Bayesian Analysis of Mixture Models with an Unknown Number of Components — an alternative to reversible jump methods

Matthew Stephens *
Department of Statistics
University of Oxford

Submitted to Annals of Statistics, December 1998
Revised August 1999

Running Head: BAYESIAN ANALYSIS OF MIXTURES

*Address for Correspondence: Department of Statistics, 1, South Parks Road, Oxford, OX1 3TG. email: stephens@stats.ox.ac.uk

Abstract

Richardson and Green (1997) present a method of performing a Bayesian analysis of data from a finite mixture distribution with an unknown number of components. Their method is a Markov Chain Monte Carlo (MCMC) approach, which makes use of the "reversible jump" methodology described by Green (1995). We describe an alternative MCMC method which views the parameters of the model as a (marked) point process, extending methods suggested by Ripley (1977) to create a Markov birth-death process with an appropriate stationary distribution. Our method is easy to implement, even in the case of data in more than one dimension, and we illustrate it on both univariate and bivariate data. There appears to be considerable potential for applying these ideas to other contexts, as an alternative to more general reversible jump methods, and we conclude with a brief discussion of how this might be achieved.

Keywords: Bayesian analysis, Birth-death process, Markov process, MCMC, Mixture model, Model Choice, Reversible Jump, Spatial point process

1 Introduction

Finite mixture models are typically used to model data where each observation is assumed to have arisen from one of k groups, each group being suitably modelled by a density from some parametric family. The density of each group is referred to as a *component* of the mixture, and is weighted by the relative frequency of the group in the population. This model provides a framework by which observations may be clustered together into groups for discrimination or classification (see for example McLachlan and Basford, 1988). For a comprehensive list of such applications see Titterington *et al.* (1985). Mixture models also provide a convenient and flexible family of distributions for estimating or approximating distributions which are not well modelled by any standard parametric family, and provide a parametric alternative to non-parametric methods of density estimation, such as kernel density estimation. See for example Roeder (1990), West (1993) and Priebe (1994).

This paper is principally concerned with the analysis of mixture models in which the number of components k is unknown. In applications where the components have a physical interpretation, inference for k may be of interest in itself. Where the mixture model is being used purely as a parametric alternative to non-parametric density estimation, the value of k chosen affects the flexibility of the model and thus the smoothness of the resulting density estimate. Inference for k may then be seen as analogous to bandwidth selection in kernel density estimation. Procedures which allow k to vary may therefore be of interest whether or not k has a physical interpretation.

Inference for k may be seen as a specific example of the very common problem of choosing a model from a given set of competing models. Taking a Bayesian approach to this problem, as we do here, has the advantage that it provides not only a way of selecting a single "best" model, but also a coherent way of combining results over different models. In the mixture model context this might include performing density estimation by taking an appropriate average of density estimates obtained using different values of k. While model choice (and model averaging) within the Bayesian framework are both theoretically straightforward, they often provide a computational challenge, particularly when (as here) the competing models are of differing dimension. The use of Markov Chain Monte Carlo (MCMC) methods (see Gilks *et al.*, 1996, for an introduction) to perform Bayesian analysis is now very common, but MCMC methods which are able to jump between models of differing dimension have become popular only recently, in particular through the use of the "reversible jump" methodology developed by Green (1995). Reversible jump methods allow the construction of an ergodic Markov chain with the joint posterior distribution of the parameters and the model as its stationary distribution. Moves between models are achieved by periodically proposing a move to a different model, and rejecting it with appropriate probability to ensure that the chain possesses the required stationary distribution. Ideally these proposed moves are designed to have a high probability of acceptance so that the algorithm explores the different models adequately, though this is not always easy to achieve in practice. As usual in MCMC methods, quantities of interest may be estimated by forming sample path averages over simulated realizations of this Markov chain. The reversible jump methodology has now been applied to a wide range of model choice problems, including change point analysis (Green, 1995), Quantitative Trait Locus analysis (Stephens and Fisch, 1998), and mixture models (Richardson and Green, 1997).

In this paper we present an alternative method of constructing an ergodic Markov chain with appropriate stationary distribution, when the number of components k is considered unknown. The method is based on the construction of a continuous time Markov birth-death process (as described by Preston, 1976) with the appropriate stationary distribution. MCMC methods based on these (and related) processes have been used extensively in the point process literature to simulate realizations of point processes which are difficult to simulate from directly; an idea which originated with Kelly

and Ripley (1976) and Ripley (1977) (see also Glötzl, 1981; Stoyan *et al.*, 1987). These realizations can then be used for significance testing (as in Ripley, 1977), or likelihood inference for the parameters of the model (see for example Geyer and Møller, 1994, and references therein). More recently such MCMC methods have been used to perform Bayesian inference for the *parameters* of a point process model, where the parameters themselves are (modelled by) a point process (see for example Baddeley and van Lieshout, 1993; Lawson, 1996).

In order to apply these MCMC methods to the mixture model context, we view the parameters of the model as a (marked) point process, with each point representing a component of the mixture. The MCMC scheme allows the number of components to vary by allowing new components to be "born" and existing components to "die". These births and deaths occur in continuous time, and the relative rates at which they occur determine the stationary distribution of the process. The relationship between these rates and the stationary distribution is formalised in Section 3 (Theorem 3.1). We then use this to construct an easily simulated process, in which births occur at a constant rate from the prior, and deaths occur at a rate which is very low for components which are critical in explaining the data, and very high for components which do not help explain the data. The accept-reject mechanism of reversible jump is thus replaced by a mechanism which allows both "good" and "bad" births to occur, but reverses bad births very quickly through a very quick death.

Our method is illustrated in Section 4, by fitting mixtures of normal (and t) distributions to univariate and bivariate data. We found that the posterior distribution of the number of components for a given data set typically depends heavily on modelling assumptions such as the form of the distribution for the components (normals or ts) and the priors used for the parameters of these distributions. In contrast, predictive density estimates tend to be relatively insensitive to these modelling assumptions. Our method appears to have similar computational expense to that of Richardson and Green (1997) in the context of mixtures of univariate normal distributions, though direct comparisons are difficult. Both methods certainly give computationally tractable solutions to the problem, with rough results available in a matter of minutes. However, our approach appears the more natural and elegant in this context, exploiting the natural nested structure of the models and exchangeability of the mixture components. As a result we remove the need for calculation of a complicated Jacobian, reducing the potential for making algebraic errors. In addition, the changes necessary to explore alternative models for the mixture components (replacing normals with t distributions for example) are trivial.

We conclude with a discussion of the potential for extending the birth-death methodology (BDMCMC) to other contexts, as an alternative to more general reversible jump (RJMCMC) methods. One interpretation of BDMCMC is as a continuous-time version of RJMCMC, with a limit on the types of moves which are permitted in order to simplify implementation. BDMCMC is easily applied to any context where the parameters of interest may be viewed as a point process, and where the likelihood of these parameters may be explicitly calculated (this latter rules out Hidden Markov Models for example). We consider briefly some examples (a multiple change-point problem, and variable selection in regression models) where these conditions are fulfilled, and discuss the difficulties of designing suitable birth-death moves. Where such moves are sufficient to achieve adequate mixing BDMCMC provides an attractive easily-implemented alternative to more general RJMCMC schemes.

2 Bayesian methods for mixtures

2.1 Notation and missing data formulation

We consider a finite mixture model in which data $x^n = x_1, \ldots, x_n$ are assumed to be independent observations from a mixture density with k (k possibly unknown but finite) components:

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \cdots + \pi_k f(x; \phi_k, \eta), \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ are the *mixture proportions* which are constrained to be non-negative and sum to unity; $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ are the (possibly vector) *component specific* parameters, with ϕ_i being specific to component i; and η is a (possibly vector) *common* parameter which is common to all components. Throughout this paper $p(\cdot \mid \cdot)$ will be used to denote both conditional densities and distributions.

It is convenient to introduce the *missing data* formulation of the model, in which each observation x_j is assumed to arise from a specific but unknown component z_j of the mixture. The model (1) can be written in terms of the missing data, with $z_1, \ldots, z_n$ assumed to be realizations of independent and identically distributed discrete random variables $Z_1, \ldots, Z_n$ with probability mass function

$$\Pr(Z_j = i \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_i \quad (j = 1, \ldots, n; \quad i = 1, \ldots, k). \tag{2}$$

Conditional on the Zs, $x_1, \ldots, x_n$ are assumed to be independent observations from the densities

$$p(x_j \mid Z_j = i, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = f(x_j; \phi_i, \eta) \quad (j = 1, \ldots, n). \tag{3}$$

Integrating out the missing data $Z_1, \ldots, Z_n$ then yields the model (1).

2.2 Hierarchical model

We assume a hierarchical model for the prior on the parameters $(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$, with $(\pi_1, \phi_1), \ldots, (\pi_k, \phi_k)$ being exchangeable. (For an alternative approach see Escobar and West (1995) who use a prior structure based on the Dirichlet process.) Specifically we assume that the prior distribution for $(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ given *hyperparameters* ω, and common component parameters η, has Radon–Nikodym derivative ("density") $r(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid \omega, \eta)$ with respect to an underlying symmetric measure $\mathcal{M}$ (defined below). For notational convenience we drop for the rest of the paper the explicit dependence of $r(\cdot \mid \omega, \eta)$ on ω and η. To ensure exchangeability we require that, for any given k, $r(\cdot)$ is invariant under relabelling of the components, in that

$$r\big(k, (\pi_1, \ldots, \pi_k), (\phi_1, \ldots, \phi_k)\big) = r\big(k, (\pi_{\epsilon(1)}, \ldots, \pi_{\epsilon(k)}), (\phi_{\epsilon(1)}, \ldots, \phi_{\epsilon(k)})\big) \tag{4}$$

for all permutations ϵ of $1, \ldots, k$.

In order to define the symmetric measure $\mathcal{M}$ we introduce some notation. Let $\mathcal{U}^{k-1}$ denote the Uniform distribution on the simplex

$$\mathcal{S}^{k-1} = \{(\pi_1, \ldots, \pi_{k-1}) : \pi_1, \ldots, \pi_{k-1} \geq 0 \cap \pi_1 + \cdots + \pi_{k-1} \leq 1\}.$$

Let Φ denote the parameter space for the ϕ_i (so $\phi_i \in \Phi$ for all i), let ν be some measure on Φ, and let ν^k be the induced product measure on Φ^k. (For most of this paper Φ will be R^m for some m, and ν can be assumed to be Lebesgue measure.) Now let $\mathcal{M}_k$ be the product measure $\nu^k \times \mathcal{U}^{k-1}$ on $\Phi^k \times \mathcal{S}^{k-1}$, and finally define $\mathcal{M}$ to be the induced measure on the disjoint union $\cup_{k=1}^{\infty}(\Phi^k \times \mathcal{S}^{k-1})$.

A special case

Given ω and η, let k have prior probability mass distribution $p(k \mid \omega, \eta)$. Suppose $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ are *a priori* independent given k, ω and η, with $\phi_1, \ldots, \phi_k$ being independent and identically distributed from a distribution with density $\widetilde{p}(\phi \mid \omega, \eta)$ with respect to ν, and $\boldsymbol{\pi}$ having a uniform distribution on the simplex $\mathcal{S}^{k-1}$. Then

$$r(k, \boldsymbol{\phi}, \boldsymbol{\pi}) = p(k \mid \omega, \eta)\widetilde{p}(\phi_1 \mid \omega, \eta) \ldots \widetilde{p}(\phi_k \mid \omega, \eta). \tag{5}$$

Note that this special case includes the specific models used by Diebolt and Robert (1994) and Richardson and Green (1997) in the context of mixtures of univariate normal distributions.

2.3 Bayesian inference *via* MCMC

Given data x^n, Bayesian inference may be performed using MCMC methods, which involve the construction of a Markov chain $\{\Theta^{(t)}\}$ with the posterior distribution $p(\theta \mid x^n)$ of the parameters $\theta = (k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ as its stationary distribution. Given suitable regularity conditions (see for example Tierney, 1996, p.65), quantities of interest may be consistently estimated by sample path averages. For example, if $\theta^{(0)}, \theta^{(1)}, \ldots$ is a sampled realization of such a Markov chain, then inference for k may be based on an estimate of the marginal posterior distribution

$$\begin{aligned} \Pr(k = i \mid x^n) &= \lim_{N \to \infty} \frac{1}{N} \#\{t : k^{(t)} = i\} \\ &\approx \frac{1}{N} \#\{t : k^{(t)} = i\} \quad (N \text{ large}), \end{aligned} \tag{6}$$

and similarly the predictive density for a future observation may be estimated by

$$p(x_{n+1} \mid x^n) \approx \frac{1}{N} \sum_{t=1}^{N} p(x_{n+1} \mid \theta^{(t)}). \tag{7}$$

More details, including details of the construction of a suitable Markov chain when k is fixed, can be found in the paper by Diebolt and Robert (1994), chapters of the books by Robert (1994) and Gelman *et al.* (1995), and the article by Robert (1996). Richardson and Green (1997) describe the construction of a suitable Markov chain when k is allowed to vary using the reversible jump methodology developed by Green (1995). We now describe an alternative approach.

