proposing to jump up is $r_{k,k+1}(x_k)$ and that the probability of proposing the *reverse move* is $r_{k+1,k}(x_{k+1})$. Then x_{k+1} is accepted with probability $\min\{1, A\}$, where

$$A = \frac{\pi(x_{k+1})}{\pi(x_k)} \frac{r_{k+1,k}(x_{k+1})}{r_{k,k+1}(x_k)q(u)} \left| \frac{\partial g(x_k, u)}{\partial(x_k, u)} \right|.$$

The reverse move from x_{k+1} to x_k occurs in the following way. Solve for u and x_k from the jump function, via the inverse of $g(\cdot)$ and accept x_k as the new state of the chain with probability $\min\{1, A^{-1}\}$.

APPENDIX C: RELABELLING ALGORITHM

We briefly describe our relabelling algorithm, which is based upon the method of Stephens (2000).

Using a short run of samples, for which no label switching is deemed to have occurred (which we determine by eye), calculate the empirical classification probabilities $\hat{r}_{ij}$ (as in Eq. (2)). Now at sweep t of our MCMC algorithm, given

that $k^{(t)} = k^*$, choose the permutation of the labelling σ_t to minimize:

$$L(R, S) = \sum_{i=1}^{n} \sum_{j=1}^{k^*} s_{ij}^{(t)} \left\{ \sigma_t \left(\theta^{(t)} \right) \right\} \log \left[\frac{s_{ij}^{(t)} \{ \sigma_t (\theta^{(t)}) \}}{\hat{r}_{ij}} \right]$$

where

$$s_{ij}^{(t)}(\boldsymbol{\theta}^{(t)}) = \frac{w_j^{(t)} f\left(y_i^{(t)}; \phi_j^{(t)}\right)}{\sum_{l=1}^{k^*} w_l^{(t)} f\left(y_i^{(t)}; \phi_l^{(t)}\right)}$$

and R and S represent the $n \times k^*$ matrices (r_{ij}) and (s_{ij}), respectively. The algorithm is performed online and the $\hat{r}_{ij}$ are never updated; see Stephens (2000) for discussion of this method.