3 Constructing a Markov chain *via* simulation of point processes

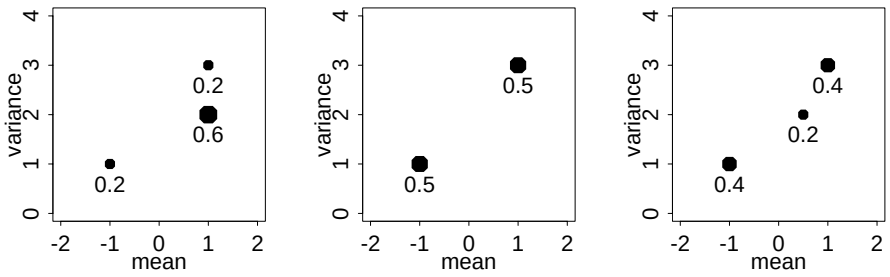
3.1 The parameters as a point process

Our strategy is to view each component of the mixture as a point in parameter space, and adapt theory from the simulation of point processes to help construct a Markov chain with the posterior distribution of the parameters as its stationary distribution. Since, for given k, the prior distribution for $(\boldsymbol{\pi}, \boldsymbol{\phi})$ defined at (4) does not depend on the labelling of the components, and the likelihood

$$L(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = p(x^n \mid k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \prod_{j=1}^{n} [\pi_1 f(x_j; \phi_1, \eta) + \cdots + \pi_k f(x_j; \phi_k, \eta)] \tag{8}$$

is also invariant under permutations of the components labels, the posterior distribution

$$p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid x^n, \omega, \eta) \propto L(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) r(k, \boldsymbol{\pi}, \boldsymbol{\phi}) \tag{9}$$



(a) (b) (c)

Figure 1: Illustration of births and deaths as defined by (10) and (11). **a)** Representation of $0.2\mathcal{N}(-1,1) + 0.6\mathcal{N}(1,2) + 0.2\mathcal{N}(1,3)$ as a set of points in parameter space. $\mathcal{N}(\mu,\sigma^2)$ denotes the univariate normal distribution with mean μ and variance σ^2. **b)** Resulting model after death of component $0.6\mathcal{N}(1,2)$ in a). **c)** Resulting model after birth of component at $0.2\mathcal{N}(0.5,2)$ in b).

will be similarly invariant. Fixing ω and η, we can thus ignore the labelling of the components and can consider any set of k parameter values $\{(\pi_1,\phi_1),\ldots,(\pi_k,\phi_k)\}$ as a set of k points in $[0,1]\times\Phi$, with the constraint that $\pi_1 + \cdots + \pi_k = 1$ (see, for example, Figure 1a.) The posterior distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi}\,|\,x^n,\omega,\eta)$ can then be seen as a (suitably constrained) distribution of points in $[0,1]\times\Phi$, or in other words a *point process* on $[0,1]\times\Phi$. Equivalently the posterior distribution can be seen as a *marked point process* in Φ, with each point ϕ_i having an associated *mark* $\pi_i \in [0,1]$, with the marks being constrained to sum to unity.

This view of the parameters as a marked point process (which is also outlined by Dawid, 1997) allows us to use methods similar to those in Ripley (1977) to construct a continuous time Markov birth-death process with stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi}\,|\,x^n,\omega,\eta)$, with ω and η kept fixed. Details of this construction are given in the next section. In Section 3.4 we combine this process with standard (fixed-dimension) MCMC update steps which allow ω and η to vary, to create a Markov chain with stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi},\omega,\eta\,|\,x^n)$.

3.2 Birth-death processes for the components of a mixture model

Let Ω_k denote the parameter space of the mixture model with k components, ignoring the labelling of the components, and let $\Omega = \bigcup_{k\geq 1}\Omega_k$. We will use set notation to refer to members of Ω, writing $y = \{(\pi_1,\phi_1),\ldots,(\pi_k,\phi_k)\} \in \Omega_k$ to represent the parameters of the model (1) keeping η fixed, and so we may write $(\pi_i,\phi_i) \in y$ for $i = 1,\ldots,k$. Note that (for given ω and η) the invariance of $L(\cdot)$ and $r(\cdot)$ under permutation of the component labels allows us to define $L(y)$ and $r(y)$ in an obvious way.

We define births and deaths on Ω as follows:

Births: If at time t our process is at $y = \{(\pi_1,\phi_1),\ldots,(\pi_k,\phi_k)\} \in \Omega_k$ and a birth is said to occur at $(\pi,\phi) \in [0,1]\times\Phi$, then the process jumps to

$$y \cup (\pi,\phi) := \Big\{(\pi_1(1-\pi),\phi_1),\ldots,(\pi_k(1-\pi),\phi_k),(\pi,\phi)\Big\} \in \Omega_{k+1}. \qquad (10)$$

Deaths: If at time t our process is at $y = \{(\pi_1,\phi_1),\ldots,(\pi_k,\phi_k)\} \in \Omega_k$ and a death is said to

occur at $(\pi_i, \phi_i) \in y$, then the process jumps to

$$y\backslash(\pi_i,\phi_i) := \Big\{ \Big(\frac{\pi_1}{1-\pi_i},\phi_1\Big),\ldots,\Big(\frac{\pi_{i-1}}{1-\pi_i},\phi_{i-1}\Big), \Big(\frac{\pi_{i+1}}{1-\pi_i},\phi_{i+1}\Big),\ldots,\Big(\frac{\pi_k}{1-\pi_i},\phi_k\Big)\Big\} \in \Omega_{k-1}. \quad (11)$$

Thus a birth increases the number of components by one, while a death decreases the number of components by one. These definitions have been chosen so that births and deaths are inverse operations to each other, and the constraint $\pi_1 + \cdots + \pi_k = 1$ remains satisfied after a birth or death; they are illustrated in Figure 1. With births and deaths thus defined, we consider the following continuous time Markov birth-death process:

When the process is at $y \in \Omega_k$, let births and deaths occur as independent Poisson processes as follows:

Births: Births occur at overall rate $\beta(y)$, and when a birth occurs it occurs at a point $(\pi, \phi) \in [0,1] \times \Phi$, chosen according to density $b\big(y;(\pi,\phi)\big)$ with respect to the product measure $\mathcal{U}^1 \times \nu$, where $\mathcal{U}^1$ is the uniform (Lebesgue) measure on $[0,1]$.

Deaths: When the process is at $y = \big\{(\pi_1,\phi_1),\ldots,(\pi_k,\phi_k)\big\}$, each point (π_j,ϕ_j) dies independently of the others as a Poisson process with rate

$$\delta_j(y) = d\big(y\backslash(\pi_j,\phi_j)\,;(\pi_j,\phi_j)\big) \quad (12)$$

for some $d : \Omega \times ([0,1]\times\Phi) \to R^+$. The overall death rate is then given by $\delta(y) = \sum_j \delta_j(y)$.

The time to the next birth/death event is then exponentially distributed, with mean $1/\ \big(\beta(y)+\delta(y)\big)$, and it will be a birth with probability $\beta(y)/(\beta(y)+\delta(y))$, and a death of component j with probability $\delta_j(y)/(\beta(y)+\delta(y))$. In order to ensure that the birth-death process doesn't jump to an area with zero "density" we impose the following conditions on b and d:

$$b\big(y;(\pi,\phi)\big) = 0 \text{ whenever } r\big(y\cup(\pi,\phi)\big)L\big(y\cup(\pi,\phi)\big) = 0, \quad (13)$$

$$d\big(y;(\pi,\phi)\big) = 0 \text{ whenever } r(y)L(y) = 0. \quad (14)$$

The following Theorem then gives sufficient conditions on b and d for this process to have stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi} \mid x^{\,n},\omega,\eta)$.

Theorem 3.1. *Assuming the general hierarchical prior on $(k,\boldsymbol{\pi},\boldsymbol{\phi})$ given in Section 2.2, and keeping ω and η fixed, the birth-death process defined above has stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi} \mid x^{\,n},\omega,\eta)$, provided b and d satisfy*

$$(k+1)\,d\big(y;(\pi,\phi)\big)\,r\big(y\cup(\pi,\phi)\big)\,L\big(y\cup(\pi,\phi)\big)\,k(1-\pi)^{k-1} = \beta(y)\,b\big(y;(\pi,\phi)\big)\,r(y)\,L(y) \quad (15)$$

for all $y \in \Omega_k$ and $(\pi,\phi) \in [0,1]\times\Phi$.

Proof. The proof is deferred to the appendix (Section 7). □

3.3 Naive algorithm for a special case

We now consider the special case described at (5), where

$$r(y) = p(k \,|\, \omega, \eta)\widetilde{p}(\phi_1 \,|\, \omega, \eta) \ldots \widetilde{p}(\phi_k \,|\, \omega, \eta). \tag{16}$$

Suppose that we can simulate from $\widetilde{p}(\cdot \,|\, \omega, \eta)$, and consider the process obtained by setting $\beta(y) = \lambda_b$ (a constant), with $b\big(y; (\pi, \phi)\big) = k(1-\pi)^{k-1}\widetilde{p}(\phi \,|\, \omega, \eta)$. Applying Theorem 3.1 we find that the process has the correct stationary distribution, provided that when the process is at $y = \big\{(\pi_1, \phi_1), \ldots, (\pi_k, \phi_k)\big\}$, each point (π_j, ϕ_j) dies independently of the others as a Poisson process with rate

$$d\big(y \backslash (\pi_j, \phi_j)\,; (\pi_j, \phi_j)\big) = \lambda_b \frac{L\big(y \backslash (\pi_j, \phi_j)\big)}{L(y)} \frac{p(k-1 \,|\, \omega, \eta)}{k p(k \,|\, \omega, \eta)} \quad (j = 1, \ldots, k). \tag{17}$$

Algorithm 3.1 below simulates this process. We note that the algorithm is very straightforward to implement, requiring *only the ability to simulate from $\widetilde{p}(\cdot \,|\, \omega, \eta)$, and to calculate the model likelihood for any given model.* The main computational burden is in calculating the likelihood, and it is important that calculations of densities are stored and reused where possible.

Algorithm 3.1. To simulate a process with appropriate stationary distribution.

Starting with initial model $y = \big\{(\pi_1, \phi_1), \ldots, (\pi_k, \phi_k)\big\} \in \Omega_k$, iterate the following steps:

1. Let the birth rate $\beta(y) = \lambda_b$.
2. Calculate the death rate for each component, the death rate for component j being given by (17):

$$\delta_j(y) = \lambda_b \frac{L\big(y \backslash (\pi_j, \phi_j)\big)}{L(y)} \frac{p(k-1 \,|\, \omega, \eta)}{k p(k \,|\, \omega, \eta)} \quad (j = 1, \ldots, k). \tag{18}$$

3. Calculate the total death rate $\delta(y) = \sum_j \delta_j(y)$.
4. Simulate the time to the next jump from an exponential distribution with mean $1/\big(\beta(y) + \delta(y)\big)$.
5. Simulate the type of jump: birth or death with respective probabilities

$$\Pr(\text{birth}) = \frac{\beta(y)}{\beta(y) + \delta(y)}, \qquad \Pr(\text{death}) = \frac{\delta(y)}{\beta(y) + \delta(y)}.$$

6. Adjust y to reflect the birth or death (as defined by (10) and (11)):

   Birth: Simulate the point (π, ϕ) at which a birth takes place from the density $b\big(y; (\pi, \phi)\big) = k(1-\pi)^{k-1}\widetilde{p}(\phi \,|\, \omega, \eta)$ by simulating π and ϕ independently from densities $k(1-\pi)^{k-1}$ and $\widetilde{p}(\phi \,|\, \omega, \eta)$ respectively. We note that the former is the Beta distribution with parameters $(1, k)$, which is easily simulated from by simulating $Y_1 \sim \Gamma(1, 1)$ and $Y_2 \sim \Gamma(k, 1)$ and setting $\pi = Y_1/(Y_1 + Y_2)$, where $\Gamma(n, \lambda)$ denotes the Gamma distribution with mean n/λ.

   Death: Select a component to die: $(\pi_j, \phi_j) \in y$ being selected with probability $\delta_j(y)/\delta(y)$ for $j = 1, \ldots, k$.

7. Return to step 2.

Remark 3.2. Algorithm 3.1 seems rather naive in that births occur (in some sense) from the prior, which may lead to many births of components which do not help to explain the data. Such components will have a high death rate (17) and so will die very quickly, which is inefficient in the same way as an accept-reject simulation algorithm is inefficient if many samples are rejected. However, in the examples we consider in the next section this naive algorithm performs reasonably well, and so we have not considered any cleverer choices of $b\left(y;(\pi,\phi)\right)$ which may allow births to occur in a less naive way (see Section 5.2 for further discussion).

3.4 Constructing a Markov Chain

If we fix ω and η then Algorithm 3.1 simulates a birth-death process with stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi}\,|\,x^n,\omega,\eta)$. This can be combined with MCMC update steps which allow ω and η to vary to create a Markov chain with stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi},\omega,\eta\,|\,x^n)$. By augmenting the data x^n by the missing data $z^n=(z_1,\ldots,z_n)$ described in Section 2.1, and assuming the existence and use of the necessary conjugate priors, we can use Gibbs sampling steps to achieve this as in Algorithm 3.2 below; Metropolis–Hastings updates could also be used, removing the need to introduce the missing data or use conjugate priors.

Algorithm 3.2. To simulate a Markov chain with appropriate stationary distribution.

Given the state $\Theta^{(t)}=\theta^{(t)}$ at time t, simulate a value for $\Theta^{(t+1)}=\theta^{(t+1)}$ as follows:

Step 1: Sample $(k^{(t)'},\boldsymbol{\pi}^{(t)'},\boldsymbol{\phi}^{(t)'})$ by running the birth-death process for a fixed time t_0, starting from $(k^{(t)},\boldsymbol{\pi}^{(t)},\boldsymbol{\phi}^{(t)})$ and fixing (ω,η) to be $(\omega^{(t)},\eta^{(t)})$. Set $k^{(t+1)}=k^{(t)'}$.

Step 2: Sample $(z^n)^{(t+1)}$ from $p(z^n\,|\,k^{(t+1)},\boldsymbol{\pi}^{(t)'},\boldsymbol{\phi}^{(t)'},\eta^{(t)},\omega^{(t)},x^n)$.

Step 3: Sample $\eta^{(t+1)},\omega^{(t+1)}$ from $p(\eta,\omega\,|\,k^{(t+1)},\boldsymbol{\pi}^{(t)'},\boldsymbol{\phi}^{(t)'},x^n,z^n)$.

Step 4: Sample $\boldsymbol{\pi}^{(t+1)},\boldsymbol{\phi}^{(t+1)}$ from $p(\boldsymbol{\pi},\boldsymbol{\phi}\,|\,k^{(t+1)},\eta^{(t+1)},\omega^{(t+1)},x^n,z^n)$.

Provided the full conditional posterior distributions for each parameter give support to all parts of the parameter space, this will define an irreducible Markov chain with stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi},\omega,\eta,z^n\,|\,x^n)$ suitable for estimating quantities of interest by forming sample path averages as in (6) and (7). The proof is straightforward and is omitted here (see Stephens, 1997, p.84). Step 1 of the algorithm involve movements between different values of k by allowing new components to be "born", and existing components to "die". Steps 2, 3 and 4 allow the parameters to vary with k kept fixed. Step 4 is not strictly necessary to ensure convergence of the Markov chain to the correct stationary distribution, but is included to improve mixing. Note that (as usual in Gibbs sampling) the algorithm remains valid if any or all of ω, η and $\boldsymbol{\phi}$ are partitioned into separate components which are updated one at a time by a Gibbs sampling step, as will be the case in our examples.

4 Examples

Our examples demonstrate the use of Algorithm 3.2 to perform inference in the context of both univariate and bivariate data x^n, which are assumed to be independent observations from a mixture of an unknown (finite) number of normal distributions:

$$p(x\,|\,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})=\pi_1\mathcal{N}_r(x;\mu_1,\Sigma_1)+\cdots+\pi_k\mathcal{N}_r(x;\mu_k,\Sigma_k). \tag{19}$$

Here $\mathcal{N}_r(x;\mu_i,\Sigma_i)$ denotes the density function of the r-dimensional multivariate normal distribution with mean μ_i and variance-covariance matrix Σ_i. In the univariate case ($r=1$) we may write σ^2 for Σ.

Prior distributions

We assume a truncated Poisson prior on the number of components k:

$$p(k) \propto \frac{\lambda^k}{k!} \quad (k = 1, \ldots, k_{\max} = 100), \tag{20}$$

where λ is a constant; we will perform analyses with several different values of λ. Conditional on k we base our prior for the model parameters on the hierarchical prior suggested by Richardson and Green (1997) in the context of mixtures of univariate normal distributions. A natural generalization of their prior to r dimensions is obtained by replacing univariate normal distributions with multivariate normal distributions, and replacing gamma distributions with Wishart distributions, to give

$$\mu_i \sim \mathcal{N}_r(\xi, \kappa^{-1}) \quad (i = 1, \ldots, k) \tag{21}$$

$$\Sigma_i^{-1} \mid \beta \sim \mathcal{W}_r(2\alpha, (2\beta)^{-1}) \quad (i = 1, \ldots, k) \tag{22}$$

$$\beta \sim \mathcal{W}_r(2g, (2h)^{-1}) \tag{23}$$

$$\boldsymbol{\pi} \sim \mathcal{D}(\gamma) \tag{24}$$

where β is a hyperparameter; κ, β and h are $r \times r$ matrices; ξ is an $r \times 1$ vector; α, γ and g are scalars; $\mathcal{D}(\gamma)$ denotes the symmetric Dirichlet distribution with parameter γ and density

$$\frac{\Gamma(k\gamma)}{\Gamma(\gamma)^k} \pi_1^{\gamma-1} \ldots \pi_{k-1}^{\gamma-1} (1 - \pi_1 - \cdots - \pi_{k-1})^{\gamma-1};$$

and $\mathcal{W}_r(m, A)$ denotes the Wishart distribution in r dimensions with parameters m and A. This last is usually introduced as the distribution of the sample covariance matrix, for a sample of size m from a multivariate normal distribution in r dimensions with covariance matrix A. Because of this interpretation m is usually taken as an integer, and for $m \geq r$ $\mathcal{W}_r(m, A)$ has density

$$\mathcal{W}_r(V; m, A) = K |A|^{-\frac{m}{2}} |V|^{\frac{m-r-1}{2}} \exp\left\{-\tfrac{1}{2}\text{tr}(A^{-1}V)\right\} I(V \text{ positive definite}) \tag{25}$$

on the space of all *symmetric* matrices ($\equiv R^{r(r+1)/2}$), where $I(\cdot)$ denotes an indicator function and

$$K^{-1} = 2^{\frac{mr}{2}} \pi^{r(r-1)/4} \prod_{s=1}^{r} \Gamma\left(\frac{m+1-s}{2}\right).$$

However, (25) also defines a density for non-integer m provided $m > r-1$. Methods of simulating from the Wishart distribution (which work for non-integer $m > r-1$) may be found in Ripley (1987). For $m \leq r-1$ we will use $\mathcal{W}_r(m, A)$ to represent the *improper* distribution with density proportional to (25). (This is not the usual definition of $\mathcal{W}_r(m, A)$ for $m \leq r-1$, which is a singular distribution confined to a subspace of symmetric matrices.) Where an improper prior distribution is used, it is important to check the integrability of the posterior.

For univariate data we follow Richardson and Green (1997), who take $(\xi, \kappa, \alpha, g, h, \gamma)$ to be (data-dependent) constants with the following values:

$$\xi = \xi_1 \qquad \kappa = \frac{1}{R_1^2} \qquad \alpha = 2$$
$$g = 0.2 \qquad h = \frac{100g}{\alpha R_1^2} \qquad \gamma = 1$$

where ξ_1 is the midpoint of the observed interval of variation of the data, and R_1 is the length of this interval. The value $\alpha = 2$ was chosen to express the belief that the variances of the components are similar, without restricting them to be equal. For bivariate data ($r = 2$) we felt that a slightly stronger constraint would be appropriate, and so increased α to 3, making a corresponding change in g and obvious generalizations for the other constants to give

$$\xi = (\xi_1, \xi_2) \qquad \kappa = \begin{pmatrix} \frac{1}{R_1^2} & 0 \\ 0 & \frac{1}{R_2^2} \end{pmatrix} \qquad \alpha = 3$$
$$g = 0.3 \qquad h = \begin{pmatrix} \frac{100g}{\alpha R_1^2} & 0 \\ 0 & \frac{100g}{\alpha R_2^2} \end{pmatrix} \qquad \gamma = 1$$

where ξ_1 and ξ_2 are the midpoints of the observed intervals of variation of the data in the first and second dimension respectively, and R_1 and R_2 are the respective lengths of these intervals. We note that the prior on β in the bivariate case

$$\beta \sim \mathcal{W}_2(0.6, (2h)^{-1})$$

is an improper distribution, but careful checking of the necessary integrals shows that the posterior distributions are proper.

In our examples we consider the following priors:

1. The *Fixed-κ* prior, which is the name we give to the prior given above. The full conditional posterior distributions required for the Gibbs sampling updates (Steps 2-4 in Algorithm 3.2) are then (using $\mid \ldots$ to denote conditioning on all other variables)

$$p(z_j = i \mid \cdots) \propto \pi_i \mathcal{N}_r(x_j; \mu_i, \Sigma_i) \tag{26}$$
$$\beta \mid \cdots \sim \mathcal{W}_r\Big(2g + 2k\alpha, \Big[2h + 2\sum_i \Sigma_i^{-1}\Big]^{-1}\Big) \tag{27}$$
$$\boldsymbol{\pi} \mid \cdots \sim \mathcal{D}(\gamma + n_1, \ldots, \gamma + n_k) \tag{28}$$
$$\mu_i \mid \cdots \sim \mathcal{N}_r\Big((n_i\Sigma_i^{-1} + \kappa)^{-1}(n_i\Sigma_i^{-1}\bar{x}_i + \kappa\xi), (n_i\Sigma_i^{-1} + \kappa)^{-1}\Big) \tag{29}$$
$$\Sigma_i^{-1} \mid \cdots \sim \mathcal{W}_r\Big(2\alpha + n_i, \Big[2\beta + \sum_{j:z_j=i}(x_j - \mu_i)(x_j - \mu_i)^T\Big]^{-1}\Big) \tag{30}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, n$, where n_i is the number of observations allocated to class i ($n_i = \#\{j : z_j = i\}$) and $\bar{x}_i$ is the mean of the observations allocated to class i ($\bar{x}_i = \sum_{j:z_j=i} x_j/n_i$.) The Gibbs sampling updates were performed in the order $\beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.

2. The *Variable-κ* prior, in which ξ and κ are also treated as hyperparameters on which we place "vague" priors. This is an attempt to represent the belief that the means will be close together when viewed on some scale, without being informative about their actual location. It is also an attempt to address some of the objections to the Fixed-κ prior discussed in Section 5.1. We chose to place an improper uniform prior distribution on ξ, and a "vague" $\mathcal{W}_r(l, (lI_r)^{-1})$ distribution on κ where I_r is the $r \times r$ identity matrix. In order to ensure the posterior distribution for κ is proper, this distribution is required to be proper, and so we require $l > r - 1$. We used $l = r - 1 + 0.001$ as our default value for l. (In general, fixing a distribution to be proper in this way is not a good idea. However, in this case it can be shown that if $l = r - 1 + \epsilon$ then inference for μ, Σ and k is not sensitive to ϵ for small ϵ, although numerical problems may occur for very small ϵ.)

   The full conditional posteriors are then as for the Fixed-κ prior, with the addition of:

$$\xi \mid \cdots \sim \mathcal{N}_r(\bar{\mu}, (k\kappa)^{-1}) \tag{31}$$

$$\kappa \mid \cdots \sim \mathcal{W}_r(\kappa; l + k, (lI_r + SS)^{-1}) \tag{32}$$

   where $\bar{\mu} = \sum_i \mu_i / k$ and $SS = \sum_i (\mu_i - \xi)(\mu_i - \xi)^T$. The Gibbs sampling updates in Algorithm 3.2 were performed in the order $\beta, \kappa, \xi, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.

These priors are both examples of the special case considered in Section 3.3, and so Algorithm 3.1 can be used. They may be viewed as convenient for the purposes of illustration, and we warn against considering them as "non-informative" or "weakly" informative. In particular we will see that inference for k can be highly sensitive to the priors used. Further discussion is deferred to Section 5.1.

Values for (t_0, λ_b)

Algorithm 3.1 requires the specification of a birth-rate λ_b, and Algorithm 3.2 requires the specification of a (virtual) time t_0 for which the birth-death process is run. Doubling λ_b is mathematically equivalent to doubling t_0, and so we are free to fix $t_0 = 1$, and specify a value for λ_b. In all our examples we used $\lambda_b = \lambda$ (the parameter of the Poisson prior in (20)), which gives a convenient form of the death rates (18) as a likelihood ratio which does not depend on λ. Larger values of λ_b will result in better mixing over k, at the cost of more computation time *per* iteration of Algorithm 3.2, and it is not clear how an optimal balance between these factors should be achieved.

4.1 Example 1: Galaxy data

As our first example we consider the galaxy data first presented by Postman *et al.* (1986), consisting of the velocities (in 10^3 km/s) of distant galaxies diverging from our own, from six well-separated conic sections of the Corona Borealis. The original data consists of 83 observations, but one of these observations (a velocity of 5.607×10^3 km/s) does not appear in the version of the data given by Roeder (1990), which has since been analyzed under a variety of mixture models by a number of authors, including Crawford (1994), Chib (1995), Carlin and Chib (1995), Escobar and West (1995), Phillips and Smith (1996) and Richardson and Green (1997). In order to make our analysis comparable with these we have chosen to ignore the missing observation. A histogram of the data overlaid with a Gaussian kernel density estimate is shown in Figure 2. The multimodality of the velocities may indicate the presence of super clusters of galaxies surrounded by large voids,

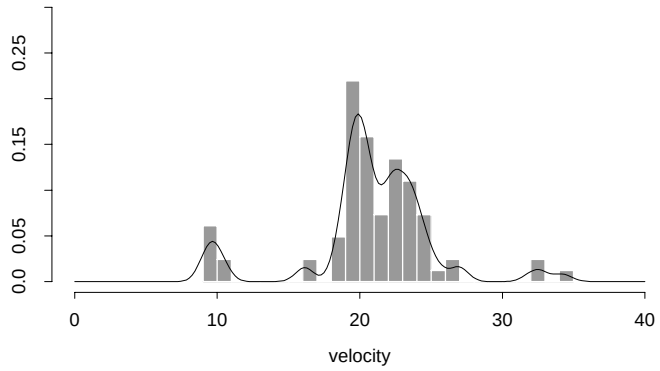


Figure 2: Histogram of the galaxy data, with bin-widths chosen by eye. Since histograms are rather unreliable density estimation devices (see for example Roeder, 1990) we have overlaid the histogram with a non-parametric density estimate using Gaussian kernel density estimation, with bandwidth chosen automatically according to a rule given by Sheather and Jones (1991), calculated using the S function `width.SJ` from Venables and Ripley (1997).

each mode representing a cluster as it moves away at its own speed (Roeder, 1990, gives more background).

We use Algorithm 3.2 to fit the following mixture models to the galaxy data:

a) A mixture of normal distributions using the Fixed-κ prior described in Section 4.

b) A mixture of normal distributions using the Variable-κ prior described in Section 4.

c) A mixture of t distributions on $p = 4$ degrees of freedom:

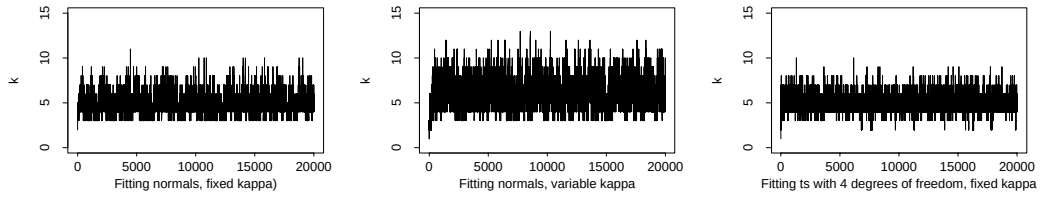
$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 t_p(x; \mu_1, \sigma_1^2) + \cdots + \pi_k t_p(x; \mu_k, \sigma_k^2), \tag{33}$$

where $t_p(x; \mu_i, \sigma_i^2)$ is the density of the t-distribution with p degrees of freedom, with mean μ_i and variance $p\sigma_i^2/(p-2)$ (see for example Gelman *et al.*, 1995, p. 476). The value $p = 4$ was chosen to give a distribution similar to the normal distribution with slightly "fatter tails", since there was some evidence when fitting the normal distributions that extra components were being used to create longer tails. We used the Fixed-κ prior for $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Adjusting the birth-death algorithm to fit t distributions is simply a matter of replacing the normal density with the t density when calculating the likelihood. The Gibbs sampling steps are performed as explained in Stephens (1997).

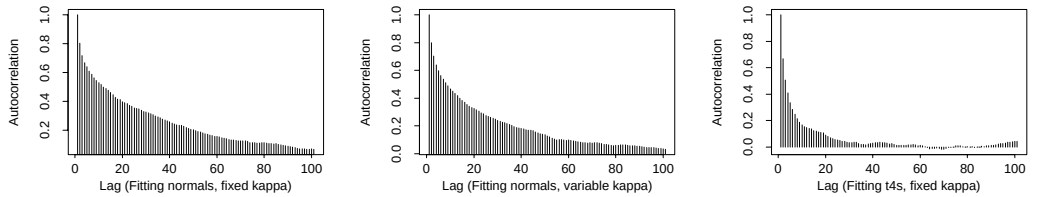
We will refer to these three models as "Normal, Fixed-κ"; "Normal, Variable-κ"; and "t_4, Fixed-κ" respectively. For each of the three models we performed the analysis with four different values of the parameter λ (the parameter of the truncated Poisson prior on k): 1,3,6 and 25. The choice of $\lambda = 25$ was considered in order to give some idea of how the method would behave as λ was allowed to get very large.

Starting points, computational expense, and mixing behaviour

For each prior we performed $20\,000$ iterations of Algorithm 3.2, with the starting point being chosen by setting $k = 1$, setting (ξ, κ) to the values chosen for the Fixed-κ prior, and sampling the



(a) Sampled values of k



(b) Autocorrelations for sampled values of k

Figure 3: Results from using Algorithm 3.2 to fit the three different models to the galaxy data using $\lambda = 3$. The columns show results for **Left:** Normals, Fixed-κ; **Middle:** Normals, Variable-κ; **Right:** t_4s, Fixed-κ.

other parameters from their joint prior distribution. In each case the sampler moved quickly from the low likelihood of the starting point to an area of parameter space with higher likelihood. The computational expense was not great. For example, the runs for $\lambda = 3$ took 150-250 seconds (CPU times on a Sun UltraSparc 200 workstation, 1997), which corresponds to about 80-130 iterations per second. Roughly the same amount of time was spent performing the Gibbs sampling steps as performing the birth-death calculations. The main expense of the birth-death process calculations is in calculating the model likelihood, and a significant saving could be made by using a look-up table for the normal density (this was not done).

In assessing the convergence and mixing properties of our algorithm we follow Richardson and Green (1997) in examining firstly the mixing over k, and then the mixing over the other parameters within k. Figure 3a shows the sampled values of k for the runs with $\lambda = 3$. A rough idea of how well the algorithm is exploring the space may be obtained from the percentages of iterations which changed k, which in this case were 36%, 52%, and 38% for models a)-c) respectively. More information can be obtained from the autocorrelation of the sampled values of k (Figure 3b) which show that successive samples have a high autocorrelation. This is due to the fact that k tends to change by at most one in each iteration, and so many iterations are required to move between small and large values of k.

In order to obtain a comparison with the performance of the reversible jump sampler of Richardson and Green (1997) we also performed runs with the prior they used for this data; namely a uniform prior on $k = 1, \ldots, 30$ and the Fixed-κ prior on the parameters. For this prior our sampler took 170 seconds and changed k in 34% of iterations, which compares favourably with the 11-18% of iterations obtained by Richardson and Green (1997) using the reversible jump sampler (their Table 1). We also tried applying the convergence diagnostic suggested by Gelman and Rubin (1992)

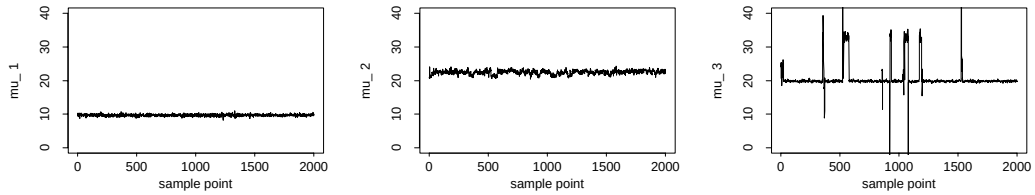


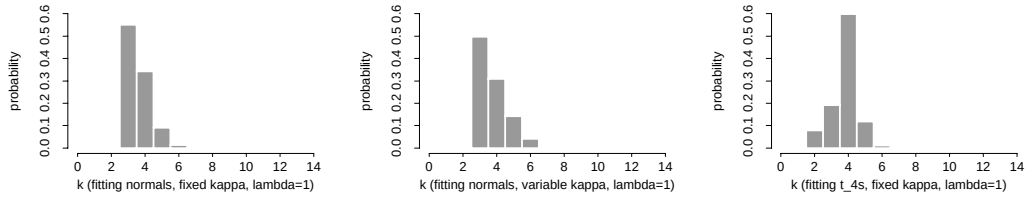
Figure 4: Sampled values of means for three components, sampled using Algorithm 3.2 when fitting a variable number of t_4 components to the galaxy data, with Fixed-κ prior, $\lambda = 1$, and conditioning the resulting output on $k = 3$. The output is essentially "unlabelled", and so labelling of the points was achieved by applying Algorithm 3.3 of Stephens (1997). The variable k sampler visits the minor mode at least 6 separate times in 1913 iterations, compared with once in $10\,000$ iterations for a fixed k sampler.

which requires more than one chain to be run from over-dispersed starting points (see the reviews by Cowles and Carlin (1996) or Brooks and Roberts (1998) for alternative diagnostics). Based on four chains of length 20 000, with two started from $k = 1$ and two started from $k = 30$, convergence was diagnosed for the output of Algorithm 3.2 within 2500 iterations.

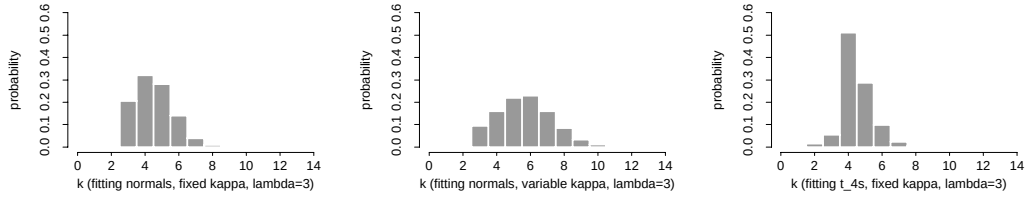
Richardson and Green (1997) note that allowing k to vary can result in much improved mixing behaviour of the sampler over the mixture model parameters *within* k. For example, if we fix k and use Gibbs sampling to fit $k = 3$ t_4 distributions to the galaxy data with the Fixed-κ prior, there are two well-separated modes (a major mode with means near 10, 20, and 23 and a minor mode with means near 10, 21 and 34). Our Gibbs sampler with fixed k struggled to move between these modes, moving from major mode to minor mode and back only once in $10\,000$ iterations (results not shown). We applied Algorithm 3.2 to this problem, using $\lambda = 1$. Of the $10\,000$ points sampled, there were 1913 visits to $k = 3$, during which the minor mode was visited on at least 6 different occasions (Figure 4). In this case the improved mixing behaviour results from the ability to move between the modes for $k = 3$ *via* states with $k = 4$: that is (roughly speaking), from the major mode to the minor mode *via* a four component model with means near 10, 20, 23 and 34. If we are genuinely only interested in the case $k = 3$ then the improved mixing behaviour of the variable k sampler must be balanced against its increased computational cost, particularly as we generated only 1913 samples from $k = 3$ in $10\,000$ iterations of the sampler. By truncating the prior on k to allow only $k = 3$ and $k = 4$, and using $\lambda = 0.1$ to favour the 3 component model strongly, we were able to increase this to 7371 samples with $k = 3$ in $10\,000$ iterations, with about 6 separate visits to the minor mode. Alternative strategies for obtaining a sample from the birth-death process conditional on a fixed value of k are given by Ripley (1977).

Inference

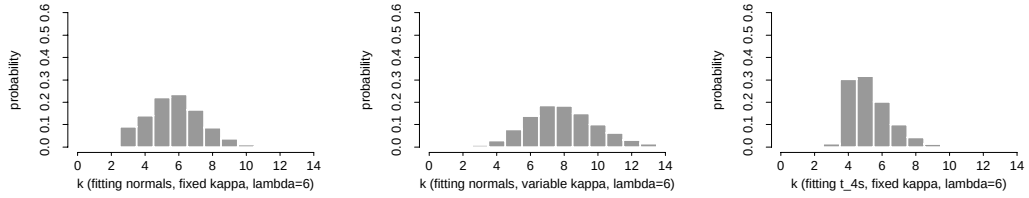
The results in this section are based on runs of length $20\,000$ with the first $10\,000$ iterations being discarded as burn-in — numbers we believe to be large enough to give meaningful results based on our investigations of the mixing properties of our chain. Estimates of the posterior distribution of k (Figure 5) show that it is highly sensitive to the prior used, both in terms of choice of λ and the prior (Variable-κ or Fixed-κ) used on the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Corresponding estimates of the predictive density (Figure 6) show that this is less sensitive to choice of model. Although the density estimates become less smooth as λ increases, even the density estimates for (the unreasonably large value of) $\lambda = 25$ do not appear to be over-fitting badly.



(a) $\lambda = 1$

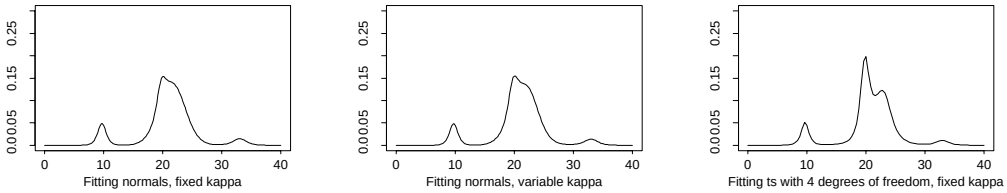


(b) $\lambda = 3$

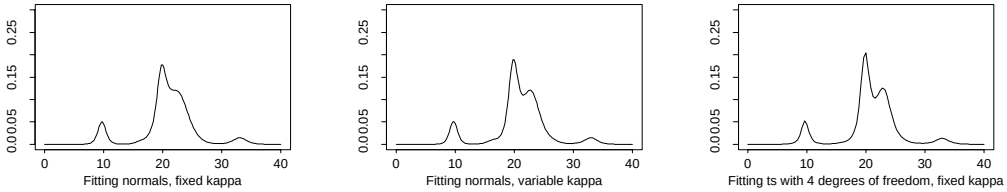


(c) $\lambda = 6$

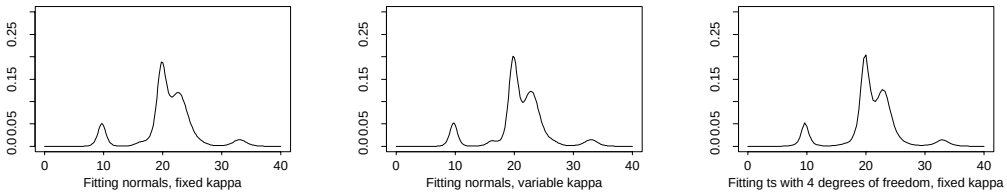
Figure 5: Graphs showing estimates (6) of $\Pr(k = i)$ for $i = 1, 2, \ldots$, for the galaxy data. These estimates are based on the values of k sampled using Algorithm 3.2 when fitting the three different models to the galaxy data with $\lambda = 1, 3, 6$, with in each case the first $10\,000$ samples having been discarded as burn-in. The three columns show results for **Left:** Normals, Fixed-κ; **Middle:** Normals, Variable-κ; **Right:** t_4s, Fixed-κ. The posterior distribution of k can be seen to depend on the type of mixture used (normal or t_4), the prior distribution for k (value of λ), and the prior distribution for $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ (Variable-κ or Fixed-κ).



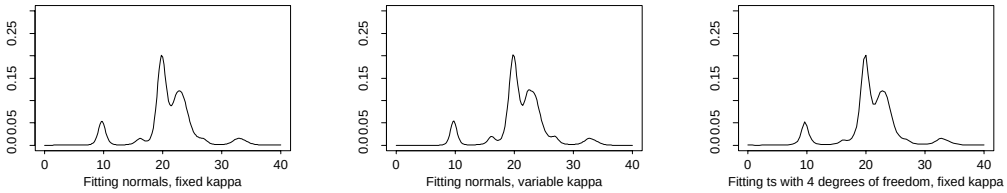
(a) $\lambda = 1$



(b) $\lambda = 3$



(c) $\lambda = 6$



(d) $\lambda = 25$

Figure 6: Predictive density estimates (7) for the galaxy data. These are based on the output of Algorithm 3.2 when fitting the three different models to the galaxy data with $\lambda = 1, 3, 6, 25$. The three columns show results for **Left:** Normals, Fixed-κ; **Middle:** Normals, Variable-κ; **Right:** t_4s, Fixed-κ. The density estimates become less smooth as λ increases, corresponding to a prior distribution which favours a larger number of components. However, the method appears to perform acceptably for even unreasonably large values of λ.

$k =$	2	3	4	5	6	> 6
$\widehat{p}(k \mid t_4, x^n)$	0.056 (0.014)	0.214 (0.009)	0.601 (0.011)	0.115 (0.005)	0.012 (0.001)	0.001 (0.000)
$\widehat{p}(k \mid \text{normal}, x^n)$	0.000	0.554 (0.014)	0.338 (0.011)	0.093 (0.004)	0.013 (0.001)	0.001 (0.000)

Table 1: Estimates of the posterior probabilities $p(k \mid t_4, x^n)$ and $p(k \mid \text{normal}, x^n)$ for the galaxy data (Fixed-κ prior, $\lambda = 1$). These are the means of the estimates from five separate runs of Algorithm 3.2, each run consisting of $20\,000$ iterations with the first $10\,000$ iterations being discarded as burn-in; the standard errors of these estimates are shown in brackets.

$k =$	2	3	4	5	6	> 6
$\widehat{p}(t_4, k \mid x^n)$	0.051	0.196	0.551	0.105	0.011	0.000
$\widehat{p}(\text{normal}, k \mid x^n)$	0.000	0.047	0.028	0.008	0.001	0.000

Table 2: Estimates of the posterior probabilities $p(t_4, k \mid x^n)$ and $p(\text{normal}, k \mid x^n)$ for the galaxy data (Fixed-κ prior, $\lambda = 1$). See text for details of how these were obtained.

The large number of normal components being fitted to the data suggests that the data is not well modelled by a mixture of normal distributions. Further investigation shows that many of these components have small weight and are being used to effectively "fatten the tails" of the normal distributions, which explains why fewer t_4 components are required to model the data. Parsimony suggests that we should prefer the t_4 model, and we can formalize this as follows. Suppose we assume that the data has arisen from either a mixture of normals or a mixture of t_4s, with $p(t_4) = p(\text{normal}) = 0.5$. For the Fixed-$\kappa$ prior with $\lambda = 1$ we can estimate $p(k \mid t_4, x^n)$ and $p(k \mid \text{normal}, x^n)$ using Algorithm 3.2 (Table 1). By Bayes theorem we have

$$p(k \mid t_4, x^n) = \frac{p(k, t_4 \mid x^n)}{p(t_4 \mid x^n)} \quad \text{for all } k \tag{34}$$

and so

$$p(t_4 \mid x^n) = \frac{p(k, t_4 \mid x^n)}{p(k \mid t_4, x^n)} = \frac{p(x^n \mid k, t_4)p(k, t_4)}{p(k \mid t_4, x^n)p(x^n)} \quad \text{for all } k, \tag{35}$$

and similarly

$$p(\text{normal} \mid x^n) = \frac{p(x^n \mid k, \text{normal})p(k, \text{normal})}{p(k \mid \text{normal}, x^n)p(x^n)} \quad \text{for all } k. \tag{36}$$

Thus if we can estimate $p(x^n \mid k, t_4)$ for *some* k and $p(x^n \mid k, \text{normal})$ for *some* k then we can estimate $p(t_4 \mid x^n)$ and $p(\text{normal} \mid x^n)$. Mathieson (1997) describes a method (a type of importance sampling which he refers to as Truncated Harmonic Mean (THM) and which is similar to the method described by DiCiccio *et al.* (1997)) of obtaining estimates for $p(x^n \mid k, t_4)$ and $p(x^n \mid k, \text{normal})$, and uses this method to obtain the estimates

$$-\log p(x^n \mid k = 3, t_4) \approx 227.64 \qquad \text{and} \qquad -\log p(x^n \mid k = 3, \text{normal}) \approx 229.08,$$

giving (using equations (35) and (36))

$$p(t_4 \mid x^n) \approx 0.916 \qquad \text{and} \qquad p(\text{normal} \mid x^n) \approx 0.084,$$

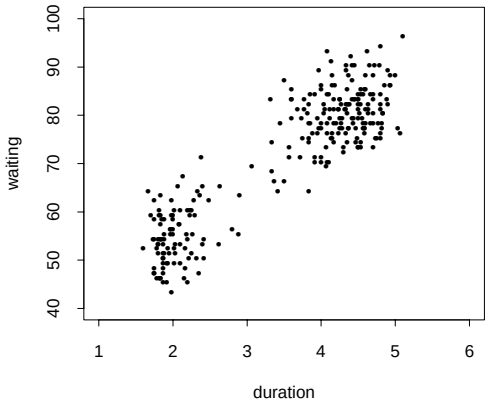


Figure 7: Scatter plot of the Old Faithful data (from Härdle, 1991). The x axis shows the duration (in minutes) of the eruption, and the y axis shows the waiting time (in minutes) before the next eruption.

from which we can estimate $p(t_4, k \mid x^n) = p(t_4 \mid x^n)p(k \mid t_4, x^n)$, and similarly for normals — the results are shown in Table 2. We conclude that for the prior distributions used, mixtures of t_4 distributions are heavily favoured over mixtures of normal distributions, with four t_4 components having the highest posterior probability. It would be relatively straightforward to modify our algorithm to fit t distributions with an unknown number of degrees of freedom, thus automating the above model choice procedure. It would also be straightforward to allow each component of the mixture to have a different number of degrees of freedom.

4.2 Example 2: Old Faithful data

For our second example, we consider the Old Faithful data (the version from Härdle, 1991, also considered by Venables and Ripley (1994)) which consists of data on 272 eruptions of the Old Faithful geyser in the Yellowstone National Park. Each observation consists of two observations: the *duration* (in minutes) of the eruption, and the *waiting* time (in minutes) before the next eruption. A scatter plot of the data in two dimensions shows two moderately separated groups (Figure 7). We used Algorithm 3.2 to fit a mixture of an unknown number of bivariate normal distributions to the data, using $\lambda = 1, 3$ and both the Fixed-κ and Variable-κ priors detailed in Section 4.

Each run consisted of $20\,000$ iterations of Algorithm 3.2, with the starting point being chosen by setting $k = 1$, setting (ξ, κ) to the values chosen for the Fixed-κ prior, and sampling the other parameters from their joint prior distribution. In each case the sampler moved quickly from the low likelihood of the starting point to an area of parameter space with higher likelihood. The runs for $\lambda = 3$ took about 7-8 minutes. Figure 8a shows the resulting sampled values of the number of components k, which can be seen to vary more rapidly for the Variable-κ model, due in part to its greater permissiveness of extra components. For the runs with $\lambda = 3$ the proportion of iterations which resulted in a change in k were 9% (Fixed-κ) and 39% (Variable-κ). For $\lambda = 1$ the corresponding figures were 3% and 10% respectively. Graphs of the autocorrelations (Figure 8b) suggest that the mixing is slightly poorer than for the galaxy data, presumably due to births of reasonable components being less likely in the two-dimensional case. This poorer mixing means that longer runs may be necessary to obtain accurate estimates of $p(k \mid x^n)$. The method of Gelman

and Rubin (1992) applied to two runs of length $20\,000$ starting from $k = 1$ and $k = 30$ diagnosed convergence within $10\,000$ iterations for the Fixed-κ prior with $\lambda = 1, 3$.

Estimates of the posterior distribution for k (Figure 8c) show that it depends heavily on the prior used, while estimates of the predictive density (Figure 8d)) are less sensitive to changes in the prior. Where more than two components are fitted to the data the extra components appear to be modelling deviations from normality in the two obvious groups, rather than interpretable extra groups.

4.3 Example 3: *Iris Virginica* data

We now briefly consider the famous *Iris* data, collected by Anderson (1935) which consists of four measurements (petal and sepal length and width) for 50 specimens of each of three species (*setosa*, *versicolor*, and *virginica*) of iris. Wilson (1982) suggests that the *virginica* and *versicolor* species may each be split into subspecies, though analysis by McLachlan (1992) using maximum likelihood methods suggests that this is not justified by the data. We investigated this question for the *virginica* species by fitting a mixture of an unknown number of bivariate normal distributions to the 50 observations of sepal length and petal length for this species, which are shown in Figure 9.

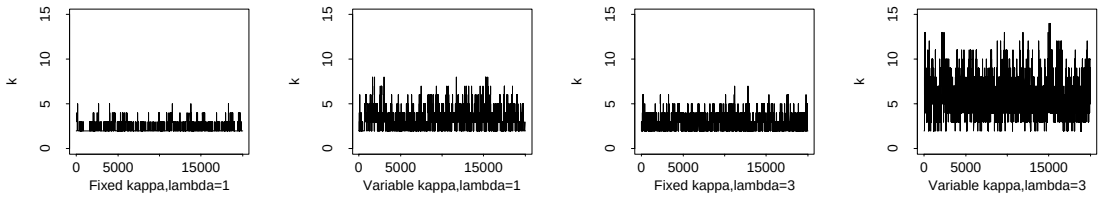
Our analysis was performed with $\lambda = 1, 3$ and with both Fixed-κ and Variable-κ priors. We applied Algorithm 3.2 to obtain a sample of size $20\,000$ from a random starting point, and discarded the first $10\,000$ observations as burn-in. The mixing behaviour of the chain over k was reasonable, with the percentages of sample points for which k changed being 6% ($\lambda = 1$) and 21% ($\lambda = 3$) for the Fixed-κ prior, and 5% ($\lambda = 1$) and 36% ($\lambda = 3$) for the Variable-κ prior. The mode of the resulting estimates for the posterior distribution of k is at $k = 1$ for at least three of the four priors used (Figure 10a) and the results seem to support the conclusion of McLachlan (1992) that the data does not support a division into subspecies (though we note that in our analysis we used only two of the four measurements available for each specimen). The full predictive density estimates in Figure 10b indicate that where more than one component is fitted to the data they are again being used to model lack of normality in the data, rather than interpretable groups in the data.

5 Discussion

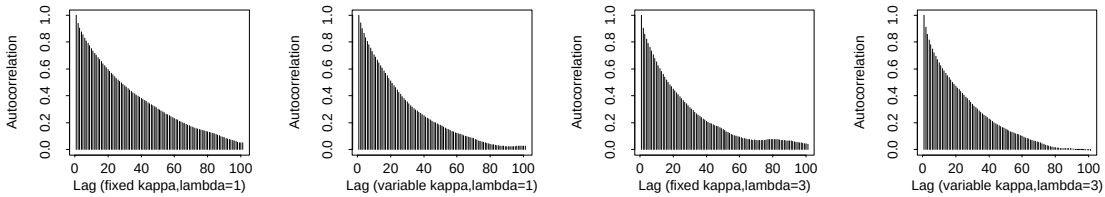
5.1 Density estimation, inference for k, and priors

Our examples demonstrate that a Bayesian approach to density estimation using mixtures of (univariate or bivariate) normal distributions with an unknown number of components is computationally feasible, and that the resulting density estimates are reasonably robust to modelling assumptions and priors used. Extension to higher dimensions is likely to provide computational challenges, but might be possible with suitable constraints on the covariance matrices (requiring them all to be equal or all to be diagonal for example).

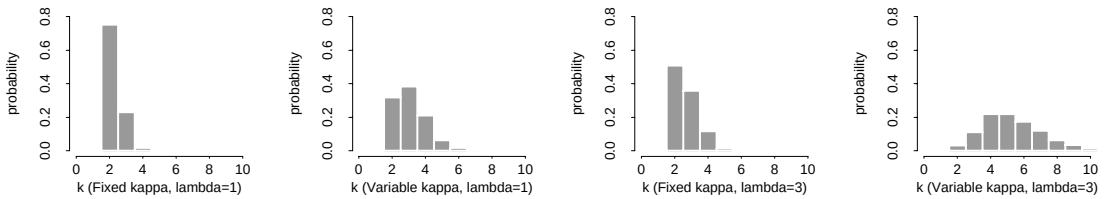
Our examples also highlight the fact that while inference for the number of components k in the mixture is also computationally feasible, the posterior distribution for k can be highly dependent on not just the prior chosen for k, but also the prior chosen for the other parameters of the mixture model. Richardson and Green (1997), in their investigation of one-dimensional data, note that when using the Fixed-κ prior, the value chosen for κ in the prior $\mathcal{N}(\xi, \kappa^{-1})$ for the means $\mu_1, \ldots, \mu_k$ has a subtle effect on the posterior distribution of k. A very large value of κ, representing a strong belief that the means lie at ξ (chosen to be the midpoint of the range of the data) will favour models with a small number of components and larger variances. Decreasing κ to represent vaguer prior knowledge about the means will initially encourage the fitting of more components with means spread across the range of the data. However, continuing to decrease κ, to represent vaguer and vaguer



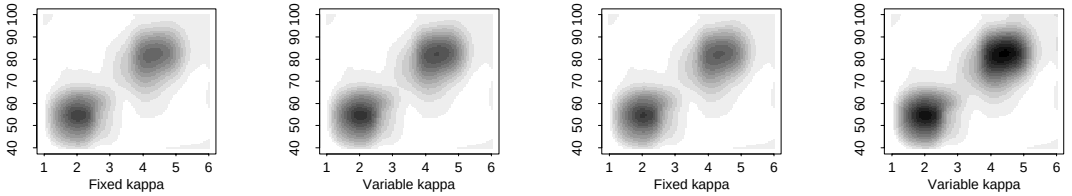
(a) Sampled values of k



(b) Autocorrelations of sampled values of k



(c) Estimates (6) of $\Pr(k = i)$



(d) Predictive density estimates (7), dark shading corresponding to regions of high density, all shaded on the same scale

Figure 8: Results for using Algorithm 3.2 to fit a mixture of normal distributions to the Old Faithful data. The columns show results for **Left:** Fixed-κ prior, $\lambda = 1$; **Left-middle:** Variable-κ prior, $\lambda = 1$; **Right-middle:** Fixed-κ prior, $\lambda = 3$; **Right:** Variable-κ prior, $\lambda = 3$. The posterior distribution of k can be seen to depend on both the prior distribution for k (value of λ), and the prior distribution for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Variable-κ or Fixed-κ). The density estimates appear to be less sensitive to choice of prior.

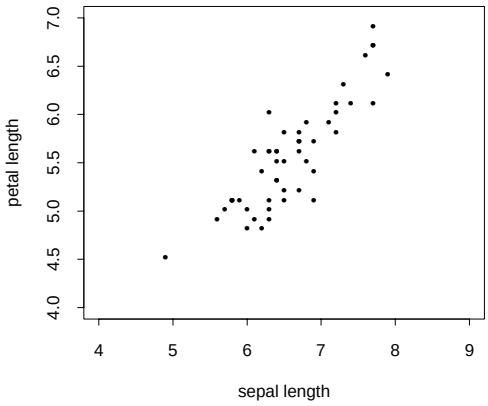
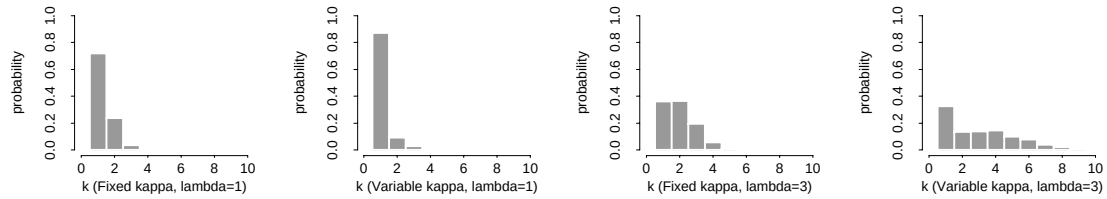
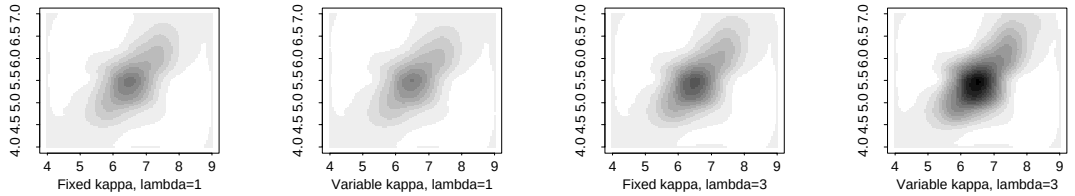


Figure 9: Scatter plot of petal length against sepal length for the *Iris Virginica* data.



(a) Estimates (6) of $\Pr(k = i)$



(b) Predictive density estimates (7), dark shading corresponding to regions of high density, all shaded on the same scale

Figure 10: Results for using Algorithm 3.2 to fit a mixture of normal distributions to the *Iris Virginica* data. The columns show results for **Left:** Fixed-κ prior, $\lambda = 1$; **Left-middle:** Variable-κ prior, $\lambda = 1$; **Right-middle:** Fixed-κ prior, $\lambda = 3$; **Right:** Variable-κ prior, $\lambda = 3$. The mode of the estimates of $\Pr(k = i)$ is $k = 1$ for at least 3 of the four priors used, and seems to indicate that the data does not support splitting the species into sub-species.

knowledge on the location of the means, eventually favours fitting fewer components. In the limit, as $\kappa \to 0$, the posterior distribution of k becomes independent of the data, and depends only on the number of observations, heavily favouring a one component model for reasonable number of observations (Stephens, 1997; Jennison, 1997). Priors which appear to be only "weakly" informative for the parameters of the mixture components may thus be highly informative for the number of components in the mixture. Since very large and very small values of κ in the Fixed-κ prior both lead to priors which are highly informative for k, it might be interesting to search for a value of κ (probably depending on the observed data) which leads to a Fixed-κ prior which is "minimally informative" for k in some well-defined way.

Where the main aim of the analysis is to define groups for *discrimination* (as in taxonomic applications such as the iris data for example) it seems natural that the priors should reflect our belief that this is a reasonable aim, and thus avoid fitting several similar components where one will suffice. This idea is certainly not captured by the priors we used here, which Richardson and Green (1997) suggest are more appropriate for "exploring heterogeneity". Inhibition priors from spatial point processes (as used by Baddeley and van Lieshout, 1993, for example) provide one way of expressing a prior belief that the components present will be somewhat distinct. Alternatively we might try distinguishing between the number of components in the model, and the number of "groups" in the data, by allowing each group to be modelled by several "similar" components. For example, group means might be *a priori* distributed on the scale of the data, and each group might consist of an unknown number of normal components, with means distributed around the group mean on a smaller scale than the data. The discussion following Richardson and Green (1997) provides a number of other avenues for further investigation of suitable priors, and we hope that the computational tools described in this paper will help make such further investigation possible.

5.2 Choice of birth distribution

The choice of birth distribution we made in Algorithm 3.1 is rather naive, and indeed we were rather surprised that we were able to make much progress with this approach. Its success in the Fixed-κ model appears to stem from the fact that the (data-dependent) independent priors on the parameters ϕ are not so vague as to never produce a reasonable birth event, and yet not so tight as to always propose components which are very similar to those already present. In the Variable-κ model the success of the naive algorithm seems to be due to the way in which the hyperparameters κ and ξ "adapt" the birth distribution to make the birth of better components more likely. Here we may have been lucky, since the priors were not chosen with these properties in mind. In general then it may be necessary to spend more effort designing sensible birth-death schemes to achieve adequate mixing. Our results suggest that a strategy of allowing the birth distribution $b(y; (\pi, \phi))$ to be independent of y, but depend on the data, may result in a simple algorithm with reasonable mixing properties. An *ad hoc* approach to improving mixing might involve simply investigating mixing behaviour for more or less "vague" choices of b. A more principled approach would be to choose a birth distribution which can be both easily calculated and simulated from directly, and which roughly approximates the (marginal) posterior distribution of a randomly chosen element of ϕ. Such an approximation might be obtained from a preliminary analysis with a naive birth mechanism, or perhaps standard fixed-dimension MCMC with large k.

A more sophisticated approach might allow the birth distribution $b\left(y; (\pi, \phi)\right)$ to depend on y. Indeed, the opposite extreme to our naive approach would be to allow all points to die at a constant rate, and find the corresponding birth distribution using (15) (as in Ripley, 1977, for example). However, much effort may then be required to calculate the birth rate $\beta(\cdot)$ (perhaps by Monte-

Carlo integration), which limits the appeal of this approach. (This problem did not arise in Ripley, 1977, where simulations were performed conditional on a fixed value of k by alternating births and deaths.) For this reason we believe that it is easier to concentrate on designing efficient birth distributions which can be simulated from directly and whose densities can be calculated explicitly so that the death rates (15) are easily computed.

5.3 Extension to other contexts

It appears from our results that, for finite mixture problems, our birth-death algorithm provides an attractive alternative to the algorithm used by Richardson and Green (1997). There seems to be considerable potential for applying similar birth-death schemes in other contexts as an alternative to more general reversible jump methods. We now attempt to give some insight into for which problems such an approach is likely to be feasible. We begin our discussion by highlighting the main differences between our Algorithm 3.1 and the algorithm used by Richardson and Green (1997).

A: Our algorithm operates in continuous time, replacing the accept-reject scheme by allowing events to occur at differing rates.

B: Our dimension-changing birth and death moves do not make use of the missing data z^n, effectively integrating out over them when calculating the likelihood.

C: Our birth and death moves take advantage of the natural nested structure of the models, removing the need for the calculation of a complicated Jacobian, and making implementation more straightforward.

D: Our birth and death moves treat the parameters as a point process, and do not make use of any constraint such as $\mu_1 < \cdots < \mu_k$ (used by Richardson and Green, 1997, in defining their split and combine moves).

We consider A to be the least important distinction. Indeed, a discrete time version of our birth-death process using an accept-reject step could be designed along the lines of Geyer and Møller (1994), or using the general reversible-jump formulation of Green (1995). (Similarly one can envision a continuous time version of the general reversible jump formulation.) We have no good intuition for whether discrete time or continuous time versions are likely to be more efficient in general, although Geyer and Møller (1994) suggests that it is easier to obtain analytical results relating to mixing for the discrete time version.

Point B raises an important requirement for application of our algorithm: we must be able to calculate the likelihood for any given parameters. This requirement makes the method difficult to apply to Hidden Markov Models, or other missing data problems where calculation of the likelihood requires knowledge of the missing data. One solution to this problem would be to introduce the missing data into the MCMC scheme, and perform births and deaths while keeping the missing data fixed (along the lines of the births and deaths of "empty" components in Richardson and Green, 1997). However, where the missing data is highly informative for k this seems likely to lead to poor mixing, and reversible jump methods which propose joint updates to the missing data and the dimension of the model appear more sensible here.

In order to take advantage of the simplicity of the birth-death methodology, we must be able to view the parameters of our model as a point process, and in particular we must be able to express our prior in terms of a Radon–Nikodym derivative, $r(\cdot)$, with respect to a symmetric measure, as in Section 2.2. This is not a particularly restrictive requirement, and we give two concrete examples

below. These examples are in many ways simpler than the mixture problem since there are no mixture proportions, and the marked point process becomes a point process on a space Φ. The analogue of Theorem 3.1 for this simpler case, (which essentially follows directly from Preston (1976) and Ripley (1977)) may be obtained by replacing the condition (15) with

$$(k+1)d(y;\phi)r(y\cup\phi)L(y\cup\phi) = \beta(y)b(y;\phi)r(y)L(y). \tag{37}$$

Provided we can calculate the likelihood $L(y)$, the viability of the birth-death methodology will depend on being able to find a birth distribution which gives adequate mixing. The comments in Section 5.2 provide some guidance here. It is clear that in some applications the use of birth and death moves alone will make it difficult to achieve adequate mixing. However, the ease with which different birth distributions may be tried, and the success of our algorithm in the mixture context with minimal effort in designing efficient birth distributions, suggests that this type of algorithm is worth trying before more complex reversible jump proposal distributions are implemented.

Example I: Change point analysis

Consider the change-point problem from Green (1995). The parameters of this model are the number of change points k, the positions $0 < s_1 < \cdots < s_k < L$ of the change points, and the heights h_i $(i = 0, \ldots, k)$ associated with the intervals $[s_i, s_{i+1}]$, where s_0 and s_{k+1} are defined to be 0 and L respectively. In order to treat the parameters of the model as a point process, we drop the requirement that $s_1 < \cdots < s_k$, and define the likelihood of the model in terms of the order statistics $s_{(1)} < \cdots < s_{(k)}$, and the corresponding heights $h_{(i)}$ $(i = 0, \ldots, k)$ associated with the intervals $[s_{(i)}, s_{(i+1)}]$, where $s_{(0)}$ and $s_{(k+1)}$ are defined to be 0 and L respectively.

Consider initially a prior in which k has prior probability mass distribution $p(k)$, and conditional on k, the s_i and h_i are assumed to be independent, with s_i uniformly distributed on [0,L], and $h_i \sim \Gamma(\alpha, \beta)$. In the notation of previous sections we take $\eta = h_{(0)}$, $\phi_i = (s_{(i)}, h_{(i)})$, $\omega = (\alpha, \beta)$, ν to be Lebesgue measure on $\Phi = [0, L] \times [0, \infty)$,

$$r(k, s, h) = p(k) \prod_{i=1}^{k} \frac{1}{L} I(s_i \in [0, L]) \Gamma(h_i; \alpha, \beta), \tag{38}$$

and π is ignored. With births and deaths on Φ defined in an obvious way, it is then straightforward to use condition (37) to create a birth-death process on $\Phi = [0, L] \times [0, \infty)$ with the posterior distribution of $\boldsymbol{\phi}$ given η as its stationary distribution. This can then be alternated with standard fixed-dimension MCMC steps (which allow $h_{(0)}$, and perhaps α and β to vary) to create an ergodic Markov chain with the posterior distribution of the parameters as its stationary distribution. The analogue of our naive algorithm for this prior would have birth distribution

$$b\big(y;(s,h)\big) = \frac{1}{L} I(s \in [0, L]) \Gamma(h; \alpha, \beta). \tag{39}$$

A more sophisticated approach would be to allow the birth of new change points to be concentrated on areas which, based on the data, seem good candidates for change points (for example, by looking at the marginal posterior distribution of the distribution of change points in a preliminary analyses using the naive birth mechanism, or fixed-dimension MCMC), and allow the birth distribution for the new h to depend on the new s, again being centred on regions which appear to be good candidates based on the data .

Now suppose that (as in Green, 1995) $s_{(1)}, \ldots, s_{(k)}$ are, given k, *a priori* distributed as the even-numbered order statistics of $2k+1$ points independently and uniformly distributed on $[0, L]$:

$$p(s_{(1)}, \ldots, s_{(k)}) = \frac{(2k+1)!}{L^{2k+1}}(s_{(1)} - 0)(s_{(2)} - s_{(1)}) \ldots (s_{(k)} - s_{(k-1)})(L - s_{(k)}) I(0 < s_{(1)} < \cdots < s_{(k)} < L). \tag{40}$$

This corresponds to $s_1, \ldots, s_k$ (which must be exchangeable) being *a priori* distributed as a random permutation of these order statistics:

$$p(s_1, \ldots, s_k) = \frac{1}{k!}\frac{(2k+1)!}{L^{2k+1}}(s_{(1)} - 0)(s_{(2)} - s_{(1)}) \ldots (s_{(k)} - s_{(k-1)})(L - s_{(k)}) \prod_{i=1}^{k} I(s_i \in [0, L]) \tag{41}$$

giving a prior which corresponds to

$$r'(k, s, h) = \frac{(2k+1)!}{k! L^{2k+1}}(s_{(1)} - 0)(s_{(2)} - s_{(1)}) \ldots (s_{(k)} - s_{(k-1)})(L - s_{(k)}) \prod_{i=1}^{k} I(s_i \in [0, L]) \Gamma(h_i; \alpha, \beta). \tag{42}$$

Given a birth-death scheme using the prior (39), it would be straightforward to modify this scheme to use this second prior (42), for example by keeping the birth distribution fixed, and modifying the calculation of the death rates by replacing r with r'. The way in which priors are so easily experimented with is one major attraction of the birth-death methodology.

Variable selection for regression models

Consider now the problem of selecting a subset of a given collection of variables to be included in a regression model (see George and McCulloch, 1996, for example). (Similar problems include deciding which terms to include in an autoregression, or which links to include in a Bayesian Belief Network.) Let there be K possible variables to include, and let variable i be associated with a parameter $\beta_i \in R$ $(i = 1, \ldots, K)$. A model which contains k of the variables can then be represented by a set of k points $\{(i_1, \beta_{i_1}), \ldots, (i_k, \beta_{i_k})\}$ in $\Phi = \{1, \ldots, K\} \times R$, where $i_1, \ldots, i_k$ are distinct integers in $\{1, \ldots, K\}$. The birth of a point (i, β_i) then corresponds to adding variable i to the regression. Note that the points are exchangeable in that the order in which they are listed is irrelevant. A suitable choice for ν in the definition of the symmetric measure $\mathcal{M}$ (Section 2.2) would be the product measure of counting measure on $\{1, \ldots, K\}$ and Lebesgue measure on R.

Suppose our prior is that variable i is present with probability p_i, independently for all i, and conditional on variable i being present, β_i has prior $p(\beta_i)$, again independent for all i. Then we have

$$r\big(k, (i_1, \beta_{i_1}), \ldots, (i_k, \beta_{i_k})\big) = \begin{cases} 0 \text{ if } i_a = i_b \text{ for some } a, b. \\ p_{i_1} p(\beta_{i_1}) \ldots p_{i_k} p(\beta_{i_k}) \text{ otherwise.} \end{cases} \tag{43}$$

The choice of birth distribution $b\big(y; (i, \beta_i)\big)$ must in this case depend on y, in order to avoid adding variables which are already present. A naive suggestion would be to set

$$b\big(y; (i, \beta_i)\big) = b_i p(\beta_i) \tag{44}$$

with $b_i \propto p_i$ for the variables i not already present in y. Again, more efficient schemes could be devised by letting the births be data-dependent, possibly through examining the marginal posterior distributions of the β_i in preliminary analyses.

6 Acknowledgements

I would like to thank my D.Phil. supervisor, Professor Brian Ripley, for suggesting this approach to the problem, and for valuable comments on earlier versions. I would also like to thank Mark Mathieson for helpful discussions on this work, and Peter Donnelly, Peter Green, two anonymous reviewers, the associate editor and the editor for helpful advice on improving the manuscript. The author was supported by an EPSRC studentship, and a grant from the University of Oxford.

7 Appendix: Proof of Theorem 3.1

Proof. Our proof draws heavily on the theory derived by Preston (1976), Section 5, for *general* Markov birth-death processes on state space $\Omega = \bigcup_k \Omega_k$ where the Ω_k are disjoint. The process evolves by jumps, of which only a finite number can occur in a finite time. The jumps are of two types: "births", which are jumps from a point in Ω_k to Ω_{k+1}, and "deaths", which are jumps from a point in Ω_k to a point in Ω_{k-1}. When the process is at $y \in \Omega_k$ the behaviour of the process is defined by the *birth rate* $\beta(y)$, the *death rate* $\delta(y)$, and the *birth and death transition kernels* $K_\beta^{(k)}(y;\cdot)$ and $K_\delta^{(k)}(y;\cdot)$ which are probability measures on Ω_{k+1} and Ω_{k-1} respectively. Births and deaths occur as independent Poisson processes, with rates $\beta(y)$ and $\delta(y)$ respectively. If a birth occurs then the process jumps to a point in Ω_{k+1}, with the probability that this point is in any particular set $F \subset \Omega_{k+1}$ being given by $K_\beta^{(k)}(y;F)$. If a death occurs then the process jumps to a point in Ω_{k-1}, with the probability that this point is in any particular set $G \subset \Omega_{k-1}$ being given by $K_\delta^{(k)}(y;G)$. Preston (1976) showed that for such a process to possess stationary distribution $\widetilde{\mu}$ it is sufficient that the following detailed balance conditions hold:

Definition 1 (Detailed Balance Conditions). $\widetilde{\mu}$ is said to satisfy detailed balance conditions if

$$\int_F \beta(y)\, d\widetilde{\mu}_k(y) = \int_{\Omega_{k+1}} \delta(z) K_\delta^{(k+1)}(z;F)\, d\widetilde{\mu}_{k+1}(z) \quad \text{for } k \geq 0,\ F \subset \Omega_k \tag{45}$$

and

$$\int_G \delta(z)\, d\widetilde{\mu}_{k+1}(z) = \int_{\Omega_k} \beta(y) K_\beta^{(k)}(y;G)\, d\widetilde{\mu}_k(y) \quad \text{for } k \geq 0,\ G \subset \Omega_{k+1}. \tag{46}$$

These have the intuitive meaning that the rate at which the process leaves any set through the occurrence of a birth is exactly matched by the rate at which the process enters that set through the occurrence of a death, and *vice-versa*.

We therefore check that $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid x^n, \omega, \eta)$ satisfies the detailed balance conditions for our process, which corresponds to the general Markov birth-death process with birth rate $\beta(y)$, death rate $\delta(y)$, and birth and death transition kernels $K_\beta^{(k)}(y;\cdot)$ and $K_\delta^{(k)}(y;\cdot)$ which satisfy

$$K_\beta^{(k)}(y;F) = \int_{(\pi,\phi):\, y \cup (\pi,\phi) \in F} b\big(y;(\pi,\phi)\big)\, d\pi\, \nu(d\phi) \tag{47}$$

and

$$\delta(y) K_\delta^{(k)}(y;F) = \sum_{(\pi,\phi) \in y:\, y\setminus(\pi,\phi) \in F} d\big(y\setminus(\pi,\phi);(\pi,\phi)\big). \tag{48}$$

We begin by introducing some notation. Let Λ_k represent the parameter space for the k-component model, with the labelling of the parameters taken into account, and let Ω_k be the corresponding space obtained by ignoring the labelling of the components. If $(\boldsymbol{\pi}, \boldsymbol{\phi}) \in \Lambda_k$, then we will write $[\boldsymbol{\pi}, \boldsymbol{\phi}]$ for the corresponding member of Ω_k. With $\Lambda = \bigcup_{k \geq 1} \Lambda_k$, let $P(\cdot)$ and $\widetilde{P}(\cdot)$ be the prior and posterior probability measures on Λ, and let $P_k(\cdot)$ and $\widetilde{P}_k(\cdot)$ denote their respective restrictions to Λ_k. The prior distribution has Radon–Nikodym derivative $r(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ with respect to $\mathcal{U}^{k-1} \times \nu^k$. Thus for $(\boldsymbol{\pi}, \boldsymbol{\phi}) \in \Lambda_k$ we have

$$dP_k\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} = r(k, \boldsymbol{\pi}, \boldsymbol{\phi})(k-1)!\, d\pi_1 \ldots d\pi_{k-1}\, \nu(d\phi_1) \ldots \nu(d\phi_k). \tag{49}$$

Also, by Bayes theorem we have

$$d\widetilde{P}\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \propto L([\boldsymbol{\pi}, \boldsymbol{\phi}])dP\{(\boldsymbol{\pi}, \boldsymbol{\phi})\}$$

and so we will write

$$d\widetilde{P}\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} = f([\boldsymbol{\pi}, \boldsymbol{\phi}])dP\{(\boldsymbol{\pi}, \boldsymbol{\phi})\}$$

for some $f([\boldsymbol{\pi}, \boldsymbol{\phi}]) \propto L([\boldsymbol{\pi}, \boldsymbol{\phi}])$.

Now let $\mu(\cdot)$ and $\widetilde{\mu}(\cdot)$ be the probability measures induced on Ω by $P(\cdot)$ and $\widetilde{P}(\cdot)$ respectively, and let $\mu_k(\cdot)$ and $\widetilde{\mu}_k(\cdot)$ denote their respective restrictions to Ω_k. Then for any function $g : \Omega \to R$ we have:

$$\int_{\Omega_k} g(y)\, d\mu_k(y) = \int_{\Lambda_k} g([\boldsymbol{\pi}, \boldsymbol{\phi}])\, dP_k\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \tag{50}$$

and

$$\begin{aligned} \int_{\Omega_k} g(y)\, d\widetilde{\mu}_k(y) &= \int_{\Lambda_k} g([\boldsymbol{\pi}, \boldsymbol{\phi}])\, d\widetilde{P}_k\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \\ &= \int g([\boldsymbol{\pi}, \boldsymbol{\phi}]) f([\boldsymbol{\pi}, \boldsymbol{\phi}])\, dP_k\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \\ &= \int_{\Omega_k} g(y) f(y)\, d\mu_k(y). \end{aligned} \tag{51}$$

We define births on Λ by

$$(\boldsymbol{\pi}, \boldsymbol{\phi}) \cup (\pi, \phi) := \Big((\pi_1(1-\pi), \phi_1), \ldots, (\pi_k(1-\pi), \phi_k), (\pi, \phi) \Big) \tag{52}$$

and will require the following Lemma (which is essentially a simple change of variable formula):

Lemma 7.1. *If* $(\boldsymbol{\pi}, \boldsymbol{\phi}) \in \Lambda_k$ *and* $(\pi, \phi) \in [0,1] \times \Phi$ *then*

$$r(k, \boldsymbol{\pi}, \boldsymbol{\phi}) dP_{k+1}\{(\boldsymbol{\pi}, \boldsymbol{\phi}) \cup (\pi, \phi)\} = r\big(k+1, (\boldsymbol{\pi}, \boldsymbol{\phi}) \cup (\pi, \phi)\big) k (1-\pi)^{k-1}\, d\pi\, \nu(d\phi)\, dP_k\{(\boldsymbol{\pi}, \boldsymbol{\phi})\}.$$

Proof.

$$
\begin{aligned}
LHS &= r(k,\boldsymbol{\pi},\boldsymbol{\phi})dP_{k+1}\big\{(\boldsymbol{\pi},\boldsymbol{\phi})\cup(\pi,\phi)\big\} \\
&= r(k,\boldsymbol{\pi},\boldsymbol{\phi})dP_{k+1}\Big\{\big((\pi_1(1-\pi),\phi_1),\ldots,(\pi_k(1-\pi),\phi_k),(\pi,\phi)\big)\Big\} \quad \text{[equation (52)]} \\
&= r(k,\boldsymbol{\pi},\boldsymbol{\phi})r\big(k+1,(\boldsymbol{\pi},\boldsymbol{\phi})\cup(\pi,\phi)\big)k!(1-\pi)^{k-1}\,d\pi_1\,\ldots\,d\pi_k\,d\pi\,\nu(d\phi_1)\,\ldots\,\nu(d\phi_k)\,\nu(d\phi) \\
&\qquad\qquad \text{[equation (49) and change of variable]} \\
&= r\big(k+1,(\boldsymbol{\pi},\boldsymbol{\phi})\cup(\pi,\phi)\big)k(1-\pi)^{k-1}\,d\pi\,\nu(d\phi)\,dP_k\big\{(\boldsymbol{\pi},\boldsymbol{\phi})\big\} \quad \text{[equation (49)]} \\
&= RHS.
\end{aligned}
$$

□

Assume for the moment that $r(y)L(y) > 0$ for all y. Let $I(\cdot)$ denote the generic indicator function, so $I(x \in A) = 1$ if $x \in A$ and 0 otherwise. We check the first part of the detailed balance conditions (45) as follows:

$$
\begin{aligned}
LHS &= \int_F \beta(y)\,d\widetilde{\mu}_k(y) \\
&= \int_{\Omega_k} I(y\in F)\,\beta(y)f(y)\,d\mu_k(y) \quad \text{[equation (51)]} \\
&= \int_{\Omega_k} I(y\in F)\,\beta(y)f(y)\int_{[0,1]}\int_\Phi b\big(y;(\pi,\phi)\big)\,d\pi\,\nu(d\phi)\,d\mu_k(y) \quad [b \text{ must integrate to 1.}] \\
RHS &= \int_{\Omega_{k+1}} \delta(z)K_\delta^{(k+1)}(z;F)\,d\widetilde{\mu}_{k+1}(z) \\
&= \int_{\Omega_{k+1}} \delta(z)K_\delta^{(k+1)}(z;F)\,f(z)\,d\mu_{k+1}(z) \quad \text{[equation (51)]} \\
&= \int_{\Omega_{k+1}} \sum_{(\pi,\phi)\in z:z\backslash(\pi,\phi)\in F} d\big(z\backslash(\pi,\phi);(\pi,\phi)\big)\,f(z)\,d\mu_{k+1}(z) \quad \text{[equation (48)]} \\
&= \int_{\Lambda_{k+1}} \sum_{i=1}^{k+1} I\big([\boldsymbol{\pi},\boldsymbol{\phi}]\backslash(\pi_i,\phi_i)\in F\big)\,d\big([\boldsymbol{\pi},\boldsymbol{\phi}]\backslash(\pi_i,\phi_i);(\pi_i,\phi_i)\big)\cdot \\
&\qquad \cdot f([\boldsymbol{\pi},\boldsymbol{\phi}])\,dP_{k+1}\big\{(\boldsymbol{\pi},\boldsymbol{\phi})\big\} \quad \text{[equation (50)]} \\
&= \int_{\Lambda_{k+1}} (k+1)I\big([\boldsymbol{\pi},\boldsymbol{\phi}]\backslash(\pi_{k+1},\phi_{k+1})\in F\big)\,d\big([\boldsymbol{\pi},\boldsymbol{\phi}]\backslash(\pi_{k+1},\phi_{k+1});(\pi_{k+1},\phi_{k+1})\big)\cdot \\
&\qquad \cdot f([\boldsymbol{\pi},\boldsymbol{\phi}])\,dP_{k+1}\big\{(\boldsymbol{\pi},\boldsymbol{\phi})\big\} \quad \text{[by symmetry of } P_{k+1}(\cdot)\text{]} \\
&= \int_{\Lambda_{k+1}} (k+1)I\big([\boldsymbol{\pi}',\boldsymbol{\phi}']\in F\big)\,d\big([\boldsymbol{\pi}',\boldsymbol{\phi}'];(\pi,\phi)\big)f\big([\boldsymbol{\pi}',\boldsymbol{\phi}']\cup(\pi,\phi)\big)\cdot \\
&\qquad \cdot dP_{k+1}\big\{(\boldsymbol{\pi}',\boldsymbol{\phi}')\cup(\pi,\phi)\big\} \quad [(\boldsymbol{\pi}',\boldsymbol{\phi}')\cup(\pi,\phi)=(\boldsymbol{\pi},\boldsymbol{\phi})] \\
&= \int_{\Lambda_k}\int_{[0,1]}\int_\Phi I([\boldsymbol{\pi}',\boldsymbol{\phi}']\in F)(k+1)\,d\big([\boldsymbol{\pi}',\boldsymbol{\phi}'];(\pi,\phi)\big)\,f\big([\boldsymbol{\pi}',\boldsymbol{\phi}']\cup(\pi,\phi)\big)\cdot \\
&\qquad \cdot \frac{r\big(k+1,(\boldsymbol{\pi}',\boldsymbol{\phi}')\cup(\pi,\phi)\big)}{r(k,\boldsymbol{\pi}',\boldsymbol{\phi}')}k(1-\pi)^{k-1}\,d\pi\,\nu(d\phi)\,dP_k\big\{(\boldsymbol{\pi}',\boldsymbol{\phi}')\big\} \quad \text{[Lemma 7.1]} \\
&= \int_{\Omega_k}\int_{[0,1]}\int_\Phi I(y\in F)(k+1)\,d\big(y;(\pi,\phi)\big)\,f\big(y\cup(\pi,\phi)\big)\cdot \\
&\qquad \cdot \frac{r\big(y\cup(\pi,\phi)\big)}{r(y)}k(1-\pi)^{k-1}\,d\pi\,\nu(d\phi)\,d\mu_k(y) \quad \text{[equation (50)]}
\end{aligned}
$$

and so $LHS = RHS$ provided

$$(k+1)d\big(y;(\pi,\phi)\big)f\big(y\cup(\pi,\phi)\big)\frac{r\big(y\cup(\pi,\phi)\big)}{r(y)}k(1-\pi)^{k-1} = \beta(y)b\big(y;(\pi,\phi)\big)f(y)$$

which is equivalent to the conditions (15) stated in the Theorem as $f(y) \propto L(y)$. The remaining detailed balance conditions (46) can be shown to hold in a similar way.

The condition that $r(y)L(y) = 0$ for all y can now be relaxed by applying the conditions (13) and (14), and restricting the spaces Λ_k and Ω_k to $\{y : r(y)L(y) > 0\}$.

□

References

Anderson, E. (1935) The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5.

Baddeley, A. J. and van Lieshout, M. N. M. (1993) Stochastic geometry models in high-level vision. In *Advances in Applied Statistics* (Eds K. V. Mardia and G. K. Kanji), volume 1, pp. 231–256. Abingdon: Carfax Publishing Company.

Brooks, S. P. and Roberts, G. O. (1998) Diagnosing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*. To appear.

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society, series B*, **57**, 473–484.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Cowles, M. K. and Carlin, B. P. (1996) Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Crawford, S. L. (1994) An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, **89**, 259–267.

Dawid, A. P. (1997) Contribution to the discussion of paper by Richardson and Green (1997). *Journal of the Royal Statistical Society, series B*, **59**, 772–773.

DiCiccio, T., Kass, R., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by posterior simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903–915.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, series B*, **56**, 363–375.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

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

Gelman, A. G., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. London: Chapman & Hall.

George, E. I. and McCulloch, R. E. (1996) Stochastic search variable selection. In *Markov Chain Monte Carlo in Practice* (Eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman & Hall.

Geyer, C. J. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Glötzl, E. (1981) Time-reversible and Gibbsian point processes I. Markovian spatial birth and death process on a general phase space. *Mathematische Nachrichten*, **102**, 217–222.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Härdle, W. (1991) *Smoothing techniques with implementation in* S. New York: Springer-Verlag-Verlag.

Jennison, C. (1997) Contribution to the discussion of paper by Richardson and Green (1997). *Journal of the Royal Statistical Society, series B*, **59**, 778–779.

Kelly, F. P. and Ripley, B. D. (1976) A note on Strauss's model for clustering. *Biometrika*, **63**(2), 357–360.

Lawson, A. B. (1996) Markov chain Monte Carlo methods for spatial cluster processes. In *Computer Science and Statistics: Proceedings of the Interface*, volume 27, pp. 314–319.

Mathieson, M. J. (1997) *Ordinal Models and Predictive Methods in Pattern Recognition*. Ph.D. thesis, University of Oxford.

McLachlan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley and Son.

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

Phillips, D. B. and Smith, A. F. M. (1996) Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice* (Eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), chapter 13, pp. 215–239. Chapman & Hall.

Postman, M., Huchra, J. P. and Geller, M. J. (1986) Probes of large-scale structure in the Corona Borealis region. *The Astronomical Journal*, **92**, 1238–1247.

Preston, C. J. (1976) Spatial birth-and-death processes. *Bulletin of the Institute of International Statistics*, **46**, 371–391.

Priebe, C. E. (1994) Adaptive mixtures. *Journal of the American Statistical Association*, **89**, 796–806.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, series B*, **59**, 731–792.

Ripley, B. D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, series B*, **39**, 172–212.

Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.

Robert, C. P. (1994) *The Bayesian Choice: a Decision-Theoretic Motivation*. Springer Texts in Statistics. New York: Springer-Verlag.

Robert, C. P. (1996) Mixtures of distributions: Inference and estimation. In *Markov Chain Monte Carlo in Practice* (Eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman & Hall.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, series B*, **53**, 683–690.

Stephens, D. A. and Fisch, R. D. (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrika*. To appear.

Stephens, M. (1997) *Bayesian Methods for Mixtures of Normal Distributions*. Ph.D. thesis, University of Oxford. Available from `http://www.stats.ox.ac.uk/~stephens`.

Stoyan, D., Kendall, W. S. and Mecke, J. (1987) *Stochastic geometry and its applications*. Wiley & Sons, first edition. 2nd Edition 1995.

Tierney, L. (1996) Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice* (Eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), chapter 4, pp. 59–74. London: Chapman & Hall.

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

Venables, W. N. and Ripley, B. D. (1994) *Modern Applied Statistics with S-Plus*. Springer-Verlag.

Venables, W. N. and Ripley, B. D. (1997) *Modern Applied Statistics with S-Plus*. Springer-Verlag, second edition.

West, M. (1993) Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society, series B*, **55**, 409–422.

Wilson, S. R. (1982) Sound and exploratory data analysis. In *COMPSTAT 1982, Proceedings in Computational Statistics* (Eds H. Caussinus, P. Ettinger and R. Tamassone), pp. 447–450, Vienna. Physica-Verlag.